$$= \begin{bmatrix} 8 & -14 \\ 6 & -13 \end{bmatrix} + \begin{bmatrix} 5 & 8 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 13 & -6 \\ 6 & -13 \end{bmatrix},$$

$$\mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} = \begin{bmatrix} 2 & 4 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} [2]$$

$$= \begin{bmatrix} 2 \\ -6 \end{bmatrix} + \begin{bmatrix} 2 \\ 0 \end{bmatrix} = \begin{bmatrix} 4 \\ -6 \end{bmatrix},$$

$$\mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} = [5, 4] \begin{bmatrix} 0 & 1 \\ 2 & -4 \end{bmatrix} + [6][5, 8]$$

$$= [8, -11] + [30, 48] = [38, 37],$$

and

$$\mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} = [5, 4] \begin{bmatrix} 3 \\ -1 \end{bmatrix} + [6][2] = [23],$$

so

$$\mathbf{AB} = \left[\begin{array}{cc|c} 13 & -6 & 4 \\ 6 & -13 & -6 \\ \hline 38 & 37 & 23 \end{array}\right] = \begin{bmatrix} 13 & -6 & 4 \\ 6 & -13 & -6 \\ 38 & 37 & 23 \end{bmatrix}, \tag{39}$$

which is the same result as obtained by the multiplication of the unpartitioned matrices $\mathbf{A}$ and $\mathbf{B}$.

COMMENT 1. By no means do we claim that partitioning made the preceding calculation easier; our aim was simply to illustrate the idea of partitioning.

COMMENT 2. The partition

$$\mathbf{B} = \left[\begin{array}{c|cc} 0 & 1 & 3 \\ 2 & -4 & -1 \\ \hline 5 & 8 & 2 \end{array}\right] = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}, \tag{40}$$

in place of (38), would also be conformable with (37) for the multiplication $\mathbf{AB}$ and would lead to the same result, (39). On the other hand, neither of the partitions

$$\mathbf{B} = \left[\begin{array}{c|cc} 0 & 1 & 3 \\ \hline 2 & -4 & -1 \\ 5 & 8 & 2 \end{array}\right] \quad \text{or} \quad \mathbf{B} = \left[\begin{array}{cc|c} 0 & 1 & 3 \\ \hline 2 & -4 & -1 \\ 5 & 8 & 2 \end{array}\right] \tag{41}$$

would be conformable with (37) for the product $\mathbf{AB}$. For example, the term $\mathbf{A}_{11}\mathbf{B}_{11}$ would not be defined then since $\mathbf{A}_{11}$ has two columns whereas the $\mathbf{B}_{11}$'s in (41) have only one row. ∎

Let us close by giving two results, for reference, that will be used later on. Both use partitioning to work out the product of two matrices, $\mathbf{A}$ and $\mathbf{B}$. First, if we partition $\mathbf{B}$ into columns $\mathbf{c}_1, \ldots, \mathbf{c}_n$, then

$$\mathbf{AB} = \mathbf{A}\left[\mathbf{c}_1, \ldots, \mathbf{c}_n\right] = \left[\mathbf{Ac}_1, \ldots, \mathbf{Ac}_n\right]. \tag{42}$$

Second, if we also partition $\mathbf{A}$ into rows $\mathbf{r}_1, \ldots, \mathbf{r}_m$, then

$$\mathbf{AB} = \begin{bmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_m \end{bmatrix} [\mathbf{c}_1, \ldots, \mathbf{c}_n] = \begin{bmatrix} \mathbf{r}_1 \cdot \mathbf{c}_1 & \cdots & \mathbf{r}_1 \cdot \mathbf{c}_n \\ \vdots & & \vdots \\ \mathbf{r}_m \cdot \mathbf{c}_1 & \cdots & \mathbf{r}_m \cdot \mathbf{c}_n \end{bmatrix}. \tag{43}$$

That is, the i, j element of $\mathbf{AB}$ is $\mathbf{r}_i$ dotted with $\mathbf{c}_j$.

Closure. We define matrices and three arithmetic operations or matrices: addition, multiplication by a scalar, and multiplication. Subtraction is accounted for by addition and multiplication by a scalar: $\mathbf{A} - \mathbf{B} = \mathbf{A} + (-1)\mathbf{B}$. But no division operation is defined for matrices.

It is emphasized that multiplication is not commutative (i.e., $\mathbf{AB} \neq \mathbf{BA}$ in general). This "failure" and several others are listed in Theorem 10.2.2. It is suggested that these shortcomings of the Cayley definition of matrix multiplication are more than offset by the fact that it permits us to express a system of m linear algebraic equations in the n unknowns $x_1, \ldots, x_n$ in compact matrix form as $\mathbf{Ax} = \mathbf{c}$.

Computer software. As mentioned in Section 8.3, the *Maple* system contains many linear algebra commands within the linalg package, among which **evalm** is especially useful. For instance, to evaluate $(\mathbf{AB})^2\mathbf{P}^3 - 5\mathbf{Q}$, where

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 & -1 \\ 0 & 2 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} 1 & -1 \\ 2 & 3 \end{bmatrix},$$

first enter

with(linalg):

and return. Then enter

$$A := \text{array}\ ([[1, 2], [3, 0]]): \tag{44}$$

and return. (If you wish the $\mathbf{A}$ matrix to be printed, type a semicolon in place of the final colon.) Similarly, for

$$B := \text{array}\ ([[1, -1], [0, 2]]): \tag{45}$$
$$P := \text{array}\ ([[1, 1], [1, 1]]): \tag{46}$$
$$Q := \text{array}\ ([[1, -1], [2, 3]]): \tag{47}$$

Then enter

evalm $((A \,\&*\, B)\hat{}2 \,\&*\, P\hat{}3 - 5 * Q)$;

and return. The printed output is the result

$$\begin{bmatrix} 11 & 21 \\ 38 & 33 \end{bmatrix}$$

Note that matrix multiplication of **A** and **B** is denoted as $A\&*B$ (not $\mathbf{A}*\mathbf{B}$), exponentiation to a positive integer power is denoted with ^, and multiplication of a matrix by a scalar is denoted by $*$. If we want $((\mathbf{AB})^2\mathbf{P}^3 - 5\mathbf{Q})^4$, we can use a quotation mark to carry the matrix $\begin{bmatrix} 11 & 21 \\ 38 & 33 \end{bmatrix}$ forward. Thus, enter

```
evalm(""^4);
```

and return. The result is

$$\begin{bmatrix} 2389489 & 2592744 \\ 4691632 & 5105697 \end{bmatrix}$$

Alternative to the array format indicated above, we can use a matrix format. For instance,

$$A := \text{matrix}\,(2, 2, [1, 2, 3, 0]):$$

is equivalent to the array format shown above, where the "2, 2" denotes that **A** is a 2×2 matrix.

EXERCISES 10.2

1. Given the matrices

$$\mathbf{A} = \begin{bmatrix} 0 & 3 \\ 2 & -5 \\ 1 & 10 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 5 & -1 \\ 0 & 2 \end{bmatrix},$$

$$\mathbf{x} = \begin{bmatrix} 4 \\ 3 \end{bmatrix}, \quad \mathbf{y} = [-1, 2],$$

work out whichever of the products $\mathbf{AB}$, $\mathbf{BA}$, $\mathbf{Ax}$, $\mathbf{xA}$, $\mathbf{Bx}$, $\mathbf{xB}$, $\mathbf{yB}$, $\mathbf{A}^2$, $\mathbf{B}^2$, $\mathbf{x}^2$, $\mathbf{xy}$, and $\mathbf{yx}$ are defined.

2. Let **A** be 6×4, **B** be 4×4, **C** be 4×3, **D** and **E** be 3×1. Determine which of the following are defined, and for those that are, give the form of the resulting matrix.

(a) $\mathbf{A}^{10}$ (b) $\mathbf{B}^{10}$
(c) $\mathbf{ABC}$ (d) $\mathbf{ABCD}$
(e) $\mathbf{ACBD}$ (f) $\mathbf{CD} + \mathbf{E}$
(g) $\mathbf{C}(2\mathbf{D} - \mathbf{E})$ (h) $\mathbf{AB} + \mathbf{AC}$
(i) $\mathbf{BC} + \mathbf{CB}$ (j) $3\mathbf{BA} - 5\mathbf{CD}$

3. Evaluate the products

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \quad \text{and}$$

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

4. Suppose that **A** is $n \times n$, **x** is $n \times 1$, and c is a scalar. Can we re-express $\mathbf{Ax} = c\mathbf{x}$ as $(\mathbf{A} - c)\mathbf{x} = \mathbf{0}$? Explain.

5. If **A** and **B** are square matrices of the same order, are the following correct? Explain.

(a) $(\mathbf{A} + \mathbf{B})^2 = \mathbf{A}^2 + 2\mathbf{AB} + \mathbf{B}^2$
(b) $(\mathbf{A} + \mathbf{B})(\mathbf{A} - \mathbf{B}) = \mathbf{A}^2 - \mathbf{B}^2$
(c) $(\mathbf{AB})^2 = \mathbf{A}^2\mathbf{B}^2$
(d) $(\mathbf{AB})^3 = \mathbf{A}^3\mathbf{B}^3$

6. (a) If p is a positive integer, does **A** need to be square for $\mathbf{A}^p$ to be defined? Explain.
(b) Let **A** be $m \times n$ and **B** be $p \times q$. What restrictions, if any, need must be satisfied by m, n, p, q if $(\mathbf{AB})^2$ is to exist (i.e., be defined)?

7. Expand each of the following [e.g., the "expanded" version of $(\mathbf{A} + \mathbf{B})\mathbf{C}$ would be $\mathbf{AC} + \mathbf{BC}$], assuming that all of the matrices are suitably conformable. Justify each step by citing the relevant equation number in Theorem 10.2.1 or 10.2.3.

(a) $(2\mathbf{A} + \mathbf{B})(\mathbf{A} + 2\mathbf{B})$ (b) $(\mathbf{A} + \mathbf{B})\mathbf{C}(\mathbf{D} + \mathbf{E})$
(c) $(\mathbf{A} + \mathbf{B})^3$ (d) $(\mathbf{A} - 3\mathbf{I})(2\mathbf{A} + \mathbf{I})$

8. Given $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, $\mathbf{C} = \begin{bmatrix} 0 & 3 \\ 0 & 0 \end{bmatrix}$, and $\mathbf{D} = \begin{bmatrix} 0 & 5 & 7 \\ 0 & 0 & 8 \\ 0 & 0 & 0 \end{bmatrix}$, evaluate each of the following.

(a) $\mathbf{A}^{100}$ (b) $\mathbf{B}^{100}$
(c) $\mathbf{C}^{100}$ (d) $\mathbf{D}^{100}$
(e) $(\mathbf{ABC})^3$ (f) $(\mathbf{CBA})^3$
(g) $\mathbf{B}^4\mathbf{C}^4$ and $(\mathbf{BC})^4$ (h) $\mathbf{C}^3\mathbf{B}^3$ and $(\mathbf{CB})^3$

9. Any diagonal matrix whose diagonal elements are all equal is called a **scalar matrix**. If

$$\mathbf{S} = \begin{bmatrix} k & 0 & \cdots & 0 \\ 0 & k & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & k \end{bmatrix}$$

is $n \times n$, and $\mathbf{A}$ is any $n \times n$ matrix, show that

$$\mathbf{AS} = \mathbf{SA} = k\mathbf{A}.$$

What can be said if, instead, $\mathbf{A}$ is $m \times n$ $(m \neq n)$?

10. If for any given vector $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_4 \end{bmatrix}$, the product $\mathbf{Ax}$ is the column vector given below, find $\mathbf{A}$.

(a) $\begin{bmatrix} x_1 - 3x_4 \end{bmatrix}$ (b) $\begin{bmatrix} x_2 + x_3 - x_4 \\ x_1 + 5x_3 \end{bmatrix}$

(c) $\begin{bmatrix} x_1 + x_2 \\ x_2 + x_3 \\ x_3 + x_4 \end{bmatrix}$ (d) $\begin{bmatrix} 2x_1 - x_3 - x_4 \\ -2x_1 + x_2 \\ x_2 + x_4 \end{bmatrix}$

(e) $\begin{bmatrix} x_4 \\ x_3 \\ x_2 \\ x_1 \end{bmatrix}$ (f) $\begin{bmatrix} x_1 + 3x_4 \\ x_2 - x_4 \\ x_3 + x_4 \\ x_4 \\ x_3 - 2x_1 \end{bmatrix}$

11. Make up a specific pair of matrices, $\mathbf{A}$ and $\mathbf{B}$, both nonzero, such that $\mathbf{AB} = \mathbf{0}$, where

(a) $\mathbf{A}$ is 2×2 and $\mathbf{B}$ is 2×2
(b) $\mathbf{A}$ is 5×2 and $\mathbf{B}$ is 2×2
(c) $\mathbf{A}$ is 1×2 and $\mathbf{B}$ is 2×4
(d) $\mathbf{A}$ is 4×3 and $\mathbf{B}$ is 3×2

12. Given the partitioned matrices $\mathbf{A}$ and $\mathbf{B}$, below, carry out the products $\mathbf{A}^2$ and $\mathbf{AB}$ for those cases in which the partitioning is suitable, i.e., conformable. If the partitioning is not suitable, explain why it is not.

(a) $\mathbf{A} = \left[\begin{array}{cc|c} 2 & 0 & -1 \\ 1 & -1 & 0 \\ \hline 5 & 2 & 4 \end{array}\right]$, $\mathbf{B} = \left[\begin{array}{cc|c} 3 & 6 & 2 \\ 0 & 1 & 0 \\ \hline 3 & -4 & 7 \end{array}\right]$

(b) $\mathbf{A} = \left[\begin{array}{cc|c} 2 & 0 & -1 \\ 1 & -1 & 0 \\ \hline 5 & 2 & 4 \end{array}\right]$, $\mathbf{B} = \left[\begin{array}{c|cc} 3 & 6 & 2 \\ 0 & 1 & 0 \\ \hline 3 & -4 & 7 \end{array}\right]$

(c) $\mathbf{A} = \left[\begin{array}{c|cc} 2 & 0 & -1 \\ \hline 1 & -1 & 0 \\ 5 & 2 & 4 \end{array}\right]$, $\mathbf{B} = \left[\begin{array}{cc|c} 3 & 6 & 2 \\ 0 & 1 & 0 \\ \hline 3 & -4 & 7 \end{array}\right]$

13. Show that $c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_n\mathbf{x}_n$ can be expressed in the form

$$[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}.$$

14. (a) If two unpartitioned matrices are not conformable for addition, can they be rendered conformable by suitable partitioning? Explain.
(b) Same as part (a), but for multiplication.

15. If there is some positive integer p such that $\mathbf{A}^p = \mathbf{0}$, then $\mathbf{A}$ is said to be **nilpotent** (mnemonic: "potentially nil"). A square matrix $\mathbf{A} = \{a_{ij}\}$ such that $a_{ij} = 0$ for all $i > j$ is said to be an **upper triangular matrix**; if $a_{ij} = 0$ for all $i < j$, then $\mathbf{A}$ is a **lower triangular matrix**. A matrix is said to be a **triangular** if it is either upper triangular or lower triangular.

(a) Every upper triangular matrix with null main diagonal (so that $a_{ij} = 0$ for all $i \geq j$) is nilpotent. Verify this result for second-, third-, and fourth-order matrices.
(b) In fact, show that *every* upper triangular matrix (of finite order) with null main diagonal is nilpotent. HINT: Use partitioning and induction.
(c) Is every lower diagonal matrix with null main diagonal nilpotent? Explain.
(d) If $\mathbf{A}^p = \mathbf{0}$, show that $(\mathbf{I}+\mathbf{A}+\mathbf{A}^2+\cdots+\mathbf{A}^{p-1})(\mathbf{I}-\mathbf{A}) = (\mathbf{I}-\mathbf{A})(\mathbf{I}+\mathbf{A}+\mathbf{A}^2+\cdots+\mathbf{A}^{p-1}) = \mathbf{I}$.

16. If $\mathbf{A}^2 = \mathbf{I}$, then $\mathbf{A}$ is called **involutory**.

(a) Show, using Theorems 10.2.1 to 10.2.3 and (27), that $\mathbf{A}$ is involutory if and only if

$$(\mathbf{I} - \mathbf{A})(\mathbf{I} + \mathbf{A}) = \mathbf{0}.$$

(b) Give an example of an involutory matrix other than $\mathbf{I}$ and $-\mathbf{I}$. Thus, observe that $\mathbf{A}^2 = \mathbf{I}$ does *not* imply that $\mathbf{A} = \pm\mathbf{I}$.
(c) Determine the most general 2×2 matrix that is involutory.

17. (a) Prove (9a) to (9c).
(b) Prove (9d) to (9f).
(c) Prove (9g) to (9i).

18. In Theorem 10.2.2, prove

(a) (i) (b) (ii) (c) (iii) (d) (iv)

19. (a) Verify (24).
(b) Verify (27).
(c) Verify (30a) and (30b).
(d) Verify (30c) and (30d).
(e) Show that (30e) follows from (30a)–(30d).

20. Show that the most general matrix that commutes with $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ is $\begin{bmatrix} \alpha - \beta & 2\beta/3 \\ \beta & \alpha \end{bmatrix}$, where α and β are arbitrary.

21. Given $\mathbf{A}$, find the most general matrix $\mathbf{B}$ such that $\mathbf{AB} = \mathbf{0}$.

(a) $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ (b) $\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 0 & 0 \end{bmatrix}$

(c) $\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 5 & 0 \end{bmatrix}$ (d) $\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 4 & 6 \end{bmatrix}$

22. Explore and discuss the advantages and disadvantages of defining $c_{ij} = a_{ij}b_{ij}$ in place of the Cayley product (13).

23. (*Transition probability matrix*) Selling their valuables, professors A and B raise \$2 apiece, and proceed to match coins at \$1 per match. There arise five possible states:

S_1	S_2	S_3	S_4	S_5
04	13	22	31	40

In state S_2, for example, A has \$1 and B has \$3. If either player is bankrupted (A is bankrupt in state S_1, B in state S_5), the game is over. Let $p_{ij}^{(n)}$ be the n-step transition probability, i.e., the probability of changing from state S_i to state S_j in n matches. For $n = 1$ we see that

$$\mathbf{P}^{(1)} = \{p_{ij}^{(1)}\} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

For example, beginning in S_2, say, one match will necessarily move us to S_1 or S_3, with 50% probability in each case; thus $p_{21}^{(1)} = p_{23}^{(1)} = \frac{1}{2}$. Precisely \$1 will change hands so that $p_{22}^{(1)} = p_{24}^{(1)} = p_{25}^{(1)} = 0$. Further, once in S_1 (or S_5) we remain there (according to the rules), so $p_{11}^{(1)} = 1$, $p_{12}^{(1)} = p_{13}^{(1)} = p_{14}^{(1)} = p_{15}^{(1)} = 0$. Show, by any convincing arguments or discussion, that

$$p_{ij}^{(2)} = \sum_{k=1}^{5} p_{ik}^{(1)} p_{kj}^{(1)}, \quad p_{ij}^{(3)} = \sum_{k=1}^{5} p_{ik}^{(2)} p_{kj}^{(1)}, \quad \text{etc.}$$

or, in matrix notation, $\mathbf{P}^{(2)} = [\mathbf{P}^{(1)}]^2$, $\mathbf{P}^{(3)} = [\mathbf{P}^{(1)}]^3$, etc. Use this result to determine $\mathbf{P}^{(2)}$ and $\mathbf{P}^{(3)}$. What is the probability that A is bankrupt after (at most) three matches if A starts with \$2? \$3? \$1? NOTE: $\mathbf{P}^{(1)}$ is an example of a **Markov matrix**. We meet Markov matrices again in Chapter 11.

24. Let

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ 3 & 0 \\ 1 & 4 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 5 & 3 & 25 \\ 2 & 0.1 & -6 \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} 9 & 1 & -1 \\ 2 & 0 & 7 \\ 0 & 4 & 6 \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} 6 & 5 \\ 4 & 1 \end{bmatrix}.$$

NOTE: The letters D, E, I, O, S, and W are "protected," in *Maple*, for other purposes. The problem: use computer software to evaluate

(a) $6\mathbf{AB} - 9\mathbf{C}$ (b) $(\mathbf{AB})^3 + 5\mathbf{C}^2$
(c) $6\mathbf{AFB} - 2\mathbf{C}^3$ (d) $4\mathbf{BAF}$
(e) $(\mathbf{BCAF})^4$ (f) $2\mathbf{CA} + 37.3\mathbf{A}$
(g) $(\mathbf{CAB})^2$ (h) $0.73\mathbf{BA} + 1.6\mathbf{F}^6$

10.3 The Transpose Matrix

We continue the development of Section 10.2 by introducing the "transpose" of a matrix. Given any $m \times n$ matrix $\mathbf{A} = \{a_{ij}\}$, we define the **transpose** of $\mathbf{A}$, denoted

as $\mathbf{A}^{\mathrm{T}}$ and read as "$\mathbf{A}$-transpose," as

$$\mathbf{A}^{\mathrm{T}} \equiv \{a_{ji}\} = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}, \tag{1}$$

that is, the $n \times m$ matrix is obtained by interchanging the rows and columns of $\mathbf{A}$: the first row of $\mathbf{A}$ becomes the first column of $\mathbf{A}^{\mathrm{T}}$, the second row of $\mathbf{A}$ becomes the second column of $\mathbf{A}^{\mathrm{T}}$, and so on. Or, the first column of $\mathbf{A}$ becomes the first row of $\mathbf{A}^{\mathrm{T}}$, and so on. That is, if we denote the i, j elements of $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$ as a_{ij} and a_{ij}^{T}, respectively, then

$$a_{ij}^{\mathrm{T}} = a_{ji}. \tag{2}$$

Be clear that $\mathbf{A}^{\mathrm{T}}$ is not $\mathbf{A}$ to the Tth power; it is the transpose of $\mathbf{A}$.

EXAMPLE 1. If

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & 1 \\ -1 & 3 & 5 \\ 4 & 6 & 7 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 2 \\ 6 \\ 7 \end{bmatrix}, \quad \text{and} \quad \mathbf{C} = [1, -8, 9],$$

then

$$\mathbf{A}^{\mathrm{T}} = \begin{bmatrix} 2 & -1 & 4 \\ 0 & 3 & 6 \\ 1 & 5 & 7 \end{bmatrix}, \quad \mathbf{B}^{\mathrm{T}} = [2, 6, 7], \quad \text{and} \quad \mathbf{C}^{\mathrm{T}} = \begin{bmatrix} 1 \\ -8 \\ 9 \end{bmatrix}. \quad \blacksquare$$

THEOREM 10.3.1 *Properties of the Transpose*

$$(\mathbf{A}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{A}, \tag{3a}$$

$$(\mathbf{A} + \mathbf{B})^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}} + \mathbf{B}^{\mathrm{T}}, \tag{3b}$$

$$(\alpha\mathbf{A})^{\mathrm{T}} = \alpha\mathbf{A}^{\mathrm{T}}, \tag{3c}$$

$$(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}, \tag{3d}$$

where it is assumed in (3b) that $\mathbf{A}$ and $\mathbf{B}$ are conformable for addition, and in (3d) that they are conformable for multiplication.

Proof: Proof of (3a)–(3c) is left for the exercises. To prove (3d), let $\mathbf{AB} \equiv \mathbf{C} = \{c_{ij}\}$. By the definition of matrix multiplication,

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}. \tag{4}$$

Thus,

$$c_{ij}^{\mathrm{T}} = c_{ji} = \sum_{k=1}^{n} a_{jk} b_{ki} = \sum_{k=1}^{n} b_{ki} a_{jk} = \sum_{k=1}^{n} b_{ik}^{\mathrm{T}} a_{kj}^{\mathrm{T}}. \tag{5}$$

Having returned, at the end of (5), to the pattern $(\ )_{ij} = \sum(\ )_{ik}(\ )_{kj}$ as in (4), we can conclude from (5) that

$$\mathbf{C}^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} \quad \text{or} \quad (\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}},$$

as was to be proved. Understand that the third equality in (5) is not equivalent to the matrix statement $\mathbf{AB} = \mathbf{BA}$ which, we recall from Section 10.2, is generally untrue. It is simply the scalar statement that $a_{jk}b_{ki} = b_{ki}a_{jk}$, which is true because the multiplication of scalars is commutative [e.g., $(2)(3) = (3)(2) = 6$]. ■

The striking feature of (3d) is the reversal in the order: $(\mathbf{AB})^{\mathrm{T}}$ on the left, $\mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$ on the right. Notice how (3d) checks "dimensionally":

$$\begin{aligned} [(m \times n)(n \times p)]^{\mathrm{T}} &= (n \times p)^{\mathrm{T}}(m \times n)^{\mathrm{T}} \\ (m \times p)^{\mathrm{T}} &= (p \times n)(n \times m) \\ (p \times m) &= (p \times m). \end{aligned} \tag{6}$$

Naturally, (6) does not *prove* (3d), but it provides a useful check, just as we check the *physical* units (such as force, mass, length, and time) of an equation to be sure that they are consistent.

EXAMPLE 2. If $\mathbf{A} = \begin{bmatrix} 4 & 2 & -5 \\ 0 & 1 & 3 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 6 \\ 0 \\ -1 \end{bmatrix}$, say, then

$$\mathbf{AB} = \begin{bmatrix} 4 & 2 & -5 \\ 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} 6 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 29 \\ -3 \end{bmatrix}, \quad \text{so} \quad (\mathbf{AB})^{\mathrm{T}} = [29, -3]$$

and

$$\mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} = [6, 0, -1] \begin{bmatrix} 4 & 0 \\ 2 & 1 \\ -5 & 3 \end{bmatrix} = [29, -3],$$

in agreement with (3d). ■

Furthermore, it follows from (3d) that

$$(\mathbf{ABC})^{\mathrm{T}} = \mathbf{C}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}, \qquad (\mathbf{ABCD})^{\mathrm{T}} = \mathbf{D}^{\mathrm{T}}\mathbf{C}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}, \tag{7}$$

and so on (Exercise 2).

Using lowercase boldface letters for matrices that happen to be column vectors from now on, let

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

Then the standard dot product

$$\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = \sum_{j=1}^{n} x_j y_j$$

can be expressed compactly, in matrix language, as*

$$\boxed{\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^{\mathrm{T}} \mathbf{y}} \tag{8}$$

or, equivalently, as $\mathbf{y}^{\mathrm{T}}\mathbf{x}$, although not as $\mathbf{x}\mathbf{y}^{\mathrm{T}}$ or $\mathbf{y}\mathbf{x}^{\mathrm{T}}$, which expressions represent $n \times n$ matrices!

EXAMPLE 3. If $\mathbf{x} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$, say, then

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^{\mathrm{T}} \mathbf{y} = [3, 1] \begin{bmatrix} 5 \\ 2 \end{bmatrix} = 15 + 2 = 17,$$

or

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{y}^{\mathrm{T}} \mathbf{x} = [5, 2] \begin{bmatrix} 3 \\ 1 \end{bmatrix} = 15 + 2 = 17,$$

whereas

$$\mathbf{x}\mathbf{y}^{\mathrm{T}} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} [5, 2] = \begin{bmatrix} 15 & 6 \\ 5 & 2 \end{bmatrix} \neq \mathbf{x} \cdot \mathbf{y},$$

and

$$\mathbf{y}\mathbf{x}^{\mathrm{T}} = \begin{bmatrix} 5 \\ 2 \end{bmatrix} [3, 1] = \begin{bmatrix} 15 & 5 \\ 6 & 2 \end{bmatrix} \neq \mathbf{x} \cdot \mathbf{y}. \quad \blacksquare$$

Finally, two more definitions: if

$$\mathbf{A}^{\mathrm{T}} = \mathbf{A} \tag{9}$$

we say that $\mathbf{A}$ is **symmetric**, and if

$$\mathbf{A}^{\mathrm{T}} = -\mathbf{A}, \tag{10}$$

*There is a small difficulty here: $\mathbf{x} \cdot \mathbf{y}$ is to be scalar, whereas $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ is a 1×1 matrix. Thus, what we really intend, by $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ in (8), is not the 1×1 matrix, but rather the scalar element inside it. That could be noted by writing $\mathbf{x} \cdot \mathbf{y} = (\mathbf{x}^{\mathrm{T}}\mathbf{y})_{11}$ instead, but it will be simpler to leave (8) intact, with the *understanding* that the right-hand side is a scalar.

we say that it is **skew-symmetric** (or antisymmetric). For either of these properties to apply $\mathbf{A}$ must be square, since otherwise $\mathbf{A}^{\mathrm{T}}$ and $\mathbf{A}$ would be of different form. And for $\mathbf{A}$ to be skew-symmetric all of its diagonal elements must be zero, since (10) implies that $a_{ji} = -a_{ij}$, or if we set $i = j$, $a_{ii} = -a_{ii}$; thus $2a_{ii} = 0$, so $a_{ii} = 0$ for each i.

It would be reasonable to imagine that the likelihood of encountering purely symmetric or skew-symmetric matrices in applications would be slim. On the contrary, we shall see that symmetric matrices arise frequently, and that their symmetry is often a consequence of fundamental physical principles, rather than chance.

Closure. The key points in this section are the defining of the transpose of any matrix, and the results $(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$ and $\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^{\mathrm{T}}\mathbf{y}$ (or $\mathbf{y}^{\mathrm{T}}\mathbf{x}$). Note that the transpose notation is sometimes used to save space. For instance, it takes less vertical space on the page to write $\mathbf{x}^{\mathrm{T}} = [7, 2]$ than $\mathbf{x} = \begin{bmatrix} 7 \\ 2 \end{bmatrix}$.

Computer software. The relevant *Maple* function, to take the transpose of a matrix $\mathbf{A}$, is the command **transpose**, within the linalg package. For instance, to obtain the transpose of $\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$ enter

with(linalg):

to access the transpose(A) command. Enter

$A :=$ array([[1, 2, 3], [4, 5, 6]]);

and return, then enter

transpose(A);

and return. The output is

$$\begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}$$

EXERCISES 10.3

1. (a) If $\mathbf{x} = [3, -3]^{\mathrm{T}}$ and $\mathbf{y} = [1, 2]^{\mathrm{T}}$, work out $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ and $\mathbf{x}\mathbf{y}^{\mathrm{T}}$.
(b) If $\mathbf{x} = [4, -1, 0]^{\mathrm{T}}$ and $\mathbf{y} = [1, 2, 3]^{\mathrm{T}}$, work out $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ and $\mathbf{x}\mathbf{y}^{\mathrm{T}}$.
(c) If $\mathbf{x} = [0, 4, -2, 1]^{\mathrm{T}}$ and $\mathbf{y} = [3, 0, 1, -2]^{\mathrm{T}}$, work out $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ and $\mathbf{x}\mathbf{y}^{\mathrm{T}}$.

2. Show that (7) follows from (3d).

3. Recall that in general $\mathbf{AB} \neq \mathbf{BA}$, and that a *necessary* (but not sufficient) condition for equality to hold is that both $\mathbf{A}$ and $\mathbf{B}$ be square and of the same order. Perhaps a *sufficient* condition is that $\mathbf{A}$ and $\mathbf{B}$ both be of the same order and *symmetric*. Prove or disprove this hypothesis.

4. Verify $(\mathbf{ABC})^{\mathrm{T}} = \mathbf{C}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$ directly, for

(a) $\mathbf{A} = \begin{bmatrix} 5 & -2 \\ 0 & 1 \\ 1 & 3 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 0 \\ 7 \end{bmatrix}$, $\mathbf{C} = [3, 1, 2, 9]$

(b) $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 5 & 4 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}, \mathbf{C} = \begin{bmatrix} 4 & -1 & 5 \\ 0 & 3 & 0 \end{bmatrix}$

(c) $\mathbf{A} = [5, 3, 0], \mathbf{B} = \begin{bmatrix} -2 & 1 \\ 6 & 4 \\ 5 & 3 \end{bmatrix}, \mathbf{C} = \begin{bmatrix} 4 \\ 8 \end{bmatrix}$

(d) $\mathbf{A} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 1 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 3 & 4 \\ 1 & 2 \\ 0 & 1 \end{bmatrix}, \mathbf{C} = \begin{bmatrix} 6 \\ 0 \end{bmatrix}$

5. Prove the properties (3a), (3b), and (3c).

6. Even if a (square) matrix is neither symmetric nor skew-symmetric it can be decomposed as the sum of a symmetric matrix and a skew-symmetric matrix. Specifically, writing

$$\begin{aligned} \mathbf{A} &= \tfrac{1}{2}(\mathbf{A} + \mathbf{A}^{\mathrm{T}}) + \tfrac{1}{2}(\mathbf{A} - \mathbf{A}^{\mathrm{T}}) \\ &\equiv \quad \mathbf{A}_1 \quad + \quad \mathbf{A}_2, \end{aligned} \tag{6.1}$$

show that $\mathbf{A}_1$ is symmetric and $\mathbf{A}_2$ is skew-symmetric. NOTE: Equation (6.1) is but one such decomposition. For instance, we saw in Chapter 9 that a vector $\mathbf{v}$ in 3-space can be decomposed into the sum of two vectors, one lying in a given plane (the "projection of $\mathbf{v}$" onto that plane) and the other perpendicular to that plane; when we study vector fields we will see that any vector field can be decomposed as the sum of two fields, one irrotational and the other solenoidal; when we study Fourier series we will see that any function $f(x)$ can be decomposed as the sum of two functions, one even and the other odd; and so on. Thus, such decompositions are not uncommon in mathematics.

7. Decompose the given matrix as the sum of two matrices, one symmetric and one skew symmetric, as explained in Exercise 6.

(a) $\begin{bmatrix} 3 & 2 \\ 1 & -5 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 0 \\ 2 & 0 \end{bmatrix}$ (d) $\begin{bmatrix} 8 & -2 \\ 4 & 0 \end{bmatrix}$

(e) $\begin{bmatrix} 9 & 8 & 7 \\ 6 & 5 & 4 \\ 3 & 2 & 1 \end{bmatrix}$ (f) $\begin{bmatrix} 1 & 0 & 1 \\ 2 & -1 & 0 \\ 3 & 0 & 6 \end{bmatrix}$

8. (*Quadratic forms*) The quadratic function $ax^2 + by^2 + cxy$ is said to be a **quadratic form** in x and y, $ax^2 + by^2 + cz^2 + dxy + exz + fyz$ is a quadratic form in x, y, z, and so on. Every quadratic form in the n variables $x_1, \ldots, x_n$ can be expressed, in matrix notation, as $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a symmetric $n \times n$ matrix. For each given quadratic form, determine the $\mathbf{A}$ matrix. NOTE: Quadratic forms will be important to us in Chapter 11.

(a) $6x_1^2 + x_2^2 - 8x_1x_2$ in x_1, x_2
(b) $x_1^2 - 3x_2^2 + 6x_1x_2$ in x_1, x_2
(c) $4x_1^2 + x_2^2 - x_3^2 + 8x_1x_2 + 3x_1x_3 - 2x_2x_3$ in x_1, x_2, x_3
(d) $x_1^2 - 4x_3^2 + 2x_1x_3 - 10x_2x_3$ in x_1, x_2, x_3
(e) $3x_2^2 + x_3^2 - 6x_1x_3$ in x_1, x_2, x_3
(f) $x_1^2 + x_2^2 + x_4^2 + x_1x_3 + x_1x_4 + x_2x_3$ in x_1, x_2, x_3, x_4

9. Show that if $\mathbf{A}$ is an $m \times n$ matrix, then $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ is symmetric. HINT: There is often an inclination to work out a problem like this using "brute force," i.e., by actually writing out the $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$ matrices, multiplying them, and examining the resulting matrix to see if it is symmetric. Whenever possible, we advise against such an approach. In this problem, for example, we wish to show that $\mathbf{C}^{\mathrm{T}} = \mathbf{C}$, where $\mathbf{C}$ is short for $\mathbf{A}\mathbf{A}^{\mathrm{T}}$; i.e., we wish to show that $(\mathbf{A}\mathbf{A}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$, and this can be done (in one or two short lines) using the properties stated in Theorem 10.3.1.

10. Prove that *the product* $\mathbf{AB}$ *need not be symmetric, even if* $\mathbf{A}$ *and* $\mathbf{B}$ *are both symmetric and of the same order.*

11. Let

$$\mathbf{A} = \begin{bmatrix} 4 & 1 & 2 \\ 0 & 5 & 7 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 & -4 & 2 \\ 8 & 1 & 4 \end{bmatrix}.$$

Use computer software, such as *Maple*, to evaluate

(a) $\mathbf{AB}^{\mathrm{T}}$ (b) $\mathbf{BA}^{\mathrm{T}}$
(c) $(\mathbf{AB}^{\mathrm{T}})^5$ (d) $(\mathbf{BA}^{\mathrm{T}})^{\mathrm{T}}$
(e) $(\mathbf{B}^{\mathrm{T}}\mathbf{A})^8$ (f) $2\mathbf{A}^{\mathrm{T}} - 7.3\mathbf{B}^{\mathrm{T}}$

10.4 Determinants

In this section we introduce a scalar quantity associated with every square matrix, the so-called "determinant" of the matrix. We denote the **determinant** of an $n \times n$

matrix $\mathbf{A} = \{a_{ij}\}$ as

$$\det\mathbf{A} = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}, \tag{1}$$

that is, with straight line braces instead of square brackets. Determinants are prominent in Chapter 3 in connection with the linear dependence or independence of sets of functions, especially with sets of solutions of linear homogeneous differential equations. More generally, they play a key role in the theory of systems of linear algebraic equations, as discussed in Section 10.5.

The determinant of an $n \times n$ matrix $\mathbf{A} = \{a_{ij}\}$ is defined by the **cofactor expansion**

$$\boxed{\det\mathbf{A} \equiv \sum_{1}^{n} a_{jk}A_{jk},} \tag{2}$$

where *the summation is carried out on j for any fixed value of k $(1 \le k \le n)$ or on k for any fixed value of j $(1 \le j \le n)$.* A_{jk} is called the **cofactor** of the a_{jk} element and is defined as

$$A_{jk} \equiv (-1)^{j+k}M_{jk}, \tag{3}$$

where M_{jk} is called the **minor** of a_{jk}, namely, the determinant of the $(n-1) \times (n-1)$ matrix that survives when the row and column containing a_{jk} (the jth row and the kth column) are struck out.

For example, if

$$\mathbf{A} = \begin{bmatrix} 4 & 7 & -2 \\ 0 & 3 & 2 \\ 1 & 5 & 6 \end{bmatrix},$$

then

$$M_{11} = \begin{vmatrix} 3 & 2 \\ 5 & 6 \end{vmatrix}, \quad M_{12} = \begin{vmatrix} 0 & 2 \\ 1 & 6 \end{vmatrix}, \quad \text{and} \quad M_{23} = \begin{vmatrix} 4 & 7 \\ 1 & 5 \end{vmatrix}.$$

Thus, if $\mathbf{A}$ is $n \times n$, then the right-hand side of (2) is a linear combination of n determinants, each of which is $(n-1) \times (n-1)$. Each of these, in turn, may be expressed as a linear combination of $(n-2) \times (n-2)$ determinants, and so on, until we have a (perhaps large) number of 1×1 determinants. Thus, the definition (2) is logically incomplete until we define a 1×1 determinant, which we do as follows:

$$\det\begin{bmatrix} a_{11} \end{bmatrix} = \begin{vmatrix} a_{11} \end{vmatrix} \equiv a_{11}. \tag{4}$$

That is, the determinant of a 1×1 matrix $\begin{bmatrix} a_{11} \end{bmatrix}$ is simply a_{11} itself. CAUTION: In the present context, the braces around a_{11}, in the middle member of (4), denote determinant, not absolute value. For instance, $\det[-6] = |-6| = -6$.

Using (2) and (4), let us work out the determinant of any 2×2 matrix. Recalling that we can sum on j with k fixed, or vice versa, let us sum on k with $j = 1$, say. Then

$$\begin{aligned}\det\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22}\end{bmatrix} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22}\end{vmatrix} &= \sum_{k=1}^{2} a_{1k}A_{1k} \\ &= a_{11}A_{11} + a_{12}A_{12} = a_{11}(-1)^{1+1}M_{11} + a_{12}(-1)^{1+2}M_{12} \\ &= a_{11}(+1)\,|a_{22}| + a_{12}(-1)\,|a_{21}| \\ &= a_{11}a_{22} - a_{12}a_{21}, \end{aligned} \tag{5}$$

which result is probably familiar to you from earlier studies. Observe that the $(-1)^{j+k}$ in (3) is simply $+1$ if $j + k$ is even, and -1 if $j + k$ is odd.

EXAMPLE 1. Let

$$\mathbf{A} = \begin{bmatrix} 0 & 2 & -1 \\ 4 & 3 & 5 \\ 2 & 0 & -4 \end{bmatrix}. \tag{6}$$

Using (2) with $j = 1$, say,

$$\begin{aligned}\det\mathbf{A} = \sum_{k=1}^{3} a_{1k}A_{1k} &= a_{11}A_{11} + a_{12}A_{12} + a_{13}A_{13} \\ &= a_{11}(+1)M_{11} + a_{12}(-1)M_{12} + a_{13}(+1)M_{13} \\ &= a_{11}M_{11} - a_{12}M_{12} + a_{13}M_{13} \\ &= (0)\begin{vmatrix} 3 & 5 \\ 0 & -4 \end{vmatrix} - (2)\begin{vmatrix} 4 & 5 \\ 2 & -4 \end{vmatrix} + (-1)\begin{vmatrix} 4 & 3 \\ 2 & 0 \end{vmatrix} \\ &= 0 - (2)(-16 - 10) + (-1)(0 - 6) = 58. \end{aligned} \tag{7}$$

This particular choice ($j = 1$, summing on k) is said to be *cofactor expansion about the first row* since (7) is the sum of the first-row elements a_{11}, a_{12}, a_{13}, multiplied by their cofactors.

According to (2), we can expand about *any* row or column. Let us illustrate that the same answer is indeed obtained if we expand about other rows or columns.

Expansion about second row [i.e., set $j = 2$ in (2), and sum on k]:

$$\begin{aligned}\det\mathbf{A} = \sum_{k=1}^{3} a_{2k}A_{2k} &= a_{21}A_{21} + a_{22}A_{22} + a_{23}A_{23} \\ &= a_{21}(-1)M_{21} + a_{22}(+1)M_{22} + a_{23}(-1)M_{23} \\ &= -(4)\begin{vmatrix} 2 & -1 \\ 0 & -4 \end{vmatrix} + (3)\begin{vmatrix} 0 & -1 \\ 2 & -4 \end{vmatrix} - (5)\begin{vmatrix} 0 & 2 \\ 2 & 0 \end{vmatrix} \\ &= -(4)(-8) + (3)(2) - (5)(-4) = 58, \end{aligned}$$

again.

Expansion about third column ($k = 3$, sum on j):

$$\begin{aligned}\det\mathbf{A} = \sum_{j=1}^{3} a_{j3}A_{j3} &= a_{13}A_{13} + a_{23}A_{23} + a_{33}A_{33} \\ &= a_{13}(+1)M_{13} + a_{23}(-1)M_{23} + a_{33}(+1)M_{33} \\ &= (-1)\begin{vmatrix} 4 & 3 \\ 2 & 0 \end{vmatrix} - (5)\begin{vmatrix} 0 & 2 \\ 2 & 0 \end{vmatrix} + (-4)\begin{vmatrix} 0 & 2 \\ 4 & 3 \end{vmatrix} \\ &= (-1)(-6) - (5)(-4) + (-4)(-8) = 58,\end{aligned}$$

once more. ∎

Since we may expand about any row or column, it is convenient in hand calculations to choose the row or column with the most zeros in it since those terms in the expansion then drop out.

Notice carefully that for large n the cofactor expansion process is exceedingly laborious. Even if $n = 10$, say, which is still quite modest, (2) gives a linear combination of ten 9×9 determinants. In turn, each of these ten is evaluated as a linear combination of nine 8×8 determinants, and, so on! Let us see just how serious this predicament is. For estimating purposes, let us count each multiplication, addition, and subtraction as one "calculation." It can be shown (Exercise 18a) that the number of calculations $N(n)$ required in the evaluation (by cofactor expansion) of an $n \times n$ determinant is

$$N(n) \sim e\,n! \tag{8}$$

as $n \to \infty$, where $e \approx 2.718$ is the base of the natural logarithm, and $n!$ is n factorial. If each calculation takes approximately one microsecond, then some time estimates are as follows. (Before reading on, we urge you to guess how long such a computer would take to evaluate a 25×25 determinant.)

n	Computing Time
5	0.0003 sec
10	10 sec
15	4×10^6 sec $\approx$ 40 days
20	7×10^{12} sec $\approx$ 210,000 years
25	4×10^{19} sec $\approx 10^{12}$ years

It is interesting that faster computers offer no hope. For instance, even a computer that is a million times as fast would still take around 10^6 years to evaluate a 25×25 determinant. And scientific calculations can easily involve determinants that are 250×250.

It is tempting to conclude that "determinants are worthless," but let us see if we can come up with a more efficient algorithm than the cofactor expansion. A logical starting point is to first determine the various properties of determinants so that we can use them to design a better algorithm.

First, we need to introduce the idea of a "triangular" matrix. A square matrix $\mathbf{A} = \{a_{ij}\}$ is **upper triangular** if $a_{ij} = 0$ for all $j < i$ and **lower triangular** if $a_{ij} = 0$ for all $j > i$. That is,

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & a_{nn} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & & \\ \vdots & & \ddots & \vdots \\ a_{n1} & & \cdots & a_{nn} \end{bmatrix} \tag{9}$$

are upper triangular and lower triangular, respectively. If a matrix is upper triangular or lower triangular it is said to be **triangular**.

Here are the properties that we will need.

PROPERTIES OF DETERMINANTS

D1. If any row (or column) of $\mathbf{A}$ is modified by adding α times the corresponding elements of another row (or column) to it, yielding a new matrix $\mathbf{B}$, then $\det\mathbf{B} = \det\mathbf{A}$.
Symbolically: $\mathbf{r}_j \to \mathbf{r}_j + \alpha\mathbf{r}_k$

D2. If any two rows (or columns) of $\mathbf{A}$ are interchanged, yielding a new matrix $\mathbf{B}$, then $\det\mathbf{B} = -\det\mathbf{A}$.
Symbolically: $\mathbf{r}_j \leftrightarrow \mathbf{r}_k$

D3. If $\mathbf{A}$ is triangular, then $\det\mathbf{A}$ is simply the product of the diagonal elements, $\det\mathbf{A} = a_{11}a_{22}\cdots a_{nn}$.

Of these, D3 is easily proved. For consider the general upper triangular matrix in (9). Doing a cofactor expansion about the first column gives a_{11} times an $(n-1)\times(n-1)$ minor determinant, which is again upper triangular. Expanding the latter about its first column gives a_{22} times an $(n-2)\times(n-2)$ minor determinant, which is again upper triangular. Repeating the process leads to $\det\mathbf{A} = a_{11}a_{22}\cdots a_{nn}$. Similarly for the general lower triangular matrix in (9), except that in that case we expand about the first row, repeatedly, rather than the first column.

Let us illustrate the use of the properties D1–D3 instead of the cofactor expansion.

EXAMPLE 2. Consider the $\mathbf{A}$ matrix of Example 1 again.

$$\det\mathbf{A} = \begin{vmatrix} 0 & 2 & -1 \\ 4 & 3 & 5 \\ 2 & 0 & -4 \end{vmatrix} = -\begin{vmatrix} 2 & 0 & -4 \\ 4 & 3 & 5 \\ 0 & 2 & -1 \end{vmatrix}$$

$$= -\begin{vmatrix} 2 & 0 & -4 \\ 0 & 3 & 13 \\ 0 & 2 & -1 \end{vmatrix} = -\begin{vmatrix} 2 & 0 & -4 \\ 0 & 3 & 13 \\ 0 & 0 & -\frac{29}{3} \end{vmatrix} = -(2)(3)\left(-\frac{29}{3}\right) = 58,$$

as obtained in Example 1. In the second equality we interchanged the first and third rows ($\mathbf{r}_1 \leftrightarrow \mathbf{r}_2$), thereby changing the sign of the determinant (D2) so we compensated by

putting the minus sign out in front. In the third equality we modified the second row by adding -2 times the first row to it ($\mathbf{r}_2 \to \mathbf{r}_2 - 2\mathbf{r}_1$), which step left the determinant unchanged (D1). In the fourth equality we modified the third row by adding $-\frac{2}{3}$ times the second row to it ($\mathbf{r}_3 \to \mathbf{r}_3 - \frac{2}{3}\mathbf{r}_2$), which step left the determinant unchanged (D1). Since those steps produced a triangular matrix, we could then use D3. ∎

The point, then, is to use some combination of D1 and D2 steps to reduce the determinant to triangular form, in which case it is evaluated easily by D3. Of course the method is quite similar to Gauss elimination, described in Section 8.3. For instance, compare D1 and D2 with the first and third elementary equation operations listed in Section 8.3. For reference purposes, we will call the method illustrated in Example 2 the method of **triangularization**.

It is hard to tell, from the 3×3 calculation in Example 2, whether the method is more efficient than the cofactor expansion. However, in Exercise 18b it is shown that using triangularization the number of calculations $N(n)$ is

$$N(n) \sim \frac{2n^3}{3} \tag{10}$$

as $n \to \infty$. Again assuming one microsecond per calculation, (10) gives a computing time of around 0.005 second for $n = 20$ and 0.01 second for $n = 25$, compared with 210,000 years and 10^{12} years, respectively! [Comparing (8) and (10), we can see how much faster $n!$ grows than n^3.]

The upshot is that except for small hand calculations we should avoid the cofactor expansion, and should use triangularization instead.

Although properties D1 – D3 suffice for the efficient calculation of determinants, other properties are sometimes useful as well, and are listed below.

ADDITIONAL PROPERTIES OF DETERMINANTS

D4. If all the elements of any row or column are zero, then $\det\mathbf{A} = 0$.

D5. If any two rows or columns are proportional to each other, then $\det\mathbf{A} = 0$.

D6. If any row (column) is a linear combination of other rows (columns), then $\det\mathbf{A} = 0$.

D7. If all the elements of any row or column are scaled by α, yielding a new matrix $\mathbf{B}$, then $\det\mathbf{B} = \alpha \det\mathbf{A}$.

D8. $\det(\alpha\mathbf{A}) = \alpha^n \det\mathbf{A}$.

D9. If any one row (or column) $\mathbf{a}$ of $\mathbf{A}$ is separated as $\mathbf{a} = \mathbf{b} + \mathbf{c}$, then

$$\det\mathbf{A}|_{\mathbf{a}} = \det\mathbf{A}|_{\mathbf{b}} + \det\mathbf{A}|_{\mathbf{c}},$$

where $\mathbf{A}|_{\mathbf{a}}$ denotes the $\mathbf{A}$ matrix with $\mathbf{a}$ intact, $\mathbf{A}|_{\mathbf{b}}$ denotes the $\mathbf{A}$ matrix with $\mathbf{b}$ in place of $\mathbf{a}$, and similarly for $\mathbf{A}|_{\mathbf{c}}$. For example,

$$\begin{vmatrix} 6+2 & -3+1 & 5+4 \\ 3 & 0 & 2 \\ 1 & -6 & 7 \end{vmatrix} = \begin{vmatrix} 6 & -3 & 5 \\ 3 & 0 & 2 \\ 1 & -6 & 7 \end{vmatrix} + \begin{vmatrix} 2 & 1 & 4 \\ 3 & 0 & 2 \\ 1 & -6 & 7 \end{vmatrix}.$$

D10. The determinant of $\mathbf{A}$ and its transpose are equal,

$$\det(\mathbf{A}^{\mathrm{T}}) = \det\mathbf{A}.$$

D11. In general,

$$\det(\mathbf{A} + \mathbf{B}) \neq \det\mathbf{A} + \det\mathbf{B}.$$

D12. The determinant of a product equals the product of their determinants,

$$\boxed{\det(\mathbf{AB}) = (\det\mathbf{A})(\det\mathbf{B}).} \tag{11}$$

These properties are not independent of each other. For example, D5 follows from D1 and D4, and D4 follows from D6. Keep in mind that det( ) is *not linear*. That is, if α, β are scalars and $\mathbf{A}$, $\mathbf{B}$ are $n \times n$ matrices, then

$$\boxed{\det(\alpha\mathbf{A} + \beta\mathbf{B}) \neq \alpha\det\mathbf{A} + \beta\det\mathbf{B},} \tag{12}$$

in general. For instance, if $\beta = 0$ is $\det(\alpha\mathbf{A}) = \alpha\det\mathbf{A}$? No, according to D7 it is $\alpha^n\det\mathbf{A}$. Or, with $\alpha = \beta = 1$, is $\det(\mathbf{A} + \mathbf{B}) = \det\mathbf{A} + \det\mathbf{B}$? Not in general, according to D11. This result may come as a surprise since we are studying "linear algebra." Also surprising is the truth of D12, if we contrast the complexity of the matrix multiplication $\mathbf{AB}$ on the left with the simplicity of the outcome, expressed on the right-hand side. This result was proved by Cauchy in 1815.*

Closure. Every $n \times n$ matrix $\mathbf{A}$ has a value associated with it called its determinant and denoted as $\det\mathbf{A}$. Though $\det\mathbf{A}$ is defined, traditionally, by the cofactor expansion (2), we find that that formula is useless, computationally, unless n is quite small. Thus, we study various properties of the determinant and put forward a computational algorithm called triangularization, based upon properties D1–D3, that is incredibly efficient compared to the cofactor expansion.

Computer software. Using *Maple*, one can evaluate determinants using the **det(A)** command. For instance, to evaluate the determinant of the matrix $\mathbf{A}$ given by (6), enter

```
with(linalg):
```

to access the det(A) command. Then enter

```
det([[0, 2, -1], [4, 3, 5], [2, 0, -4]]);
```

and return. The output is 58. Alternatively, the sequence

```
with(linalg):
A := array([[0, 2, -1], [4, 3, 5], [2, 0, -4]]) :
det(A);
```

**Augustin–Louis Cauchy* (1789–1857) is among the great mathematicians. Unlike his contemporary, Gauss who published little of his work, Cauchy published more than 700 papers. Among the subjects on which he worked were determinants, ordinary and partial differential equations, complex variable theory, and the wave theory of light.

gives the same result.

EXERCISES 10.4

1. In (5) we evaluated the determinant of a general 2×2 matrix using a cofactor expansion about the first row. Evaluate it again, using a cofactor expansion about the second row instead, then about the first column, and then about the second column, showing that the answer is the same in each case.

2. Evaluate each, using a cofactor expansion about the first and last rows, and also about the last column.

(a) $\begin{vmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \\ 1 & 1 & 1 \end{vmatrix}$ (b) $\begin{vmatrix} 2 & -3 & 0 \\ 1 & 4 & 2 \\ -6 & 1 & 5 \end{vmatrix}$

(c) $\begin{vmatrix} -4 & 1 & 0 \\ 3 & 2 & 0 \\ 1 & 5 & 7 \end{vmatrix}$ (d) $\begin{vmatrix} 3 & 3 & 12 \\ 0 & 6 & -1 \\ 4 & 0 & 0 \end{vmatrix}$

(e) $\begin{vmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \end{vmatrix}$ (f) $\begin{vmatrix} -5 & 2 & 1 & 0 \\ 4 & 0 & 3 & 0 \\ 2 & 4 & 7 & 1 \\ 0 & 0 & -2 & 6 \end{vmatrix}$

(g) $\begin{vmatrix} 2 & 0 & 1 & 0 \\ 0 & 3 & 1 & -1 \\ 0 & 4 & 5 & 0 \\ 1 & 2 & 3 & 6 \end{vmatrix}$ (h) $\begin{vmatrix} 0 & 1 & 2 & 0 \\ 3 & -1 & 1 & 4 \\ 5 & 6 & -7 & 1 \\ 0 & 2 & 1 & 0 \end{vmatrix}$

(i) $\begin{vmatrix} a & 0 & 0 & 0 \\ 0 & b & c & 0 \\ 0 & d & e & 0 \\ 0 & 0 & 0 & f \end{vmatrix}$ (j) $\begin{vmatrix} a & b & c & 0 \\ d & e & f & 0 \\ g & h & i & 0 \\ 0 & 0 & 0 & k \end{vmatrix}$

3. (a)–(j) Same as Exercise 2, but expanding about the second row, and about the first column.

4. (a)–(h) Same as Exercise 2, but using the method of triangularization.

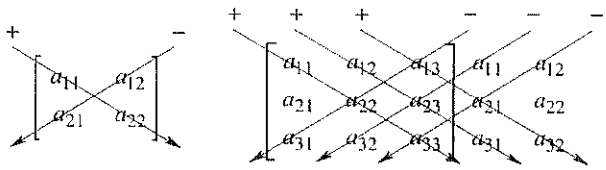
5. (a)–(h) Same as Exercise 2, but using computer software.

6. Evaluate, by any means other than computer software, showing your steps or logic. You may use any of the properties D1–D12.

(a) $\begin{vmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \\ 0 & 1 & -3 & 5 \end{vmatrix}$ (b) $\begin{vmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 10 & 11 & 12 \\ 13 & 14 & 15 & 0 \end{vmatrix}$

(c) $\begin{vmatrix} 0 & 0 & a \\ 0 & b & c \\ d & e & f \end{vmatrix}$ (d) $\begin{vmatrix} a & b & c \\ d & e & 0 \\ f & 0 & 0 \end{vmatrix}$

7. A mnemonic device often put forward for evaluating 2×2 and 3×3 determinants is as shown below.



In other words, the determinants are the sums of the indicated products, with each product carrying the indicated sign. For example, in the 2×2 case this device gives

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = +(a_{11}a_{22}) - (a_{21}a_{12}),$$

which does agree with (5). We now state the problem: write out the mnemonic result for the 3×3 case, and verify (by cofactor expansion) that it is correct. CAUTION: This device does *not* hold, in general, for $n \times n$ determinants if $n \geq 4$.

8. Let an $n \times n$ matrix $\mathbf{A} = \{a_{ij}\}$ be diagonal. Show that

$$\det\mathbf{A} = a_{11}a_{22}\cdots a_{nn}. \tag{8.1}$$

9. (a) Suppose that an $n \times n$ matrix $\mathbf{A}$ can be partitioned into the **block-diagonal** form

$$\mathbf{A} = \begin{bmatrix} \boxed{\mathbf{A}_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boxed{\mathbf{A}_2} & \cdots & \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & & \cdots & \boxed{\mathbf{A}_m} \end{bmatrix},$$

where $\mathbf{A}_1, \ldots, \mathbf{A}_m$ are all square, although not necessarily all of the same order. Show that

$$\det\mathbf{A} = (\det\mathbf{A}_1)(\det\mathbf{A}_2)\cdots(\det\mathbf{A}_m). \tag{9.1}$$

This result may be regarded as a generalization of (8.1), above, wherein $\mathbf{A}_1, \ldots, \mathbf{A}_m$ were all 1×1's.
(b) Does (9.1) still hold if the elements above the m blocks are

nonzero? Explain.
(c) Does (9.1) still hold if the elements below the m blocks are nonzero? Explain.
(d) To which determinants, in Exercise 2, can (9.1) be applied? In each of those cases use (9.1) to evaluate det$\mathbf{A}$.

10. Deduce, from property D12, that if $\mathbf{A}_1, \ldots, \mathbf{A}_k$ are $n \times n$ matrices, then

$$\det(\mathbf{A}_1\mathbf{A}_2 \cdots \mathbf{A}_k) = (\det\mathbf{A}_1)(\det\mathbf{A}_2)\cdots(\det\mathbf{A}_k). \quad (10.1)$$

11. (a) Derive the property D7. HINT: Write out the cofactor expansion about the row or column in question.
(b) Then, show that D8 follows from D7.

12. Prove property D6, using any of the other listed properties.

13. Prove property D9. HINT: Write out the cofactor expansion about the row (or column) $\mathbf{a}$.

14. (*Routh–Hurwitz criterion*) First, review Section 3.4.5, on stability. According to the **Routh–Hurwitz criterion**, necessary and sufficient conditions for the stability of the system governed by equation (65) in Section 3.4.5 (i.e., for all the roots of its characteristic equation to have negative real parts) are that $a_j > 0$ for each $j = 1, \ldots, n$, and that $\Delta_j > 0$ for each $j = 1, \ldots, n$, where

$$\Delta_j = \begin{vmatrix} a_1 & 1 & 0 & 0 & 0 & \cdots & 0 \\ a_3 & a_2 & a_1 & 1 & 0 & \cdots & 0 \\ a_5 & a_4 & a_3 & a_2 & a_1 & \cdots & 0 \\ \vdots & & & & & & \vdots \\ a_{2j-1} & a_{2j-2} & a_{2j-3} & a_{2j-4} & & \cdots & a_j \end{vmatrix}.$$

Zeros are entered for any a_k's that are called for where $k > n$. For example, if $n = 3$, then

$$\Delta_1 = a_1, \quad \Delta_2 = \begin{vmatrix} a_1 & 1 \\ a_3 & a_2 \end{vmatrix}, \quad \Delta_3 = \begin{vmatrix} a_1 & 1 & 0 \\ a_3 & a_2 & a_1 \\ 0 & 0 & a_3 \end{vmatrix};$$

expanding Δ_3 about the third row yields the simplification $\Delta_3 = a_3\Delta_2$. Here is the problem: Use the Routh–Hurwitz criterion to determine whether or not the systems associated with the following characteristic equations are stable.

(a) $\lambda^4 + 6\lambda^3 + 5\lambda^2 + 4\lambda + 1 = 0$
(b) $\lambda^4 + 2\lambda^3 + 7\lambda^2 + 4\lambda + 8 = 0$
(c) $\lambda^4 + 2\lambda^3 + 5\lambda^2 + 8\lambda + 12 = 0$
(d) $\lambda^4 + \lambda^3 + 4\lambda + 8 = 0$
(e) $\lambda^5 + \lambda^4 + \lambda^3 + \lambda^2 + \lambda + 1 = 0$
(f) $\lambda^5 + \lambda^4 + \lambda^3 + \lambda^2 + \lambda + 8 = 0$
(g) $\lambda^5 + 2\lambda^4 + 3\lambda^3 + 4\lambda^2 + 5\lambda + 6 = 0$

15. It can be shown that the equations

$$\begin{aligned} a_0x^2 + a_1x + a_2 &= 0, \qquad (a_0 \neq 0) \\ b_0x^2 + b_1x + b_2 &= 0 \qquad (b_0 \neq 0) \end{aligned}$$

have a common root if and only if

$$\begin{vmatrix} a_0 & a_1 & a_2 & 0 \\ 0 & a_0 & a_1 & a_2 \\ b_0 & b_1 & b_2 & 0 \\ 0 & b_0 & b_1 & b_2 \end{vmatrix} = 0.$$

[Similar results for two algebraic equations of degrees m and n, say, were put forward by *James Joseph Sylvester* (1814–1897).] Use this result to determine whether or not the following equation pairs have any common roots.

(a) $3x^2 + 2x - 5 = 0$, $3x^2 + 3x - 2 = 0$ (b) $3x^2 + 2x - 5 = 0$, $x^2 + x + 1 = 0$

16. (a) Suppose that the elements a_{ij} of an $n \times n$ matrix $\mathbf{A}$ are differentiable functions of some parameter t. Regarding det$\mathbf{A}$ as a function of the n^2 variables $a_{11}, a_{12}, \ldots, a_{nn}$, show that

$$\frac{\partial}{\partial a_{ij}}(\det\mathbf{A}) = A_{ij}, \quad (16.1)$$

where A_{ij} is the cofactor of a_{ij}. Then use (16.1) and chain differentiation to show that

$$\frac{d}{dt}(\det\mathbf{A}) = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}\frac{da_{ij}}{dt}, \quad (16.2)$$

a formula first given by *Carl Gustav Jacob Jacobi* (1804–1851) in 1841. By the $\sum\sum$ notation we mean

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} \equiv \sum_{i=1}^{n}\left(\sum_{j=1}^{n} c_{ij}\right).$$

For example, if $n = 2$, then

$$\sum_{i=1}^{2}\sum_{j=1}^{2} c_{ij} = \sum_{i=1}^{2}\left(\sum_{j=1}^{2} c_{ij}\right) = \sum_{i=1}^{2}(c_{i1} + c_{i2})$$
$$= c_{11} + c_{21} + c_{12} + c_{22}.$$

Observe that (16.2) is equivalent to the statement

$$\frac{d}{dt}(\det\mathbf{A}) = \begin{vmatrix} \frac{da_{11}}{dt} & \cdots & \frac{da_{1n}}{dt} \\ a_{21} & \cdots & a_{2n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{vmatrix}$$

$$+\cdots+\begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n-1,1} & \cdots & a_{n-1,n} \\ \dfrac{da_{n1}}{dt} & \cdots & \dfrac{da_{nn}}{dt} \end{vmatrix}. \tag{16.3}$$

(b) Thus, evaluate $\dfrac{d}{dt}(\det\mathbf{A})$ if

$$\mathbf{A} = \begin{bmatrix} t^2 & t & 2 \\ 0 & 3t & 1 \\ 4 & 0 & \sin t \end{bmatrix},$$

and check your result by working out $\det\mathbf{A}$ and taking its t derivative.

17. (*Vandermonde determinant*) First, review Theorem 3.4.1 and its proof. The determinant in (43),

$$\begin{vmatrix} 1 & \cdots & 1 \\ \lambda_1 & \cdots & \lambda_n \\ \vdots & & \vdots \\ \lambda_1^{n-1} & \cdots & \lambda_n^{n-1} \end{vmatrix}, \tag{17.1}$$

is known as a **Vandermonde determinant**. It can be shown that it equals $(-1)^{n(n-1)/2}\Pi$, where Π denotes the product of all factors $\lambda_j - \lambda_k$ with $j < k\ (\leq n)$. For example, if $n = 3$ then $\Pi = (\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3)(\lambda_2 - \lambda_3)$. The key property of any $n \times n$ Vandermonde determinant, which can be seen from this result, is that it is nonzero if and only if all of the λ_j's are distinct. The problem that we pose is for you to verify that the determinant is equal to $(-1)^{n(n-1)/2}\Pi$, as claimed above, simply by working out the determinant, for the cases where $n = 2$ and 3.

18. (a) Derive (8). HINT: Show that $N(n) = nN(n-1) + 2n-1$ for $n \geq 2$. Let us use the subscript notation $N(n) = P_n$, which is generally used for functions of a discrete integer variable. Thus,

$$P_n - nP_{n-1} = 2n - 1, \qquad (n \geq 2) \tag{18.1}$$

with the initial condition $P_2 = 3$ (two multiplications and one subtraction). Seeking P_n in the form $n!Q_n$, show that Q_n satisfies the difference equation

$$Q_n - Q_{n-1} = \frac{2n-1}{n!} \tag{18.2}$$

with initial condition $Q_2 = 3/2$, which admits the solution

$$Q_n = \frac{3}{2} + 2\sum_{j=2}^{n-1}\frac{1}{j!} - \sum_{j=3}^{n}\frac{1}{j!}. \tag{18.3}$$

[You need merely verify (18.3).] Finally, show from (18.3) that $Q_n \sim e$ as $n \to \infty$.

(b) Derive (10). HINT: $n(n-1) + (n-1)(n-2) + \cdots + (2)(1) = n(n^2-1)/3$.

10.5 Rank; Application to Linear Dependence and to Existence and Uniqueness for Ax=c

With determinants defined, we can now introduce one more concept, the "rank" of a matrix, which concept will enable us to obtain important results regarding linear dependence, and also regarding the existence and uniqueness of solutions of the linear equation $\mathbf{Ax} = \mathbf{c}$.

10.5.1. Rank. First, we say that any matrix obtained from a given $m \times n$ matrix $\mathbf{A}$ by deleting at most $m-1$ rows and at most $n-1$ columns from $\mathbf{A}$

is a **submatrix** of $\mathbf{A}$. For instance, the 2×3 matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}$$

has 21 submatrices: one 2×3 ($\mathbf{A}$ itself), three 2×2's, two 1×3's, three 2×1's, six 1×2's, and six 1×1's. For instance,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}, \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \begin{bmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{bmatrix}, \begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix}, [a_{21}, a_{23}],$$

and $[a_{13}]$ are all submatrices of $\mathbf{A}$.

Then the rank of a matrix is defined as follows.

DEFINITION 10.5.1 *Rank*
A matrix $\mathbf{A}$, not necessarily square, is of **rank** r, or $r(\mathbf{A})$, if it contains at least one $r \times r$ submatrix with nonzero determinant but no square submatrix larger than $r \times r$ with nonzero determinant. A matrix is of rank 0 if it is a zero matrix.

EXAMPLE 1. Let

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 1 & 0 \\ 0 & 3 & 3 & 6 \\ 1 & 4 & 5 & 9 \end{bmatrix}. \tag{1}$$

Certainly, r is at most 3 in this case since the largest possible square submatrix of $\mathbf{A}$ is 3×3. (More generally, if $\mathbf{A}$ is $m \times n$, then r is at most equal to the smaller of m and n.) However, upon calculation, we find that all four of the 3×3 submatrices have zero determinant so that r is at most 2. In fact, there are a number of 2×2 submatrices with nonzero determinant such as

$$\begin{vmatrix} 2 & -1 \\ 0 & 3 \end{vmatrix} = 6 \neq 0,$$

but even if there were only one such submatrix that would still be all we need to conclude that $r(\mathbf{A}) = 2$. ∎

EXAMPLE 2. The ranks of

$$\mathbf{A} = \begin{bmatrix} 5 \\ 6 \\ 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 5 & -2 \\ 6 & 3 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 5 & -2 & 0 \\ 6 & 3 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} 3 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 9 & 0 & 3 & 0 \end{bmatrix}$$

are 1, 2, 2, and 2, respectively. In $\mathbf{D}$, for example, every 3×3 submatrix contains a column of zeros and hence has a vanishing determinant, but the 2×2 submatrix $\begin{bmatrix} 3 & 1 \\ 0 & 1 \end{bmatrix}$ has a

nonvanishing determinant, so $r(\mathbf{D}) = 2$. ∎

We may regard the rows of an $m \times n$ matrix $\mathbf{A} = \{a_{ij}\}$ as n-dimensional vectors, which we call the **row vectors** of $\mathbf{A}$ and which we denote as $\mathbf{r}_1, \ldots, \mathbf{r}_m$. Similarly, the columns are m-dimensional vectors, which we call the **column vectors** of $\mathbf{A}$ and which we denote as $\mathbf{c}_1, \ldots, \mathbf{c}_n$. Further, we define the vector spaces span $\{\mathbf{r}_1, \ldots, \mathbf{r}_m\}$ and span $\{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$ as the **row** and **column spaces** of $\mathbf{A}$, respectively. From the definition of dimension, the dimensions of the row and column spaces are equal to the number of LI row vectors and the number of LI column vectors, respectively.

It will be important to be able to calculate the rank of a given matrix efficiently. With that purpose in mind, we recall the **elementary row operations** defined in Section 8.3:

1. Addition of a multiple of one row to another
   Symbolically: $\mathbf{r}_j \to \mathbf{r}_j + \alpha \mathbf{r}_k$
2. Multiplication of a row by a nonzero constant
   Symbolically: $\mathbf{r}_j \to \alpha \mathbf{r}_j$
3. Interchange of two rows
   Symbolically: $\mathbf{r}_j \leftrightarrow \mathbf{r}_k$

Furthermore, we defined (in Section 8.3) two matrices to be **row equivalent** if one can be obtained from the other by finitely many elementary row operations. The following theorem provides an efficient means of calculating the rank of a matrix.

THEOREM 10.5.1 *Elementary Row Operations and Rank*
Row equivalent matrices have the same rank. That is, elementary row operations do not alter the rank of a matrix.

Proof: If matrices $\mathbf{A}$ and $\mathbf{B}$ are row equivalent, then $\mathbf{B}$ can be obtained from $\mathbf{A}$ by a finite number of elementary row operations. It follows that each row vector of $\mathbf{B}$ must be a linear combination of the row vectors of $\mathbf{A}$ so the row space of $\mathbf{B}$ must be a subspace of the row space of $\mathbf{A}$. Similarly, the row space of $\mathbf{A}$ must be a subspace of the row space of $\mathbf{B}$. Thus, the row space of $\mathbf{A}$ is identical to the row space of $\mathbf{B}$, and hence the dimension of the row space of $\mathbf{A}$ (which is the rank of $\mathbf{B}$) must equal the dimension of the row space of $\mathbf{B}$ (which is the rank of $\mathbf{B}$). ∎

EXAMPLE 3. Let

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & -3 & 4 \\ 2 & 4 & -2 & 5 \\ 0 & 3 & 1 & 3 \\ 2 & 1 & -3 & -2 \end{bmatrix}. \tag{2}$$

Using elementary row operations,

$$\mathbf{A} \to \begin{bmatrix} 2 & 1 & -3 & 4 \\ 0 & 3 & 1 & 1 \\ 0 & 3 & 1 & 3 \\ 0 & 0 & 0 & -6 \end{bmatrix} \to \begin{bmatrix} 2 & 1 & -3 & 4 \\ 0 & 3 & 1 & 1 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & -6 \end{bmatrix} \to \begin{bmatrix} 2 & 1 & -3 & 4 \\ 0 & 3 & 1 & 1 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{3}$$

where the operations were as follows: in the first step $\mathbf{r}_2 \to \mathbf{r}_2 + (-1)\mathbf{r}_1$ and $\mathbf{r}_4 \to \mathbf{r}_4 + (-1)\mathbf{r}_1$; in the second step $\mathbf{r}_3 \to \mathbf{r}_3 + (-1)\mathbf{r}_2$; and in the final step $\mathbf{r}_4 \to \mathbf{r}_4 + 3\mathbf{r}_3$. Clearly, the rank of the final matrix is 3 because (deleting the fourth row and third column)

$$\begin{vmatrix} 2 & 1 & 4 \\ 0 & 3 & 1 \\ 0 & 0 & 2 \end{vmatrix} = 12 \neq 0.$$

Thus, by Theorem 10.5.1, $r(\mathbf{A}) = 3$. ∎

The idea, then, is to reduce a given matrix $\mathbf{A}$ to row echelon form by means of elementary row operations.* It can be seen that, in that form, the nonzero rows are LI. In Example 3, for instance, several conclusions follow from (3): $r(\mathbf{A}) = 3$, the number of LI vectors among the rows of $\mathbf{A}$ is 3; the dimension of the row space of $\mathbf{A}$ is 3 and a basis for that row space is given by the vectors $[2, 1, -3, 4]$, $[0, 3, 1, 1]$, and $[0, 0, 0, 2]$.

There is a connection between the rank of a matrix and the linear dependence of a set of vectors, studied in Chapter 9:

THEOREM 10.5.2 *Rank and Linear Dependence*
For any matrix $\mathbf{A}$, the number of LI row vectors is equal to the number of LI column vectors and these, in turn, equal the rank of $\mathbf{A}$.†

Thus, if we wish to determine how many vectors in a given vector set $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ are LI we can form a matrix $\mathbf{A}$ with $\mathbf{u}_1, \ldots, \mathbf{u}_k$ as the rows (or columns) and then use elementary row operations to determine the rank of $\mathbf{A}$.

EXAMPLE 4. How many LI vectors are contained in $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4\}$, where

$$\mathbf{u}_1 = \begin{bmatrix} 2 \\ 2 \\ 0 \\ 2 \end{bmatrix}, \quad \mathbf{u}_2 = \begin{bmatrix} 1 \\ 4 \\ 3 \\ 1 \end{bmatrix}, \quad \mathbf{u}_3 = \begin{bmatrix} -3 \\ -2 \\ 1 \\ -3 \end{bmatrix}, \quad \mathbf{u}_4 = \begin{bmatrix} 4 \\ 5 \\ 3 \\ -2 \end{bmatrix}?$$

*If we scale the first three rows in the final matrix in (3) by $\frac{1}{2}$, $\frac{1}{3}$, and $\frac{1}{2}$, respectively, so as to begin the nonzero rows with leading ones, then we would say that the resulting matrix is in *reduced* row echelon form.

†For proof of the first part of this theorem, see Theorem 3.5.5 in Steven J. Leon's *Linear Algebra*, 3rd ed. (New York: Macmillan, 1990).

If we construct a matrix having these vectors as columns, then we have the $\mathbf{A}$ matrix (2). Using elementary row operations, we saw, in Example 3, that $r(\mathbf{A}) = 3$. Hence, there are three LI vectors in $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4\}$ or, put differently, $\dim[\text{span}\,\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4\}] = 3$.

COMMENT. The ordering of the columns is immaterial. For instance, we could make $\mathbf{u}_1$ the third column, $\mathbf{u}_2$ the first, $\mathbf{u}_3$ the fourth, and $\mathbf{u}_4$ the second because rank depends upon the zeroness or nonzeroness of determinants whereas the interchanging of columns (or rows) merely changes the *sign* of a determinant. ∎

EXAMPLE 5. *Application to Stoichiometry.* To model the combustion of gasoline in an automobile engine, one can begin by writing down a list of well over 100 simultaneous chemical reactions involving the various hydrocarbons, oxygen, nitrogen, and so on. In turn, these reactions can be modeled by ODE's governing the amount of each chemical species as a function of time, and one can solve the resulting system of ODE's by methods such as those discussed in Chapter 6. It is easy to appreciate that solving around 100 coupled ODE's is a difficult undertaking. Thus, it is important to reduce the list of reactions insofar as possible, and we can do this by eliminating ones that are redundant. For instance, if $A + B \to C$ and $A + C \to D$, then a third statement, $2A + B + C \to C + D$, is redundant in that it is implied by the first two.

To illustrate the reduction process, consider the burning of a mixture of CO, H_2, and CH_4 in a furnace, producing CO, CO_2, and H_2O.* Writing all possible reactions that we can think of gives the list

$$CO + \frac{1}{2}O_2 \to CO_2, \tag{4a}$$

$$H_2 + \frac{1}{2}O_2 \to H_2O, \tag{4b}$$

$$CH_4 + \frac{3}{2}O_2 \to CO + 2H_2O, \tag{4c}$$

$$CH_4 + 2O_2 \to CO_2 + 2H_2O, \tag{4d}$$

where (4c) and (4d) represent the partial and complete combustion of CH_4, respectively. How many of these reactions are independent? It is convenient to re-express them symbolically in the equation format

$$\begin{aligned} CO + \tfrac{1}{2}O_2 - CO_2 &= 0, \\ H_2 + \tfrac{1}{2}O_2 - H_2O &= 0, \\ CH_4 + \tfrac{3}{2}O_2 - CO - 2H_2O &= 0, \\ CH_4 + 2O_2 - CO_2 - 2H_2O &= 0, \end{aligned} \tag{5}$$

where the elements of the coefficient matrix

*This example is discussed by Ben Noble in his book *Applications of Undergraduate Mathematics in Engineering* (New York: Macmillan, 1967). In turn, he notes that the problem was contributed by John Mahoney, Department of Chemical Engineering, West Virginia University, Morgantown, WV.

$$\mathbf{A} = \begin{array}{c} \begin{array}{cccccc} CO & O_2 & CO_2 & H_2 & H_2O & CH_4 \end{array} \\ \begin{bmatrix} 1 & \frac{1}{2} & -1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 1 & -1 & 0 \\ -1 & \frac{3}{2} & 0 & 0 & -2 & 1 \\ 0 & 2 & -1 & 0 & -2 & 1 \end{bmatrix} \end{array} \tag{6}$$

are known as stoichiometric coefficients. To determine a minimum set of independent reactions we reduce $\mathbf{A}$ by elementary row operations and obtain

$$\begin{array}{c} \begin{array}{cccccc} CO & O_2 & CO_2 & H_2 & H_2O & CH_4 \end{array} \\ \begin{bmatrix} 1 & \frac{1}{2} & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 2 & -2 & 0 \\ 0 & 0 & -1 & -4 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}, \tag{7}$$

the rank of which is three. Thus, there are three independent reactions such as the list

$$\begin{aligned} CO + \tfrac{1}{2}O_2 - CO_2 &= 0, \\ O_2 + 2H_2 - 2H_2O &= 0, \\ -CO_2 - 4H_2 + 2H_2O + CH_4 &= 0, \end{aligned} \tag{8}$$

implied by (7). That is, $CO + \frac{1}{2}O_2 \to CO_2$, and so on. ∎

10.5.2. Application of rank to the system $\mathbf{Ax} = \mathbf{c}$. In Chapter 8 we use the method of Gauss elimination both to solve systems of linear algebraic equations and to study the questions of the existence and uniqueness of solutions. Having developed vector and matrix concepts now, we can return to the important problem $\mathbf{Ax} = \mathbf{c}$ and bring these additional concepts to bear. In doing so, it is convenient to have a representative example to refer to.

EXAMPLE 6. Consider the system $\mathbf{Ax} = \mathbf{c}$ given by

$$\begin{aligned} x_1 - x_2 + x_3 + 3x_4 \qquad + 2x_6 &= 4, \\ x_1 \qquad + 3x_3 + 3x_4 - x_5 + 6x_6 &= 3, \\ 2x_1 - x_2 + 2x_3 + x_4 - x_5 + 7x_6 &= 9, \\ x_1 \qquad + 5x_3 + 8x_4 - x_5 + 7x_6 &= \underline{1}. \end{aligned} \tag{9}$$

Carrying out Gauss elimination by applying elementary row operations to the augmented matrix

$$\mathbf{A}|\mathbf{c} = \begin{bmatrix} 1 & -1 & 1 & 3 & 0 & 2 & 4 \\ 1 & 0 & 3 & 3 & -1 & 6 & 3 \\ 2 & -1 & 2 & 1 & -1 & 7 & 9 \\ 1 & 0 & 5 & 8 & -1 & 7 & 1 \end{bmatrix} \tag{10}$$

gives the row echelon result

$$\begin{bmatrix} 1 & -1 & 1 & 3 & 0 & 2 & 4 \\ 0 & 1 & 2 & 0 & -1 & 4 & -1 \\ 0 & 0 & 2 & 5 & 0 & 1 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & \underline{0} \end{bmatrix} \tag{11}$$

and hence the three-parameter family of solutions

$$\begin{gathered} x_6 = \alpha_1, \qquad x_5 = \alpha_2, \qquad x_4 = \alpha_3, \qquad x_3 = -1 - \tfrac{1}{2}\alpha_1 - \tfrac{5}{2}\alpha_3, \\ x_2 = 1 - 3\alpha_1 + \alpha_2 + 5\alpha_3, \qquad x_1 = 6 - \tfrac{9}{2}\alpha_1 + \alpha_2 + \tfrac{9}{2}\alpha_3, \end{gathered} \tag{12}$$

where the parameters $\alpha_1, \alpha_2, \alpha_3$ are arbitrary.

It is illuminating to express (12) in vector form as

$$\mathbf{x} = \begin{bmatrix} 6 - \frac{9}{2}\alpha_1 + \alpha_2 + \frac{9}{2}\alpha_3 \\ 1 - 3\alpha_1 + \alpha_2 + 5\alpha_3 \\ -1 - \frac{1}{2}\alpha_1 - \frac{5}{2}\alpha_3 \\ \alpha_3 \\ \alpha_2 \\ \alpha_1 \end{bmatrix} = \begin{bmatrix} 6 \\ 1 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \alpha_1 \begin{bmatrix} -\frac{9}{2} \\ -3 \\ -\frac{1}{2} \\ 0 \\ 0 \\ 1 \end{bmatrix} + \alpha_2 \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + \alpha_3 \begin{bmatrix} \frac{9}{2} \\ 5 \\ -\frac{5}{2} \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

$$= \mathbf{x}_0 + \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \alpha_3\mathbf{x}_3. \tag{13}$$

Observe that $\mathbf{x}_0$ is a *particular solution* of $\mathbf{Ax} = \mathbf{c}$ (i.e., $\mathbf{Ax}_0 = \mathbf{c}$), and $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ are *homogeneous solutions* (i.e., $\mathbf{Ax}_1 = \mathbf{0}$, $\mathbf{Ax}_2 = \mathbf{0}$, $\mathbf{Ax}_3 = \mathbf{0}$), by the following reasoning. Since (13) is a solution of $\mathbf{Ax} = \mathbf{c}$ for any α_j's, we can set $\alpha_1 = \alpha_2 = \alpha_3 = 0$ and conclude that $\mathbf{Ax}_0 = \mathbf{c}$. Next, put (13) into $\mathbf{Ax} = \mathbf{c}$:

$$\mathbf{A}\left(\mathbf{x}_0 + \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \alpha_3\mathbf{x}_3\right) = \mathbf{c}, \tag{14}$$

hence, $\mathbf{Ax}_0 + \alpha_1\mathbf{Ax}_1 + \alpha_2\mathbf{Ax}_2 + \alpha_3\mathbf{Ax}_3 = \mathbf{c}$ or, since $\mathbf{Ax}_0 = \mathbf{c}$,

$$\alpha_1\mathbf{Ax}_1 + \alpha_2\mathbf{Ax}_2 + \alpha_3\mathbf{Ax}_3 = \mathbf{0}. \tag{15}$$

The choice $\alpha_1 = 1, \alpha_2 = \alpha_3 = 0$ reveals that $\mathbf{Ax}_1 = \mathbf{0}$; $\alpha_2 = 1, \alpha_1 = \alpha_3 = 0$ reveals that $\mathbf{Ax}_2 = \mathbf{0}$; and $\alpha_3 = 1, \alpha_1 = \alpha_2 = 0$ reveals that $\mathbf{Ax}_3 = \mathbf{0}$, as claimed.

Observe further that $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ are LI, for the rank of

$$[\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3] = \begin{bmatrix} -\frac{9}{2} & 1 & \frac{9}{2} \\ -3 & 1 & 5 \\ -\frac{1}{2} & 0 & -\frac{5}{2} \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \tag{16}$$

is 3, as is readily seen from the bottom three rows. ∎

Generalizing the results of Example 6, suppose that a system

$$\mathbf{A}\mathbf{x} = \mathbf{c}, \tag{17}$$

where $\mathbf{A}$ is $m \times n$, has a p-parameter family of solutions

$$\mathbf{x} = \mathbf{x}_0 + \alpha_1\mathbf{x}_1 + \cdots + \alpha_p\mathbf{x}_p. \tag{18}$$

Then $\mathbf{x}_0$ is necessarily a particular solution, and $\mathbf{x}_1, \ldots, \mathbf{x}_p$ are necessarily LI homogeneous solutions. We call span $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ the **solution space** of the homogeneous equation $\mathbf{A}\mathbf{x} = \mathbf{0}$, or the **null space** of $\mathbf{A}$. The dimension of that null space is called the **nullity** of $\mathbf{A}$.

It is helpful to see that if $\{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$ denote the columns of $\mathbf{A}$, then $\mathbf{A}\mathbf{x} = \mathbf{c}$ can be expressed as

$$x_1\mathbf{c}_1 + x_2\mathbf{c}_2 + \cdots + x_n\mathbf{c}_n = \mathbf{c}, \tag{19}$$

from which we can see that (17) is consistent if and only if $\mathbf{c}$ happens to be in the column space of $\mathbf{A}$ [namely, span $\{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$]. Or, in terms of rank, (17) is consistent if and only if the rank of the augmented matrix $\mathbf{A}|\mathbf{c}$ equals the rank of the coefficient matrix $\mathbf{A}$: $r(\mathbf{A}|\mathbf{c}) = r(\mathbf{A})$.

In Example 6, for instance, we see from (11) that $r(\mathbf{A}|\mathbf{c}) = 3$ and (by covering up the last column, which is $\mathbf{c}$) that $r(\mathbf{A}) = 3$ as well. Thus, $r(\mathbf{A}|\mathbf{c}) = r(\mathbf{A})$ and, sure enough, (9) is consistent, solutions being given by (12). However, if we modify (9) by changing the underlined 1 to a 2, say, then the underlined 0 in (11) becomes a 1. In that case $r(\mathbf{A}|\mathbf{c}) = 4$ and $r(\mathbf{A}) = 3$ are unequal and there is no solution because the bottom row of (11) would then be equivalent to $0x_1 + \cdots + 0x_6 = 1$, which cannot be satisfied by any combination of x_j's.

Finally, what can be said about p in (18)? In Example 6, we see from (11) that $p = 3$ for there are three arbitrary x_j values (as seen from the third row), and that that value arises as the difference between the number of unknowns $n = 6$ and the rank $r = 3$.

Let us summarize, for any system $\mathbf{A}\mathbf{x} = \mathbf{c}$.

THEOREM 10.5.3 *Existence and Uniqueness for* $\mathbf{A}\mathbf{x} = \mathbf{c}$
Consider the linear system

$$\mathbf{A}\mathbf{x} = \mathbf{c}, \tag{20}$$

where $\mathbf{A}$ is $m \times n$. There is

1. no solution if and only if $r(\mathbf{A}|\mathbf{c}) \neq r(\mathbf{A})$,
2. a unique solution if and only if $r(\mathbf{A}|\mathbf{c}) = r(\mathbf{A}) = n$,
3. an $(n - r)$-parameter family of solutions if and only if $r(\mathbf{A}|\mathbf{c}) = r(\mathbf{A}) \equiv r$ is less than n.

The essential ideas required to prove these three results were developed above, so the proofs are left for the exercises.

Naturally, the **homogeneous** system

$$\mathbf{Ax} = \mathbf{0} \tag{21}$$

is but a special case of (20), hence, it is already covered by Theorem 10.5.3. However, it is such an important case that it deserves special attention. In (21), the augmented matrix $r(\mathbf{A}|\mathbf{c})$ is the $\mathbf{A}$ matrix augmented by a column of zeros so it is surely true that $r(\mathbf{A}|\mathbf{c}) = r(\mathbf{A})$ and, according to Theorem 10.5.3, it must be true that (21) is consistent. That result is no great surprise since (21) always admits the trivial solution $\mathbf{x} = \mathbf{0}$. Hence, the significant question about (21) is not whether or not it is consistent, but whether $\mathbf{x} = \mathbf{0}$ is the *only* solution. That is, does (21) admit *non*trivial solutions as well? That question is answered by parts 1 and 2 of Theorem 10.5.3 so we can state the following more specialized results.

THEOREM 10.5.4 *Homogeneous Case Where* $\mathbf{A}$ *is* $m \times n$
If $\mathbf{A}$ is $m \times n$, then

$$\mathbf{Ax} = \mathbf{0} \tag{22}$$

1. is consistent,
2. admits the trivial solution $\mathbf{x} = \mathbf{0}$,
3. admits the unique solution $\mathbf{x} = \mathbf{0}$ if and only if, $r(\mathbf{A}) = n$,
4. admits an $(n - r)$-parameter family of nontrivial solutions, in addition to the trivial solution, if and only if $r(\mathbf{A}) \equiv r < n$.

THEOREM 10.5.5 *Homogeneous Case Where* $\mathbf{A}$ *is* $n \times n$
If $\mathbf{A}$ is $n \times n$, then

$$\mathbf{Ax} = \mathbf{0} \tag{23}$$

admits nontrivial solutions, besides the trivial solution $\mathbf{x} = \mathbf{0}$, if and only if $\det\mathbf{A} = 0$.

As a final example, consider an interesting application of these concepts to "dimensional analysis."

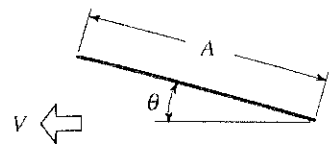


Figure 1. Flat plate airfoil.

EXAMPLE 7. *Dimensional Analysis.* Consider a rectangular flat plate (i.e., a flat rectangular wing, or "airfoil") in steady motion through otherwise-undisturbed air as shown in Fig. 1; V is the flight speed, θ is the incidence or angle of attack of the airfoil, A is the chord length, and B is the span (the dimension normal to the paper). Equivalently, it is experimentally more convenient to keep the airfoil fixed (in a wind tunnel) and to blow air

past it, at a speed V. Imagine that our object is to conduct an experimental determination of the lift force ℓ generated on the airfoil, that is, to experimentally determine the functional dependence of ℓ on the various relevant quantities. What quantities *are* relevant? Surely A, B, θ, V are important, as well as the air density ρ (for instance we expect a much greater lift in water than in air, and no lift at all in a vacuum). A reasonable list of the relevant variables is given in Table 1. Other variables come to mind, such as the ambient temperature,

Table 1. Relevant variables.

Variable	Symbol	Fundamental Units
Chord	A	L
Span	B	L
Incidence	θ	$M^0L^0T^0$
Flight velocity	V	LT^{-1}
Velocity of sound in air	V_0	LT^{-1}
Air density	ρ	ML^{-3}
Absolute viscosity	μ	$ML^{-1}T^{-1}$
Lift	ℓ	MLT^{-2}

but if we expect ℓ to be only weakly dependent on them we can leave them out.

Major difficulties are now apparent. If ℓ depends upon the seven variables listed in Table 1, and we measure ℓ for five different values of each variable, then we need to conduct $5^7 = 78{,}125$ experimental runs, and then present the results (using graphs, tables, or whatever) in a user-friendly way. Furthermore, whereas some variables are easily varied (such as A, B, θ, V) others are not (such as ρ, μ). The principal object of the following "dimensional analysis" is to reduce the number of variables as much as possible.

To begin, we express each variable in terms of the fundamental units M (mass), L (length), and T (time), in the right-hand column. We do not need to include F (force) as a fundamental unit because, according to Newton's second law, $F = MLT^{-2}$, dimensionally speaking. Also, observe that θ is dimensionless.*

Next, we seek all possible dimensionless products of the form

$$A^a B^b \theta^c V^d V_0^e \rho^f \mu^g \ell^h. \tag{24}$$

That is, we seek the exponents $a, \ldots, h$ such that

$$\begin{aligned}&(L)^a\,(L)^b\,(M^0L^0T^0)^c\,(LT^{-1})^d\,(LT^{-1})^e\,(ML^{-3})^f\,(ML^{-1}T^{-1})^g\,(MLT^{-2})^h\\&= M^0L^0T^0.\end{aligned} \tag{25}$$

Equating exponents of L, T, M on both sides, we see that $a, \ldots, h$ must satisfy the homo-

*Recall that angle is defined by the formula $s = r\theta$, where s is the arc length of a circular arc of radius r, subtended by an angle θ, measured in radians. Thus, $\theta = s/r$ = length / length = dimensionless.

geneous linear system

$$\begin{aligned} a + b + \quad d + e - 3f - g + \quad h &= 0, \\ -d - e \qquad\quad - g - 2h &= 0, \\ f + g + \quad h &= 0. \end{aligned} \tag{26}$$

Solving (26) by Gauss elimination gives the five-parameter family of solutions

$$\begin{bmatrix} a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \end{bmatrix} = \alpha_1 \begin{bmatrix} -2 \\ 0 \\ 0 \\ -2 \\ 0 \\ -1 \\ 0 \\ 1 \end{bmatrix} + \alpha_2 \begin{bmatrix} -1 \\ 0 \\ 0 \\ -1 \\ 0 \\ -1 \\ 1 \\ 0 \end{bmatrix} + \alpha_3 \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \alpha_4 \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \alpha_5 \begin{bmatrix} -1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \tag{27}$$

where $\alpha_1, \ldots, \alpha_5$ are arbitrary constants. With $\alpha_1 = 1$ and $\alpha_2 = \cdots = \alpha_5 = 0$, say, (27) gives $a = -2, b = c = 0$, etc., and hence the nondimensional parameter $A^{-2}B^0\theta^0V^{-2}V_0^0\rho^{-1}\mu^0\ell^1$, namely, the nondimensional lift $\ell/(\rho V^2A^2)$.* Similarly, setting $\alpha_2 = 1$ and the other α_j's $= 0$ gives the nondimensional parameter $\rho AV/\mu$, well known in fluid mechanics as the *Reynolds number* and denoted as Re; setting $\alpha_3 = -1$ and the other α_j's $= 0$ gives the *Mach number* V/V_0, denoted as $\mathcal{M}$; setting $\alpha_4 = 1$ and the other α_j's $= 0$ gives the *incidence* θ (which was nondimensional to begin with); and setting $\alpha_5 = 1$ and the other α_j's $= 0$ gives B/A, known as the *aspect ratio*, denoted typically as $\mathcal{AR}$.

The upshot is that rather than seek a functional relationship on the eight variables listed in the table, we can seek a relationship on the five nondimensional variables $[\ell/(\rho V^2A^2), \mathrm{Re}, \mathcal{M}, \theta, \mathcal{AR}]$. Or, singling out the nondimensional lift, we can express

$$\frac{\ell}{\rho V^2 A^2} = f\,(\mathrm{Re}, \mathcal{M}, \theta, \mathcal{AR}) \tag{28}$$

and determine f experimentally by measuring $\ell/(\rho V^2A^2)$ for various combinations of Re, $\mathcal{M}$, θ, and $\mathcal{AR}$ values.

COMMENT 1. A trained fluid mechanicist could probably simplify the problem even further. For example, it is known (from the governing equations of fluid mechanics) that the effect of the Mach number $\mathcal{M}$ will be negligible if $\mathcal{M}^2 \ll 1$. Thus, if we have flight speeds V less than 300 miles per hour in mind, then $\mathcal{M}$ can, to a good approximation, be dropped from (28).†

COMMENT 2. We mentioned, above, the practical difficulty in carrying out the experiment for a range of values of the fluid density. For instance, we could use air and water, but

*A fluid mechanicist would probably change this to $\ell/(\rho V^2AB)$ (corresponding to $\alpha_1 = 1, \alpha_2 = \alpha_3 = \alpha_4 = 0, \alpha_5 = -1$), or to $\ell/(\frac{1}{2}\rho V^2AB)$ because $\frac{1}{2}\rho V^2$ has physical significance as the stagnation pressure, and AB is the area of the airfoil.

†At ground level, the speed of sound is 762 mph, so if $V = 300$ mph then $\mathcal{M} = (300/762)^2 = 0.155$ is indeed small compared to 1.

their densities are widely different and we would need both wind tunnel and water tunnel facilities. In the right-hand side of (28), however, ρ shows up only within the Reynolds number $\mathrm{Re} = \rho AV/\mu$, which can be varied readily by varying the wind speed V.

COMMENT 3. Of course, (27) gives an *infinite* number of nondimensional parameters. However, there are only five independent ones, such as the ones named above. For instance, we could choose $\alpha_3 = \alpha_5 = 1$ and $\alpha_1 = \alpha_2 = \alpha_4 = 0$, but the resulting nondimensional parameter, $V_0B/(VA)$, is merely the aspect ratio divided by the Mach number. ∎

Closure. The rank $r(\mathbf{A})$ is defined as the size of the largest nonvanishing determinant within $\mathbf{A}$. Because the rank of a matrix is unaffected by elementary row operations, we can determine the rank of a given matrix efficiently by reducing it to row echelon form, in which form the rank can be seen by inspection. Principal applications of the concept of rank include the calculation of the number of LI vectors (n-tuple vectors, that is) within a given set, and the theory of the existence and uniqueness of solutions of systems of linear algebraic equations.

Computer software. Using *Maple*, we can evaluate the rank of a given matrix using the **rank(A)** command. For instance, to evaluate $r(\mathbf{A})$ where the rows of $\mathbf{A}$ are $[1, 2, 3, 4]$, $[2, 4, 6, 8]$, and $[1, 1, 1, 1]$, respectively, enter

```
with(linalg):
```

to access the rank(A) command. Then enter

```
rank(array([[1,2,3,4,], [2,4,6,8], [1,1,1,1]]);
```

and return. The output is 2. Alternatively, the sequence

```
A := array([[1,2,3,4,], [2,4,6,8], [1,1,1,1]]):
rank(A);
```

gives the same result.

EXERCISES 10.5

1. Determine the rank, nullity, number of LI rows, and number of LI columns for the given matrix.

(a) $[0, 0, 2, 0]$

(b) $[1, 2, 3]$

(c) $\begin{bmatrix} 5 & 7 \\ 4 & 9 \end{bmatrix}$

(d) $\begin{bmatrix} 4 & 8 & 0 \\ 3 & 6 & 0 \end{bmatrix}$

(e) $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$

(f) $\begin{bmatrix} 3 & 2 & 1 \\ -1 & 1 & 4 \\ 1 & 4 & 9 \end{bmatrix}$

(g) $\begin{bmatrix} 5 & 0 & 0 \\ 3 & 0 & 4 \\ 2 & 0 & 0 \end{bmatrix}$

(h) $\begin{bmatrix} 1 & 3 & 2 \\ 2 & 6 & 4 \\ 3 & 9 & 6 \end{bmatrix}$

(i) $\begin{bmatrix} 6 & 5 & 2 & 0 \\ 0 & 2 & 3 & 0 \\ 0 & 0 & 0 & 6 \end{bmatrix}$

(j) $\begin{bmatrix} 1 & 0 & 3 & 0 \\ 0 & 1 & 0 & 3 \\ 1 & 0 & 3 & 1 \end{bmatrix}$

(k) $\begin{bmatrix} 1 & 0 & 2 \\ 2 & 1 & -1 \\ 0 & 1 & 3 \\ 1 & 2 & 4 \end{bmatrix}$ (l) $\begin{bmatrix} 0 & 4 & -1 & 1 \\ 1 & 1 & 5 & -1 \\ 1 & 5 & 4 & 0 \\ 2 & 6 & 9 & -1 \end{bmatrix}$

(m) $\begin{bmatrix} 3 & 1 & -4 & 4 \\ 0 & 2 & 2 & 9 \\ 2 & 2 & 1 & 0 \\ -1 & 3 & 7 & 5 \end{bmatrix}$ (n) $\begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & -1 & 0 & 5 \\ 6 & -1 & 2 & 8 \\ 7 & -1 & 0 & 7 \\ 8 & -1 & -3 & 5 \end{bmatrix}$

2. (a)–(n) Use computer software to determine the rank of the matrix given in the corresponding part of Exercise 1.

3. (a)–(n) Consider the problem $\mathbf{Ax} = \mathbf{c}$, where $\mathbf{A}$ is the $m \times n$ matrix given in the corresponding part of Exercise 1. In each case let $\mathbf{c}$ be the m-dimensional vector $[1, 1, \ldots, 1]^T$. Use Theorem 10.5.3, and suitable rank calculations, to determine whether or not the system is consistent. If consistent, determine whether it admits a unique solution or a p-parameter family of solutions. If the latter, determine p. Do not solve the system; merely use the concept of rank and Theorem 10.5.3.

4. If two matrices of the same form have the same rank, need they be row equivalent? Prove or disprove.

5. Show, by carrying out suitable row operations, that the following pairs of matrices are row equivalent.

(a) $\begin{bmatrix} 1 & 3 & -3 & 0 \\ 2 & 1 & 0 & 4 \end{bmatrix}$ and $\begin{bmatrix} 4 & 2 & 0 & 8 \\ 3 & 4 & -3 & 4 \end{bmatrix}$

(b) $\begin{bmatrix} 2 & 1 & 0 \\ 2 & 2 & 3 \\ 4 & 4 & 9 \\ 4 & 4 & 6 \end{bmatrix}$ and $\begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & 3 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix}$

(c) $\begin{bmatrix} 5 & 2 \\ 0 & -3 \\ 1 & 4 \end{bmatrix}$ and $\begin{bmatrix} 9 & 2 \\ -6 & -3 \\ 9 & 4 \end{bmatrix}$

6. Show that the following pairs of matrices are *not* row equivalent.

(a) $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ and $\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$

(b) $\begin{bmatrix} 0 & 2 & 0 \\ 0 & 0 & 4 \\ 6 & 0 & 0 \end{bmatrix}$ and $\begin{bmatrix} 3 & -1 & 5 \\ 1 & 5 & -5 \\ 2 & 2 & 0 \end{bmatrix}$

7. Exercises 5 and 6 are simple enough to be solvable by inspection. More generally, inspection may not suffice. Put forward a systematic procedure for determining whether or not two given matrices are row equivalent, and apply that procedure to the given matrices.

(a) $\begin{bmatrix} 1 & 2 & 3 & -1 \\ 2 & 4 & -1 & 1 \\ 0 & 5 & 6 & 3 \\ 4 & -2 & 0 & 6 \end{bmatrix}$ and $\begin{bmatrix} 1 & -1 & 2 & 1 \\ 3 & 0 & 5 & 1 \\ 2 & 2 & 1 & 3 \\ 3 & 1 & 3 & 4 \end{bmatrix}$

(b) $\begin{bmatrix} 4 & -1 & 2 & 3 \\ 0 & 1 & 1 & 2 \\ -1 & 0 & 2 & 0 \end{bmatrix}$ and $\begin{bmatrix} 3 & 0 & 5 & 5 \\ 6 & 0 & -1 & 5 \\ 2 & 0 & 7 & 5 \end{bmatrix}$

8. Determine whether the following set of vectors is LI or LD by computing the rank of a suitable matrix and invoking the relevant theorem.

(a) $[2, 0, 1, -1]$, $[0, 3, 0, 3]$, $[4, 3, 2, 1]$
(b) $[4, 1, 2]$, $[2, 2, 1]$, $[2, -1, 1]$, $[4, 7, 2]$, $[0, 1, 0]$
(c) $[1, 3, 2, 4, 5]$, $[2, 3, 1, 5, 4]$, $[4, 5, 3, 1, 2]$
(d) $[2, 1, 1]$, $[4, 2, 2]$, $[0, 1, 2]$, $[1, 0, 0]$
(e) $[1, -2, 0, 1]$, $[0, 1, 1, 2]$, $[1, 0, 2, 5]$, $[2, -7, -3, -4]$

9. Prove that $r(\mathbf{A}^{\mathrm{T}}) = r(\mathbf{A})$ for every $m \times n$ matrix using any results given in this section.

10. The property $\det(\mathbf{AB}) = (\det\mathbf{A})(\det\mathbf{B})$, of determinants, where $\mathbf{A}$ and $\mathbf{B}$ are both $n \times n$, might seem to imply that $r(\mathbf{AB}) = r(\mathbf{A})r(\mathbf{A})$. Is the latter true? Prove or disprove.

11. Is $r(\mathbf{A} + \mathbf{B}) = r(\mathbf{A}) + r(\mathbf{B})$ true? Prove or disprove.

12. Prove that if $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times p$, then $r(\mathbf{AB}) \leq n$. HINT: Partition $\mathbf{B}$ into rows, and write

$$\mathbf{AB} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} \mathbf{r}_1 \\ \vdots \\ \mathbf{r}_n \end{bmatrix} = \begin{bmatrix} a_{11}\mathbf{r}_1 + \cdots + a_{1n}\mathbf{r}_n \\ \vdots \\ a_{m1}\mathbf{r}_1 + \cdots + a_{mn}\mathbf{r}_n \end{bmatrix}.$$

13. (a) Below (18), we stated that $\mathbf{x}_1, \ldots, \mathbf{x}_p$ are necessarily LI. Prove that claim. HINT: Pattern your proof after the discussion in Example 6.
(b) Show that $\mathbf{x}_0$ cannot be in the span of $\mathbf{x}_1, \ldots, \mathbf{x}_p$.

14. Although we made a case for the truth of Theorem 10.5.3, we did not provide a detailed proof.

(a) Prove part 1. (b) Prove part 2. (c) Prove part 3.

15. This exercise refers to Example 6 and the discussion following that example. For definiteness, let $\mathbf{A}$ be 3×3.

(a) Suppose that $\mathbf{Ax} = \mathbf{c}$ admits a one-parameter family of solutions

$$\mathbf{x} = \mathbf{x}_0 + \alpha_1 \mathbf{x}_1. \tag{15.1}$$

Explain, with the help of a labeled sketch, the geometrical significance of $\mathbf{x}_0, \mathbf{x}_1$ and $\mathbf{x}_0 + \alpha_1 \mathbf{x}_1$.
(b) Suppose that $\mathbf{A}\mathbf{x} = \mathbf{c}$ admits a two-parameter family of solutions

$$\mathbf{x} = \mathbf{x}_0 + \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2. \tag{15.2}$$

Explain, with the help of a labeled sketch, the geometrical significance of $\mathbf{x}_0$ and $\mathbf{x}_0 + \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2$.

16. (*Stoichiometry*) Determine how many of the following reactions are independent, and give such a set of independent reactions.

(a)
$$\begin{aligned} H_2 + O_2 &\rightleftharpoons 2OH, \\ H_2 + \tfrac{1}{2}O_2 &\rightleftharpoons H_2O, \\ H + OH &\rightleftharpoons H_2 + O, \\ H_2 &\rightleftharpoons 2H, \\ O_2 &\rightleftharpoons 2O \end{aligned}$$

(b)
$$\begin{aligned} H_2 + Cl_2 &\rightleftharpoons 2HCl, \\ Cl + H_2 &\rightleftharpoons HCl + H, \\ H_2 &\rightleftharpoons 2H, \\ Cl_2 &\rightleftharpoons 2Cl \end{aligned}$$

(c)
$$\begin{aligned} O_2 + CH_3 &\rightarrow CH_3OO, \\ CH_4 + CH_3OO &\rightarrow CH_3 + CH_3OOH, \\ CH_3OOH &\rightarrow CO + 2H_2 + O, \\ CH_4 + O &\rightarrow CH_3 + OH, \\ CH_4 + O_2 &\rightarrow CH_3OOH \end{aligned}$$

17. (*Dimensional analysis*) In studying the drag force on a sphere moving beneath a water surface, the tabulated variables are deemed relevant. Proceeding along the same lines as in Example 7, obtain the following relevant dimensionless parameters: the dimensionless drag force $D/(\rho V^2 R^2)$, the *Reynolds number* $\mathrm{Re} = \rho R V/\mu$, the *Froude number* $V^2/(Rg)$, and the two length ratios λ/R and d/R.

Variable	Symbol	Fundamental Units
Radius of sphere	R	L
Depth below water surface	d	L
Velocity of sphere	V	LT^{-1}
Water density	ρ	ML^{-3}
Absolute viscosity	μ	$ML^{-1}T^{-1}$
Gravity	g	LT^{-2}
Wavelength of free surface waves	λ	L
Drag force	D	MLT^{-2}

10.6 Inverse Matrix, Cramer's Rule, Factorization

There exist important methods for solving a linear system $\mathbf{A}\mathbf{x} = \mathbf{c}$ besides Gauss elimination. In this section we study three: the inverse matrix method, Cramer's rule, and LU factorization.

10.6.1. Inverse matrix. Having introduced matrix notation so that a system of linear algebraic equations can be expressed compactly as

$$\mathbf{A}\mathbf{x} = \mathbf{c}, \tag{1}$$

the form of (1) itself suggests other solution strategies. For how would we solve the simple scalar equation $3x = 12$? We could divide both sides by 3 and obtain $x = \frac{12}{3} = 4$. However, if we try to carry that idea over to the matrix case (1), we obtain $\mathbf{x} = \dfrac{\mathbf{c}}{\mathbf{A}}$, and need to know how to divide one matrix into another. However, matrix division has not been (and will not be) defined. Alternatively, we can solve

$3x = 12$ by multiplying both sides by $\frac{1}{3}$, for that step gives $\frac{1}{3}3x = \frac{1}{3}12$, or $1x = 4$, and hence $x = 4$. That idea does carry over to (1) because matrix *multiplication* is defined.

The idea, then, is to seek a matrix "$\mathbf{A}^{-1}$" having the property that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ for then

$$\mathbf{A}^{-1}\mathbf{A}\mathbf{x} = \mathbf{A}^{-1}\mathbf{c} \tag{2}$$

becomes

$$\mathbf{I}\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}, \tag{3}$$

and since $\mathbf{I}\mathbf{x} = \mathbf{x}$, we have the solution

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{c} \tag{4}$$

of (1). Note that $\mathbf{A}^{-1}$ does not mean $1/\mathbf{A}$ or $\mathbf{I}/\mathbf{A}$; it is simply the name of the matrix having the property

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}, \tag{5}$$

if one exists. We call it the **inverse** of $\mathbf{A}$, or "$\mathbf{A}$-inverse" for brevity.

Consider an exploratory example.

EXAMPLE 1. Let $\mathbf{A}\mathbf{x} = \mathbf{c}$ be the system

$$\begin{aligned} -x_1 - 2x_2 &= 1 \\ x_1 + x_2 &= 1 \\ x_2 &= 1 \end{aligned} \quad \text{or} \quad \begin{bmatrix} -1 & -2 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}. \tag{6}$$

We find, for example by trial and error, that

$$\begin{bmatrix} 1 & 2 & 0 \\ 1 & 1 & 2 \end{bmatrix} \begin{bmatrix} -1 & -2 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{I}, \tag{7}$$

so that the pre-multiplication of (6) by the matrix $\begin{bmatrix} 1 & 2 & 0 \\ 1 & 1 & 2 \end{bmatrix}$ yields

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 0 \\ 1 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \tag{8}$$

and hence the (tentative) solution $x_1 = 3, x_2 = 4$. Yet the latter does not satisfy (6), so the method is in some way incorrect. [In fact, Gauss elimination reveals that (6) is inconsistent, has no solution.] ∎

In other words, we must proceed with caution. The idea is that the pre-multiplication of $\mathbf{A}\mathbf{x} = \mathbf{c}$ by a matrix $\mathbf{A}^{-1}$ having the property that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ *does not necessarily lead to an equivalent system.*

However, we now show that if $\mathbf{A}$ is square, say $n \times n$, and a matrix $\mathbf{A}^{-1}$ can be found such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, then the pre-multiplication of $\mathbf{Ax} = \mathbf{c}$ by $\mathbf{A}^{-1}$ *does* lead to an equivalent system, namely, the unique solution $\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}$.

First, assuming that $\mathbf{A}$ is $n \times n$, observe from

$$\underbrace{\mathbf{A}^{-1}}_{m\times p}\ \underbrace{\mathbf{A}}_{n\times n} = \mathbf{I} \tag{9}$$

that p (the number of columns in $\mathbf{A}^{-1}$) must equal n for $\mathbf{A}^{-1}$ to be conformable for multiplication with $\mathbf{A}$, and that m (the number of rows in $\mathbf{A}^{-1}$) must equal n for the product $\mathbf{A}^{-1}\mathbf{A}$ to be square. Thus, $\mathbf{A}^{-1}$ is necessarily square too, of order n.

Second, it follows from (9) that

$$\det(\mathbf{A}^{-1}\mathbf{A}) = \det\mathbf{I} = 1$$

or, since the determinant of a product equals the product of the determinants (property D12 in Section 10.4),

$$(\det\mathbf{A}^{-1})(\det\mathbf{A}) = 1. \tag{10}$$

Equation (10) cannot possibly be satisfied if $\det\mathbf{A} = 0$. Hence, if a matrix $\mathbf{A}^{-1}$ satisfying (9) is to exist, it is *necessary* that $\det\mathbf{A} \neq 0$.

Assuming that that is the case, that $\det\mathbf{A} \neq 0$, let us seek to determine $\mathbf{A}^{-1}$. Our starting point is the cofactor expansion from Section 10.4,

$$\det\mathbf{A} = \sum a_{jk}A_{jk}, \tag{11}$$

where A_{jk} is the cofactor of the a_{jk} element, and the sum is either on j (for any fixed value of k) or on k (for any fixed value of j). Let us take the sum to be on j. Observe that

$$\sum_j a_{jk}A_{ji} = \begin{cases} \det\mathbf{A} & \text{if } i = k, \\ 0 & \text{if } i \neq k \end{cases} \tag{12}$$

since if $i = k$, then (11) applies, and if $i \neq k$, then the left-hand side of (12) is again a cofactor expansion, this time an expansion about the ith column – but with the ith column replaced by the kth column; thus, it is a determinant containing two identical columns and according to property D5 in Section 10.4, it must therefore be zero. Rearranging (12) by dividing through by $\det\mathbf{A}$ (which is permissible since we have assumed that $\det\mathbf{A} \neq 0$) and using the Kronecker delta notation,*

$$\sum_j \left(\frac{A_{ji}}{\det\mathbf{A}}\right) a_{jk} = \delta_{ik}. \tag{13}$$

*Defined in Section 9.10.2, δ_{ik} is simply 1 if $i = k$ and 0 if $i \neq k$. Thus, a matrix $\{\delta_{ik}\}$ is an identity matrix $\mathbf{I}$.

This scalar statement (which holds for $1 \le i \le n$ and $1 \le j \le n$) is equivalent, according to the definition of matrix multiplication,† to the matrix equation

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}, \tag{14}$$

where the desired **inverse matrix** $\mathbf{A}^{-1}$ is‡

$$\mathbf{A}^{-1} = \{\alpha_{ij}\} = \left\{\frac{A_{ji}}{\det\mathbf{A}}\right\}. \tag{15}$$

Or, written out,

$$\mathbf{A}^{-1} = \frac{1}{\det\mathbf{A}}\begin{bmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \vdots & & & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{bmatrix}. \tag{16}$$

CAUTION: The A_{ji} in (15) is not a misprint; the j, i indices simply turn out to be in the reverse order of those in α_{ij}. For instance, the $1, 2$ element of $\mathbf{A}^{-1}$ is $A_{21}/\det\mathbf{A}$, not $A_{12}/\det\mathbf{A}$, where A_{21} is the *cofactor* of the $2, 1$ element in $\mathbf{A}$, not the $2, 1$ *element* of $\mathbf{A}$ (which is a_{21}).

The matrix in (16) is called the **adjoint of A** and is denoted as $\operatorname{adj}\mathbf{A}$ so

$$\boxed{\mathbf{A}^{-1} = \frac{1}{\det\mathbf{A}}\operatorname{adj}\mathbf{A}.} \tag{17}$$

To form the adjoint of a given square matrix $\mathbf{A}$ we replace each element by its cofactor, and take the transpose of the resulting matrix.

The upshot, then, is that if $\det\mathbf{A} \neq 0$, then $\mathbf{A}^{-1}$ exists and is given by (17). In that case we say that $\mathbf{A}$ is **invertible**. If $\det\mathbf{A} = 0$, then $\mathbf{A}^{-1}$ does not exist, and we say that $\mathbf{A}$ is **singular**.

EXAMPLE 2. Determine the inverse of

$$\mathbf{A} = \begin{bmatrix} 3 & 2 & -1 \\ 0 & 1 & 4 \\ 1 & 5 & -2 \end{bmatrix}, \tag{18}$$

if it exists. It does exist because $\det\mathbf{A} = -57 \neq 0$ and is given by (17). Since $\operatorname{adj}\mathbf{A}$ is the transpose of the cofactor matrix, we have

$$\operatorname{adj}\mathbf{A} = \begin{bmatrix} \begin{vmatrix} 1 & 4 \\ 5 & -2 \end{vmatrix} & -\begin{vmatrix} 0 & 4 \\ 1 & -2 \end{vmatrix} & \begin{vmatrix} 0 & 1 \\ 1 & 5 \end{vmatrix} \\ -\begin{vmatrix} 2 & -1 \\ 5 & -2 \end{vmatrix} & \begin{vmatrix} 3 & -1 \\ 1 & -2 \end{vmatrix} & -\begin{vmatrix} 3 & 2 \\ 1 & 5 \end{vmatrix} \\ \begin{vmatrix} 2 & -1 \\ 1 & 4 \end{vmatrix} & -\begin{vmatrix} 3 & -1 \\ 0 & 4 \end{vmatrix} & \begin{vmatrix} 3 & 2 \\ 0 & 1 \end{vmatrix} \end{bmatrix}^{\mathrm{T}}$$

†Recall that if $\mathbf{BC} = \mathbf{D}$, then $d_{ij} = \sum_k b_{ik}c_{kj}$ or, what is equivalent, $d_{ik} = \sum_j b_{ij}c_{jk}$.

‡It is tempting to let the i, j element of $\mathbf{A}^{-1}$ be denoted as a_{ij}^{-1}, but this quantity could be misunderstood to be $1/a_{ij}$. Thus, let us use α_{ij}.

$$= \begin{bmatrix} -22 & 4 & -1 \\ -1 & -5 & -13 \\ 9 & -12 & 3 \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} -22 & -1 & 9 \\ 4 & -5 & -12 \\ -1 & -13 & 3 \end{bmatrix} \tag{19}$$

so

$$\mathbf{A}^{-1} = \frac{1}{\det\mathbf{A}}\,\mathrm{adj}\mathbf{A} = -\frac{1}{57}\begin{bmatrix} -22 & -1 & 9 \\ 4 & -5 & -12 \\ -1 & -13 & 3 \end{bmatrix}. \tag{20}$$

It is readily verified from (20) and (18) that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. ∎

Besides having $\mathbf{A}^{-1}$ satisfy $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, as stated in (14), it is crucial, as we shall see, to have $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$ as well. To show that $\mathbf{A}\mathbf{A}^{-1}$ does equal $\mathbf{I}$, write

$$\mathbf{A}\mathbf{A}^{-1} = \left\{\sum_k a_{ik}\alpha_{kj}\right\} = \left\{\sum_k a_{ik}\frac{A_{jk}}{\det\mathbf{A}}\right\} = \frac{1}{\det\mathbf{A}}\left\{\sum_k a_{ik}A_{jk}\right\}, \tag{21}$$

where the first equality follows from the definition of matrix multiplication, and the second follows from (15). Now,

$$\sum_k a_{ik}A_{jk} = \begin{cases} \det\mathbf{A}, & i = j \\ 0, & i \neq j \end{cases} \tag{22}$$

because if $i = j$ then (11) applies, and if $i \neq j$ then the left-hand side of (22) is again a cofactor expansion, this time an expansion about the jth row – but with the jth row replaced by the ith row. Thus, it is a determinant containing two identical rows and, according to property D5, it must therefore be zero. Hence (21) becomes

$$\mathbf{A}\mathbf{A}^{-1} = \{\delta_{ij}\} = \mathbf{I}, \tag{23}$$

as claimed. Thus,

$$\boxed{\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}.} \tag{24}$$

In view of the first equality in (24), we see that $\mathbf{A}^{-1}$ and $\mathbf{A}$ necessarily commute.

To understand the significance of (24), let us review the solution of $\mathbf{A}\mathbf{x} = \mathbf{c}$ by the inverse matrix method. The steps (2) and (3) gave $\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}$, provided that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. To verify that $\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}$ does indeed satisfy $\mathbf{A}\mathbf{x} = \mathbf{c}$, let us put $\mathbf{A}^{-1}\mathbf{c}$ into that equation in place of $\mathbf{x}$:

$$\mathbf{A}\left(\mathbf{A}^{-1}\mathbf{c}\right) = \mathbf{c} \tag{25}$$

or, by the associative property of matrix multiplication,

$$\left(\mathbf{A}\mathbf{A}^{-1}\right)\mathbf{c} = \mathbf{c}, \tag{26}$$

which is indeed true because $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$.*

Let us pull these results together.

THEOREM 10.6.1 *Inverse Matrix*
Let $\mathbf{A}$ be $n \times n$. If $\det\mathbf{A} \neq 0$, then there exists a unique matrix $\mathbf{A}^{-1}$, also $n \times n$, called the inverse of $\mathbf{A}$, such that

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}. \tag{27}$$

$\mathbf{A}$ is then said to be invertible, and its inverse is given by (17). If $\det\mathbf{A} = 0$, then a matrix $\mathbf{A}^{-1}$ satisfying (27) does not exist, and $\mathbf{A}$ is said to be singular.*

Proof: In the discussion preceding the theorem we proved all but the uniqueness of $\mathbf{A}^{-1}$. To prove uniqueness, let $\mathbf{B}$ and $\mathbf{C}$ both be inverses of $\mathbf{A}$. Then $\mathbf{B}\mathbf{A} = \mathbf{I}$ and $\mathbf{C}\mathbf{A} = \mathbf{I}$. Subtracting, $\mathbf{B}\mathbf{A} - \mathbf{C}\mathbf{A} = \mathbf{0}$ or $(\mathbf{B} - \mathbf{C})\mathbf{A} = \mathbf{0}$. And post-multiplying this last equation by $\mathbf{A}^{-1}$ (which exists by assumption), we have $(\mathbf{B} - \mathbf{C})\mathbf{A}\mathbf{A}^{-1} = \mathbf{0}\mathbf{A}^{-1}$ or $(\mathbf{B} - \mathbf{C})\mathbf{I} = \mathbf{0}$. Thus, $\mathbf{B} - \mathbf{C} = \mathbf{0}$, and hence $\mathbf{B} = \mathbf{C}$. ■

Finally, we return to the application of $\mathbf{A}^{-1}$ in the solution of $\mathbf{A}\mathbf{x} = \mathbf{c}$.

THEOREM 10.6.2 *Solution of* $\mathbf{A}\mathbf{x} = \mathbf{c}$
If $\mathbf{A}$ is $n \times n$ and $\det\mathbf{A} \neq 0$, then $\mathbf{A}\mathbf{x} = \mathbf{c}$ admits the unique solution $\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}$.

Proof: That $\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}$ satisfies $\mathbf{A}\mathbf{x} = \mathbf{c}$ was shown just above Theorem 10.6.1. That the solution is unique follows from Theorem 10.5.3 because $\det\mathbf{A} \neq 0$ implies that $r(\mathbf{A}|\mathbf{c}) = r(\mathbf{A}) = n$. ■

There are several useful properties of inverse matrices.

---

*Now we can understand that the failure in Example 1 occurred because

$$\begin{bmatrix} -1 & -2 \\ 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 \\ 1 & 1 & 2 \end{bmatrix} = \begin{bmatrix} -3 & -4 & -4 \\ 2 & 3 & 2 \\ 1 & 1 & 2 \end{bmatrix} \neq \mathbf{I}.$$

*There exist interesting generalizations of the notion of the inverse matrix for matrices that are not strictly invertible (perhaps not even square). Such are the *Moore–Penrose generalized inverse* and the *pseudoinverse*. See, for example, Gilbert Strang, *Linear Algebra and Its Applications* (New York: Academic Press, 1976), Chap. 3.

PROPERTIES OF INVERSES

I1. If $\mathbf{A}$ and $\mathbf{B}$ are of the same order, and invertible, then $\mathbf{AB}$ is too, and

$$\boxed{(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}.} \tag{28}$$

I2. If $\mathbf{A}$ is invertible, then

$$\boxed{\left(\mathbf{A}^{\mathrm{T}}\right)^{-1} = \left(\mathbf{A}^{-1}\right)^{\mathrm{T}}} \tag{29}$$

and

$$\det\left(\mathbf{A}^{-1}\right) = \frac{1}{\det\mathbf{A}}. \tag{30}$$

I3. If $\mathbf{A}$ is invertible, then $\left(\mathbf{A}^{-1}\right)^{-1} = \mathbf{A}$ and $(\mathbf{A}^m)^n = \mathbf{A}^{mn}$ for any (positive, negative, or zero) integers m and n.

I4. If $\mathbf{A}$ is invertible, then $\mathbf{AB} = \mathbf{AC}$ implies that $\mathbf{B} = \mathbf{C}$, $\mathbf{BA} = \mathbf{CA}$ implies that $\mathbf{B} = \mathbf{C}$, $\mathbf{AB} = \mathbf{0}$ implies that $\mathbf{B} = \mathbf{0}$, and $\mathbf{BA} = \mathbf{0}$ implies that $\mathbf{B} = \mathbf{0}$.

Of these, let us prove (28) and (29) and leave the remaining proofs as exercises. First (28): Since $\mathbf{A}$ and $\mathbf{B}$ are invertible, by assumption, $\det\mathbf{A} \neq 0$ and $\det\mathbf{B} \neq 0$. Thus, $\det(\mathbf{AB}) = (\det\mathbf{A})(\det\mathbf{B}) \neq 0$ so $\mathbf{AB}$ is invertible too. Let us denote $(\mathbf{AB})^{-1}$ as $\mathbf{C}$. Then $\mathbf{ABC} = \mathbf{I}$, $\mathbf{A}^{-1}\mathbf{ABC} = \mathbf{A}^{-1}\mathbf{I}$, $\mathbf{BC} = \mathbf{A}^{-1}$, $\mathbf{B}^{-1}\mathbf{BC} = \mathbf{B}^{-1}\mathbf{A}^{-1}$, hence, $\mathbf{C} = \mathbf{B}^{-1}\mathbf{A}^{-1}$, as claimed. As a mnemonic device, note the resemblence of (28) to the transpose formula $(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$.

To prove (29), begin with $\left(\mathbf{AA}^{-1}\right)^{\mathrm{T}} = \mathbf{I}^{\mathrm{T}} = \mathbf{I}$. But $\left(\mathbf{AA}^{-1}\right)^{\mathrm{T}} = \left(\mathbf{A}^{-1}\right)^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$. Hence, $\left(\mathbf{A}^{-1}\right)^{\mathrm{T}} = \left(\mathbf{A}^{\mathrm{T}}\right)^{-1}$.

10.6.2. Application to a mass-spring system. To illustrate a number of these ideas with a physical application, consider the arrangement of masses and springs shown in Fig. 1. The three masses are in static equilibrium under the action of

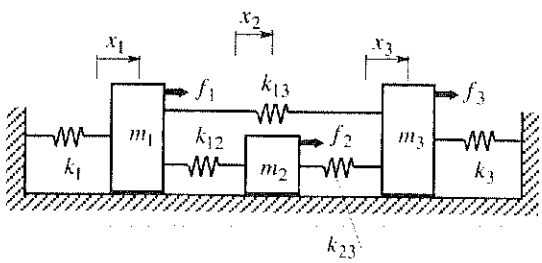


Figure 1. Mass-spring system.

prescribed applied forces f_1, f_2, f_3, and the k's denote the stiffnesses of the various springs. For instance, k_{12} denotes the stiffness of the spring connecting mass number 1 and mass number 2. Mass-spring systems are discussed in Section 1.3,

and in Example 3 of Section 3.9.1, which discussions should be reviewed if the following is not clear.

The free-body diagrams (i.e., the force diagrams) of the three masses are shown in Fig. 2, where it has been assumed, simply for definiteness, that $x_1 > x_2 > x_3 > 0$.

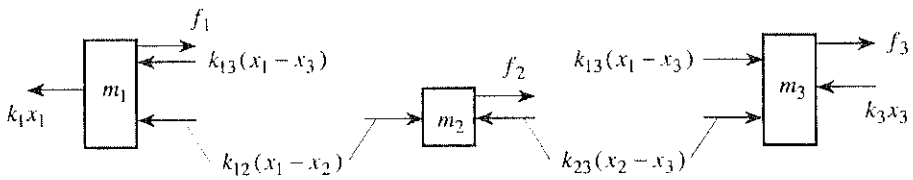


Figure 2. Free-body diagrams.

From that assumption it follows that each spring is in compression except for the left-hand spring of stiffness k_1.

From Fig. 2 and Newton's second law, we obtain the equations of motion

$$\begin{aligned} m_1x_1'' &= f_1 - k_1x_1 - k_{12}(x_1 - x_2) - k_{13}(x_1 - x_3), \\ m_2x_2'' &= f_2 + k_{12}(x_1 - x_2) - k_{23}(x_2 - x_3), \\ m_3x_3'' &= f_3 + k_{13}(x_1 - x_3) + k_{23}(x_2 - x_3) - k_3x_3, \end{aligned} \tag{31}$$

where primes denote differentiation with respect to the time t. Since the system is in static equilibrium by assumption, $x_1'' = x_2'' = x_3'' = 0$, and (31) becomes, in matrix form,

$$\begin{bmatrix} (k_1 + k_{12} + k_{13}) & -k_{12} & -k_{13} \\ -k_{12} & (k_{12} + k_{23}) & -k_{23} \\ -k_{13} & -k_{23} & (k_{13} + k_{23} + k_3) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} \tag{32}$$

or

$$\mathbf{K}\mathbf{x} = \mathbf{f}, \tag{33}$$

where we will call $\mathbf{K}$ the *stiffness matrix*. We see that (33) is a matrix generalization of the simple Hooke's law $f = kx$ for a single spring.

Is there a unique solution $\mathbf{x} = \mathbf{K}^{-1}\mathbf{f}$? Experience and physical intuition probably tells us there is. Mathematically, everything hinges upon $\det\mathbf{K}$. If $\det\mathbf{K} \neq 0$, there is a unique solution for $\mathbf{x}$, and if $\det\mathbf{K} = 0$, then there is either no solution or an infinity of solutions. With five parameters within $\mathbf{K}$ (namely, $k_1, k_{12}, k_{13}, k_{23}, k_3$) it is hard to imagine that we cannot have $\det\mathbf{K} = 0$ for some choice(s) of those parameters. Let us see. Working out the determinant of $\mathbf{K}$, we find that

$$\begin{aligned} \det\mathbf{K} = {} & k_1\left(k_{12}k_{13} + k_{12}k_{23} + k_{12}k_3 + k_{23}k_{13} + k_{23}k_3\right) \\ & + k_3\left(k_{12}k_{23} + k_{12}k_{13}\right). \end{aligned} \tag{34}$$

Since each sign is positive, and the k's are positive, we see that $\det\mathbf{K} \neq 0$ so there is indeed a unique solution for $\mathbf{x}$, namely, $\mathbf{x} = \mathbf{K}^{-1}\mathbf{f}$.

However, suppose we degrade the system by removing one or more springs. We can see from (34) that even if we set any one k value equal to zero (i.e., remove that spring), $\det\mathbf{K}$ is still positive. If we are willing to remove two springs, then we can obtain $\det\mathbf{K} = 0$ in either of two ways, by setting $k_1 = k_3 = 0$ or by setting $k_{12} = k_{23} = 0$. Let us consider the former, and leave the latter for the exercises.

With $k_1 = k_3 = 0$, $\mathbf{K}$ is singular (noninvertible) and (33) admits either no solution or an infinity of them. Which is it, and how is the result to be understood physically? Setting $k_1 = k_3 = 0$, (32) reduces to

$$\begin{bmatrix} (k_{12}+k_{13}) & -k_{12} & -k_{13} \\ -k_{12} & (k_{12}+k_{23}) & -k_{23} \\ -k_{13} & -k_{23} & (k_{13}+k_{23}) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} \tag{35}$$

and, in augmented matrix form, Gauss elimination gives

$$\begin{bmatrix} 1 & \dfrac{k_{23}}{k_{13}} & -1-\dfrac{k_{23}}{k_{13}} & -\dfrac{f_3}{k_{13}} \\ 0 & 1+\dfrac{k_{23}}{k_{12}}+\dfrac{k_{23}}{k_{13}} & -1-\dfrac{k_{23}}{k_{12}}-\dfrac{k_{23}}{k_{13}} & \dfrac{f_3}{k_{13}}-\dfrac{f_2}{k_{12}} \\ 0 & 0 & 0 & f_1+f_2+f_3 \end{bmatrix}, \tag{36}$$

which result reveals two possibilities.

(i) If $f_1+f_2+f_3 \neq 0$, then there is *no* solution. That mathematical result makes perfect sense physically, because with the end springs removed $f_1+f_2+f_3$ is the net lateral force on the three-mass system, and if that net force is nonzero, then the system cannot be in static equilibrium, as was assumed when we set $x_1'' = x_2'' = x_3'' = 0$ in (31)!

(ii) If $f_1+f_2+f_3 = 0$, then we see from (36) that there is an *infinity* of solutions, of the form

$$x_3 = \alpha, \quad x_2 = \alpha + \text{etc.}, \quad x_1 = \alpha + \text{etc.},$$

where α is an arbitrary constant and the two etc.'s involve the f's and k's. That is, the solution is nonunique because of the arbitrary *translation* α. Again, that result makes sense physically because with $k_1 = k_3 = 0$ there are no end springs to restrain the three-mass system laterally.

Let us make one more important point. Observe from (32) that the $\mathbf{K}$ matrix is symmetric. Yet the system (Fig. 1) is not *physically* symmetric; that is, in general, $k_1 \neq k_3$ and $k_{12} \neq k_{23}$. Thus, the mathematical symmetry is somewhat unexpected and mysterious. In fact, we state without proof that for *any* number of masses interconnected with springs the resulting $\mathbf{K}$ matrix will be symmetric.

There is a striking consequence of the symmetry of $\mathbf{K}$, which we now explain. Property I2 gives $(\mathbf{K}^{-1})^{\mathrm{T}} = (\mathbf{K}^{\mathrm{T}})^{-1} = \mathbf{K}^{-1}$ since $\mathbf{K}^{\mathrm{T}} = \mathbf{K}$. Let us denote

$\mathbf{K} = \{\alpha_{ij}\}$, say, and let us compare the displacement x_3 of m_3 due to a unit load $f_1 = 1$ on m_1 (with $f_2 = f_3 = 0$) with the displacement x_1 of m_1 due to a unit load $f_3 = 1$ on m_3 (with $f_1 = f_2 = 0$). In the first case,

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

gives $x_3 = \alpha_{31}$ and, in the second case,

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

gives $x_1 = \alpha_{13}$. But these are the same ($\alpha_{31} = \alpha_{13}$) because $\mathbf{K}$ is symmetric. In this manner we find that

$$\begin{pmatrix} \text{displacement } x_j \text{ of mass } m_j \\ \text{due to unit load on mass } m_k \end{pmatrix} = \begin{pmatrix} \text{displacement } x_k \text{ of mass } m_k \\ \text{due to unit load on mass } m_j \end{pmatrix}. \tag{37}$$

The latter "reciprocity" result can be generalized so as to apply to any linear elastic system and is known as *Maxwell reciprocity*.*

There is an electrical analog of the mechanical system shown in Fig. 1, a circuit containing resistors and voltage sources (such as batteries), and discussion of that case is left for the exercises.

10.6.3. Cramer's rule. We have seen that if $\mathbf{A}$ is $n \times n$ and $\det\mathbf{A} \neq 0$, then $\mathbf{Ax} = \mathbf{c}$ has the unique solution

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}. \tag{38}$$

To focus on the individual components of $\mathbf{x}$, rather than the entire $\mathbf{x}$ vector, let us write out (38):

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \vdots & \vdots & & \vdots \\ \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{nn} \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} \sum_j \alpha_{1j}c_j \\ \vdots \\ \sum_j \alpha_{nj}c_j \end{bmatrix}. \tag{39}$$

Equating the ith component on the left with the ith component on the right, we have the scalar statement

$$x_i = \sum_j \alpha_{ij}c_j \tag{40}$$

*By a linear elastic system is meant one obeying Hooke's law. For a statement and proof of Maxwell's reciprocity theorem see, for instance, Den Hartog's *Advanced Strength of Materials* (New York: McGraw–Hill, 1952).

for any desired i ($1 \le i \le n$). Or, recalling (17),

$$x_i = \sum_j \left(\frac{A_{ji}}{\det\mathbf{A}}\right) c_j = \frac{\sum_j A_{ji} c_j}{\det\mathbf{A}}. \tag{41}$$

Now, if the numerator on the right-hand side were $\sum_j A_{ji} a_{ji}$, instead, it would be recognizable as the determinant of $\mathbf{A}$, namely, the cofactor expansion about the ith column. But the a_{ji}'s are replaced, in (41), by the c_j's, so the numerator of (41) amounts to a determinant but not the determinant of $\mathbf{A}$; rather, it is the determinant of the $\mathbf{A}$ matrix with its ith column replaced by the column of c_j's (or the $\mathbf{c}$ vector, if you like).

The result, known as **Cramer's rule**, after *Gabriel Cramer* (1704–1752), is as follows.

THEOREM 10.6.3 *Cramer's Rule*
If $\mathbf{Ax} = \mathbf{c}$ where $\mathbf{A}$ is invertible, then each component x_i of $\mathbf{x}$ may be computed as the ratio of two determinants; the denominator is $\det\mathbf{A}$, and the numerator is also the determinant of the $\mathbf{A}$ matrix but with the ith column replaced by $\mathbf{c}$.

EXAMPLE 3. Let us solve the system

$$\begin{bmatrix} 1 & 3 & 0 \\ -2 & 3 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 1 \\ -2 \end{bmatrix} \tag{42}$$

for x_1 and x_2, say, using Cramer's rule. In this case $\det\mathbf{A} = 8 \neq 0$ so the method is, first of all, applicable. Thus,

$$x_1 = \frac{\begin{vmatrix} \mathbf{5} & 3 & 0 \\ \mathbf{1} & 3 & 1 \\ \mathbf{-2} & 1 & 1 \end{vmatrix}}{\begin{vmatrix} 1 & 3 & 0 \\ -2 & 3 & 1 \\ 0 & 1 & 1 \end{vmatrix}} = \frac{1}{8} \quad \text{and} \quad x_2 = \frac{\begin{vmatrix} 1 & \mathbf{5} & 0 \\ -2 & \mathbf{1} & 1 \\ 0 & \mathbf{-2} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & 3 & 0 \\ -2 & 3 & 1 \\ 0 & 1 & 1 \end{vmatrix}} = \frac{13}{8}, \tag{43}$$

where we have printed the "replacement columns" as boldface, for emphasis. ∎

Cramer's rule, like the inverse matrix solution (38) from which it comes, has the advantage of being an explicit *formula*, rather than a *method*. It is also useful in that it permits us to focus on any single component of $\mathbf{x}$ without having to compute the entire $\mathbf{x}$ vector.

10.6.4. Evaluation of $\mathbf{A}^{-1}$ by elementary row operations. Equation (16) gives $\mathbf{A}^{-1}$ in terms of $\det\mathbf{A}$ and n^2 cofactors, each of which is ± 1 times an $(n-1) \times$

$(n-1)$ minor determinant. Each of these determinants can be evaluated by the cofactor expansion definition or, especially if n is large, by a faster method – such as triangularization. Alternatively, we can bypass (16) altogether, and determine $\mathbf{A}^{-1}$ efficiently as follows.

Whether we are seeking $\mathbf{A}^{-1}$ in order to solve a system $\mathbf{Ax} = \mathbf{c}$, or whether we are simply seeking the inverse of a given matrix $\mathbf{A}$, observe that if we solve a system $\mathbf{Ax} = \mathbf{c}$ of n equations in n unknowns, or equivalently $\mathbf{Ax} = \mathbf{Ic}$, by Gauss–Jordan reduction, the result is the form $\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}$, or equivalently $\mathbf{Ix} = \mathbf{A}^{-1}\mathbf{c}$. Symbolically, then, the sequence of elementary row operations effects the following transformation:

$$\begin{array}{c} \mathbf{Ax} = \mathbf{Ic} \\ \downarrow \\ \mathbf{Ix} = \mathbf{A}^{-1}\mathbf{c}. \end{array} \tag{44}$$

That is, at the same time that the row operations are transforming $\mathbf{A}$ to $\mathbf{I}$ they are also transforming $\mathbf{I}$ to $\mathbf{A}^{-1}$. Thus, we can skip $\mathbf{x}$ and $\mathbf{c}$ altogether, put $\mathbf{A}$ and $\mathbf{I}$ "side by side" as an augmented matrix $\mathbf{A}|\mathbf{I}$, and carry out elementary row operations on $\mathbf{A}|\mathbf{I}$ so as to reduce $\mathbf{A}$, on the left, to $\mathbf{I}$. When that has been accomplished, the matrix on the right will be $\mathbf{A}^{-1}$.

EXAMPLE 4. To illustrate, let us find the inverse of

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & 0 \\ -2 & 3 & 1 \\ 0 & 1 & 1 \end{bmatrix}. \tag{45}$$

Then

$$\begin{aligned}
\mathbf{A}|\mathbf{I} &= \left[\begin{array}{ccc|ccc} 1 & 3 & 0 & 1 & 0 & 0 \\ -2 & 3 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right] \to \left[\begin{array}{ccc|ccc} 1 & 3 & 0 & 1 & 0 & 0 \\ 0 & 9 & 1 & 2 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right] \\
&\to \left[\begin{array}{ccc|ccc} 1 & 3 & 0 & 1 & 0 & 0 \\ 0 & 9 & 1 & 2 & 1 & 0 \\ 0 & 0 & \frac{8}{9} & -\frac{2}{9} & -\frac{1}{9} & 1 \end{array}\right] \to \left[\begin{array}{ccc|ccc} 1 & 3 & 0 & 1 & 0 & 0 \\ 0 & 9 & 1 & 2 & 1 & 0 \\ 0 & 0 & 1 & -\frac{1}{4} & -\frac{1}{8} & \frac{9}{8} \end{array}\right] \\
&\to \left[\begin{array}{ccc|ccc} 1 & 3 & 0 & 1 & 0 & 0 \\ 0 & 9 & 0 & \frac{9}{4} & \frac{9}{8} & -\frac{9}{8} \\ 0 & 0 & 1 & -\frac{1}{4} & -\frac{1}{8} & \frac{9}{8} \end{array}\right] \to \left[\begin{array}{ccc|ccc} 1 & 3 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & \frac{1}{4} & \frac{1}{8} & -\frac{1}{8} \\ 0 & 0 & 1 & -\frac{1}{4} & -\frac{1}{8} & \frac{9}{8} \end{array}\right] \\
&\to \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & \frac{1}{4} & -\frac{3}{8} & \frac{3}{8} \\ 0 & 1 & 0 & \frac{1}{4} & \frac{1}{8} & -\frac{1}{8} \\ 0 & 0 & 1 & -\frac{1}{4} & -\frac{1}{8} & \frac{9}{8} \end{array}\right]
\end{aligned} \tag{46}$$

so

$$\mathbf{A}^{-1} = \begin{bmatrix} \frac{1}{4} & -\frac{3}{8} & \frac{3}{8} \\ \frac{1}{4} & \frac{1}{8} & -\frac{1}{8} \\ -\frac{1}{4} & -\frac{1}{8} & \frac{9}{8} \end{bmatrix}. \quad \blacksquare \tag{47}$$

Can this method *fail* to work? Yes indeed, it had *better* fail if $\det\mathbf{A} = 0$, for then $\mathbf{A}$ is not invertible. The way that circumstance would show up is that the elementary row operations would produce one or more rows of zeros on the left so that $\mathbf{A}$ cannot be converted to $\mathbf{I}$.

10.6.5. LU-factorization. This final subsection is not really about the inverse matrix or about the inverse matrix method of solving $\mathbf{Ax} = \mathbf{c}$. Rather, it is about an alternative method of solution that is based upon the factorization of an $n \times n$ matrix $\mathbf{A}$ as a lower triangular matrix $\mathbf{L}$ times an upper triangular matrix $\mathbf{U}$:

$$\mathbf{A} = \mathbf{LU} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}, \tag{48}$$

where we have taken $n = 3$ simply for compactness. If we carry out the multiplication on the right and equate the nine elements of LU to the corresponding elements of $\mathbf{A}$ we obtain nine equations in the 12 unknown l_{ij}'s and u_{ij}'s. Since we have more unknowns than equations, there is some flexibility in implementing the idea. Hence there are various versions of LU-factorization.

According to **Doolittle's method** we can set each $l_{jj} = 1$ in $\mathbf{L}$ (i.e., the diagonal elements) and solve uniquely for the remaining l_{ij}'s and the u_{ij}'s. With $\mathbf{L}$ and $\mathbf{U}$ determined, we then solve $\mathbf{Ax} = \mathbf{LUx} = \mathbf{c}$ by setting $\mathbf{Ux} = \mathbf{y}$ so that $\mathbf{L}(\mathbf{Ux}) = \mathbf{c}$ breaks into the two problems

$$\mathbf{Ly} = \mathbf{c}, \tag{49a}$$
$$\mathbf{Ux} = \mathbf{y}, \tag{49b}$$

each of which is simple because $\mathbf{L}$ and $\mathbf{U}$ are triangular. We solve (49a) for $\mathbf{y}$, put that $\mathbf{y}$ into (49b), and then solve (49b) for $\mathbf{x}$. Let us illustrate the procedure.

EXAMPLE 5. To solve

$$\begin{bmatrix} 2 & -3 & 3 \\ 6 & -8 & 7 \\ -2 & 6 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -2 \\ -3 \\ 3 \end{bmatrix} \tag{50}$$

by the Doolittle LU-factorization method, we first need to determine $\mathbf{L}$ and $\mathbf{U}$ by equating

$$\begin{aligned} \begin{bmatrix} 2 & -3 & 3 \\ 6 & -8 & 7 \\ -2 & 6 & -1 \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix} \\ &= \begin{bmatrix} \underline{u_{11}} & \underline{u_{12}} & \underline{u_{13}} \\ \underline{l_{21}}u_{11} & l_{21}u_{12} + \underline{u_{22}} & l_{21}u_{13} + \underline{u_{23}} \\ \underline{l_{31}}u_{11} & l_{31}u_{12} + \underline{l_{32}}u_{22} & l_{31}u_{13} + l_{32}u_{23} + \underline{u_{33}} \end{bmatrix}. \end{aligned} \tag{51}$$

Matching $a_{11}, a_{12}, a_{13}, a_{21}, \ldots, a_{32}, a_{33}$ to the corresponding terms on the right gives a sequence of equations for $u_{11}, u_{12}, u_{13}, l_{21}, u_{22}, u_{23}, l_{31}, l_{32}$, and u_{33} (i.e., the underlined

entries) in turn:

$$
\begin{aligned}
u_{11} &= 2, \\
u_{12} &= -3, \\
u_{13} &= 3, \\
l_{21} &= 6/u_{11} = 6/2 = 3, \\
u_{22} &= -8 - l_{21}u_{12} = -8 - (3)(-3) = 1, \\
u_{23} &= 7 - l_{21}u_{13} = 7 - (3)(3) = -2, \\
l_{31} &= -2/u_{11} = -2/2 = -1, \\
l_{32} &= (6 - l_{31}u_{12})/u_{22} = [6 - (-1)(-3)]/1 = 3, \\
u_{33} &= -1 - l_{31}u_{13} - l_{32}u_{23} = -1 - (-1)(3) - (3)(-2) = 8.
\end{aligned} \tag{52}
$$

Then (49a) becomes

$$
\begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ -1 & 3 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} -2 \\ -3 \\ 3 \end{bmatrix},
$$

which gives $\mathbf{y} = [-2, 3, -8]^{\mathrm{T}}$. Finally, (49a) becomes

$$
\begin{bmatrix} 2 & -3 & 3 \\ 0 & 1 & -2 \\ 0 & 0 & 8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -2 \\ -3 \\ 3 \end{bmatrix},
$$

which gives the final solution $\mathbf{x} = [2, 1, -1]^{\mathrm{T}}$. ∎

The beauty of the method is that the l_{ij}'s and u_{ij}'s are found not by solving simultaneous coupled equations but by solving a sequence of linear equations in only one unknown [as illustrated in (52)]. With $\mathbf{L}$ and $\mathbf{U}$ thus determined, the solution of (49a) and (49b) is likewise simple since $\mathbf{L}$ and $\mathbf{U}$ are triangular. In fact, the method is around twice as fast as Gauss–Jordan elimination.

Closure. The inverse of a matrix $\mathbf{A}$, denoted as $\mathbf{A}^{-1}$, exists if and only if $\mathbf{A}$ is square $(n \times n)$ and $\det\mathbf{A} \neq 0$. If it exists it is given uniquely as

$$
\mathbf{A}^{-1} = \frac{1}{\det\mathbf{A}} \operatorname{adj}\mathbf{A}, \tag{53}
$$

where the matrix $\operatorname{adj}\mathbf{A}$ is the adjoint of $\mathbf{A}$ (the transpose of the cofactor matrix), and is such that

$$
\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}. \tag{54}
$$

The case where $\mathbf{A}$ is not invertible (i.e., is singular) is the exceptional case; in the generic case a given $n \times n$ matrix is invertible. Besides (54), several useful properties of inverses are given as I1–I4 in Section 10.6.1.

If $\mathbf{A}$ is invertible, then $\mathbf{A}\mathbf{x} = \mathbf{c}$ admits the unique solution

$$
\mathbf{x} = \mathbf{A}^{-1}\mathbf{c} = \frac{1}{\det\mathbf{A}} (\operatorname{adj}\mathbf{A})\, \mathbf{c}, \tag{55}
$$

which result gives us Cramer's rule, whereby each component of $\mathbf{x}$ is expressed as the ratio of $n \times n$ determinants.

Notice that the equation $ax = c$ has a unique solution if and only if $a \neq 0$. For $\mathbf{Ax} = \mathbf{c}$, where $\mathbf{A}$ is $n \times n$, that condition generalizes not to $\mathbf{A} \neq \mathbf{0}$ but to $\det\mathbf{A} \neq 0$.

In contrast with Gauss–Jordan reduction and LU-factorization, which are solution *methods*, (55) and Cramer's rule are explicit *formulas* for the solution (when a unique solution does exist).

Finally, we urge you to be careful with the sequencing of matrices because of the general absence of commutativity under multiplication. For instance, $\mathbf{Ax} = \mathbf{c}$ implies $\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}$ (if $\mathbf{A}$ is invertible), NOT $\mathbf{x} = \mathbf{cA}^{-1}$. Indeed, the product $\mathbf{cA}^{-1}$ is not even defined (unless $n = 1$) since $\mathbf{c}$ is $n \times 1$ and $\mathbf{A}^{-1}$ is $n \times n$.

Computer software. Using *Maple*, the relevant command is **inverse(A)**, within the linalg package. For instance, to invert

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix},$$

enter

```
with(linalg):
```

and return. Then the steps

```
A := array ([[1,2], [3,4]]):
```

and

```
inverse(A);
```

give the result

$$\begin{bmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{bmatrix}$$

EXERCISES 10.6

1. Use (17) to evaluate the inverse matrix. If the matrix is not invertible, state that.

(a) $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, where $ad - bc \neq 0$

(b) $\begin{bmatrix} 5 & 4 \\ 3 & 2 \end{bmatrix}$

(c) $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

(d) $\begin{bmatrix} 2 & 1 \\ 2 & -1 \end{bmatrix}$

(e) $\begin{bmatrix} 3 & 0 \\ 0 & 4 \end{bmatrix}$

(f) $\begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 3 \\ 0 & 3 & -1 \end{bmatrix}$

(g) $\begin{bmatrix} 0 & 1 & 0 \\ 2 & 0 & 5 \\ 0 & 0 & 3 \end{bmatrix}$

(h) $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$

(i) $\begin{bmatrix} 0 & 0 & 2 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$

(j) $\begin{bmatrix} 3 & 1 & -2 \\ 1 & 2 & 1 \\ 1 & -3 & -3 \end{bmatrix}$

(k) $\begin{bmatrix} 7 & 1 & 3 \\ 2 & -1 & 1 \\ 0 & 1 & 4 \end{bmatrix}$

(l) $\begin{bmatrix} 1 & 0 & 2 \\ 0 & 3 & 0 \\ 4 & 0 & 5 \end{bmatrix}$ (m) $\begin{bmatrix} 2 & 3 & -5 \\ 12 & 1 & 0 \\ 3 & 4 & 1 \end{bmatrix}$

(n) $\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 4 & 1 \end{bmatrix}$ (o) $\begin{bmatrix} 2 & 3 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 4 & 1 \end{bmatrix}$

(p) $\begin{bmatrix} 7 & 1 & 3 & 0 \\ 2 & -1 & 1 & 0 \\ 0 & 1 & 4 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ (q) $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$

(r) $\begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$ (s) $\begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix}$

2. (c)–(o) Evaluate the inverse of the matrix given in the corresponding part of Exercise 1 using elementary row operations, as we did in Example 4.

3. (c)–(o) Evaluate the inverse of the matrix given in the corresponding part of Exercise 1 using computer software.

4. (*Block-diagonal matrices*) If an $n \times n$ matrix $\mathbf{A}$ can be partitioned as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & & \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & & \cdots & \mathbf{A}_k \end{bmatrix} \qquad (k \le n)$$

it is said to be **block diagonal**. All of the $\mathbf{A}_j$ submatrices need to be square, although not necessarily of equal order, with their main diagonals coinciding with the main diagonal of $\mathbf{A}$. For instance,

$$\mathbf{A} = \left[\begin{array}{cc|ccc|c} 2 & 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & -1 & 0 & 3 & 0 \\ 0 & 0 & 2 & 5 & 1 & 0 \\ 0 & 0 & 3 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 4 \end{array}\right] = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_3 \end{bmatrix} \tag{4.1}$$

is block diagonal. Such matrices exhibit essentially the same simple features as diagonal matrices.

(a) Show that $\mathbf{A}$ is invertible if and only if $\mathbf{A}_1, \ldots, \mathbf{A}_k$ are. HINT: Recall equation (9.1) in Exercise 9, Section 10.4.

(b) Assuming that $\mathbf{A}_1, \ldots, \mathbf{A}_k$ are invertible, verify that

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}_1^{-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2^{-1} & & \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & & \cdots & \mathbf{A}_k^{-1} \end{bmatrix}.$$

(c) Use the latter result to evaluate $\mathbf{A}^{-1}$, where $\mathbf{A}$ is given in (4.1).

5. Solve for x_1 and x_2 by Cramer's rule.

(a) $\begin{aligned} x_1 + 4x_2 &= 0 \\ 3x_1 - x_2 &= 6 \end{aligned}$

(b) $\begin{aligned} ax_1 + bx_2 &= c \\ dx_1 + ex_2 &= f \end{aligned}$

(c) $\begin{aligned} x_1 - 2x_2 + x_3 &= 4 \\ 2x_1 + 3x_2 + x_3 &= -7 \\ 4x_1 + x_2 + 2x_3 &= 0 \end{aligned}$

(d) $\begin{aligned} x_1 + 2x_2 + 3x_3 &= 9 \\ x_1 + 4x_2 \qquad &= 6 \\ x_1 \qquad - 5x_3 &= 2 \end{aligned}$

(e) $\begin{aligned} 2x_1 + x_2 \qquad\qquad &= 1 \\ x_1 + 2x_2 + x_3 \qquad &= 0 \\ x_2 + 2x_3 + x_4 &= 0 \\ x_3 + 2x_4 &= 0 \end{aligned}$

(f) $\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ x_1 + 2x_2 + 3x_3 &= 0 \\ x_1 - x_2 + 4x_3 &= 0 \end{aligned}$

6. (a) Given a certain 3×3 matrix $\mathbf{A}$, we find its inverse to be

$$\mathbf{A}^{-1} = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 0 \end{bmatrix}.$$

Can that result be correct? Explain.

(b) Same as (a), for

$$\mathbf{A}^{-1} = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 3 & 0 & 0 \end{bmatrix}.$$

7. If $\mathbf{A}^{-1}$ is the given matrix, find $\mathbf{A}$.

(a) $\begin{bmatrix} 3 & -1 \\ 3 & -2 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 2 & 1 \\ 0 & 3 & 1 \\ 1 & 1 & 0 \end{bmatrix}$ (d) $\begin{bmatrix} 2 & 1 & 1 \\ 4 & 1 & 1 \\ 6 & 2 & 1 \end{bmatrix}$

8. Suppose that $\mathbf{Ax} = \mathbf{c}$ is a linear system of order 3, and that to the $\mathbf{c}$ vectors

$$\mathbf{c} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

there correspond unique solutions

$$\mathbf{x} = \begin{bmatrix} 2 \\ 5 \\ -1 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}, \quad \begin{bmatrix} 3 \\ 0 \\ 4 \end{bmatrix},$$

respectively. Then what is the solution $\mathbf{x}$ corresponding to $\mathbf{c} = [4, 3, -1]^{\mathrm{T}}$? Is it unique? Explain.

9. Suppose that Gauss elimination gives the solution of a linear system $\mathbf{Ax} = \mathbf{c}$ as $\mathbf{x} = \mathbf{x}_0 + \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2$, where $\mathbf{A}$ is 6×6 and α_1 and α_2 are arbitrary. Is $\mathbf{A}$ invertible? Explain.

10. (*Nilpotent matrices*) If there is some positive integer p such that $\mathbf{A}^p = \mathbf{0}$, then $\mathbf{A}$ said to be **nilpotent** (i.e., potentially nil).

(a) Show that a nilpotent matrix is necessarily singular.
(b) If $\mathbf{A}$ is nilpotent, with $\mathbf{A}^p = \mathbf{0}$, show that

$$(\mathbf{I} - \mathbf{A})^{-1} = \mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \cdots + \mathbf{A}^{p-1}. \qquad (10.1)$$

11. First, read Exercise 10. Use (10.1) to find the inverse of the given matrix. HINT: You will need to identify $\mathbf{A}$.

(a) $\begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 8 \\ 0 & 0 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 2 & 7 & 1 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 5 & 1 & 0 \\ 0 & 1 & 2 & 7 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 0 & 0 & 0 \\ -4 & 1 & 0 & 0 \\ 0 & 3 & 1 & 0 \\ 2 & 1 & 6 & 1 \end{bmatrix}$

12. First, read Exercise 10. Use (10.1) to find the inverse of the given matrix

(a) $\begin{bmatrix} 2 & 8 & 10 \\ 0 & 3 & 12 \\ 0 & 0 & 4 \end{bmatrix}$ HINT:

$$\begin{bmatrix} 2 & 8 & 10 \\ 0 & 3 & 12 \\ 0 & 0 & 4 \end{bmatrix}^{-1} = \left[\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix}\begin{bmatrix} 1 & 4 & 5 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix}\right]^{-1}$$
$$= \begin{bmatrix} 1 & 4 & 5 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix}^{-1}\begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{4} \end{bmatrix}.$$

(b) $\begin{bmatrix} 3 & 0 & 0 \\ 4 & -2 & 0 \\ 10 & 0 & 2 \end{bmatrix}$ (c) $\begin{bmatrix} 2 & 0 & 0 \\ 4 & 2 & 0 \\ 1 & 6 & 2 \end{bmatrix}$

13. (*Ill-conditioned systems*) Consider the system

$$\begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \quad \text{or} \quad \mathbf{Ax} = \mathbf{c}. \qquad (13.1)$$

Here $\mathbf{A}$ is a third-order **Hilbert matrix**, named after *David Hilbert* (1862–1943).

(a) Evaluate $\mathbf{A}^{-1}$, by any analytical means that you wish, and show that $\mathbf{x} = \mathbf{A}^{-1}\mathbf{c} = [-3, -12, 30]^{\mathrm{T}}$.
(b) To simulate the effects of roundoff error, consider in place of (13.1) the rounded off system

$$\begin{bmatrix} 1 & 0.5 & 0.33 \\ 0.5 & 0.33 & 0.25 \\ 0.33 & 0.25 & 0.2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}. \qquad (13.2)$$

Solving (13.2) by any means you wish (computer software being the easiest), show that the solution of (13.2) is $\mathbf{x} \approx [11.1, -84.1, 96.8]^{\mathrm{T}}$. NOTE: In this example we see that only a slight error in $\mathbf{A}$ leads to a disproportionately large error in the solution. Hence, (13.1) is said to be **ill-conditioned**. (Ill-conditioned systems are also mentioned in Exercise 13, Section 8.3.) In applications it is important to know if a given system is ill-conditioned so that steps can be taken to obtain a sufficiently accurate solution. According to one criterion in the literature, an $n \times n$ matrix $\mathbf{A}$ may be considered as ill-conditioned if

$$\frac{|\det \mathbf{A}|}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2}} \ll 1. \qquad (13.3)$$

For the Hilbert matrix in (13.1), the left-hand side of (13.3) is 0.00033, which is indeed much smaller than unity.
(c) In place of (13.2), use the more accurate rounded off system

$$\begin{bmatrix} 1 & 0.5 & 0.333 \\ 0.5 & 0.333 & 0.25 \\ 0.333 & 0.25 & 0.2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \qquad (13.4)$$

and see how much closer the solution of (13.4) comes to the exact solution $\mathbf{x} = [-3, -12, 30]^{\mathrm{T}}$ of (13.1).

14. Prove property I3.

15. Prove property I4.

16. In Section 10.6.2 we stated that if we are willing to remove two springs then we can have $\det\mathbf{K} = 0$ either by setting $k_1 = k_3 = 0$ or by setting $k_{12} = k_{23} = 0$. We discuss the former choice, $k_1 = k_3 = 0$, both mathematically and physically as well. Do the same for the latter choice, $k_{12} = k_{23} = 0$.

17. (*A dc circuit*) Application of Kirchhoff's laws to the circuit shown in Exercise 10 of Section 8.3 produced the five equations (10.1) on the currents i_1, i_2, i_3. Of these equations, the second is the same as the first, and the fifth is the fourth minus the third. Thus, deleting those two redundant equations leaves the system

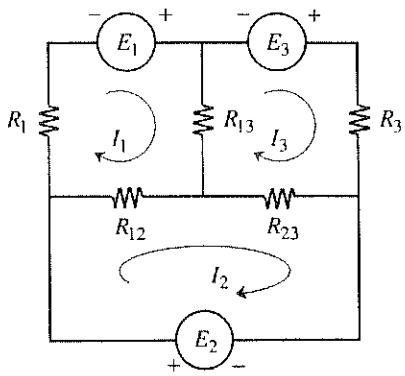
$$\begin{aligned} i_1 - \quad i_2 - \quad i_3 &= 0 \\ R_2 i_2 - R_3 i_3 &= 0 \\ R_1 i_1 + R_2 i_2 \qquad &= E. \end{aligned} \tag{17.1}$$

(a) Show that the determinant of the coefficient matrix in (17.1) is necessarily nonzero, so that the system $\mathbf{Ri} = \mathbf{e}$ given by (17.1) necessarily admits the unique solution $\mathbf{i} = \mathbf{R}^{-1}\mathbf{e}$. NOTE: $R_1 > 0, R_2 > 0$, and $R_3 > 0$.
(b) Solve for $\mathbf{i}$ by the inverse matrix method. Also, solve for i_1, i_2, i_3 by Cramer's rule and verify that the results are the same.
(c) Suppose, instead, that we allow one or more of the resistances to be zero so that $R_1 \geq 0, R_2 \geq 0, R_3 \geq 0$. Show that if any two, or all three, of the resistances are zero, then the determinant *does* vanish so that equations (17.1) admit either no solution or an infinity of solutions. For each of these four "singular" cases determine whether there is no solution or an infinity of solutions by applying Gauss elimination. State the *physical* significance of each of these results insofar as possible.

18. (*Circuit analog*) The electrical circuit analog of the mass-spring system shown in Fig. 1 is shown below.
(a) Applying Kirchhoff's voltage law, show that

$$\begin{bmatrix} (R_1 + R_{12} + R_{13}) & -R_{12} & -R_{13} \\ -R_{12} & (R_{12} + R_{23}) & -R_{23} \\ -R_{13} & -R_{23} & (R_{13} + R_{23} + R_3) \end{bmatrix} \begin{bmatrix} I_1 \\ I_2 \\ I_3 \end{bmatrix} = \begin{bmatrix} E_1 \\ E_2 \\ E_3 \end{bmatrix}, \tag{18.1}$$

which is the analog of (32) under the correspondence $R_{ij} \leftrightarrow k_{ij}$, $I_j \leftrightarrow x_j$, $E_j \leftrightarrow f_j$.
(b) Discuss the existence and uniqueness of solutions of (18.1) in the same way that we did that for the mass-spring system in Section 10.6.2, including a reciprocity result analogous to (37).



19. Solve by the Doolittle LU-factorization method.

(a) $\begin{bmatrix} 2 & 3 \\ 8 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -4 \\ 10 \end{bmatrix}$

(b) $\begin{bmatrix} 2 & -1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 7 \\ 13 \end{bmatrix}$

(c) $\begin{bmatrix} 2 & 5 & 1 \\ 2 & 8 & 0 \\ 8 & 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ -7 \\ 10 \end{bmatrix}$

(d) $\begin{bmatrix} 1 & 3 & -1 \\ 2 & 2 & 0 \\ 3 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 6 \\ 12 \end{bmatrix}$

20. (*Dual or reciprocal set*) In Exercise 13 of Section 9.9 we introduced the concept of a set of dual or reciprocal vectors $\{\mathbf{e}_1^*, \ldots, \mathbf{e}_n^*\}$ corresponding to a basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ that is not necessarily orthogonal, or ON. Having learned more about the solution of systems of linear algebraic equations in Sections 10.5 and 10.6, we can now return to that exercise and prove the claim made in part (a). Specifically, prove the dual set exists, is unique, and is itself a basis for $\mathbb{R}^n$.

10.7 Change of Basis (Optional)

In a given problem one selects what appears to be the most convenient basis, but at some stage of the analysis it may be desirable to switch to some other basis. For example, in studying the aerodynamics of a propeller it is generally most convenient to carry out the analysis (of the propeller-induced pressure field, for example) with respect to a propeller-fixed basis, one that rotates with the propeller, although eventually we may wish to relate quantities back to a stationary (nonrotating) basis. How do the coordinates (i.e., the components) of a given vector change as we change the basis? It is that question which we address in this section.

Let $B = \{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ be a given basis for the vector space V under consideration so that any given vector $\mathbf{x}$ in V can be expanded as

$$\mathbf{x} = x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n. \tag{1}$$

If we switch to some other basis $B' = \{\mathbf{e}'_1, \ldots, \mathbf{e}'_n\}$, then we may, similarly, expand the same vector $\mathbf{x}$ as

$$\mathbf{x} = x'_1\mathbf{e}'_1 + \cdots + x'_n\mathbf{e}'_n. \tag{2}$$

How are the x'_j coordinates related to the x_j coordinates? Since B' is a basis, we may expand each of the $\mathbf{e}_j$'s in terms of B':

$$\begin{aligned} \mathbf{e}_1 &= q_{11}\mathbf{e}'_1 + \cdots + q_{n1}\mathbf{e}'_n, \\ &\vdots \\ \mathbf{e}_n &= q_{1n}\mathbf{e}'_1 + \cdots + q_{nn}\mathbf{e}'_n. \end{aligned} \tag{3}$$

Putting (3) into (1) gives

$$\begin{aligned} \mathbf{x} &= x_1\left(q_{11}\mathbf{e}'_1 + \cdots + q_{n1}\mathbf{e}'_n\right) + \cdots + x_n\left(q_{1n}\mathbf{e}'_1 + \cdots + q_{nn}\mathbf{e}'_n\right) \\ &= \left(x_1q_{11} + \cdots + x_nq_{1n}\right)\mathbf{e}'_1 + \cdots + \left(x_1q_{n1} + \cdots + x_nq_{nn}\right)\mathbf{e}'_n, \end{aligned} \tag{4}$$

and a comparison of (2) and (4) gives the desired relations

$$\begin{aligned} x'_1 &= q_{11}x_1 + \cdots + q_{1n}x_n, \\ &\vdots \\ x'_n &= q_{n1}x_1 + \cdots + q_{nn}x_n \end{aligned} \tag{5}$$

or, in matrix notation,

$$\boxed{[\mathbf{x}]_{B'} = \mathbf{Q}\,[\mathbf{x}]_B\,,} \tag{6}$$

where

$$\mathbf{Q} = \begin{bmatrix} q_{11} & \cdots & q_{1n} \\ \vdots & & \vdots \\ q_{n1} & \cdots & q_{nn} \end{bmatrix} \tag{7}$$

and

$$[\mathbf{x}]_B = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad [\mathbf{x}]_{B'} = \begin{bmatrix} x'_1 \\ \vdots \\ x'_n \end{bmatrix}. \tag{8}$$

We call $[\mathbf{x}]_B$ the **coordinate vector** of the vector $\mathbf{x}$ with respect to the ordered basis B, and similarly for $[\mathbf{x}]_{B'}$, and we call $\mathbf{Q}$ the **coordinate transformation matrix** from B to B'.

Thus far, our results apply whether the bases are orthogonal or not. *In the remainder of this section we assume that both bases, B and B', are ON.* Thus, let us rewrite (3) as

$$\begin{aligned} \hat{\mathbf{e}}_1 &= q_{11}\hat{\mathbf{e}}'_1 + \cdots + q_{n1}\hat{\mathbf{e}}'_n, \\ &\vdots \\ \hat{\mathbf{e}}_n &= q_{1n}\hat{\mathbf{e}}'_1 + \cdots + q_{nn}\hat{\mathbf{e}}'_n, \end{aligned} \tag{9}$$

where the carets denote unit vectors, as usual. If we dot $\hat{\mathbf{e}}'_1$ into both sides of the first equation in (9), and remember that B' is ON, we obtain $q_{11} = \hat{\mathbf{e}}'_1 \cdot \hat{\mathbf{e}}_1$. Dotting $\hat{\mathbf{e}}'_2$ gives $q_{21} = \hat{\mathbf{e}}'_2 \cdot \hat{\mathbf{e}}_1, \ldots$, dotting $\hat{\mathbf{e}}'_n$ gives $q_{n1} = \hat{\mathbf{e}}'_n \cdot \hat{\mathbf{e}}_1$, and similarly for the second through nth equation in (9). The result is the formula

$$\boxed{q_{ij} = \hat{\mathbf{e}}'_i \cdot \hat{\mathbf{e}}_j,} \tag{10}$$

which tells us how to compute the transformation matrix $\mathbf{Q}$.

There are two properties of the $\mathbf{Q}$ matrix to address before turning to an example. To obtain the first of these, observe that

$$\begin{aligned} \mathbf{Q}^{\mathrm{T}}\mathbf{Q} &= \begin{bmatrix} q_{11} & \cdots & q_{n1} \\ \hline q_{12} & \cdots & q_{n2} \\ \hline & \vdots & \\ \hline q_{1n} & \cdots & q_{n2} \end{bmatrix} \left[\begin{array}{c|c|c|c} q_{11} & q_{12} & & q_{1n} \\ \vdots & \vdots & \cdots & \vdots \\ q_{n1} & q_{n2} & & q_{nn} \end{array}\right] \\ &= \begin{bmatrix} \hat{\mathbf{e}}_1 \cdot \hat{\mathbf{e}}_1 & \cdots & \hat{\mathbf{e}}_1 \cdot \hat{\mathbf{e}}_n \\ \vdots & & \vdots \\ \hat{\mathbf{e}}_n \cdot \hat{\mathbf{e}}_1 & \cdots & \hat{\mathbf{e}}_n \cdot \hat{\mathbf{e}}_n \end{bmatrix} = \mathbf{I} \end{aligned} \tag{11}$$

so that

$$\boxed{\mathbf{Q}^{-1} = \mathbf{Q}^{\mathrm{T}}.} \tag{12}$$

It was useful to partition the $\mathbf{Q}^{\mathrm{T}}$ and $\mathbf{Q}$ matrices in (11) because we can see from (9) that the columns of $\mathbf{Q}$ (and hence the rows of $\mathbf{Q}^{\mathrm{T}}$) are actually the $\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_n$ vectors (in n-tuple form). Thus, from the way matrix multiplication is defined, we can see that the elements of the product matrix are dot products. Specifically, the i, j element of $\mathbf{Q}^{\mathrm{T}}\mathbf{Q}$ is $\hat{\mathbf{e}}_i \cdot \hat{\mathbf{e}}_j$, which is the Kronecker delta δ_{ij}. Hence, $\mathbf{Q}^{\mathrm{T}}\mathbf{Q}$ equals the identity matrix $\mathbf{I}$, so $\mathbf{Q}^{\mathrm{T}}$ must be the inverse of $\mathbf{Q}$, as stated in (12).

That result makes it easy for us to reverse equation (6) – that is, to solve for $[\mathbf{x}]_B$ in terms of $[\mathbf{x}]_{B'}$ for then $[\mathbf{x}]_B = \mathbf{Q}^{-1}[\mathbf{x}]_{B'} = \mathbf{Q}^{\mathrm{T}}[\mathbf{x}]_{B'}$. In other words, we do not need to face up to the evaluation of $\mathbf{Q}^{-1}$ since $\mathbf{Q}^{-1}$ is merely $\mathbf{Q}^{\mathrm{T}}$.

Any matrix with the useful property (12) is known as an **orthogonal matrix** because it follows from (12) that the column vectors in $\mathbf{Q}$ are orthonormal.

As the second property of $\mathbf{Q}$, observe that it also follows from (12) that

$$\boxed{\det\mathbf{Q} = \pm 1,} \tag{13}$$

that is, either $+1$ or -1 since $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$ implies that $\det(\mathbf{Q}^{\mathrm{T}}\mathbf{Q}) = \det\mathbf{I} = 1$. But $\det(\mathbf{Q}^{\mathrm{T}}\mathbf{Q}) = (\det\mathbf{Q}^{\mathrm{T}})(\det\mathbf{Q}) = (\det\mathbf{Q})(\det\mathbf{Q}) = (\det\mathbf{Q})^2$. Hence, $\det\mathbf{Q}$ must be $+1$ or -1.

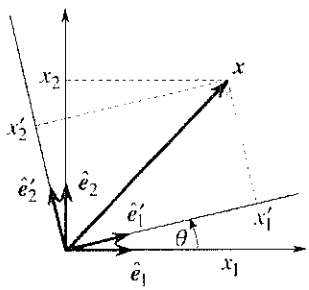


Figure 1. Rotation in the plane.

EXAMPLE 1. *Rotation in the Plane.* Consider the vector space $\mathbb{R}^2$, with the ON bases $B = \{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2\}$ and $B' = \{\hat{\mathbf{e}}_1', \hat{\mathbf{e}}_2'\}$ shown in Fig. 1. B' is obtained from B by a counterclockwise rotation through an angle θ (or clockwise if θ is negative). From the figure,

$$\begin{aligned} q_{11} &= \hat{\mathbf{e}}_1' \cdot \hat{\mathbf{e}}_1 = (1)(1)\cos\theta = \cos\theta, \\ q_{12} &= \hat{\mathbf{e}}_1' \cdot \hat{\mathbf{e}}_2 = (1)(1)\cos\left(\frac{\pi}{2} - \theta\right) = \sin\theta, \\ q_{21} &= \hat{\mathbf{e}}_2' \cdot \hat{\mathbf{e}}_1 = (1)(1)\cos\left(\frac{\pi}{2} + \theta\right) = -\sin\theta, \\ q_{22} &= \hat{\mathbf{e}}_2' \cdot \hat{\mathbf{e}}_2 = (1)(1)\cos\theta = \cos\theta, \end{aligned} \tag{14}$$

so that the coordinate transformation matrix is

$$\mathbf{Q} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}. \tag{15}$$

Hence,

$$\begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \tag{16}$$

Or, the other way around,

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}^{-1} \begin{bmatrix} x_1' \\ x_2' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} x_1' \\ x_2' \end{bmatrix}$$

so

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1' \\ x_2' \end{bmatrix}. \tag{17}$$

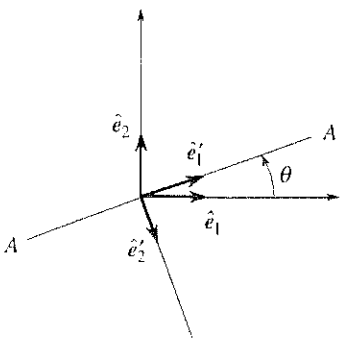


Figure 2. Rotation plus reflection.

COMMENT 1. It is easy enough to check (16) and (17) for one or two special cases. For example, if $\theta = 0$ then the two bases coincide so we should have $x_1' = x_1$ and $x_2' = x_2$, and that is what (16) and (17) give. Also, if $\theta = \pi/2$, say, we should have $x_1' = x_2$ and $x_2' = -x_1$ and, again, that is what (16) and (17) give.

COMMENT 2. Two ON bases in a plane are not necessarily related through a rotation. In this example, for instance, if we reverse the direction of $\hat{\mathbf{e}}_2'$, then $\{\hat{\mathbf{e}}_1', \hat{\mathbf{e}}_2'\}$ is still ON

(Fig. 2), but is not obtainable from $\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2\}$ by means of a rotation alone. Rather, we need a rotation *and a reflection*, a counterclockwise rotation through an angle θ, and then a reflection about AA (or, first a reflection about the $\hat{\mathbf{e}}_1$ axis and then a counterclockwise rotation through an angle θ). In this case

$$\begin{aligned} q_{11} &= \hat{\mathbf{e}}_1' \cdot \hat{\mathbf{e}}_1 = \cos\theta, \\ q_{12} &= \hat{\mathbf{e}}_1' \cdot \hat{\mathbf{e}}_2 = \cos\left(\frac{\pi}{2} - \theta\right) = \sin\theta, \\ q_{21} &= \hat{\mathbf{e}}_2' \cdot \hat{\mathbf{e}}_1 = \cos\left(\frac{\pi}{2} - \theta\right) = \sin\theta, \\ q_{22} &= \hat{\mathbf{e}}_2' \cdot \hat{\mathbf{e}}_2 = \cos(\pi - \theta) = -\cos\theta \end{aligned}$$

so that $\mathbf{Q}$ is the orthogonal matrix

$$\mathbf{Q} = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}. \tag{18}$$

Recall from (13) that $\det\mathbf{Q}$ is either $+1$ or -1. For the case where B and B' are related through a pure rotation [$\mathbf{Q}$ given by (15)] $\det\mathbf{Q} = +1$, and for the case where they are related through a reflection and a rotation [$\mathbf{Q}$ given by (18)] $\det\mathbf{Q} = -1$. ∎

Closure. In this brief section we study the relationship between the components, or coordinates, of any given vector $\mathbf{x}$ expanded in terms of two different bases B and B'. We find the linear relationship (6), where the q_{ij} elements of the coordinate transformation matrix $\mathbf{Q}$ are the expansion coefficients of $\mathbf{e}_j$ in terms of $\mathbf{e}_1', \ldots, \mathbf{e}_n'$, as indicated by (3).

If B and B' are ON, then the q_{ij}'s are computed, simply, from (10), and $\mathbf{Q}$ admits the properties that $\mathbf{Q}^{\mathrm{T}} = \mathbf{Q}^{-1}$ and that $\det\mathbf{Q}$ is $+1$ or -1. *Any* matrix $\mathbf{Q}$ having the property $\mathbf{Q}^{\mathrm{T}} = \mathbf{Q}^{-1}$ has ON column vectors and is called an orthogonal matrix.

EXERCISES 10.7

1. Given that $\mathbf{e}_1 = \mathbf{e}_1' + 2\mathbf{e}_2'$ and $\mathbf{e}_2 = \mathbf{e}_1' - \mathbf{e}_2'$, find the coordinate transformation matrix $\mathbf{Q}$. Is $\mathbf{Q}$ orthogonal? If $[\mathbf{x}]_B = [5, -1]^{\mathrm{T}}$, find $[\mathbf{x}]_{B'}$. If $[\mathbf{x}]_{B'} = [2, 3]^{\mathrm{T}}$, find $[\mathbf{x}]_B$.

2. Given that $\mathbf{e}_1 = \mathbf{e}_1' + \mathbf{e}_2' - \mathbf{e}_3'$, $\mathbf{e}_2 = \mathbf{e}_1' - \mathbf{e}_2' + \mathbf{e}_3'$, and $\mathbf{e}_3 = -\mathbf{e}_1' + \mathbf{e}_2' + \mathbf{e}_3'$, find the coordinate transformation matrix $\mathbf{Q}$. Is $\mathbf{Q}$ orthogonal? If $[\mathbf{x}]_B = [4, 1, -2]^{\mathrm{T}}$, find $[\mathbf{x}]_{B'}$. If $[\mathbf{x}]_{B'} = [1, 0, 2]^{\mathrm{T}}$, find $[\mathbf{x}]_B$.

3. Let $\hat{\mathbf{e}}_1 = [1, 0]^{\mathrm{T}}$, $\hat{\mathbf{e}}_2 = [0, 1]^{\mathrm{T}}$, and $\hat{\mathbf{e}}_1' = \frac{1}{\sqrt{5}}[2, 1]^{\mathrm{T}}$, $\hat{\mathbf{e}}_2' = \frac{1}{\sqrt{5}}[1, -2]^{\mathrm{T}}$.

(a) Find the coordinate transformation matrix $\mathbf{Q}$. Is $\mathbf{Q}$ orthogonal?
(b) If $[\mathbf{x}]_B = [8, -6]^{\mathrm{T}}$, find $[\mathbf{x}]_{B'}$.
(c) If $[\mathbf{x}]_{B'} = [1, 3]^{\mathrm{T}}$, find $[\mathbf{x}]_B$.

4. Let $\hat{\mathbf{e}}_1 = [1, 0, 0, 0]^{\mathrm{T}}$, $\hat{\mathbf{e}}_2 = [0, 1, 0, 0]^{\mathrm{T}}$, $\hat{\mathbf{e}}_3 = [0, 0, 1, 0]^{\mathrm{T}}$, $\hat{\mathbf{e}}_4 = [0, 0, 0, 1]^{\mathrm{T}}$, and $\hat{\mathbf{e}}_1' = \frac{1}{\sqrt{2}}[1, 1, 0, 0]^{\mathrm{T}}$, $\hat{\mathbf{e}}_2' = [0, 0, 1, 0]^{\mathrm{T}}$, $\hat{\mathbf{e}}_3' = \frac{1}{\sqrt{3}}[1, -1, 0, 1]^{\mathrm{T}}$, $\hat{\mathbf{e}}_4' = \frac{1}{\sqrt{6}}[1, -1, 0, -2]^{\mathrm{T}}$.

(a) Find the coordinate transformation matrix $\mathbf{Q}$. Is $\mathbf{Q}$ orthogonal?
(b) If $[\mathbf{x}]_B = [1, 1, 2, 5]^{\mathrm{T}}$, find $[\mathbf{x}]_{B'}$.

(c) If $[\mathbf{x}]_{B'} = [1, 1, 2, 5]^{\mathrm{T}}$, find $[\mathbf{x}]_B$.

5. Show whether or not these matrices are orthogonal.

(a) $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$ (d) $\begin{bmatrix} 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \end{bmatrix}$

(e) $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$ (f) $\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{3}} & -\frac{2}{\sqrt{6}} \end{bmatrix}$

6. For the case of rotation in a plane, the transformation $\mathbf{Q}$ corresponded to a counterclockwise rotation of the basis through an angle θ. Does $\mathbf{Q}^{-1}$ correspond to the reverse of this, a *clockwise* rotation θ ? Prove or disprove.

7. (*Rotation and reflection*) (a) Show that *every* orthogonal coordinate transformation matrix of order 2 is of one of the following two types:

$$\mathbf{Q}_1 = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, \quad \mathbf{Q}_2 = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix},$$

i.e., as given by the pure rotation (15) or by the rotation plus reflection (18).
(b) Show that these two cases can be distinguished by the sign of the determinant, specifically, that $\det\mathbf{Q} = +1$ if $\mathbf{Q}$ corresponds to pure rotation, and that $\det\mathbf{Q} = -1$ if $\mathbf{Q}$ corresponds to rotation plus reflection.

8. (a) Prove that if $\mathbf{Q}$ is orthogonal, then so is $\mathbf{Q}^{\mathrm{T}}$.
(b) Prove that if $\mathbf{Q}$ is orthogonal, then so is $\mathbf{Q}^{-1}$.

9. Evaluate $\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}^n$.

10.8 Vector Transformation (Optional)

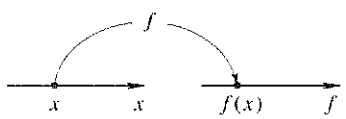


Figure 1. Function f as a transformation.

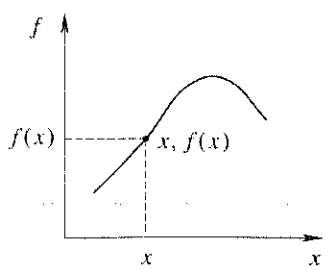


Figure 2. The graph of f.

Recall that a real-valued function f of a real variable x is a rule that assigns a uniquely determined value $f(x)$ to each specified value x as illustrated in Fig. 1. Thus, f is a transformation, or mapping, from points on an x axis to points on an f axis, and we view x as the "input" and $f(x)$ as the "output." [A more familiar graphical display of f, called the *graph* of f, can be obtained if, following *Descartes* (1596–1650), we arrange the x and f axes at right angles to each other and plot the set of points $x, f(x)$ as illustrated in Fig. 2.]

In this section we reconsider vectors and matrices from this transformation point of view. Specifically, we consider vector-valued functions $\mathbf{F}$ of a vector variable $\mathbf{x}$. That is, the "input" is now a vector $\mathbf{x}$ from some vector space V, and the function $\mathbf{F}$ assigns a uniquely determined "output" vector $\mathbf{F}(\mathbf{x})$ in some vector space W. We call $\mathbf{F}$ a **transformation**, or **mapping**, from V into W, and denote it as

$$\mathbf{F} : V \to W.$$

We call V the **domain of F** and W the **range of definition of F**. W may, but need not, be identical to V. If it *is* identical, then $\mathbf{F} : V \to V$ is called an **operator** on V.

EXAMPLE 1. To illustrate, consider the transformation $\mathbf{F} : \mathbb{R}^4 \to \mathbb{R}^3$ defined by

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} x_1 - 2x_2 + x_4 \\ x_1 + x_3 \\ x_1 + 2x_2 + 2x_3 - x_4 \end{bmatrix}. \tag{1}$$

Here V is $\mathbb{R}^4$, W is $\mathbb{R}^3$, and the input vector $\mathbf{x}$ and the output vector $\mathbf{F}(\mathbf{x})$ are

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \text{ in } \mathbb{R}^4, \quad \text{and} \quad \mathbf{F}(\mathbf{x}) = \begin{bmatrix} x_1 - 2x_2 + x_4 \\ x_1 + x_3 \\ x_1 + 2x_2 + 2x_3 - x_4 \end{bmatrix} \text{ in } \mathbb{R}^3. \tag{2}$$

For example, if $\mathbf{x} = [2, 3, 6, -1]^{\mathrm{T}}$, then $\mathbf{F}(\mathbf{x}) = [-5, 8, 21]^{\mathrm{T}}$. This transformation is not an operator since $W = \mathbb{R}^3$, whereas $V = \mathbb{R}^4$. ∎

We say that the vector $\mathbf{F}(\mathbf{x})$ in W is the **image** of the vector $\mathbf{x}$ in V under the transformation $\mathbf{F}$, and that $\mathbf{x}$ is the **inverse image** of $\mathbf{F}(\mathbf{x})$.

Since V and W are vector spaces, each must contain a zero vector. We will denote these zero vectors as $\mathbf{0}_V$ and $\mathbf{0}_W$, respectively. Finally, we define the image of V in W as the **range** R of $\mathbf{F}$, and we define the inverse image of $\mathbf{0}_W$ in V as the **nullspace** or **kernel** K of $\mathbf{F}$. That is, the kernel K is the part of V that maps to the zero vector $\mathbf{0}_W$ in W.

EXAMPLE 2. Let us find the range and kernel of the transformation $\mathbf{F}$ given in Example 1. First, the range. The range of $\mathbf{F}$ is the set of all vectors $\mathbf{c}$ in W for which the equation $\mathbf{F}(\mathbf{x}) = \mathbf{c}$ is consistent, that is, has at least one solution $\mathbf{x}$ in V. In the present case $\mathbf{F}(\mathbf{x}) = \mathbf{c}$ is

$$\begin{bmatrix} x_1 - 2x_2 + x_4 \\ x_1 + x_3 \\ x_1 + 2x_2 + 2x_3 - x_4 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} \tag{3}$$

or, in scalar form,

$$\begin{aligned} x_1 - 2x_2 \qquad\quad + x_4 &= c_1, \\ x_1 \qquad\quad + x_3 \qquad &= c_2, \\ x_1 + 2x_2 + 2x_3 - x_4 &= c_3. \end{aligned} \tag{4}$$

Applying elementary operations to (4), we obtain the equivalent system

$$\begin{aligned} x_1 - 2x_2 \qquad\quad + x_4 &= c_1, \\ 2x_2 + x_3 - x_4 &= c_2 - c_1, \\ 0 &= c_3 - 2c_2 + c_1. \end{aligned} \tag{5a,b,c}$$

This system is consistent if and only if $\mathbf{c}$ lies in the plane (through the origin) defined by $c_3 - 2c_2 + c_1 = 0$. That plane is a two-dimensional subspace of $\mathbb{R}^3$ and is the range R of $\mathbf{F}$.

Turning to the kernel K of $\mathbf{F}$, that is, the inverse image of $[0,0,0]^{\mathrm{T}}$ in $\mathbb{R}^4$, we need merely set $c_1 = c_2 = c_3 = 0$ in (5) and solve for $\mathbf{x}$. That solution gives

$$\mathbf{x} = \alpha_1 \begin{bmatrix} 0 \\ 1 \\ 0 \\ 2 \end{bmatrix} + \alpha_2 \begin{bmatrix} -2 \\ -1 \\ 2 \\ 0 \end{bmatrix} \equiv \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2,$$

where α_1, α_2 are arbitrary constants. Then $K = \operatorname{span}\{\mathbf{x}_1, \mathbf{x}_2\}$ and $\dim K = 2$. These results are summarized, schematically, in Fig. 3. ∎

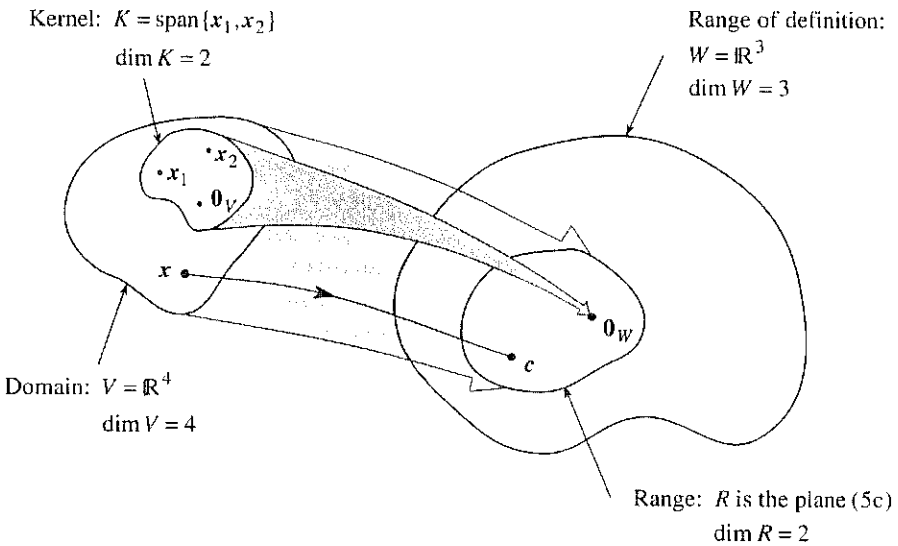


Figure 3. The transformation $\mathbf{F}$.

Just as linearity, or the absence of it, was crucial in the theory of ordinary differential equations, it is likewise crucial here. We distinguish transformations as linear or nonlinear as follows.

DEFINITION 10.8.1 *Linear Transformation*
We say that $\mathbf{F} : V \to W$ is a **linear** transformation if

$$\mathbf{F}(\alpha \mathbf{u} + \beta \mathbf{v}) = \alpha \mathbf{F}(\mathbf{u}) + \beta \mathbf{F}(\mathbf{v}) \tag{6}$$

for every choice of vectors $\mathbf{u}, \mathbf{v}$ in V, and scalars α, β; otherwise, $\mathbf{F}$ is said to be **nonlinear**.

EXAMPLE 3. To illustrate, consider the transformation $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^2$, where

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} x_1 - 2x_2 + x_3 \\ x_2 + 5x_3 \end{bmatrix}. \tag{7}$$

Let us see if $\mathbf{F}$ is linear.

$$\begin{aligned} \mathbf{F}(\alpha\mathbf{u} + \beta\mathbf{v}) &= \begin{bmatrix} (\alpha u_1 + \beta v_1) - 2(\alpha u_2 + \beta v_2) + (\alpha u_3 + \beta v_3) \\ (\alpha u_2 + \beta v_2) + 5(\alpha u_3 + \beta v_3) \end{bmatrix} \\ &= \alpha \begin{bmatrix} u_1 - 2u_2 + u_3 \\ u_2 + 5u_3 \end{bmatrix} + \beta \begin{bmatrix} v_1 - 2v_2 + v_3 \\ v_2 + 5v_3 \end{bmatrix} \\ &= \alpha\,\mathbf{F}(\mathbf{u}) + \beta\,\mathbf{F}(\mathbf{v}) \end{aligned} \tag{8}$$

for every choice of $\mathbf{u}, \mathbf{v}, \alpha, \beta$, so $\mathbf{F}$ is indeed linear. Notice that the key step in (8), the second equality, follows from the definitions of the addition and scalar multiplication of matrices. ∎

EXAMPLE 4. Consider $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$, where

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} x_1^2 \\ x_1 + 2x_2 \end{bmatrix}. \tag{9}$$

Then

$$\begin{aligned} &\mathbf{F}(\alpha\mathbf{u} + \beta\mathbf{v}) \\ &\quad = \begin{bmatrix} (\alpha u_1 + \beta v_1)^2 \\ (\alpha u_1 + \beta v_1) + 2(\alpha u_2 + \beta v_2) \end{bmatrix} \\ &\quad = \alpha \begin{bmatrix} u_1^2 \\ u_1 + 2u_2 \end{bmatrix} + \beta \begin{bmatrix} v_1^2 \\ v_1 + 2v_2 \end{bmatrix} + \begin{bmatrix} (\alpha^2 - \alpha)u_1^2 + (\beta^2 - \beta)v_1^2 + 2\alpha\beta u_1 v_1 \\ 0 \end{bmatrix} \\ &\quad = \alpha\,\mathbf{F}(\mathbf{u}) + \beta\,\mathbf{F}(\mathbf{v}) + \text{deviation}, \end{aligned} \tag{10}$$

where the "deviation,"

$$\mathbf{F}(\alpha\mathbf{u} + \beta\mathbf{v}) - \alpha\,\mathbf{F}(\mathbf{u}) - \beta\,\mathbf{F}(\mathbf{v}) = \begin{bmatrix} (\alpha^2 - \alpha)u_1^2 + (\beta^2 - \beta)v_1^2 + 2\alpha\beta u_1 v_1 \\ 0 \end{bmatrix} \tag{11}$$

is obviously not zero for all choices of $\alpha, \beta, \mathbf{u}, \mathbf{v}$. For instance, if $\alpha = \beta = u_1 = v_1 = 1$, then (for any u_2 and v_2) the deviation vector is $[2, 0]^{\mathrm{T}}$. Thus, (6) does not hold for all choices of $\alpha, \beta, \mathbf{u}, \mathbf{v}$, so $\mathbf{F}$ is nonlinear. ∎

If $\mathbf{F}$ is linear, then besides (6) we have

$$\begin{aligned} \mathbf{F}(\alpha\mathbf{u} + \beta\mathbf{v} + \gamma\mathbf{w}) &= \mathbf{F}(\alpha\mathbf{u} + (\beta\mathbf{v} + \gamma\mathbf{w})) \\ &= \alpha\,\mathbf{F}(\mathbf{u}) + \mathbf{F}(\beta\mathbf{v} + \gamma\mathbf{w}) = \alpha\,\mathbf{F}(\mathbf{u}) + \beta\,\mathbf{F}(\mathbf{v}) + \gamma\,\mathbf{F}(\mathbf{w}) \end{aligned}$$

or, more generally,

$$\boxed{\mathbf{F}(\alpha_1\mathbf{u}_1 + \cdots + \alpha_n\mathbf{u}_n) = \alpha_1\mathbf{F}(\mathbf{u}_1) + \cdots + \alpha_n\mathbf{F}(\mathbf{u}_n).} \tag{12}$$

Observe that (7) can be expressed in matrix notation as

$$\mathbf{F}(\mathbf{x}) = \begin{bmatrix} 1 & -2 & 1 \\ 0 & 1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}. \tag{13}$$

That is, the action of $\mathbf{F}$ on $\mathbf{x}$ is equivalent to multiplication by $\mathbf{A}$, where $\mathbf{A}$ is the 2×3 matrix in (13),

$$\mathbf{F}(\mathbf{x}) = \mathbf{A}\mathbf{x}. \tag{14}$$

Note that we do not say that "the transformation $\mathbf{F}$ is the matrix $\mathbf{A}$." Rather, we say that $\mathbf{F}$ is the transformation *multiplication by* $\boldsymbol{A}$. For that reason we call $\mathbf{F}$ a **matrix transformation**. The point that we wish to make here is that the linear transformation given in Example 3 happens to be a matrix transformation. We now show that that correspondence is no accident.

THEOREM 10.8.1 *Matrix Transformation*
A transformation $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ is linear if and only if it is a matrix transformation.

Proof: First, we show that if $\mathbf{F}$ is a matrix transformation [i.e., if there is an $m \times n$ matrix $\mathbf{A}$ such that $\mathbf{F}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ for each $\mathbf{x}$ in $\mathbb{R}^n$], then $\mathbf{F}$ is linear. That is easy since $\mathbf{F}(\alpha\mathbf{u} + \beta\mathbf{v}) = \mathbf{A}(\alpha\mathbf{u} + \beta\mathbf{v}) = \alpha\mathbf{A}\mathbf{u} + \beta\mathbf{A}\mathbf{v} = \alpha\mathbf{F}(\mathbf{u}) + \beta\mathbf{F}(\mathbf{v})$ for all $\mathbf{u}, \mathbf{v}$ in $\mathbb{R}^n$ and for all scalars α, β. To prove the converse, let $\{\mathbf{i}_1, \ldots, \mathbf{i}_n\}$ and $\{\mathbf{j}_1, \ldots, \mathbf{j}_m\}$ be bases for $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. We may express

$$\mathbf{x} = \sum_{j=1}^{n} x_j \mathbf{i}_j \quad \text{and} \quad \mathbf{y} = \sum_{k=1}^{m} y_k \mathbf{j}_k.$$

Then

$$\mathbf{F}(\mathbf{x}) = \mathbf{F}\left(\sum_{j=1}^{n} x_j \mathbf{i}_j\right) = \sum_{j=1}^{n} x_j \mathbf{F}(\mathbf{i}_j)$$

by (12). Since $\mathbf{F}(\mathbf{i}_j)$ is in $\mathbb{R}^m$, it can be expressed in the form

$$\mathbf{F}(\mathbf{i}_j) = \sum_{k=1}^{m} a_{kj} \mathbf{j}_k$$

so

$$\mathbf{F}(\mathbf{x}) = \sum_{j=1}^{n} x_j \sum_{k=1}^{m} a_{kj} \mathbf{j}_k = \sum_{k=1}^{m} \left(\sum_{j=1}^{n} a_{kj} x_j\right) \mathbf{j}_k. \tag{15}$$

But we also have

$$\mathbf{F}(\mathbf{x}) = \mathbf{y} = \sum_{k=1}^{m} y_k \mathbf{j}_k, \tag{16}$$

and it follows from (15) and (16), and the linear independence of the $\mathbf{j}_k$'s, that

$$y_k = \sum_{j=1}^{n} a_{kj} x_j$$

or $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{A} = \{a_{kj}\}$ is $m \times n$, as was to be proved. ■

We now introduce some additional terminology. First, recall that $\mathbf{F}$ is understood to be single valued. That is, to each vector $\mathbf{x}$ in the domain V of $\mathbf{F}$ there corresponds a unique image $\mathbf{F}(\mathbf{x})$ in the range R of $\mathbf{F}$. If, in addition to each vector in R there corresponds a unique inverse image in V, then $\mathbf{F}$ is said to be **one-to-one**. Notice that we do not say "to each vector in W there corresponds a unique inverse image in V" since R may not be all of W, in which case those vectors which are in W but not in R have no inverse image at all. If R does turn out to be all of W, then $\mathbf{F}$ is said to be **onto**; that is, $\mathbf{F}$ maps V "onto" W rather than "into" W. Finally, if $\mathbf{F}$ is both one-to-one and onto, it is said to be **invertible** for then every vector in W has a unique image in V. This inverse transformation, from W onto V, is called the **inverse of F** and is denoted as $\mathbf{F}^{-1}$.

EXAMPLE 5. Consider the matrix transformation $\mathbf{F}$ in Example 2. There, $\mathbf{F}(\mathbf{x}) = \mathbf{A}\mathbf{x}$, with

$$\mathbf{A} = \begin{bmatrix} 1 & -2 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 2 & 2 & -1 \end{bmatrix}. \tag{17}$$

From (5c) we see that R is only the two-dimensional subspace of W $(= \mathbb{R}^3)$ consisting of the plane $c_3 - 2c_2 + c_1 = 0$ so that $\mathbf{F}$ is not *onto*. That result is illustrated schematically in Fig. 3, where R is shown to be only a part of W. Furthermore, if $\mathbf{c}$ is in R [i.e., if (5c) is satisfied], then (5) yields a nonunique solution for $\mathbf{x}$, so $\mathbf{F}$ is not one-to-one. Summarizing, $\mathbf{F}$ is neither one-to-one nor onto and is therefore not invertible. ■

EXAMPLE 6. Consider the matrix transformation $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^3$ associated with the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & -2 \\ 1 & 1 \\ 2 & -1 \end{bmatrix}. \tag{18}$$

Applying elementary operations to the system $\mathbf{A}\mathbf{x} = \mathbf{c}$, namely, to

$$\begin{bmatrix} 1 & -2 \\ 1 & 1 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}, \tag{19}$$

we obtain the equivalent system

$$\begin{aligned} x_1 - 2x_2 &= c_1, \\ 3x_2 &= c_2 - c_1, \\ 0 &= c_3 - c_2 - c_1, \end{aligned} \tag{20a,b,c}$$

from which it is seen that $\mathbf{F}$ is *not* onto (why not?), although it is one-to-one (why?). Thus $\mathbf{F}$ is not invertible. ∎

EXAMPLE 7. Consider the matrix transformation $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^2$ associated with the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 0 & -3 \end{bmatrix}. \tag{21}$$

Applying elementary operations to the system $\mathbf{Ax} = \mathbf{c}$, namely, to

$$\begin{bmatrix} 1 & 1 & 2 \\ 2 & 0 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}, \tag{22}$$

we obtain the equivalent system

$$\begin{aligned} x_1 + \ \ x_2 + 2x_3 &= c_1, \\ 2x_2 + 7x_3 &= 2c_1 - c_2, \end{aligned} \tag{23a,b}$$

from which it is seen that $\mathbf{F}$ *is* onto (why?), although *not* one-to-one (why not?). Thus, $\mathbf{F}$ is not invertible. ∎

EXAMPLE 8. Consider the matrix transformation $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^3$ associated with the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 1 \\ 0 & 3 & 1 \\ 1 & 1 & 0 \end{bmatrix}. \tag{24}$$

Applying elementary operations to the system $\mathbf{Ax} = \mathbf{c}$, we obtain

$$\begin{aligned} 2x_1 - \ \ x_2 + \ \ x_3 &= c_1, \\ 3x_2 + \ \ x_3 &= c_2, \\ -\ 2x_3 &= 2c_3 - c_2 - c_1, \end{aligned} \tag{25a,b,c}$$

from which it is seen that $\mathbf{F}$ is both onto (so that $\mathbf{F}$ is an operator) and one-to-one, and is therefore invertible; completion of the Gauss–Jordan reduction of (25) reveals that the inverse operator $\mathbf{F}^{-1}$ is the matrix operator associated with the matrix

$$\begin{bmatrix} \frac{1}{6} & -\frac{1}{6} & \frac{2}{3} \\ -\frac{1}{6} & \frac{1}{6} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} & -1 \end{bmatrix}, \tag{26}$$

which, of course, is the inverse of the $\mathbf{A}$ matrix, $\mathbf{A}^{-1}$. ∎

Closure. Recall that we develop n-space in Chapter 9, and then generalize the vector space concept in Section 9.6. The role of the present section is analogous in

that here we have generalize the concept of matrix, developed earlier in this chapter, to that of *transformations* on vector spaces.

Besides establishing the concept, together with the standard mathematical terminology, the key result is given in Theorem 10.8.1, that a transformation $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ is linear if and only if it is a matrix transformation.

EXERCISES 10.8

1. In general, the effect of a transformation on the input vector varies from one input vector to another. For example, let $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$ be a matrix transformation $\mathbf{F}(\mathbf{x}) = \mathbf{A}\mathbf{x}$, where $\mathbf{A} = \begin{bmatrix} 3 & 2 \\ 0 & 1 \end{bmatrix}$. In 2-space the "effect" of a transformation on a nonzero vector $\mathbf{x}$ amounts to the resulting rotation in the plane, and the dilation (i.e., $\|\mathbf{A}\mathbf{x}\| / \|\mathbf{x}\|$). For the transformation $\mathbf{F}$ given above, show that these effects are as follows for the given input vectors, and notice that the effect of $\mathbf{F}$ varies from one $\mathbf{x}$ to another.

	Input	Effect of $\mathbf{F}$
(a)	$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$	rotation $= 0$ radians dilation $= 3$
(b)	$\mathbf{x} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$	rotation ≈ 1.1 radians dilation $= \sqrt{5}$
(c)	$\mathbf{c} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$	rotation $= 0$ radians dilation $= 1$

2. Determine whether $\mathbf{F}$ is linear or nonlinear by determining whether or not the deviation $\mathbf{F}(\alpha\,\mathbf{u} + \beta\,\mathbf{v}) - \alpha\mathbf{F}(\mathbf{u}) - \beta\mathbf{F}(\mathbf{u})$ is necessarily zero.

(a) $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$, $\mathbf{F}(\mathbf{x}) = \begin{bmatrix} x_1^2 \\ x_1 + x_2 \end{bmatrix}$

(b) $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$, $\mathbf{F}(\mathbf{x}) = \begin{bmatrix} 3x_1 \\ x_1 + x_2 \end{bmatrix}$

(c) $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^2$, $\mathbf{F}(\mathbf{x}) = \begin{bmatrix} x_1 x_2 \\ x_3 \end{bmatrix}$

(d) $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^1$, $\mathbf{F}(\mathbf{x}) = \begin{bmatrix} 3x_2 \end{bmatrix}$

(e) $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$, $\mathbf{F}(\mathbf{x}) = \begin{bmatrix} \sin x_2 \\ 4x_1 \end{bmatrix}$

(f) $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$, $\mathbf{F}(\mathbf{x}) = \begin{bmatrix} x_1 + 1 \\ x_2 + 1 \end{bmatrix}$

3. (*Identity operator*) We say that $\mathbf{I} : V \to V$ is an **identity operator** if $\mathbf{I}(\mathbf{x}) = \mathbf{x}$ for each $\mathbf{x}$ in V.

(a) Suppose that $\mathbf{F} : V \to V$ is linear and that $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a basis for V. Show that if $\mathbf{F}(\mathbf{v}_1) = \mathbf{v}_1, \ldots, \mathbf{F}(\mathbf{v}_n) = \mathbf{v}_n$, then $\mathbf{F} = \mathbf{I}$.

(b) Determine the matrix $\mathbf{A}$ corresponding to the identity operator $\mathbf{I} : \mathbb{R}^n \to \mathbb{R}^n$, i.e., such that $\mathbf{I}(\mathbf{x}) = \mathbf{A}\mathbf{x}$.

4. (*Zero transformation*) We say that $\mathbf{\Phi} : V \to W$ is a **zero transformation** if $\mathbf{\Phi}(\mathbf{x}) = \mathbf{0}$ for all $\mathbf{x}$'s in V.

(a) Suppose that $\mathbf{F} : V \to W$ is linear and that $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a basis for V. Show that if $\mathbf{F}(\mathbf{v}_1) = \mathbf{0}, \ldots, \mathbf{F}(\mathbf{v}_n) = \mathbf{0}$, then $\mathbf{F} = \mathbf{\Phi}$.

(b) Determine the matrix $\mathbf{A}$ corresponding to the zero transformation $\mathbf{\Phi} : \mathbb{R}^n \to \mathbb{R}^m$, i.e., such that $\mathbf{\Phi}(\mathbf{x}) = \mathbf{A}\mathbf{x}$.

5. In each case $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ is the matrix transformation corresponding to the given $m \times n$ matrix $\mathbf{A}$. Determine $\dim R$, $\dim K$, and $\dim V$. Is $\mathbf{F}$ onto? One-to-one? Invertible? Explain. Put forward any basis for K and any basis for R (if, indeed, they have bases; see Exercise 7 in Section 9.9).

(a) $\begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 1 \\ 4 & 3 & 3 \end{bmatrix}$

(c) $\begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix}$ (d) $\begin{bmatrix} 4 & 1 \\ 3 & 2 \\ 0 & -1 \end{bmatrix}$

(e) $\begin{bmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{bmatrix}$ (f) $\begin{bmatrix} 0 & 0 & 6 & 0 \\ 0 & -2 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$

(g) $\begin{bmatrix} 2 & -1 & 1 \\ 0 & 0 & 3 \\ 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}$ (h) $\begin{bmatrix} 1 & 4 \\ 2 & 1 \\ 0 & 1 \\ 0 & 3 \end{bmatrix}$

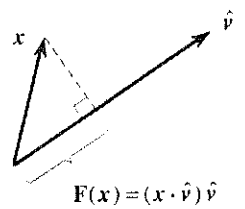
6. Make up any example of a matrix transformation $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^m$ that is one-to-one but not onto, one that is onto but not one-to-one, one that is neither one-to-one nor onto, and one that is both one-to-one and onto. If such an example is impossible, explain why that is so.

(a) $n = 2, m = 2$ (b) $n = 3, m = 2$
(c) $n = 2, m = 3$ (d) $n = 3, m = 3$

7. (*Projection operators*) Let $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^3$ be the transformation

$$\mathbf{F}(\mathbf{x}) = (\mathbf{x} \cdot \hat{\mathbf{v}})\hat{\mathbf{v}}, \tag{7.1}$$

where $\hat{\mathbf{v}} = [v_1, v_2, v_3]^{\mathrm{T}}$ is a prescribed unit vector. In geometric terms, $\mathbf{F}(\mathbf{x})$ is the vector orthogonal projection of $\mathbf{x}$ onto the line of action of $\hat{\mathbf{v}}$, as illustrated below. Hence, $\mathbf{F}$ in (7.1) is known as a **projection operator**. We now state the problem: Show that $\mathbf{F}$ is linear, so that one can express $\mathbf{F}(\mathbf{x})$ as $\mathbf{A}\mathbf{x}$. Then determine the nine elements of the $\mathbf{A}$ matrix. (They will depend on v_1, v_2, v_3.)



8. (*More about projection operators*) Let $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^3$ be the transformation

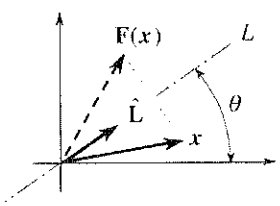
$$\mathbf{F}(\mathbf{x}) = (\mathbf{x} \cdot \hat{\mathbf{v}}_1)\hat{\mathbf{v}}_1 + (\mathbf{x} \cdot \hat{\mathbf{v}}_2)\hat{\mathbf{v}}_2, \tag{8.1}$$

where $\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2$ are prescribed ON vectors. Then $\mathbf{F}(\mathbf{x})$ is the vector orthogonal projection of $\mathbf{x}$ onto the *plane* spanned by $\{\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2\}$.

(a) Given that $\hat{\mathbf{v}}_1 = [1, 0, 0]^{\mathrm{T}}$ and $\hat{\mathbf{v}}_2 = [0, 1, 0]^{\mathrm{T}}$, work out $\mathbf{F}(\mathbf{x})$ for $\mathbf{x} = [2, 3, 4]^{\mathrm{T}}$, and draw an informative, labeled picture, analogous to the one shown in Exercise 7.
(b) Show that $\mathbf{F}$ is linear so that one can express $\mathbf{F}(\mathbf{x})$ as $\mathbf{A}\mathbf{x}$. Then determine the nine elements of the $\mathbf{A}$ matrix, in terms of the components v_{11}, v_{12}, v_{13} of $\hat{\mathbf{v}}_1$ and v_{21}, v_{22}, v_{23} of $\hat{\mathbf{v}}_2$.

9. Show that $\mathbf{F}(\alpha\mathbf{u} + \beta\mathbf{v}) = \alpha\mathbf{F}(\mathbf{u}) + \beta\mathbf{F}(\mathbf{v})$, in Definition 10.8.1, is equivalent to the two conditions $\mathbf{F}(\mathbf{u} + \mathbf{v}) = \mathbf{F}(\mathbf{u}) + \mathbf{F}(\mathbf{v})$ and $\mathbf{F}(\alpha\mathbf{u}) = \alpha\mathbf{F}(\mathbf{u})$.

10. (*Reflection about a line*) Let $\mathbf{F} : \mathbb{R}^2 \to \mathbb{R}^2$ reflect any given vector $\mathbf{x}$ about the line L, as shown in the accompanying figure.



(a) Show that $\mathbf{F}(\mathbf{x}) = \mathbf{x} + 2[(\mathbf{x} \cdot \hat{\mathbf{L}})\hat{\mathbf{L}} - \mathbf{x}]$.
(b) Show that $\mathbf{F}$ is linear and determine the matrix $\mathbf{A}$ such that $\mathbf{F}(\mathbf{x}) = \mathbf{A}\mathbf{x}$.
(c) Work out $\mathbf{A}^2$ and show that $\mathbf{A}^2 = \mathbf{I}$ (Exercise 3). Why should it have been obvious, without working out $\mathbf{A}^2$, that $\mathbf{A}^2 = \mathbf{I}$?

11. (*Linear combination and composition*) If $\mathbf{F}$ and $\mathbf{G}$ are transformations from V into W, then we define the **linear combination** of $\mathbf{F}$ and $\mathbf{G}$, $(\alpha\mathbf{F} + \beta\mathbf{G}) : V \to W$, by

$$(\alpha\mathbf{F} + \beta\mathbf{G})(\mathbf{x}) \equiv \alpha\mathbf{F}(\mathbf{x}) + \beta\mathbf{G}(\mathbf{x}) \tag{11.1}$$

for all $\mathbf{x}$ in V. Given transformations $\mathbf{F} : U \to V$ and $\mathbf{G} : V \to W$, we define the **composition** of $\mathbf{F}$ and $\mathbf{G}$, $(\mathbf{GF}) : U \to W$, by

$$(\mathbf{GF})(\mathbf{x}) \equiv \mathbf{G}(\mathbf{F}(\mathbf{x})) \tag{11.2}$$

for all $\mathbf{x}$ in U.

(a) Let $\mathbf{F} : \mathbb{R}^4 \to \mathbb{R}^2$ and $\mathbf{G} : \mathbb{R}^2 \to \mathbb{R}^2$ be matrix transformations with matrices

$$\mathbf{A} = \begin{bmatrix} 1 & 5 & 3 & -2 \\ 2 & 5 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 2 & 0 \\ 4 & 0 \end{bmatrix},$$

respectively. Evaluate $(\mathbf{GF})(\mathbf{x})$ for $\mathbf{x} = [3, 1, -2, 6]^{\mathrm{T}}$. Find a single matrix corresponding to the composite transformation $\mathbf{GF}$.
(b) Let $\mathbf{F} : V \to V$ be a linear operator, and define the composite transformation $\mathbf{F}^2 : V \to V$ by $\mathbf{F}^2(\mathbf{x}) \equiv \mathbf{F}(\mathbf{F}(\mathbf{x}))$. Show that $\mathbf{F}^2$ is linear, too.

12. Show that the *translation operator* $\mathbf{F}(\mathbf{x}) = \mathbf{x} + \mathbf{c}$, where $\mathbf{c}$ is a constant vector, is nonlinear.

13. (*Applications to computer graphics*) It is basic, in computer graphics, to be able to move points about, in 3-space, by combinations of translations and rotations. Translation is easily accomplished by the operator

$$\mathbf{F}(\mathbf{X}) = \mathbf{X} + \Delta\mathbf{X}, \tag{13.1}$$

where $\mathbf{X} = [x, y, z]^{\mathrm{T}}$ is the position vector to the point and $\Delta\mathbf{X} = [\Delta x, \Delta y, \Delta z]^{\mathrm{T}}$ is the translation. However, whereas it is convenient, in the computer software, to express all translations and rotations as matrix transformations, the operator $\mathbf{F} : \mathbb{R}^3 \to \mathbb{R}^3$ is not linear (Exercise 12), and hence not expressible as a matrix transformation. To circumvent this difficulty we define $\mathbf{X} = [x, y, z, 1]^{\mathrm{T}}$, instead, where the fourth

component, unity, is included for convenience. Then we can express

$$\mathbf{F}(\mathbf{X}) = \mathbf{T}\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & \Delta x \\ 0 & 1 & 0 & \Delta y \\ 0 & 0 & 1 & \Delta z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} x + \Delta x \\ y + \Delta y \\ z + \Delta z \\ 1 \end{bmatrix}, \tag{13.2}$$

which, if we pay attention to only the first three components, effects the translation by means of multiplication by $\mathbf{T}$, where $\mathbf{T}$ is the 4×4 matrix in (13.2).

(a) Show that

$$\mathbf{F}(\mathbf{X}) = \mathbf{R}_z\mathbf{X} = \begin{bmatrix} c_z & -s_z & 0 & 0 \\ s_z & c_z & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{13.3}$$

effects a rotation about the z axis through an angle θ_z, taken according to the right-hand rule, where c_z, s_z are shorthand for $\cos\theta_z$, $\sin\theta_z$, respectively. HINT: Letting $x = r\cos\theta$, $y = r\sin\theta$, show that

$$\mathbf{F}(\mathbf{X}) = \mathbf{R}_z\mathbf{X} = \begin{bmatrix} r\cos(\theta + \theta_z) \\ r\sin(\theta + \theta_z) \\ z \\ 1 \end{bmatrix}.$$

NOTE: Similarly, rotations about the x axis through an angle θ_x, and about the y axis through an angle θ_y, are effected by

$$\mathbf{F}(\mathbf{X}) = \mathbf{R}_y\mathbf{X} = \begin{bmatrix} c_y & 0 & -s_y & 0 \\ 0 & 1 & 0 & 0 \\ s_y & 0 & c_y & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{13.4}$$

and

$$\mathbf{F}(\mathbf{X}) = \mathbf{R}_x\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & c_x & -s_x & 0 \\ 0 & s_x & c_x & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}, \tag{13.5}$$

respectively, where c_x, s_x, c_y, s_y denote $\cos\theta_x, \sin\theta_x, \cos\theta_y, \sin\theta_y$, respectively.

(b) Show that a rotation about the z axis, followed by a translation, is effected by the composite transformation (see Exercise 11)

$$\mathbf{F}(\mathbf{X}) = \mathbf{T}\mathbf{R}_z\mathbf{X} = \begin{bmatrix} c_z & -s_z & 0 & \Delta x \\ s_z & c_z & 0 & \Delta y \\ 0 & 0 & 1 & \Delta z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}.$$

Does the order of the operations matter? That is, is $\mathbf{T}\mathbf{R}_z = \mathbf{R}_z\mathbf{T}$?

(c) Compute $\mathbf{F}(\mathbf{X}) = \mathbf{T}\mathbf{R}_x\mathbf{R}_y\mathbf{R}_z\mathbf{X}$ for $\mathbf{X} = [1, 1, 0, 1]^{\mathrm{T}}$, $\theta_z = -\pi/4$, $\theta_y = \pi/2$, $\theta_x = \pi$, $\Delta x = 2$, $\Delta y = 1$, $\Delta z = -3$. Verify the result by drawing the coordinate axes, identifying the initial point $\mathbf{X}$, and then carrying out each rotation and translation graphically in a neat sketch.

(d) Let the point and eraser of a pencil be located, initially, by $\mathbf{X}_p = [0, 1, 1, 1]^{\mathrm{T}}$ and $\mathbf{X}_e = [0, 1, 0, 1]^{\mathrm{T}}$, respectively. Locate the point and the eraser following the composite transformation

$$\mathbf{F}(\mathbf{X}) = \mathbf{T}\mathbf{R}_x\mathbf{R}_y\mathbf{R}_z\mathbf{X},$$

where $\theta_x = 0.2$, $\theta_y = 0.3$, $\theta_z = -0.6$, $\Delta x = 1$, $\Delta y = 3$, $\Delta z = 1$. In a neat sketch, show the pencil in its initial and final configurations, and verify that its length has remained the same.

(e) Repeat part (d), with $\theta_x = \pi/2$, $\theta_y = 0$, $\theta_z = -\pi$, $\Delta x = \Delta y = \Delta z = 1$.

(f) Repeat part (d), with $\theta_x = -\pi/2$, $\theta_y = \theta_z = \pi/2$, $\Delta x = \Delta y = 1$, $\Delta z = 0$.

Chapter 10 Review

The following review is limited to a number of isolated results and formulas that should be both understood and memorized.

Matrix multiplication:

$$\mathbf{AB} \neq \mathbf{BA} \qquad \text{(in general)}$$

Transpose:

$$(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$$

Determinants:

$$\det(\alpha\mathbf{A} + \beta\mathbf{B}) \neq \alpha \det\mathbf{A} + \beta \det\mathbf{B} \qquad \text{(in general)}$$
$$\det(\mathbf{AB}) = (\det\mathbf{A})(\det\mathbf{B})$$

Rank:

$$\begin{aligned} r(\mathbf{A}) &= \text{number of LI columns in } \mathbf{A} \\ &= \text{number of LI rows in } \mathbf{A} \end{aligned}$$

Systems of linear algebraic equations, $\mathbf{Ax} = \mathbf{c}$ (where $\mathbf{A}$ is $m \times n$):

Inconsistent:

$$\text{No solution if} \quad r(\mathbf{A}|\mathbf{c}) \neq r(\mathbf{A})$$

Consistent:

$$\text{Unique solution if} \quad r(\mathbf{A}|\mathbf{c}) = r(\mathbf{A}) = n$$
$$(n-r)\text{-parameter family of solutions if} \quad r(\mathbf{A}|\mathbf{c}) = r(\mathbf{A}) = r < n$$

The case where $m = n$:

$$\text{Unique solution} \quad \left(\mathbf{x} = \mathbf{A}^{-1}\mathbf{c}\right) \text{ if and only if } \det\mathbf{A} \neq 0 \quad [\text{i.e., } r(\mathbf{A}) = n]$$

Inverse matrix, $\mathbf{A}^{-1}$:

Exists, and is unique, if and only if $\det\mathbf{A} \neq 0$.

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$
$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$
$$(\mathbf{A}^{-1})^{\mathrm{T}} = (\mathbf{A}^{\mathrm{T}})^{-1}$$

Orthogonal matrices:

A matrix $\mathbf{Q}$ is orthogonal if it is square and its columns are ON.

Chapter 11

The Eigenvalue Problem

11.1 Introduction

In this chapter we study the problem

$$\boxed{\mathbf{A}\mathbf{x} = \lambda\mathbf{x},} \tag{1}$$

where $\mathbf{A}$ is a given $n \times n$ matrix, $\mathbf{x}$ is an unknown $n \times 1$ vector, and λ is an unknown scalar. If we re-express (1) as $\mathbf{A}\mathbf{x} = \lambda\mathbf{I}\mathbf{x}$ (where $\mathbf{I}$ is an $n \times n$ identity matrix), then subtraction of $\lambda\mathbf{I}\mathbf{x}$ from both sides gives the equivalent equation*

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0, \tag{2}$$

which is a homogeneous system of n equations in the n unknown x_j's, where the coefficient matrix $\mathbf{A} - \lambda\mathbf{I}$ contains the parameter λ.

To be sure that (1) and (2) are clear, let us write them out in scalar form, for $n = 3$, for example. Then (1) is the system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= \lambda x_1, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= \lambda x_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= \lambda x_3. \end{aligned}$$

Subtracting the terms on the right from those on the left gives

$$\begin{aligned} (a_{11} - \lambda)x_1 + \quad a_{12}x_2 + \quad a_{13}x_3 &= 0, \\ a_{21}x_1 + (a_{22} - \lambda)x_2 + \quad a_{23}x_3 &= 0, \\ a_{31}x_1 + \quad a_{32}x_2 + (a_{33} - \lambda)x_3 &= 0, \end{aligned}$$

which, in matrix form, is equation (2).

*Of course we don't need to insert the $\mathbf{I}$. We could re-express (1), correctly, as $\mathbf{A}\mathbf{x} - \lambda\mathbf{x} = 0$, but it would *not* follow from the latter that $(\mathbf{A} - \lambda)\mathbf{x} = 0$ because subtraction of a scalar (λ) from a matrix ($\mathbf{A}$) is not defined. Hence the need to insert $\mathbf{I}$.

From Chapter 10, we know that (2) is consistent because it necessarily admits the "trivial" solution $\mathbf{x} = \mathbf{0}$. However, our interest in (2) shall be in the search for *nontrivial* solutions, and we anticipate that whether or not nontrivial solutions exist will depend upon the value of λ. Thus, the problem of interest is as follows: given the $n \times n$ matrix $\mathbf{A}$, find the value(s) of λ (if any) such that (2) admits nontrivial solutions, and find those nontrivial solutions. The latter is called the **eigenvalue problem** and is the focus of this chapter. The λ's that lead to nontrivial solutions for $\mathbf{x}$ are called the **eigenvalues** (or **characteristic values**), and the corresponding nontrivial solutions for $\mathbf{x}$ are called the **eigenvectors** (or **characteristic vectors**).

The eigenvalue problem (1) [or, equivalently, (2)] occurs in a wide variety of applications such as vibration theory, chemical kinetics, stability of equilibria, buckling of structures, convergence of iterative techniques, and systems of coupled ordinary differential equations. To place the eigenvalue problem (1) in perspective, recall that in Chapter 8 and 10 we studied the problem

$$\mathbf{A}\mathbf{x} = \mathbf{c} \tag{3}$$

of m linear algebraic equations in n unknowns (i.e., $\mathbf{A}$ was $m \times n$). In general, $\mathbf{c} \neq \mathbf{0}$, in which case (3) was said to be nonhomogeneous. The eigenvalue problem is by no means unrelated to (3); it amounts to a special case, where $\mathbf{c} = \mathbf{0}$ (i.e., it is homogeneous), where $m = n$, and where the coefficient matrix "$\mathbf{A}$" $= \mathbf{A} - \lambda\mathbf{I}$ contains the parameter λ. Thus, to solve the eigenvalue problem we will be able to use results already established in Chapter 10.

11.2 Solution Procedure and Applications

11.2.1. Solution and applications. The eigenvalue problem

$$\boxed{(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}} \tag{1}$$

has the unique trivial solution $\mathbf{x} = \mathbf{0}$ if $\det(\mathbf{A} - \lambda\mathbf{I}) \neq 0$, and nontrivial solutions (in addition to the trivial solution) if and only if

$$\boxed{\det(\mathbf{A} - \lambda\mathbf{I}) = 0.} \tag{2}$$

The latter is not a vector or matrix equation; it is an algebraic equation in λ, known as the **characteristic equation** corresponding to the matrix $\mathbf{A}$, and its left-hand side is an nth degree polynomial known as the **characteristic polynomial**. According to the fundamental theorem of algebra, such an equation has precisely n roots in the complex plane. Since one or more of these roots can be repeated, we can say that there is at least one eigenvalue λ, and at most n distinct eigenvalues λ, corresponding to any given $n \times n$ matrix $\mathbf{A}$.

As in Chapter 10, we continue to consider only real matrices. However, even if $\mathbf{A}$ is real (so that the coefficients of the characteristic polynomial are too), the characteristic equation can still have complex roots. That case will not be very

important to us. Thus, we avoid it entirely in Chapter 11, and consider it briefly in Chapter 12.

This is not the first time we have run into the need to solve polynomial equations. In Section 3.4 we sought solutions to linear, homogeneous, constant-coefficient differential equations by seeking $y(x) = e^{\lambda x}$. Putting that solution form into the nth order differential equation gave an nth degree polynomial equation on λ. In fact, even the terminology was the same: the equation was called the **characteristic equation** of the differential equation, and the nth degree polynomial was called the **characteristic polynomial**. If $n = 2$ we can solve the characteristic equation by the quadratic formula. For larger n's we can, if necessary, use computer software such as the *Maple* fsolve command discussed in Section 3.4. Thus, let us consider (2) to have been solved for the eigenvalues, for the moment, and let us designate them as $\lambda_1, \ldots, \lambda_k$ $(1 \le k \le n)$.

Next, set $\lambda = \lambda_1$ in (1). Since $\det(\mathbf{A} - \lambda_1\mathbf{I}) = 0$, it is guaranteed that $(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{x} = \mathbf{0}$ will have nontrivial solutions. We can find those solutions by Gauss elimination, and we designate them as $\mathbf{e}_1$, where the letter e is for eigenvector. The $\mathbf{e}_1$ solution space is called the **eigenspace** corresponding to the eigenvalue λ_1.* Next, we set $\lambda = \lambda_2, \ldots, \lambda_k$ and repeat the process until the k eigenspaces have been found.

EXAMPLE 1. Determine all eigenvalues and eigenspaces of

$$\mathbf{A} = \begin{bmatrix} 2 & 2 & 1 \\ 1 & 3 & 1 \\ 1 & 2 & 2 \end{bmatrix}. \tag{3}$$

The characteristic equation is

$$\begin{aligned} \det(\mathbf{A} - \lambda\mathbf{I}) &= \begin{vmatrix} 2-\lambda & 2 & 1 \\ 1 & 3-\lambda & 1 \\ 1 & 2 & 2-\lambda \end{vmatrix} = \lambda^3 - 7\lambda^2 + 11\lambda - 5 \\ &= (\lambda - 5)(\lambda - 1)^2 = 0 \end{aligned} \tag{4}$$

so the eigenvalues of $\mathbf{A}$ are $\lambda_1 = 5$ and $\lambda_2 = 1$ (or vice versa since the order is immaterial), with $\lambda_2 = 1$ called a *repeated eigenvalue* – specifically, an eigenvalue of *multiplicity* 2 because it is a double root of the characteristic equation (4).

Next, find the eigenspaces.

$\lambda_1 = 5$: Then $(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{x} = \mathbf{0}$ becomes

$$\begin{bmatrix} 2-5 & 2 & 1 \\ 1 & 3-5 & 1 \\ 1 & 2 & 2-5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -3 & 2 & 1 \\ 1 & -2 & 1 \\ 1 & 2 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}; \tag{5}$$

*Note that the eigenspace corresponding to an eigenvalue λ_j is not quite the same as the set of eigenvectors corresponding to λ_j, it is that set *plus* the trivial solution (which is NOT itself an eigenvector). The reason we define the eigenspace corresponding to λ_j as the entire solution space of $(\mathbf{A} - \lambda_j\mathbf{I})\mathbf{x} = 0$ (i.e., including the zero solution) is so that the eigenspace will be a *vector space*, for recall that a vector space must contain a zero vector.

Gauss elimination of which gives

$$\begin{bmatrix} -3 & 2 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{6}$$

The solution is $x_3 = \alpha$ (arbitrary), $x_2 = \alpha$, $x_1 = \alpha$ so, using $\mathbf{e}$ in place of $\mathbf{x}$,

$$\mathbf{e} = \begin{bmatrix} \alpha \\ \alpha \\ \alpha \end{bmatrix} = \alpha \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}. \tag{7}$$

Thus, the eigenspace corresponding to $\lambda_1 = 5$ is span $\{[1, 1, 1]^{\mathrm{T}}\}$, the latter being a one-dimensional subspace of $\mathbb{R}^3$, namely, the line through the origin given by (7).

$\lambda_2 = 1$: Then $(\mathbf{A} - \lambda_2\mathbf{I})\mathbf{x} = \mathbf{0}$ becomes

$$\begin{bmatrix} 2-1 & 2 & 1 \\ 1 & 3-1 & 1 \\ 1 & 2 & 2-1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 2 & 1 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \tag{8}$$

Gauss elimination of which gives

$$\begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{9}$$

The solution is $x_3 = \beta$ (arbitrary), $x_2 = \gamma$ (arbitrary), $x_1 = -\beta - 2\gamma$ so

$$\mathbf{e} = \begin{bmatrix} -\beta - 2\gamma \\ \gamma \\ \beta \end{bmatrix} = \beta \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} + \gamma \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}. \tag{10}$$

Thus, the eigenspace corresponding to $\lambda_2 = 1$ is span $\{[-1, 0, 1]^{\mathrm{T}}, [-2, 1, 0]^{\mathrm{T}}\}$, the latter being a two-dimensional subspace of $\mathbb{R}^3$, namely, the plane through the origin, spanned by $[-1, 0, 1]^{\mathrm{T}}$ and $[-2, 1, 0]^{\mathrm{T}}$. In fact, the equation of that plane is seen, in (9), as $x_1 + 2x_2 + x_3 = 0$.

COMMENT 1. We can determine eigenvectors only up to arbitrary scale factors [such as the α in (7)] because, if a vector $\mathbf{e}$ satisfies $\mathbf{A}\mathbf{e} = \lambda\mathbf{e}$, then so does any scalar multiple of $\mathbf{e}$. Along these lines, observe that it would be correct to write $\mathbf{e} = \beta[1, 0, -1]^{\mathrm{T}} + \gamma[-2, 1, 0]^{\mathrm{T}}$, say, since the scale factor of -1 in the first vector can be absorbed by the arbitrary β.

COMMENT 2. In the language of Section 10.5, the rank of the $\mathbf{A} - \lambda_1\mathbf{I}$ coefficient matrix in (6) is 2 so $n - r = 3 - 2 = 1$, and (6) admits a one-parameter family of solutions. That is, the nullity of $\mathbf{A} - \lambda_1\mathbf{I}$ is 1 and the $\mathbf{e}_1$ eigenspace is one-dimensional. Similarly, the rank of $\mathbf{A} - \lambda_2\mathbf{I}$ in (9) is 1, so $n - r = 3 - 1 = 2$, and (9) admits a two-parameter family of solutions. That is, the nullity of $\mathbf{A} - \lambda_2\mathbf{I}$ is 2, and the $\mathbf{e}_2$ eigenspace is two-dimensional. However, there is no reason to believe that the multiplicity of an eigenvalue necessarily

equals the dimension of the corresponding eigenspace, even though it happens to be true in this example. ∎

EXAMPLE 2. The matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{11}$$

has the characteristic equation

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 1-\lambda & 0 & 1 \\ 1 & 1-\lambda & 0 \\ 0 & 0 & 1-\lambda \end{vmatrix} = (1-\lambda)^3 = 0 \tag{12}$$

with roots $\lambda = 1, 1, 1$. That is, $\lambda_1 = 1$ is a root of multiplicity three. To find the eigenspace, write out $(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{x} = \mathbf{0}$ as

$$\begin{bmatrix} 1-1 & 0 & 1 \\ 1 & 1-1 & 0 \\ 0 & 0 & 1-1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{13}$$

The solution is $x_3 = 0$, $x_1 = 0$, $x_2 = \alpha$ (arbitrary) so

$$\mathbf{e} = \begin{bmatrix} 0 \\ \alpha \\ 0 \end{bmatrix} = \alpha \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}. \tag{14}$$

Thus, in this case an eigenvalue of multiplicity three gave rise to an eigenspace of dimension one. ∎

With the "mechanics" of the eigenvalue problem explained in the first two examples, let us devote the next examples to applications.

EXAMPLE 3. *Solution of Differential Equations.* How can we solve the coupled differential equations

$$\begin{aligned} x' &= x + 4y, \\ y' &= x + y \end{aligned} \tag{15}$$

on $x(t)$ and $y(t)$? We could use the method of elimination (Section 3.9) to uncouple them, or solve by the Laplace transform method (Chapter 5). Here, we pursue a different approach, that will lead to an eigenvalue problem.

Since (15) is linear, constant-coefficient, and homogeneous, we can find exponential solutions. Thus, seek x, y in the form

$$x(t) = q_1 e^{rt}, \quad y(t) = q_2 e^{rt}, \tag{16}$$

where q_1, q_2, r are constants that are to be determined. Putting (16) into (15) gives

$$\begin{aligned} rq_1e^{rt} &= q_1e^{rt} + 4q_2e^{rt}, \\ rq_2e^{rt} &= q_1e^{rt} + q_2e^{rt}, \end{aligned} \tag{17}$$

or, cancelling the e^{rt}'s (because they are nonzero) and expressing the result in matrix form, gives

$$\begin{bmatrix} 1 & 4 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = r \begin{bmatrix} q_1 \\ q_2 \end{bmatrix}$$

or

$$\mathbf{A}\mathbf{q} = r\mathbf{q}, \tag{18}$$

which is an eigenvalue problem with $\lambda = r$.

We are not interested in the trivial solution $\mathbf{q} = \mathbf{0}$ because it gives the trivial particular solution $x(t) = y(t) = 0$, of (15), whereas we seek the general solution.

Proceeding as above, we obtain these eigenvalues and eigenspaces:

$$\lambda_1 = 3, \ \mathbf{e}_1 = \alpha \begin{bmatrix} 2 \\ 1 \end{bmatrix}; \qquad \lambda_2 = -1, \ \mathbf{e}_2 = \beta \begin{bmatrix} -2 \\ 1 \end{bmatrix}. \tag{19}$$

Denoting $\mathbf{x}(t) = [x(t), y(t)]^{\mathrm{T}}$, each "eigenpair" gives a solution of (15) as

$$\mathbf{x}(t) = \alpha \begin{bmatrix} 2 \\ 1 \end{bmatrix} e^{3t} \quad \text{and} \quad \mathbf{x}(t) = \beta \begin{bmatrix} -2 \\ 1 \end{bmatrix} e^{-t}. \tag{20}$$

By the linearity of (15), we can superimpose these solutions and thereby obtain the general solution

$$\mathbf{x}(t) = \alpha \begin{bmatrix} 2 \\ 1 \end{bmatrix} e^{3t} + \beta \begin{bmatrix} -2 \\ 1 \end{bmatrix} e^{-t} \tag{21}$$

or, in scalar form,

$$\begin{aligned} x(t) &= 2\alpha e^{3t} - 2\beta e^{-t}, \\ y(t) &= \alpha e^{3t} + \beta e^{-t}. \end{aligned} \tag{22}$$

Of course, α and β are the integration constants (usually denoted as A, B or C_1, C_2 in the ODE chapters).

COMMENT 1. Since we use q_1 and q_2 in (16), it would be natural to wonder why we don't also allow for different exponents, and seek $x(t) = q_1 \exp(r_1 t)$ and $y(t) = q_2 \exp(r_2 t)$. The reason is that unless $r_1 = r_2$ we obtain only the trivial solution $q_1 = q_2 = 0$ (Exercise 1).

COMMENT 2. The method illustrated in this example can be used for any system of coupled, linear, constant-coefficient homogeneous differential equations. However, it will fail to produce a general solution if $\mathbf{A}$ has a repeated eigenvalue of multiplicity k if the dimension of the corresponding eigenspace is less than k. ∎

EXAMPLE 4. *Markov Process.* Suppose that there is a population exchange between Delaware, Maryland, and Pennsylvania such that, each year, 20% of Delaware's residents move to Maryland and 8% move to Pennsylvania; 12% of Maryland's residents move to

Delaware and 10% to Pennsylvania; 10% of Pennsylvania's residents move to Delaware and 3% move to Maryland. For simplicity, let us ignore gains in population due to births and losses due to deaths – or, equivalently, suppose that these effects are nonzero, but equal and opposite, so as to cancel. Further, let us suppose that the three states are a closed system; that is, they exchange populations only among themselves.

If we denote the populations at the end of the nth year, in DE, MD, and PA as x_n, y_n, z_n, respectively, then (see Fig. 1)

$$\begin{aligned} x_{n+1} &= x_n - (0.2 + 0.08)x_n + 0.12y_n + 0.1z_n, \\ y_{n+1} &= y_n + 0.2x_n - (0.12 + 0.1)y_n + 0.03z_n, \\ z_{n+1} &= z_n + 0.08x_n + 0.1y_n - (0.1 + 0.03)z_n, \end{aligned} \tag{23}$$

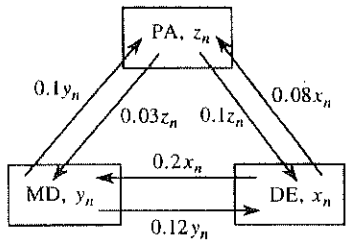


Figure 1. Population exchange model.

which are coupled *difference equations*. Difference equations were studied in Section 6.5.3 within the context of differential equations. Here, however, let us consider (23) as a matrix equation,

$$\begin{bmatrix} x_{n+1} \\ y_{n+1} \\ z_{n+1} \end{bmatrix} = \begin{bmatrix} 0.72 & 0.12 & 0.1 \\ 0.2 & 0.78 & 0.03 \\ 0.08 & 0.1 & 0.87 \end{bmatrix} \begin{bmatrix} x_n \\ y_n \\ z_n \end{bmatrix} \tag{24}$$

or,

$$\mathbf{p}_{n+1} = \mathbf{A}\mathbf{p}_n, \tag{25}$$

where $\mathbf{p}_n = [x_n, y_n, z_n]^{\mathrm{T}}$ is the "population vector."

The first problem that we pose is to find the population $\mathbf{p}_n$ as a function of n, n being essentially a discrete time variable, given some initial population $\mathbf{p}_0$. That's easy, because (25) gives $\mathbf{p}_1 = \mathbf{A}\mathbf{p}_0$, $\mathbf{p}_2 = \mathbf{A}\mathbf{p}_1 = \mathbf{A}(\mathbf{A}\mathbf{p}_0) = \mathbf{A}^2\mathbf{p}_0$, $\mathbf{p}_3 = \mathbf{A}\mathbf{p}_2 = \mathbf{A}(\mathbf{A}^2\mathbf{p}_0) = \mathbf{A}^3\mathbf{p}_0$, and so on, so

$$\mathbf{p}_n = \mathbf{A}^n\mathbf{p}_0. \tag{26}$$

We wonder whether $\mathbf{A}^n\mathbf{p}_0$ keeps changing as n increases, or whether it settles down and approaches an equilibrium (or steady-state) vector, say $\mathbf{P}$. If there is such an equilibrium vector then, by definition of equilibrium, $\mathbf{p}_{n+1} = \mathbf{p}_n = \mathbf{P}$, so (25) becomes $\mathbf{P} = \mathbf{A}\mathbf{P}$. Surely, $\mathbf{P} = \mathbf{0}$ satisfies

$$\mathbf{A}\mathbf{P} = \mathbf{P}, \tag{27}$$

but the interesting question is whether or not there exist *non*trivial $\mathbf{P}$'s. In fact, (27) is an eigenvalue problem with $\lambda = 1$, so we can say that nontrivial equilibrium vectors exist if and only if 1 is an eigenvalue of $\mathbf{A}$. As explained at the end of this section, we can use *Maple* to obtain the following eigenvalues and eigenspaces of $\mathbf{A}$:

$$\begin{gathered} \lambda_1 = 1,\ \mathbf{e}_1 = \alpha \begin{bmatrix} -0.56 \\ -0.62 \\ -0.82 \end{bmatrix}; \qquad \lambda_2 = 0.77,\ \mathbf{e}_2 = \beta \begin{bmatrix} -0.09 \\ -0.70 \\ 0.79 \end{bmatrix}; \\ \lambda_3 = 0.60,\ \mathbf{e}_3 = \gamma \begin{bmatrix} -0.62 \\ 0.70 \\ -0.07 \end{bmatrix}. \end{gathered} \tag{28}$$

(Actually, *Maple* gave the λ_j's and $\mathbf{e}_j$'s to nine and ten significant figures, respectively, but we have rounded off for brevity.)

Sure enough, $\lambda = 1$ is among the eigenvalues of $\mathbf{A}$ so there is an equilibrium population vector $\mathbf{P}$ given by the corresponding eigenvector

$$\mathbf{P} = \alpha \begin{bmatrix} 0.56 \\ 0.62 \\ 0.82 \end{bmatrix}, \tag{29}$$

where we have absorbed a factor of -1 into the scale factor α, to avoid the appearance of negative populations. If desired, we can compute α by conserving the total population: $0.56\alpha + 0.62\alpha + 0.82\alpha = x_0 + y_0 + z_0$, so $\alpha = (x_0 + y_0 + z_0)/2.0$.*

Finally, it is important to determine whether or not the equilibrium is stable for it will be observed only if it is stable, just as marbles are found in valleys but not on hilltops. To address the question of stability, let us use the set of LI eigenvectors $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, with any nonzero values of α, β, and γ, as a basis for $\mathbb{R}^3$, and expand the initial vector $\mathbf{p}_0$ in terms of that basis as

$$\mathbf{p}_0 = c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + c_3\mathbf{e}_3. \tag{30}$$

Then (25) gives

$$\begin{aligned}
\mathbf{p}_1 = \mathbf{A}\mathbf{p}_0 &= c_1\mathbf{A}\mathbf{e}_1 + c_2\mathbf{A}\mathbf{e}_2 + c_3\mathbf{A}\mathbf{e}_3 \\
&= c_1\lambda_1\mathbf{e}_1 + c_2\lambda_2\mathbf{e}_2 + c_3\lambda_3\mathbf{e}_3, \\
\mathbf{p}_2 = \mathbf{A}\mathbf{p}_1 &= c_1\lambda_1\mathbf{A}\mathbf{e}_1 + c_2\lambda_2\mathbf{A}\mathbf{e}_2 + c_3\lambda_3\mathbf{A}\mathbf{e}_3 \\
&= c_1\lambda_1^2\mathbf{e}_1 + c_2\lambda_2^2\mathbf{e}_2 + c_3\lambda_3^2\mathbf{e}_3, \\
&\vdots \\
\mathbf{p}_n = \mathbf{A}\mathbf{p}_{n-1} &= c_1\lambda_1^n\mathbf{e}_1 + c_2\lambda_2^n\mathbf{e}_2 + c_3\lambda_3^n\mathbf{e}_3 \\
&= c_1\mathbf{e}_1 + c_2(0.77)^n\mathbf{e}_2 + c_3(0.60)^n\mathbf{e}_3 \\
&\sim c_1\mathbf{e}_1
\end{aligned} \tag{31}$$

as $n \to \infty$, provided that $c_1 \neq 0$. In fact, c_1 cannot be zero because if it were zero then (31) would give $\mathbf{p}_n \to \mathbf{0}$ as $n \to \infty$ and, since the total population is conserved, that could happen only in the uninteresting case where $\mathbf{p}_0 = \mathbf{0}$. Thus, we see from (31) that $\mathbf{p}_n$ inevitably tends to a multiple of $\mathbf{e}_1$, namely, to the equilibrium vector $\mathbf{P}$.

The upshot is that the population history is given by (26), and that $\mathbf{p}_n$ inevitably tends to a unique steady state which is some scalar multiple of $\mathbf{e}_1$, the multiple being fixed by the conservation of the total population.

COMMENT. This example incorporates a number of linear algebra concepts: matrix multiplication in expressing (23) compactly as (25) and in deriving the solution (26) for $\mathbf{p}_n$; the eigenvalue problem in regard to the possibility of a steady-state solution $\mathbf{P}$; and bases and expansions in assessing the stability of that steady state, in (30)–(31). Consider how effective are these linear algebra concepts and methods in providing a systematic approach to solving this problem, especially in determining the stability of the steady state. The same

*Recall that we built into (24) the assumption that any births and deaths cancel, in number, as revealed by adding the three scalar equations in (24), for that step gives $x_{n+1} + y_{n+1} + z_{n+1} = x_n + y_n + z_n$.

approach applies whether the system includes only three states, as in the present example, or 30 states. ∎

The matrix $\mathbf{A}$ in Example 4 is an example of a "Markov" matrix. An $n \times n$ matrix $\mathbf{A} = \{a_{ij}\}$ is called a **Markov (or stochastic) matrix** if $a_{ij} \geq 0$ for each i, j, and if the elements of each column sum to unity, or if the elements of each row sum to unity. [In the case of the $\mathbf{A}$ given in (24) its columns sum to unity.] It was no coincidence that $\lambda = 1$ was an eigenvalue, in Example 4 since $\lambda = 1$ is an eigenvalue of *every* Markov matrix.

To prove that claim, suppose that $\mathbf{A}$ is a Markov matrix. The value $\lambda = 1$ will be among the eigenvalues of $\mathbf{A}$ if and only if $\mathbf{Ax} = \mathbf{x}$ has nontrivial solutions for $\mathbf{x}$ or, equivalently, if the rows or columns of $\mathbf{A} - \mathbf{I}$ are linearly dependent. Since $\mathbf{A}$ is a Markov matrix, either the elements of each of its columns sum to unity or the elements of each of its rows sum to unity. It follows that either the elements of each of the columns of $\mathbf{A} - \mathbf{I}$ sum to zero (in which case the rows of $\mathbf{A} - \mathbf{I}$ are linearly dependent) or the elements of each of the rows of $\mathbf{A} - \mathbf{I}$ sum to zero (in which case the columns of $\mathbf{A} - \mathbf{I}$ are linearly dependent), or both. Thus, $\mathbf{A} - \mathbf{I}$ is singular and our claim is proved.

11.2.2. Application to elementary singularities in the phase plane. If you studied Chapter 7, you will recall the fundamental role of the *elementary singularities* in the x, y phase plane, where $x(t)$ and $y(t)$ satisfy the linear ODE's

$$\begin{aligned} x' &= ax + by \\ y' &= cx + dy. \end{aligned} \tag{32}$$

For instance, the system (15) is of that form, so let us reconsider the result, given by (21) and (22), in terms of the x, y phase plane. If $\beta = 0$, then $x(t) = 2\alpha e^{3t}$ and $y(t) = \alpha e^{3t}$, so the phase trajectory is the line $y = x/2$, and if $\alpha = 0$ then $x(t) = -2\beta e^{-t}$ and $y(t) = \beta e^{-t}$, so the phase trajectory is the line $y = -x/2$. These are shown in Fig. 2. The directions of the arrows follow from the fact that e^{3t} increases with t and e^{-t} decreases (since the λ's are of opposite sign), and they imply that (15) has a *saddle* at the origin.

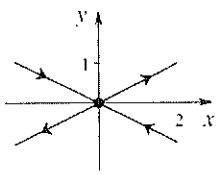


Figure 2. Significance of the eigenvectors.

More generally, observe that we can classify the singularity directly from the eigenvalues of the $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ matrix in (32). Let a, b, c, d be real. Then the eigenvalues are either real or they are complex conjugates. In general, then, we can write $\lambda = \alpha \pm i\beta$, and these λ's contribute solutions of the form $e^{(\alpha \pm i\beta)t}$. We have these possibilities:

λ's real and of the same sign:

λ's $< 0 \Rightarrow$ stable node

λ's $> 0 \Rightarrow$ unstable node

λ's real and of opposite sign $\Rightarrow$ saddle

$$\begin{aligned} &\lambda\text{'s complex}(\lambda = \alpha \pm i\beta):\\ &\qquad \alpha < 0 \;\Rightarrow\; \text{stable focus}\\ &\qquad \alpha = 0 \;\Rightarrow\; \text{center}\\ &\qquad \alpha > 0 \;\Rightarrow\; \text{unstable focus} \end{aligned}$$

In the case of an improper node or saddle, the eigenvectors give the stable and/or unstable manifolds, as mentioned above for the system (15), for which case the stable and unstable manifolds are shown in Fig. 2.

Closure. It is important to remember that the eigenvalue problem $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ is *homogeneous* [since it is equivalent to $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$] and that the whole point is to find *non*trivial solutions. If, for a given eigenvalue λ, you solve $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ by Gauss elimination and obtain $\mathbf{e} = \mathbf{0}$, then your calculations are incorrect: either your eigenvalue is incorrect and/or the Gauss elimination is incorrect.

Observe that our solution strategy uncouples the calculation of the eigenvalues and the eigenvectors: first we solve the characteristic polynomial equation $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$ for the λ's, and then for each λ we solve $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$, by Gauss elimination, for the corresponding eigenvectors.

Computer software. In *Maple*, the relevant commands are **eigenvals** and **eigenvects**, both of which are in the linalg package. The command eigenvals gives just the eigenvalues, and eigenvects gives the eigenvalues, their multiplicity, and a basis for each eigenspace. For instance, let $\mathbf{A}$ be the matrix in Example 1. First, enter

```
with(linalg):
```

and return. Then type

```
A := matrix (3, 3, [2, 2, 1, 1, 3, 1, 1, 2, 2]):
```

because $\mathbf{A}$ is 3×3 and its rows are $2, 2, 1$, and $1, 3, 1$, and $1, 2, 2$, in turn. Then

```
eigenvals(A);
```

gives the eigenvalues as

$$5, 1, 1$$

and

```
eigenvects(A);
```

gives both the eigenvalues and the eigenvectors as

$$[5, 1, \{[1, 1, 1]\}], \quad [1, 2, \{[-2, 1, 0], [-1, 0, 1]\}]$$

In place of the eigenvals command, one can use fsolve to obtain the roots of the characteristic equation, but that is less convenient since to obtain the characteristic equation one needs to expand the $n \times n$ determinant of $\mathbf{A} - \lambda\mathbf{I}$.

Suppose we want the eigenvalues of $\mathbf{A}^{20}$. Enter

```
A := matrix (3, 3, [2, 2, 1, 1, 3, 1, 1, 2, 2]):
```

then

```
evalm(A^20);
```

and then

```
eigenvals(");
```

The quotation mark saves us the trouble of entering the matrix $\mathbf{A}^{20}$ that was calculated in the preceding step.

EXERCISES 11.2

1. (*Example 3*) (a) Derive the eigenvalues and eigenspaces given in (19).
(b) Show that if we assume the forms $x(t) = q_1 \exp(r_1 t)$ and $y(t) = q_2 \exp(r_2 t)$, then we obtain only the trivial solution unless $r_1 = r_2$, as claimed in COMMENT 1.

2. (*Example 4*) To determine the stability of the equilibrium solution $\mathbf{P}$, we expanded $\mathbf{p}_0$ in terms of the basis consisting of the eigenvectors of $\mathbf{A}$. Explain why that choice is particularly convenient. HINT: If necessary, you could try using a different basis, such as $\{[1,0,0]^{\mathrm{T}}, [0,1,0]^{\mathrm{T}}, [0,0,1]^{\mathrm{T}}\}$, instead.

3. Find the eigenvalues and eigenspaces, as well as a basis for each eigenspace.

(a) $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & -3 \\ 0 & 0 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ (d) $\begin{bmatrix} -3 & 2 \\ 6 & -4 \end{bmatrix}$

(e) $\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ (f) $\begin{bmatrix} 0 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

(g) $\begin{bmatrix} 2 & 0 & 0 \\ 0 & -5 & 0 \\ 0 & 0 & 4 \end{bmatrix}$ (h) $\begin{bmatrix} 2 & 1 & 6 \\ 0 & -5 & 3 \\ 0 & 0 & 4 \end{bmatrix}$

(i) $\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$ (j) $\begin{bmatrix} 4 & 4 & 4 \\ 4 & 4 & 4 \\ 4 & 4 & 4 \end{bmatrix}$

(k) $\begin{bmatrix} 1 & 0 & 2 \\ 1 & 0 & 2 \\ 1 & 0 & 2 \end{bmatrix}$ (l) $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$

(m) $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 4 & 0 & 4 \end{bmatrix}$ (n) $\begin{bmatrix} 0 & 0 & 3 \\ 0 & 0 & 1 \\ 1 & -13 & 7 \end{bmatrix}$

(o) $\begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ (p) $\begin{bmatrix} 0 & 2 & 0 \\ 0 & 3 & 0 \\ 0 & 4 & 0 \end{bmatrix}$

(q) $\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}$ (r) $\begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 4 & 4 & 4 & 4 \end{bmatrix}$

4. (a)–(r) Use computer software to find the eigenvalues and eigenspaces for the matrix in the corresponding part of Exercise 3.

5. Is the following an eigenvector of the matrix $\mathbf{A}$? Explain.

$$\mathbf{A} = \begin{bmatrix} 1 & 8 & 5 & 3 \\ 2 & 16 & 10 & 6 \\ 5 & -14 & -11 & -3 \\ -1 & -8 & -5 & -3 \end{bmatrix}.$$

(a) $[1,2,-1,3]^{\mathrm{T}}$ (b) $[1,2,-4,-1]^{\mathrm{T}}$ (c) $[1,2,1,1]^{\mathrm{T}}$
(d) $[1,0,1,-2]^{\mathrm{T}}$ (e) $[1,0,1,-1]^{\mathrm{T}}$ (f) $[1,1,0,-3]^{\mathrm{T}}$
(g) $[1,2,-1,-1]^{\mathrm{T}}$ (h) $[2,1,0,1]^{\mathrm{T}}$ (i) $[2,1,1,-5]^{\mathrm{T}}$

6. The given matrix has $\lambda = 2$ among its eigenvalues. Find the eigenspace corresponding to that eigenvalue.

(a) $\begin{bmatrix} 3 & 2 & 2 & 1 \\ 2 & 3 & 1 & 2 \\ -1 & 1 & 2 & 0 \\ 2 & 4 & 3 & 5 \end{bmatrix}$ (b) $\begin{bmatrix} 3 & 1 & 2 & 1 \\ -1 & 3 & 1 & 2 \\ 0 & 2 & 5 & 3 \\ 1 & 3 & 5 & 6 \end{bmatrix}$

(c) $\begin{bmatrix} 3 & 1 & 1 & 1 \\ 1 & 3 & 1 & 1 \\ 1 & 1 & 3 & 1 \\ 1 & 1 & 1 & 3 \end{bmatrix}$ (d) $\begin{bmatrix} 3 & 0 & 1 & 1 \\ 1 & 3 & 0 & 1 \\ -1 & 1 & 3 & -1 \\ 1 & 2 & 2 & 3 \end{bmatrix}$

7. It is known that the $n \times n$ tridiagonal matrix

$$\mathbf{A} = \begin{bmatrix} b & c & 0 & 0 & & & \cdots & 0 \\ a & b & c & 0 & & & & \vdots \\ 0 & a & b & c & & & & \\ & & & & \ddots & & & \vdots \\ \vdots & & & & & a & b & c \\ 0 & \cdots & & & \cdots & 0 & a & b \end{bmatrix}$$

has eigenvalues

$$\lambda_j = b + 2\sqrt{ac}\cos\frac{j\pi}{n+1} \tag{7.1}$$

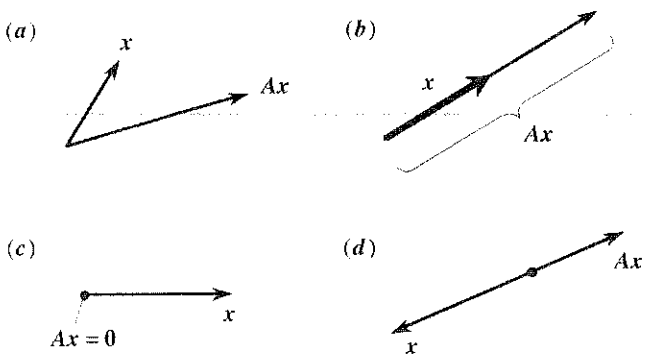
for $j = 1, 2, \ldots, n$. ($\mathbf{A}$ is called **tridiagonal** because all elements are zero except for those on the main diagonal and the two adjacent diagonals.)

(a) Verify (7.1) by calculating the eigenvalues for $n = 1$ and $n = 2$.
(b) Verify (7.1) by using computer software to determine the eigenvalues for $n = 1$ and 2 and 3.
(c) Verify (7.1) by using computer software to determine the eigenvalues for $n = 4$ and $a = 1, b = 2, c = 1$.
(d) Same as (c), for $n = 4$ and $a = 2, b = 3, c = -1$.
(e) Same as (c), for $n = 5$ and $a = 1, b = 5, c = 3$.

8. Is it possible for a matrix to have no eigenvalues? Explain.

9. We saw in Example 1 that a given eigenvalue can have more than one LI eigenvector. Can a given eigenvector correspond to more than one eigenvalue? Explain.

10. Let $\mathbf{x}$ and $\mathbf{Ax}$ be as shown. Is $\mathbf{x}$ an eigenvector of $\mathbf{A}$? If so, estimate the corresponding eigenvalue; if not, explain why not.



11. Show that the eigenvalues of $k\mathbf{A}$, for any scalar k, are k times those of $\mathbf{A}$. Are the corresponding eigenspaces the same? Explain.

12. Show that the eigenvalues of $\mathbf{A}^{\mathrm{T}}$ are the same as those of $\mathbf{A}$. Is the eigenspace corresponding to an eigenvalue λ of $\mathbf{A}$ the same as the eigenspace corresponding to the same eigenvalue λ of $\mathbf{A}^{\mathrm{T}}$? Prove or disprove.

13. If λ, $\mathbf{e}$ are an eigenvalue and corresponding eigenvector of a matrix $\mathbf{A}$, show that $\lambda = (\mathbf{e}^{\mathrm{T}}\mathbf{Ae})/(\mathbf{e}^{\mathrm{T}}\mathbf{e})$. HINT: Recall that the dot product of two column vectors $\mathbf{u}$ and $\mathbf{v}$ is $\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^{\mathrm{T}}\mathbf{v}$.

14. Show that if $\mathbf{A}$ is triangular, its eigenvalues are simply the diagonal elements of $\mathbf{A}$.

15. Show that if λ is an eigenvalue of $\mathbf{A}$, with a corresponding eigenvector $\mathbf{e}$, then λ^n is an eigenvalue of $\mathbf{A}^n$, with the same eigenvector $\mathbf{e}$, for any integer n. (Of course, if n is negative, $\mathbf{A}$ needs to be nonsingular if $\mathbf{A}^n$ is to exist in the first place.) HINT: Pre-multiply $\mathbf{Ae} = \lambda\mathbf{e}$ by $\mathbf{A}, \mathbf{A}^2, \ldots$.

16. Use the results stated in the preceding two exercises to determine the eigenvalues and eigenspaces of $\mathbf{A}^{10}$ for each of the following $\mathbf{A}$ matrices. Check your results by working out $\mathbf{A}^{10}$ and its eigenvalues and eigenvectors.

(a) $\begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 1 & 3 \end{bmatrix}$ (b) $\begin{bmatrix} 3 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & -2 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 1 & 2 \\ 0 & -1 & 2 \\ 0 & 0 & 2 \end{bmatrix}$

17. For the given $\mathbf{A}$ matrix, use computer software to determine its eigenvalues and eigenspaces. Then, use computer software to obtain $\mathbf{A}^5$ and to determine its eigenvalues and eigenspaces. Then, verify the result stated in Exercise 15, for this case.

(a) $\begin{bmatrix} 2 & 2 & 0 \\ 0 & 2 & 0 \\ 2 & 0 & 2 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 3 & 3 & 3 \end{bmatrix}$

18. (*Similar matrices*) (a) Suppose that $\mathbf{Ax} = \mathbf{y}$, where $\mathbf{A}$ is square. Setting $\mathbf{x} = \mathbf{Q}\tilde{\mathbf{x}}$ and $\mathbf{y} = \mathbf{Q}\tilde{\mathbf{y}}$, where $\mathbf{Q}$ is invertible, show that

$$\tilde{\mathbf{A}}\tilde{\mathbf{x}} = \tilde{\mathbf{y}}, \tag{18.1}$$

where

$$\tilde{\mathbf{A}} = \mathbf{Q}^{-1}\mathbf{AQ}. \tag{18.2}$$

Given any invertible matrix $\mathbf{Q}$, matrices $\mathbf{A}$ and $\widetilde{\mathbf{A}}$ related by (18.2) are said to be **similar**.

(b) Show that if $\mathbf{A}$ and $\widetilde{\mathbf{A}}$ are similar, then they have the same characteristic polynomials and hence the same eigenvalues.

19. (*The characteristic polynomial*) Let us write the characteristic equation $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$ in the standard form

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \vdots & & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix}$$
$$\equiv (-1)^n[\lambda^n - \beta_1\lambda^{n-1} + \beta_2\lambda^{n-2} - \cdots + (-1)^n\beta_n] = 0. \tag{19.1}$$

If we denote the n roots (which need not be real) as $\lambda_1, \lambda_2, \ldots, \lambda_n$, numbering repeated roots separately, we may factor

$$\lambda^n - \beta_1\lambda^{n-1} + \cdots + (-1)^n\beta_n = (\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_n). \tag{19.2}$$

Multiplying out the right-hand side of (19.2) and equating coefficients of like powers of λ, on both sides of the equation, yields the relations

$$\begin{aligned} \beta_1 &= \lambda_1 + \lambda_2 + \cdots + \lambda_n, \\ \beta_2 &= \lambda_1\lambda_2 + \lambda_1\lambda_3 + \cdots + \lambda_{n-1}\lambda_n, \\ \beta_3 &= \lambda_1\lambda_2\lambda_3 + \cdots + \lambda_{n-2}\lambda_{n-1}\lambda_n, \\ &\vdots \\ \beta_n &= \lambda_1\lambda_2\cdots\lambda_n. \end{aligned} \tag{19.3}$$

For example, if $n = 4$, then $\beta_2 = \lambda_1\lambda_2 + \lambda_1\lambda_3 + \lambda_1\lambda_4 + \lambda_2\lambda_3 + \lambda_2\lambda_4 + \lambda_3\lambda_4$ and $\beta_3 = \lambda_1\lambda_2\lambda_3 + \lambda_1\lambda_2\lambda_4 + \lambda_1\lambda_3\lambda_4 + \lambda_2\lambda_3\lambda_4$.

Alternatively, one may expand the determinant in (19.1) directly, and identify the coefficients of the various powers of λ in terms of certain subdeterminants of $\mathbf{A}$, and hence determine the β_j's in terms of these subdeterminants. The result, we state without proof, is that β_j (for $j = 1, \ldots, n$) *is the sum of all the jth-order principal minors of* $\mathbf{A}$. (By a "principal minor" we mean the determinant of a submatrix of $\mathbf{A}$, whose main diagonal lies along that of $\mathbf{A}$.) Thus,

$$\begin{aligned} \beta_1 &= a_{11} + a_{22} + \cdots + a_{nn} \equiv \mathrm{tr}\mathbf{A}, \\ &\vdots \\ \beta_n &= \det\mathbf{A}, \end{aligned} \tag{19.4}$$

where the sum $a_{11} + a_{22} + \cdots + a_{nn}$ is called the **trace** of $\mathbf{A}$ and is denoted here as $\mathrm{tr}\mathbf{A}$.

(a) Verify equations (19.4) for $n = 2$ by expanding $\det(\mathbf{A} - \lambda\mathbf{I})$.

(b) Verify equations (19.4) for $n = 3$ by expanding $\det(\mathbf{A} - \lambda\mathbf{I})$; i.e., verify that

$$\beta_1 = a_{11} + a_{22} + a_{33},$$
$$\beta_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} + \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix} + \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix},$$
$$\beta_3 = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}.$$

(c) Finally, comparing equations (19.3) with equations (19.4), show that

$$\begin{aligned} \lambda_1 + \lambda_2 + \cdots + \lambda_n &= \mathrm{tr}\mathbf{A}, \\ &\vdots \\ \lambda_1\lambda_2\cdots\lambda_n &= \det\mathbf{A}, \end{aligned} \tag{19.5}$$

and hence that $\mathbf{A}$ is singular if and only if at least one of its eigenvalues is zero.

20. Show that if two $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ have the same eigenvalues $\lambda_1, \ldots, \lambda_n$ and the same n LI eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$, then it must be true that $\mathbf{A} = \mathbf{B}$.

21. Can an $n \times n$ matrix have more than n LI eigenvectors? Explain.

22. (*Markov matrices*) Recall the definition of a Markov matrix, and the fact every Markov matrix contains $\lambda = 1$ among its eigenvalues. You may use the result stated in Exercise 11, if you need it.

(a) Find the eigenvalues of

$$\mathbf{A} = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

(b) Find one eigenvalue of

$$\mathbf{A} = \begin{bmatrix} 8 & 10 & 12 \\ 9 & 10 & 11 \\ 10 & 10 & 10 \end{bmatrix}.$$

(c) Find one eigenvalue of

$$\mathbf{A} = \begin{bmatrix} 3 & 1 & 2 \\ 2 & 2 & 1 \\ 0 & 2 & 2 \end{bmatrix}.$$

(d) Determine the eigenvalues and a basis for each eigenspace for the 20×20 matrix $\mathbf{A}$ having unity for each of its 400 elements.
(e) Determine the eigenvalues and a basis for each eigenspace for the 30×30 matrix $\mathbf{A}$ having 2 for each of its 900 elements.

23. In each case, use the same method as in Example 3 to find the general solution of the given system of coupled differential equations, if possible. If your solution falls short of being a general solution, explain why that happened. Primes denote d/dt.

(a) $x' = x + 2y$, $y' = 3x + 6y$
(b) $x' = x + y$, $y' = x + y$
(c) $x'' = 2x + y$, $y'' = 9x + 2y$
(d) $x'' = x + y$, $y'' = x + y$
(e) $x' = 4x + y$, $y' = -5x + z$, $z' = x - y - z$
(f) $x' = y - z$, $y' = -5x + 4y + z$, $z' = x - z$
(g) $x' = 2x - y$, $y' = y - z$, $z' = -x + y$
(h) $x'' = -x + y$, $y'' = -x + z$, $z'' = x - 2z$

24. (*Cayley–Hamilton theorem*) The **Cayley–Hamilton theorem** states that if the characteristic equation of any square matrix $\mathbf{A}$ is $\lambda^n + \alpha_1\lambda^{n-1} + \cdots + \alpha_{n-1}\lambda + \alpha_n = 0$, then $\mathbf{A}^n + \alpha_1\mathbf{A}^{n-1} + \cdots + \alpha_{n-1}\mathbf{A} + \alpha_n\mathbf{I} = \mathbf{0}$; i.e., $\mathbf{A}$ satisfies its characteristic equation.

(a) Prove this theorem for the general 2×2 case, $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$.

(b) If $\mathbf{A} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, show that $\mathbf{A}^2 - 4\mathbf{A} + 3\mathbf{I} = \mathbf{0}$ so that

$$\mathbf{A}^{-1} = \frac{4}{3}\mathbf{I} - \frac{1}{3}\mathbf{A} = \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} \end{bmatrix}.$$

25. (*Generalized eigenvalue problem*) If $\mathbf{B} \neq \mathbf{I}$, then $\mathbf{A}\mathbf{x} = \lambda\mathbf{B}\mathbf{x}$ is called a **generalized eigenvalue problem**. It should be easy to see that in this case the characteristic equation is $\det(\mathbf{A} - \lambda\mathbf{B}) = 0$, and that the eigenvectors then follow as the nontrivial solutions of $(\mathbf{A} - \lambda\mathbf{B})\mathbf{x} = \mathbf{0}$. Find the eigenvalues and eigenspaces in each case.

(a) $\begin{bmatrix} 5 & 1 \\ 1 & 5 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda\begin{bmatrix} 8 & -4 \\ -4 & 8 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

(b) $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

(c) $\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

(d) $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \lambda\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$

26. In seeking the eigenvalues and eigenvectors of a given matrix $\mathbf{A}$, is it permissible first to simplify $\mathbf{A}$ by means of some elementary row operations? (That is, are the eigenvalues and eigenvectors of $\mathbf{A}$ invariant with respect to elementary row operations?) Explain.

27. In each case, given the values of a, b, c, d in (32), use the eigenvalues to classify the singularity as an unstable node, stable node, saddle, stable focus, center, or unstable focus. If applicable, use the eigenvectors to give a labeled sketch of any stable and/or unstable manifolds.

(a) $a = 1, b = 1, c = 3, d = -1$
(b) $a = -1, b = -3, c = 1, d = 1$
(c) $a = 4, b = 1, c = 1, d = 4$
(d) $a = -3, b = 1, c = 1, d = -3$
(e) $a = 3, b = 1, c = -1, d = 3$
(f) $a = -2, b = -2, c = 2, d = -2$
(g) $a = 1, b = 2, c = 3, d = 4$
(h) $a = 5, b = 1, c = -8, d = 1$

11.3 Symmetric Matrices

In applications, symmetric matrices arise surprisingly often, and their symmetry leads to important results regarding their eigenvalues and eigenvectors. Thus, it is important to treat this case separately.

11.3.1. Eigenvalue problem $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$. We have the following three important

results. We shall prove the third, but defer proofs for the first two to the exercises.

THEOREM 11.3.1 *Real Eigenvalues*
If $\mathbf{A}$ is symmetric ($\mathbf{A}^{\mathrm{T}} = \mathbf{A}$), then all of its eigenvalues are real.

THEOREM 11.3.2 *Dimension of Eigenspace*
If an eigenvalue λ of a symmetric matrix $\mathbf{A}$ is of multiplicity k, then the eigenspace corresponding to λ is of dimension k.

THEOREM 11.3.3 *Orthogonality of Eigenvectors*
If $\mathbf{A}$ is symmetric, then eigenvectors corresponding to distinct eigenvalues are orthogonal.

Proof: Let $\mathbf{e}_j$ and $\mathbf{e}_k$ be eigenvectors corresponding to distinct eigenvalues λ_j and λ_k, respectively. Thus,

$$\mathbf{A}\mathbf{e}_j = \lambda_j \mathbf{e}_j \quad \text{and} \quad \mathbf{A}\mathbf{e}_k = \lambda_k \mathbf{e}_k. \tag{1a,b}$$

Next, recall that the dot product of n-dimensional column vectors $\mathbf{x}$ and $\mathbf{y}$ is $\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^{\mathrm{T}}\mathbf{y}$, and that $(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$ for any matrices $\mathbf{A}$ and $\mathbf{B}$ that are conformable for multiplication. Then, if we dot $\mathbf{e}_k$ into each side of (1a) and dot each side of (1b) into $\mathbf{e}_j$ [i.e., we "pre-dot" (1a) with $\mathbf{e}_k$ and "post-dot" (1b) with $\mathbf{e}_j$], we obtain

$$\begin{aligned} \mathbf{e}_k \cdot (\mathbf{A}\mathbf{e}_j) &= \mathbf{e}_k \cdot (\lambda_j \mathbf{e}_j) \\ \mathbf{e}_k^{\mathrm{T}}\mathbf{A}\mathbf{e}_j &= \lambda_j \mathbf{e}_k^{\mathrm{T}}\mathbf{e}_j \end{aligned} \quad \left| \quad \begin{aligned} (\mathbf{A}\mathbf{e}_k) \cdot \mathbf{e}_j &= (\lambda_k \mathbf{e}_k) \cdot \mathbf{e}_j \\ (\mathbf{A}\mathbf{e}_k)^{\mathrm{T}}\mathbf{e}_j &= \lambda_k \mathbf{e}_k^{\mathrm{T}}\mathbf{e}_j \\ \mathbf{e}_k^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{e}_j &= \lambda_k \mathbf{e}_k^{\mathrm{T}}\mathbf{e}_j. \end{aligned} \tag{2}$$

But $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$ by assumption so if we subtract the bottom equations on the left and right of the vertical divider, we obtain

$$0 = (\lambda_j - \lambda_k)\mathbf{e}_k^{\mathrm{T}}\mathbf{e}_j. \tag{3}$$

Finally, $\lambda_j - \lambda_k \neq 0$ since λ_j and λ_k were assumed to be distinct so it follows from (3) that $\mathbf{e}_k^{\mathrm{T}}\mathbf{e}_j = 0$. Thus, $\mathbf{e}_k \cdot \mathbf{e}_j = 0$, as claimed. ■

As usual, be careful not to read converses into theorems. For instance, Theorem 11.3.1 says that if $\mathbf{A}$ is symmetric, then its eigenvalues are real. It does *not* say that the eigenvalues of $\mathbf{A}$ are real if and only if $\mathbf{A}$ is symmetric. For instance, in

each of the four examples of Section 11.2 the λ's are real, yet none of the matrices is symmetric.

EXAMPLE 1. For the symmetric matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}, \tag{4}$$

we find (Exercise 1) the eigenvalues and eigenspaces

$$\lambda_1 = 4,\ \mathbf{e}_1 = \alpha \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \qquad \lambda_2 = 1,\ \mathbf{e}_2 = \beta \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} + \gamma \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \tag{5}$$

where $\lambda_2 = 1$ is of multiplicity two [i.e., the characteristic polynomial can be factored as $-(\lambda-4)(\lambda-1)^2$]. Since $\mathbf{A}$ is symmetric, the theorems apply. In accordance with Theorem 11.3.1, the λ's are real; in accordance with Theorem 11.3.2, λ_1 is of multiplicity 1 and its eigenspace is one-dimensional (namely, span$\{[1,1,1]^{\mathrm{T}}\}$), and λ_2 is of multiplicity 2 and its eigenspace is two-dimensional (namely, span$\{[-1,0,1]^{\mathrm{T}}, [-1,1,0]^{\mathrm{T}}\}$); and in accordance with Theorem 11.3.3, $\mathbf{e}_1 \cdot \mathbf{e}_2 = 0$ for all choices of α, β, γ. The eigenspace $\mathbf{e}_2$ is the plane through the origin (in 3-space) that is spanned by $[-1,0,1]^{\mathrm{T}}$ and $[-1,1,0]^{\mathrm{T}}$, and the eigenspace $\mathbf{e}_1$ is the line through the origin that is spanned by $[1,1,1]^{\mathrm{T}}$ and is normal to the plane.

The vectors $[-1,0,1]^{\mathrm{T}}$ and $[-1,1,0]^{\mathrm{T}}$ in $\mathbf{e}_2$ are LI and a basis for $\mathbf{e}_2$, but happen not to be orthogonal. Their lack of orthogonality does not violate Theorem 11.3.3 since they come from the same λ, not from distinct λ's. Nonetheless, we can "trade those vectors in" for two within $\mathbf{e}_2$ that *are* orthogonal (in a nonunique way, in fact, for there is an infinite number of pairs of orthogonal vectors within that plane). For instance, we can choose

$$\mathbf{e}_2 = [-1,0,1]^{\mathrm{T}} \tag{6}$$

[i.e., by setting $\beta = 1$ and $\gamma = 0$ in (5)] and seek

$$\mathbf{e}_3 = \beta[-1,0,1]^{\mathrm{T}} + \gamma[-1,1,0]^{\mathrm{T}} \tag{7}$$

such that

$$\mathbf{e}_2 \cdot \mathbf{e}_3 = (-1)(-\beta-\gamma) + (0)(\gamma) + (1)(\beta) = 2\beta + \gamma = 0. \tag{8}$$

Choosing $\beta = 1$, say, then $\gamma = -2$, and (7) gives

$$\mathbf{e}_3 = [1,-2,1]^{\mathrm{T}}, \tag{9}$$

and the vectors given by (6) and (9) constitute an orthogonal basis for the eigenspace corresponding to the eigenvalue 1.

And since (with $\alpha = 1$, say)

$$\mathbf{e}_1 = [1,1,1]^{\mathrm{T}} \tag{10}$$

is orthogonal to each of those vectors, it follows that the eigenvectors given by (6), (9), and (10) constitute an orthogonal basis for 3-space. That is, among the eigenvectors of the 3×3 symmetric $\mathbf{A}$ given by (4) we can find an orthogonal basis for 3-space.

COMMENT 1. The $\mathbf{A}$ matrix in (4) happens to be symmetric about the other diagonal as well as about the main diagonal. That symmetry is irrelevant; by symmetry we always mean that $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$, which is symmetry about the *main* diagonal (from upper left to bottom right).

COMMENT 2. You might be thinking "Of course the eigenvalues are real, for the $\mathbf{A}$ matrix is real." No, $\mathbf{A}$ being real implies only that the coefficients are real in its characteristic equation, and a polynomial equation with real coefficients *can* have complex roots. For instance, the real but nonsymmetric matrix $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$ has the characteristic equation $\lambda^2 - 2\lambda + 2 = 0$ and the complex eigenvalues $\lambda = 1 \pm i$.

COMMENT 3. The procedure that we used to obtain the orthogonal set $\{\mathbf{e}_2, \mathbf{e}_3\}$ from the LI set $\{[-1, 0, 1]^{\mathrm{T}}, [-1, 1, 0]^{\mathrm{T}}\}$ is essentially the Gram–Schmidt orthogonalization procedure explained in Exercise 11 of Section 9.9. ∎

Generalization of the ideas contained in Example 1 yields the following theorem.

THEOREM 11.3.4 *Orthogonal Basis*
If an $n \times n$ matrix $\mathbf{A}$ is symmetric, then its eigenvectors provide an orthogonal basis for n-space.

Proof: If all of $\mathbf{A}$'s n eigenvalues are distinct then, according to Theorem 11.3.3, the n eigenspaces are orthogonal (each being a one-dimensional line in n-space) and therefore provide n orthogonal vectors, which necessarily constitute a basis for n-space. What if the eigenvalues are not distinct? Suppose that all are distinct except for one, say λ, which is of multiplicity k. Then the $n - k$ eigenvectors corresponding to the other eigenvalues are orthogonal to each other and also to all vectors in the eigenspace corresponding to λ. Further, λ's eigenspace is k-dimensional (Theorem 11.3.2) and hence contains k orthogonal vectors. Altogether, then we have $(n - k) + k = n$ orthogonal eigenvectors and hence an orthogonal basis for n-space. A similar argument applies if there is more than one repeated eigenvalue. ∎

EXAMPLE 2. *Free Vibration of a Two-Mass System.* Consider the system of two masses subjected to forces $f_1(t)$ and $f_2(t)$ and restrained laterally by springs and supported vertically by a frictionless table as shown in Fig. 1. The equations of motion, already derived in Example 3 of Section 3.9.1, are

$$\begin{aligned} m_1 x_1'' + (k_1 + k_{12})x_1 - k_{12}x_2 &= f_1(t), \\ m_2 x_2'' - k_{12}x_1 + (k_2 + k_{12})x_2 &= f_2(t). \end{aligned} \tag{11}$$

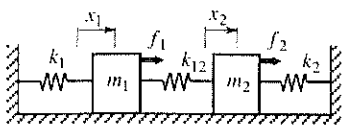


Figure 1. Two-mass system.

Let $m_1 = m_2 = k_1 = k_{12} = k_2 = 1$, say, for definiteness, and consider the *free vibration*, where $f_1(t) = f_2(t) = 0$. Then (11) becomes

$$\begin{aligned} x_1'' + 2x_1 - x_2 &= 0 \\ x_2'' - x_1 + 2x_2 &= 0. \end{aligned} \tag{12}$$

The system (12) is solved in Example 8 in Section 3.9.3, by the method of elimination, and we urge you to review that solution and to compare it with the following matrix eigenvalue problem approach. [We could also solve (12) by the Laplace transform method.]

Let us follow the same line of approach that was put forward in Example 2 of Section 11.2. Namely, seek

$$\begin{aligned} x_1(t) &= q_1 e^{\lambda t}, \\ x_2(t) &= q_2 e^{\lambda t}. \end{aligned} \tag{13}$$

Actually, on physical grounds we expect the solution to be a vibration (i.e., the λ's will turn out to be purely imaginary) so it seems more sensible to seek

$$\begin{aligned} x_1(t) &= q_1 \sin(\omega t + \phi), \\ x_2(t) &= q_2 \sin(\omega t + \phi), \end{aligned} \tag{14}$$

where the q_j's are amplitudes, ω is the frequency, and ϕ is the phase angle.* Putting (14) into (12) and canceling the $\sin(\omega t + \phi)$ factors gives

$$\begin{aligned} -\omega^2 q_1 + 2q_1 - q_2 &= 0, \\ -\omega^2 q_2 - q_1 + 2q_2 &= 0 \end{aligned}$$

or, equivalently,

$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = \omega^2 \begin{bmatrix} q_1 \\ q_2 \end{bmatrix}, \tag{15}$$

which is a matrix eigenvalue problem

$$\mathbf{A}\mathbf{q} = \lambda \mathbf{q}, \tag{16}$$

with $\lambda = \omega^2$ as the eigenvalue. Solving for the eigenvalues and eigenspaces as explained in Section 11.2 we obtain

$$\lambda_1 = 1, \ \mathbf{e}_1 = \alpha \begin{bmatrix} 1 \\ 1 \end{bmatrix}; \qquad \lambda_2 = 3, \ \mathbf{e}_2 = \beta \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{17}$$

Each "eigenpair" gives us a solution of the form (14). The first gives $\omega = \sqrt{\lambda_1} = 1$, and†

$$\mathbf{x} = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \alpha \begin{bmatrix} 1 \\ 1 \end{bmatrix} \sin(t + \phi_1), \tag{18}$$

where α and ϕ_1 are arbitrary. The second gives $\omega = \sqrt{\lambda_2} = \sqrt{3}$, and

$$\mathbf{x} = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \beta \begin{bmatrix} 1 \\ -1 \end{bmatrix} \sin(\sqrt{3}\, t + \phi_2), \tag{19}$$

*Along the lines of COMMENT 1 in Example 2 of Section 11.2, if we do not use the same ω's and the same ϕ's in (14), then we will obtain only the trivial solution $x_1(t) = x_2(t) = 0$.

†The other root, $\omega = -1$, would yield no additional information.

where β and ϕ_2 are arbitrary, and there is no reason why ϕ_2 should be the same as ϕ_1. One can verify that (18) satisfies (12) for any α and any ϕ_1 and that (19) also satisfies (12) for any β and any ϕ_2. Since (12) is linear and homogeneous, it follows that the linear combination

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \alpha \begin{bmatrix} 1 \\ 1 \end{bmatrix} \sin(t + \phi_1) + \beta \begin{bmatrix} 1 \\ -1 \end{bmatrix} \sin(\sqrt{3}\,t + \phi_2) \tag{20}$$

is also a solution; indeed, it is a general solution of (12). Or, returning to scalar form, we have

$$\begin{aligned} x_1(t) &= \alpha \sin(t + \phi_1) + \beta \sin(\sqrt{3}\,t + \phi_2), \\ x_2(t) &= \alpha \sin(t + \phi_1) - \beta \sin(\sqrt{3}\,t + \phi_2). \end{aligned} \tag{21}$$

Here $\alpha, \beta, \phi_1, \phi_2$ are the constants of integration [just as A, B are the constants of integration in the general solution $x(t) = Ae^{3t} + Be^{-3t}$ of $x'' - 9x = 0$], and are determined from the initial conditions $x_1(0), x_2(0), x_1'(0), x_2'(0)$.

COMMENT 1. Note the central and organizing role of the eigenvalue problem. Each eigenpair defines a vibrational "mode," the eigenvalue gives the vibrational frequency ($\omega = \sqrt{\lambda}$) and the eigenvector gives the mode shape or configuration. The frequencies are called the **eigenfrequencies**, or **natural frequencies** (natural in that they correspond to the free, unforced vibration). The two terms on the right-hand side of (20) are called the **orthogonal modes** of vibration, orthogonal because $[1, 1]^{\mathrm{T}} \cdot [1, -1]^{\mathrm{T}} = 0$, that orthogonality being a consequence of the symmetry of $\mathbf{A}$. The first term is called the **low mode** because it occurs at the lower of the two natural frequencies, and the second term is called the **high mode** because it is at the higher of those two frequencies.

COMMENT 2. Depending upon the initial conditions, we can excite either one of those modes or both of them. For instance, the conditions $x_1(0) = x_2(0) = 0$ and $x_1'(0) = x_2'(0) = 1$ give $\beta = 0, \alpha = 1, \phi_1 = 0$ (ϕ_2 is irrelevant because $\beta = 0$) and hence the low mode motion

$$\begin{aligned} x_1(t) &= \sin t, \\ x_2(t) &= \sin t, \end{aligned} \tag{22}$$

the conditions $x_1'(0) = x_2'(0) = 0$ and $x_1(0) = 1, x_2(0) = -1$ give $\alpha = 0, \beta = 1, \phi_2 = \pi/2$ (ϕ_1 is irrelevant because $\alpha = 0$) and hence the high mode motion

$$\begin{aligned} x_1(t) &= \sin(\sqrt{3}\,t + \frac{\pi}{2}) = \cos\sqrt{3}\,t, \\ x_2(t) &= -\sin(\sqrt{3}\,t + \frac{\pi}{2}) = -\cos\sqrt{3}\,t, \end{aligned} \tag{23}$$

and the conditions $x_1(0) = 1, x_2(0) = x_1'(0) = x_2'(0) = 0$ give a motion containing both modes,

$$\begin{aligned} x_1(t) &= \frac{1}{2}\sin(t + \frac{\pi}{2}) + \frac{1}{2}\sin(\sqrt{3}\,t + \frac{\pi}{2}) = \frac{1}{2}\cos t + \frac{1}{2}\cos\sqrt{3}\,t, \\ x_2(t) &= \frac{1}{2}\sin(t + \frac{\pi}{2}) - \frac{1}{2}\sin(\sqrt{3}\,t + \frac{\pi}{2}) = \frac{1}{2}\cos t - \frac{1}{2}\cos\sqrt{3}\,t. \end{aligned} \tag{24}$$

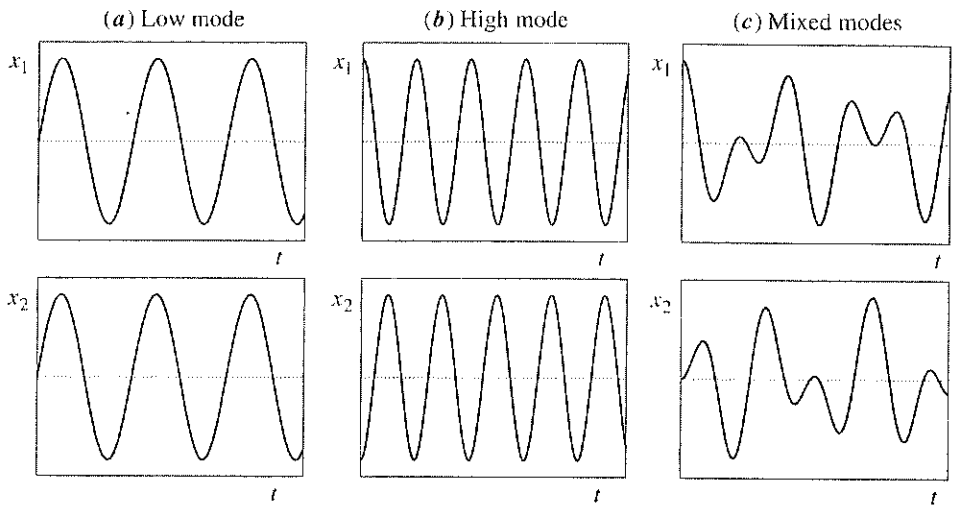


Figure 2. Low mode (22), high mode (23), mixed modes (24).

These three motions are contrasted in Fig. 2. Observe that each of the individual modes is simple and "clean," but the mixed mode motion is not. In the low mode the masses vibrate in unison and at the low frequency $\omega = 1$, and in the high mode their motions are opposite and at the high frequency $\omega = \sqrt{3}$.

To be sure it is clear how to apply the initial conditions, let us derive (24). We have the four equations

$$\begin{aligned} x_1(0) &= 1 = \alpha \sin\phi_1 + \beta \sin\phi_2, \\ x_2(0) &= 0 = \alpha \sin\phi_1 - \beta \sin\phi_2, \\ x_1'(0) &= 0 = \alpha \cos\phi_1 + \sqrt{3}\beta \cos\phi_2, \\ x_2'(0) &= 0 = \alpha \cos\phi_1 - \sqrt{3}\beta \cos\phi_2 \end{aligned} \tag{25}$$

in $\alpha, \beta, \phi_1, \phi_2$. They happen not to be linear algebraic equations [i.e., of the form $(\ )\alpha + (\ )\beta + (\ )\phi_1 + (\ )\phi_2 = (\ )$, where the parentheses contain constants], but they are readily solved. For instance, adding the first two and last two gives

$$\alpha \sin\phi_1 = \frac{1}{2}, \quad \alpha \cos\phi_1 = 0, \tag{26}$$

and these give $\phi_1 = \pi/2$ and $\alpha = 1/2$.* Similarly, subtracting the second from the first and the fourth from the third results in $\phi_2 = \pi/2$ and $\beta = 1/2$.

COMMENT 3. In this example there were two masses and two "degrees of freedom," $x_1(t)$ and $x_2(t)$. Consequently, $\mathbf{A}$ was 2×2 and the motion of each mass was found to consist of a linear combination of two eigenmodes. More generally, if there are n masses and n degrees of freedom $x_1(t), \ldots, x_n(t)$, then the motion of each mass consists of a linear

*These values are not uniquely determined. For instance $\phi_1 = 3\pi/2$ and $\alpha = -1/2$ satisfy (26) too. However, such differences do not lead to different solutions $x_1(t)$ and $x_2(t)$.

combination of n eigenmodes of an $n \times n$ matrix. ∎

11.3.2. Nonhomogeneous problem $\mathbf{Ax} = \Lambda\mathbf{x} + \mathbf{c}$. (Optional) In Chapters 8 and 10 we studied the general nonhomogeneous equation $\mathbf{Ax} = \mathbf{c}$, and in this chapter we have studied the eigenvalue problem, which is the homogeneous equation $\mathbf{Ax} = \lambda\mathbf{x}$. Next, we wish to show that eigenvalue problem concepts can be used to solve *non*homogeneous equations.

Specifically, consider the nonhomogeneous equation

$$\mathbf{Ax} = \Lambda\mathbf{x} + \mathbf{c}, \tag{27}$$

where the scalar Λ is a parameter.* We ask of $\mathbf{A}$ that it be $n \times n$, and that its eigenvectors provide a basis (not necessarily orthogonal) for $\mathbb{R}^n$. That will surely be the case if $\mathbf{A}$ is symmetric since then its eigenvectors provide an orthogonal basis for $\mathbb{R}^n$.

Of course, since Λ is considered as known we could subtract $\Lambda\mathbf{x}$ from $\mathbf{Ax}$ and absorb Λ into the $\mathbf{A}$ matrix. However, if Λ is a design parameter and we wish to see the explicit effect of Λ on the solution, then it is best to leave the $\Lambda\mathbf{x}$ term intact.

The idea is that to solve (27) we first take a "time-out" and solve the eigenvalue problem for $\mathbf{A}$ (i.e., $\mathbf{Ax} = \lambda\mathbf{x}$); that is, solve for the eigenvalues and eigenvectors of $\mathbf{A}$, which we denote as $\lambda_1, \ldots, \lambda_n$ (not necessarily distinct) and $\mathbf{e}_1, \ldots, \mathbf{e}_n$. Next, expand both $\mathbf{x}$ and $\mathbf{c}$, in (27), in terms of the $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ basis:

$$\mathbf{x} = \sum_{j=1}^{n} a_j\mathbf{e}_j \quad \text{and} \quad \mathbf{c} = \sum_{j=1}^{n} c_j\mathbf{e}_j. \tag{28a,b}$$

The c_j's are known (*i.e.*, they can be computed) since we know $\mathbf{c}$ and the $\mathbf{e}_j$ base vectors so the a_j's are our unknowns. To evaluate them, put (28) into (27):

$$\mathbf{A}\sum_{1}^{n} a_j\mathbf{e}_j = \Lambda\sum_{1}^{n} a_j\mathbf{e}_j + \sum_{1}^{n} c_j\mathbf{e}_j. \tag{29}$$

But

$$\mathbf{A}\sum_{1}^{n} a_j\mathbf{e}_j = \sum_{1}^{n} a_j\mathbf{A}\mathbf{e}_j = \sum_{1}^{n} a_j\lambda_j\mathbf{e}_j, \tag{30}$$

so that (29) can be re-expressed as

$$\sum_{1}^{n}(\lambda_j - \Lambda)a_j\mathbf{e}_j = \sum_{1}^{n} c_j\mathbf{e}_j. \tag{31}$$

Finally, since the $\mathbf{e}_j$'s are LI (for they are a basis) it follows from (31) that

$$(\lambda_j - \Lambda)a_j = c_j. \quad (j = 1, \ldots, n) \tag{32}$$

At this point we need to be careful to distinguish these cases:

*By a *parameter* we mean a constant, the value of which we are free to specify. In the equation $x'' + 9x = F\sin\Omega t$, for instance, F and Ω are parameters.

(i) Suppose that none of the λ_j's equals Λ. Then we can divide both sides of (32) by $\lambda_j - \Lambda$ and obtain $a_j = c_j/(\lambda_j - \Lambda)$, and the *unique solution*

$$\mathbf{x} = \sum_1^n \frac{c_j}{\lambda_j - \Lambda}\mathbf{e}_j \tag{33}$$

of (27).

(ii) Next, suppose that $\Lambda = \lambda_1$, where λ_1 is an eigenvalue of multiplicity 1. Then (32) becomes $(0)a_1 = c_1$ for $j = 1$, and two possibilities exist: if $c_1 \neq 0$, then there exists no a_1 satisfying $(0)a_1 = c_1$, and there is *no solution* of (27); but if $c_1 = 0$, then a_1 is arbitrary, and (27) admits a *nonunique* solution, the one-parameter family of solutions

$$\mathbf{x} = a_1\mathbf{e}_1 + \sum_2^n \frac{c_j}{\lambda_j - \Lambda}\mathbf{e}_j. \quad (a_1 \text{ arbitrary}) \tag{34}$$

(iii) Similarly if $\Lambda = \lambda_1$ is an eigenvalue of multiplicity p: if $c_1, \ldots, c_p$ are not all zero, then there is *no solution* of (27); and if $c_1 = \cdots = c_p = 0$, then there is a *nonunique* solution, the p-parameter family of solutions

$$\mathbf{x} = a_1\mathbf{e}_1 + \cdots + a_p\mathbf{e}_p + \sum_{p+1}^n \frac{c_j}{\lambda_j - \Lambda}\mathbf{e}_j. \quad (a_1, \ldots, a_p \text{ arbitrary}) \tag{35}$$

It is illuminating to compare (35) with the solution (18) in Section 10.5, to the problem $\mathbf{Ax} = \mathbf{c}$ (i.e., $\Lambda = 0$ in that case). There, $\mathbf{x}_0$ is a particular solution (i.e., $\mathbf{Ax}_0 = \mathbf{c}$), and $\mathbf{x}_1, \ldots, \mathbf{x}_p$ were homogeneous solutions (i.e., $\mathbf{Ax}_1 = \cdots = \mathbf{Ax}_p = \mathbf{0}$). In (35), the $\sum_{p+1}^n$ term corresponds to $\mathbf{x}_0$, and $\mathbf{e}_1, \ldots, \mathbf{e}_p$ are homogeneous solutions [i.e., solutions of (27) with $\mathbf{c}$ removed].*

EXAMPLE 3. *Forced Vibration of the Two-Mass System.* To illustrate, consider the mechanical system of Example 2 again, but this time with the forcing functions $f_1(t) = F_1 \sin \Omega t$ and $f_2(t) = F_2 \sin \Omega t$. Then in place of (12) we have

$$\begin{aligned} x_1'' + 2x_1 - x_2 &= F_1 \sin \Omega t, \\ x_2'' - x_1 + 2x_2 &= F_2 \sin \Omega t. \end{aligned} \tag{36}$$

We already found the general solution of the homogeneous equations (12), so in this example let us seek only a particular solution (the total solution then being the sum of the two). Let us seek a particular solution in the form†

$$\begin{aligned} x_1(t) &= q_1 \sin \Omega t, \\ x_2(t) &= q_2 \sin \Omega t. \end{aligned} \tag{37}$$

*Of course there is no reason why the $\mathbf{e}_1, \ldots, \mathbf{e}_p$ need to equal $\mathbf{x}_1, \ldots, \mathbf{x}_p$, but it *is* true that $\text{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_p\}$ and $\text{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ are identical.

†The form (37) is inspired by the method of undetermined coefficients. Actually, we might try $x_1(t) = q_1 \sin \Omega t + r_1 \cos \Omega t$ and $x_2(t) = q_2 \sin \Omega t + r_2 \cos \Omega t$, but we can anticipate that the $\cos \Omega t$ terms will not be needed (i.e., we will find that $r_1 = r_2 = 0$) because there are no x_1' or x_2' terms on the left-hand side of (36).

Putting (37) into (36), canceling the $\sin \Omega t$ factors, and expressing the equations in matrix form, we have

$$\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = \Omega^2 \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} + \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} \tag{38}$$

or

$$\mathbf{A}\mathbf{q} = \Lambda \mathbf{q} + \mathbf{c}, \tag{39}$$

where

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \quad \Lambda = \Omega^2, \quad \mathbf{c} = \begin{bmatrix} F_1 \\ F_2 \end{bmatrix}.$$

Recall from (17) that the eigenvalues and eigenvectors of $\mathbf{A}$ are (with $\alpha = \beta = 1$, say)

$$\lambda_1 = 1, \ \mathbf{e}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}; \qquad \lambda_2 = 3, \ \mathbf{e}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{40}$$

Suppose that $\Lambda \neq \lambda_1$ and $\Lambda \neq \lambda_2$; that is, the driving frequency Ω does not equal either of the natural frequencies 1 and $\sqrt{3}$. Then we have the unique solution (33), where "$\mathbf{x}$" is $\mathbf{q}$, $c_1 = (\mathbf{c} \cdot \mathbf{e}_1)/(\mathbf{e}_1 \cdot \mathbf{e}_1) = (F_1 + F_2)/2$, $c_2 = (\mathbf{c} \cdot \mathbf{e}_2)/(\mathbf{e}_2 \cdot \mathbf{e}_2) = (F_1 - F_2)/2$, $\lambda_1 = 1$, $\lambda_2 = 3$, and $\Lambda = \Omega^2$.* Thus,

$$\begin{aligned} \mathbf{q} &= \frac{F_1 + F_2}{2(1-\Omega^2)} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \frac{F_1 - F_2}{2(3-\Omega^2)} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \\ &= \frac{1}{(1-\Omega^2)(3-\Omega^2)} \begin{bmatrix} (2-\Omega^2) F_1 + F_2 \\ F_1 + (2-\Omega^2) F_2 \end{bmatrix} \end{aligned} \tag{41}$$

so the desired particular solution (37), of (36), is

$$x_1(t) = \frac{(2-\Omega^2) F_1 + F_2}{(1-\Omega^2)(3-\Omega^2)} \sin \Omega t, \tag{42a}$$

$$x_2(t) = \frac{F_1 + (2-\Omega^2) F_2}{(1-\Omega^2)(3-\Omega^2)} \sin \Omega t. \tag{42b}$$

In Section 3.8 we studied the forced vibration of a single mass, and stressed the importance to an engineer of the amplitude- and phase-response curves. Here too, let us plot the amplitude-response curves (Fig. 3) for the representative case where $F_1 = 1$ and $F_2 = 0$. The amplitudes $A_1(\Omega)$ and $A_2(\Omega)$ are, from (42a) and (42b),

$$A_1(\Omega) = \left| \frac{(2-\Omega^2)}{(1-\Omega^2)(3-\Omega^2)} \right| \quad \text{and} \quad A_2(\Omega) = \left| \frac{1}{(1-\Omega^2)(3-\Omega^2)} \right|, \tag{43}$$

and we observe that these tend to infinity as Ω tends to either of the natural frequencies, 1 and $\sqrt{3}$.

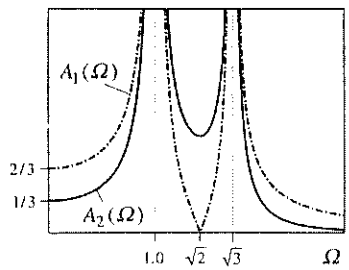


Figure 3. Amplitude-response curves, equation (43).

What if Ω *equals* 1 or $\sqrt{3}$? Then our derivation of (43) does not hold since it is based on the assumption that $\Lambda \neq \lambda_1$ and $\Lambda \neq \lambda_2$. If we do have $\Lambda = \lambda_1$ (i.e., $\Omega = 1$) say,

*If the formula $c_j = (\mathbf{c} \cdot \mathbf{e}_j)/(\mathbf{e}_j \cdot \mathbf{e}_j)$ is unfamiliar to you, we urge you to review Section 9.9, especially equation (23) therein.

then [according to case (ii) above] there is no solution if $c_1 \neq 0$ and there is an infinity of solutions if $c_1 = 0$. Since

$$c_1 = \frac{[F_1, F_2]^{\mathrm{T}} \cdot \mathbf{e}_1}{\mathbf{e}_1 \cdot \mathbf{e}_1}, \tag{44}$$

the idea is that there will be no solution unless the forcing vector $[F_1, F_2]^{\mathrm{T}}$ does not excite that particular eigenmode – namely, unless $[F_1, F_2]^{\mathrm{T}}$ is orthogonal to $\mathbf{e}_1$! Similarly, if $\Lambda = \lambda_2$ (i.e., $\Omega = \sqrt{3}$).

What do we mean when we say there is no solution if $\Lambda = \lambda_1$ and $c_1 \neq 0$ (or $\Lambda = \lambda_2$ and $c_2 \neq 0$)? We do not mean that there exists no particular solution of (36) but only that there is no solution of the assumed form (37). A modified version of (37) is needed, but we will not pursue that point. ∎

The method presented here for solving (27) is known as the **eigenvector expansion method**. What is the advantage in using the eigenvectors of $\mathbf{A}$ as our basis; why not use *any* basis for $\mathbb{R}^n$? The idea is that the vector $\mathbf{A}\mathbf{e}_j$ in the middle of (30) is simply $\lambda_j \mathbf{e}_j$ if $\mathbf{e}_j$ is an eigenvector of $\mathbf{A}$, whereas it would be a linear combination of all the base vectors if some other basis were used. In that case we would end up with a *coupled* system of linear algebraic equations for the a_j's, rather than the uncoupled (and hence readily solved) system (32). This same comment applies to Example 4 in Section 11.2 where, to study the stability of the equilibrium population, we expanded the initial population vector $\mathbf{p}_0$ in terms of the eigenvectors of $\mathbf{A}$.

Closure. Symmetric matrices arise frequently in applications, and their symmetry leads to several important results regarding their eigenvalues and eigenvectors: their eigenvalues are real, eigenvectors corresponding to distinct eigenvalues are orthogonal, and the eigenvectors of an $n \times n$ symmetric matrix provide an orthogonal basis for $\mathbb{R}^n$.

We discussed an important application to multimass mechanical systems, and found that the free oscillation can be represented as the superposition of orthogonal modes, with the eigenvalues giving the modal frequencies and the eigenvectors giving the modal configurations.

The special importance of symmetric matrices is further revealed in the remaining sections of this chapter.

In the second half of this section we return to the nonhomogeneous equation $\mathbf{A}\mathbf{x} = \mathbf{c}$, actually to the form $\mathbf{A}\mathbf{x} = \Lambda\mathbf{x} + \mathbf{c}$ where, Λ is a parameter, and develop a line of approach known as the eigenvector expansion method. The idea behind that method is to compute first the eigenvalues and eigenvectors of $\mathbf{A}$. Assuming that we can obtain a basis of $\mathbb{R}^n$ from the eigenvectors of $\mathbf{A}$ (as is always possible for symmetric $\mathbf{A}$'s which, indeed, provide us with *orthogonal* bases), we use that basis, which is the most convenient or "natural" basis to expand $\mathbf{x}$ and $\mathbf{c}$. Finally, equating coefficients of the various base vectors on both sides of the equation gives uncoupled linear algebraic equations on the unknown coefficients in the expansion of $\mathbf{x}$.

The eigenvector expansion method is applicable to other cases as well, such as ordinary and partial differential equations. We will meet it again when we come to

partial differential equations.

EXERCISES 11.3

1. From the eigenvectors of the given $n \times n$ matrix obtain an orthogonal basis for $\mathbb{R}^n$.

(a) $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix}$

(c) $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ (d) $\begin{bmatrix} 0 & 2 & 0 \\ 2 & 0 & 0 \\ 0 & 0 & -2 \end{bmatrix}$

(e) $\begin{bmatrix} 4 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 4 \end{bmatrix}$ (f) $\begin{bmatrix} 7 & 4 & -4 \\ 4 & 1 & 8 \\ -4 & 8 & 1 \end{bmatrix}$

(g) $\begin{bmatrix} 0 & 0 & 0 & 4 \\ 0 & 0 & 4 & 0 \\ 0 & 4 & 0 & 0 \\ 4 & 0 & 0 & 0 \end{bmatrix}$ (h) $\begin{bmatrix} 0 & 3 & 3 & 0 \\ 3 & 0 & 3 & 0 \\ 3 & 3 & 0 & 0 \\ 0 & 0 & 0 & -3 \end{bmatrix}$

(i) $\begin{bmatrix} 6 & 0 & 0 & 0 \\ 0 & 2 & 2 & 2 \\ 0 & 2 & 2 & 2 \\ 0 & 2 & 2 & 2 \end{bmatrix}$ (j) $\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$

2. Determine whether or not the eigenvectors of the given *non*symmetric 2×2 matrix provide an orthogonal basis for $\mathbb{R}^2$.

(a) $\begin{bmatrix} 0 & 0 \\ 3 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 9 \\ 1 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 0 & 0 \\ 9 & 0 \end{bmatrix}$

3. (*Proof of Theorem 11.3.1*) (a) Prove Theorem 11.3.1 for the simple case where $\mathbf{A}$ is merely 2×2, i.e., of the form $\begin{bmatrix} a & b \\ b & d \end{bmatrix}$.

(b) We will now supply the "skeleton" of a general proof of Theorem 11.3.1 and ask you to answer questions about the steps. Overhead bars will denote complex conjugates: if $z = x + iy$, then $\overline{z} = x - iy$. Proof:

$$\mathbf{A}\mathbf{e} = \lambda \mathbf{e}, \tag{3.1}$$

so

$$\mathbf{A}\overline{\mathbf{e}} = \overline{\lambda}\overline{\mathbf{e}}. \tag{3.2}$$

Dot both sides of (3.1) into $\overline{\mathbf{e}}$, and dot $\mathbf{e}$ into both sides of (3.2):

$$(\mathbf{A}\mathbf{e}) \cdot \overline{\mathbf{e}} = (\lambda \mathbf{e}) \cdot \overline{\mathbf{e}} \quad \text{and} \quad \mathbf{e} \cdot (\mathbf{A}\overline{\mathbf{e}}) = \mathbf{e} \cdot (\overline{\lambda}\overline{\mathbf{e}}). \tag{3.3}$$

Thus,

$$\mathbf{e}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\overline{\mathbf{e}} = \lambda \mathbf{e}^{\mathrm{T}}\overline{\mathbf{e}} \quad \text{and} \quad \mathbf{e}^{\mathrm{T}}\mathbf{A}\overline{\mathbf{e}} = \overline{\lambda}\mathbf{e}^{\mathrm{T}}\overline{\mathbf{e}}, \tag{3.4}$$

so

$$(\lambda - \overline{\lambda})\mathbf{e}^{\mathrm{T}}\overline{\mathbf{e}} = 0, \tag{3.5}$$

and hence

$$\lambda = \overline{\lambda} \tag{3.6}$$

so λ is real. Questions: How does (3.2) follow from (3.1)? (3.4) from (3.3)? (3.5) from (3.4)? (3.6) from (3.5)?

4. (*Proof of Theorem 11.3.2*) Prove Theorem 11.3.2 for the simple case where $n = 2$.

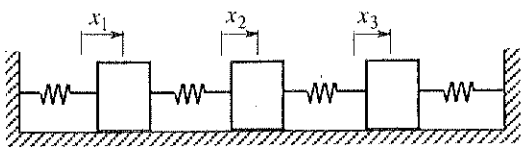
5. (*Post-dotting versus pre-dotting*) Observe that the top left equation in (2) was obtained by dotting $\mathbf{e}_k$ into both sides of (1a); i.e., we "pre-dotted" (1a) with $\mathbf{e}_k$. However, the top right equation in (2) was obtained by dotting both sides of (1b) into $\mathbf{e}_j$; i.e., we "post-dotted" $\mathbf{e}_j$ into (1b). Now, post-dotting or pre-dotting doesn't matter, in the sense that the dot product is commutative: $\mathbf{x} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{x}$. Nevertheless, one may be more convenient than the other. Specifically, show that if we write the top left equation in (2) as $(\mathbf{A}\mathbf{e}_j) \cdot \mathbf{e}_k = (\lambda_j \mathbf{e}_j) \cdot \mathbf{e}_k$, instead of $\mathbf{e}_k \cdot (\mathbf{A}\mathbf{e}_j) = \mathbf{e}_k \cdot (\lambda_j \mathbf{e}_j)$, then it is more difficult to obtain (3) and hence the desired result.

6. Use any theorem(s) from Chapter 3 to show that the solution to (12), with the initial conditions $x_1(0), x_2(0), x_1'(0)$, and $x_2'(0)$ specified, is unique.

7. Beginning with (21), complete the solution for the following initial conditions:

(a) $x_1(0) = 2, x_2(0) = 3, x_1'(0) = x_2'(0) = 0$
(b) $x_1(0) = 1, x_2(0) = x_1'(0) = 0, x_2'(0) = -3$
(c) $x_1(0) = x_2(0) = x_1'(0) = 0, x_2'(0) = 5$
(d) $x_1(0) = -2, x_2(0) = x_1'(0) = 0, x_2'(0) = 3$

8. Consider the three-mass system shown.



(a) Derive the equations of motion for the free vibration, taking all masses and spring stiffnesses to be 1, say.
(b) Following the same lines as in Example 2, find the orthogonal modes (i.e., the eigenvectors) and their corresponding natural frequencies, proceeding either by hand or by using computer software.
(c) Give any set of initial conditions that will excite the low frequency mode only; the middle mode only; the high mode only.
(d) Find the solution $x_1(t), x_2(t), x_3(t)$ corresponding to the initial condition $x_1(0) = 1, x_2(0) = x_3(0) = x_1'(0) = x_2'(0) = x_3'(0) = 0$.

9. Consider a mass-spring system like the one shown in Exercise 8, but with five masses and six springs. If all the masses and spring stiffnesses are 1, say, then the equations of motion are these, in matrix form:

$$\begin{bmatrix} x_1'' \\ x_2'' \\ x_3'' \\ x_4'' \\ x_5'' \end{bmatrix} + \begin{bmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{9.1}$$

or $\mathbf{x}'' + \mathbf{A}\mathbf{x} = \mathbf{0}$. Observe that the $\mathbf{A}$ matrix is **tridiagonal** (i.e., all elements are zero except for the main diagonal and the two neighboring diagonals). Physically, this result correponds to **nearest-neighbor coupling** whereby each mass feels only its immediate neighbors. Nearest-neighbor coupling occurs in other systems as well. For instance, in modeling single-lane traffic flow, each driver accelerates or decelerates according to the motion of the cars immediately ahead and immediately behind, so the resulting coupled differential equations exhibit nearest-neighbor coupling. The problem that we pose is for you to use computer software to determine the natural frequencies and corresponding mode shapes (eigenvectors).

10. (*Beats*) Suppose in Example 2 that $m_1 = m_2 = k_{12} = 1$, $k_1 = k_2 = 20$ so that the coupling of m_1 and m_2 is weak since k_{12} is much smaller than k_1 and k_2. (In the limit, if k_{12} were zero, there would be no coupling at all, and the motions of m_1 and m_2 would be entirely independent.) Supposing further that $x_1(0) = 1, x_2(0) = x_1'(0) = x_2'(0) = 0$, proceed as in Example 2 and show that, for the free vibration,

$$\begin{aligned} x_1(t) &= \tfrac{1}{2}\left(\cos\sqrt{20}\,t + \cos\sqrt{22}\,t\right), \\ x_2(t) &= \tfrac{1}{2}\left(\cos\sqrt{20}\,t - \cos\sqrt{22}\,t\right). \end{aligned} \tag{10.1}$$

Next, use the trigonometric identities

$$\begin{aligned} \cos A + \cos B &= 2\cos\tfrac{A+B}{2}\cos\tfrac{A-B}{2}, \\ \cos A - \cos B &= -2\sin\tfrac{A+B}{2}\sin\tfrac{A-B}{2} \end{aligned} \tag{10.2}$$

to show that

$$\begin{aligned} x_1(t) &\approx \cos 4.58t \cos 0.11t, \\ x_2(t) &\approx \sin 4.58t \sin 0.11t. \end{aligned} \tag{10.3}$$

Use (10.3) to sketch $x_1(t)$ and $x_2(t)$ versus t in separate graphs, one below the other, labeling key values. Observe the slow transfer of energy back and forth between m_1 and m_2. In vibration theory this phenomenon is known as **beats**.

11. (*Rayleigh's quotient*) Let $\mathbf{A}$ be a symmetric $n \times n$ matrix. Dotting any eigenvector $\mathbf{e}$ of $\mathbf{A}$ into both sides of $\mathbf{A}\mathbf{e} = \lambda\mathbf{e}$ and solving for λ, gives

$$\lambda = \frac{\mathbf{e}\cdot\mathbf{A}\mathbf{e}}{\mathbf{e}\cdot\mathbf{e}} = \frac{\mathbf{e}^{\mathrm{T}}\mathbf{A}\mathbf{e}}{\mathbf{e}^{\mathrm{T}}\mathbf{e}}. \tag{11.1}$$

More generally, if $\mathbf{x}$ is any vector, not necessarily an eigenvector of $\mathbf{A}$, then the number

$$\boxed{R(\mathbf{x}) = \frac{\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}}} \tag{11.2}$$

is known as **Rayleigh's quotient**, after *Lord Rayleigh* (*John William Strutt*; 1842–1919). Imagine putting randomly chosen $\mathbf{x}$ vectors, one after another, into Rayleigh's quotient. If $\mathbf{x}$ happens to coincide with an eigenvector, then, according to (11.1), $R(\mathbf{x})$ gives the corresponding eigenvalue. In any case,

$$|R(\mathbf{x})| = \left|\frac{\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}}{\mathbf{x}^{\mathrm{T}}\mathbf{x}}\right| \le |\lambda_1|, \tag{11.3}$$

where the eigenvalues are ordered so that $|\lambda_1| \ge |\lambda_2| \ge \cdots \ge |\lambda_n|$. That is, $|R(\mathbf{x})|$ provides a **lower bound** on the magnitude of the largest eigenvalue of $\mathbf{A}$, where $\mathbf{x}$ is any vector [i.e., any nonzero vector, since $R(\mathbf{0}) = 0/0$ is undefined]. Upper and lower bounds on eigenvalues are sometimes important, and Rayleigh's quotient is used again in Exercise 12.

(a) Prove the inequality in (11.3). HINT: Since $\mathbf{A}$ is symmetric, it has n orthogonal eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$, corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_n$. Expand $\mathbf{x}$ in terms of the orthogonal basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$:

$$\mathbf{x} = \sum_{j=1}^{n} a_j \mathbf{e}_j. \tag{11.4}$$

(b) Verify (11.3) for the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & -1 & -2 \\ 0 & -2 & -1 \end{bmatrix} \tag{11.5}$$

by taking $\mathbf{x} = [1,0,0]^{\mathrm{T}}, [0,1,0]^{\mathrm{T}}, [0,0,1]^{\mathrm{T}}, [1,2,3]^{\mathrm{T}}$ and $[3,-1,4]^{\mathrm{T}}$, say, verifying that $|R(\mathbf{x})| \le |\lambda_1|$ in each case.
(c) Evaluate $R(\mathbf{x})$ for $\mathbf{x} = [0.4, 0.3, 0.3]^{\mathrm{T}}$, $[0.6, 0.2, 0.2]^{\mathrm{T}}$, $[0.8, 0.1, 0.1]^{\mathrm{T}}$, $[0.96, 0.02, 0.02]^{\mathrm{T}}$, $[1,0,0]^{\mathrm{T}}$, and for $\mathbf{x} = [0,1,1.4]^{\mathrm{T}}$, $[0,1,1.1]^{\mathrm{T}}$, $[0,1,1]^{\mathrm{T}}$, and discuss your results in the light of (11.1) and (11.3).

12. (*The power method*) There exists a simple *iterative* procedure for calculating eigenvalues and eigenvectors, which is known as the **power method**. To begin, one selects any nonzero vector $\mathbf{x}^{(0)}$ and then computes $\mathbf{x}^{(1)} \equiv \mathbf{A}\mathbf{x}^{(0)}$, $\mathbf{x}^{(2)} \equiv \mathbf{A}\mathbf{x}^{(1)}$, and so on. That is,

$$\mathbf{x}^{(k+1)} \equiv \mathbf{A}\mathbf{x}^{(k)} \quad (k = 0, 1, 2, \ldots). \tag{12.1}$$

Before analyzing the situation, let us apply (12.1) and see what happens. Let

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \tag{12.2}$$

for example. Choosing $\mathbf{x}^{(0)} = [1,0,0]^{\mathrm{T}}$, say, successive application of (12.1) gives

$$\begin{array}{cccccc} & \mathbf{x}^{(0)} & \mathbf{x}^{(1)} & \mathbf{x}^{(2)} & \mathbf{x}^{(3)} & \mathbf{x}^{(4)} & \mathbf{x}^{(5)} \\ \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} & \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} & \begin{bmatrix} 3 \\ 1 \\ 1 \end{bmatrix} & \begin{bmatrix} 5 \\ 3 \\ 3 \end{bmatrix} & \begin{bmatrix} 11 \\ 5 \\ 5 \end{bmatrix} & \begin{bmatrix} 21 \\ 11 \\ 11 \end{bmatrix}, \end{array} \tag{12.3}$$

and so on. Now, observe, from the very nature of the eigenvalue problem $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, that if $\mathbf{x}^{(0)}$ were an eigenvector of $\mathbf{A}$, then $\mathbf{A}\mathbf{x}^{(0)}$ would be some scalar multiple of $\mathbf{x}^{(0)}$. But $\mathbf{A}\mathbf{x}^{(0)} \equiv \mathbf{x}^{(1)}$, and $\mathbf{x}^{(1)}$ is seen from (12.3) *not* to be a scalar multiple of $\mathbf{x}^{(0)}$. Hence, $\mathbf{x}^{(0)}$ is not an eigenvector of $\mathbf{A}$. Similarly, $\mathbf{x}^{(1)}$ is not an eigenvector because $\mathbf{x}^{(2)}$ is not a multiple of $\mathbf{x}^{(1)}$, and so on. Nevertheless, we see that with each successive step $\mathbf{x}^{(k+1)}$ draws closer and closer to being a multiple of $\mathbf{x}^{(k)}$ so that the sequence $\mathbf{x}^{(k)}$ is evidently *approaching* an eigenvector of $\mathbf{A}$. In fact, $\mathbf{x}^{(5)}$ is very close to being a multiple of $\mathbf{x}^{(4)}$ so that there is an eigenvector

$$\mathbf{e} \approx \begin{bmatrix} 21 \\ 11 \\ 11 \end{bmatrix}. \tag{12.4}$$

What is the corresponding λ, $\lambda \approx 21/11$? $\lambda \approx 11/5$? An average of the two? We state without proof that one does well to use the Rayleigh quotient from Exercise 11:

$$\lambda \approx \frac{\mathbf{x}^{(4)\mathrm{T}}\mathbf{A}\mathbf{x}^{(4)}}{\mathbf{x}^{(4)\mathrm{T}}\mathbf{x}^{(4)}} = \frac{\mathbf{x}^{(4)\mathrm{T}}\mathbf{x}^{(5)}}{\mathbf{x}^{(4)\mathrm{T}}\mathbf{x}^{(4)}} = \frac{341}{171} = 1.994. \tag{12.5}$$

How do (12.4) and (12.5) compare with exact values? The eigenvalues and eigenvectors of $\mathbf{A}$ are

$$\lambda_1 = 2,\ \mathbf{e}_1 = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}, \qquad \lambda_2 = -1,\ \mathbf{e}_2 = \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix},$$
$$\lambda_3 = 0,\ \mathbf{e}_3 = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} \tag{12.6}$$

so the iteration (12.3) is evidently converging to $\mathbf{e}_1$. [It is striking that whereas (12.4) is accurate to only around one part in 20, (12.5) is accurate to around one part in 200. This enhancement of accuracy, using Rayleigh's quotient to determine λ, is not a coincidence and can be explained theoretically. See, for instance, Stephen H. Crandall, *Engineering Analysis* (New York: McGraw–Hill, 1956), Chap. 2.]

To see what is going on, suppose that $\mathbf{A}$ is a symmetric matrix of order n (although symmetry is more than we need; it would suffice for $\mathbf{A}$ to have n LI eigenvectors) and let its eigenvalues be ordered so that $|\lambda_1| \ge |\lambda_2| \ge \cdots \ge |\lambda_n|$ as in (12.6). Since $\mathbf{A}$ is symmetric, its eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$ are on orthogonal basis for n-space. Hence, our initial vector $\mathbf{x}^{(0)}$ must be expressible as

$$\mathbf{x}^{(0)} = \sum_{j=1}^{n} a_j \mathbf{e}_j. \tag{12.7}$$

Of course, we cannot *compute* the a_j's since we do not know the $\mathbf{e}_j$'s yet, but that is no problem; it is the *form* of (12.7) that is important here. It follows from (12.1) and (12.7) that

$$\begin{aligned} \mathbf{x}^{(k)} &= \sum_{1}^{n} a_j \lambda_j^k \mathbf{e}_j \\ &= \lambda_1^k \left[a_1 \mathbf{e}_1 + a_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k \mathbf{e}_2 + \cdots + a_n \left(\frac{\lambda_n}{\lambda_1}\right)^k \mathbf{e}_n \right]. \end{aligned}$$

(12.8)

If λ_1 is, in fact, *dominant*, i.e., $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$, then $(\lambda_2/\lambda_1)^k, \ldots, (\lambda_n/\lambda_1)^k$ all tend to zero as $k \to \infty$, so that $\mathbf{x}^{(k)} \sim \lambda_1^k a_1 \mathbf{e}_1$ as $k \to \infty$ (provided that $a_1 \neq 0$, i.e., provided that $\mathbf{x}^{(0)}$ does not happen to be orthogonal to $\mathbf{e}_1$). Eigenvectors can be scaled arbitrarily so the $\lambda_1^k a_1$ factor is of little interest; the point is that $\mathbf{x}^{(k)}$ converges to $\mathbf{e}_1$, the eigenvector corresponding to the dominant eigenvalue. That is precisely what was found in the preceding illustration, wherein the dominant eigenvalue is $\lambda_1 = 2$. NOTE: Once again we see that the most convenient basis to use is the basis provided by the $\mathbf{A}$ matrix itself.

(a) Show that (12.8) follows from (12.1) and (12.7).
(b) Determine the dominant eigenvalue and corresponding eigenvector by the power method for

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

Also, evaluate the eigenvalues and eigenvectors exactly, either by hand or by using computer software. NOTE: Observe that even though your iteration converges, you cannot be certain that the eigenvalue obtained is the *dominant* one, λ_1, since your chosen $\mathbf{x}^{(0)}$ may (without your knowing it) be orthogonal to $\mathbf{e}_1$, as mentioned in the sentence following (12.8). Hence, we recommend that you carry out the iteration three times, once with $\mathbf{x}^{(0)} = [1,0,0]^{\mathrm{T}}$, once with $\mathbf{x}^{(0)} = [0,1,0]^{\mathrm{T}}$, and once with $\mathbf{x}^{(0)} = [0,0,1]^{\mathrm{T}}$ since there is no way that all *three* of these $\mathbf{x}^{(0)}$'s can be orthogonal to $\mathbf{e}_1$. (Do you see why this is so?) Go as far as $\mathbf{x}^{(4)}$ in each case, and use the Rayleigh quotient to estimate λ, as we did in (12.5).

13. The same as Exercise 12(b), for the given matrix

(a) $\begin{bmatrix} 2 & 1 & -1 \\ 1 & 4 & 3 \\ -1 & 3 & 4 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 3 & 0 \\ 1 & 0 & 1 \end{bmatrix}$

(e) $\begin{bmatrix} 0 & 1 & -1 \\ 1 & 3 & 4 \\ -1 & 4 & 3 \end{bmatrix}$ (f) $\begin{bmatrix} 4 & 1 & 3 \\ 1 & 0 & -1 \\ 3 & -1 & 4 \end{bmatrix}$

(g) $\begin{bmatrix} 2 & 1 & 1 & 2 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 2 & 1 & 1 & 2 \end{bmatrix}$ (h) $\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$

14. For the $\mathbf{A}$ matrix given in part (b) of Exercise 13, work out $\mathbf{A}^4$. Apply the power method (Exercise 13) both to $\mathbf{A}$ and $\mathbf{A}^4$, beginning with $\mathbf{x}^{(0)} = [0,1,0]^{\mathrm{T}}$, say. Go as far as $\mathbf{x}^{(2)}$ and use the Rayleigh quotient to estimate λ. Explain why the iteration converges more rapidly for $\mathbf{A}^4$ than for $\mathbf{A}$, and show how to recover the eigenvalue and eigenvector of $\mathbf{A}$ from those of $\mathbf{A}^4$. (See Exercise 15 of Section 11.2.)

15. Consider the problem $\mathbf{A}\mathbf{x} = \Lambda\mathbf{x} + \mathbf{c}$, where $\mathbf{A}, \Lambda, \mathbf{c}$ are given below (along with the eigenvalues of $\mathbf{A}$, for your convenience). Solve for $\mathbf{x}$ by the eigenfunction expansion method; if no solution exists, state that. You may use any of equations (33)–(35) without deriving them.

(a) $\mathbf{A}$ given in Exercise 13(a) ($\lambda = 7,3,0$), $\Lambda = 2$, $\mathbf{c} = [1,2,3]^{\mathrm{T}}$
(b) $\mathbf{A}$ given in Exercise 13(a) ($\lambda = 7,3,0$), $\Lambda = 3$, $\mathbf{c} = [2,2,0]^{\mathrm{T}}$
(c) $\mathbf{A}$ given in Exercise 13(a) ($\lambda = 7,3,0$), $\Lambda = 3$, $\mathbf{c} = [1,1,3]^{\mathrm{T}}$
(d) $\mathbf{A}$ given in Exercise 13(g) ($\lambda = 4,2,0,0$), $\Lambda = 1$, $\mathbf{c} = [3,-1,1,0]^{\mathrm{T}}$
(e) $\mathbf{A}$ given in Exercise 13(g) ($\lambda = 4,2,0,0$), $\Lambda = 4$, $\mathbf{c} = [1,2,0,3]^{\mathrm{T}}$
(f) $\mathbf{A}$ given in Exercise 13(g) ($\lambda = 4,2,0,0$), $\Lambda = 4$, $\mathbf{c} = [2,0,1,-2]^{\mathrm{T}}$
(g) $\mathbf{A}$ given in Exercise 13(g) ($\lambda = 4,2,0,0$), $\Lambda = 0$, $\mathbf{c} = [1,3,3,1]^{\mathrm{T}}$
(h) $\mathbf{A}$ given in Exercise 13(g) ($\lambda = 4,2,0,0$), $\Lambda = 0$, $\mathbf{c} = [1,2,3,4]^{\mathrm{T}}$
(i) $\mathbf{A}$ given in Exercise 13(g) ($\lambda = 4,2,0,0$), $\Lambda = 0$, $\mathbf{c} = [1,3,3,2]^{\mathrm{T}}$

16. To solve the nonhomogeneous equation (27), we first solve for the eigenvalues λ_j and eigenvectors $\mathbf{e}_j$ of $\mathbf{A}$, then we use those eigenvectors as a basis to expand $\mathbf{x}$ and $\mathbf{c}$. Why do we go to the extra trouble of solving for the eigenvalues and eigenvectors of $\mathbf{A}$ when it is easy to make up bases [such as $(1,0,\ldots,0), (0,1,0,\ldots,0), \ldots, (0,\ldots,0,1)$] without trouble? That is, explain what advantage there is in using the eigenvectors of $\mathbf{A}$ as our basis.

17. (*Generalized eigenvalue problem*) With $f_1(t) = f_2(t) = 0$, $x_1(t) = q_1 \sin(\omega t + \phi)$, and $x_2(t) = q_2 \sin(\omega t + \phi)$, (11) becomes the eigenvalue problem

$$\begin{bmatrix} \dfrac{k_1 + k_{12}}{m_1} & -\dfrac{k_{12}}{m_1} \\ -\dfrac{k_{12}}{m_2} & \dfrac{k_2 + k_{12}}{m_2} \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = \omega^2 \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} \qquad (17.1)$$

or $\mathbf{A}\mathbf{q} = \lambda\mathbf{q}$, where $\lambda = \omega^2$. In Example 2 we took $m_1 = m_2 = 1$ so $\mathbf{A}$ is symmetric and the eigenmodes are orthogonal. In general, however, $m_1 \neq m_2$ and $\mathbf{A}$ is not

symmetric. Nonetheless, observe that we can reexpress (16.1) as

$$\begin{bmatrix} k_1 + k_{12} & -k_{12} \\ -k_{12} & k_2 + k_{12} \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = \omega^2 \begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} \tag{17.2}$$

or

$$\mathbf{Kq} = \lambda \mathbf{Mq}, \tag{17.3}$$

where $\lambda = \omega^2$ again. Equation (17.3) is called a **generalized eigenvalue problem**, "generalized" because of the presence of $\mathbf{M}$. (The generalized eigenvalue problem is introduced in Exercise 20 of Section 11.2.) The eigenvalues are determined by $\det(\mathbf{K} - \lambda\mathbf{M}) = 0$, and then the corresponding eigenvectors follow as the nontrivial solutions of $(\mathbf{K} - \lambda\mathbf{M})\mathbf{q} = \mathbf{0}$.

Here is the problem:

(a) Let $\mathbf{K}$ and $\mathbf{M}$ be symmetric [as is the case in (17.2), although there is no need for $\mathbf{M}$ to be diagonal as well]. Show that if $\mathbf{e}_j$ and $\mathbf{e}_k$ are eigenvectors corresponding to distinct eigenvalues λ_j and λ_k, respectively, then $\mathbf{e}_j$ and $\mathbf{e}_k$ satisfy the **generalized orthogonality relation**

$$\mathbf{e}_j \cdot (\mathbf{Me}_k) = 0. \tag{17.4}$$

That is, $\mathbf{e}_j$ is orthogonal to $\mathbf{e}_k$ "relative to $\mathbf{M}$."
(b) Verify the truth of (17.4) for the case

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix} = \lambda \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} q_1 \\ q_2 \end{bmatrix}.$$

11.4 Diagonalization

We have seen that diagonal matrices are particularly straightforward. For instance, the solution of $\mathbf{Ax} = \mathbf{c}$, where $\mathbf{A}$ is $n \times n$, is generally tedious if n is large but is simple if $\mathbf{A}$ is diagonal, for then the scalar equations are not coupled. Similarly, raising $\mathbf{A}$ to the mth power is generally tedious if m is large but is simple if $\mathbf{A}$ is diagonal. Likewise, the solution of a system of differential equations

$$\mathbf{x}'(t) = \mathbf{Ax}(t) \tag{1}$$

is generally tedious but is simple if $\mathbf{A}$ is diagonal, for then the scalar equations are uncoupled.

To introduce the idea of diagonalization, let us focus on the application of diagonalization to the solution of the system of differential equations given by (1), where we assume that $\mathbf{A}$ is constant (i.e., its elements do not vary with t). We have already studied several methods for the solution of (1): the method of elimination (Section 3.9), which is essentially Gauss elimination but where coefficients are differential operators; the Laplace transform method (Section 5.4), which would reduce (1) to n linear algebraic equations in $\bar{x}_1(s), \ldots, \bar{x}_n(s)$; and seeking $\mathbf{x}(t) = \mathbf{q}e^{\lambda t}$ and obtaining an eigenvalue problem (Sections 11.2 and 11.3).

We begin the solution of (1) by diagonalization by making a linear change of variables from $x_1, \ldots, x_n$ to $\tilde{x}_1, \ldots, \tilde{x}_n$:

$$\mathbf{x} = \mathbf{Q}\tilde{\mathbf{x}}, \tag{2}$$

where $\mathbf{Q}$ is a constant matrix. Written out,

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} q_{11} & \cdots & q_{1n} \\ \vdots & & \vdots \\ q_{n1} & \cdots & q_{nn} \end{bmatrix} \begin{bmatrix} \widetilde{x}_1 \\ \vdots \\ \widetilde{x}_n \end{bmatrix}. \tag{3}$$

Putting (2) into (1) [and observing that $(\mathbf{Q}\widetilde{\mathbf{x}})' = \mathbf{Q}\widetilde{\mathbf{x}}'$ because the matrix $\mathbf{Q}$ is constant] gives

$$\mathbf{Q}\widetilde{\mathbf{x}}' = \mathbf{A}\mathbf{Q}\widetilde{\mathbf{x}}. \tag{4}$$

Since the choice of $\mathbf{Q}$ is ours, we can ask $\mathbf{Q}$ to be invertible. Then, multiplying (4) by $\mathbf{Q}^{-1}$ gives $\mathbf{Q}^{-1}\mathbf{Q}\widetilde{\mathbf{x}}' = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}\widetilde{\mathbf{x}}$ or

$$\widetilde{\mathbf{x}}' = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}\widetilde{\mathbf{x}}. \tag{5}$$

Given the $\mathbf{A}$ matrix, the idea is to try to find a $\mathbf{Q}$ matrix so that

$$\boxed{\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{D}} \tag{6}$$

is diagonal because then the differential equations within (5) will be *uncoupled*. If there does exist such a $\mathbf{Q}$ we say that $\mathbf{A}$ is **diagonalizable** and that $\mathbf{Q}$ **diagonalizes** $\mathbf{A}$.

Two questions present themselves: given $\mathbf{A}$, does there exist such a $\mathbf{Q}$ and, if so, how do we find it? (There is also a question of uniqueness, but we are not especially interested in whether or not $\mathbf{Q}$ is unique; we'll be happy to find *any* $\mathbf{Q}$ that diagonalizes $\mathbf{A}$.)

THEOREM 11.4.1 *Diagonalization*
Let $\mathbf{A}$ be $n \times n$.

1. $\mathbf{A}$ is diagonalizable if and only if it has n LI eigenvectors.
2. If $\mathbf{A}$ has n LI eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$ and we make these the columns of $\mathbf{Q}$, so that $\mathbf{Q} = [\mathbf{e}_1, \ldots, \mathbf{e}_n]$, then $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{D}$ is diagonal and the jth diagonal element of $\mathbf{D}$ is the jth eigenvalue of $\mathbf{A}$.

Proof: First, by $\mathbf{Q} = [\mathbf{e}_1, \ldots, \mathbf{e}_n]$ we mean that $\mathbf{Q}$ is partitioned into columns, the columns being $\mathbf{e}_1, \ldots, \mathbf{e}_n$.

Let us prove that if $\mathbf{A}$ is diagonalizable, then it has n LI eigenvectors. If $\mathbf{A}$ is diagonalizable, then there is an invertible matrix $\mathbf{Q}$ such that

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{D} = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & d_n \end{bmatrix}. \tag{7}$$

Pre-multiplying both sides of (7) by $\mathbf{Q}$ gives $\mathbf{AQ} = \mathbf{QD}$:

$$\mathbf{AQ} = \begin{bmatrix} q_{11} & \cdots & q_{1n} \\ \vdots & & \vdots \\ q_{n1} & \cdots & q_{nn} \end{bmatrix} \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & d_n \end{bmatrix}$$

$$= \begin{bmatrix} d_1 q_{11} & \cdots & d_n q_{1n} \\ \vdots & & \vdots \\ d_1 q_{n1} & \cdots & d_n q_{nn} \end{bmatrix} \equiv [d_1\mathbf{q}_1, \ldots, d_n\mathbf{q}_n], \tag{8}$$

where the vector $\mathbf{q}_j$ simply denotes the jth column of $\mathbf{Q}$. Alternatively,

$$\mathbf{AQ} = \mathbf{A}[\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n] = [\mathbf{Aq}_1, \mathbf{Aq}_2, \ldots, \mathbf{Aq}_n] \tag{9}$$

and, comparing (8) and (9), we see that

$$\mathbf{Aq}_1 = d_1\mathbf{q}_1, \quad \ldots, \quad \mathbf{Aq}_n = d_n\mathbf{q}_n. \tag{10}$$

Does (10) imply that the $\mathbf{q}_j$'s are eigenvectors of $\mathbf{A}$? Only if we can be sure that they are nonzero. Since we have assumed $\mathbf{A}$ to be diagonalizable, $\mathbf{Q}$ must be invertible. Hence, none of its columns $\mathbf{q}_j$ can be $\mathbf{0}$. Thus, the d_j's and $\mathbf{q}_j$'s are the eigenvalues λ_j and eigenvectors $\mathbf{e}_j$ of $\mathbf{A}$. Furthermore, the rank of $\mathbf{Q}$ must be n since $\mathbf{Q}$ is to be invertible, so (Theorem 10.5.2) its columns must be LI.

Thus far we have proved half of item 1, that if $\mathbf{A}$ is diagonalizable, then it has n LI eigenvectors. In doing so we have also proved item 2. It remains to prove the rest of 1, that if $\mathbf{A}$ has n LI eigenvectors, then it is diagonalizable. To do so, let us take $\mathbf{Q}$ to be made up of columns which are the eigenvectors of $\mathbf{A}$, so $\mathbf{Q} = [\mathbf{e}_1, \ldots, \mathbf{e}_n]$. Then

$$\mathbf{AQ} = [\mathbf{Ae}_1, \ldots, \mathbf{Ae}_n] = [\lambda_1\mathbf{e}_1, \ldots, \lambda_n\mathbf{e}_n] = \begin{bmatrix} \lambda_1 e_{11} & \cdots & \lambda_n e_{n1} \\ \vdots & & \vdots \\ \lambda_1 e_{1n} & \cdots & \lambda_n e_{nn} \end{bmatrix}$$

$$= \begin{bmatrix} e_{11} & \cdots & e_{n1} \\ \vdots & & \vdots \\ e_{1n} & \cdots & e_{nn} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & \lambda_n \end{bmatrix} = \mathbf{QD}. \tag{11}$$

Finally, $\mathbf{Q}$ is invertible since its columns are LI, so pre-multiplying (11) by $\mathbf{Q}^{-1}$ gives $\mathbf{Q}^{-1}\mathbf{AQ} = \mathbf{D}$. Hence $\mathbf{A}$ is diagonalizable, and the proof is complete. ■

Since the columns of $\mathbf{Q}$ are LI eigenvectors of $\mathbf{A}$, $\mathbf{Q}$ is called a **modal matrix** of $\mathbf{A}$.

Theorem 11.4.1 relates the diagonalizability of $\mathbf{A}$ to the eigenvectors of $\mathbf{A}$. With the help of Theorem 11.4.2, we will be able to relate the diagonalizability of $\mathbf{A}$ to the eigenvalues of $\mathbf{A}$ as well. When that is done we will turn to applications.

THEOREM 11.4.2 *Distinct Eigenvalues, LI Eigenvectors*
If an $n \times n$ matrix $\mathbf{A}$ has distinct eigenvalues $\lambda_1, \ldots, \lambda_n$, then the corresponding eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$ are LI.

Proof: We need to show that

$$c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + \cdots + c_n\mathbf{e}_n = \mathbf{0} \tag{12}$$

holds only if $c_1 = c_2 = \cdots = c_n = 0$. Multiplying (12) by $\mathbf{A}$ and noting that $\mathbf{A}\mathbf{e}_j = \lambda_j\mathbf{e}_j$, we have

$$c_1\lambda_1\mathbf{e}_1 + c_2\lambda_2\mathbf{e}_2 + \cdots + c_n\lambda_n\mathbf{e}_n = \mathbf{0}. \tag{13}$$

Repeating the process gives

$$\begin{aligned} c_1\lambda_1^2\mathbf{e}_1 + c_2\lambda_2^2\mathbf{e}_2 + \cdots + c_n\lambda_n^2\mathbf{e}_n &= \mathbf{0}, \\ &\vdots \\ c_1\lambda_1^{n-1}\mathbf{e}_1 + c_2\lambda_2^{n-1}\mathbf{e}_2 + \cdots + c_n\lambda_n^{n-1}\mathbf{e}_n &= \mathbf{0}. \end{aligned} \tag{14}$$

Expressed in matrix form,

$$\begin{bmatrix} 1 & \cdots & 1 \\ \lambda_1 & \cdots & \lambda_n \\ \vdots & & \vdots \\ \lambda_1^{n-1} & \cdots & \lambda_n^{n-1} \end{bmatrix} \begin{bmatrix} c_1\mathbf{e}_1 \\ c_2\mathbf{e}_2 \\ \vdots \\ c_n\mathbf{e}_n \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}. \tag{15}$$

The determinant of the coefficient matrix is a *Vandermonde determinant*, which (see Exercise 17 in Section 10.4) is nonzero if the λ_j's are distinct. Since λ_j's are indeed distinct by assumption, (15) admits the unique trivial solution $c_1\mathbf{e}_1 = \mathbf{0}, c_2\mathbf{e}_2 = \mathbf{0}, \ldots, c_n\mathbf{e}_n = \mathbf{0}$, and since the $\mathbf{e}_j$'s are nonzero (because they are eigenvectors) it follows that $c_1 = c_2 = \cdots = c_n = 0$, so $\mathbf{e}_1, \ldots, \mathbf{e}_n$ are LI. ■

From Theorems 11.4.1 and 11.4.2 we can draw the following conclusion.

THEOREM 11.4.3 *Diagonalizability*
If an $n \times n$ matrix has n distinct eigenvalues, then it is diagonalizable.

As usual, be careful not to read converses into theorems when they are not stated. Specifically, Theorem 11.4.3 does *not* say that an $n \times n$ matrix is diagonalizable *if and only if* it has n distinct eigenvalues.

Consider an application.

EXAMPLE 1. *A Problem in Chemical Kinetics.* We consider, here, a special class of chemical reactions known as *first-order reactions*. These reactions are governed by systems of linear, coupled, first-order ordinary differential equations. Specifically, suppose that $X_1, \ldots, X_n$ are the chemical names of n reacting species (elements or molecules), that $x_j(t)$ denotes the concentration of X_j (in suitable units) as a function of the time t, and that the *rate constant* for the conversion of X_i to X_j is the positive constant k_{ji}. For a two-component reaction, for example, denoted schematically in Fig. 1, this means that

$$\begin{tikzcd} X_1 \arrow[r, shift right, "k_{12}"'] & X_2 \arrow[l, shift right, "k_{21}"'] \end{tikzcd}$$

Figure 1. Two-component reaction.

$$x_1' = -k_{21}x_1 + k_{12}x_2, \tag{16a}$$

$$x_2' = k_{21}x_1 - k_{12}x_2. \tag{16b}$$

The first term on the right-hand side of (16a) accounts for the loss of X_1 due to the $X_1 \to X_2$ reaction; it is proportional to the concentration of X_1, namely x_1, and the constant of proportionality is the relevant rate constant k_{21}. The second term on the right-hand side of (16a) accounts for the rate of gain of X_1 due to the reverse reaction $X_2 \to X_1$. A similar accounting holds for the terms in (16b).

The difficulty in solving (16), and similar systems for n-component reactions where $n > 2$, is due to the coupling. Equations (16) are coupled due to the $k_{12}x_2$ term in (16a) and the $k_{21}x_1$ term in (16b).

Let us solve (16) by diagonalization, if that is possible. In matrix form (16) is

$$\mathbf{x}' = \mathbf{A}\mathbf{x}, \quad \text{where} \quad \mathbf{A} = \begin{bmatrix} -k_{21} & k_{12} \\ k_{21} & -k_{12} \end{bmatrix}. \tag{17}$$

The eigenvalues and eigenvectors of $\mathbf{A}$ are readily found to be

$$\lambda_1 = 0, \ \mathbf{e}_1 = \alpha \begin{bmatrix} k_{12} \\ k_{21} \end{bmatrix}; \qquad \lambda_2 = -(k_{12} + k_{21}), \ \mathbf{e}_2 = \beta \begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{18}$$

The λ_j's are distinct, because $k_{12} > 0$ and $k_{21} > 0$, so Theorem 11.4.3 guarantees that $\mathbf{A}$ is diagonalizable. Alternatively, observe that the $\mathbf{e}_j$'s are necessarily LI because for them to be LD we would need $k_{21} = -k_{12}$, which is impossible since $k_{12} > 0$ and $k_{21} > 0$. Their linear independence implies that $\mathbf{A}$ is diagonalizable, by Theorem 11.4.1.

Thus, if we set $\mathbf{x} = \mathbf{Q}\tilde{\mathbf{x}}$, where (with $\alpha = \beta = 1$, say)

$$\mathbf{Q} = [\mathbf{e}_1, \mathbf{e}_2] = \begin{bmatrix} k_{12} & 1 \\ k_{21} & -1 \end{bmatrix}, \tag{19}$$

then the preceding analysis assures us that

$$\tilde{\mathbf{x}}' = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}\tilde{\mathbf{x}} = \mathbf{D}\tilde{\mathbf{x}} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \tilde{\mathbf{x}}. \tag{20}$$

Thus, we have the *uncoupled* system (which, of course, is our objective)

$$\begin{aligned} \tilde{x}_1' &= \lambda_1 \tilde{x}_1, \\ \tilde{x}_2' &= \lambda_2 \tilde{x}_2, \end{aligned} \tag{21}$$

the general solution of which is

$$\begin{aligned} \tilde{x}_1 &= C_1 e^{\lambda_1 t} = C_1, \\ \tilde{x}_2 &= C_2 e^{\lambda_2 t} = C_2 e^{-(k_{12}+k_{21})t}. \end{aligned} \tag{22}$$

Finally, putting these expressions into $\mathbf{x} = \mathbf{Q}\tilde{\mathbf{x}}$ gives

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} k_{12} & 1 \\ k_{21} & -1 \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 e^{-(k_{12}+k_{21})t} \end{bmatrix} \tag{23}$$

or

$$\begin{aligned} x_1(t) &= C_1 k_{12} + C_2 e^{-(k_{12}+k_{21})t}, \\ x_2(t) &= C_1 k_{21} - C_2 e^{-(k_{12}+k_{21})t}. \end{aligned} \tag{24}$$

COMMENT 1. Here we have emphasized the mathematics rather than the chemistry and have assumed the rate constants to be known. A problem of importance to the chemist is the determination of those constants. Such determination normally involves a blend of the foregoing theory with suitable experiments.

COMMENT 2. The numbering of the eigenvalues and eigenvectors is immaterial. For instance, we could just as well take $\lambda_1 = -(k_{12} + k_{21})$ and $\lambda_2 = 0$. The final result, (24), would be the same. ∎

Theorem 11.4.3 revealed that diagonalizability is the typical case, the generic case, because an nth degree algebraic equation (namely, the characteristic equation of $\mathbf{A}$) typically has distinct roots. Furthermore, every *symmetric* matrix is diagonalizable:

THEOREM 11.4.4 *Symmetric Matrices*
Every symmetric matrix is diagonalizable.

Proof: Theorem 11.4.1 states that $\mathbf{A}$ is diagonalizable if and only if it has n LI eigenvectors, and Theorem 11.3.4 assures us that every $n \times n$ symmetric matrix has n orthogonal (and hence LI) eigenvectors. ∎

Suppose that for a symmetric matrix $\mathbf{A}$ we use the *normalized* eigenvectors of $\mathbf{A}$ to form its modal matrix $\mathbf{Q}$ so that

$$\mathbf{Q} = [\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_n]. \tag{25}$$

Then, observe that

$$\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \begin{bmatrix} \hat{\mathbf{e}}_1^{\mathrm{T}} \\ \vdots \\ \hat{\mathbf{e}}_n^{\mathrm{T}} \end{bmatrix} [\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_n] = \begin{bmatrix} \hat{\mathbf{e}}_1^{\mathrm{T}}\hat{\mathbf{e}}_1 & \cdots & \hat{\mathbf{e}}_1^{\mathrm{T}}\hat{\mathbf{e}}_n \\ \vdots & & \vdots \\ \hat{\mathbf{e}}_n^{\mathrm{T}}\hat{\mathbf{e}}_1 & \cdots & \hat{\mathbf{e}}_n^{\mathrm{T}}\hat{\mathbf{e}}_n \end{bmatrix}$$

$$= \begin{bmatrix} \hat{\mathbf{e}}_1 \cdot \hat{\mathbf{e}}_1 & \cdots & \hat{\mathbf{e}}_1 \cdot \hat{\mathbf{e}}_n \\ \vdots & & \vdots \\ \hat{\mathbf{e}}_n \cdot \hat{\mathbf{e}}_1 & \cdots & \hat{\mathbf{e}}_n \cdot \hat{\mathbf{e}}_n \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} = \mathbf{I} \tag{26}$$

so that

$$\boxed{\mathbf{Q}^{-1} = \mathbf{Q}^{\mathrm{T}}.} \tag{27}$$

Be sure to understand each step in (26). $\mathbf{Q}$ starts out as an $n \times n$ matrix, but when we partition it into columns, as $[\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_n]$, it is then a $1 \times n$ matrix with elements $\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_n$. To form $\mathbf{Q}^{\mathrm{T}}$, we make the jth column of $\mathbf{Q}$, namely $\hat{\mathbf{e}}_j$, the jth row of $\mathbf{Q}^{\mathrm{T}}$, and to put it into row format we need to write it as $\hat{\mathbf{e}}_j^{\mathrm{T}}$ rather than $\hat{\mathbf{e}}_j$. Thus, working with the partitioned $\mathbf{Q}$ and $\mathbf{Q}^{\mathrm{T}}$ matrices, the product to the right of the first equal sign in (26) is an $n \times 1$ matrix times a $1 \times n$ matrix, which product gives the $n \times n$ matrix to the right of the second equal sign.

Understand also that (27) has nothing to do with the $\hat{\mathbf{e}}_j$'s being eigenvectors. The steps in (26) rely only on the fact that the columns of $\mathbf{Q}$ are ON. *Any* square matrix, the columns of which are ON, satisfies (27) and is called an **orthogonal matrix**.* Of course, the property (27) is very nice because if we ever need the inverse of $\mathbf{Q}$ it is simply $\mathbf{Q}^{\mathrm{T}}$.

Getting back to diagonalization, note that if $\mathbf{A}$ is symmetric then, $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{D}$ is diagonal whether or not the columns of the modal matrix $\mathbf{Q}$ are normalized. However, let us agree (at least within this text) to always normalize them, if $\mathbf{A}$ is symmetric, so as to have access to the property (27) if we need it.

Let us close with one more application.

EXAMPLE 2. *A Free-Vibration Problem.* Consider a mass m constrained by two mechanical springs, of stiffnesses k_1 and k_2, as sketched in Fig. 2. Imagine Fig. 2 as a view looking down on the apparatus, which lies in a horizontal plane on a frictionless table. In the configuration shown, the springs are neither stretched nor compressed, and m is at rest in static equilibrium. However, if some initial displacement and/or velocity is imparted to m, some motion, no doubt vibrational, will result, and it is that motion that we wish to determine.

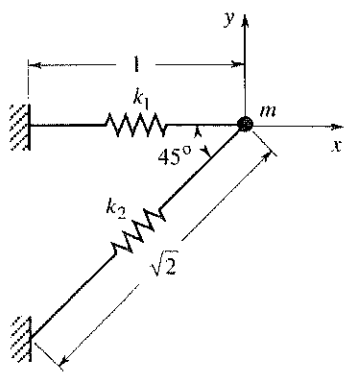


Figure 2. Mass-spring system; view from above.

The first step in the formulation is to introduce a coordinate system. A reasonable choice is the Cartesian system shown in Fig. 2, with its origin at the equilibrium position of the mass m (which we regard as a "point mass").

If m is at some point x, y other than the origin, then one or both springs will be stretched or compressed and will exert forces $\mathbf{F}_1$ and $\mathbf{F}_2$ on m (Fig.2). The magnitude

*Recall our encounter with orthogonal matrices in the optional Section 10.7.

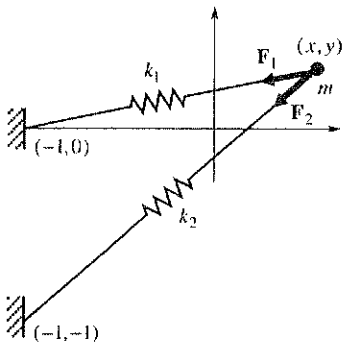


Figure 3. The forces on m.

of $\mathbf{F}_1$ is

$$\begin{aligned}\|\mathbf{F}_1\| &= k_1 \text{ times the stretch in spring \# 1}\\ &= k_1\left\{\sqrt{[x-(-1)]^2+(y-0)^2}-1\right\}\\ &= k_1\left\{\sqrt{(x+1)^2+y^2}-1\right\}.\end{aligned} \tag{28}$$

If we multiply this magnitude by a unit vector directed from (x,y) toward $(-1,0)$, we will have $\mathbf{F}_1$. The vector from (x,y) to $(-1,0)$ is $(-1-x)\hat{\mathbf{i}}+(0-y)\hat{\mathbf{j}}$, where $\hat{\mathbf{i}},\hat{\mathbf{j}}$ are unit base vectors in the x,y directions, respectively. Normalizing that vector gives the desired unit vector

$$\hat{\mathbf{F}}_1 = -\frac{(1+x)\hat{\mathbf{i}}+y\hat{\mathbf{j}}}{\sqrt{(1+x)^2+y^2}}, \tag{29}$$

so

$$\mathbf{F}_1 = \|\mathbf{F}_1\|\,\hat{\mathbf{F}}_1 = -k_1\left[\frac{\sqrt{(x+1)^2+y^2}-1}{\sqrt{(x+1)^2+y^2}}\right]\left[(x+1)\hat{\mathbf{i}}+y\hat{\mathbf{j}}\right]. \tag{30}$$

In like manner, we find that

$$\mathbf{F}_2 = \|\mathbf{F}_2\|\,\hat{\mathbf{F}}_2 = -k_2\left[\frac{\sqrt{(x+1)^2+(y+1)^2}-\sqrt{2}}{\sqrt{(x+1)^2+(y+1)^2}}\right]\left[(x+1)\hat{\mathbf{i}}+(y+1)\hat{\mathbf{j}}\right]. \tag{31}$$

According to Newton's second law,

$$mx'' = F_x \quad\text{and}\quad my'' = F_y, \tag{32}$$

where F_x is the sum of the x components of $\mathbf{F}_1$ and $\mathbf{F}_2$, and F_y is the sum of the y components of $\mathbf{F}_1$ and $\mathbf{F}_2$. Thus, the governing equations of motion are

$$\begin{aligned}mx'' &= -k_1\left[\frac{\sqrt{(x+1)^2+y^2}-1}{\sqrt{(x+1)^2+y^2}}\right](x+1)\\ &\quad -k_2\left[\frac{\sqrt{(x+1)^2+(y+1)^2}-\sqrt{2}}{\sqrt{(x+1)^2+(y+1)^2}}\right](x+1),\end{aligned} \tag{33a}$$

$$\begin{aligned}my'' &= -k_1\left[\frac{\sqrt{(x+1)^2+y^2}-1}{\sqrt{(x+1)^2+y^2}}\right]y\\ &\quad -k_2\left[\frac{\sqrt{(x+1)^2+(y+1)^2}-\sqrt{2}}{\sqrt{(x+1)^2+(y+1)^2}}\right](y+1).\end{aligned} \tag{33b}$$

The latter, coupled, nonlinear differential equations are, clearly, quite intractable analytically. Two possibilities present themselves. First, if we assign numerical values to $m, k_1, k_2, x(0), y(0), x'(0), y'(0)$, then we can generate $x(t)$ and $y(t)$ by one of the numerical methods studied in Chapter 6 (such as fourth-order Runge–Kutta integration) or using computer software (such as the *Maple* dsolve command).

Second, we can limit our attention to small motions, motions that remain close to the equilibrium position at the origin: $|x| \ll 1$ and $|y| \ll 1$. In that case we can simplify (33) in essentially the same way that we can simplify the nonlinear differential equation

$$x'' + \frac{g}{l}\sin x = 0 \tag{34}$$

governing the motion of a pendulum (Fig. 4): for small motions, near the equilibrium point $x = 0$,

$$\begin{aligned}\sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots \qquad \text{(Taylor series)}\\ &\approx x\end{aligned}$$

for $|x| \ll 1$ so (34) can be approximated by the simple linear equation

$$x'' + \frac{g}{l}x = 0. \tag{35}$$

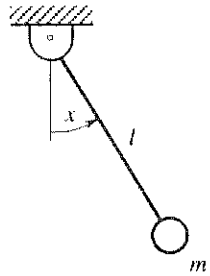


Figure 4. Pendulum, equations (34) and (35).

We will follow the same steps for (33), but instead of the Taylor series in one variable (about $x = 0$),

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \cdots, \tag{36}$$

we will need the Taylor series in two variables (about $x = y = 0$):*

$$\begin{aligned}f(x,y) &= f(0,0) + f_x(0,0)x + f_y(0,0)y\\ &\quad + \frac{1}{2!}\left[f_{xx}(0,0)x^2 + 2f_{xy}(0,0)xy + f_{yy}(0,0)y^2\right] + \cdots\end{aligned} \tag{37}$$

because the right-hand sides of (33a,b) are functions of x and y. First, let $f(x,y)$ in (37) be the right-hand side of (33a). Then (37) gives

$$\begin{aligned}f(x,y) &= 0 + \left(-k_1 - \frac{k_2}{2}\right)x + \left(-\frac{k_2}{2}\right)y + \cdots\\ &\approx -\left(k_1 + \frac{k_2}{2}\right)x - \frac{k_2}{2}y.\end{aligned} \tag{38}$$

Next, let $f(x,y)$ in (37) be the right-hand side of (33b). Then (37) gives

$$\begin{aligned}f(x,y) &= 0 - \frac{k_2}{2}x - \frac{k_2}{2}y + \cdots\\ &\approx -\frac{k_2}{2}x - \frac{k_2}{2}y.\end{aligned} \tag{39}$$

With these approximations (linearizations) of the right-hand sides of (33a,b), we have the *linearized* equations

$$mx'' = -\left(k_1 + \frac{k_2}{2}\right)x - \frac{k_2}{2}y, \tag{40a}$$

$$my'' = -\frac{k_2}{2}x - \frac{k_2}{2}y \tag{40b}$$

*Taylor series in more than one variable is discussed in Chapter 13.

or

$$\mathbf{x}'' + \mathbf{A}\mathbf{x} = \mathbf{0}, \tag{41}$$

where

$$\mathbf{x} = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \dfrac{2k_1 + k_2}{2m} & \dfrac{k_2}{2m} \\ \dfrac{k_2}{2m} & \dfrac{k_2}{2m} \end{bmatrix}. \tag{42}$$

Let us solve (41) by diagonalization. Observe that $\mathbf{A}$ is symmetric (even though there is no "physical" symmetry to be seen in Fig. 2). For definiteness, let us set

$$m = 1, \; k_1 = 3, \; k_2 = 4. \tag{43}$$

Then the eigenvalues and normalized eigenvectors of $\mathbf{A}$ are

$$\lambda_1 = 1, \; \hat{\mathbf{e}}_1 = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 \\ -2 \end{bmatrix}; \qquad \lambda_2 = 6, \; \hat{\mathbf{e}}_2 = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 \\ 1 \end{bmatrix}. \tag{44}$$

With

$$\mathbf{Q} = [\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2] = \begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix}, \tag{45}$$

set $\mathbf{x} = \mathbf{Q}\tilde{\mathbf{x}}$ in (41). Thus,

$$\mathbf{Q}\tilde{\mathbf{x}}'' + \mathbf{A}\mathbf{Q}\tilde{\mathbf{x}} = \mathbf{0} \tag{46}$$

and hence

$$\tilde{\mathbf{x}}'' + \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}\tilde{\mathbf{x}} = \mathbf{0} \tag{47}$$

or

$$\tilde{\mathbf{x}}'' + \mathbf{D}\tilde{\mathbf{x}} = \mathbf{0}, \tag{48}$$

where

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 6 \end{bmatrix}. \tag{49}$$

In scalar form, (48) gives the *uncoupled* equations

$$\begin{aligned} \tilde{x}'' + \tilde{x} &= 0, \\ \tilde{y}'' + 6\tilde{y} &= 0 \end{aligned} \tag{50}$$

with general solution [expressed in the $A \sin(\omega t + \phi)$ form]:

$$\begin{aligned} \tilde{x} &= A_1 \sin(t + \phi_1), \\ \tilde{y} &= A_2 \sin(\sqrt{6}\, t + \phi_2), \end{aligned} \tag{51}$$

where the amplitudes A_1, A_2 and phase angles ϕ_1, ϕ_2 are the four constants of integration. To return to the original x, y variables, write

$$\begin{aligned} \begin{bmatrix} x \\ y \end{bmatrix} = \mathbf{Q}\tilde{\mathbf{x}} &= \begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} \\ &= A_1 \begin{bmatrix} \dfrac{1}{\sqrt{5}} \sin(t + \phi_1) \\ -\dfrac{2}{\sqrt{5}} \sin(t + \phi_1) \end{bmatrix} + A_2 \begin{bmatrix} \dfrac{2}{\sqrt{5}} \sin(\sqrt{6}\, t + \phi_2) \\ \dfrac{1}{\sqrt{5}} \sin(\sqrt{6}\, t + \phi_2) \end{bmatrix} \\ &= A_1 \hat{\mathbf{e}}_1 \sin(t + \phi_1) + A_2 \hat{\mathbf{e}}_2 \sin(\sqrt{6}\, t + \phi_2) \end{aligned} \tag{52}$$

or, in scalar form,

$$\begin{aligned} x(t) &= C_1 \sin(t + \phi_1) + 2C_2 \sin(\sqrt{6}\,t + \phi_2), \\ y(t) &= -2C_1 \sin(t + \phi_1) + C_2 \sin(\sqrt{6}\,t + \phi_2), \end{aligned} \tag{53}$$

where C_1 is $A_1/\sqrt{5}$ and C_2 is $A_2/\sqrt{5}$, for brevity.

COMMENT 1. It is seen from (52) that the general solution is a linear combination of two orthogonal modes, as in Example 2 of Section 11.3. The low mode is a vibration along the $\hat{\mathbf{e}}_1$ direction, at a frequency that is the square root of λ_1, and the high mode is a vibration along the $\hat{\mathbf{e}}_2$ direction, at a frequency that is the square root of λ_2, as summarized in Fig. 5. In this example the orthogonality of the modes is geometric since the low- and high-mode motions are 90° apart; in Example 2 of Section 11.3, the orthogonality is more mathematical ($\mathbf{e}_1 \cdot \mathbf{e}_2$ being zero) than geometric.

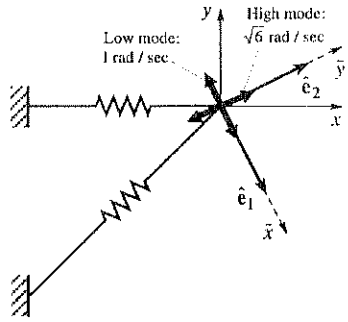


Figure 5. The orthogonal modes.

COMMENT 2. Why are the directions of the low and high modes, shown in Fig. 5, physically reasonable? Recall that the natural frequency of the classical harmonic mechanical oscillator, governed by the equation $mx'' + kx = 0$, is $\sqrt{k/m}$, which increases as the stiffness k increases. It should be clear intuitively that the $\mathbf{e}_2$ axis is the line of maximum stiffness (of the two-spring system) and the $\mathbf{e}_1$ axis is the line of minimum stiffness, at least roughly speaking. Strikingly, the mathematics reveals that these directions are necessarily 90° apart.

COMMENT 3. Why do we show the positive $\tilde{x}$ and $\tilde{y}$ coordinate axes as being in the $\hat{\mathbf{e}}_1$ and $\hat{\mathbf{e}}_2$ directions, respectively, in Fig. 5? Because if we set $\tilde{x} = 1$ and $\tilde{y} = 0$ in

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix}$$

we get $[x, y]^{\mathrm{T}} = [1/\sqrt{5}, -2/\sqrt{5}]^{\mathrm{T}} = \hat{\mathbf{e}}_1$, and if we set $\tilde{x} = 0$ and $\tilde{y} = 1$ we get $[x, y]^{\mathrm{T}} = [2/\sqrt{5}, 1/\sqrt{5}]^{\mathrm{T}} = \hat{\mathbf{e}}_2$. In fact, if you studied the optional Section 10.7, you will appreciate that the effect of the change of variables $\mathbf{x} = \mathbf{Q}\tilde{\mathbf{x}}$, where $\mathbf{Q}$ is an *orthogonal* matrix with its determinant equal to $+1$, as here, is a pure *rotation* of axes. Thus, we have the vivid visual image of the elements of the coupling matrix varying as we rotate the Cartesian coordinate system (somewhat like looking into a kaleidoscope), until the off-diagonal terms become zero and the equations uncouple.

COMMENT 4. In principle, it would have been best to choose the $\tilde{x}, \tilde{y}$ coordinate system in the first place, but its orientation was not known. Thus, we chose any x, y system, to get started, and then used the method of diagonalization to find the optimal $\tilde{x}, \tilde{y}$ coordinate system. ∎

Closure. From a mathematical viewpoint, this section is about finding an $n \times n$ matrix $\mathbf{Q}$, given an $n \times n$ matrix $\mathbf{A}$, such that $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{D}$ is diagonal. We find that in the generic case $\mathbf{A}$ is diagonalizable: it is diagonalizable if and only if it has n LI eigenvectors, and it is diagonalizable if it has n distinct eigenvalues or is symmetric. $\mathbf{Q}$ can be made up of columns which are the eigenvectors of $\mathbf{A}$, and the diagonal elements of $\mathbf{D}$ are the corresponding eigenvalues. In the event that $\mathbf{A}$

is symmetric, we suggest always normalizing the eigenvectors that are the columns of $\mathbf{Q}$, so that $\mathbf{Q}$ will admit the useful property $\mathbf{Q}^{-1} = \mathbf{Q}^{\mathrm{T}}$; that is, so that $\mathbf{Q}$ will be an orthogonal matrix.

From an applications standpoint, we look only at the use of digonalization in uncoupling systems of coupled differential equations, but additional applications are to be found in the exercises and in the next two sections.

It turns out that even if an $n \times n$ matrix $\mathbf{A}$ cannot be diagonalized, it can be triangularized. That is, a generalized modal matrix $\mathbf{P}$ can be found for $\mathbf{A}$ so that

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J} \tag{54}$$

is triangular. Called the **Jordan normal form**, or simply the **Jordan form**, for $\mathbf{A}$, $\mathbf{J}$ is upper triangular, with zeros above its main diagonal – except for 1's immediately above one or more diagonal elements. This case is discussed briefly in the exercises.

EXERCISES 11.4

1. Diagonalize each of the given $\mathbf{A}$ matrices. That is, determine matrices $\mathbf{Q}$ and $\mathbf{D}$ such that $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{D}$ is diagonal. Also, work out $\mathbf{Q}^{-1}$ and verify that $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$ is diagonal and that its diagonal elements are the eigenvalues of $\mathbf{A}$. If $\mathbf{A}$ is not diagonalizable, state that and give the reason.

(a) $\begin{bmatrix} 2 & -3 \\ 0 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 4 \\ -1 & -2 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$ (d) $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$

(e) $\begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$ (f) $\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$

(g) $\begin{bmatrix} 2 & 1 & -1 \\ 1 & 4 & 3 \\ -1 & 3 & 4 \end{bmatrix}$ (h) $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

(i) $\begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$ (j) $\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$

2. Use the method of diagonalization to obtain the general solution of the given system of differential equations, where primes denote d/dt.

(a) $x' = x + y$, $y' = x + y$ (b) $x' = x + 4y$, $y' = x + y$

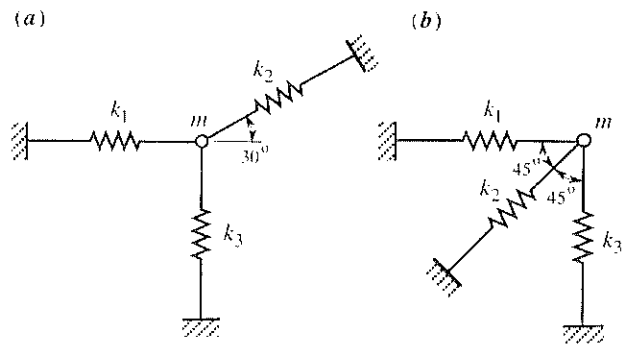
(c) $x'' = 2x + 4y$, $y'' = x - y$ (d) $x' + 2x + y = 0$, $y' + x + 2y + z = 0$, $z' + y + 2z = 0$

(e) $x' = 4x + y + 3z$, $y' = x - z$, $z' = 3x - y + 4z$ (f) $x'' = -y + z$, $y'' = -x - 3y - 4z$, $z'' = x - 4y - 3z$

3. Can a singular matrix be diagonalized? Explain.

4. We see from Fig. 5 that the line of action of the high mode falls between the two springs. Show that that situation holds for all possible combinations of stiffnesses k_1 and k_2.

5. Determine (as in Example 2) the natural frequencies and mode shapes for each of the systems shown below. You need carry only three or four significant figures. Each spring is of unit length.



$m = k_1 = k_2 = 1$, $k_3 = 10$ $m = k_1 = 4$, $k_2 = 3$, $k_3 = 1$

6. Show why the second equality in (9) is true.

7. (*Application to exponentiation*) Diagonalization can be helpful in raising a square matrix to a large power. Specifically, show that if $\mathbf{A}$ is diagonalizable so that $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{D}$, then

$$\mathbf{A}^m = \mathbf{Q}\mathbf{D}^m\mathbf{Q}^{-1}, \tag{7.1}$$

the advantage being that $\mathbf{D}^m$ is simply

$$\mathbf{D}^m = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix}^m = \begin{bmatrix} \lambda_1^m & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n^m \end{bmatrix}. \tag{7.2}$$

8. Use (7.1), above, to evaluate $\mathbf{A}^{1000}$, where

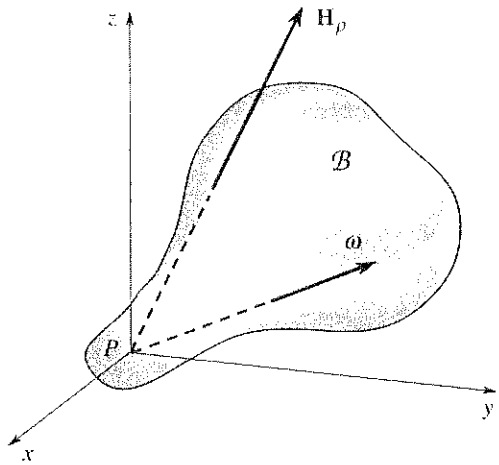
(a) $\mathbf{A} = \begin{bmatrix} 0 & 2 & 2 \\ 2 & 0 & 2 \\ 2 & 2 & 0 \end{bmatrix}$ (b) $\mathbf{A} = \begin{bmatrix} 2 & 2 & 1 \\ 1 & 3 & 1 \\ 1 & 2 & 2 \end{bmatrix}$

9. (*Application to principal inertias and principal axes*) Two vectors of importance in studying the dynamics of a rigid body $\mathcal{B}$ are the *moment of momentum* $\mathbf{H}_p$ and the *angular velocity* $\boldsymbol{\omega}$, of $\mathcal{B}$. These are related according to $\mathbf{H}_p = \mathcal{I}\boldsymbol{\omega}$, where $\mathcal{I}$ is the *inertia matrix*. Written out, we have

$$\begin{bmatrix} (H_p)_x \\ (H_p)_y \\ (H_p)_z \end{bmatrix} = \begin{bmatrix} I_{xx} & -I_{xy} & -I_{xz} \\ -I_{yx} & I_{yy} & -I_{yz} \\ -I_{zx} & -I_{zy} & I_{zz} \end{bmatrix} \begin{bmatrix} \omega_x \\ \omega_y \\ \omega_z \end{bmatrix}, \tag{9.1}$$

where P is the origin of a Cartesian x, y, z coordinate system (see the figure), and

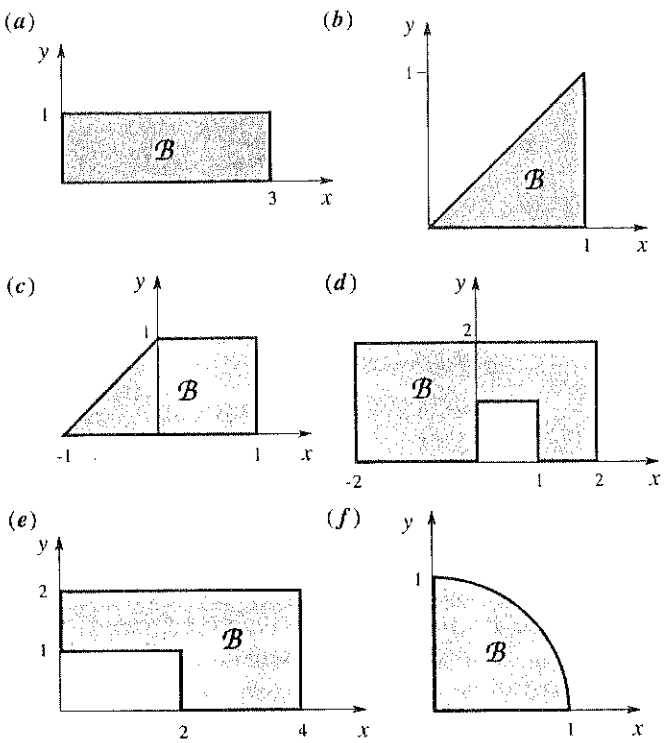
$$\begin{aligned} I_{xx} &= \int_{\mathcal{B}} (y^2 + z^2)\, dm, & I_{xy} &= I_{yx} = \int_{\mathcal{B}} xy\, dm \\ I_{yy} &= \int_{\mathcal{B}} (x^2 + z^2)\, dm, & I_{xz} &= I_{zx} = \int_{\mathcal{B}} xz\, dm \\ I_{zz} &= \int_{\mathcal{B}} (x^2 + y^2)\, dm, & I_{yz} &= I_{zy} = \int_{\mathcal{B}} yz\, dm. \end{aligned} \tag{9.2}$$



I_{xx}, I_{yy}, I_{zz} are known as the *moments of inertia* of $\mathcal{B}$ about the x, y, z axes, respectively, and I_{xy}, I_{xz}, I_{yz} are known as the *products of inertia* of $\mathcal{B}$; dm in (9.2) is "d(mass)." Now, the relation (9.1) and hence the subsequent dynamic analysis (which will be of no concern here) will be simplest if $\mathcal{I}$ is diagonal, i.e., if all of the products of inertia are zero. In general, it is too difficult to see, by inspection, how to orient the coordinate axes to achieve this result. Instead, we go ahead and choose *some* x, y, z reference frame, compute the nine inertia components, and then rotate to a new Cartesian $\tilde{x}, \tilde{y}, \tilde{z}$ frame so as to diagonalize $\mathcal{I}$. That is, if $\mathbf{x} = \mathbf{Q}\tilde{\mathbf{x}}$, where $\mathbf{x} = [x, y, z]^{\mathrm{T}}$ and $\tilde{\mathbf{x}} = [\tilde{x}, \tilde{y}, \tilde{z}]^{\mathrm{T}}$, then $\mathbf{H}_p = \mathbf{Q}\tilde{\mathbf{H}}_p$ and $\boldsymbol{\omega} = \mathbf{Q}\tilde{\boldsymbol{\omega}}$ so that $\mathbf{H}_p = \mathcal{I}\boldsymbol{\omega}$ becomes

$$\mathbf{Q}\tilde{\mathbf{H}}_p = \mathcal{I}\mathbf{Q}\tilde{\boldsymbol{\omega}} \quad \text{and} \quad \tilde{\mathbf{H}}_p = \left(\mathbf{Q}^{-1}\mathcal{I}\mathbf{Q}\right)\tilde{\boldsymbol{\omega}}, \tag{9.3}$$

where $\mathbf{Q}^{-1}\mathcal{I}\mathbf{Q} = \tilde{\mathcal{I}}$ is diagonal. That such diagonalization is possible follows from the fact that $\mathcal{I}$ is *symmetric* since $I_{xy} = I_{yx}$, $I_{xz} = I_{zx}$, and $I_{yz} = I_{zy}$. $I_{\tilde{x}\tilde{x}}$, $I_{\tilde{y}\tilde{y}}$, and $I_{\tilde{z}\tilde{z}}$ are called the **principal inertias** of $\mathcal{B}$ (with respect to coordinates with origin at P), and the $\tilde{x}, \tilde{y}, \tilde{z}$ axes are called the **principal axes**. We now state the problem: compute the principal inertias and determine the principal axes for each of the following bodies; sketch the principal axes. In each case $\mathcal{B}$ can be assumed, for simplicity, to be infinitely thin, with mass density σ mass units per unit area.



10. (*Jordan form*) Recall that an $n \times n$ matrix $\mathbf{A}$ is diagonalizable if and only if it has n LI eigenvectors. Thus, not all matrices can be diagonalized. However, our experience in Chapters 8–11 has shown that triangular matrices are "almost as nice" as diagonal ones, and it turns out that even if $\mathbf{A}$ cannot be diagonalized, it can be triangularized.

More specifically, if $\mathbf{A}$ is diagonalizable and its modal matrix is $\mathbf{Q} = [\mathbf{e}_1, \ldots, \mathbf{e}_n]$, where $\mathbf{e}_j$'s are the LI eigenvectors of $\mathbf{A}$, then $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{D}$ is diagonal, with d_{jj} equal to the jth eigenvalue of $\mathbf{A}$. But suppose $\mathbf{A}$ is *not* diagonalizable. Though it does not have n LI eigenvectors, it has n LI **generalized eigenvectors** $\mathbf{e}_1, \ldots, \mathbf{e}_n$ (defined below) and if its **generalized modal matrix** is $\mathbf{P} = [\mathbf{e}_1, \ldots, \mathbf{e}_n]$, then $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{J}$, where $\mathbf{J}$ is the **Jordan form** for $\mathbf{A}$; $\mathbf{J}$ will be upper triangular, with the eigenvalues of $\mathbf{A}$ on its diagonal and zeros everywhere else except perhaps immediately above repeated eigenvalues. For details, we refer you to M. Greenberg, *Advanced Engineering Mathematics*, 1st ed. (Englewood Cliffs, NJ: Prentice Hall, 1988). Here, we will merely go through an example with you and, at the end, ask you to supply various steps.

Consider, as a representative case,

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 2 & 0 \\ 0 & 3 & -1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & -3 & 5 \end{bmatrix}. \tag{10.1}$$

Calculation reveals these eigenvalues and eigenvectors:

$$\begin{aligned} &\lambda_1 = \lambda_2 = \lambda_3 = 2, && \mathbf{e} = [1,0,0,0]^{\mathrm{T}} \equiv \mathbf{e}_1, \\ &\lambda_4 = 5, && \mathbf{e} = [0,0,0,1]^{\mathrm{T}} \equiv \mathbf{e}_4. \end{aligned} \tag{10.2}$$

In this case the eigenvalue of multiplicity three, $\lambda = 2$, contributes only one LI eigenvector, instead of three, so we end up with only two LI eigenvectors, $\mathbf{e}_1$ and $\mathbf{e}_4$, instead of the four that are needed for diagonalization. Thus, $\mathbf{A}$ is not diagonalizable. But we can find "generalized eigenvectors" $\mathbf{e}_2$ and $\mathbf{e}_3$, associated with $\lambda = 2$, such that $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4$ are LI. Noting that $\mathbf{e}_1$ satisfies

$$(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{e}_1 = \mathbf{0}, \tag{10.3}$$

we introduce $\mathbf{e}_2, \mathbf{e}_3$ so as to satisfy

$$(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{e}_2 = \mathbf{e}_1, \tag{10.4}$$

$$(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{e}_3 = \mathbf{e}_2. \tag{10.5}$$

With λ_1 and $\mathbf{e}_1$ given in (10.2), solution of (10.4) and (10.5) by Gauss elimination gives

$$\mathbf{e}_2 = [\alpha, 1, 1, 2/3]^{\mathrm{T}}, \tag{10.6}$$

$$\mathbf{e}_3 = [\beta, \alpha + 2, \alpha + 1, 2\alpha/3 + 5/9]^{\mathrm{T}}, \tag{10.7}$$

where α, β are arbitrary. Take $\alpha = \beta = 0$, say. Then, with the generalized modal matrix

$$\mathbf{P} = [\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 2/3 & 5/9 & 1 \end{bmatrix}, \tag{10.8}$$

we obtain

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \begin{bmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 1 & 0 \\ 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix} = \mathbf{J}. \tag{10.9}$$

(a) Derive the eigenvalues and eigenvectors given in (10.2).

(b) Use Gauss elimination to derive (10.6) from (10.4), and (10.7) from (10.5).

(c) Use (10.3), (10.4), (10.5), and $(\mathbf{A} - \lambda_4\mathbf{I})\mathbf{e}_4 = \mathbf{0}$ to show that $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4$ are LI, as claimed above (10.3). HINT: Suppose that

$$\alpha_1\mathbf{e}_1 + \alpha_2\mathbf{e}_2 + \alpha_3\mathbf{e}_3 + \alpha_4\mathbf{e}_4 = \mathbf{0}. \tag{10.10}$$

Then multiply each term in (10.10) by $\mathbf{A} - \lambda_1\mathbf{I}$. Then multiply each term in the resulting equation by $\mathbf{A} - \lambda_1\mathbf{I}$. Again multiply each term in the resulting equation by $\mathbf{A} - \lambda_1\mathbf{I}$. Explain why the resulting set of four equations implies that $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$. NOTE: More generally, let λ be an eigenvalue of the given matrix $\mathbf{A}$, of multiplicity K, and let the eigenspace corresponding to λ be of dimension k, where $k < K$. Within that eigenspace there can be found k LI eigenvectors of $\mathbf{A}$, say $\mathbf{e}_1$ through $\mathbf{e}_k$. Vectors $\mathbf{e}_{k+1}$ through $\mathbf{e}_K$ satisfying

$$\begin{aligned}(\mathbf{A} - \lambda\mathbf{I})\mathbf{e}_{k+1} &= \mathbf{e}_k,\\ (\mathbf{A} - \lambda\mathbf{I})\mathbf{e}_{k+2} &= \mathbf{e}_{k+1},\\ &\vdots\\ (\mathbf{A} - \lambda\mathbf{I})\mathbf{e}_K &= \mathbf{e}_{K-1}\end{aligned} \tag{10.11}$$

are the **generalized eigenvectors** corresponding to λ.

(d) Consider the system of differential equations

$$\mathbf{x}' = \mathbf{A}\mathbf{x}, \tag{10.12}$$

where the prime denotes d/dt and $\mathbf{A}$ is given by (10.1). Under the change of variables $\mathbf{x} = \mathbf{P}\widetilde{\mathbf{x}}$, reduce (10.12) to the Jordan form

$$\widetilde{\mathbf{x}}' = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}\widetilde{\mathbf{x}} = \mathbf{J}\widetilde{\mathbf{x}}, \tag{10.13}$$

where $\mathbf{J}$ is given in (10.9). Solve the triangular system (10.13) for $\widetilde{\mathbf{x}}(t)$, and thus obtain the general solution $\mathbf{x}(t) = \mathbf{P}\widetilde{\mathbf{x}}(t)$ of (10.12).

11.5 Application to First-Order Systems with Constant Coefficients (Optional)

In Section 11.4 we studied diagonalization, and showed how to use that method to uncouple and solve systems of differential equations. In this section we continue that discussion, but this time our emphasis is on the theory of differential equations rather than on diagonalization.

Consider the initial-value problem

$$\begin{aligned} x_1' &= a_{11}x_1 + \cdots + a_{1n}x_n + f_1(t); \qquad x_1(0) = c_1\\ &\vdots\\ x_n' &= a_{n1}x_1 + \cdots + a_{nn}x_n + f_n(t); \qquad x_n(0) = c_n\end{aligned} \tag{1}$$

where the a_{ij}'s are constants. Or, in matrix form,

$$\mathbf{x}' = \mathbf{A}\mathbf{x} + \mathbf{f}(t); \qquad \mathbf{x}(0) = \mathbf{c}. \tag{2}$$

Unlike Section 11.4, here we allow for forcing functions $f_1(t), \ldots, f_n(t)$.

To solve the first-order system (2) by diagonalization, we suppose that $\mathbf{A}$ has n LI eigenvectors, a modal matrix $\mathbf{Q}$, and eigenvalues $\lambda_1, \ldots, \lambda_n$ that are not necessarily distinct. Setting

$$\mathbf{x}(t) = \mathbf{Q}\widetilde{\mathbf{x}}(t), \tag{3}$$

(2) becomes

$$\mathbf{Q}\widetilde{\mathbf{x}}' = \mathbf{A}\mathbf{Q}\widetilde{\mathbf{x}} + \mathbf{f}(t); \qquad \mathbf{Q}\widetilde{\mathbf{x}}(0) = \mathbf{c} \tag{4}$$

or, equivalently,

$$\widetilde{\mathbf{x}}' = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}\widetilde{\mathbf{x}} + \mathbf{Q}^{-1}\mathbf{f}(t); \qquad \widetilde{\mathbf{x}}(0) = \mathbf{Q}^{-1}\mathbf{c}, \tag{5}$$

where $\mathbf{Q}^{-1}$ necessarily exists because the columns of $\mathbf{Q}$ are LI. We know that

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & \lambda_n \end{bmatrix}, \tag{6}$$

and if we denote $\mathbf{Q}^{-1}\mathbf{f}(t)$ as $\widetilde{\mathbf{f}}(t)$ and $\mathbf{Q}^{-1}\mathbf{c}$ as $\widetilde{\mathbf{c}}$, for brevity, then

$$\widetilde{\mathbf{x}}' = \mathbf{D}\widetilde{\mathbf{x}} + \widetilde{\mathbf{f}}(t); \qquad \widetilde{\mathbf{x}}(0) = \widetilde{\mathbf{c}} \tag{7}$$

or, returning to scalar form,

$$\begin{aligned} \widetilde{x}_1' &= \lambda_1\widetilde{x}_1 + \widetilde{f}_1(t); \qquad \widetilde{x}_1(0) = \widetilde{c}_1 \\ &\vdots \\ \widetilde{x}_n' &= \lambda_n\widetilde{x}_n + \widetilde{f}_n(t); \qquad \widetilde{x}_n(0) = \widetilde{c}_n. \end{aligned} \tag{8}$$

Each of these equations is first-order linear, like equation (2) in Section 2.2, and its solution is given by (24) in Section 2.2.2 [with $x \to t, y \to \widetilde{x}_j, p(x) \to -\lambda_j$, and so on]:

$$\begin{aligned} \widetilde{x}_1(t) &= \widetilde{c}_1 e^{\lambda_1 t} + \int_0^t e^{\lambda_1(t-\tau)}\widetilde{f}_1(\tau)d\tau, \\ &\vdots \\ \widetilde{x}_n(t) &= \widetilde{c}_n e^{\lambda_n t} + \int_0^t e^{\lambda_n(t-\tau)}\widetilde{f}_n(\tau)d\tau. \end{aligned} \tag{9}$$

If we define

$$e^{\mathbf{D}t} \equiv \begin{bmatrix} e^{\lambda_1 t} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2 t} & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & e^{\lambda_n t} \end{bmatrix}, \tag{10}$$

then we can express (9) as

$$\widetilde{\mathbf{x}}(t) = e^{\mathbf{D}t}\widetilde{\mathbf{c}} + \int_0^t e^{\mathbf{D}(t-\tau)}\widetilde{\mathbf{f}}(\tau)d\tau. \tag{11}$$

Finally, since $\widetilde{\mathbf{x}} = \mathbf{Q}^{-1}\mathbf{x}$, $\widetilde{\mathbf{c}} = \mathbf{Q}^{-1}\mathbf{c}$, and $\widetilde{\mathbf{f}} = \mathbf{Q}^{-1}\mathbf{f}$,

$$\boxed{\mathbf{x}(t) = \mathbf{Q}e^{\mathbf{D}t}\mathbf{Q}^{-1}\mathbf{c} + \int_0^t \mathbf{Q}e^{\mathbf{D}(t-\tau)}\mathbf{Q}^{-1}\mathbf{f}(\tau)d\tau} \tag{12}$$

is the unique solution of (1). Note the order: $e^{\mathbf{D}t}\widetilde{\mathbf{c}}$, not $\widetilde{\mathbf{c}}e^{\mathbf{D}t}$, in (11).

EXAMPLE 1. Use (12) to solve the system

$$\begin{aligned} x' &= x + 4y - 4t^2 - 3; \qquad & x(0) &= 2 \\ y' &= x + y - t^2 + 2t - 3; \qquad & y(0) &= 3. \end{aligned} \tag{13}$$

Then

$$\mathbf{A} = \begin{bmatrix} 1 & 4 \\ 1 & 1 \end{bmatrix}, \quad \mathbf{f}(t) = \begin{bmatrix} -4t^2 - 3 \\ -t^2 + 2t - 3 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}. \tag{14}$$

The eigenvalues and eigenvectors of $\mathbf{A}$ are found to be

$$\lambda_1 = 3,\ \mathbf{e}_1 = \alpha \begin{bmatrix} 2 \\ 1 \end{bmatrix}; \qquad \lambda_2 = -1,\ \mathbf{e}_1 = \beta \begin{bmatrix} -2 \\ 1 \end{bmatrix}. \tag{15}$$

Since these $\mathbf{e}_j$'s are LI, $\mathbf{A}$ is diagonalizable so (12) applies. With

$$\mathbf{D} = \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} 2 & -2 \\ 1 & 1 \end{bmatrix}, \quad \mathbf{Q}^{-1} = \begin{bmatrix} \frac{1}{4} & \frac{1}{2} \\ -\frac{1}{4} & \frac{1}{2} \end{bmatrix}, \tag{16}$$

(12) gives

$$\begin{aligned} \mathbf{x}(t) &= \begin{bmatrix} 2 & -2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} e^{3t} & 0 \\ 0 & e^{-t} \end{bmatrix} \begin{bmatrix} \frac{1}{4} & \frac{1}{2} \\ -\frac{1}{4} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} \\ &\quad + \int_0^t \begin{bmatrix} 2 & -2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} e^{3(t-\tau)} & 0 \\ 0 & e^{-(t-\tau)} \end{bmatrix} \begin{bmatrix} \frac{1}{4} & \frac{1}{2} \\ -\frac{1}{4} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} -4\tau^2 - 3 \\ -\tau^2 + 2\tau - 3 \end{bmatrix} d\tau \\ &= \begin{bmatrix} 4e^{3t} - 2e^{-t} \\ 2e^{3t} + e^{-t} \end{bmatrix} + \begin{bmatrix} 3 - \frac{3}{2}e^{3t} - \frac{3}{2}e^{-t} \\ t^2 - \frac{3}{4}e^{3t} + \frac{3}{4}e^{-t} \end{bmatrix} \end{aligned} \tag{17}$$

so

$$\begin{aligned} x(t) &= 3 + \frac{5}{2}e^{3t} - \frac{7}{2}e^{-t}, \\ y(t) &= t^2 + \frac{5}{4}e^{3t} + \frac{7}{4}e^{-t} \end{aligned} \tag{18}$$

is the desired solution. ∎

EXAMPLE 2. Can (12) be used for the system

$$x'' + 3x' + 2x - y = e^t \tag{19a}$$

$$y'' - y' + 5x' - x - 6y = t\ ? \tag{19b}$$

Yes, if we can reduce (19) to a *first*-order system. How to do that is described in Section 3.9 (see Example 5, therein, and the paragraph preceding that example). Specifically, set $x' = u$ and $y' = v$. Then (19) can be re-expressed as

$$\begin{aligned} x' &= u, && \text{(by definition)} \\ u' &= -3u - 2x + y + e^t, && \text{[from (19a)]} \\ y' &= v, && \text{(by definition)} \\ v' &= v - 5u + x + 6y + t && \text{[from (19b)]} \end{aligned} \tag{20}$$

or, in matrix form, as

$$\begin{bmatrix} x' \\ u' \\ y' \\ v' \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -2 & -3 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & -5 & 6 & 1 \end{bmatrix} \begin{bmatrix} x \\ u \\ y \\ v \end{bmatrix} + \begin{bmatrix} 0 \\ e^t \\ 0 \\ t \end{bmatrix}. \tag{21}$$

The latter is a first-order system, to which (12) could be applied. ∎

In deriving (12) an interesting quantity arose, the **exponential matrix function** $e^{\mathbf{B}}$, where $\mathbf{B} = \{b_{ij}\}$ is an $n \times n$ diagonal matrix:

$$\boxed{e^{\mathbf{B}} \equiv \begin{bmatrix} e^{b_{11}} & 0 & \cdots & 0 \\ 0 & e^{b_{22}} & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & e^{b_{nn}} \end{bmatrix}.} \quad (\mathbf{B} \text{ diagonal}) \tag{22}$$

More generally, let $\mathbf{B}$ be any $n \times n$ matrix, not necessarily diagonal. Following the familiar Taylor series formula,

$$e^x = 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + \cdots,$$

it seems reasonable to define

$$\boxed{e^{\mathbf{B}} \equiv \mathbf{I} + \frac{1}{1!}\mathbf{B} + \frac{1}{2!}\mathbf{B}^2 + \cdots.} \tag{23}$$

Fine, but observe that the right-hand side of (23) is an *infinite series of matrices*, which we have not yet defined. Mimicing the usual definition of convergence for series of scalars, we define an infinite series of matrices $\sum_{j=1}^{\infty} \mathbf{A}_j$ as the limit of the sequence of partial sums $\mathbf{S}_N$, where $\mathbf{S}_N = \sum_{j=1}^{N} \mathbf{A}_j$. That is,

$$\sum_{j=1}^{\infty} \mathbf{A}_j \equiv \lim_{N\to\infty} \mathbf{S}_N. \tag{24}$$

The infinite series is said to **converge** if the limit on the right exists and to diverge if that limit does not exist. Finally, observe that $\lim_{N\to\infty} \mathbf{S}_N$ is the *limit of a sequence of matrices*, which we have not yet defined so we are not done.

Let $\mathbf{C}_1, \mathbf{C}_2, \ldots$ be a sequence of $m \times n$ matrices, with $(c_{ij})_n$ as the i, j element of $\mathbf{C}_n$. We say that the sequence **converges** to a matrix $\mathbf{C} = \{c_{ij}\}$ if

$$\lim_{n\to\infty} (c_{ij})_n = c_{ij}$$

for each i, j, and we denote such convergence by writing either $\lim_{n\to\infty} \mathbf{C}_n = \mathbf{C}$ or $\mathbf{C}_n \to \mathbf{C}$ as $n \to \infty$. If the sequence does not converge, then it is said to **diverge**.

EXAMPLE 3.

$$\begin{bmatrix} 2 + \dfrac{3}{n} & e^{-n} \\ 1 & \dfrac{4}{n+2} \end{bmatrix} \to \begin{bmatrix} 2 & 0 \\ 1 & 0 \end{bmatrix}, \tag{25}$$

but

$$\begin{bmatrix} 7 - 3n & 2 + e^{-4n} \\ 3 & \dfrac{1}{n} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 2 & 5 \\ \sin n & \dfrac{1}{n+2} \end{bmatrix} \tag{26}$$

diverge because $\lim_{n\to\infty}(7 - 3n)$ and $\lim_{n\to\infty} \sin n$ do not exist. ∎

With (23), let us return to (12) and consider the combination $\mathbf{Q}e^{\mathbf{D}t}\mathbf{Q}^{-1}$:

$$\begin{aligned} \mathbf{Q}e^{\mathbf{D}t}\mathbf{Q}^{-1} &= \mathbf{Q}\left(\mathbf{I} + \mathbf{D}t + \frac{1}{2!}\mathbf{D}^2t^2 + \cdots\right)\mathbf{Q}^{-1} \\ &= \mathbf{I} + \mathbf{Q}\mathbf{D}\mathbf{Q}^{-1}t + \frac{1}{2!}\mathbf{Q}\mathbf{D}^2\mathbf{Q}^{-1}t^2 + \cdots \\ &= \mathbf{I} + \mathbf{Q}\mathbf{D}\mathbf{Q}^{-1}t + \frac{1}{2!}\mathbf{Q}\mathbf{D}\mathbf{Q}^{-1}\mathbf{Q}\mathbf{D}\mathbf{Q}^{-1}t^2 + \cdots, \end{aligned} \tag{27}$$

and recalling that $\mathbf{Q}\mathbf{D}\mathbf{Q}^{-1} = \mathbf{A}$, this series gives

$$\mathbf{Q}e^{\mathbf{D}t}\mathbf{Q}^{-1} = \mathbf{I} + \mathbf{A}t + \frac{1}{2!}\mathbf{A}^2t^2 + \cdots = e^{\mathbf{A}t}. \tag{28}$$

Similarly, $\mathbf{Q}e^{\mathbf{D}(t-\tau)}\mathbf{Q}^{-1} = e^{\mathbf{A}(t-\tau)}$ so (12) can be expressed in the equivalent form

$$\boxed{\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{c} + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{f}(\tau)d\tau.} \tag{29}$$

How are we to evaluate the exponentials in (29)? We could use the series formula (23), but it is more efficient to reverse (28) and to use $e^{\mathbf{A}t} = \mathbf{Q}e^{\mathbf{D}t}\mathbf{Q}^{-1}$ and $e^{\mathbf{A}(t-\tau)} = \mathbf{Q}e^{\mathbf{D}(t-\tau)}\mathbf{Q}^{-1}$ because $e^{\mathbf{D}t}$ is given simply by (10), and similarly

for $e^{\mathbf{D}(t-\tau)}$. That is, computationally, we continue to rely on (12) rather than (29). Nonetheless, (29) is of considerable interest because it shows how the solution

$$x(t) = e^{At}c + \int_0^t e^{A(t-\tau)} f(\tau)d\tau \tag{30}$$

of the single initial-value problem

$$x'(t) = Ax + f(t); \qquad x(0) = c \tag{31}$$

can be generalized to the solution (29) of (2) using the exponential matrix function.*

Closure. There were two objectives in this section. One was to derive the matrix solution (12) of the initial-value problem (2), and the other was to introduce the exponential matrix function $e^{\mathbf{B}}$ for any $n \times n$ matrix $\mathbf{B}$. In fact, one can introduce other functions of square matrices as well: sines, cosines, fractional powers, and so on. For discussion of such functions we refer you to Wylie and Barrett.†

Computer software. Using *Maple*, the exponential matrix function can be evaluated by means of the **exponential** command within the linalg package. For instance, to evaluate $e^{\mathbf{A}t}$, where

$$\mathbf{A} = \begin{bmatrix} 1 & 4 \\ 1 & 1 \end{bmatrix},$$

enter

with(linalg):

and return, then

B := matrix (2, 2, [t, 4 * t, t, t]):

and return. Then,

exponential(B);

and return gives the output

$$\begin{bmatrix} \frac{1}{2}e^{(-t)} + \frac{1}{2}e^{(3t)} & e^{(3t)} - e^{(-t)} \\ \frac{1}{4}e^{(3t)} - \frac{1}{4}e^{(-t)} & \frac{1}{2}e^{(-t)} + \frac{1}{2}e^{(3t)} \end{bmatrix}$$

Actually, when you enter the $\mathbf{B}$ matrix you may prefer to end with a semicolon rather than a colon because then the $\mathbf{B}$ matrix will be printed, and you can inspect it to see if its elements were entered correctly.

---

*Of course, it is more natural to write the first term on the right-hand side of (30) as ce^{At}, but we have written it as $e^{At}c$ to emphasize the correspondence between (29) and (30).

†C. R. Wylie and L. C. Barrett, *Advanced Engineering Mathematics*, 5th ed. (New York: McGraw–Hill, 1982).

EXERCISES 11.5

1. Solve, using (12). If (12) does not apply, explain why it does not. NOTE: You may need to first re-express the problem in the standard form (1).

(a) $x' = x + y; \quad x(0) = 0$
$y' = x + y + e^{3t}; \quad y(0) = 0$

(b) $x' = 2x + 4y + 1; \quad x(0) = 0$
$y' = x - y; \quad y(0) = 0$

(c) $x' = 2x - y - 3t + 1; \quad x(0) = 1$
$y' = -x + 2y + 3; \quad y(0) = 0$

(d) $x' = x + 2y - t - 1; \quad x(0) = 0$
$y' = 4x + 8y - 4t - 8; \quad y(0) = 3$

(e) $x' = \sin t - y; \quad x(0) = 0$
$y' = -9x + 4; \quad y(0) = 1$

(f) $x' = x - 8y; \quad x(0) = 0$
$y' = -x - y - 3t^2; \quad y(0) = 0$

(g) $x' - 3x + y = 4e^t; \quad x(0) = 1$
$y' + 3x - y = 6; \quad y(0) = -1$

(h) $x' - x - y = 3t; \quad x(0) = 5$
$x' + y' - 5x - 2y = 5; \quad y(0) = 3$

(i) $x'' + 3x' + 2x = e^t; \quad x(0) = x'(0) = 0$

(j) $x'' + 2x' + x = 0; \quad x(0) = 2, \; x'(0) = 9$

(k) $x_1' = 4x_1 + x_2 + 3x_3; \quad x_1(0) = 1$
$x_2' = x_1 - x_3 + 6; \quad x_2(0) = 0$
$x_3' = 3x_1 - x_2 + 4x_3; \quad x_3(0) = 0$

(l) $x' = x + y + z - 1; \quad x(0) = 0$
$y' = x + y + z; \quad y(0) = 0$
$z' = x + y + z; \quad z(0) = 0$

2. (a)–(l) Same as Exercise 1, but drop the initial conditions and find the general solution instead.

3. Evaluate $e^{\mathbf{A}}$ for the given $\mathbf{A}$ matrix. HINT: If $\mathbf{A}$ is diagonalizable you can use the formula

$$e^{\mathbf{A}} = \mathbf{Q}e^{\mathbf{D}}\mathbf{Q}^{-1} \tag{3.1}$$

[i.e., equation (28) with $t = 1$]. Whether or not $\mathbf{A}$ is diagonalizable, you can use the series definition

$$e^{\mathbf{A}} = \mathbf{I} + \frac{1}{1!}\mathbf{A} + \frac{1}{2!}\mathbf{A}^2 + \cdots. \tag{3.2}$$

Generally, (3.2) is unwieldy because one needs to sum an infinite series of matrices. However, if $\mathbf{A}$ happens to be nilpotent then the series reduces to a finite number of terms.

(a) $\begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$

(c) $\begin{bmatrix} 0 & 4 \\ 9 & 0 \end{bmatrix}$ (d) $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$

(e) $\begin{bmatrix} 2 & 0 \\ 7 & 3 \end{bmatrix}$ (f) $\begin{bmatrix} 5 & 0 \\ 0 & 0 \end{bmatrix}$

(g) $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ (h) $\begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix}$

(i) $\begin{bmatrix} 2 & 2 & 1 \\ 1 & 3 & 1 \\ 1 & 2 & 2 \end{bmatrix}$ (j) $\begin{bmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \\ 2 & 2 & 3 \end{bmatrix}$

(k) $\begin{bmatrix} 0 & 1 & 2 & 3 \\ 0 & 0 & 4 & 5 \\ 0 & 0 & 0 & 6 \\ 0 & 0 & 0 & 0 \end{bmatrix}$ (l) $\begin{bmatrix} 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 1 & 2 & 1 & 0 \end{bmatrix}$

4. (a)–(l) Evaluate $e^{\mathbf{A}}$ using computer software, where $\mathbf{A}$ is given in the corresponding part of Exercise 3.

11.6 Quadratic Forms (Optional)

A function of the form

$$f(x_1, x_2) = a_{11}x_1^2 + a_{22}x_2^2 + 2a_{12}x_1x_2 \tag{1}$$

is called a **quadratic form** in the variables x_1 and x_2. Similarly,

$$f(x_1, x_2, x_3) = a_{11}x_1^2 + a_{22}x_2^2 + a_{33}x_3^2 + 2a_{12}x_1x_2 + 2a_{13}x_1x_3 + 2a_{23}x_2x_3 \tag{2}$$

is a quadratic form in x_1, x_2, x_3, and so on for any number of variables $x_1, \ldots, x_n$. The a_{ij}'s are constants (where a_{ij} is the coefficient of the product $x_i x_j$), and the 2's are included for our subsequent convenience.

It turns out that a quadratic form $f(x_1, \ldots, x_n)$ can be expressed concisely, in matrix notation, as

$$\boxed{f(x_1, \ldots, x_n) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x},} \tag{3}$$

where $\mathbf{x} = [x_1, \ldots, x_n]^{\mathrm{T}}$, and $\mathbf{A} = \{a_{ij}\}$ is an $n \times n$ matrix. For instance, suppose that $n = 2$ and

$$f(x_1, x_2) = 2x_1^2 - x_2^2 + 6x_1x_2. \tag{4}$$

Writing out the right-hand side of (3) gives

$$\begin{aligned}\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} &= [x_1, x_2]\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = [x_1, x_2]\begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{bmatrix} \\ &= a_{11}x_1^2 + a_{22}x_2^2 + (a_{12} + a_{21})x_1x_2. \end{aligned} \tag{5}$$

Comparing (4) and (5), we see that $a_{11} = 2, a_{22} = -1$, and $a_{12} + a_{21} = 6$. Since the sum $a_{12} + a_{21}$ is prescribed, but not a_{12} and a_{21} individually, we are free to ask a_{12} and a_{21} to be equal, so that $a_{12} = a_{21} = 3$. Then the right-hand side of (5) becomes $a_{11}x_1^2 + a_{22}x_2^2 + 2a_{12}x_1x_2$, as written in (1). The benefit in asking a_{12} to equal a_{21} is that then the $\mathbf{A}$ matrix in (3) is *symmetric*, and symmetric matrices have advantages over nonsymmetric matrices, as we have seen in Sections 11.3 and 11.4. Similarly for $n = 3, 4, \ldots$.

Thus, if we ask the $\mathbf{A}$ matrix in (3) to be symmetric, for convenience, then (3) gives (1) for $n = 2$, (2) for $n = 3$, and so on.

EXAMPLE 1. Let

$$f(x_1, \ldots, x_4) = x_2^2 + 5x_3^2 + 6x_1x_3 - x_2x_3 + 10x_3x_4.$$

Then $a_{22} = 1, a_{33} = 5, 2a_{13} = 6$ so $a_{13} = 3, 2a_{23} = -1$ so $a_{23} = -\frac{1}{2}, 2a_{34} = 10$ so $a_{34} = 5$, and all the other a_{ij}'s are zero. Thus

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 3 & 0 \\ 0 & 1 & -\frac{1}{2} & 0 \\ 3 & -\frac{1}{2} & 5 & 5 \\ 0 & 0 & 5 & 0 \end{bmatrix}. \quad \blacksquare$$

A quadratic form is said to be **canonical** if all mixed terms (such as x_1x_2, x_1x_3, and x_2x_3) are absent, that is, if $a_{ij} = 0$ for $i \neq j$. Thus,

$$f(x_1, \ldots, x_n) = a_{11}x_1^2 + a_{22}x_2^2 + \cdots + a_{nn}x_n^2 \tag{6}$$

is canonical, and its associated matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & a_{nn} \end{bmatrix}$$

is diagonal. For instance, (4) is not canonical, but $f(x_1, x_2) = 5x_1^2 - 7x_2^2$ and $f(x_1, x_2, x_3) = 6x_2^2 + x_3^2$ are.

There is interest in being able to reduce a given quadratic form to canonical form (i.e., to its simplest form) through a linear change of variables

$$\mathbf{x} = \mathbf{Q}\widetilde{\mathbf{x}} \tag{7}$$

from $x_1, \ldots, x_n$ to $\widetilde{x}_1, \ldots, \widetilde{x}_n$. Putting (7) into (3) gives

$$f = (\mathbf{Q}\widetilde{\mathbf{x}})^{\mathrm{T}} \mathbf{A} (\mathbf{Q}\widetilde{\mathbf{x}}) = \widetilde{\mathbf{x}}^{\mathrm{T}} \left(\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q}\right) \widetilde{\mathbf{x}}. \tag{8}$$

Thus, given $\mathbf{A}$, we wish to determine a $\mathbf{Q}$ matrix such that $\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q}$ is diagonal, for then (8) will be canonical in the new variables $\widetilde{x}_1, \ldots, \widetilde{x}_n$. Theorem 11.4.1 tells us that $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$ will be a diagonal matrix, with its diagonal elements being the eigenvalues of $\mathbf{A}$, if $\mathbf{A}$ has n LI eigenvectors and if these eigenvectors are used as the columns of $\mathbf{Q}$.* However, (8) contains $\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q}$, not $\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$. But since $\mathbf{A}$ is symmetric it has n orthogonal (and hence LI) eigenvectors, and if these are normalized and used as the columns of $\mathbf{Q}$ then $\mathbf{Q}^{\mathrm{T}} = \mathbf{Q}^{-1}$, and

$$\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q} = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & \lambda_n \end{bmatrix} \equiv \mathbf{D}, \tag{9}$$

where the λ_j's are the eigenvalues of $\mathbf{A}$, so (8) is the desired canonical form of f,

$$f = \lambda_1 \widetilde{x}_1^2 + \cdots + \lambda_n \widetilde{x}_n^2. \tag{10}$$

EXAMPLE 2. Reduce

$$f(x_1, x_2) = 3x_1^2 + 3x_2^2 + 2x_1x_2 \tag{11}$$

to canonical form. First, identify $\mathbf{A}$:

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}. \tag{12}$$

*We do not claim that $\mathbf{Q}$ *must* be a modal matrix of $\mathbf{A}$ for $\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q}$ to be diagonal.

Next, determine the eigenvalues and normalized eigenvectors of $\mathbf{A}$. These are found to be

$$\lambda_1 = 4,\ \hat{\mathbf{e}}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}; \qquad \lambda_2 = 2,\ \hat{\mathbf{e}}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}. \tag{13}$$

(Whether we call $\lambda_1 = 4$ and $\lambda_2 = 2$, or vice versa, will not matter.) According to (10), the desired canonical form of f is then

$$f = \lambda_1 \widetilde{x}_1^2 + \lambda_2 \widetilde{x}_2^2 = 4\widetilde{x}_1^2 + 2\widetilde{x}_2^2. \tag{14}$$

It looks like we are done, so why do we need to know $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2$? To know the connection between x_1, x_2 and $\widetilde{x}_1, \widetilde{x}_2$. Specifically,

$$\mathbf{x} = \mathbf{Q}\widetilde{\mathbf{x}} = [\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2]\widetilde{\mathbf{x}},$$

or

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} \widetilde{x}_1 \\ \widetilde{x}_2 \end{bmatrix}$$

so that

$$\begin{aligned} x_1 &= \frac{1}{\sqrt{2}}\widetilde{x}_1 + \frac{1}{\sqrt{2}}\widetilde{x}_2 \\ x_2 &= \frac{1}{\sqrt{2}}\widetilde{x}_1 - \frac{1}{\sqrt{2}}\widetilde{x}_2. \end{aligned} \tag{15}$$

Or, if we wish to have it the other way around, we can use $\widetilde{\mathbf{x}} = \mathbf{Q}^{-1}\mathbf{x} = \mathbf{Q}^{\mathrm{T}}\mathbf{x}$. ∎

Let us summarize:

THEOREM 11.6.1 *Canonical Form*
A quadratic form $f(x_1, \ldots, x_n) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ can be reduced to the canonical form $\lambda_1\widetilde{x}_1^2 + \cdots + \lambda_n\widetilde{x}_n^2$ by introducing the change of variables $\mathbf{x} = \mathbf{Q}\widetilde{\mathbf{x}}$, where the λ_j's are the eigenvalues of $\mathbf{A}$ and the columns of $\mathbf{Q}$ are the corresponding normalized eigenvectors of $\mathbf{A}$. The reverse transformation is given by $\widetilde{\mathbf{x}} = \mathbf{Q}^{\mathrm{T}}\mathbf{x}$.

A quadratic form $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ is classified as **positive definite** (i.e., "definitely positive") if $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$, and as **negative definite** if $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} < 0$ for all $\mathbf{x} \neq \mathbf{0}$. Likewise, $\mathbf{A}$ is classified as positive (negative) definite if the quadratic form $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ is positive (negative) definite.

THEOREM 11.6.2 *Definiteness*
Let $\mathbf{A}$ be symmetric. Then $\mathbf{A}$ and its quadratic form $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ are positive (negative) definite if every eigenvalue of $\mathbf{A}$ is positive (negative).

Proof: It is simplest to work with the canonical form $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} = \lambda_1\widetilde{x}_1^2 + \cdots + \lambda_n\widetilde{x}_n^2$. Since the latter is a sum of squares, it is evident that if $\lambda_1, \ldots, \lambda_n$ are all positive (negative), then $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ is positive (negative) for all $\mathbf{x} \neq \mathbf{0}$. It remains to show that $\widetilde{\mathbf{x}} \neq \mathbf{0}$ if and only if $\mathbf{x} \neq \mathbf{0}$. Since $\mathbf{x} = \mathbf{Q}\widetilde{\mathbf{x}}$ and $\widetilde{\mathbf{x}} = \mathbf{Q}^{-1}\mathbf{x} = \mathbf{Q}^{\mathrm{T}}\mathbf{x}$ (i.e., $\mathbf{Q}$ is nonsingular), $\mathbf{x} = \mathbf{0}$ implies that $\mathbf{x} = \mathbf{Q}\mathbf{0} = \mathbf{0}$, and $\mathbf{x} = \mathbf{0}$ implies that $\widetilde{\mathbf{x}} = \mathbf{Q}^{\mathrm{T}}\mathbf{0} = \mathbf{0}$. ∎

For instance, f and $\mathbf{A}$ in Example 2 are positive definite because $\lambda_1 = 4 > 0$ and $\lambda_2 = 2 > 0$.

EXAMPLE 3. If

$$f(x_1, x_2, x_3) = 2x_1^2 + 4x_2^2 + 4x_3^2 + 2x_1x_2 - 2x_1x_3 + 6x_2x_3, \tag{16}$$

then

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & -1 \\ 1 & 4 & 3 \\ -1 & 3 & 4 \end{bmatrix} \quad \text{and} \quad \lambda_1 = 7,\ \lambda_2 = 3,\ \lambda_3 = 0. \tag{17}$$

It might appear that f is positive definite but it is not, because λ_3 is not positive. Being zero is not good enough, for remember that being positive definite means that $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$ or, equivalently, $\widetilde{\mathbf{x}}^{\mathrm{T}}(\mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q})\widetilde{\mathbf{x}} > 0$ for all $\mathbf{x} \neq \mathbf{0}$. Yet,

$$f = 7\widetilde{x}_1^2 + 3\widetilde{x}_2^2 + 0\widetilde{x}_3^2 \tag{18}$$

is zero if

$$\widetilde{\mathbf{x}} = \begin{bmatrix} \widetilde{x}_1 \\ \widetilde{x}_2 \\ \widetilde{x}_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 8 \end{bmatrix}, \tag{19}$$

for instance. ∎

Let us conclude this section with a physical application.

EXAMPLE 4. *Buckling Load.* Three rigid rods, each of length L, are pinned at their ends, with the end A constrained to move in a frictionless vertical slot, as shown in Fig. 1. The three lateral springs, each of stiffness k, are unstretched and uncompressed when the system is undeflected (i.e., when $x = y = 0$), and the middle spring is attached at the middle of the middle rod. We find that as the load P is slowly increased the deflection remains zero ($x = y = z = 0$) so that $x = y = z = 0$ constitutes an equilibrium state of the system. Eventually, however, when P reaches a critical value, say P_{cr}, the system "buckles," that is, collapses. The problem posed here is to determine P_{cr} in terms of the given quantities. It will be convenient to work with the potential energy V of the system, and to use the physical principle that the system will arrange itself so as to minimize its potential V.

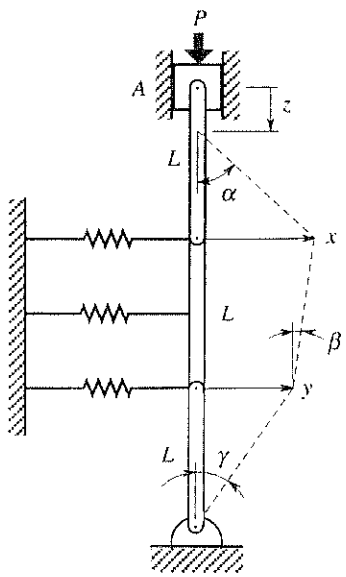


Figure 1. Buckling.

Recall, from elementary physics, that the potential energy stored in a spring of stiffness k, deflected by an amount x, is $\frac{1}{2}kx^2$. Further, note that we may associate a potential

with the load P by regarding it as being due to a weight P sitting on the slider. Then the gravitational potential of the weight is $-Pz$, if we define the potential to be zero when $z = 0$. Thus,

$$V = \frac{1}{2}kx^2 + \frac{1}{2}ky^2 + \frac{1}{2}k\left(\frac{x+y}{2}\right)^2 - Pz. \tag{20}$$

But x, y, z are not independent variables. For instance, we can express z in terms of x and y through the geometry. Specifically, if we introduce the angles α, β, γ, shown in Fig. 1, then

$$z + L\cos\alpha + L\cos\beta + L\cos\gamma = 3L. \tag{21}$$

For the onset of the buckling x, y, z are all infinitesimal (arbitrarily small), so

$$\begin{aligned}\cos\alpha &= \sqrt{1-\sin^2\alpha} = \sqrt{1-\left(\frac{x}{L}\right)^2} \\ &= 1 - \frac{1}{2}\left(\frac{x}{L}\right)^2 - \frac{1}{8}\left(\frac{x}{L}\right)^4 - \cdots \quad \text{(Taylor series)} \\ &\sim 1 - \frac{1}{2}\left(\frac{x}{L}\right)^2 \quad \text{as } x \to 0.\end{aligned} \tag{22}$$

Similarly for $\cos\beta$ and $\cos\gamma$ so (21) becomes

$$z + L\left[1 - \frac{1}{2}\left(\frac{x}{L}\right)^2\right] + L\left[1 - \frac{1}{2}\left(\frac{x-y}{L}\right)^2\right] + L\left[1 - \frac{1}{2}\left(\frac{y}{L}\right)^2\right] \sim 3L \tag{23}$$

or, upon simplification,

$$z \sim \frac{x^2 - xy + y^2}{L} \tag{24}$$

so that

$$V(x,y) \sim \left(\frac{5k}{8} - \frac{P}{L}\right)x^2 + \left(\frac{5k}{8} - \frac{P}{L}\right)y^2 + \left(\frac{k}{4} + \frac{P}{L}\right)xy. \tag{25}$$

The right-hand side of (25) is a quadratic form

$$f(x,y) = \left(\frac{5k}{8} - \frac{P}{L}\right)x^2 + \left(\frac{5k}{8} - \frac{P}{L}\right)y^2 + \left(\frac{k}{4} + \frac{P}{L}\right)xy, \tag{26}$$

with the associated matrix

$$\mathbf{A} = \begin{bmatrix} \frac{5k}{8} - \frac{P}{L} & \frac{k}{8} + \frac{P}{2L} \\ \frac{k}{8} + \frac{P}{2L} & \frac{5k}{8} - \frac{P}{L} \end{bmatrix}. \tag{27}$$

The crucial point is whether or not f is positive definite for if $f(x,y)$ is positive definite, then $V(x,y)$ has a minimum at $x = y = 0$ and the undeflected equilibrium configuration $x = y = 0$ is stable. Otherwise, the equilibrium configuration will be unstable and the system will buckle.

To assess the positive definiteness of f, we need merely evaluate the eigenvalues of $\mathbf{A}$. These are found to be

$$\lambda_1 = \frac{3k}{2}\left(\frac{1}{3} - \frac{P}{kL}\right), \qquad \lambda_2 = \frac{k}{2}\left(\frac{3}{2} - \frac{P}{kL}\right) \tag{28}$$

from which we see that f is positive definite (i.e., both λ_j's are positive) if and only if $P/(kL) < \frac{1}{3}$. Thus, the critical load, the "buckling load," is $P_{cr} = kL/3$.

COMMENT 1. If $P/(kL) < \frac{1}{3}$, then both λ_j's are positive in the canonical version $\lambda_1\tilde{x}^2 + \lambda_2\tilde{y}^2$ of V so the graph of V has a "valley" at the origin ($\tilde{x} = \tilde{y} = 0$ or, equivalently $x = y = 0$). If $P/(kL) > \frac{3}{2}$, then both λ_j's are negative so the graph of V has a "hill" at the origin. And if $\frac{1}{3} < P/(kL) < \frac{3}{2}$, then $\lambda_1 < 0$ and $\lambda_2 > 0$ so the graph of V has a "saddle" at the origin (i.e., like an equestrian saddle it has a valley in one direction and a hill in the other). In the case of the valley V has a *minimum* at the origin so the equilibrium solution $x = y = 0$ is *stable* and buckling does not occur. In the other two cases (hill and saddle) V does not have a minimum at the origin, so the equilibrium solution $x = y = 0$ is unstable and buckling occurs. The borderline between these two cases gives the critical buckling load.

COMMENT 2. Physically, the idea is that as the system begins to deflect, the springs gain potential energy while the weight P loses potential energy. For instability (buckling) we need the loss to exceed the gain. Since the loss is proportional to P, one anticipates the existence of a critical value P_{cr} such that if $P = P_{cr}$, the loss just balances the gain, and if $P > P_{cr}$, then the loss exceeds the gain. ∎

Closure. The chief result of this section is given in Theorem 11.6.1, that a quadratic form can be reduced to canonical form by a normalized modal matrix. In addition, Theorem 11.6.2 tells us that a quadratic form $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ is positive definite if every eigenvalue of $\mathbf{A}$ is positive, and negative definite if every eigenvalue is negative. Positive and negative definiteness is central when we discuss the maxima and minima of functions of several variables in Chapter 13.

EXERCISES 11.6

1. For each of the following quadratic forms in n variables determine the associated symmetric $\mathbf{A}$ matrix.

(a) $2x_1^2 + 4x_2^2 + x_1x_2 \quad (n = 2)$
(b) $x_3^2 + 2x_1x_3 + 2x_2x_3 \quad (n = 3)$
(c) $x_1^2 + x_2^2 + x_1x_2 \quad (n = 2)$
(d) $x_1^2 - 4x_3^2 + 3x_1x_2 \quad (n = 3)$
(e) $x_1^2 + x_4^2 + 2x_1x_4 \quad (n = 4)$
(f) $4x_1x_4 + 4x_2x_3 \quad (n = 4)$
(g) $3x_1x_2 \quad (n = 2)$
(h) $4x_1x_2 - 2x_3^2 \quad (n = 3)$
(i) $-2x_1x_3 \quad (n = 3)$
(j) $x_3^2 + x_1x_2 + x_1x_3 + x_2x_3 \quad (n = 3)$
(k) $6x_2x_3 \quad (n = 3)$
(l) $2x_1x_2 + 2x_1x_3 + 2x_2x_3 - 4x_4^2 \quad (n = 4)$

2. (a)–(l) Reduce each quadratic form in Exercise 1 to the canonical form (10) by means of a normalized modal matrix transformation $\mathbf{Q}$. Further, classify the quadratic form and the associated symmetric $\mathbf{A}$ matrix as positive definite or negative definite, where applicable

3. We state, without proof, that a *necessary and sufficient con-*

dition for a quadratic form $f = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$, and its associated symmetric matrix $\mathbf{A}$, to be positive definite is that

$$a_{11} > 0, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0, \quad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} > 0, \quad \ldots, \quad \det\mathbf{A} > 0.$$

Apply this condition to each of the following $\mathbf{A}$ matrices, and compare your conclusion with that drawn from direct examination of the eigenvalues.

(a) $\begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

(c) $\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$ (d) $\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$

(e) $\begin{bmatrix} 6 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$ (f) $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 2 & 0 \\ 1 & 0 & 0 \end{bmatrix}$

(g) $\begin{bmatrix} -2 & 1 & 2 \\ 1 & -2 & 2 \\ 2 & 2 & 1 \end{bmatrix}$ (h) $\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 9 \end{bmatrix}$

4. Reduce the following quadratic equations to the canonical form $a\widetilde{x}^2 + b\widetilde{y}^2 = c$, and hence decide whether they correspond to *ellipses* or *hyperbolas*. Sketch their graphs, showing both the x, y and $\widetilde{x}, \widetilde{y}$ axes, as well as the intercepts on the $\widetilde{x}, \widetilde{y}$ axes. HINT: Recall that $(\widetilde{x}/A)^2 + (\widetilde{y}/B)^2 = 1$ is the equation of an ellipse with intercepts at $\pm A$ on the $\widetilde{x}$ axis and $\pm B$ on the $\widetilde{y}$ axis (or a circle of radius A if $A = B$). Further, $(\widetilde{x}/A)^2 - (\widetilde{y}/B)^2 = 1$ is the equation of a hyperbola with intercepts at $\pm A$ on the $\widetilde{x}$ axis, and $(\widetilde{y}/B)^2 - (\widetilde{x}/A)^2 = 1$ is the equation of a hyperbola with intercepts at $\pm B$ on the $\widetilde{y}$ axis.

(a) $3x^2 + 2y^2 - 2xy = 6$
(b) $xy = 1$
(c) $4y^2 + 3xy = 1$
(d) $x^2 + y^2 - 10xy = 4$
(e) $x^2 + y^2 + 2xy = 4$ (This case will be found to correspond to a limiting case. Explain what we mean by that.)

5. (*Completing squares*) The successful $\mathbf{Q}$ matrix, in (7), is not necessarily a modal matrix of $\mathbf{A}$. Given $f = x_1^2 + x_2^2 + x_1x_2$, for example, suppose that we proceed, instead, by "completing squares:"

$$\begin{aligned} f &= x_1^2 + x_2^2 + x_1x_2 \\ &= \left(x_1^2 + x_1x_2 + \tfrac{1}{4}x_2^2\right) + x_2^2 - \tfrac{1}{4}x_2^2 \\ &= \left(x_1 + \tfrac{1}{2}x_2\right)^2 + \tfrac{3}{4}x_2^2, \end{aligned} \tag{5.1}$$

so the transformation

$$\widetilde{x}_1 = x_1 + \tfrac{1}{2}x_2, \quad \widetilde{x}_2 = x_2 \tag{5.2}$$

reduces f to the canonical form $f = \widetilde{x}_1^2 + \frac{3}{4}\widetilde{x}_2^2$.

(a) Show that the matrix $\mathbf{Q}$ corresponding to (5.2) is *not* a modal matrix.
(b) Reduce $f = x_1^2 + x_2^2 + 4x_3^2 + 2x_1x_2 - x_1x_3$ to canonical form by completing squares.
(c) Repeat part (b), for $f = x_1^2 + 4x_1x_2 + x_2x_3$.
(d) Repeat part (b), for $f = 4x_1^2 + 2x_1x_3 + x_2x_3$.
(e) Repeat part (b), for $f = x_3^2 + 2x_1x_3$.
(f) Does the method of completing squares work for $f = 2x_1x_2$? Explain.

6. Rework Example 4 for the case where the middle spring is removed. Show that the buckling load is then $P_{cr} = \frac{1}{3}kL$. Note that this result is the same as in Example 4, where the middle spring is included. Explain, in physical terms, why this is so.

Chapter 11 Review

The matrix eigenvalue problem is the search for nontrivial solutions of $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ or, equivalently,

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}; \tag{1}$$

that is, solutions other than the trivial solution $\mathbf{x} = \mathbf{0}$. The eigenvalues are found

by setting

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0, \tag{2}$$

which condition guarantees the existence of nontrivial solutions of (1). Known as the characteristic equation of $\mathbf{A}$, (2) is an nth-degree polynomial equation, which always has at least one and at most n distinct roots. For each eigenvalue λ_j thus found, the solution of $(\mathbf{A} - \lambda_j\mathbf{I})\mathbf{x} = \mathbf{0}$ by Gauss elimination gives the corresponding eigenvectors $\mathbf{e}_j$.

Rather than being rare, the case of symmetric matrices is common in applications (as noted, for instance, in Example 2 of Section 11.3, Example 1 in Section 11.4, and in all of Section 11.6). If an $n \times n$ matrix $\mathbf{A}$ is symmetric, then all of its eigenvalues are real, eigenvectors corresponding to distinct eigenvalues are orthogonal, and its eigenvectors provide an orthogonal basis for n-space.

A pattern emerges insofar as choice of basis. Namely, when a basis is needed, to expand vectors, the most convenient basis to use is probably the basis provided by the eigenvectors of the $\mathbf{A}$ matrix to be found within the given problem. For instance, we do that to study the stability of the equilibrium solution of the Markov process (Example 4, Section 11.2), to solve the nonhomogeneous equation $\mathbf{Ax} = \Lambda\mathbf{x} + \mathbf{c}$ (Section 11.3.2), and to prove the convergence of the power method (Exercise 12, Section 11.3). The reason that eigenvector bases are convenient is that if we multiply a vector equation by $\mathbf{A}$, then we need to evaluate the vectors $\mathbf{Ae}_j$. If the $\mathbf{e}_j$'s are eigenvectors of $\mathbf{A}$, then $\mathbf{Ae}_j$ is simply the single term $\lambda_j\mathbf{e}_j$; if not, it is a linear combination of the n base vectors, $\mathbf{e}_1, \ldots, \mathbf{e}_n$.

In Section 11.4 we study the diagonalization of an $n \times n$ matrix $\mathbf{A}$. There, Theorem 11.4.1 is the most significant result because it gives a necessary and sufficient condition for $\mathbf{A}$ to be diagonalizable (namely, that it have n LI eigenvectors), and it tells us how to choose $\mathbf{Q}$ so that

$$\mathbf{Q}^{-1}\mathbf{AQ} = \mathbf{D} \tag{3}$$

is diagonal. Namely, if we use the (n LI) eigenvectors of $\mathbf{A}$ as the columns of $\mathbf{Q}$, then $\mathbf{D} = \{d_{ij}\}$ is diagonal, with $d_{jj} = \lambda_j$. Further, Theorem 11.4.3 gives a sufficient condition for diagonalizability, that $\mathbf{A}$ have n distinct eigenvalues. Since the generic case is for the characteristic equation to have distinct roots, the generic case is for a given $n \times n$ matrix to be diagonalizable. If $\mathbf{A}$ is symmetric, then it is diagonalizable whether or *not* it has n distinct eigenvalues because every $n \times n$ symmetric matrix has n orthogonal (and hence LI) eigenvectors. For symmetric matrices we urge you to use the *normalized* eigenvectors of $\mathbf{A}$ to form its modal matrix $\mathbf{Q}$ because then $\mathbf{Q}$ admits the useful property that

$$\mathbf{Q}^{-1} = \mathbf{Q}^{\mathrm{T}}, \tag{4}$$

that is, the inverse of $\mathbf{Q}$ is simply its transpose.

Even if $\mathbf{A}$ is not diagonalizable, it can be reduced to Jordan normal form. That is, a generalized modal matrix $\mathbf{P}$ can be found so that

$$\mathbf{P}^{-1}\mathbf{AP} = \mathbf{J} \tag{5}$$

is triangular; $\mathbf{J}$ is upper triangular with zeros above its main diagonal except for 1's immediately above one or more diagonal elements. That case is left for the exercises.

Whereas in Section 11.4 we concentrate on the concept of diagonalization, and show how to use diagonalization to uncouple systems of differential equations, in Section 11.5 we shift our focus to differential equation theory itself, using matrix theory and diagonalization as tools, and solve the general system of n first-order nonhomogeneous linear differential equations with constant coefficients.

In the final section, 11.6, we show that a quadratic form can be expressed in matrix terminology as $\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is symmetric, and can be reduced to canonical form $\lambda_1\widetilde{x}_1^2 + \cdots + \lambda_n\widetilde{x}_n^2$ by the change of variables $\mathbf{x} = \mathbf{Q}\widetilde{\mathbf{x}}$, where $\mathbf{Q}$ is a normalized modal matrix of $\mathbf{A}$. That is, $\mathbf{Q} = [\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_n]$, where $\hat{\mathbf{e}}_j$ is a normalized eigenvector corresponding to the eigenvalue λ_j.

Applications of the eigenvalue problem are extensive. Examples studied in this chapter are drawn from the areas of oscillation theory, systems of coupled differential equations, and population dynamics, with other applications covered in the exercises. In Chapter 13 we use the theory of quadratic forms to help us classify the extrema of functions of several variables.

Chapter 12

Extension to Complex Case (Optional)

PREREQUISITE: Familiarity with the algebra of complex numbers, covered in Section 21.2.

12.1 Introduction

In Chapters 8–11 all scalars are understood to be real: the coefficients in systems of linear equations, the components of vectors, the elements of matrices, and so on. In some applications, however, perhaps more so in physics and chemistry than in engineering, complex numbers enter. For example, you may have met the *Euler angles* θ, ϕ, ψ used to specify the orientation of a rigid body such as a gyroscope. Alternatively, it is sometimes advantageous to employ so-called *Cayley–Klein parameters*, and in doing so one meets the *Pauli spin matrices*

$$\sigma_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \sigma_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad \sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \tag{1}$$

of which σ_y is seen to have complex elements.

The purpose of this chapter is to indicate the changes that need to be made in extending our vector and matrix systems so as to include complex numbers. Although this could have been done from the start, it was felt that the gain in simplicity, in the preceding chapters, offsets the need for this special chapter, which we make quite brief.

12.2 Complex n-Space

All scalars in the vector space $\mathbb{R}^n$ defined in Section 9.4 (namely, the scalars that multiplied vectors and the scalar components of the vectors themselves) are real.

If we allow these scalars to be complex, then in place of $\mathbb{R}^n$ we have the **complex n-space** denoted here as $\mathbb{C}^n$:

$$\mathbb{C}^n = \{(a_1, \ldots, a_n) \mid a_1, \ldots, a_n \text{ complex numbers }\}. \tag{1}$$

The definitions $\mathbf{u} + \mathbf{v} \equiv (u_1 + v_1, \ldots, u_n + v_n)$, $\alpha\mathbf{u} \equiv (\alpha u_1, \ldots, \alpha u_n)$, $\mathbf{0} \equiv (0, \ldots, 0)$, and $-\mathbf{u} \equiv (-u_1, \ldots, -u_n)$ are the same as for $\mathbb{R}^n$, except that now the scalars are complex numbers. From these definitions the same properties follow as for $\mathbb{R}^n$ ($\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$, etc.) as in (10) in Section 9.4.

However, the Euclidean dot product $\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + \cdots + u_n v_n$, which we adopted for real vector space, is unacceptable because the resulting norm $\|\mathbf{u}\| = \sqrt{\mathbf{u} \cdot \mathbf{u}} = \sqrt{u_1^2 + \cdots + u_n^2}$ fails to display the key properties expected of a norm, in particular, the nonnegativeness condition

$$\begin{aligned} \|\mathbf{u}\| &> 0 \quad \text{for all } \mathbf{u} \neq \mathbf{0}, \\ &= 0 \quad \text{for } \mathbf{u} = \mathbf{0}. \end{aligned} \tag{2}$$

For example, if $\mathbf{u} = [2, 2i, 0, 0]$, then $\|\mathbf{u}\| = \sqrt{(2)^2 + (2i)^2 + (0)^2 + (0)^2} = 0$ even though $\mathbf{u} \neq \mathbf{0}$; and if $\mathbf{u} = [0, 2i, 0, 0]$, then $\|\mathbf{u}\| = \sqrt{0 - 4 + 0 + 0} = 2i$ is not even real so it cannot satisfy the condition $\|\mathbf{u}\| > 0$.*

To avoid this problem with the norm, we adopt the modified dot product

$$\boxed{\mathbf{u} \cdot \mathbf{v} \equiv u_1 \overline{v}_1 + u_2 \overline{v}_2 + \cdots + u_n \overline{v}_n = \sum_{j=1}^{n} u_j \overline{v}_j} \tag{3}$$

where the overhead bar denotes complex conjugate for then

$$\boxed{\|\mathbf{u}\| \equiv \sqrt{\mathbf{u} \cdot \mathbf{u}} = \sqrt{\sum_{j=1}^{n} u_j \overline{u}_j} = \sqrt{\sum_{j=1}^{n} |u_j|^2}} \tag{4}$$

does satisfy the nonnegativeness condition (2). [Recall that if $z = a + ib$, then $z\overline{z} = (a + ib)(a - ib) = a^2 + b^2 = |z|^2$.]

EXAMPLE 1. If $\mathbf{u} = [2, 3 - 5i, 0, 4i]$, then

$$\begin{aligned} \|\mathbf{u}\| &= \sqrt{(2)(2) + (3 - 5i)(3 + 5i) + (0)(0) + (4i)(-4i)} \\ &= \sqrt{4 + 34 + 16} = \sqrt{54}, \end{aligned}$$

*Recall from Section 22.2 (which is the prerequisite for this chapter) that inequalities such as $z > 0$ and $z < 0$ are not meaningful if z is complex; see the paragraph below equation (12) in that section.

which *is* real and positive. ∎

Properties of the dot product. Observe that

$$\begin{aligned}\overline{\mathbf{v}\cdot\mathbf{u}} &= \overline{v_1\overline{u}_1 + \cdots + v_n\overline{u}_n}\\ &= \overline{v_1\overline{u}_1} + \cdots + \overline{v_n\overline{u}_n}\\ &= \overline{v}_1\overline{\overline{u}}_1 + \cdots + \overline{v}_n\overline{\overline{u}}_n\\ &= \overline{v}_1 u_1 + \cdots + \overline{v}_n u_n\\ &= u_1\overline{v}_1 + \cdots + u_n\overline{v}_n\\ &= \mathbf{u}\cdot\mathbf{v}.\end{aligned}$$

Thus, in place of the commutativity condition $\mathbf{u}\cdot\mathbf{v} = \mathbf{v}\cdot\mathbf{u}$, satisfied in the real case, we have the so-called conjugate commutativity $\mathbf{u}\cdot\mathbf{v} = \overline{\mathbf{v}\cdot\mathbf{u}}$ in the complex case. In fact, the properties

$$\begin{aligned}&\textit{Conjugate Commutative}: & \mathbf{u}\cdot\mathbf{v} &= \overline{\mathbf{v}\cdot\mathbf{u}}, & &(5a)\\ &\textit{Nonnegative}: & \mathbf{u}\cdot\mathbf{u} &> 0 & &\text{for all } \mathbf{u}\neq\mathbf{0},\\ & & &= 0 & &\text{for } \mathbf{u}=\mathbf{0}, & &(5b)\\ &\textit{Linear}: & (\alpha\mathbf{u}+\beta\mathbf{v})\cdot\mathbf{w} &= \alpha(\mathbf{u}\cdot\mathbf{w}) + \beta(\mathbf{v}\cdot\mathbf{w}), & &(5c)\end{aligned}$$

of the complex dot product are the same as in the real case [(12) in Section 9.5.2] except for the complex conjugate bar in (5a).

EXAMPLE 2. If $\mathbf{u} = [1+2i, -4]$ and $\mathbf{v} = [i, 3-i]$, then $\mathbf{u}\cdot\mathbf{v} = (1+2i)(-i) + (-4)(3+i) = -10-5i$, and $\mathbf{v}\cdot\mathbf{u} = (i)(1-2i) + (3-i)(-4) = -10+5i$, which does equal $\overline{\mathbf{u}\cdot\mathbf{v}}$, in accordance with (5a). ∎

The **Schwarz inequality**,

$$|\mathbf{u}\cdot\mathbf{v}| \le \|\mathbf{u}\|\,\|\mathbf{v}\|, \tag{6}$$

is found to hold (Exercise 7), as in the real case, except that here $|\mathbf{u}\cdot\mathbf{v}|$ is the modulus of a complex number rather than the absolute value of a real number.

Properties of the norm. The norm (4) admits the properties

$$\begin{aligned}&\textit{Scaling}: & \|\alpha\mathbf{u}\| &= |\alpha|\,\|\mathbf{u}\|, & &(7a)\\ &\textit{Nonnegative}: & \|\mathbf{u}\| &> 0 & &\text{for all } \mathbf{u}\neq\mathbf{0},\\ & & &= 0 & &\text{for } \mathbf{u}=\mathbf{0}, & &(7b)\\ &\textit{Triangle Inequality}: & \|\mathbf{u}+\mathbf{v}\| &\le \|\mathbf{u}\| + \|\mathbf{v}\|. & &(7c)\end{aligned}$$

These properties are identical to those for the real case [(17) in Section 9.5.3] but here $|\alpha|$ is the modulus of a complex number rather than the absolute value of a real number. Also, the proofs are slightly different due to the presence of complex numbers. To illustrate, consider (7c):

$$\begin{aligned}
\|\mathbf{u}+\mathbf{v}\|^2 &= (\mathbf{u}+\mathbf{v})\cdot(\mathbf{u}+\mathbf{v}) \\
&= \mathbf{u}\cdot\mathbf{u}+\mathbf{u}\cdot\mathbf{v}+\mathbf{v}\cdot\mathbf{u}+\mathbf{v}\cdot\mathbf{v} \\
&= \|\mathbf{u}\|^2+\mathbf{u}\cdot\mathbf{v}+\overline{\mathbf{u}\cdot\mathbf{v}}+\|\mathbf{v}\|^2 \\
&= \|\mathbf{u}\|^2+2\,\mathrm{Re}(\mathbf{u}\cdot\mathbf{v})+\|\mathbf{v}\|^2, \\
\|\mathbf{u}+\mathbf{v}\|^2 &\le \|\mathbf{u}\|^2+2\,|\mathrm{Re}(\mathbf{u}\cdot\mathbf{v})|+\|\mathbf{v}\|^2, \\
\|\mathbf{u}+\mathbf{v}\|^2 &\le \|\mathbf{u}\|^2+2\,|\mathbf{u}\cdot\mathbf{v}|+\|\mathbf{v}\|^2, \\
\|\mathbf{u}+\mathbf{v}\|^2 &\le \|\mathbf{u}\|^2+2\,\|\mathbf{u}\|\,\|\mathbf{v}\|+\|\mathbf{v}\|^2 \\
&= (\|\mathbf{u}\|+\|\mathbf{v}\|)^2
\end{aligned}$$

so that

$$\|\mathbf{u}+\mathbf{v}\| \le \|\mathbf{u}\|+\|\mathbf{v}\|.$$

Proofs of (7a) and (7b) are left for the exercises.

Thus far, then, the only substantial change (beyond the fact that the scalars are now complex) is the change in the dot product from $\mathbf{u}\cdot\mathbf{v} \equiv u_1v_1+\cdots+u_nv_n$ to $\mathbf{u}\cdot\mathbf{v} \equiv u_1\overline{v}_1+\cdots+u_n\overline{v}_n$. Furthermore, whereas we defined the angle θ between $\mathbf{u}$ and $\mathbf{v}$ as $\theta = \cos^{-1}(\mathbf{u}\cdot\mathbf{v}/\|\mathbf{u}\|\,\|\mathbf{v}\|)$ in the real case, this definition is awkward in the complex case since $\mathbf{u}\cdot\mathbf{v}$ (and hence the argument of the arccosine) is complex. Instead of trying to patch things up, we simply choose not to define θ for the complex case, although we do retain a notion of perpendicularity; that is, we still say that $\mathbf{u}$ and $\mathbf{v}$ are orthogonal if $\mathbf{u}\cdot\mathbf{v}=0$.

Finally,

THEOREM 12.2.1 *Dimension of* $\mathbb{C}^n$
The dimension of $\mathbb{C}^n$ is n.

You may have expected the dimension to be $2n$, on the grounds that each of the n vector components has both real and imaginary parts. However, observe that the ON set

$$\begin{aligned}
\hat{\mathbf{e}}_1 &= [1,0,\ldots,0], \\
\hat{\mathbf{e}}_2 &= [0,1,0,\ldots,0], \\
&\vdots \\
\hat{\mathbf{e}}_n &= [0,\ldots,0,1]
\end{aligned} \tag{8}$$

is a basis for $\mathbb{C}^n$, just as it is for $\mathbb{R}^n$ since the set is LI and spans the space. That it spans $\mathbb{C}^n$ follows from the fact that every vector $\mathbf{u}=(u_1,\ldots,u_n)$ in $\mathbb{C}^n$ can be

expanded as $\mathbf{u} = u_1\hat{\mathbf{e}}_1 + \cdots + u_n\hat{\mathbf{e}}_n$. And since the basis (8) contains n vectors, it follows from the definition of dimension in Section 9.9.2 that $\mathbb{C}^n$ is n-dimensional.

EXERCISES 12.2

1. Normalize each of the following vectors

(a) $[1, i]$ (b) $[1+i, 1-i]$
(c) $[1, 3, -2, 0]$ (d) $[2, 1-3i, 0, 5]$
(e) $[2+3i, 1-i, 4i]$ (f) $[i, 0, 0, -i]$
(g) $[x, y, z, ict]$ This vector (in which x, y, z are Cartesian coordinates, c is the speed of light, and t is the time) arises in the *special theory of relativity*. In that application, the space $\mathbb{C}^4$ is known as *world space*, or *Minkowski space*.

2. Show whether the following vector sets are bases for $\mathbb{C}^3$.

(a) $[i, 2, 1+i]$, $[0, 1, 2+i]$, $[4, 1, -i]$
(b) $[1, 0, 2]$, $[3, 2, -2]$, $[0, 0, 4]$
(c) $[4, 1-2i, 0]$, $[3+i, i, -2i]$, $[-2-2i, 1-4i, 4i]$
(d) $[1, 0, 0]$, $[0, 1, 0]$, $[0, 0, i]$
(e) $[i, 1, 2]$, $[3, 1-i, -i]$, $[3+2i, 3-i, 4-i]$, $[3-i, -i, -2-i]$

3. Show that the set $\mathbf{e}_1 = [i, 1, 0]$, $\mathbf{e}_2 = [2, 2i, 1]$, $\mathbf{e}_3 = [1, i, -4]$ is an orthogonal basis for $\mathbb{C}^3$.

4. Show that (23) in Section 9.9 (with $k = n$) is valid for $\mathbb{C}^n$, and use it to expand each vector in terms of the bases given in Exercise 3.

(a) $[2+4i, -3i, 2]$ (b) $[0, 0, 1]$
(c) $[1, 1, 1]$ (d) $[i, 2i, 3i]$
(e) $[0, i, 0]$ (f) $[1-i, 0, 0]$

5. Show that

$$(\alpha\mathbf{x}) \cdot \mathbf{y} = \alpha(\mathbf{x} \cdot \mathbf{y}), \tag{5.1}$$

whereas

$$\mathbf{x} \cdot (\alpha\mathbf{y}) = \overline{\alpha}(\mathbf{x} \cdot \mathbf{y}). \tag{5.2}$$

6. The property (5a) was proved in the text. Prove (5b) and (5c).

7. Prove the Schwarz inequality (6).

8. Prove the properties (7a) and (7b).

9. Vectors in $\mathbb{R}^1$, $\mathbb{R}^2$, and $\mathbb{R}^3$ can be displayed, graphically, as arrow vectors. Is the same true for $\mathbb{C}^1$, $\mathbb{C}^2$, $\mathbb{C}^3$? Explain.

12.3 Complex Matrices

All of Chapter 10 (on matrices, determinants, and linear equations) holds even if the scalars are allowed to be complex. To illustrate, consider a representative example.

EXAMPLE 1. Given

$$\mathbf{A} = \begin{bmatrix} 2 & 1+i \\ 0 & i \end{bmatrix}, \tag{1}$$

compute $\mathbf{A}^{-1}$. We proceed as usual, although the numbers are now complex:

$$\det\mathbf{A} = (2)(i) - (0)(1+i) = 2i \quad (\neq 0, \text{ so } \mathbf{A}^{-1} \text{ exists})$$

$$\begin{aligned} \text{Minors: } M_{11} &= i & \text{Cofactors: } A_{11} &= i \\ M_{12} &= 0 & A_{12} &= 0 \\ M_{21} &= 1+i & A_{21} &= -1-i \\ M_{22} &= 2 & A_{22} &= 2 \end{aligned}$$

so

$$\mathbf{A}^{-1} = \frac{1}{2i}\begin{bmatrix} i & -1-i \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{-1+i}{2} \\ 0 & -i \end{bmatrix}, \tag{2}$$

as is verified by showing that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. ∎

However, in Chapter 11, on the eigenvalue problem, dot products begin to appear so the impact of the change in the dot product (from $\mathbf{x}\cdot\mathbf{y} = x_1y_1 + \cdots + x_ny_n$ to $\mathbf{x}\cdot\mathbf{y} = x_1\overline{y}_1 + \cdots + x_n\overline{y}_n$) begins to be felt. Expressed in matrix form, the dot product is now

$$\boxed{\mathbf{x}\cdot\mathbf{y} = \mathbf{x}^{\mathrm{T}}\overline{\mathbf{y}}} \quad \text{or} \quad \overline{\mathbf{y}}^{\mathrm{T}}\mathbf{x}. \tag{3}$$

We begin with two definitions. Given an $m \times n$ matrix $\mathbf{A} = \{a_{ij}\}$, we define the **complex conjugate** of $\mathbf{A}$ as

$$\overline{\mathbf{A}} \equiv \{\overline{a}_{ij}\} \tag{4}$$

and the **Hermitian conjugate** of $\mathbf{A}$ as

$$\mathbf{A}^* \equiv \{\overline{a}_{ji}\}, \quad \text{i.e.,} \quad \boxed{\mathbf{A}^* \equiv \overline{\mathbf{A}}^{\mathrm{T}}.} \tag{5}$$

If $\overline{\mathbf{A}} = \mathbf{A}$ then $\mathbf{A}$ is **real**, and if $\mathbf{A}^* = \mathbf{A}$ then $\mathbf{A}$ is **Hermitian**.† If $\mathbf{A}$ is not square then it cannot be Hermitian.

EXAMPLE 2. If

$$\mathbf{A} = \begin{bmatrix} 2+i & 0 & 3-5i \\ 7 & 1 & 4i \\ 2 & i & 3 \end{bmatrix},$$

then

$$\overline{\mathbf{A}} = \begin{bmatrix} 2-i & 0 & 3+5i \\ 7 & 1 & -4i \\ 2 & -i & 3 \end{bmatrix} \quad \text{and} \quad \mathbf{A}^* = \begin{bmatrix} 2-i & 7 & 2 \\ 0 & 1 & -i \\ 3+5i & -4i & 3 \end{bmatrix}.$$

Since $\mathbf{A}^* \neq \mathbf{A}$, $\mathbf{A}$ is not Hermitian. ∎

EXAMPLE 3. If

$$\mathbf{A} = \begin{bmatrix} 3 & 1+4i \\ 1-4i & 0 \end{bmatrix},$$

†*Charles Hermite* (1822–1901), a professor at the Sorbonne and at the Ecole Polytechnique, contributed to the theory of elliptic functions and is also well known for his introduction of the Hermite polynomials.

then

$$\overline{\mathbf{A}} = \begin{bmatrix} 3 & 1-4i \\ 1+4i & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{A}^* = \begin{bmatrix} 3 & 1+4i \\ 1-4i & 0 \end{bmatrix}.$$

Since $\mathbf{A}^* = \mathbf{A}$, $\mathbf{A}$ is Hermitian. ∎

Some properties of the complex conjugate and Hermitian conjugate matrices are as follows:

$$\overline{\overline{\mathbf{A}}} = \mathbf{A}, \tag{6a}$$

$$\overline{\mathbf{A}+\mathbf{B}} = \overline{\mathbf{A}} + \overline{\mathbf{B}}, \tag{6b}$$

$$\overline{\mathbf{A}\mathbf{B}} = \overline{\mathbf{A}}\,\overline{\mathbf{B}}, \tag{6c}$$

$$(\mathbf{A}^*)^* = \mathbf{A}, \tag{6d}$$

$$(\mathbf{A}+\mathbf{B})^* = \mathbf{A}^* + \mathbf{B}^*, \tag{6e}$$

$$(\mathbf{A}\mathbf{B})^* = \mathbf{B}^*\mathbf{A}^*. \tag{6f}$$

These properties are readily verified. In addition, the key property of the Hermitian conjugate matrix $\mathbf{A}^*$ is that

$$\boxed{(\mathbf{A}\mathbf{x})\cdot\mathbf{y} = \mathbf{x}\cdot(\mathbf{A}^*\mathbf{y})} \tag{7}$$

holds for all vectors $\mathbf{x}$ and $\mathbf{y}$; specifically, if $\mathbf{A}$ is $m \times n$, $\mathbf{x}$ is any $n \times 1$ vector and $\mathbf{y}$ is any $m \times 1$ vector. To prove (7), observe that

$$(\mathbf{A}\mathbf{x})\cdot\mathbf{y} = (\mathbf{A}\mathbf{x})^{\mathrm{T}}\overline{\mathbf{y}} = \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\overline{\mathbf{y}} = \mathbf{x}^{\mathrm{T}}\overline{\overline{\mathbf{A}}^{\mathrm{T}}\mathbf{y}} = \mathbf{x}^{\mathrm{T}}\overline{\mathbf{A}^*\mathbf{y}} = \mathbf{x}\cdot(\mathbf{A}^*\mathbf{y}). \tag{8}$$

Hermitian matrices. Recall from Chapters 10 and 11 that matrices arising in applications are often *symmetric* ($\mathbf{A}^{\mathrm{T}} = \mathbf{A}$) and that such matrices exhibit several useful properties concerning the eigenvalue problem (Theorems 11.3.1–11.3.4). Likewise, when complex matrices arise in applications they are often *Hermitian* ($\mathbf{A}^* = \overline{\mathbf{A}}^{\mathrm{T}} = \mathbf{A}$), and such matrices exhibit analogous useful properties, given by Theorems 12.3.1–12.3.4 below.

EXAMPLE 4. *Lorentz Transformation.* In the special theory of relativity one considers the vector $[x, y, z, ict]$, where x, y, z are Cartesian coordinates, c is the speed of light, and t is the time. If the corresponding vector, referred to an x', y', z' system which is translating in the z direction with constant speed v, is denoted as $[x', y', z', ict']$, it turns out that these vectors are related according to

$$\begin{bmatrix} x' \\ y' \\ z' \\ ict' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \dfrac{1}{\sqrt{1-\beta^2}} & \dfrac{i\beta}{\sqrt{1-\beta^2}} \\ 0 & 0 & \dfrac{-i\beta}{\sqrt{1-\beta^2}} & \dfrac{1}{\sqrt{1-\beta^2}} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ ict \end{bmatrix}, \tag{9}$$

where $\beta = v/c\,(< 1)$. This is the important *Lorentz transformation*, and it is seen that the transformation matrix is Hermitian. ∎

THEOREM 12.3.1 *Real Eigenvalues*
If $\mathbf{A}$ is Hermitian ($\overline{\mathbf{A}}^{\mathrm{T}} = \mathbf{A}$), then all of its eigenvalues are real.

Proof: Let us both pre-dot and post-dot both sides of $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ with $\mathbf{x}$:

$$\begin{aligned} \text{Pre-dot:}\quad \mathbf{x}\cdot(\mathbf{A}\mathbf{x}) &= \mathbf{x}\cdot(\lambda\mathbf{x}) \\ &= \overline{\lambda}(\mathbf{x}\cdot\mathbf{x}) \quad \text{[according to (3)]} && (10a) \\ \text{Post-dot:}\quad (\mathbf{A}\mathbf{x})\cdot\mathbf{x} &= (\lambda\mathbf{x})\cdot\mathbf{x} \\ &= \lambda(\mathbf{x}\cdot\mathbf{x}) \quad \text{[according to (3)]}. && (10b) \end{aligned}$$

But the left-hand sides are equal, by virtue of property (7) together with the assumption that $\mathbf{A}$ is Hermitian. Thus, subtracting (10b) from (10a) gives

$$(\overline{\lambda} - \lambda)(\mathbf{x}\cdot\mathbf{x}) = 0. \tag{11}$$

Now, $\mathbf{x}\cdot\mathbf{x} = \|\mathbf{x}\|^2 \neq 0$ since $\mathbf{x}$ is an eigenvector, so it follows from (11) that $\overline{\lambda} - \lambda = 0$, or $\overline{\lambda} = \lambda$. Thus, λ is real. ∎

THEOREM 12.3.2 *Dimension of Eigenspace*
If an eigenvalue λ of an Hermitian matrix $\mathbf{A}$ is of multiplicity k, then the eigenspace corresponding to λ is of dimension k.

THEOREM 12.3.3 *Orthogonality of Eigenvectors*
If $\mathbf{A}$ is Hermitian, then eigenvectors corresponding to distinct eigenvalues are orthogonal.

Proof: Let $\mathbf{e}_j$ and $\mathbf{e}_k$ be eigenvectors corresponding to distinct eigenvalues λ_j and λ_k, respectively. That is,

$$\mathbf{A}\mathbf{e}_j = \lambda_j\mathbf{e}_j \quad \text{and} \quad \mathbf{A}\mathbf{e}_k = \lambda_k\mathbf{e}_k. \tag{12a,b}$$

Pre-dotting both sides of (12a) with $\mathbf{e}_k$, and post-dotting both sides of (12b) with $\mathbf{e}_j$, and using the fact that $\overline{\lambda} = \lambda$, we have

$$\begin{aligned} \mathbf{e}_k\cdot(\mathbf{A}\mathbf{e}_j) &= \mathbf{e}_k\cdot(\lambda_j\mathbf{e}_j) \\ &= \overline{\lambda}_j(\mathbf{e}_k\cdot\mathbf{e}_j) \\ &= \lambda_j(\mathbf{e}_k\cdot\mathbf{e}_j) \end{aligned} \quad \Bigg| \quad \begin{aligned} (\mathbf{A}\mathbf{e}_k)\cdot\mathbf{e}_j &= (\lambda_k\mathbf{e}_k)\cdot\mathbf{e}_j \\ &= \lambda_k(\mathbf{e}_k\cdot\mathbf{e}_j). \end{aligned} \tag{13}$$

But (7) tells us that $(\mathbf{A}\mathbf{e}_k)\cdot\mathbf{e}_j = \mathbf{e}_k\cdot(\mathbf{A}^*\mathbf{e}_j)$ and, since $\mathbf{A}^* = \mathbf{A}$ by assumption, the equation on the right side of (13) becomes $\mathbf{e}_k\cdot(\mathbf{A}\mathbf{e}_j) = \lambda_k(\mathbf{e}_k\cdot\mathbf{e}_j)$. Subtracting that equation from the one on the left side of (13) gives

$$0 = (\lambda_j - \lambda_k)(\mathbf{e}_k\cdot\mathbf{e}_j). \tag{14}$$

Finally, $\lambda_j - \lambda_k \neq 0$ since λ_j and λ_k are distinct, by assumption, so it follows from (14) that $\mathbf{e}_k\cdot\mathbf{e}_j = 0$, as was to be shown. ■

THEOREM 12.3.4 *Orthogonal Basis*
If an $n\times n$ matrix $\mathbf{A}$ is Hermitian, then its eigenvectors provide an orthogonal basis for n-space.

Proof of Theorem 12.3.4 follows the same lines as the proof of Theorem 11.3.4.

EXAMPLE 5. Find the eigenvalues and eigenvectors of

$$\mathbf{A} = \begin{bmatrix} 3 & 2i \\ -2i & 0 \end{bmatrix}.$$

Setting $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$ gives the characteristic equation

$$(3-\lambda)(-\lambda) - (2i)(-2i) = \lambda^2 - 3\lambda - 4 = (\lambda+1)(\lambda-4) = 0.$$

Hence,

$$\lambda = -1 \quad \text{and} \quad \lambda = 4.$$

Observe that $\mathbf{A}$ is Hermitian and, sure enough, in accordance with Theorem 12.3.1, the λ's are real.

To find the eigenvectors, solve $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$.

$$\lambda = \lambda_1 = -1: \quad \begin{bmatrix} 4 & 2i \\ -2i & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \text{or} \quad \begin{aligned} 4x_1 + 2ix_2 &= 0 \\ -2ix_1 + x_2 &= 0. \end{aligned}$$

Solving, $x_2 = 2ix_1$ so the eigenspace corresponding to λ_1 is

$$\mathbf{e}_1 = \alpha\begin{bmatrix} 1 \\ 2i \end{bmatrix}.$$

Similarly, we find that the eigenspace corresponding to λ_2 is

$$\mathbf{e}_1 = \beta\begin{bmatrix} 2i \\ 1 \end{bmatrix}.$$

Since $\mathbf{A}$ is Hermitian and λ_1 and λ_2 are distinct, the eigenvectors should be orthogonal (Theorem 12.3.3). Let us see: with $\alpha = \beta = 1$, say, $\mathbf{e}_1\cdot\mathbf{e}_2 = (1)(\overline{2i}) + (2i)(\overline{1}) = -2i + 2i = 0$ so they are, indeed, orthogonal. ■

Hermitian forms, diagonalization, and unitary matrices. The analog of the quadratic form $f(x_1, \ldots, x_n) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is symmetric and $\mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$, is the **Hermitian form**

$$f = \overline{\mathbf{x}}^{\mathrm{T}}\mathbf{A}\mathbf{x}, \tag{15}$$

where $\mathbf{A}$ is Hermitian. For example, if $n = 2$, then (15) becomes

$$\begin{aligned} f &= [\overline{x}_1, \overline{x}_1] \begin{bmatrix} a_{11} & a_{12} \\ \overline{a}_{12} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= a_{11}x_1\overline{x}_1 + a_{22}x_2\overline{x}_2 + a_{12}\overline{x}_1x_2 + \overline{a}_{12}x_1\overline{x}_2. \end{aligned} \tag{16}$$

The right-hand side of (16) is seen to be real since a_{11} and a_{22} are diagonal elements of an Hermitian matrix and hence are real, $x_1\overline{x}_1 = |x_1|^2$ and $x_2\overline{x}_2 = |x_2|^2$ are real, and the fourth term is the conjugate of the third term. In fact, f is real for *all* values $n \geq 1$:

$$\overline{f} = \overline{\overline{\mathbf{x}}^{\mathrm{T}}\mathbf{A}\mathbf{x}} = \mathbf{x}^{\mathrm{T}}\overline{\mathbf{A}\mathbf{x}} = \overline{(\mathbf{A}\mathbf{x})^{\mathrm{T}}}\mathbf{x} = \overline{\mathbf{x}}^{\mathrm{T}}\overline{\mathbf{A}}^{\mathrm{T}}\mathbf{x} = \overline{\mathbf{x}}^{\mathrm{T}}\mathbf{A}\mathbf{x} = f. \tag{17}$$

(See Exercise 9.)

To reduce f to canonical form, set $\mathbf{x} = \mathbf{U}\widetilde{\mathbf{x}}$ in (15). Then

$$f = \overline{(\mathbf{U}\widetilde{\mathbf{x}})^{\mathrm{T}}}\mathbf{A}\mathbf{U}\widetilde{\mathbf{x}} = \overline{\widetilde{\mathbf{x}}}^{\mathrm{T}}(\mathbf{U}^*\mathbf{A}\mathbf{U})\widetilde{\mathbf{x}} \tag{18}$$

so $\mathbf{U}$ is to be chosen so that

$$\mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{D} \tag{19}$$

is diagonal. Recalling our discussion of the real case (Section 11.4), it is not surprising that a suitable $\mathbf{U}$ matrix is a normalized modal matrix of $\mathbf{A}$ (Exercise 10). With that choice, the diagonal elements of $\mathbf{D}$ are the eigenvalues of $\mathbf{A}$, and f reduces to the canonical form

$$\begin{aligned} f &= \lambda_1\widetilde{x}_1\overline{\widetilde{x}}_1 + \lambda_2\widetilde{x}_2\overline{\widetilde{x}}_2 + \cdots + \lambda_n\widetilde{x}_n\overline{\widetilde{x}}_n \\ &= \lambda_1 |\widetilde{x}_1|^2 + \lambda_2 |\widetilde{x}_2|^2 + \cdots + \lambda_n |\widetilde{x}_n|^2 . \end{aligned} \tag{20}$$

In the real case the normalized modal matrix was denoted as $\mathbf{Q}$, and it turned out that $\mathbf{Q}$ was orthogonal; that is, $\mathbf{Q}^{\mathrm{T}} = \mathbf{Q}^{-1}$. Analogously, we find (Exercise 11), in the present case, that

$$\overline{\mathbf{U}}^{\mathrm{T}} = \mathbf{U}^{-1} \quad \text{or} \quad \boxed{\mathbf{U}^* = \mathbf{U}^{-1},} \tag{21}$$

and such a matrix is said to be **unitary**.

THEOREM 12.3.5 *Eigenvalues of Unitary Matrix*
If λ is an eigenvalue of a unitary matrix, then $|\lambda| = 1$.

Proof: Let $\mathbf{U}$ be unitary, with an eigenvalue λ and corresponding eigenvector $\mathbf{e}$:

$$\mathbf{U}\mathbf{e} = \lambda\mathbf{e} \qquad (\mathbf{e} \neq \mathbf{0}). \tag{22}$$

Seeking to employ (21), take the conjugate transpose of both sides:

$$(\overline{\mathbf{U}\mathbf{e}})^{\mathrm{T}} = (\overline{\lambda\mathbf{e}})^{\mathrm{T}} \quad \text{or} \quad \overline{\mathbf{e}}^{\mathrm{T}}\mathbf{U}^{-1} = \overline{\lambda}\overline{\mathbf{e}}^{\mathrm{T}}. \tag{23}$$

And post-multiplying the left with $\mathbf{U}\mathbf{e}$ and the right with $\lambda\mathbf{e}$ [which are equal by (22) and nonzero], (23) gives

$$\overline{\mathbf{e}}^{\mathrm{T}}\mathbf{U}^{-1}\mathbf{U}\mathbf{e} = \overline{\lambda}\overline{\mathbf{e}}^{\mathrm{T}}\lambda\mathbf{e} \quad \text{or} \quad \overline{\mathbf{e}}^{\mathrm{T}}\mathbf{e} = |\lambda|^2\,\overline{\mathbf{e}}^{\mathrm{T}}\mathbf{e}. \tag{24}$$

But $\overline{\mathbf{e}}^{\mathrm{T}}\mathbf{e} = \|\mathbf{e}\|^2 \neq 0$ so (24) implies that $|\lambda|^2 = 1$, or $|\lambda| = 1$, as claimed. ■

EXAMPLE 6. Reduce

$$f = 3x_1\overline{x}_1 + 2i\overline{x}_1x_2 - 2ix_1\overline{x}_2 \tag{25}$$

to a canonical form. Comparing (25) with (16), we see that $a_{11} = 3$, $a_{22} = 0$, and $a_{12} = 2i$, so that

$$\mathbf{A} = \begin{bmatrix} 3 & 2i \\ -2i & 0 \end{bmatrix}.$$

This is the same $\mathbf{A}$ as in Example 5 so

$$\lambda_1 = -1, \quad \hat{\mathbf{e}}_1 = \frac{1}{\sqrt{5}}\begin{bmatrix} 1 \\ 2i \end{bmatrix}; \qquad \lambda_2 = 4, \quad \hat{\mathbf{e}}_2 = \frac{1}{\sqrt{5}}\begin{bmatrix} 2i \\ 1 \end{bmatrix}.$$

Thus, the desired canonical form is

$$f = \lambda_1\,|\widetilde{x}_1|^2 + \lambda_2\,|\widetilde{x}_2|^2 = -\,|\widetilde{x}_1|^2 + 4\,|\widetilde{x}_2|^2,$$

where x_1, x_2 and $\widetilde{x}_1, \widetilde{x}_2$ are related according to

$$\mathbf{x} = \mathbf{U}\widetilde{\mathbf{x}} = [\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2]\widetilde{\mathbf{x}}$$

or

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{\sqrt{5}}\begin{bmatrix} 1 & 2i \\ 2i & 1 \end{bmatrix}\begin{bmatrix} \widetilde{x}_1 \\ \widetilde{x}_2 \end{bmatrix}.$$

Or, if we prefer it the other way around,

$$\widetilde{\mathbf{x}} = \mathbf{U}^{-1}\mathbf{x} = \mathbf{U}^{*}\mathbf{x} \quad \text{or} \quad \begin{bmatrix} \widetilde{x}_1 \\ \widetilde{x}_2 \end{bmatrix} = \frac{1}{\sqrt{5}}\begin{bmatrix} 1 & -2i \\ -2i & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \quad ■$$

Computer software. Using *Maple*, for instance, no new commands are needed; just type I's for i's. For instance, to find the eigenvalues and eigenvectors of

$$\mathbf{A} = \begin{bmatrix} 3 & 2i \\ -2i & 0 \end{bmatrix},$$

type

```
with(linalg):
```

and return. Then enter

```
A := matrix(2, 2, [3, 2 * I, -2 * I, 0]) :
```

and the command

```
eigenvects(A);
```

The result,

$$[4, 1, \{[2I, 1]\}],\ [-1, 1, \{[-\frac{1}{2}I, 1]\}]$$

is the same as found, by hand, in Example 5. The 1's following the eigenvalues 4 and -1 indicate the multiplicity of those eigenvalues.

EXERCISES 12.3

1. Invert each of the following matrices. If the matrix is not invertible (i.e., if it is singular), state that.

(a) $\begin{bmatrix} 2 & i \\ 1 & 4 \end{bmatrix}$ (b) $\begin{bmatrix} 1+i & 3i \\ 1 & 2-i \end{bmatrix}$

(c) $\begin{bmatrix} 3 & 1+i \\ 1-i & 2 \end{bmatrix}$ (d) $\begin{bmatrix} 1 & 3 & i \\ 0 & 1+i & 1 \\ 0 & 0 & 2 \end{bmatrix}$

(e) $\begin{bmatrix} 0 & i & 0 \\ -i & 0 & 2+2i \\ 0 & 2-2i & 0 \end{bmatrix}$ (f) $\begin{bmatrix} 4 & 0 & -i \\ 0 & 4 & 0 \\ i & 0 & 4 \end{bmatrix}$

2. (a)–(f) Same as Exercise 1, but using computer software.

3. (a)–(f) Determine the eigenvalues and eigenvectors of each of the matrices in Exercise 1, and show that the results are in accord with the relevant theorems in this section.

4. (a)–(f) Same as Exercise 3, but using computer software.

5. Determine the eigenvalues and eigenvectors in each case.

(a) $\begin{bmatrix} 1 & 2 \\ 1 & 0 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & 1 \\ -1 & 2 \end{bmatrix}$

(c) $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ (d) $\begin{bmatrix} 4 & -5 \\ 1 & 2 \end{bmatrix}$

6. Give necessary and sufficient conditions on a, b, c, d such that $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ will have complex eigenvalues, i.e., such that $\operatorname{Im} \lambda \neq 0$.

7. Reduce the following to the canonical form (20), and give the matrix transformations from $\mathbf{x}$ to $\tilde{\mathbf{x}}$, and from $\tilde{\mathbf{x}}$ to $\mathbf{x}$.

(a) $f = 2x_1\overline{x}_1 + 3x_2\overline{x}_2 + (1+i)\overline{x}_1x_2 + (1-i)x_1\overline{x}_2$
(b) $f = 2x_1\overline{x}_1 + 5x_2\overline{x}_2 - 2i\overline{x}_1x_2 + 2ix_1\overline{x}_2$
(c) $f = i\overline{x}_1x_2 - ix_1\overline{x}_2 + 2(1+i)\overline{x}_2x_3 + 2(1-i)x_2\overline{x}_3$
(d) $f = 4x_1\overline{x}_1 + 4x_2\overline{x}_2 + 4x_3\overline{x}_3 + ix_1\overline{x}_3 - i\overline{x}_1x_3$

8. Given each of the following $\mathbf{A}$ matrices, evaluate $\mathbf{A}^{1000}$.

(a) $\begin{bmatrix} 2 & 1-i \\ 1+i & 3 \end{bmatrix}$ (b) $\begin{bmatrix} 2 & -2i \\ 2i & 5 \end{bmatrix}$

(c) $\begin{bmatrix} 0 & -i \\ i & 2 \end{bmatrix}$ (d) $\begin{bmatrix} 3 & 2-i \\ 2+i & -1 \end{bmatrix}$

9. Justify the second through fifth equalities in (17), citing relevant equation numbers or theorems.

10. Verify that if $\mathbf{U}$ is a normalized modal matrix of $\mathbf{A}$, then f does reduce to the canonical form (20), as claimed.

11. Verify that a normalized modal matrix $\mathbf{U}$ satisfies (21).

12. Is the Lorentz transformation matrix, in (9), unitary? Orthogonal?

13. If $\mathbf{x} \cdot (\mathbf{A}\mathbf{x}) = \mathbf{x} \cdot (\mathbf{B}\mathbf{x})$ for all $\mathbf{x}$, where $\mathbf{A}$ and $\mathbf{B}$ are Hermitian, does it follow that $\mathbf{A} = \mathbf{B}$? Explain.

14. (*Necessary condition for existence of solutions of* $\mathbf{A}\mathbf{x} = \mathbf{c}$) Prove that a necessary condition for the *existence* of solution(s) to $\mathbf{A}\mathbf{x} = \mathbf{c}$ (where $\mathbf{A}$ is $m \times n$, $\mathbf{x}$ is $n \times 1$, and $\mathbf{c}$ is $m \times 1$) is that $\mathbf{c}$ be orthogonal to every solution $\mathbf{z}$ of the associated homogeneous equation $\mathbf{A}^*\mathbf{z} = \mathbf{0}$. HINT: Dot both sides of $\mathbf{A}\mathbf{x} = \mathbf{c}$ with $\mathbf{z}$, and then use (7).

15. Use the result stated in Exercise 14 to determine necessary conditions, if any, on the components of $\mathbf{c}$, for the given system to be consistent. Compare your result with those obtained by direct application of Gauss elimination to the given system.

(a) $\begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$

(b) $\begin{bmatrix} 2 & 1 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$

(c) $\begin{bmatrix} 2 & 1 \\ 4 & 2 \\ i & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$

(d) $\begin{bmatrix} 1 & 3 & 2 & 1 \\ 2 & -1 & 1 & 0 \\ 3 & 2 & 3 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$

(e) $\begin{bmatrix} 1 & 3 & 2 & i \\ 2 & -1 & 1 & 0 \\ 3 & 2 & 3 & i \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$

(f) $\begin{bmatrix} 2 & 1 & 1 \\ i & 1 & 2i-3 \\ 1 & 2 & -4 \\ -1 & 1 & -5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix}$

16. (*Decomposition*) Given any $m \times n$ matrix $\mathbf{A}$, Hermitian or not, show that $\mathbf{A}$ can be split as $\mathbf{A} = \mathbf{B} + i\mathbf{C}$, where $\mathbf{B}$ and $\mathbf{C}$ are each Hermitian. Show that $\mathbf{B} = (\mathbf{A}^* + \mathbf{A})/2$ and $\mathbf{C} = i(\mathbf{A}^* - \mathbf{A})/2$.

17. (a) If $\mathbf{A}$ is Hermitian, is $i\mathbf{A}$ Hermitian? Explain.
(b) If $\mathbf{A}$ is Hermitian, is $\mathbf{A}^2$ Hermitian? Explain.

Chapter 12 Review

This chapter is so compact that it hardly warrants review. But, let us stress three points.

First, the key difference between $\mathbb{R}^n$ and $\mathbb{C}^n$ is in the *dot product*, which is $\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^{\mathrm{T}}\mathbf{y}$ for $\mathbb{R}^n$ and

$$\boxed{\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^{\mathrm{T}}\overline{\mathbf{y}}} \tag{1}$$

for $\mathbb{C}^n$, the complex conjugate being introduced so that the dot product $\mathbf{x} \cdot \mathbf{y}$ and the norm $\|\mathbf{x}\| = \sqrt{\mathbf{x} \cdot \mathbf{x}}$ satisfy the properties listed in equations (12) and (17), respectively, in Section 9.5. (If you studied the optional Section 9.6 you will recall that these properties were elevated to requirements, or axioms, for any normed inner product space.)

Second, just as real matrices are found, in Section 11.3, to admit special useful properties (e.g., the eigenvalues are real, eigenvectors corresponding to distinct eigenvalues are orthogonal, and the eigenvectors provide an orthogonal basis for the n-space) if they are symmetric ($\mathbf{A}^{\mathrm{T}} = \mathbf{A}$), we find in this chapter that complex matrices admit the same properties if they are *Hermitian* ($\mathbf{A}^* = \mathbf{A}$, where $\mathbf{A}^*$ is $\overline{\mathbf{A}}^{\mathrm{T}}$).

Finally, we call your attention to the result

$$\boxed{(\mathbf{A}\mathbf{x}) \cdot \mathbf{y} = \mathbf{x} \cdot (\mathbf{A}^*\mathbf{y}),} \tag{2}$$

given as equation (7) in Section 12.3. For real matrices the latter becomes

$$(\mathbf{A}\mathbf{x}) \cdot \mathbf{y} = \mathbf{x} \cdot (\mathbf{A}\mathbf{y}) \tag{3}$$

if $\mathbf{A}$ is symmetric, and for complex matrices (3) holds if $\mathbf{A}$ is Hermitian; (3) is the key relation used in proving that eigenvectors corresponding to distinct eigenvalues are orthogonal, which result, in turn, is needed in proving that the eigenvectors provide an orthogonal basis for the n-space. We will meet a function-space version of (3) known as *Lagrange's identity*, in Chapter 17, when we study the *Sturm–Liouville theory*. Expansions in terms of bases consisting of orthogonal eigenvectors (or "eigenfunctions" in the function space case) will be of great importance to us, and therefore the underlying relation (3) is of great importance as well.

Chapter 13

Differential Calculus of Functions of Several Variables

13.1 Introduction

In Chapters 8 – 12 on Linear Algebra most of our interest was in n-space. In Chapters 13–16 on Multivariable Calculus and Field Theory we return to physical two- and three-dimensional space. The heart of this group of chapters is Chapter 16, on scalar and vector field theory. The three preceding chapters prepare the way by covering a number of topics from the subject area generally known as "advanced calculus."

In Chapter 13 we consider (real valued) functions of more than one (real) variable, and lay some of the groundwork for our subsequent development of field theory and partial differential equations. Much of the development parallels that in the calculus of functions of a single variable: derivatives, chain differentiation, Taylor's formula, the mean value theorem, and maxima and minima.

Chapter 14 is not a continuation of the development begun in Chapter 13; it covers the fundamentals of vectors and their manipulation: the dot and cross product, base vectors, differentiation of vectors, Cartesian and non-Cartesian coordinate systems, and the representation of curves in space. Chapter 15 returns to the calculus of functions of several variables, begun in Chapter 13, with introductory material on double and triple integrals, and on surfaces and volumes. Finally, in Chapter 16 we consider both scalar and vector fields; we introduce the divergence, gradient, and curl, and consider several integral theorems and a number of physical applications from the subjects of heat conduction, fluid flow, and electrostatics.

Discussion of non-Cartesian coordinates (polar, cylindrical, and spherical) is confined to OPTIONAL sections so that it can be omitted in a shorter course.

13.2 Preliminaries

13.2.1. Functions. We will be considering real-valued functions of n real variables, say $f(x_1, \ldots, x_n)$. The function is a **mapping**, or **transformation**, from a point $\mathbf{x} = (x_1, \ldots, x_n)$ on an f axis, $\mathbb{R}^1$, as illustrated in Fig. 1a for $n = 1$. To obtain

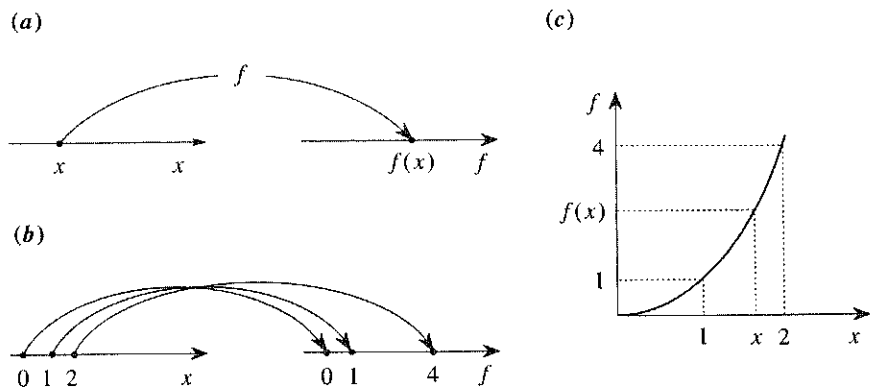


Figure 1. The function f.

a graphical display of the mapping one can draw arrows from a number of x points to their image points on the f axis as we have done in Fig. 1b for the function $f(x) = x^2$. However, Descartes had a better idea: place the x and f axes at right angles to each other and plot all points $(x, f(x))$ for the desired x interval. That procedure gives the familiar **graph** of f, as in Fig. 1c.

Similarly, if $n = 2$ we can obtain the graph of f as a surface in three-dimensional Cartesian x_1, x_2, f space. Of course, if $n \geq 3$ the idea fails because we would need four or more dimensions in which to present the graph.

In Chapters 13–16, n will usually be 2 or 3, but there is no need for any such limitation in the present discussion.

13.2.2. Point set theory definitions. The physical nature of the independent variables $x_1, \ldots, x_n$ is irrelevant here; they may be pressures, distances, or whatever. As noted above, we will regard $x_1, \ldots, x_n$ as the coordinates of a point in an n-dimensional space and will denote that point as $\mathbf{x} = (x_1, \ldots, x_n)$. Further, we define the **distance** $d(\mathbf{x}, \mathbf{x}')$ between $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{x}' = (x_1', \ldots, x_n')$ as

$$\boxed{d(\mathbf{x}, \mathbf{x}') \equiv \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2 + \cdots + (x_n - x_n')^2}.} \tag{1}$$

If $n = 1$, then $\mathbf{x}$ and $\mathbf{x}'$ are two points x_1 and x_1' on an x axis, and (1) becomes

$$d(\mathbf{x}, \mathbf{x}') = |x_1 - x_1'|. \tag{2a}$$

For $n = 2$

$$d(\mathbf{x}, \mathbf{x}') = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2}, \tag{2b}$$

and for $n = 3$

$$d(\mathbf{x}, \mathbf{x}') = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2 + (x_2 - x_3')^2}. \tag{2c}$$

These formulas are familiar to us: (2b) and (2c) are the Pythagorean formulas for the distance between points in two- and three-dimensional Euclidean space, and for $n > 3$ the definition (1) amounts to an n-dimensional generalization of these Pythagorean formulas.*

Next, we define a **neighborhood** $N(\mathbf{x}'; r)$ of a point $\mathbf{x}'$ as the set of all points $\mathbf{x}$ closer to $\mathbf{x}'$ than r, namely, such that

$$d(\mathbf{x}, \mathbf{x}') < r. \qquad (r > 0) \tag{3}$$

In one dimension (3) becomes $|x' - x| < r$, which corresponds to the *interval* shown in Fig. 2a (not including the endpoints $x' + r$ and $x' - r$), in two dimensions N is the circular *disk* shown in Fig. 2b (not including the points on its edge), in three dimensions N is a *sphere* (not including the points on its surface), and if $n > 3$ we speak of N as an n-dimensional *hypersphere*.

Next, we say that a point set $\mathcal{S}$ is **connected** if each pair of points in $\mathcal{S}$ can be joined by an unbroken line consisting of a finite number of straight segments, each contained entirely within $\mathcal{S}$.†

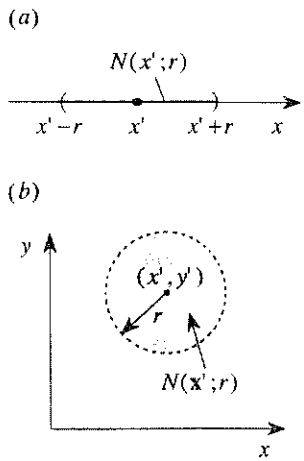


Figure 2. One- and two-dimensional neighborhoods.

EXAMPLE 1. Each of the point sets $\mathcal{S}_1$ and $\mathcal{S}_2$ in the two-dimensional plane, shown in Fig. 3, is connected. The points $\mathbf{x}_1, \mathbf{x}_2$, for example, may be joined as shown. In contrast,

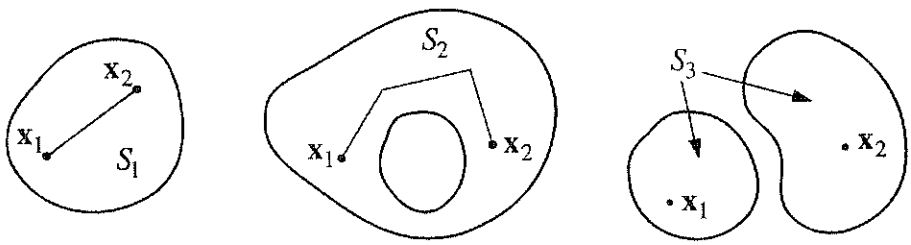


Figure 3. Connectedness.

*Recall from Section 9.5 that (1) is the Euclidean norm of the vector from $\mathbf{x}$ to $\mathbf{x}'$ in n-space. It is not the *only* definition of distance possible (just as the Euclidean norm is not the only viable norm in n-space), but it is the one most commonly adopted. For instance, $d(\mathbf{x}, \mathbf{x}') \equiv |x_1 - x_1'| + \cdots + |x_n - x_n'|$ is also acceptable and is sometimes preferred because it is algebraically simpler.

†A straight line may be defined in n-dimensional space, even if $n > 3$, by parametric equations $x_1 = a_1 + b_1 t, x_2 = a_2 + b_2 t, \ldots, x_n = a_n + b_n t$, where $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ are constants, and where t is the parameter.

the set $\mathcal{S}_3$ is not connected since not *all* pairs of points in $\mathcal{S}_3$ can be joined by linear segments lying entirely within $\mathcal{S}_3$. The points $\mathbf{x}_1, \mathbf{x}_2$, for example, cannot be so joined. ∎

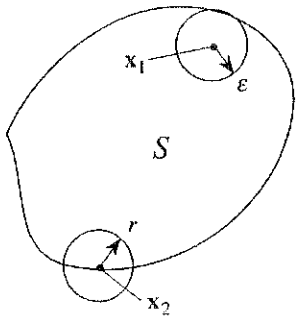


Figure 4. Interior point $\mathbf{x}_1$, boundary point $\mathbf{x}_2$.

A point $\mathbf{x}$ is said to be a **boundary point** of a point set $\mathcal{S}$ if *every* neighborhood $N(\mathbf{x}; r)$ (i.e., no matter how small we take r to be) contains points in $\mathcal{S}$ *as well as* points not in $\mathcal{S}$. Thus, a boundary point is just what it sounds like. To illustrate, the point $\mathbf{x}_1$ in the two-dimensional point set $\mathcal{S}$ shown in Fig. 4 is *not* a boundary point since the disk $N(\mathbf{x}_1; r)$ contains no points outside $\mathcal{S}$ if we choose r to be smaller than ϵ. The point $\mathbf{x}_2$, on the other hand, *is* a boundary point since $N(\mathbf{x}_2; r)$ evidently contains points in $\mathcal{S}$ *and* points not in $\mathcal{S}$, no matter how small we take r to be. (Recall, from the definition of neighborhood, that $r > 0$, so r cannot be zero.)

A point $\mathbf{x}$ of a set $\mathcal{S}$ is said to be an **interior point** of $\mathcal{S}$ if there exists some neighborhood of $\mathbf{x}$ lying entirely within $\mathcal{S}$. For instance, the point $\mathbf{x}_1$ in Fig. 4 is an interior point of $\mathcal{S}$ because the disk $N(P; r)$ lies entirely within $\mathcal{S}$ for any $r < \epsilon$.

Finally, we call a connected set $\mathcal{S}$ an **open region** or **domain** if it contains *none* of its boundary points, and a **closed region** if it contains *all* of its boundary points.

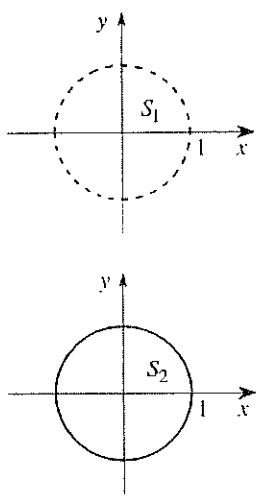


Figure 5. Open and closed regions.

EXAMPLE 2. Let $\mathcal{S}_1$ be the set of points (x, y), in Cartesian 2-space, such that $x^2 + y^2 < 1$ as shown in Fig. 5. We write $\mathcal{S}_1 = \{(x, y) | x^2 + y^2 < 1\}$. Evidently, $\mathcal{S}_1$ is connected. Its boundary points are the points on the unit circle $x^2 + y^2 = 1$. None of these points belongs to $\mathcal{S}_1$ (as conveyed in the figure by the use of dashed lines), so $\mathcal{S}_1$ is an open domain. We call this domain the *open unit disk*. Every point in $\mathcal{S}_1$ is an interior point.

Next, let $\mathcal{S}_2 = \{(x, y) | x^2 + y^2 \leq 1\}$ as shown in Fig. 5. This time all the boundary points (those on the circle $x^2 + y^2 = 1$) are contained in $\mathcal{S}_2$, as emphasized by the solid boundary line in the figure, so $\mathcal{S}_2$ is a closed region, the *closed unit disk*. All points (x, y) such that $x^2 + y^2 < 1$ are interior points of $\mathcal{S}_2$. ∎

Of course, if $\mathcal{S}$ contains some but not all of its boundary points, then it is neither an open region *nor* a closed region. However, the regions that we shall work with will generally be either open or closed. Further, the terms "open" and "closed" are not mutually exclusive. For example, the connected set $\mathcal{S} = \{x | -\infty < x < \infty\}$, namely, the entire real axis, has no boundary points. Thus it is at once open (since it contains none of its boundary points) and closed (since it contains all of its boundary points).

13.2.3. Limits and continuity. To define limits and continuity for functions of n variables, let us first review these concepts, from the calculus, for a function of a single variable.

We say that $f(x)$ has a **limit** L as x tends to x', and write

$$\lim_{x \to x'} f(x) = L \quad \text{or} \quad f(x) \to L \quad \text{as } x \to x', \tag{4}$$

if to each $\epsilon > 0$ (i.e., no matter how small) there corresponds a $\delta(\epsilon, x') > 0$ such that $|f(x) - L| < \epsilon$ whenever $0 < |x - x'| < \delta$. That is, $f(x)$ can be made arbitrarily close to L by making x sufficiently close to x'.

EXAMPLE 3. Consider $f(x) = 1/x$ over $0.1 \le x \le 1$. At the x' shown in Fig. 6, the limit exists and is equal to the L shown, namely, $1/x'$. To prove that claim we need to put forward a suitable function $\delta(\epsilon, x')$. Draw an arbitrarily small ϵ band about L and, where it intersects the graph (at A and B), drop verticals to the x axis. Observing that $a < b$, we can choose $\delta = a$. Then $|x - x'| < \delta$ is the centered interval denoted by the small parentheses. Surely, if x is closer to x' than δ (i.e., $0 < |x - x'| < \delta$), then $f(x)$ will be closer to L than ϵ, as desired. To determine the functional form of δ, write $f(x' - \delta) - f(x') = \epsilon$. That is, $1/(x' - \delta) - 1/x' = \epsilon$. Solving for δ,

$$\delta(\epsilon, x') = x' - \frac{x'}{\epsilon x' + 1}. \tag{5}$$

Of course any smaller (nonzero) value, such as $\frac{1}{5}[x' - x'/(\epsilon x' + 1)]$, will do just as well.

COMMENT. Practically speaking, we don't really need to put forward a $\delta(\epsilon, x')$ to convince ourselves that the limit of $f(x)$, as $x \to x'$, is L (in Fig. 6). Our purpose, in doing so here, has been to clarify the meaning of the ϵ, δ definition of limit, given above. ∎

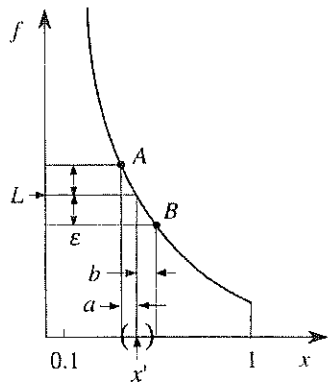


Figure 6. $\lim f(x)$ for $f(x) = 1/x$.

Note carefully that because of the "0 <" in $0 < |x - x'| < \delta$, in our definition of $\lim_{x \to x'} f(x)$, $f(x')$ need not equal L. For instance, if

$$f(x) = \begin{cases} x^2, & 0 \le x \le 3 \quad \text{but } x \ne 2 \\ 12, & x = 2 \end{cases} \tag{6}$$

then (Fig. 7) $\lim_{x \to 2} f(x) = 4$, not 12. [Similarly, the limit of $f(x)$ is 0 as $x \to 0$, 1 as $x \to 1$, 6.25 as $x \to 2.5$, and 9 as $x \to 3$.]*

If, *in addition* to having $\lim_{x \to x'} f(x) = L$ we also have $f(x') = L$, hence

$$\lim_{x \to x'} f(x) = f(x'), \tag{7}$$

then we say that $f(x)$ is **continuous** at x'. In ϵ, δ language, *$f(x)$ is continuous at x' if to each $\epsilon > 0$ (i.e., no matter how small) there corresponds a $\delta(\epsilon, x') > 0$ such that $|f(x) - f(x')| < \epsilon$ whenever $|x - x'| < \delta$.* If $f(x)$ is not continuous, then it is **discontinuous**.

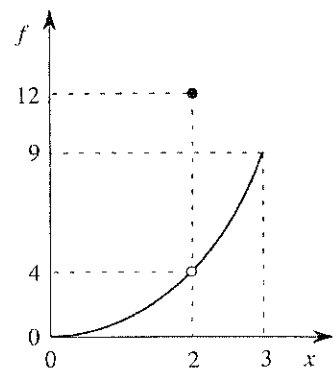


Figure 7. The function defined by (6).

For instance, the function $f(x) = 1/x$ is continuous on any x interval not containing the origin, such as $1 \le x < \infty$, $0 < x \le 5$, or $-\infty < x < 0$. The functions e^x and $\sin x$ are continuous on $-\infty < x < \infty$, and the function defined

*In $0 < |x - x'| < \delta$, within the italicized definition of limit, it is understood that x also needs to be within the interval of definition of the function. Thus, we say that $f(x) \to 6.25$ as $x \to 2.5$ is a *two-sided limit*, since x can approach 2.5 from the right or from the left, but $f(x) \to 9$ as $x \to 3$ is a *one-sided limit*, or a *left-hand limit*, since x can approach 3 only from the left. Similarly, $f(x) \to 0$ as $x \to 0$ is a one-sided or *right-hand limit*.

by (6) and shown in Fig. 7 is continuous on $0 \leq x \leq 3$, except at $x = 2$. The Heaviside step function, defined by

$$H(x) = \begin{cases} 1, & x > 0 \\ 0, & x < 0 \end{cases} \tag{8}$$

with $H(0) = \frac{1}{2}$, say, is continuous on any interval not containing the origin. At the origin it is discontinuous because (7) does not hold there: $f(x') = f(0) = \frac{1}{2}$, but the left-hand side of (7) does not exist because $f(x) \to 1$ as $x \to 0$ from the right, whereas $f(x) \to 0$ as $x \to 0$ from the left. For $\lim_{x \to x'} f(x)$ to exist, the limit from the left (the *left-hand limit*) and the limit from the right (the *right-hand limit*) must exist and equal each other. In the present example both of those limits exist, but they do not equal each other.

Similarly for the functions of several variables. We say that $f(\mathbf{x})$ is continuous at x' if

$$\lim_{\mathbf{x} \to \mathbf{x}'} f(\mathbf{x}) = f(\mathbf{x}'), \tag{9}$$

where $\mathbf{x} = (x_1, \ldots, x_n), \mathbf{x}' = (x_1', \ldots, x_n')$, and where we use $f(\mathbf{x})$ for brevity to denote $f(x_1, \ldots, x_n)$, and similarly for $f(\mathbf{x}')$. In ϵ, δ language, $f(\mathbf{x})$ is continuous at $\mathbf{x}'$ if to each $\epsilon > 0$ (i.e., no matter how small) there corresponds a $\delta(\epsilon, \mathbf{x}')$ such that $|f(\mathbf{x}) - f(\mathbf{x}')| < \epsilon$ whenever $d(\mathbf{x}, \mathbf{x}') < \delta$, where $d(\mathbf{x}, \mathbf{x}')$ was defined by (1). If it is not continuous, then it is discontinuous.

For instance, $f(x_1, x_2) = x_1^2 + 3\sin(x_1 x_2)$ is continuous for all values (i.e., for all finite values) of x_1 and x_2, and $g(x_1, x_2, x_3) = 1/(x_1^2 + x_2^2 + x_3^2)$ is continuous everywhere except at $x_1 = x_2 = x_3 = 0$. Generally, the functions that arise in applications are either continuous everywhere, or almost everywhere. Of the example functions f and g, for instance, f is continuous everywhere and g is continuous everywhere except at the origin.

Closure. To study limits and continuity one first needs to establish basic concepts of point set theory. For instance, if we wish $\lim_{\mathbf{x} \to \mathbf{x}'} f(\mathbf{x}) = L$ to mean that $f(\mathbf{x})$ can be made arbitrarily close to L by making $\mathbf{x}$ sufficiently close to $\mathbf{x}'$, then we need to make the concept of closeness precise by defining the *distance* between $\mathbf{x}$ and $\mathbf{x}'$. We adopt the n-dimensional version

$$d(\mathbf{x}, \mathbf{x}') = \sqrt{(x_1 - x_1')^2 + \cdots + (x_n - x_n')^2}$$

of the familiar two- and three-dimensional Euclidean distance. Other terms defined here, that will come up in these chapters, include neighborhood, connected, boundary point, interior point, open region (or domain), and closed region.

EXERCISES 13.2

1 Determine the distance $d(P, P')$ in each case.

(a) $P = (3, 1, 0)$, $P' = (-1, 2, 2)$
(b) $P = (1, -1, 5, 0)$, $P' = (0, 4, 3, 2)$
(c) $P = (1, -4)$, $P' = (6, 5)$
(d) $P = (0, 4, 3, 0)$, $P' = (0, 0, 0, 5)$
(e) $P = (-6)$, $P' = (-2)$
(f) $P = (8, 0, 1, 0)$, $P' = (0, 2, -6, 7)$

2. Is $P = (3.2, 3.7)$ in $N(P_0; r)$ if $P_0 = (3, 4)$ and $r = 0.4$?

3. Is $P = (4.3, 1.1)$ in $N(P_0; r)$ if $P_0 = (4.2, 1)$ and $r = 0.2$?

4. Is $P = (2, 5, 7)$ in $N(P_0; r)$ if $P_0 = (3, 4, 5)$ and $r = 3$?

5. Is $P = (0.99, 1.96, -0.95, 0.03)$ in $N(P_0; r)$ if $P_0 = (1, 2, -1, 0)$ and $r = 0.01$?

6. In each case identify the boundary points and indicate whether or not the set is connected, an open region, a closed region, or neither.

(a) $\{x \mid 0 < x \leq 1\}$ (b) $\{x \mid -3 \leq x < 5\}$
(c) $\{x \mid 0 < x < \infty\}$ (d) $\{x \mid 0 \leq x < \infty\}$
(e) $\{x \mid 0 < \sin x \leq 1\}$ (f) $\{x \mid \sin x = 0\}$
(g) $\{(x_1, x_2) \mid 2 < x_1 < 3, 1 < x_2 < 5\}$
(h) $\{(x_1, x_2) \mid -\infty < x_1 < \infty, 0 < x_2 < 1\}$
(i) $\{(x_1, x_2) \mid 6 < x_1 < 8, x_2 = 0\}$
(j) $\{(x_1, x_2) \mid 1 < x_1^2 + x_2^2 \leq 4\}$
(k) $\{(x_1, x_2) \mid 0 < x_1^2 + x_2^2 < 1\}$
(l) $\{(x_1, x_2) \mid -2 < x_1 x_2 < 2\}$
(m) $\{(x_1, x_2, x_3) \mid 1 < x_1^2 + x_3^2 < 2, x_2 = 0\}$
(n) $\{(x_1, x_2, x_3, x_4) \mid 3 \leq x_1^2 + x_2^2 + x_3^2 + x_4^2 < \infty\}$
(o) $\{(x_1, x_2, x_3, x_4) \mid 0 < x_1 < 1, 0 < x_2 < 1, 0 < x_3 < 1, 0 < x_4 < 1\}$

7. For the stated limit, put forward a suitable $\delta(\epsilon)$, as we did in Example 3. NOTE: In Example 3, x' was any point in $0.1 \leq x \leq 1$, so δ depended not only on ϵ but also on x'. In this exercise x' is a specific point, so δ will depend only on ϵ.

(a) $\lim_{x \to 2} x^2 = 4$ (b) $\lim_{x \to 5} 3x = 15$
(c) $\lim_{x \to 1} \sin x = \sin 1$ (d) $\lim_{x \to 2} \sin x = \sin 2$
(e) $\lim_{x \to 0} e^x = 1$ (f) $\lim_{x \to 0} e^{-x} = 1$

8. (a) Does $\lim_{x \to 0} \sin(1/x)$ exist? If so, give its value; if not, explain why not. HINT: To sketch the graph of $\sin(1/x)$ it is useful to re-express it as $\sin[(1/x^2)x]$ since then we have the more familiar form $\sin \omega x$, where in this case the frequency $\omega = 1/x^2$ is a function of x. Alternatively, setting $t = 1/x$ observe that $\lim_{x \to 0} \sin(1/x) = \lim_{t \to \infty} \sin t$.
(b) Show that $\lim_{x \to 0} x^2 \sin(1/x) = 0$, by putting forward a suitable $\delta(\epsilon)$.

9. (a) Prove that

$$|A + B| \leq |A| + |B|. \tag{9.1}$$

(b) Show that it follows from (9.1) that

$$|A_1 + \cdots + A_n| \leq |A_1| + \cdots + |A_n|. \tag{9.2}$$

(c) Prove that $\lim_{x \to a} f(x) = A$ is equivalent to the statement

$$\lim_{x \to a} [f(x) - A] = 0. \tag{9.3}$$

(d) Prove that

$$\lim_{x \to a} [f(x) + g(x)] = \lim_{x \to a} f(x) + \lim_{x \to a} g(x), \tag{9.4}$$

provided that each of the limits on the right exists. HINT: Let $\lim_{x \to a} f(x) = A$ and $\lim_{x \to a} g(x) = B$. Then, for any $\epsilon > 0$ there is a δ_1 such that $|f(x) - A| < \epsilon$ whenever $0 < |x - a| < \delta_1$, and a δ_2 such that $|g(x) - B| < \epsilon$ whenever $0 < |x - a| < \delta_2$. Then, express $|[f(x) + g(x)] - (A + B)|$ as $|[f(x) - A] + [g(x) - B]|$ and use (9.1).
(e) Prove that

$$\lim_{x \to a} [Cf(x)] = C \lim_{x \to a} f(x), \tag{9.5}$$

provided that $\lim_{x \to a} f(x)$ exists.
(f) Show that it follows from (9.4) and (9.5) that the limit is **linear**:

$$\lim_{x \to a} [\alpha f(x) + \beta g(x)] = \alpha \lim_{x \to a} f(x) + \beta \lim_{x \to a} g(x) \tag{9.6}$$

for arbitrary constants α, β, provided that each of the limits on the right exists.
(g) Prove that

$$\lim_{x \to a} [f(x) g(x)] = [\lim_{x \to a} f(x)][\lim_{x \to a} g(x)], \tag{9.7}$$

provided that each of the limits on the right exists. HINT: Let $\lim_{x \to a} f(x) = A$ and $\lim_{x \to b} g(x) = B$. Then, for any $\epsilon > 0$ there is a δ_1 such that $|f(x) - A| < \epsilon$ whenever $0 < |x - a| < \delta_1$, and there is a δ_2 such that $|g(x) - B| < \epsilon$

whenever $0 < |x - a| < \delta_2$. Let ϵ be less than unity. Then, $|f(x)| = |A + [f(x) - A]| < |A| + \epsilon < |A| + 1$ so

$$\begin{aligned} |f(x)g(x) - AB| &= |f(x)g(x) - Bf(x) + Bf(x) - AB| \\ &\le |f(x)||g(x) - B| + |B||f(x) - A| \\ &< (|A| + 1)\epsilon + |B|\epsilon. \end{aligned} \tag{9.8}$$

Provide the remaining steps in the proof.

10. We state in Exercise 9 that (9.4) holds, "provided that each of the limits on the right exists." Give a counterexample, where (9.4) fails to hold because the limit on the left exists, yet those on the right do not. That is, give $f(x)$, $g(x)$, and a.

11. We stated that (9.5) in Exercise 9 holds, "provided that $\lim_{x\to a} f(x)$ exists." Give a counterexample, where (9.5) fails to hold because the limit on the left exists, yet the one on the right does not. That is, give $f(x)$, C, and a.

12. Identify any points or point sets (for instance, the plane $x + y - 2z = 5$), if any, where f is discontinuous.

(a) $f(x) = \dfrac{x+1}{x(x^2 + x - 2)^2}$
(b) $f(x) = \tan x$
(c) $f(x, y) = \tan(x + y)$
(d) $f(x, y) = e^x/(y + 3)$
(e) $f(x, y) = (x + y)/(x^2 + y)^2$
(f) $f(x, y, z) = x/(yz)$
(g) $f(x, y, z) = 5/(x^2 + y^2 + z^2 - 1)$

13.3 Partial Derivatives

To review partial differentiation, it should suffice to consider a function f of two variables, say x and y. Suppose $f(x, y)$ is defined throughout some neighborhood of a point $\mathbf{x}_0 = (x_0, y_0)$. Let us hold y fixed at y_0, so that $f(x, y) = f(x, y_0)$ is then a function of x alone. If the x-derivative of this function exists at x_0, it is called the **partial derivative** of f with respect to x, at (x_0, y_0), and is usually denoted either as $\partial f/\partial x$ or as f_x. Thus,

$$f_x = \frac{\partial f}{\partial x} = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x, y_0) - f(x_0, y_0)}{\Delta x} \tag{1a}$$

at (x_0, y_0). Similarly,

$$f_y = \frac{\partial f}{\partial y} = \lim_{\Delta y \to 0} \frac{f(x_0, y_0 + \Delta y) - f(x_0, y_0)}{\Delta y} \tag{1b}$$

at (x_0, y_0). For example, if $f(x, y) = x^2y^3$, then $f_x = 2xy^3$ and $f_y = 3x^2y^2$ at any given point (x, y).

In geometrical terms, (1a) is the slope of the tangent line A, and (1b) is the slope of the tangent line B in Fig. 1. For the case shown in the figure, $f_x > 0$ and $f_y < 0$ at $\mathbf{x}_0$, that is, at (x_0, y_0).

If the partial derivatives f_x and f_y exist not only at $\mathbf{x}_0$, but throughout some neighborhood of $\mathbf{x}_0$, then they are functions of x and y which may, in turn, admit further derivatives, namely, the *second-order partial derivatives*

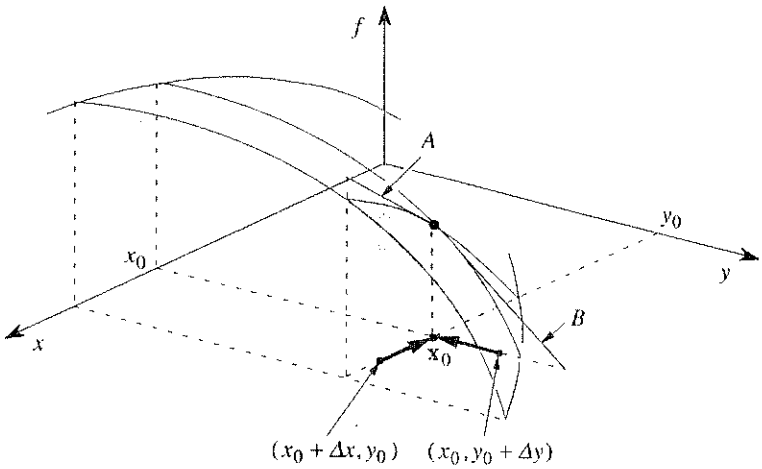


Figure 1. f_x and f_y at $\mathbf{x}_0$.

$$\begin{aligned} \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) &= \frac{\partial^2 f}{\partial x^2} = f_{xx}, & \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) &= \frac{\partial^2 f}{\partial x \partial y} = f_{yx}, \\ \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) &= \frac{\partial^2 f}{\partial y \partial x} = f_{xy}, & \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right) &= \frac{\partial^2 f}{\partial y^2} = f_{yy}. \end{aligned} \tag{2}$$

Similarly, f may admit *third-order partial derivatives*, such as

$$\frac{\partial}{\partial x}\left(\frac{\partial^2 f}{\partial y^2}\right) = \frac{\partial^3 f}{\partial x \partial y^2} = f_{yyx},$$

and so on.

Note the order of the subscripts. For example, f_{yx} means $(f_y)_x$. Thus,

$$f_{yx} = (f_y)_x = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right), \qquad f_{xyy} = ((f_x)_y)_y = \frac{\partial}{\partial y}\left(\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)\right),$$

and so on. Does the order really matter? That is, for a "mixed" derivative such as f_{xy} is it true that $f_{xy} = f_{yx}$ always? Consider two examples.

EXAMPLE 1. Let $f(x, y) = x \sin(xy^2)$. Then

$$\begin{aligned} f_x &= \sin(xy^2) + xy^2 \cos(xy^2), \\ f_{xy} &= 2xy \cos(xy^2) + 2xy \cos(xy^2) - 2x^2y^3 \sin(xy^2), \\ f_y &= 2x^2 y \cos(xy^2), \\ f_{yx} &= 4xy \cos(xy^2) - 2x^2y^3 \sin(xy^2), \end{aligned}$$

so that $f_{xy} = f_{yx}$, in this case, for all values of x and y. ∎

EXAMPLE 2. Let

$$f(x,y) = \begin{cases} \dfrac{xy(x^2-y^2)}{x^2+y^2} & \text{if } (x,y) \neq (0,0) \\ 0 & \text{if } (x,y) = (0,0). \end{cases} \tag{3a,b}$$

[The separate specification (3b) is needed because at the origin (3a) gives $0/0$, which is not defined; (3b) was *not* obtained from (3a) by letting x and y tend to zero, it was stated as the author's choice, as part of the *definition* of $f(x,y)$.] If we stay away from the origin, and work only with (3a), we would find, as in Example 1, that $f_{xy} = f_{yx}$ at each point (x,y) in the plane. We will demonstrate that $f_{xy} = f_{yx}$ does *not* hold, however, at the origin $(0,0)$.

First, take $\partial/\partial x$ of f. That step gives

$$f_x(x,y) = \begin{cases} \dfrac{y(x^4+4x^2y^2-y^4)}{(x^2+y^2)^2} & \text{if } (x,y) \neq (0,0) \\ 0 & \text{if } (x,y) = (0,0). \end{cases} \tag{4a,b}$$

Obtaining (4a) from (3a) was straightforward, but where did we get (4b)? We obtained (4b) directly from (1a), as follows:

$$\begin{aligned} f_x(0,0) &= \lim_{\Delta x \to 0} \frac{f(\Delta x, 0) - f(0,0)}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{0-0}{\Delta x} = \lim_{\Delta x \to 0} 0 = 0, \end{aligned}$$

where $f(\Delta x, 0) = 0$ followed from (3a) and $f(0,0) = 0$ from (3b). Continuing,

$$\begin{aligned} f_{xy}(0,0) &= \lim_{\Delta y \to 0} \frac{f_x(0,\Delta y) - f_x(0,0)}{\Delta y} \\ &= \lim_{\Delta y \to 0} \frac{-\Delta y - 0}{\Delta y} \quad \text{from (4a) and (4b)} \\ &= -1. \end{aligned} \tag{5}$$

Similarly, we find that

$$f_y(x,y) = \begin{cases} \dfrac{x(x^4-4x^2y^2-y^4)}{(x^2+y^2)^2} & \text{if } (x,y) \neq (0,0) \\ 0 & \text{if } (x,y) = (0,0) \end{cases} \tag{6a,b}$$

and

$$\begin{aligned} f_{yx}(0,0) &= \lim_{\Delta x \to 0} \frac{f_y(\Delta x, 0) - f_y(0,0)}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{\Delta x - 0}{\Delta x} \quad \text{from (6a) and (6b)} \\ &= 1. \end{aligned} \tag{7}$$

Comparing (5) and (7), we see that $f_{xy}(0,0) \neq f_{yx}(0,0)$ in this example. ∎

In general, then, the order of differentiation does matter. The following theorem gives conditions which guarantee that the mixed partial derivatives f_{xy} and

f_{yx} of a given function $f(x, y)$ are equal.

THEOREM 13.3.1 *Equality of Mixed Partials*
If f_x, f_y, f_{xy}, and f_{yx} are continuous in some neighborhood of (x_0, y_0), then $f_{yx} = f_{xy}$ at (x_0, y_0).*

Evidently, the function $f(x, y)$ in Example 2 must not have met all of the conditions of Theorem 13.3.1; see Exercise 6.

Closure. Theorems analogous to Theorem 13.3.1 can be obtained for mixed partial derivatives of higher order as well, but we will not go into the details here because the vast majority of functions that are met in applications are sufficiently well behaved to ensure that the order of differentiation is immaterial. In that case, why discuss the matter at all? To bring this important theoretical question to light and, at the very least, to answer it for the important case of mixed partial derivatives of second order.

Further, Theorem 13.3.1 is representative of numerous theorems that deal with whether or not it is permissible to interchange the order of two *limit* operations. For instance, is

$$\iint_{\mathcal{D}} f(x, y)\, dxdy = \iint_{\mathcal{D}} f(x, y)\, dydx\,? \tag{8}$$

Is

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_{ij} = \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} a_{ij}\,? \tag{9}$$

Is

$$\frac{d}{dt}\int_a^b f(x, t)\, dx = \int_a^b \frac{\partial f}{\partial t}(x, t)\, dx\,? \tag{10}$$

That is, does it matter if we integrate first on x and then on y or vice versa? Or if we sum first on j and then on i or vice versa? Or if we first integrate with respect to x and then differentiate the result with respect to t or vice versa? Each of these operations is indeed a limit operation: integration involves the limit of a sequence of Riemann sums, an infinite sum involves the limit of a sequence of partial sums, and differentiation involves the limit of a difference quotient.

The general idea is that the interchange can be performed with impunity if the quantity involved [$f(x, y)$ in (8), a_{ij} in (9), $f(x, t)$ in (10)] is sufficiently well behaved. For instance, (10) holds if $f(x, t)$ and $\partial f(x, t)/\partial t$ are continuous over the

*For proof, see J. E. Marsden and A. J. Tromba, *Vector Calculus* (San Francisco: W. H. Freeman, 1976), pg. 120. They ask a bit more in their theorem, namely, that f_{xx} and f_{yy} be continuous too. But they do not use those conditions in their proof, which therefore holds for our theorem as well.

region of interest in the x, t plane. The relevant theorems can be found in textbooks on advanced calculus.*

EXERCISES 13.3

1. Evaluate f_x, f_y, f_{xy}, f_{yx}, f_{xx}, and f_{yy}. Are there any points (x, y) at which $f_{xy} \neq f_{yx}$? Explain.

(a) $f = x^3y^5$ (b) $f = y\sin(3x^2y)$
(c) $f = 1/(x^2 + y^2 + 1)$ (d) $f = x^3 + y^5$
(e) $f = (x^2 + y^2)^{1/2}$ (f) $f = (x^2 + y^2)^{3/2}$

2. Verify that $f_{xxy} = f_{xyx} = f_{yxx}$ everywhere in the x, y plane.

(a) $f = x^4y^3$ (b) $f = \cos(x^2y)$
(c) $f = \exp(x^2 + y^2)$ (d) $f = \sin(x - y^2)$

3. Recall that PDE is shorthand for partial differential equation. Show that

(a) $f = \ln(x^2 + y^2)$ satisfies the PDE $f_{xx} + f_{yy} = 0$, except at $(0, 0)$
(b) $f = \ln[(x - x_0)^2 + (y - y_0)^2]$ satisfies the PDE $f_{xx} + f_{yy} = 0$, except at (x_0, y_0)
(c) $f = r^n \sin n\theta$ satisfies the PDE $r^2 f_{rr} + r f_r + f_{\theta\theta} = 0$ for $n = 0, \pm 1, \pm 2, \ldots$ everywhere, except at $r = 0$ for the case where n is negative
(d) $f = r^n \cos n\theta$ satisfies the PDE $r^2 f_{rr} + r f_r + f_{\theta\theta} = 0$ for $n = 0, \pm 1, \pm 2, \ldots$ everywhere, except at $r = 0$ for the case where n is negative
(e) $f = 1/[(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2]$ satisfies the PDE $f_{xx} + f_{yy} + f_{zz} = 0$, except at (x_0, y_0, z_0)
(f) $f = \sin \kappa x \exp(-\kappa^2 t)$ satisfies the PDE $f_{xx} = f_t$ for all (x, t), where κ is any constant
(g) $f = \cos \kappa x \exp(-\kappa^2 t)$ satisfies the PDE $f_{xx} = f_t$ for all (x, t), where κ is any constant
(h) $f = \sin(x - ct)$ satisfies the PDE $c^2 f_{xx} = f_{tt}$ for all (x, t), where c is any constant

4. Determine the allowable values, if any, of the constant α such that $f = (x_1^2 + \cdots + x_n^2)^\alpha$ is a solution of the PDE

$$f_{x_1x_1} + \cdots + f_{x_nx_n} = 0 \tag{4.1}$$

[everywhere in $(x_1, \cdots, x_n)$ space, except possibly at the origin].

5. Let $f(x, y) = x$ for $y = 0$ and 0 for $y \neq 0$. Evaluate each of the following; if they do not exist, state that.

(a) $f_x(0, 0)$ (b) $f_x(0, 2)$ (c) $f_x(2, 0)$
(d) $f_x(3, 4)$ (e) $f_y(6, 0)$ (f) $f_y(0, 2)$
(g) $f_y(0, 0)$ (h) $f_{yx}(0, 0)$ (i) $f_{xy}(0, 0)$
(j) $f_{xy}(3, 0)$ (k) $f_{yxx}(3, 0)$ (l) $f_{xxy}(3, 0)$

6. In Example 2 we found that $f_{xy}(0, 0) \neq f_{yx}(0, 0)$. Evidently, the function $f(x, y)$ defined by (3) does not meet all the requirements stated in Theorem 13.3.1.

(a) Specifically, show that f_{xy} is discontinuous at $(0, 0)$. HINT: Evaluate $f_{xy}(x, y)$ for $(x, y) \neq (0, 0)$. Setting $y = \alpha x$ in that result, show that

$$f_{xy}(x, y) = \frac{1 + 9\alpha^2 - 9\alpha^4 - \alpha^6}{(1 + \alpha^2)^3} \tag{6.1}$$

on each ray $y = \alpha x$ so that the limit of $f_{xy}(x, y)$ as we approach the origin depends on the direction of approach. Thus, conclude that $\lim_{(x,y)\to(0,0)} f_{xy}(x, y) = f_{xy}(0, 0)$ cannot hold.
(b) Proceeding as in part (a), show that $f_{yx}(x, y)$ is discontinuous at $(0, 0)$.
(c) Show that $f_x(x, y)$ given by (4) *is* continuous at $(0, 0)$.
(d) Show that $f_y(x, y)$ given by (6) *is* continuous at $(0, 0)$.

7. If $f_x(x, y)$ and $f_y(x, y)$ exist at (x_0, y_0), does that result imply that f is continuous at (x_0, y_0)? Explain. HINT: Consider the function

$$f(x, y) = \begin{cases} 1, & x = 0 \text{ or } y = 0 \\ 0, & x \neq 0 \text{ and } y \neq 0. \end{cases}$$

*See, for instance, T. M. Apostol, *Mathematical Analysis* (Reading, MA: Addison-Wesley, 1957).

13.4 Composite Functions and Chain Differentiation

Let $f(x)$ be differentiable over some interval $X = \{x|\, a < x < b\}$. If $x = x(t)$ is, in turn, a differentiable function of the variable t over some interval $T = \{t|\, \alpha < t < \beta\}$, such that the value $x(t)$ is in X whenever t is in T, then we speak of $f(x) = f(x(t)) \equiv F(t)$ as a **composite function** of t. As noted in a first course in calculus, we can compute dF/dt by the **chain rule**,

$$\frac{dF}{dt} = \frac{df}{dx}\frac{dx}{dt}. \tag{1}$$

EXAMPLE 1. If $f(x) = \sin x$ and $x = t^2$ (over $-\infty < t < \infty$, say), then

$$f(x(t)) = \sin(t^2) \equiv F(t) \qquad \text{and} \qquad \frac{dF}{dt} = \frac{df}{dx}\frac{dx}{dt} = (\cos x)(2t) = 2t\cos(t^2). \quad \blacksquare$$

Just as the chain rule is indispensible in the calculus of functions of a single variable, its extension to functions of several variables is indispensible here. For the case $f(x(t), y(t))$, we have the following basic theorem.

THEOREM 13.4.1 *Chain Rule*
Let $f(x, y)$, $f_x(x, y)$ and $f_y(x, y)$ be continuous at each point of an open region $\mathcal{R}$ in the x, y plane. And let $x = x(t)$, $y = y(t)$ be differentiable functions of t over some open interval T on the t axis, such that the point $(x(t), y(t))$ is in $\mathcal{R}$ whenever t is in T. Then the composite function $f(x(t), y(t)) \equiv F(t)$ is a differentiable function of t for all t in T, and

$$\boxed{\frac{dF}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt}.} \tag{2}$$

Proof: Let t be in T, and let Δt be small enough so that $t + \Delta t$ is in T as well. Define

$$\Delta x \equiv x(t + \Delta t) - x(t) \qquad \text{and} \qquad \Delta y \equiv y(t + \Delta t) - y(t). \tag{3}$$

Then

$$\begin{aligned}
\Delta F &= F(t + \Delta t) - F(t) \\
&= f(x(t + \Delta t), y(t + \Delta t)) - f(x(t), y(t)) \\
&= f(x + \Delta x, y + \Delta y) - f(x, y) \\
&= [f(x, y + \Delta y) - f(x, y)] + [f(x + \Delta x, y + \Delta y) - f(x, y + \Delta y)] \\
&= f|_P^Q + f|_Q^R,
\end{aligned} \tag{4}$$

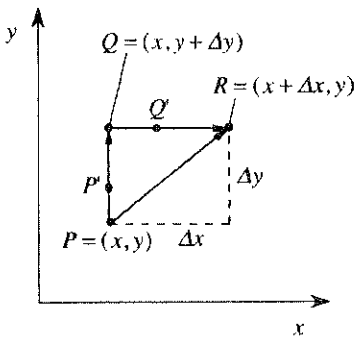


Figure 1. Following constant-coordinate lines.

where P, Q, R are the points shown in Fig. 1.* That is, instead of moving from P to R, we break the trip into two parts, from P to Q and then from Q to R, in order to follow constant-coordinate lines.

Next, we apply the mean value theorem of the differential calculus,† to each of the terms $f|_P^Q$ and $f|_Q^R$ in (4):

$$\Delta F = \left.\frac{\partial f}{\partial y}\right|_{P'} \Delta y + \left.\frac{\partial f}{\partial x}\right|_{Q'} \Delta x, \tag{5}$$

where Q' is some point between Q and R, and P' is some point between P and Q, as shown in Fig. 1. (This step is justified since it was assumed that f is continuous and that f_x and f_y exist within $\mathcal{R}$.)

Finally, divide (5) by Δt and let $\Delta t \to 0$. As $\Delta t \to 0$, we see from (3) that $\Delta x \to 0$ and $\Delta y \to 0$; hence $Q' \to P$ and $P' \to P$. Thus

$$\lim_{\Delta t \to 0} \frac{\Delta F}{\Delta t} = \lim_{\Delta t \to 0} \left(\left.\frac{\partial f}{\partial x}\right|_{Q'} \frac{\Delta x}{\Delta t} + \left.\frac{\partial f}{\partial y}\right|_{P'} \frac{\Delta y}{\Delta t} \right)$$

becomes‡

$$\frac{dF}{dt} = \left(\lim_{Q' \to P} \frac{\partial f}{\partial x} \right) \frac{dx}{dt} + \left(\lim_{P' \to P} \frac{\partial f}{\partial y} \right) \frac{dy}{dt}. \tag{6}$$

And if the partials f_x and f_y are continuous, as assumed, then the limit of f_x as $Q' \to P$ is equal to f_x *at* P, and similarly for f_y. Thus, (6) yields

$$\frac{dF}{dt} = f_x(x, y)\frac{dx}{dt} + f_y(x, y)\frac{dy}{dt}$$

as was to be shown. ■

Notice that the assumed continuity of the partials f_x, f_y is finally called upon in the last step of the proof. Our need for continuity of the partials, at the point in question, is quite reasonable since (2) is essentially an *interpolation* formula, wherein the change in F is computed as a linear combination of the rates of change in the perpendicular x and y directions, for interpolation is viable only if the quantities involved are continuous.

EXAMPLE 2. Let $r(x, y) = x^2 y - e^{2y}$, where

$$x = 3t^2, \qquad y = \sin t. \qquad (1 < t < 4)$$

*We use the symbol $(\ \ )|_P$ to denote $(\ \ )$ evaluated at P, and $(\ \ )|_P^Q$ to denote $(\ \ )|_Q - (\ \ )|_P$. Incidentally, observe that in the fourth line of (4) we have added and subtracted the quantity $f(x, y + \Delta y)$. This simple idea, the adding and subtracting of a quantity, is often useful.

†The **mean value theorem**: If $f(x)$ is continuous over $a \le x \le b$, and $f'(x)$ exists over $a < x < b$, then there is at least one point x_1 between a and b such that $f(b) - f(a) = f'(x_1)(b - a)$.

‡Recall from the calculus that $\lim(A + B) = \lim A + \lim B$ and $\lim(AB) = (\lim A)(\lim B)$ if $\lim A$ and $\lim B$ both exist. Thus $\lim(AB + CD) = \lim(AB) + \lim(CD) = (\lim A)(\lim B) + (\lim C)(\lim D)$ if the four limits on the right-hand side exist.

Then $r = r(x(t), y(t)) \equiv R(t)$, and

$$\begin{aligned}\frac{dR}{dt} &= \frac{\partial r}{\partial x}\frac{dx}{dt} + \frac{\partial r}{\partial y}\frac{dy}{dt}\\ &= (2xy)(6t) + (x^2 - 2e^{2y})(\cos t)\\ &= 36t^3 \sin t + (9t^4 - 2e^{2\sin t})\cos t\end{aligned}$$

for $1 < t < 4$. ∎

EXAMPLE 3. Let $f(u, v) = uv^2$, where

$$u = 3x - y, \qquad v = x^2 y.$$

Then $f(u, v) = f(u(x, y), v(x, y)) \equiv F(x, y)$. To compute $\partial F/\partial x$, say, use the chain rule (2):

$$\begin{aligned}\frac{\partial F}{\partial x} &= \frac{\partial f}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial x}\\ &= (v^2)(3) + (2uv)(2xy)\\ &= 3x^4y^2 + 4x^3y^2(3x - y)\\ &= 15x^4y^2 - 4x^3y^3. \qquad (7)\end{aligned}$$

COMMENT 1. The chain rule (2) applies even though u and v are functions of more than one variable (x and y) because y is held fixed when we compute $\partial F/\partial x$ so that $u(x, y)$ and $v(x, y)$ are, for the moment, considered as functions of x alone.

COMMENT 2. Is it clear which variables are held fixed when doing the various partial derivatives? There are two sets of variables, $\{x, y\}$ and $\{u, v\}$; $\partial/\partial x$ means the derivative with respect to x with all other variables in the x set (namely, y) fixed, $\partial/\partial u$ means the derivative with respect to u with all other variables in the u set (namely, v) fixed, and so on. ∎

Extension to more than two variables should be evident. Namely, if $f[x_1(t), \ldots, x_n(t)] \equiv F(t)$, then the chain rule (2) becomes

$$\boxed{\frac{dF}{dt} = \frac{\partial f}{\partial x_1}\frac{dx_1}{dt} + \frac{\partial f}{\partial x_2}\frac{dx_2}{dt} + \cdots + \frac{\partial f}{\partial x_n}\frac{dx_n}{dt},} \qquad (8)$$

subject to conditions analogous to those stated in Theorem 13.4.1.

Before closing this section we need to discuss the notation that we have been using. If f is a function of x and, in turn, x is a function of t, then we distinguish $f(x)$ from $f(x(t))$ by denoting the latter as $F(t)$. That is, it is necessary to introduce a new function name (F, say) because the f and F functions are different (in general). To illustrate, let $f(x) = \sin x$ and $x = t^2$ (as in Example 1), so then $F(t) = \sin(t^2)$. With $x = 3$, for instance, $f(3) = \sin 3$ whereas $F(3) = \sin 9$, and these values are not the same.

However, beware that in the engineering and science literature this notational distinction is usually not observed. For instance, if $f(x) = \sin x$ and $x = t^2$, then the writer will probably follow with $f(t) = \sin(t^2)$.* As explained above, the latter is incorrect and can cause problems. To illustrate, suppose f is a function of x, y, u, v, and that u, v are, in turn, functions of x and y. Then we should distinguish the function $f(x, y, u, v)$ of x, y, u, v from the function $F(x, y) = f(x, y, u(x, y), v(x, y))$ of x and y. The chain rule gives

$$\frac{\partial F}{\partial x} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial x}, \tag{9}$$

and similarly for $\partial F/\partial y$. Notice that if we had not introduced a separate function name, F, then the left side of (9) would be $\partial f/\partial x$ and we might cancel the two $\partial f/\partial x$ terms. That step would be incorrect since the one on the left is the partial derivative of $f(\underline{x}, y, u(\underline{x}, y), v(\underline{x}, y))$ with respect to all of the underlined x dependence (with y fixed), whereas the one on the right is the partial derivative of $f(\underline{x}, y, u, v)$ with respect to only the first of the four arguments (the underlined x).

Closure. The key point is the chain rule (2) and its extension to n variables, given by (8). We stress that the chain rule is, essentially, an interpolation formula and, as such, it requires continuity of the partial derivatives [$\partial f/\partial x$ and $\partial f/\partial y$ in (2), and $\partial f/\partial x_1, \ldots, \partial f/\partial x_n$ in (8)].

EXERCISES 13.4

1. Let $f(x, y) = \sin(x^4 + 3y)$, where $x = 5t$ and $y = t^2 + 1$, and denote $f(x(t), y(t)) = F(t)$. Evaluate dF/dt using the chain rule,

$$\frac{dF}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt}. \tag{1.1}$$

NOTE: Actually, (1.1) is not the end of the "chain differentiation story," for in computing $\partial f/\partial x$ we set $x^4 + 3y = u$, so that, again applying chain differentiation,

$$\frac{\partial f}{\partial x} = \frac{d}{du}(\sin u)\frac{\partial u}{\partial x} = \text{etc.},$$

and similarly for $\partial f/\partial y$.

2. Let $f(x, y) = e^{xy}$, and denote $f(x(t), y(t)) = F(t)$. Evaluate dF/dt in each case, using the chain rule.

(a) $x = t^2 - 1, \quad y = \sin 3t$
(b) $x = \sqrt{t+1}, \quad \cos t$
(c) $x = t^2, \quad y = 1/(t^2 + 1)$
(d) $x = \ln t, \quad y = t \quad (t > 0)$
(e) $x = \sin t, \quad y = \cos t$
(f) $x = 3t - 1, \quad y = 2t + 5$

3. Let $g(u, v) = \sqrt{u^2 - v}$ and denote $g(u(s), v(s)) = G(s)$. Evaluate dG/ds in each case, using the chain rule. Next, evaluate dG/ds directly; i.e., put $u(s)$ and $v(s)$ into $g(u, v)$, to obtain $G(s)$, and then compute dG/ds. Show that your two answers are the same.

(a) $u = \sin s, \quad v = \cos 6s$
(b) $u = 4s^2, \quad v = e^{-3s}$
(c) $u = s + 1, \quad v = s^3$
(d) $u = \sin 5s, \quad v = 4$
(e) $u = 3s, \quad v = s^2 + 2$
(f) $u = \sin s, \quad v = \cos^2 s$

*Similarly, engineering and science writers typically write $\int_0^x f(x)\,dx$ rather than introducing a new letter for the dummy variable of integration and writing $\int_0^x f(\xi)\,d\xi$. Though use of the former is not likely to lead to an incorrect result, it is important to keep in mind that the x's in $f(x)\,dx$ are not the same as the x in the upper limit. The upper limit x is a fixed endpoint, whereas the x in $f(x)\,dx$ is a dummy variable that varies over the interval of integration.

4. If $r = \sqrt{x^2(t) + y^2(t) + z^2(t)} = R(t)$ is the distance of a particle from the origin of a Cartesian x, y, z coordinate system and t is the time, use the chain rule to determine the radial speed dR/dt in each case, at $t = 2$.

(a) $x = t, \quad y = \sin t, \quad z = 4$
(b) $x = 2t, \quad y = t^2, \quad z = t^3 + 1$
(c) $x = \cos 2t, \quad y = 3, \quad z = \sin 2t$
(d) $x = e^t, \quad y = 4t, \quad z = e^{-t}$
(e) $x = \cos t, \quad y = \sin t, \quad z = 6t$
(f) $x = 3, \quad y = 1 + 2t, \quad z = t^2$

5. (a) Let the Cartesian coordinates of a fish be x, y, z. If the fish swims so that $x = 6t, \quad y = t + 2, \quad z = e^{-t}$, where t is the time, and the temperature distribution in the water is given by

$$T(x, y, z) = \left(5 + \frac{1}{x^2 + y^2 + z^2}\right) e^{-z},$$

determine the time rate of change of temperature experienced by the fish when $t = 0$.
(b) Repeat part (a), with

$$T(x, y, z, t) = \frac{60e^z(1 + 0.1 \sin t)}{x^2 + y^2 + 2}.$$

6. (*Euler's theorem*) $f(x_1, \ldots, x_n)$ is said to be **homogeneous of degree** k if

$$f(\lambda x_1, \ldots, \lambda x_n) = \lambda^k f(x_1, \ldots, x_n). \tag{6.1}$$

(a) In each case, show whether the function is homogeneous and, if it is, indicate its degree: $f = x^2 + 3xy$, $g = \ln(x^2 + y^2)$, $h = (x^2 - xy)/(2x + y)$, and $p = x^3 e^{x/2y}$. (Assume that $x^2 + y^2 \neq 0$ in g, that $2x + y \neq 0$ in h, and that $y \neq 0$ in p.)
(b) If $f(u_1, \ldots, u_m)$ is a homogeneous function of $u_1, \ldots, u_m$ of degree p, and $u_1(x_1, \ldots, x_n), \ldots, u_m(x_1, \ldots, x_n)$ are homogeneous functions of $x_1, \ldots, x_n$ of degree q, show whether f is necessarily a homogeneous function of $x_1, \ldots, x_n$. If so, of what degree?
(c) Show that if $f(x, y, z)$ is homogeneous of degree k, and has partial derivatives $\partial f/\partial x$, $\partial f/\partial y$, $\partial f/\partial z$, then these partial derivatives are homogeneous of degree $k - 1$.
(d) (*Euler's theorem on homogeneous functions*) Prove **Euler's theorem**, that if $f(x_1, \ldots, x_n)$ is homogeneous of degree k and has continuous first-order partial derivatives, then

$$x_1 \frac{\partial f}{\partial x_1} + x_2 \frac{\partial f}{\partial x_2} + \cdots + x_n \frac{\partial f}{\partial x_n} = kf. \tag{6.2}$$

HINT: Differentiate equation (6.1) with respect to λ, then set $\lambda = 1$.

7. Verify Euler's theorem [(6.2) in Exercise 6] for these cases.

(a) $f(x, y) = \sqrt{x^4 + 2y^4} \sin(3x/y), \quad y \neq 0$
(b) $f(x, y) = (x^3 - 2x^2y + 5y^3)/(x^2 + y^2), \quad x^2 + y^2 \neq 0$

8. Show that $u = f(x + ct) + g(x - ct)$ satisfies the partial differential equation $c^2 u_{xx} - y_{tt} = 0$ (the *wave equation*), where c is a constant and f and g are arbitrary twice-differentiable functions.

9. The differential equation $xy'' + y' + xy = 0$, known as *Bessel's equation of order zero*, has the general solution $y = AJ_0(x) + BY_0(x)$, where J_0 is the *Bessel function of the first kind and order zero*, and Y_0 is the *Bessel function of the second kind and order zero*.

(a) Solve $xy'' + y' + k^2xy = 0$ in terms of Bessel functions. HINT: Set $x = \alpha t$ [and $y(x) \equiv u(t)$], and choose α so that the differential equation on $u(t)$ is the Bessel equation.
(b) Solve $xy'' + y' + k^2y = 0$ in terms of Bessel functions. HINT: Set $x = \alpha t^\beta$ [and $y(x) = y(\alpha t^\beta) \equiv u(t)$], and choose α, β so that the new differential equation is the Bessel equation.

13.5 Taylor's Formula and Mean Value Theorem

Just as the Taylor series of a function of a single real variable is indispensible in applied mathematics, so is the Taylor series of a function of several real variables. We expect that you have already studied the former in the calculus but not the latter.

In Section 13.5.1 we review Taylor's formula, Taylor series, and the mean value theorem, and in Section 13.5.2 we extend those results to functions of several real variables.

13.5.1. Taylor's formula and Taylor series for $f(x)$. To develop Taylor's formula about an initial point a, we begin with the identity

$$\int_a^x f'(x)\,dx = f(x) - f(a), \tag{1}$$

where in this discussion it will prove more convenient to forego using different letters for the dummy variables of integration. Solving (1) for $f(x)$,

$$f(x) = f(a) + \int_a^x f'(x)\,dx. \tag{2}$$

Just as (2) holds for $f(x)$, it holds if we replace $f(x)$ by $f'(x)$:

$$f'(x) = f'(a) + \int_a^x f''(x)\,dx. \tag{3}$$

[We obtained (3) by changing f to f' in (2), not by differentiating (2).] Putting this expression for $f'(x)$ into the integral in (2) gives

$$\begin{aligned} f(x) &= f(a) + \int_a^x \left[f'(a) + \int_a^x f''(x)\,dx \right] dx \\ &= f(a) + f'(a)(x-a) + \int_a^x \int_a^x f''(x)\,dx\,dx. \end{aligned} \tag{4}$$

Next, replacing f' in (3) by f'' gives

$$f''(x) = f''(a) + \int_a^x f'''(x)\,dx \tag{5}$$

and putting that expression into the integral in (4) gives

$$\begin{aligned} f(x) &= f(a) + f'(a)(x-a) + \int_a^x \int_a^x \left[f''(a) + \int_a^x f'''(x)\,dx \right] dx\,dx \\ &= f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 \\ &\qquad + \int_a^x \int_a^x \int_a^x f'''(x)\,dx\,dx\,dx. \end{aligned} \tag{6}$$

Repeating this process (assuming that f is sufficiently differentiable*) gives

$$\boxed{\begin{aligned} f(x) = f(a) &+ f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 \\ &+ \cdots + \frac{f^{(n-1)}(a)}{(n-1)!}(x-a)^{n-1} + R_n(x), \end{aligned}} \tag{7a}$$

*For instance, $\sin x$ is infinitely differentiable (i.e., it admits derivatives of all orders) whereas $x^3 H(x)$ (where H is the Heaviside function) is only twice differentiable (on any interval containing $x = 0$).

where the **remainder term** is

$$R_n(x) = \int_a^x \cdots \int_a^x f^{(n)}(x)\,(dx)^n \tag{7b}$$

and $(dx)^n$ denotes $dx \cdots dx$, n times; (7a) is known as **Taylor's formula**, with the remainder term given in **integral form** by (7b).

To obtain an alternative form for R_n, suppose that $f^{(n)}(x)$ has a minimum value m and a maximum value M on $[a, x]$.* Surely, then,

$$\int_a^x \cdots \int_a^x m\,(dx)^n \le R_n(x) \le \int_a^x \cdots \int_a^x M\,(dx)^n \tag{8}$$

or, upon carrying out the integrations,

$$\frac{m}{n!}(x-a)^n \le R_n(x) \le \frac{M}{n!}(x-a)^n. \tag{9}$$

If we assume that $f^{(n)}(x)$ is continuous on $[a, x]$, then (it can be shown that) it must take on all values, from its minimum m to its maximum M, over the interval. It therefore follows from (9) that we must be able to express

$$R_n(x) = \frac{f^{(n)}(\xi)}{n!}(x-a)^n \tag{10}$$

for some suitable point ξ in $[a, x]$; (10) gives the **Lagrange form** of R_n, which is often more convenient than the integral form (7b).

If we put (10) into (7a) it appears that the resulting right-hand side is an $(n-1)$th-degree polynomial representation of $f(x)$. Something seems amiss because we know that functions such as $f(x) = e^x$, $\sin x$, and the Bessel function $J_0(x)$ cannot be expressed as finite-degree polynomials. The catch is that ξ, in (10), is not merely a constant, it is a function of the endpoint x. For instance, since ξ is somewhere in $[a, x]$, then it follows that if we let x approach a, then ξ must approach a too; thus, ξ depends upon x. That is, (10) really means

$$R_n(x) = \frac{f^{(n)}(\xi(x))}{n!}(x-a)^n, \tag{11}$$

and (in general) we do not know $\xi(x)$, except that it is somewhere in $[a, x]$. Thus, if we put (10) into (7a) then the resulting right-hand side is not really a polynomial.

Nonetheless, Taylor's formula with Lagrange remainder is valuable because we can write, from (7a),

$$f(x) \approx f(a) + f'(a)(x-a) + \cdots + \frac{f^{(n-1)}(a)}{(n-1)!}(x-a)^{n-1}, \tag{12}$$

*Recall that $[\alpha, \beta]$ denotes the **closed interval** $\alpha \le x \le \beta$. Actually, x can be to the right *or* left of the point a, so when we write $[a, x]$ we will mean $[a, x]$ if $x > a$, and $[x, a]$ if $x < a$.

where the error in the approximation is $R_n(x)$, and we can use the form (10) to "bound" that error. Thus, the point is that Taylor's formula (7a) is about *approximation*, the approximation of a given function $f(x)$ by a finite-degree polynomial, with $R_n(x)$ enabling us to obtain a bound on the error thereby incurred.

EXAMPLE 1. Use (7a) to approximate the function $f(x) = e^{-x}$ over $0.7 \le x \le 1.3$, by a third-degree polynomial, and obtain a bound on the error. We obtain, from (7a),

$$e^{-x} = e^{-1} - \frac{e^{-1}}{1!}(x-1) + \frac{e^{-1}}{2!}(x-1)^2 - \frac{e^{-1}}{3!}(x-1)^3 + R_4(x) \tag{13}$$

or

$$e^{-x} \approx e^{-1} - \frac{e^{-1}}{1!}(x-1) + \frac{e^{-1}}{2!}(x-1)^2 - \frac{e^{-1}}{3!}(x-1)^3, \tag{14}$$

where the error incurred in (14) is

$$R_4(x) = \frac{e^{-\xi}}{4!}(x-1)^4 \tag{15}$$

for some ξ in the interval [0.7,1.3]. Even without knowing ξ we can bound $R_4(x)$ as

$$|R_4(x)| \le \frac{e^{-0.7}}{4!}(0.3)^4 = 0.000168 \tag{16}$$

because the greatest value of $e^{-\xi}$ on [0.7,1.3] occurs if $\xi = 0.7$, and the greatest value of $(x-1)^4$ occurs if $x = 0.7$ or 1.3 and is $(0.3)^4$.

Thus, if we confine x to the interval $0.7 \le x \le 1.3$, then the error incurred by the approximation (14) is at most ± 0.000168. In fact, calculation reveals that the actual absolute magnitude of the error is 0.000132 at $x = 0.7$, 0 at $x = 1$ and 0.000117 at $x = 1.3$. ∎

Suppose we let $n \to \infty$ in (12). The infinite series that results is called the **Taylor series of f**, about the point a, and we denote that series as TS $f|_a$:

$$\boxed{\text{TS } f|_a \equiv \sum_{j=0}^{\infty} \frac{f^{(j)}(a)}{j!}(x-a)^j,} \tag{17}$$

where $(x-a)^0 \equiv 1$ even if $x = a$, and $0! \equiv 1$ so that the first term of the series is inevitably $f(a)$, as in (12). For the Taylor series of f to exist, about $x = a$, we need f to be infinitely differentiable at $x = a$ so that the coefficients $f^{(j)}(a)/j!$ exist (for $j = 0, 1, 2, \ldots$).

However, for TS $f|_a$ to be *useful* we need two things: we need the series to *converge* on some x interval, and we need the sum function (i.e., the function to which the series converges) to *equal* $f(x)$ on some x interval I. Then we say that TS $f|_a$ **represents** f on I.

Why do we need to explicitly ask that the sum function equal $f(x)$? Is it possible for TS $f|_a$ to converge, on some interval containing $x = a$, and yet not converge to $f(x)$? Let us see.

EXAMPLE 2. Consider the Taylor series of

$$f(x) = \begin{cases} e^{-1/x^2}, & x \neq 0 \\ 0, & x = 0 \end{cases} \tag{18}$$

about $x = 0$.* (See Fig. 1.) It can be shown (Exercise 7) that $f'(0) = f''(0) = \cdots = 0$ so that

$$\text{TS } f|_0 = 0 + 0x + 0x^2 + \cdots. \tag{19}$$

Surely the latter series converges for all x, but its sum function is identically zero, not the function $f(x)$ defined in (18). Of course the sum function and $f(x)$ do agree at the point of expansion, $x = 0$, as will always be true, but that is not good enough, and we conclude that the Taylor series of f, about $x = 0$, does not represent f on any interval I even though it converges for all x!

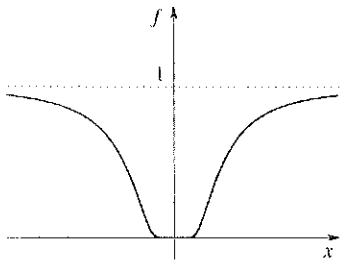


Figure 1. Graph of f.

COMMENT. The failure of TS $f|_0$, as a representation of f, would be reasonable if f were badly behaved at $x = 0$, the point of expansion, but f appears (from Fig. 1) to be well behaved there. To expose the source of the difficulty, we need to examine the behavior of

$$f(z) = \begin{cases} e^{-1/z^2}, & z \neq 0 \\ 0, & z = 0 \end{cases}$$

in the complex $z = x + iy$ plane. As we approach the origin along the real axis (on which $z = x + i0 = x$)

$$\lim_{z \to 0} f(z) = \lim_{x \to 0} e^{-1/x^2} = 0$$

(as can also be seen from Fig. 1), but suppose we approach the origin along the imaginary axis (on which $z = 0 + iy = iy$). In that case

$$\lim_{z \to 0} f(z) = \lim_{y \to 0} e^{-1/(iy)^2} = \lim_{y \to 0} e^{+1/y^2} = \infty,$$

so $f(z)$ is indeed badly behaved at the origin, although that difficulty cannot be observed by looking only along the real axis. The moral of the story is that to fully understand the theory of Taylor series for functions $f(x)$ of a real variable x one must study the theory of Taylor series for functions $f(z)$ of a *complex variable* z. We will do that, but not until Chapter 24. ∎

Meanwhile, we can say the following. If TS $f|_a$ is to represent f, then

$$f(x) = \text{TS } f|_a = \sum_{j=0}^{\infty} \frac{f^{(j)}(a)}{j!}(x-a)^j = \lim_{N \to \infty} \sum_{j=0}^{N} \frac{f^{(j)}(a)}{j!}(x-a)^j \tag{20}$$

*Why do we need to define $f(0) = 0$ separately? Does not $e^{-1/x^2} = e^{-1/0} = e^{-\infty} = 0$ at $x = 0$? No, $e^{-1/0}$ is simply undefined because $1/0$ is undefined. Of course we could define $f(0)$ to be any value we wish but, in making up this example, we chose 0 so that the resulting function would be continuous at $x = 0$.

or, equivalently,

$$\lim_{N\to\infty}\left[f(x)-\sum_{j=0}^{N}\frac{f^{(j)}(a)}{j!}(x-a)^j\right]=0. \tag{21}$$

But, from (7a), we see that (21) is the same as $\lim_{N\to\infty} R_{N+1}(x)=0$ or, equivalently,

$$\lim_{N\to\infty} R_N(x)=0. \tag{22}$$

The condition (22), on some interval I, is both necessary and sufficient for the Taylor series of f to represent f on I. Not only was the condition (22) not satisfied in Example 2, but (7a) became

$$f(x)=0+0+\cdots+0+R_n(x),$$

so that $R_n(x)$ did not tend to zero as $n\to\infty$, it was actually equal to $f(x)$ for each n.

Let us summarize. If we expand a given function $f(x)$ in a Taylor series about some point $x=a$ in the hope that that series will represent f over some x interval, then it does not suffice to test the series for convergence, because it is possible for the series to converge but for $R_n(x)$ not to tend to zero as $n\to\infty$. For if the series converges, but not to $f(x)$, then it does not provide the desired representation of f, as occurred in Example 2. Rather, to show that the Taylor series of a given function f represents f, we need to show that $R_n(x)\to 0$ over some x interval.

EXAMPLE 3. Consider the Taylor series of $f(x)=e^{-x}$ about $x=0$,

$$\mathrm{TS}\, e^{-x}|_0=1-x+\frac{1}{2!}x^2-\frac{1}{3!}x^3+\cdots. \tag{23}$$

Does that series represent e^{-x} over some x interval so that we can write

$$e^{-x}=1-x+\frac{1}{2!}x^2-\frac{1}{3!}x^3+\cdots? \tag{24}$$

The ratio test (Section 4.2) shows that (23) converges for all x but, as emphasized above, that convergence does not suffice to establish the equality in (24). Everything hinges on the remainder term

$$R_n(x)=\frac{f^{(n)}(\xi)}{n!}(x-a)^n=\frac{(-1)^n e^{-\xi}}{n!}x^n. \tag{25}$$

Since ξ is some point in $[0,x]$ if $x>0$, and in $[x,0]$ if $x<0$, the $e^{-\xi}$ in (25) is at most $e^{|x|}$, so

$$|R_n(x)|\le\frac{e^{|x|}}{n!}|x|^n. \tag{26}$$

Since $|x|^n/n!\to 0$ as $n\to\infty$, for each fixed value of x (no matter how large), it follows that $R_n(x)\to 0$ for every (finite) value of x and, hence, that the equality holds in (24) for all x, for $-\infty<x<\infty$. (See Exercise 8.)

Further, let us explore the convergence of the right-hand side of (24) to e^{-x} in graphical terms. The partial sums of the series are $s_1(x) = 1$, $s_2(x) = 1 - x$, $s_3(x) = 1 - x + x^2/2$, and so on, and the first few are shown in Fig. 2. The first partial sum $s_1(x)$ matches the value of e^{-x} at $x = 0$ (i.e., the point of expansion); $s_2(x)$ matches both the value of e^{-x} and its slope at $x = 0$; $s_3(x)$ matches the value of e^{-x}, its slope, and its second derivative at $x = 0$; and so on. It's true that for each fixed n the discrepancy between $s_n(x)$ and e^{-x} becomes infinite as $x \to \infty$, but that fact is irrelevant insofar as convergence is concerned because by the convergence of $s_n(x)$ to e^{-x} we mean that $s_n(x)$ tends to e^{-x}, at a *fixed value of* x, as $n \to \infty$.

Just as a rodeo rider endeavors to stay on the horse, we can imagine the polynomials $s_1(x)$, $s_2(x)$, ... endeavoring to "stay on the function:" $s_2(x)$ does a better job than $s_1(x)$, $s_3(x)$ does a better job than $s_2(x)$, and so on, but eventually , as x increases, they all "bite the dust." ∎

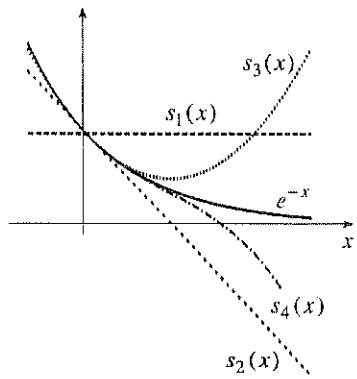


Figure 2. Convergence of the sequence of partial sums to e^{-x}.

The preceding example notwithstanding, it is generally not practical to determine whether or not the Taylor series of a given function f converges to $f(x)$ by determining whether or not $R_n(x) \to 0$ as $n \to \infty$. The reason is that in most examples $R_n(x)$ becomes unwieldy as n increases. Not so in Example 3 because $f^{(n)}(\xi)$ is simply equal to $(-1)^n e^{-\xi}$ if $f(x)$ is e^{-x}, but in general $f^{(n)}(\xi)$, and hence $R_n(x)$, is too unwieldy to bound.

However, functions such as the one in Example 2, whose Taylor series converge, but not to the given function, are rarely encountered in applications. Practically speaking, then, one can merely test the Taylor series of a given function $f(x)$, about a given point $x = a$, for convergence, using Theorem 4.2.2, and assuming (i.e., hoping) that if it converges in $|x - a| < R$ (with the radius of convergence R determined according to the theorem) then it converges to $f(x)$ in that interval and represents $f(x)$ in that interval. This approach is flawed, as discussed above, but should suffice until Taylor series are clarified completely in Chapter 24.

One final application:

EXAMPLE 4. Consider the representation of $f(x) = 1/(1 + x^2)$ by its Taylor series about $x = 0$. To generate the Taylor series of f about $x = 0$ we can use the Taylor series formula (17), but it is easier to recall the geometric series

$$\frac{1}{1-t} = 1 + t + t^2 + \cdots, \tag{27}$$

which holds for $|t| < 1$ (see Exercises 5 and 6 in Section 4.2). That is, identifying t as $-x^2$, we have, from (27),

$$\begin{aligned} f(x) = \frac{1}{1+x^2} = \frac{1}{1-(-x^2)} &= 1 + (-x^2) + (-x^2)^2 + (-x^2)^3 + \cdots \\ &= 1 - x^2 + x^4 - x^6 + \cdots \end{aligned} \tag{28}$$

for $|-x^2| < 1$, that is, for $|x| < 1$. From the graph of f (Fig. 3), we wonder why (28) converges only in $|x| < 1$ since $f(x) = 1/(1 + x^2)$ is evidently very well behaved for all

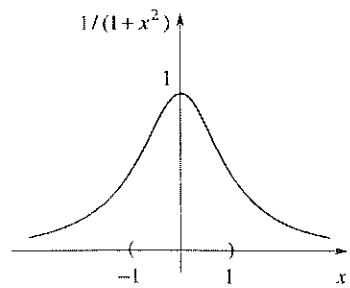


Figure 3. Graph of $f(x) = 1/(1 + x^2)$.

x. As in the case of Example 2, to understand this apparent paradox we need to consider not $f(x) = 1/(1+x^2)$ along the x axis, but $f(z) = 1/(1+z^2)$ in the complex z plane. Specifically, and we don't expect this comment to be completely clear until you study Chapter 24, $f(z) = 1/(1+z^2)$ is singular (becomes infinite) at $z = +i$ and at $z = -i$. Complex Taylor series converge in *disks*, and the Taylor series

$$f(z) = 1 - z^2 + z^4 - z^6 + \cdots \tag{29}$$

converges in the unit disk (Fig. 4), its radius being limited by the singularities of $f(z)$ at $\pm i$. The intersection of the *disk* $|z| < 1$ with the real axis gives the *interval* of convergence $|x| < 1$ of (28).

Thus, the interval of convergence on the x axis, of the Taylor series of a function $f(x)$, may be limited by singularities of $f(z)$ off of the real axis in the complex z plane. ∎

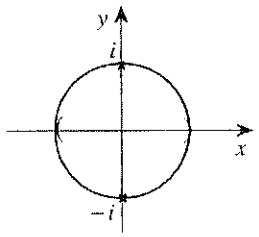


Figure 4. Disk of convergence of (29).

There is one last point before we extend these ideas to functions of more than one variable. Namely, for the special case where $n = 1$ and where the remainder term $R_1(x)$ is expressed in the Lagrange form (10), Taylor's formula (7a) gives $f(x) = f(a) + f'(\xi)(x-a)$, which result is known as the **mean value theorem.** Let us state that theorem explicitly, for reference.

THEOREM 13.5.1 *Mean Value Theorem for f(x)*
Let f be continuous on the closed interval $[a, x]$* and differentiable on the open interval (a, x). Then there exists a number ξ in (a, x) such that

$$\boxed{f(x) = f(a) + f'(\xi)(x-a).} \tag{30}$$

13.5.2. Extension to functions of more than one variable. You may recall that we have already encountered Taylor series of functions of more than one variable in Example 2 in Section 11.4. We suggest that you review that example either now or when you have completed the present section.

The basic idea behind the Taylor series TS $f|_a$ of a function $f(x)$ is extrapolation: knowing "all about" $f(x)$ at $x = a$ [i.e., knowing $f(a)$, $f'(a)$, $f''(a)$, ...], can we extrapolate that knowledge and predict what $f(x)$ will be at some other point x? Consider the extrapolation problem in *two* variables. That is, let $f(x, y)$ be defined in an open region $\mathcal{R}$ in the x, y plane, and suppose that all the values of f, f_x, f_y, f_{xx}, f_{xy}, f_{yy}, ..., up to nth order are known at a point (a, b) in $\mathcal{R}$. Can we, from these data, extrapolate from (a, b) and determine the value of f at some other point (x, y) in $\mathcal{R}$?

Often, in mathematics, new ideas are introduced as limits or extensions of old ones. Accordingly, let us try to reduce the stated problem to one involving a single independent variable, since that case is already in hand. Thus, draw a

*As noted earlier, we mean $[a, x]$ if $x > a$, and $[x, a]$ if $x < a$.

straight line from the initial point (a, b) to any desired final point (x_0, y_0), as in Fig. 5, and regard (x, y) as a variable point along that line. The straight line may be parametrized as

$$x = a + (x_0 - a)t, \tag{31a}$$
$$y = b + (y_0 - b)t, \tag{31b}$$

where t is the parameter; for $t = 0$ the point (x, y) coincides with (a, b), and for $t = 1$ it coincides with (x_0, y_0). Then, and this is the point,

$$f(x, y) = f\left(a + (x_0 - a)t, b + (y_0 - b)t\right) \equiv F(t) \tag{32}$$

is a function of the *single variable* t, for $0 \le t \le 1$, so if F is sufficiently differentiable we can write Taylor's formula,

$$F(t) = F(0) + F'(0)t + \frac{F''(0)}{2!}t^2 + \cdots + \frac{F^{(n-1)}(0)}{(n-1)!}t^{n-1} + R_n(t), \tag{33a}$$

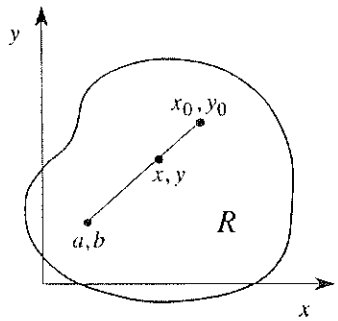


Figure 5. Points of interest.

with Lagrange remainder

$$R_n(t) = \frac{F^{(n)}(\tau)}{n!}t^{n+1}, \tag{33b}$$

where τ is some point in $0 \le \tau \le 1$. Here we have expanded about $t = 0$ because $t = 0$ corresponds to $x = a$, $y = b$, which is the point in the x, y plane about which we are expanding.

We need to evaluate the coefficients $F(0)$, $F'(0)$, ..., in (33). By chain differentiation,

$$\frac{d}{dt} = \frac{\partial}{\partial x}\frac{dx}{dt} + \frac{\partial}{\partial y}\frac{dy}{dt} = (x_0 - a)\frac{\partial}{\partial x} + (y_0 - b)\frac{\partial}{\partial y} \equiv D \tag{34}$$

so

$$F'(0) = \frac{d}{dt}F(t)|_{t=0} = Df(x, y)|_{a,b}, \tag{35a}$$
$$F''(0) = \frac{d}{dt}\frac{d}{dt}F(t)|_{t=0} = D^2 f(x, y)|_{a,b}, \tag{35b}$$

and so on. Putting these expressions into (33) and setting $t = 1$, since $F(1) = f(x_0, y_0)$, gives

$$f(x_0, y_0) = f(a, b) + \frac{1}{1!}Df|_{a,b} + \cdots + \frac{1}{(n-1)!}D^{n-1}f|_{a,b} + R_n(t) \tag{36a}$$

with

$$R_n(t) = \frac{1}{n!}D^n f|_{\xi,\eta}, \tag{36b}$$

where ξ, η are given by the right-hand side of (31) when $t = \tau$ (Fig. 6). Equation (36) is **Taylor's formula** in two variables, with Lagrange remainder. It holds if $f(x, y)$ has continuous partial derivatives (since we used chain differentiation to derive it), through nth order in $\mathcal{R}$, and if the straight line from (a, b) to (x_0, y_0) lies entirely within $\mathcal{R}$.

Remember that the D's in (36) are defined by (34), so

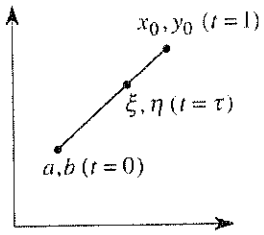


Figure 6. The point (ξ, η).

$$Df = [(x_0 - a)\frac{\partial}{\partial x} + (y_0 - b)\frac{\partial}{\partial y}]f = (x_0 - a)f_x + (y_0 - b)f_y,$$

$$\begin{aligned} D^2 f = D(Df) &= [(x_0 - a)\frac{\partial}{\partial x} + (y_0 - b)\frac{\partial}{\partial y}][(x_0 - a)f_x + (y_0 - b)f_y] \\ &= (x_0 - a)^2 f_{xx} + 2(x_0 - a)(y_0 - b)f_{xy} + (y_0 - b)^2 f_{yy}, \end{aligned} \tag{37}$$

and so on. With Df, $D^2 f$, ... in hand, we can write out (36) for any desired n. With $n = 1$, for example, we obtain, with the help of (37),

$$\begin{aligned} f(x_0, y_0) &= f(a, b) + Df|_{\xi,\eta} \\ &= f(a, b) + f_x(\xi, \eta)(x_0 - a) + f_y(\xi, \eta)(y_0 - b). \end{aligned} \tag{38}$$

Notice that x, y were "transition variables" that carried us along the straight line from (a, b) to (x_0, y_0). In the final result, such as (38), x and y no longer appear, only the initial point (a, b) and the final point (x_0, y_0) appear. Since we no longer need to distinguish between x, y and x_0, y_0, we can, for notational simplicity, drop the subscripted zeros and re-express (38) as

$$\boxed{f(x, y) = f(a, b) + f_x(\xi, \eta)(x - a) + f_y(\xi, \eta)(y - b).} \tag{39}$$

The latter is the two-variable version of the **mean value theorem**, which we state, for reference.

THEOREM 13.5.2 *Mean Value Theorem for f(x,y)*
Let $f(x, y)$ and its first-order partial derivatives be continuous in an open region $\mathcal{R}$, and let (a, b) and (x, y) be points in $\mathcal{R}$ such that the straight line joining these points lies entirely within $\mathcal{R}$. Then there exists a point (ξ, η) on that line, between the endpoints, such that

$$f(x, y) = f(a, b) + f_x(\xi, \eta)(x - a) + f_y(\xi, \eta)(y - b). \tag{40}$$

Further, we define the **Taylor series** of $f(x, y)$ about (a, b) as

$$\text{TS}\, f|_{a,b} = f(a, b) + \frac{1}{1!}Df|_{a,b} + \frac{1}{2!}D^2 f|_{a,b} + \cdots$$

or, with the help of (37) (with x_0 changed to x and y_0 changed to y),

$$\boxed{\begin{aligned}\mathrm{TS}\, f|_{a,b} = f(a,b) &+ \frac{1}{1!}\left[f_x(a,b)(x-a) + f_y(a,b)(y-b)\right] \\ &+ \frac{1}{2!}\left[f_{xx}(a,b)(x-a)^2 + 2f_{xy}(a,b)(x-a)(y-b)\right. \\ &\left. + f_{yy}(a,b)(y-b)^2\right] + \cdots .\end{aligned}} \tag{41}$$

We say that $(x-a)^m(y-b)^n$ is of **order** $m+n$, so the dots in (41) denote terms of third order and higher.

EXAMPLE 5. Derive the Taylor series of $f(x,y) = e^{xy}$ about the point $(1,2)$, up to and including second-order terms. Then $f_x = ye^{xy}$, $f_y = xe^{xy}$, $f_{xx} = y^2e^{xy}$, $f_{xy} = (1+xy)e^{xy}$, and $f_{yy} = x^2e^{xy}$, so (41) gives

$$\begin{aligned}\mathrm{TS}\, e^{xy}|_{1,2} &= e^2 + 2e^2(x-1) + e^2(y-2) \\ &\quad + 2e^2(x-1)^2 + 3e^2(x-1)(y-2) + \frac{e^2}{2}(y-2)^2 + \cdots\end{aligned}$$

up to terms of second order. ∎

Extension of these ideas to functions of more than two variables should be straightforward and is left to the exercises.

Before closing this section, let us introduce some notation that will prove useful in what follows. Specifically, if all of the partial derivatives of f, through nth order, are continuous in a region $\mathcal{R}$, then f is said to be **of class C^n** in $\mathcal{R}$.* For example, $f(x,y)$ is of class C^1 in $\mathcal{R}$ if f_x and f_y are continuous in $\mathcal{R}$, and it is of class C^2 in $\mathcal{R}$ if f_x, f_y, f_{xx}, f_{xy}, and f_{yy} are continuous in $\mathcal{R}$. Generally, one omits the words "of class" and simply says that f is C^1 in $\mathcal{R}$, C^2 in $\mathcal{R}$, and so on.

Closure. We begin this section by reviewing and extending the theory of Taylor's formula, Taylor series, and the mean value theorem for functions f of a single variable x. *Taylor's formula* has the advantage that it is the sum of a finite number of terms so convergence of infinite series is not an issue. However, the remainder term, be it in integral form or in Lagrange form, is not known explicitly; for instance, the ξ in (10) is not known. Nevertheless, we can drop the remainder term and *approximate* $f(x)$ by the *polynomial*

$$f(x) \approx f(a) + f'(a)(x-a) + \cdots + \frac{f^{(n-1)}(a)}{(n-1)!}(x-a)^{n-1}$$

*It can be shown that the continuity of the first-order partial derivatives of f imply the continuity of f. Thus, it would be equivalent to say that if f *and* all of the partial derivatives of f, through nth order, are continuous in $\mathcal{R}$, then f is C^n in $\mathcal{R}$.

and use the remainder term $R_n(x)$ to bound the error. *Taylor series*, on the other hand, is an infinite series, so we need to be concerned with whether or not it converges and, if it does, we need to know the x interval over which the convergence is achieved. Further, we emphasized that establishing the convergence of TS $f|_a$ is not the end of the story because it is possible for TS $f|_a$ to converge on an interval I, but not to *represent* f on I. Rather, for TS $f|_a$ to represent f on I it is necessary and sufficient that $R_n(x) \to 0$ on I as $n \to \infty$. Since it is generally difficult to show that $R_n(x) \to 0$ on some interval we are in an awkward position. However, when we study Chapter 24 on complex variable theory we will learn how to establish the interval over which TS $f|_a$ represents $f(x)$, simply by inspection of the function $f(z)$ (where $z = x + iy$). Until then, we suggested merely checking TS $f|_a$ for convergence and not worrying about R_n, on the grounds that in practical applications one rarely encounters functions for which TS $f|_a$ converges without R_n tending to zero as $n \to \infty$.

Computer software. For functions of a single variable, see the computer section at the end of Section 4.2. For multivariable Taylor expansions the *Maple* command is **mtaylor**. For instance, to expand e^{2xy} about $x = 1$, $y = 4$, through terms of second order, enter

readlib(mtaylor):

Then,

mtaylor(exp(2 * x * y), [x = 1, y = 4], 3);

and obtain the result

$$e^8 + 8e^8(x-1) + 2e^8(y-4) + 32e^8(x-1)^2 + 18e^8(x-1)(y-4) + 2e^8(y-4)^2$$

EXERCISES 13.5

1. Expand the given function about the indicated point a, through third-order terms. NOTE: $(x-a)^n$ is of nth order.

(a) e^{-2x}, $a = 0$ (b) e^{-2x}, $a = 5$
(c) e^{-2x}, $a = -3$ (d) $\ln x$, $a = 2$
(e) $1/(1+x^2)$, $a = 1$ (f) $1/(1+x^2)$, $a = -1$
(g) $\sin x$, $a = 2$ (h) $\cos 2x$, $a = \pi$
(i) $x(x-1)^2$, $a = 1$ (j) $x^3(x^4-1)+5$, $a = 0$

2. Obtain the first four nonvanishing terms in the Taylor series of the given function about $x = 0$.

(a) $1/(1-x^5)$ (b) $1/(2+x^{10})$
(c) $\sin x^{20}$ (d) $\cos x^{20}$

3. Using computer software, generate the Taylor series of the given function, about the indicated point a, through fourth-order terms, and obtain plots of the first five partial sums [$s_1(x)$ through $s_5(x)$], as well as the given function, over the indicated interval.

(a) e^x, $a = 0$, $[-2, 2]$
(b) $\cos x$, $a = 0$, $[-8, 8]$
(c) $1/(4+x)$, $a = 0$, $[-3, 3]$
(d) $1/x$, $a = 5$, $[1, 9]$
(e) x^2e^{-x}, $a = 2$, $[-4, 8]$
(f) $(2x-1)^2$, $a = 0$, $[-4, 4]$
(g) $1 + x + x^2 + x^3$, $a = 1$, $[-3, 5]$
(h) x^3, $a = -1$, $[-4, 2]$

4. (a) Expanding $\sin x$ about $x = 0$, use (7a) and (10) to show that

$$\sin x = x + R_3(x); \qquad R_3(x) = -\frac{\cos\xi}{3!}x^3,$$

where ξ is some number between 0 and x. Supposing that the interval of interest is $0 \le x \le 0.5$, show that a bound on the error in the approximation $\sin x \approx x$ is

$$|\sin x - x| < 0.021 \quad \text{over} \quad 0 \le x \le 0.5.$$

(b) Similarly, show that

$$\left|\sin x - \left(x - \frac{x^3}{3!}\right)\right| < 0.000261 \quad \text{over} \quad 0 \le x \le 0.5.$$

(c) Similarly, show that

$$\left|\sin x - \left(x - \frac{x^3}{3!} + \frac{x^5}{5!}\right)\right| < 0.000011 \quad \text{over} \quad 0 \le x \le 0.5.$$

5. Expanding $\sin x$ about $x = 2$, use (7a) and (10) to show that

$$\sin x = \sin 2 + (\cos 2)(x-2) - \left(\frac{\sin 2}{2}\right)(x-2)^2 + R_3(x),$$

where $R_3(x) = -(\cos\xi)(x-2)^3/3!$ and ξ is some number between 2 and x. Supposing that the interval of interest is $1.5 \le x \le 2.5$, show that

$$\left|\sin x - \left[\sin 2 + (\cos 2)(x-2) - \frac{\sin 2}{2}(x-2)^2\right]\right| < 0.0167$$

over $1.5 \le x \le 2.5$.

6. In the sentence following (9), we speak of assuming the continuity of the derivative $f^{(n)}(x)$. That comment begs a question as to how a derivative can *fail* to be continuous. For instance, suppose $f(x)$ is the "ramp function" $xH(x)$, where H is the Heaviside step function. Then $f'(x) = H(x)$ has a jump discontinuity at $x = 0$. However, at that point the original function $f(x) = xH(x)$ has a "kink" and is not differentiable. Thus, it is hard to imagine how a derivative $f'(x)$ can exist at a point x, yet fail to be continuous there. Nonetheless, the function

$$g(x) = \begin{cases} x^2 \sin\dfrac{1}{x}, & x \ne 0 \\ 0, & x = 0 \end{cases} \tag{6.1}$$

shows that such behavior is possible because $g'(x)$ exists at $x = 0$ but is not continuous there. The problem that we pose is for you to verify that $g'(x)$ exists at $x = 0$ but is not continuous there. HINT: Evaluate $g'(0)$, and $g'(x)$ for $x \ne 0$, and show that $\lim_{x\to 0} g'(x)$ does not equal $g'(0)$.

7. In Example 2, we stated that $f'(0) = f''(0) = \cdots = 0$. Verify that

(a) $f'(0) = 0$ (b) $f''(0) = 0$ (c) $f'''(0) = 0$

HINT: It will be necessary fo fall back on the limit-of-the-difference-quotient definition of derivative.

8. Below (26), we stated that $|x|^n/n! \to 0$ as $n \to \infty$. Prove that claim. HINT: Use **Stirling's formula**,

$$\boxed{n! \sim \sqrt{2\pi n}\, n^n e^{-n} \quad \text{as} \quad n \to \infty.} \tag{8.1}$$

9. (*Tangent plane*) Let a surface S be defined by an equation $f(x,y,z) = 0$ and let (a,b,c) be a point on S. The tangent plane to S, at (a,b,c), is obtained by replacing $f(x,y,z)$ in $f(x,y,z) = 0$ by its linear approximation $f(a,b,c) + f_x(a,b,c)(x-a) + f_y(a,b,c)(y-b) + f_z(a,b,c)(z-c)$ [i.e., by the first three terms of its Taylor series about (a,b,c)]. Obtain the tangent plane at $(1,3,-2)$:

(a) $x^3yz = -6$ (b) $xyz = -6$
(c) $3x^2 + y^2 + z^2 = 16$ (d) $\sin(x - y - z) = 0$
(e) $\sin(x^2 + y^2 + z) = \sin 8$ (f) $x^4 + y^2 + z^4 = 26$

10. The essential idea, in our derivation of the two-variable Taylor series (41), was the reduction to a Taylor series in *one* variable by introducing a parametrized line, given by (31), from the initial point (a,b) to the final point (x_0, y_0). Actually, *any* parametrized curve between those two points should give the same final result (41); we chose the straight line (31) as the simplest. In place of (31), use the parametrization

$$x = a + (x_0 - a)t, \qquad y = b + (y_0 - b)t^2, \tag{10.1}$$

which is a parabola, and show that you obtain the same final result (41).

11. (*Expanding in one variable at a time*) An easy way to derive Taylor series in more than one variable is to expand in one variable at a time. Use this method to expand the given function about the given point, through third-order terms.

(a) $x^5y^4 - y$, (1,2) (b) e^{xy}, (1,2)
(c) $\sin(3xy)$, $(1,-1)$ (d) $1/(x^2+y^2)$, (2,1)
(e) $1/(x^2-y^2)$, (3,1) (f) $1/(1+x^2y^2)$, (1,1)

12. (a) – (f) Use computer software to obtain the Taylor series in the corresponding part of exercise 11.

13.6 Implicit Functions and Jacobians

13.6.1. Implicit function theorem. An equation

$$f(x, y) = 0 \tag{1}$$

is said to constitute a **relation** on x and y. In (1), x and y have the same status; they are independent variables. Depending on the context, it may be desirable to change our point of view and regard (1) as implicitly defining y as a function of x (or vice versa). In that case we re-express (1) as

$$f(x, y(x)) = 0, \tag{2}$$

because x is now the independent variable and y is the dependent variable.

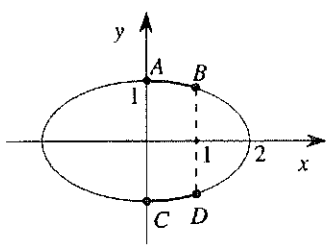


Figure 1. The ellipse $x^2 + 4y^2 - 4 = 0$.

EXAMPLE 1. Consider the relation

$$x^2 + 4y^2 - 4 = 0, \tag{3}$$

satisfied by the points on the ellipse shown in Fig. 1. That (3) is an implicit definition of a function $y(x)$ is clear because we can solve (3) for y, by algebra, as

$$y = \pm\sqrt{1 - \left(\frac{x}{2}\right)^2}. \tag{4}$$

If we are interested in the interval $0 \le x \le 1$, for instance, then we see that there are actually two continuous functions $y(x)$, the positive square root corresponding to the upper branch AB in Fig. 1, and the negative square root corresponding to the lower branch CD. If, instead, we are interested in the interval $3 \le x \le 8$, say, then there is *no* (real valued) function $y(x)$ implied by (3). ∎

In Example 1 we were able to solve (3) for $y(x)$ by simple algebra. In other cases we are not as fortunate. For instance, $2xy + \sin y = 3$ is a transcendental equation* and cannot be solved explicitly for y. Thus, there is the following important question. Given an x, y pair (x_0, y_0) satisfying the relation (1), does there exist an **implicit function** $y(x)$ [i.e., a function $y(x)$ implicitly defined by (1)] and, if so, is it unique? Sufficient conditions for that to be true are given by the **implicit function theorem**, which we state without proof.

THEOREM 13.6.1 *Implicit Function Theorem*
Let $f(x, y) = 0$ be satisfied by a pair of real numbers x_0, y_0 so that $f(x_0, y_0) = 0$, and suppose that $f(x, y)$ is C^1 in some neighborhood of (x_0, y_0) with

$$\boxed{\frac{\partial f}{\partial y}(x_0, y_0) \neq 0.} \tag{5}$$

*In Section 8.2 we define an equation as **algebraic** if it is a finite-degree polynomial equation and **transcendental** otherwise.

Then $f(x, y) = 0$ uniquely implies a function $y(x)$ in some neighborhood N of x_0 such that $y(x_0) = y_0$, where $y(x)$ is differentiable in N.*

Observe that the implicit function theorem is a local one. That is, it assures us that there exists a unique differentiable function $y(x)$ in some neighborhood of x_0, but it does not tell us how large that neighborhood is.

EXAMPLE 2. Consider again the relation (3) in Example 1,

$$f(x, y) = x^2 + 4y^2 - 4 = 0. \tag{6}$$

Noticing that the latter is satisfied by $x = 1$ and $y = -\sqrt{3}/2$, say, let (x_0, y_0) be $(1, -\sqrt{3}/2)$. Clearly, $f_x(x, y) = 2x$ and $f_y(x, y) = 8y$ are continuous in some neighborhood of $(1, -\sqrt{3}/2)$, because they are continuous everywhere in the x, y plane. Thus, f is C^1 in some neighborhood of $(1, -\sqrt{3}/2)$. Further, $f_y(1, -\sqrt{3}/2) = -4\sqrt{3} \neq 0$, so Theorem 13.6.1 tells us that there is, indeed, a unique differentiable function $y(x)$ through the point $(1, -\sqrt{3}/2)$. The graph of that function is the lower half of the ellipse (Fig. 2), over $-2 < x < 2$.

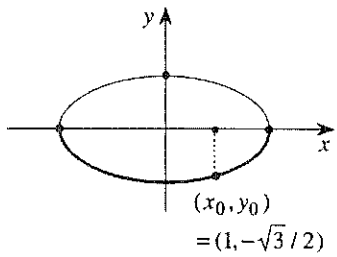


Figure 2. Implicit function $y(x)$ through $(1, -\sqrt{3}/2)$.

COMMENT. If we select (x_0, y_0) to be at one of the two ends, say $(2, 0)$, then the theorem would provide no information since $f_y(2, 0) = (8)(0) = 0$ so that condition (5) is not satisfied. ∎

Whereas Example 2 served to illustrate the implicit function theorem, it did not reveal the power of that theorem because it was such a simple case.

EXAMPLE 3. Consider the relation

$$(y - 2x)e^y - x^2 + 1 = 0. \tag{7}$$

We see by inspection that (7) is satisfied by $x = 1$ and $y = 2$ so let us take $(x_0, y_0) = (1, 2)$. Does there exist a function $y(x)$, implied by (7), through that point? The derivatives

$$f_x = -2e^y - 2x \qquad \text{and} \qquad f_y = (y - 2x + 1)e^y$$

are continuous everywhere in the x, y plane and

$$f_y(1, 2) = e^2 \neq 0$$

so, according to Theorem 13.6.1, there does exist such an implicit function $y(x)$. ∎

How to obtain that function $y(x)$ is another matter. Equation (7) is a transcendental equation in y and cannot be solved for y in closed form. Giving up on the

*$f(x, y)$ being C^1 means that the first-order partial derivatives f_x and f_y are continuous; see Section 13.5, below Example 5. Also, note that the neighborhood N of x_0 is an open *interval* on the x axis, whereas the neighborhood of (x_0, y_0) is an open *disk* in the x, y plane.

hope of obtaining a closed form solution, perhaps we can at least develop the Taylor series of $y(x)$ about x_0 (i.e., an "open" form solution).

Turning from Example 3 to the general case

$$f(x, y(x)) = 0, \tag{8}$$

suppose that we know an x_0 and a y_0 such that $f(x_0, y_0) = 0$, and let us seek the Taylor series of y about x_0,

$$y(x) = y(x_0) + y'(x_0)(x - x_0) + \frac{y''(x_0)}{2!}(x - x_0)^2 + \cdots. \tag{9}$$

The first coefficient, $y(x_0) = y_0$, is already known, but we need to evaluate $y'(x_0)$, $y''(x_0)$, and so on. Applying the chain rule to (8) gives

$$\frac{d}{dx} f(x, y(x)) = f_x + f_y y' = 0 \tag{10}$$

so

$$y'(x) = -\frac{f_x(x, y(x))}{f_y(x, y(x))}. \tag{11}$$

[It is interesting that (10) is a linear algebraic equation for y' even if (8) is a transcendental equation in y.] Continuing, differentiate (11) to find y'':

$$y'' = -\frac{f_{xx} + f_{xy}y'}{f_y} + \frac{f_x(f_{yx} + f_{yy}y')}{f_y^2}. \tag{12}$$

Using (11) for y', and assuming that $f_{xy} = f_{yx}$, (12) can be expressed as

$$y'' = \frac{2f_x f_y f_{xy} - f_x^2 f_{yy} - f_y^2 f_{xx}}{f_y^3}. \tag{13}$$

Similarly for y''', and so on. It remains merely to evaluate the expressions for y', y'', ... at (x_0, y_0).

Now we can appreciate the significance of the condition (5) in Theorem 13.6.1, for it is evident that increasingly higher powers of f_y build up in the denominators of y', y'', ... so that if $f_y(x_0, y_0) = 0$, then these derivatives fail to exist and, consequently, the Taylor series (9) fails to exist.

EXAMPLE 4. Let us return to Example 3 and obtain the Taylor series of $y(x)$ about $x_0 = 1$. Rather than merely evaluating the right-hand sides of (11) and (13), it might be more instructive to carry out steps (10)–(13) for the given equation

$$(y - 2x)e^y - x^2 + 1 = 0. \tag{14}$$

First, differentiating (14) with respect to x gives

$$(y' - 2)e^y + (y - 2x)e^y y' - 2x = 0$$

so

$$y' = \frac{2 + 2xe^{-y}}{y - 2x + 1}. \tag{15}$$

Next, differentiation of (15) gives

$$y'' = \frac{(2 - 2xy')e^{-y}}{y - 2x + 1} + \frac{(2 + 2xe^{-y})(-1)(y' - 2)}{(y - 2x + 1)^2}, \tag{16}$$

and so on. Evaluating (15) and (16) at $(1, 2)$ gives $y'(1, 2) = 2 + 2e^{-2}$ and $y''(1, 2) = -6e^{-2} - 8e^{-4}$ so

$$\begin{aligned} y(x) &= 2 + (2 + 2e^{-2})(x - 1) + \frac{1}{2!}(-6e^{-2} - 8e^{-4})(x - 1)^2 + \cdots \\ &= 2 + 2.271(x - 1) - 0.479(x - 1)^2 + \cdots \end{aligned}$$

is the desired Taylor series. ∎

Before extending these ideas to the case of two or more independent variables and two or more dependent variables, let us observe that **inverse functions** are but a special case of implicit functions.

EXAMPLE 5. Given the function

$$y = \sin x \tag{17}$$

on $-\infty < x < \infty$, is there an inverse function $x(y)$ through the point $x = 0, y = 0$? Yes, known as $\sin^{-1} y$, its graph is as shown in Fig. 3b and is obtained from Fig. 3a simply by reversing the horizontal and vertical axes. The x interval is limited to $-\pi/2 < x < \pi/2$ because if it were made broader then multivaluedness would result (Fig. 3c), whereas functions are required to be single valued.

Seeking an inverse function $x(y)$ is equivalent to re-expressing (17) as

$$f(x, y) = y - \sin x = 0 \tag{18}$$

and seeking an implicit function $x(y)$. Since the roles of x and y are reversed here, compared to their roles in this section prior to this example, let us re-express (18) as $f(y, x) = y - \sin x = 0$. Since $f_x = -\cos x$ and $f_y = 1$ are continuous everywhere in the x, y plane and $f_x = -\cos x = -1 \neq 0$ at $(x_0, y_0) = (0, 0)$, Theorem 13.6.1 assures us that there does exist a unique differentiable implicit function $x(y)$ (namely, the inverse function $\sin^{-1} y$) in some neighborhood of y_0 (namely, $-1 < y < 1$). ∎

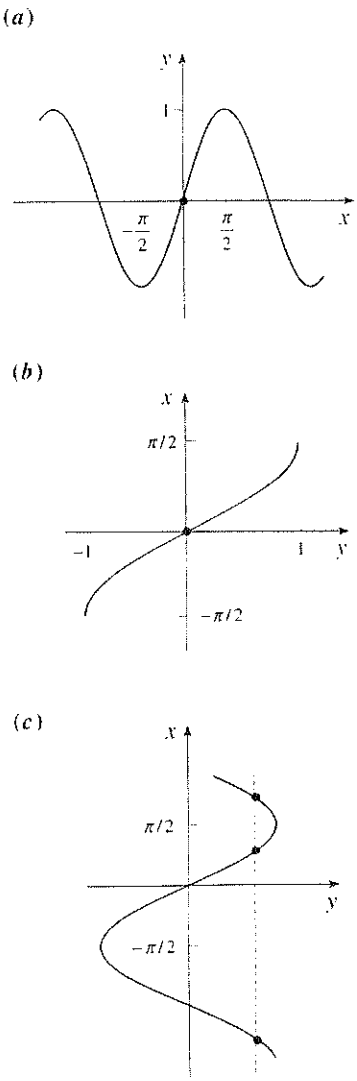


Figure 3. The inverse function $x = \sin^{-1} y$.

13.6.2. Extension to multivariable case. Next, consider *two* relations,

$$\begin{aligned} f(x, y, u, v) &= 0, \\ g(x, y, u, v) &= 0, \end{aligned} \tag{19}$$

on x, y, u, v. Supposing that we know a point (x_0, y_0, u_0, v_0) in four-dimensional x, y, u, v space such that $f(x_0, y_0, u_0, v_0) = 0$ and $g(x_0, y_0, u_0, v_0) = 0$, we wonder if (19) implicitly defines functions $u(x, y), v(x, y)$ throughout some neighborhood of x_0, y_0. To explore that possibility, let us re-express (19) as

$$\begin{aligned} f(x, y, u(x, y), v(x, y)) &= 0, \\ g(x, y, u(x, y), v(x, y)) &= 0 \end{aligned} \tag{20}$$

and seek Taylor series of u and v,

$$\begin{aligned} u(x, y) &= u(x_0, y_0) + u_x(x_0, y_0)(x - x_0) + u_y(x_0, y_0)(y - y_0) + \cdots, \\ v(x, y) &= v(x_0, y_0) + v_x(x_0, y_0)(x - x_0) + v_y(x_0, y_0)(y - y_0) + \cdots, \end{aligned} \tag{21}$$

about x_0, y_0. We already know $u(x_0, y_0) = u_0$ and $v(x_0, y_0) = v_0$, but we need to compute $u_x(x_0, y_0)$, $u_y(x_0, y_0)$, ..., and $v_x(x_0, y_0)$, $v_y(x_0, y_0)$, ... To find u_x and v_x, apply the chain rule to (20):

$$\begin{aligned} \frac{\partial}{\partial x} f(x, y, u(x, y), v(x, y)) &= f_x + f_u u_x + f_v v_x = 0, \\ \frac{\partial}{\partial x} g(x, y, u(x, y), v(x, y)) &= g_x + g_u u_x + g_v v_x = 0 \end{aligned} \tag{22}$$

or

$$\begin{aligned} f_u u_x + f_v v_x &= -f_x, \\ g_u u_x + g_v v_x &= -g_x. \end{aligned} \tag{23}$$

Solving (23) by Cramer's rule (Section 10.6.3 or Appendix B) gives

$$u_x = \frac{\begin{vmatrix} -f_x & f_v \\ -g_x & g_v \end{vmatrix}}{\begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix}} = -\frac{\begin{vmatrix} f_x & f_v \\ g_x & g_v \end{vmatrix}}{\begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix}}, \tag{24a}$$

$$v_x = \frac{\begin{vmatrix} f_u & -f_x \\ g_u & -g_x \end{vmatrix}}{\begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix}} = -\frac{\begin{vmatrix} f_u & f_x \\ g_u & g_x \end{vmatrix}}{\begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix}}. \tag{24b}$$

Similarly, to find u_y and v_y take $\partial/\partial y$ of (20) and obtain

$$u_y = \frac{\begin{vmatrix} -f_y & f_v \\ -g_y & g_v \end{vmatrix}}{\begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix}} = -\frac{\begin{vmatrix} f_y & f_v \\ g_y & g_v \end{vmatrix}}{\begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix}}, \tag{25a}$$

$$v_y = \frac{\begin{vmatrix} f_u & -f_y \\ g_u & -g_y \end{vmatrix}}{\begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix}} = -\frac{\begin{vmatrix} f_u & f_y \\ g_u & g_y \end{vmatrix}}{\begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix}}. \tag{25b}$$

These results are analogous to those obtained above for the Taylor series expansion of a function $y(x)$ that is implicitly defined by $f(x, y(x)) = 0$. There, we saw from (11) – (13) that the coefficients $y'(x_0)$, $y''(x_0)$, ... in (9) exist if $f_y(x_0, y_0) \neq 0$, and fail to exist if $f_y(x_0, y_0) = 0$. From (24) and (25) we see, analogously, that the derivatives u_x, v_x, u_y, v_y needed in the Taylor series expansions (21) exist if the *determinant*

$$\begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix} \tag{26}$$

does not vanish (i.e., equal zero) at x_0, y_0, u_0, v_0, and that they fail to exist if the determinant does vanish at that point. Similarly, we would find that the same determinant condition applies for the existence of the higher-order partial derivatives needed in (21).

Thus, just as the condition

$$\left.\frac{\partial f}{\partial y}\right|_{x_0, y_0} \neq 0 \tag{27}$$

is crucial [recall (5)] if

$$f(x, y) = 0 \tag{28}$$

is to imply the existence of an implicit function $y(x)$, the condition

$$\left.\begin{vmatrix} f_u & f_v \\ g_u & g_v \end{vmatrix}\right|_{x_0, y_0, u_0, v_0} \neq 0 \tag{29}$$

is crucial if

$$f(x, y, u, v) = 0, \tag{30a}$$

$$g(x, y, u, v) = 0 \tag{30b}$$

are to imply the existence of implicit functions $u(x, y)$ and $v(x, y)$. If we realize that the $\partial f / \partial y$ in (27) is actually a "one-by-one determinant," then we can see how the condition (27) generalizes to the nonvanishing of a two-by-two determinant when there are two dependent variables and two independent variables and, similarly, to the nonvanishing of an n-by-n determinant when there are n dependent variables and n independent variables. Thus, the generalized version of Theorem 13.6.1 is as follows.

THEOREM 13.6.2 *Implicit Function Theorem, Multivariable Case*
Let the system of n equations

$$\begin{aligned} f_1(x_1, \ldots, x_n, u_1, \ldots, u_n) &= 0 \\ &\vdots \\ f_n(x_1, \ldots, x_n, u_1, \ldots, u_n) &= 0 \end{aligned} \tag{31}$$

be satisfied by the real numbers $x_{1_0}, \ldots, x_{n_0}, u_{1_0}, \ldots, u_{n_0}$, and suppose that the functions $f_1(x_1, \ldots, x_n, u_1, \ldots, u_n)$ through $f_n(x_1, \ldots, x_n, u_1, \ldots, u_n)$ are C^1 in some neighborhood of $(x_{1_0}, \ldots, x_{n_0}, u_{1_0}, \ldots, u_{n_0})$, with

$$\boxed{\left|\begin{matrix} \dfrac{\partial f_1}{\partial u_1} & \dfrac{\partial f_1}{\partial u_2} & \cdots & \dfrac{\partial f_1}{\partial u_n} \\ \vdots & & & \vdots \\ \dfrac{\partial f_n}{\partial u_1} & \dfrac{\partial f_n}{\partial u_2} & \cdots & \dfrac{\partial f_n}{\partial u_n} \end{matrix}\right|_{x_{1_0}, \ldots, x_{n_0}, u_{1_0}, \ldots, u_{n_0}} \neq 0.} \tag{32}$$

Then (31) uniquely implies functions $u_1(x_1, \ldots, x_n), \ldots, u_n(x_1, \ldots, x_n)$ in some neighborhood N of $(x_{1_0}, \ldots, x_{n_0})$ such that $u_1(x_{1_0}, \ldots, x_{n_0}) = u_{1_0}, \ldots, u_n(x_{1_0}, \ldots, x_{n_0}) = u_{n_0}$, where $u_1(x_1, \ldots, x_n), \ldots, u_n(x_1, \ldots, x_n)$ are C^1 in N.

Although we motivated Theorem 13.6.2, above, we did not prove it.*

EXAMPLE 6. Consider the familiar change of variables from the Cartesian x, y coordinates to the polar coordinates r, θ (Fig. 4),

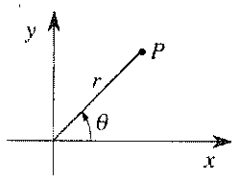


Figure 4. Polar coordinates.

$$x = r\cos\theta, \tag{33a}$$
$$y = r\sin\theta \tag{33b}$$

or, equivalently,

$$\begin{aligned} f_1(x, y, r, \theta) &= x - r\cos\theta = 0, \\ f_2(x, y, r, \theta) &= y - r\sin\theta = 0. \end{aligned} \tag{34}$$

Do these relations define implicit functions $r(x, y)$ and $\theta(x, y)$ throughout some neighborhood of any given point P (Fig. 4)? Put differently, if (33) gives x and y as functions of r and θ, do those relations implicitly define inverse functions $r(x, y)$ and $\theta(x, y)$?

To answer that queston using Theorem 13.6.2 let us identify x, y, r, θ as x_1, x_2, u_1, u_2, respectively. (Whether we identify x as x_1 and y as x_2, or vice versa, will not matter; similarly for r, θ and u_1, u_2.) All first-order partial derivatives of f_1 and f_2 are continuous everywhere so f_1 and f_2 are C^1 everywhere. Further,

$$\begin{vmatrix} \dfrac{\partial f_1}{\partial r} & \dfrac{\partial f_1}{\partial \theta} \\ \dfrac{\partial f_2}{\partial r} & \dfrac{\partial f_2}{\partial \theta} \end{vmatrix} = \begin{vmatrix} -\cos\theta & r\sin\theta \\ -\sin\theta & -r\cos\theta \end{vmatrix} = -r \tag{35}$$

is nonzero everywhere except at the origin ($r = 0$). From Theorem 13.6.2 it follows that inverse functions $r(x, y)$ and $\theta(x, y)$ do exist (and are unique and of class C^1) in some neighborhood of any given point P (Fig. 4) except if P is at the origin, in which case the

*See, for instance, I. S. Sokolnikoff, *Advanced Calculus* (New York: McGraw-Hill, 1939), Chapter 12, or W. Maak, *Modern Calculus* (New York: Holt, Rinehart and Winston, 1963), Chapter 9.

theorem gives no information. This singular result is not surprising since there is a problem at the origin; as we can see from the figure, θ is *not defined* there. ∎

13.6.3. Jacobians. The determinant in (32) is known as the **Jacobian** of $f_1, \ldots, f_n$ with respect to $u_1, \ldots, u_n$ and is denoted either by the notation $\dfrac{\partial(f_1, \ldots, f_n)}{\partial(u_1, \ldots, u_n)}$ or as $J(u_1, \ldots, u_n)$:

$$\begin{vmatrix} \dfrac{\partial f_1}{\partial u_1} & \dfrac{\partial f_1}{\partial u_2} & \cdots & \dfrac{\partial f_1}{\partial u_n} \\ \vdots & & & \vdots \\ \dfrac{\partial f_n}{\partial u_1} & \dfrac{\partial f_n}{\partial u_2} & \cdots & \dfrac{\partial f_n}{\partial u_n} \end{vmatrix} \equiv \frac{\partial(f_1, \ldots, f_n)}{\partial(u_1, \ldots, u_n)} \quad \text{or} \quad J(u_1, \ldots, u_n) \tag{36}$$

and is a function of $u_1, \ldots, u_n$.

Jacobian determinants arise not only in the Implicit Function Theorem but also in the evaluation of the various partial derivatives of implicit functions. For instance, consider the functions $u(x, y)$ and $v(x, y)$ defined implicitly by (20). The steps (22)–(25), which we urge you to review, give these results for the partial derivatives of u and v with respect to x and y:

$$u_x = -\frac{\dfrac{\partial(f, g)}{\partial(x, v)}}{\dfrac{\partial(f, g)}{\partial(u, v)}}, \qquad u_y = -\frac{\dfrac{\partial(f, g)}{\partial(y, v)}}{\dfrac{\partial(f, g)}{\partial(u, v)}}, \tag{37a,b}$$

$$v_x = -\frac{\dfrac{\partial(f, g)}{\partial(u, x)}}{\dfrac{\partial(f, g)}{\partial(u, v)}}, \qquad v_y = -\frac{\dfrac{\partial(f, g)}{\partial(u, y)}}{\dfrac{\partial(f, g)}{\partial(u, v)}}. \tag{38a,b}$$

EXAMPLE 7. Consider, once again, the change of variables from x, y to r, θ, given by (33). We will refer to x_r, x_θ, y_r and y_θ as the "forward" partial derivatives, since x and y are given explicitly by (33) as functions of r and θ. The forward derivatives follow readily from (33) as

$$x_r = \cos\theta, \qquad x_\theta = -r\sin\theta, \qquad y_r = \sin\theta, \qquad y_\theta = r\cos\theta, \tag{39}$$

but to evaluate the reverse derivatives is harder; we can carry out the steps (22)–(25) or we can simply use the final results, reproduced (in terms of Jacobians) in (37) and (38). To use (37) and (38), let u be r and let v be θ. Then

$$\begin{aligned} f(x, y, r, \theta) &= x - r\cos\theta = 0, \\ g(x, y, r, \theta) &= y - r\sin\theta = 0 \end{aligned} \tag{40}$$

so (37) gives

$$r_x = -\frac{\begin{vmatrix} f_x & f_\theta \\ g_x & g_\theta \end{vmatrix}}{\begin{vmatrix} f_r & f_\theta \\ g_r & g_\theta \end{vmatrix}} = -\frac{\begin{vmatrix} 1 & r\sin\theta \\ 0 & -r\cos\theta \end{vmatrix}}{\begin{vmatrix} -\cos\theta & r\sin\theta \\ -\sin\theta & -r\cos\theta \end{vmatrix}} = \cos\theta, \tag{41a}$$

$$r_y = -\frac{\begin{vmatrix} f_y & f_\theta \\ g_y & g_\theta \end{vmatrix}}{\begin{vmatrix} f_r & f_\theta \\ g_r & g_\theta \end{vmatrix}} = -\frac{\begin{vmatrix} 0 & r\sin\theta \\ 1 & -r\cos\theta \end{vmatrix}}{\begin{vmatrix} -\cos\theta & r\sin\theta \\ -\sin\theta & -r\cos\theta \end{vmatrix}} = \sin\theta, \tag{41b}$$

and (38) gives

$$\theta_x = -\frac{\begin{vmatrix} f_r & f_x \\ g_r & g_x \end{vmatrix}}{\begin{vmatrix} f_r & f_\theta \\ g_r & g_\theta \end{vmatrix}} = -\frac{\begin{vmatrix} -\cos\theta & 1 \\ -\sin\theta & 0 \end{vmatrix}}{r} = -\frac{\sin\theta}{r}, \tag{42a}$$

$$\theta_y = -\frac{\begin{vmatrix} f_r & f_y \\ g_r & g_y \end{vmatrix}}{\begin{vmatrix} f_r & f_\theta \\ g_r & g_\theta \end{vmatrix}} = -\frac{\begin{vmatrix} -\cos\theta & 0 \\ -\sin\theta & 1 \end{vmatrix}}{r} = \frac{\cos\theta}{r}. \tag{42b}$$

Of course these results can be expressed in terms of x, y if we wish. For instance, $\theta_y = r\cos\theta/r^2 = x/(x^2+y^2)$.

COMMENT. In this case we could have solved (33) explicitly for $r(x,y)$ and $\theta(x,y)$, for squaring and adding those equations gives r and dividing one by the other gives θ:

$$r = \sqrt{x^2+y^2}, \qquad \theta = \tan^{-1}\frac{y}{x}, \tag{43}$$

from which we can compute the "forward" derivatives r_x, r_y, θ_x, and θ_y. However, we cannot always solve for the inverse functions analytically. Hence, we have used this example to illustrate the use of the Jacobians in solving for the reverse derivatives. ∎

Why all the fuss? Could we not have solved for the reverse derivatives by finding the forward derivatives (which is simple) and "turning them upside down"? For instance, knowing that $\partial x/\partial r = \cos\theta$, could we not say that $\partial r/\partial x = \frac{1}{\partial x/\partial r} = \frac{1}{\cos\theta}$? Evidently not, since (41a) gave $\partial r/\partial x = \cos\theta$, not $1/\cos\theta$.*

To understand this important point, observe that if $x = x(u)$ has an inverse function $u = u(x)$, then du/dx *does* equal $(dx/du)^{-1}$ or, equivalently,

$$\frac{du}{dx}\frac{dx}{du} = 1. \tag{44}$$

*The fact that, besides not being numerical inverses of each other, r_x and x_r are actually equal to each other is just a coincidence.

For instance, if $x = u^2$ then $dx/du = 2u$ and $du/dx = d(\sqrt{x})/dx = 1/(2\sqrt{x}) = 1/(2u) = (dx/du)^{-1}$. However, if

$$x = x(u, v) \quad \text{and} \quad y = y(u, v) \tag{45a}$$

have inverse functions

$$u = u(x, y) \quad \text{and} \quad v = v(x, y), \tag{45b}$$

then (in general)

$$\frac{\partial u}{\partial x}\frac{\partial x}{\partial u} \neq 1, \quad \frac{\partial u}{\partial y}\frac{\partial y}{\partial u} \neq 1, \quad \frac{\partial v}{\partial x}\frac{\partial x}{\partial v} \neq 1, \quad \frac{\partial v}{\partial y}\frac{\partial y}{\partial v} \neq 1. \tag{46}$$

Rather, what follows from (45) is the *Jacobian* generalization of (44),

$$\frac{\partial(u, v)}{\partial(x, y)}\frac{\partial(x, y)}{\partial(u, v)} = 1, \tag{47}$$

rather than (46). Similarly, if $x_1, \ldots, x_n$ are functions of $u_1, \ldots, u_n$, and vice versa, then

$$\boxed{\frac{\partial(u_1, \ldots, u_n)}{\partial(x_1, \ldots, x_n)}\frac{\partial(x_1, \ldots, x_n)}{\partial(u_1, \ldots, u_n)} = 1.} \tag{48}$$

Summarizing, while it is *not* true (in general) that the individual partial derivatives are numerical inverses of each other [e.g., $\dfrac{\partial u_2}{\partial x_3} \neq \left(\dfrac{\partial x_3}{\partial u_2}\right)^{-1}$], it *is* true that the *Jacobians* are numerical inverses of each other:

$$\frac{\partial(u_1, \ldots, u_n)}{\partial(x_1, \ldots, x_n)} = \left(\frac{\partial(x_1, \ldots, x_n)}{\partial(u_1, \ldots, u_n)}\right)^{-1}. \tag{49}$$

Rather than prove the general case (48), let it suffice to prove (47), where $n = 2$. We begin by writing out

$$\begin{aligned}
\frac{\partial(u, v)}{\partial(x, y)}\frac{\partial(x, y)}{\partial(u, v)} &= \begin{vmatrix} u_x & u_y \\ v_x & v_y \end{vmatrix} \begin{vmatrix} x_u & x_v \\ y_u & y_v \end{vmatrix} \\
&= (u_x v_y - u_y v_x)(x_u y_v - x_v y_u) \\
&= u_x v_y x_u y_v - u_y \underline{v_x x_u} y_v \\
&\qquad - \underline{u_x} v_y \underline{x_v} y_u + u_y v_x x_v y_u.
\end{aligned} \tag{50}$$

Next, use the chain derivative results*

$$\frac{\partial v}{\partial u} = 0 = \frac{\partial v}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial v}{\partial y}\frac{\partial y}{\partial u} \quad \text{so} \quad v_x x_u = -v_y y_u,$$

$$\frac{\partial u}{\partial v} = 0 = \frac{\partial u}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial u}{\partial y}\frac{\partial y}{\partial v} \quad \text{so} \quad u_x x_v = -u_y y_v$$

*In case it is not clear why $\partial v/\partial u = 0$, for instance, think of $\{x, y\}$ and $\{u, v\}$ as two "families" of variables. By $\partial/\partial u$ we mean differentiation with respect to u, holding all other members of the u family (namely, v) fixed. Because v is regarded as fixed, $\partial v/\partial u = 0$.

to replace the underlined products in (50). Then the right-hand side of (50) becomes

$$\begin{aligned}\text{RHS} &= u_x v_y x_u y_v + u_y y_v v_y y_u \\ &\qquad + v_y y_u u_y y_v + u_y v_x x_v y_u \\ &= y_v v_y (u_x x_u + u_y y_u) + y_u u_y (v_y y_v + v_x x_v) \\ &= y_v v_y u_u + y_u u_y v_v \\ &= y_v v_y + y_u u_y = \frac{\partial y}{\partial y} \\ &= 1, \end{aligned} \tag{51}$$

as claimed. Naturally, we assumed that all the first-order partial derivatives are continuous so that the chain rule could be used.

13.6.4. Application to change of variables. It sometimes happens that it is convenient to change from one set of independent variables to another, and that step involves the concepts discussed above.

To illustrate, suppose that we wish to determine the temperature distribution $T(x, y)$ in a semicircular plate, with the temperature maintained at 100° along the circular edge and at 0° along the straight edge, as depicted in Fig. 5. It turns out, as we will find in Chapter 16, that $T(x, y)$ must satisfy the *Laplace equation*, namely, the partial differential equation

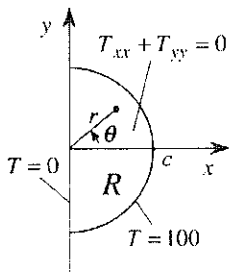


Figure 5. Heat conduction problem.

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0 \tag{52}$$

in the semicircular region $\mathcal{R}$, together with the boundary conditions that $T = 100$ on the circular edge, and $T = 0$ on the straight edge.

It is much better to re-express the problem in terms of the polar coordinates r, θ, because then the boundary conditions will be on constant-coordinate curves: $T = 100$ on $r = c$, $T = 0$ on $\theta = -\pi/2$, and $T = 0$ on $\theta = \pi/2$. Thus, let us re-express (52) in terms of r and θ. First, consider the T_{xx} term. Since $T_{xx} = (\partial/\partial x)(\partial/\partial x)T$, we need to express the differential operator $\partial/\partial x$ in terms of r and θ. By chain differentiation,

$$\frac{\partial(\ )}{\partial x} = \frac{\partial(\ )}{\partial r}\frac{\partial r}{\partial x} + \frac{\partial(\ )}{\partial \theta}\frac{\partial \theta}{\partial x}. \tag{53}$$

The reverse derivatives $r_x, \theta_x, r_y, \theta_y$ are given in (41) and (42) so (53) becomes

$$\frac{\partial}{\partial x} = c\frac{\partial}{\partial r} - \frac{s}{r}\frac{\partial}{\partial \theta}, \tag{54}$$

where we use the convenient shorthand $c \equiv \cos\theta$ and $s \equiv \sin\theta$. Thus,

$$T_{xx} = \left(c\frac{\partial}{\partial r} - \frac{s}{r}\frac{\partial}{\partial \theta}\right)\left(c\frac{\partial}{\partial r} - \frac{s}{r}\frac{\partial}{\partial \theta}\right)T$$

$$= \left(c\frac{\partial}{\partial r} - \frac{s}{r}\frac{\partial}{\partial \theta}\right)\left(cT_r - \frac{s}{r}T_\theta\right)$$

$$= c\frac{\partial}{\partial r}(cT_r) - c\frac{\partial}{\partial r}\left(\frac{s}{r}T_\theta\right) - \frac{s}{r}\frac{\partial}{\partial \theta}(cT_r) + \frac{s}{r}\frac{\partial}{\partial \theta}\left(\frac{s}{r}T_\theta\right)$$

$$= c^2T_{rr} + \frac{cs}{r^2}T_\theta - \frac{cs}{r}T_{\theta r} + \frac{s^2}{r}T_r - \frac{sc}{r}T_{r\theta} + \frac{sc}{r^2}T_\theta + \frac{s^2}{r^2}T_{\theta\theta}. \tag{55}$$

Similarly,

$$\frac{\partial(\ )}{\partial y} = \frac{\partial(\ )}{\partial r}\frac{\partial r}{\partial y} + \frac{\partial(\ )}{\partial \theta}\frac{\partial \theta}{\partial y} = s\frac{\partial}{\partial r} + \frac{c}{r}\frac{\partial}{\partial \theta}, \tag{56}$$

and (Exercise 13)

$$T_{yy} = s^2T_{rr} - \frac{cs}{r^2}T_\theta + \frac{cs}{r}T_{\theta r} + \frac{c^2}{r}T_r + \frac{cs}{r}T_{r\theta} - \frac{cs}{r^2}T_\theta + \frac{c^2}{r^2}T_{\theta\theta}. \tag{57}$$

Finally, adding (55) and (57) and recalling that $c^2 + s^2 = 1$ gives

$$\frac{\partial^2 T}{\partial r^2} + \frac{1}{r}\frac{\partial T}{\partial r} + \frac{1}{r^2}\frac{\partial^2 T}{\partial \theta^2} = 0 \tag{58}$$

as the Laplace equation in polar coordinates.

For notational simplicity we used the same letter, T, for the temperature, independent of whether it was regarded as a function of x, y or as a function of r, θ. Strictly speaking, of course, we should have used different letters [as discussed below equation (8) in Section 13.4]: $T(x, y) = T(x(r, \theta), y(r, \theta)) \equiv U(r, \theta)$, for instance.

Closure. We begin this section by considering a relation $f(x, y) = 0$ on x and y, to determine whether it implies a function $y(x)$ over some neighborhood of x_0, through a point (x_0, y_0) that satisfies the relation. Typically, we cannot solve $f(x, y) = 0$ for y by algebra, so we give up on the idea of a closed form expression of $y(x)$ and seek $y(x)$ in the form of a Taylor series about x_0. We find that all goes well if $f_y(x_0, y_0)$ does not vanish. With that motivational introduction completed, we then state the Implicit Function Theorem 13.6.1, which includes the key stipulation that

$$f_y(x_0, y_0) \neq 0. \tag{59}$$

Generalizing from one relation on two variables to two relations on four variables,

$$\begin{aligned} f(x, y, u, v) &= 0, \\ g(x, y, u, v) &= 0, \end{aligned} \tag{60}$$

we find that the condition (59) generalizes [from the one-by-one determinant on the left-hand side of (59)] to the determinant condition

$$\left| \begin{matrix} f_u & f_v \\ g_u & g_v \end{matrix} \right| \Bigg|_{x_0, y_0, u_0, v_0} \neq 0,$$

that is, the nonvanishing of the Jacobian determinant of f and g with respect to u and v at (x_0, y_0, u_0, v_0). Similarly for n relations on $2n$ variables, which case is covered in Theorem 13.6.2.

The Jacobians themselves are of great importance. We see that the "reverse derivatives," such as u_y, where u and v are defined implicitly as functions of x and y by $x = x(u, v)$ and $y = y(u, v)$, are expressible as ratios of Jacobian determinants. Further, we see that the result

$$\frac{dy}{dx}\frac{dx}{dy} = 1 \quad \text{or} \quad \frac{dx}{dy} = \left(\frac{dy}{dx}\right)^{-1},$$

for functions $y(x)$ and $x(y)$, generalizes *not* to $\dfrac{\partial x}{\partial u} = \left(\dfrac{\partial u}{\partial x}\right)^{-1}$ but to the *Jacobian* statements

$$\frac{\partial(x,y)}{\partial(u,v)}\frac{\partial(u,v)}{\partial(x,y)} = 1 \quad \text{or} \quad \frac{\partial(u,v)}{\partial(x,y)} = \left(\frac{\partial(x,y)}{\partial(u,v)}\right)^{-1}, \tag{61}$$

$$\frac{\partial(x,y,z)}{\partial(u,v,w)}\frac{\partial(u,v,w)}{\partial(x,y,z)} = 1 \quad \text{or} \quad \frac{\partial(u,v,w)}{\partial(x,y,z)} = \left(\frac{\partial(x,y,z)}{\partial(u,v,w)}\right)^{-1}, \tag{62}$$

and so on. In fact, in Chapter 15 we will meet Jacobians again, in deriving expressions for area and volume elements for surface and volume integrals.

EXERCISES 13.6

1. Given $f(x, y) = 0$ and a point (x_0, y_0) such that $f(x_0, y_0) = 0$, see if the conditions of Theorem 13.6.1 are met. If so, develop the implicit function $y(x)$ in a Taylor series about x_0, through second-order terms, as we did in Example 4.

(a) $x^2 + y^2 - 2y = 0$; $(1, 1)$
(b) $x^2 + 4y^2 - 4 = 0$; $(0, 1)$
(c) $x^2 + 4y^2 - 4 = 0$; $(0, -1)$
(d) $x(\cos \pi y + 1) + x^3 y + 8 = 0$; $(-2, 1)$
(e) $x(\cos \pi y + 1) + (x^3 + 8)y = 0$; $(-2, 1)$
(f) $x - y - \sin y = 0$; $(0, 0)$
(g) $x - y + \sin y = 0$; $(0, 0)$
(h) $(y - 1)e^y - x^2 + 1 = 0$; $(1, 1)$

2. In each case, find $y'(x)$ and $y''(x)$.

(a) $xy - y^3 = 1$
(b) $xe^y + y = 3$
(c) $y + y^2 + y^3 = x$
(d) $xye^y = 1$
(e) $xy + \sin y = 3x^2$
(f) $y \cos y = x^3$

3. Solve for $y_x|_z$ (i.e., holding z fixed) and $z_x|_y$.

(a) $xy + \sin(x + z) - z^2 = 5$
(b) $xe^y - y^2 - z^2 \sin z = 0$
(c) $e^x + e^y + e^z = 3$
(d) $xy^3 - 2x^3 z + y^5 z^5 = 1$

4. Apply Theorem 13.6.2 to see if implicit functions $u(x, y)$ and $v(x, y)$ exist in some neighborhood of (x_0, y_0), given the values (x_0, y_0, u_0, v_0) satisfying the two relations.

(a) $x - u \cos v = 0$, $y - u \sin v = 0$; $(0, 0, 0, 0)$
(b) $x - u \cos v = 0$, $y - u \sin v = 0$; $(0, 2, 2, \pi/2)$
(c) $x \sin u + y^2 - v^2 = 0$,
$(x + v)^2 - \sin(uy) = 0$; $(1, 1, 0, -1)$
(d) $x \cos u + y + v^2 = 0$, $x - y + \sin(u^2 v) = 0$; $(1, 1, \pi, 0)$
(e) $y \cos u + x - v^3 = 0$, $v^2 + \sin(x - y) = 0$; $(1, 1, \pi, 0)$
(f) $xe^y - 2u^2 + v + 7 = 0$, $\sin y - x^2 \sin v = 0$, $(1, 0, 2, 0)$

5. Evaluate u_y.

(a) $x - y + u^2 + v^2 = 1$, $x + y + u^3 e^v = 2$
(b) $x - u^2 \cos v = 0$, $y - ue^v = 6$

(c) $xe^u - uy + v = 0, \quad x^2u - y^3 + v^3 = 1$
(d) $x - u^2 \sin v = 0, \quad y - u^2 \cos v = 1$

6. Evaluate the indicated Jacobian(s).

(a) $f(u,v) = 3uv^2, \quad g(u,v) = u^2 - v^2; \quad \dfrac{\partial(f,g)}{\partial(u,v)}$
(b) $f(u,v,w) = uw^3, \quad g(u,v,w) = 2v - w,$
$h(u,v,w) = e^{uv}; \quad \dfrac{\partial(f,g,h)}{\partial(u,v,w)}$
(c) $F(p,q) = p^2 + 2, \quad G(p,q) = p \sin q;$
$\dfrac{\partial(F,G)}{\partial(p,q)}, \quad \dfrac{\partial(G,F)}{\partial(p,q)}, \quad \dfrac{\partial(G,F)}{\partial(q,p)}$
(d) $P(x,y) = y^3, \quad Q(x,y) = x^2 - y^2;$
$\dfrac{\partial(P,Q)}{\partial(x,y)}, \quad \dfrac{\partial(P,Q)}{\partial(y,x)}$
(e) $F(x,y,z) = x + y + z, \quad G(x,y,z) = x^2 + y^2 + z^2,$
$H(x,y,z) = x^3 + y^3 + z^3; \quad \dfrac{\partial(F,G,H)}{\partial(x,y,z)}$

7. Show that

(a) $\dfrac{\partial(f,g)}{\partial(x,y)} = -\dfrac{\partial(f,g)}{\partial(y,x)}$ (b) $\dfrac{\partial(f,g)}{\partial(x,y)} = \dfrac{\partial(g,f)}{\partial(y,x)}$

8. Verify (44), given the relation

(a) $xe^u - x^3u - 5 = 0$
(b) $xu^2 - ue^{-x} - u^5 = 0$
(c) $x + u + x^3 + u^3 - 9 = 0$
(d) $e^{xu} - \sin(x + u) - 1 = 0$

9. Verify (48) for these cases, by working out the eight partial derivatives, multiplying the determinants, and showing that the result is unity.

(a) $x = u + v, \quad y = u - v$
(b) $x = u + v, \quad y = u^3 - v^3$
(c) $x = u + e^v, \quad y = ve^u$
(d) $xe^y - u + e^v = 2, \quad x^2 + y^2 + u^2 + v^2 = 1$
(e) $x + y + u + v = 1, \quad x^2 + y^2 + u^2 + v^2 = 4$

10. (*Chain rule*) Recall from the calculus that if $u = u(x(s))$, then

$$\frac{du}{dx}\frac{dx}{ds} = \frac{du}{ds}, \tag{10.1}$$

which result is an example of the **chain rule**; equation (44) is a special case of (10.1), where s is u. If u and v are functions of x and y, and x and y, in turn, are functions of r and s, then (10.1) generalizes to

$$\boxed{\frac{\partial(u,v)}{\partial(x,y)}\frac{\partial(x,y)}{\partial(r,s)} = \frac{\partial(u,v)}{\partial(r,s)}.} \tag{10.2}$$

Similarly,

$$\boxed{\frac{\partial(u,v,w)}{\partial(x,y,z)}\frac{\partial(x,y,z)}{\partial(r,s,t)} = \frac{\partial(u,v,w)}{\partial(r,s,t)},} \tag{10.3}$$

and so on. Prove (10.2). HINT: Proceed essentially as we did in (50)-(51).

11. Verify (10.2), above, for these cases.

(a) $x = u\cos v, \ y = u\sin v,$ and $u = r + s, \quad v = r^2 + s^2$
(b) $x = u + v, \ y = u - v,$ and $u = r^2 + 2s, \ v = r\cos s$
(c) $x = u, \ y = u + v,$ and $u = r\cos s, \ v = r\sin s$
(d) $x = u^2 + 4v, \ y = u - v,$ and $u = r - s, \ v = r^2 + 4s$

12. One speaks of a relation

$$f(p,T,v) = 0 \tag{12.1}$$

on the pressure p, the absolute temperature T, and the specific volume v of a gas as an *equation of state*.

(a) If we think of (12.1) as implicitly defining $v(p,T)$, show that

$$\frac{\partial v}{\partial p} = -\frac{f_p}{f_v} \quad \text{and} \quad \frac{\partial v}{\partial T} = -\frac{f_T}{f_v}, \quad \text{if} \quad f_v \neq 0. \tag{12.2}$$

(b) Is it true, in this case, that

$$\frac{\partial v}{\partial p} = \left(\frac{\partial p}{\partial v}\right)^{-1} \quad \text{and} \quad \frac{\partial v}{\partial T} = \left(\frac{\partial T}{\partial v}\right)^{-1}? \tag{12.3}$$

Explain.
(c) One well known equation of state is the **van der Waals equation**

$$\left(p + \frac{a}{v^2}\right)(v - b) = RT, \tag{12.4}$$

where a, b, R are constants. For this case, compute v_p and v_T, using (12.2) and, again, using (12.3). Show that the results agree.

13. Derive (57), just as we derived (55).

14. In Section 13.6.4 we show how to transform the Laplace equation (52) under the change of variables $x = r\cos\theta, \ y = r\sin\theta$. Do likewise, for the given change of variables.

(a) $x = 2u + v, \quad y = 2u - v$
(b) $x = 2u + v, \quad y = 3u + v$
(c) $x = e^v, \quad y = u - v^2$

(d) $x = u^2 + v^2, \quad y = u + v$
(e) $x = u^2 \cos v, \quad y = u^2 \sin v$
(f) $x = u^2 + v, \quad y = u^2 - v$

15. In general, if we transform the Laplace equation (52) according to a change of variables

$$x = x(u, v), \quad y = y(u, v),$$

the equation gets changed; i.e., it is *not* of the form

$$T_{uu} + T_{vv} = 0. \tag{15.1}$$

However, show that if the transformation is such that

$$u_x = v_y, \quad u_y = -v_x, \quad \text{and} \quad u_x^2 + u_y^2 \neq 0,$$

then the Laplace equation is *preserved*; i.e., the result *is* (15.1). NOTE: This result is at the heart of the method of **conformal mapping**, to which we devote a later chapter.

13.7 Maxima and Minima

PREREQUISITE: Section 11.6 on quadratic forms is a prerequisite for Section 13.7.2.

13.7.1. Single variable case. Although our interest in this section is in functions of more than one variable, let us begin by reviewing from the calculus the theory of maxima and minima of functions of a single real variable.

We say that a function f has an **absolute maximum** at x_0 if $f(x) \leq f(x_0)$ for all x in the domain of definition of f, and an **absolute minimum** at x_0 if $f(x) \geq f(x_0)$ for all x in the domain. Further, it has a **local maximum** at a point x_0 in its domain if $f(x) \leq f(x_0)$ for all x in some neighborhood of x_0, and a **local minimum** at x_0 if $f(x) \geq f(x_0$ for all x in some neighborhood of x_0.* We use the term **extremum** to mean either a maximum or a minimum. For instance, a local extremum is either a local maximum or a local minimum.

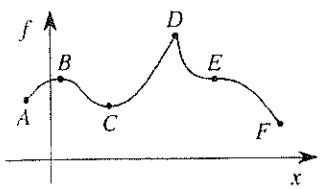


Figure 1. Maxima and minima.

To illustrate, consider the function f, the graph of which is shown in Fig. 1. We can see that f has extrema at A, B,C, D, and F (but not at E): a local minimum at A, a local maximum at B, a local minimum at C, a local and absolute maximum at D, and a local and absolute minimum at F. In this section we will study extrema such as B and C, at which the derivative of f exists and is zero.

THEOREM 13.7.1 *Vanishing Derivative, for Local Extremum*
For a function f of a single real variable x to have a local extremum at a point X in its domain of definition, where f is differentiable at X, it is necessary that

$$f'(X) = 0. \tag{1}$$

Proof: Suppose f has a local maximum at X. Then $f(x) \leq f(X)$ for all x's within some neighborhood N of X. Since f is differentiable at X,

*The term *relative* is sometimes used in place of *local*.

$$\lim_{x \to X} \frac{f(x) - f(X)}{x - X} \tag{2}$$

exists, independent of the manner in which x approaches X. Since $x \to X$, we can assume that x is within N, so $f(x) \leq f(X)$. If $x \to X$ from the right, then

$$\lim_{x \to X+} \frac{f(x) - f(X)}{x - X} = f'(X) \leq 0 \tag{3}$$

because $f(x) - f(X) \leq 0$ and $x - X > 0$, and if $x \to X$ from the left, then

$$\lim_{x \to X-} \frac{f(x) - f(X)}{x - X} = f'(X) \geq 0 \tag{4}$$

because $f(x) - f(X) \leq 0$ and $x - X < 0$. Comparison of (3) and (4) reveals that $f'(X) = 0$. A similar argument applies if f has a local minimum at X. ■

That the condition (1) is not also sufficient is evident from Fig. 1, for $f'(x) = 0$ at E, yet f has neither a local maximum at E nor a local minimum, it has a **horizontal inflection point** there. If the derivative $f'(x)$ is found to vanish at a point X but we do not have the graph of f to examine, how can we determine whether f has a local maximum, local minimum, or horizontal inflection point there?

THEOREM 13.7.2 *Maximum, Minimum, Horizontal Inflection Point*
Suppose that

$$f'(X) = f''(X) = \cdots = f^{(n-1)}(X) = 0,$$

but $f^{(n)}(X) \neq 0$, and that $f^{(n)}(x)$ is continuous in some neighborhood of X, where $n \geq 2$. If n is even, then f has a local maximum at X if $f^{(n)}(X) < 0$ and a local minimum at X if $f^{(n)}(X) > 0$. If n is odd, then f has a horizontal inflection point at X.

Proof: Since by assumption $f^{(n)}(x)$ is continuous in some neighborhood of X, there must exist a neighborhood $N(X)$ throughout which $f^{(n)}(x)$ has the same sign as $f^{(n)}(X)$.* By Taylor's formula with Lagrange remainder, we can express

$$f(x) = f(X) + 0 + \cdots + 0 + \frac{f^{(n)}(\xi)}{n!}(x - X)^n,$$

*If $F(x)$ is continuous at X, then, according to the definition of continuity, to each number $\epsilon > 0$ there corresponds a number $\delta > 0$ such that $|F(x) - F(X)| < \epsilon$ whenever $|x - X| < \delta$. Thus, if we choose ϵ to be smaller than $|F(X)|$, then there exists a $\delta > 0$ such that $F(x)$ has the same sign as $F(X)$ for all x in the neighborhood $|x - X| < \delta$.

where ξ is some point between X and x. If we require that x be in $N(X)$, then ξ must be in $N(X)$ as well. Thus, knowing that $f^{(n)}(\xi)$ has the same sign as $f^{(n)}(X)$ and that

$$f(x) - f(X) = \frac{f^{(n)}(\xi)}{n!}(x - X)^n, \tag{5}$$

then we can see from (5) that:

(1) if n is even and $f^{(n)}(X) > 0$, then $f(x) - f(X) \geq 0$ in $N(X)$, so f has a local minimum at X;

(2) if n is even and $f^{(n)}(X) < 0$, then $f(x) - f(X) \leq 0$ in $N(X)$, so f has a local maximum at X;

(3) and if n is odd, then [independent of the sign of $f^{(n)}(X)$] $f(x) - f(X)$ is positive to one side of X and negative to the other, so f has a horizontal inflection point at X,

which step completes the proof. ■

NOTE: It is sometimes stated that if $f'(X) = 0$, then f has a local maximum, local minimum, or a horizontal inflection point at X if $f''(X)$ is negative, positive, or zero, respectively. No, if $f''(X) = 0$ then no conclusion can be drawn, according to Theorem 13.7.2, until we determine the first nonvanishing derivative, $f^{(n)}(X)$.

EXAMPLE 1. Let $f(x) = (2\sqrt{x} - x - 1)^2$ over $0 < x < \infty$. Then

$$f'(x) = 2\left(\frac{1}{\sqrt{x}} - 1\right)(2\sqrt{x} - x - 1) = -\frac{2}{\sqrt{x}}(1 - \sqrt{x})^3, \tag{6}$$

which vanishes at $x = 1$. Differentiating further, at $x = 1$,

$$f'(1) = 0, \quad f''(1) = 0, \quad f'''(1) = 0, \quad f''''(1) = \frac{3}{2}.$$

Thus, $n = 4$, which is even, and $f''''(1) > 0$, so f has a minimum at $x = 1$. ▮

EXAMPLE 2. Let $f(x) = (x - 1)^2 \ln x$ over $0 < x < \infty$. Again $f'(1) = f''(1) = 0$, but $f'''(1) = 6$. This time n is odd so f has a horizontal inflection point at $x = 1$. ▮

13.7.2. Multivariable case. Let us extend these ideas to functions $f(x_1, \ldots, x_n)$ of n independent real variables, $x_1, \ldots, x_n$. We continue to be interested in *local* extrema at *interior* points of the domain of f, and we adopt the vector notation $\mathbf{x}$ as shorthand for the point $(x_1, \ldots, x_n)$ in the n-dimensional space, and $f(\mathbf{x})$ as shorthand for $f(x_1, \ldots, x_n)$. We say that f has a **local maximum** at a point $\mathbf{X}$ of its domain if $f(\mathbf{x}) \leq f(\mathbf{X})$ for all points $\mathbf{x}$ in some neighborhood of $\mathbf{X}$,

and a **local minimum** at $\mathbf{X}$ if $f(\mathbf{x}) \geq f(\mathbf{X})$ for all points $\mathbf{x}$ in some neighborhood of $\mathbf{X}$.

THEOREM 13.7.3 *Vanishing Partial Derivatives for Local Extremum*
For a function f of n real variables $x_1, \ldots, x_n$ to have a local extremum at a point $\mathbf{X} = (X_1, \ldots, X_n)$ in its domain of definition, where f is C^1 in some neighborhood of $\mathbf{X}$, it is necessary that

$$\frac{\partial f}{\partial x_1} = 0, \quad \frac{\partial f}{\partial x_2} = 0, \ \ldots, \quad \frac{\partial f}{\partial x_n} = 0 \tag{7}$$

at $\mathbf{X}$.

Proof: Let $x_1 = x_1(\tau), \ldots, x_n = x_n(\tau)$ be parametric equations of any curve through $\mathbf{x} = \mathbf{X}$ in n-space, with $\mathbf{x} = \mathbf{X}$ at $\tau = 0$, say. Further, let $F(\tau) \equiv f(\mathbf{x}(\tau))$. Since f has been assumed to be C^1, we can use the chain rule to express

$$\frac{dF}{d\tau} = \frac{\partial f}{\partial x_1}\frac{dx_1}{d\tau} + \cdots + \frac{\partial f}{\partial x_n}\frac{dx_n}{d\tau}.$$

Since $f(\mathbf{x})$ has an extremum at $\mathbf{X}$, $F(\tau)$ has an extremum at $\tau = 0$ so, according to Theorem 13.7.1, it is necessary that

$$\frac{dF}{d\tau}(0) = \frac{\partial f}{\partial x_1}(\mathbf{X})\frac{dx_1}{d\tau}(0) + \cdots + \frac{\partial f}{\partial x_n}(\mathbf{X})\frac{dx_n}{d\tau}(0) = 0. \tag{8}$$

Since (8) must hold for *every* path $\mathbf{x} = \mathbf{x}(\tau)$ through $\mathbf{X}$, it follows from (8) that $\frac{\partial f}{\partial x_1}(\mathbf{X}) = 0, \ldots, \frac{\partial f}{\partial x_n}(\mathbf{X}) = 0$, as was to be shown. That is, since $\mathbf{x}(\tau)$ is arbitrary we can choose it so that $\frac{dx_1}{d\tau}(0)$ is nonzero and the other derivatives, $\frac{dx_2}{d\tau}(0), \ldots, \frac{dx_n}{d\tau}(0)$, are zero, in which case we learn from (8) that $\frac{\partial f}{\partial x_1}(\mathbf{X}) = 0$. Similarly, we can let $\frac{dx_2}{d\tau}(0)$ be nonzero and the other derivatives be zero, and learn from (8) that $\frac{\partial f}{\partial x_2}(\mathbf{X}) = 0$, and so on. ■

Before continuing, there is a point of rigor to address. Namely, why do we bother introducing the space curve $\mathbf{x} = \mathbf{x}(\tau)$ in the preceding proof? Why not use the simpler looking *differential* form of the chain rule

$$df = \frac{\partial f}{\partial x_1}(\mathbf{X})dx_1 + \cdots + \frac{\partial f}{\partial x_n}(\mathbf{X})dx_n = 0 \tag{9}$$

in place of the *derivative* form (8)? We could argue that since $x_1, \ldots, x_n$ are independent variables, the increments $dx_1, \ldots, dx_n$ are independent as well. Thus, letting dx_1 be nonzero and $dx_2 = \cdots = dx_n = 0$, we learn from (9) that $\dfrac{\partial f}{\partial x_1}(\mathbf{X}) = 0$, and so on, as before. The problem with that approach is that the differential version (9) treats df, dx_1, ... , dx_n as though they are finite numbers, whereas they are not. Nonetheless, in Section 13.7.3 we will use (9) for simplicity, with the understanding that we can always change to the (impeccable) derivative form if we wish to.

Let us continue. Any point $\mathbf{X}$ at which (7) holds is called a **critical point** of f. As emphasized below Theorem 13.7.1 for the case of functions of a single variable, the condition (7) [like the condition (1) in Theorem 13.7.1] is necessary but not sufficient for the existence of an extremum of f at $\mathbf{X}$. For instance, if

$$f(x_1, x_2) = x_1^2 - x_2^2, \tag{10}$$

then $\partial f/\partial x_1 = 2x_1 = 0$ all along the line $x_1 = 0$ (i.e., the x_2 axis) in the x_1, x_2 plane, and $\partial f/\partial x_2 = -2x_2 = 0$ all along the line $x_2 = 0$, so both are zero at the intersection of those lines, namely, at the origin, $\mathbf{X} = (0, 0)$. However, along the x_1 axis $f(x_1, 0) = x_1^2$ has a "valley," a minimum, whereas along the x_2 axis $f(0, x_2) = -x_2^2$ has a "mountain," a maximum, so the flat spot on the f surface at $(0, 0)$ is neither a local maximum nor a local minimum. Rather, we call it a "saddle" because it resembles an equestrian saddle, a valley one way and a mountain the other (Fig. 2).

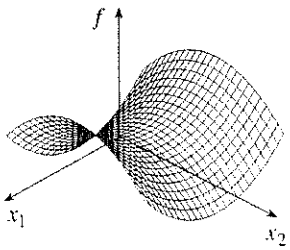


Figure 2. Saddle at $(0, 0)$.

Extending this idea to higher dimensions, we say that $f(\mathbf{x})$ has a **saddle** at a point $\mathbf{X}$ in n-dimensional space if f is C^1 in some neighborhood of $\mathbf{X}$, if $\partial f/\partial x_1 = \cdots = \partial f/\partial x_n = 0$ at $\mathbf{X}$, and if $f(\mathbf{x}) - f(\mathbf{X})$ takes on both positive and negative values in every neighborhood (no matter how small) of $\mathbf{X}$. Of course, if $n = 1$ the saddle is actually a horizontal inflection point, and if $n \geq 3$ we cannot display the graph of f as we have done in Fig. 2 since we would need four or more dimensions.

Given a function $f(x_1, \ldots, x_n)$, suppose that, using (7), we determine a critical point $\mathbf{X}$ of f. How can we determine whether f has a local maximum, a local minimum, or a saddle there? Let us explain the essential idea briefly and heuristically.

For a function of a single variable, $f(x)$, with $f'(X) = 0$, we used Taylor's formula to write

$$f(x) = f(X) + \frac{1}{2} f''(\xi)(x - X)^2 \tag{11}$$

for some ξ between X and x so

$$f(x) - f(X) \sim \frac{1}{2} f''(X)(x - X)^2 \tag{12}$$

as $x \to X$. If $f''(X) < 0$ then the right-hand side of (12) represents a "mountain" and f has a local maximum at X; if $f''(X) > 0$ then the right side represents a "valley" and f has a local minimum; if $f''(X) = 0$ no information is obtained and

we need to include more terms in Taylor's formula, namely, the first nonvanishing term beyond the initial $f(X)$ term.

For functions of more than one variable the idea is the same. Consider two variables, for simplicity. Given $f(x_1, x_2)$, with $\partial f/\partial x_1 = \partial f/\partial x_2 = 0$ at $\mathbf{X} = (X_1, X_2)$, use Taylor's formula to write

$$f(x_1, x_2) - f(X_1, X_2) \sim \frac{1}{2!}\left[f_{x_1x_1}(\mathbf{X})(x_1 - X_1)^2 + f_{x_2x_2}(\mathbf{X})(x_2 - X_2)^2 + 2f_{x_1x_2}(\mathbf{X})(x_1 - X_1)(x_2 - X_2)\right]. \tag{13}$$

Observe that the right side of (13) is a *quadratic form* in $x_1 - X_1$ and $x_2 - X_2$ so it can be reduced to canonical form as was explained in Section 11.6. Thus, (13) can be re-expressed as

$$f(x_1, x_2) - f(X_1, X_2) \sim \lambda_1\widetilde{x}_1^2 + \lambda_2\widetilde{x}_2^2, \tag{14}$$

where λ_1 and λ_2 are the eigenvalues of the matrix

$$\mathbf{A} = \frac{1}{2}\begin{bmatrix} f_{x_1x_1}(\mathbf{X}) & f_{x_1x_2}(\mathbf{X}) \\ f_{x_1x_2}(\mathbf{X}) & f_{x_2x_2}(\mathbf{X}) \end{bmatrix}, \tag{15}$$

and where $\widetilde{\mathbf{x}} = [\widetilde{x}_1, \widetilde{x}_2]^T$ and $\mathbf{x} = [x_1 - X_1, x_2 - X_2]^T$ are related through the modal matrix transformation $\mathbf{x} = \mathbf{Q}\widetilde{\mathbf{x}}$. Since (14) expresses $f(x_1, x_2) - f(X_1, X_2)$ as a sum of squares [just as (12) expresses $f(x) - f(X)$ as a square] it reveals whether f has a local maximum, local minimum, or saddle at $\mathbf{X}$: if both λ_1 and λ_2 are negative then f has a local maximum there; if both are positive then f has a local minimum; and if one is positive and one is negative then f has a saddle.

In fact, we have the following theorem.

THEOREM 13.7.4 *Maximum, Minimum, Saddle*
For a function f of n variables $x_1, \ldots, x_n$, let

$$\frac{\partial f}{\partial x_1} = 0, \quad \frac{\partial f}{\partial x_2} = 0, \quad \ldots, \quad \frac{\partial f}{\partial x_n} = 0 \tag{16}$$

at a point $\mathbf{X}$ in f's domain of definition, let f be C^2 in some neighborhood of $\mathbf{X}$, and define the $n \times n$ matrix

$$\mathbf{A} = \begin{bmatrix} f_{x_1x_1}(\mathbf{X}) & f_{x_1x_2}(\mathbf{X}) & \cdots & f_{x_1x_n}(\mathbf{X}) \\ f_{x_2x_1}(\mathbf{X}) & f_{x_2x_2}(\mathbf{X}) & \cdots & f_{x_2x_n}(\mathbf{X}) \\ \vdots & & & \vdots \\ f_{x_nx_1}(\mathbf{X}) & f_{x_nx_2}(\mathbf{X}) & \cdots & f_{x_nx_n}(\mathbf{X}) \end{bmatrix}. \tag{17}$$

Let $\det\mathbf{A}$ be nonzero. If $\mathbf{A}$ is positive definite then f has a local minimum at $\mathbf{X}$, if it is negative definite then f has a local maximum at $\mathbf{X}$, and if it has at least one positive eigenvalue and at least one negative eigenvalue then f has a saddle at $\mathbf{X}$.

Proof: Before beginning the proof let us make several remarks. First, the 1/2 present in (15) has been omitted in (17) because the theorem concerns itself only with the *sign* of the eigenvalues, not with their magnitude. Also, observe that $f_{x_j x_k}(\mathbf{X}) = f_{x_k x_j}(\mathbf{X})$ because f has been assumed to be C^2 in some neighborhood of $\mathbf{X}$; thus, $\mathbf{A}$ is symmetric. Finally, why do we ask $\det\mathbf{A}$ to be nonzero? If $\det\mathbf{A} = 0$, then $\lambda = 0$ is among the eigenvalues of $\mathbf{A}$, which borderline case we wish to avoid. For recall from Theorem 13.7.2 (for the single-variable case) that if $f'(X) = 0$ *and* $f''(X) = 0$ then we need to keep going until we come to the first *non*vanishing derivative at X before we can tell if f has a maximum, minimum, or saddle at X. For the more difficult multivariable case we prefer to avoid such additional complications. In any case, $\det\mathbf{A} \neq 0$ is the generic case, so we are not giving up very much.

To prove the theorem we begin with Taylor's formula

$$f(\mathbf{x}) - f(\mathbf{X}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} f_{x_i x_j}(\xi)(x_i - X_i)(x_j - X_j) \tag{18}$$

with the first-order terms omitted by virtue of (16), where $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{X} = (X_1, \ldots, X_n)$, and $\xi = (\xi_1, \ldots, \xi_n)$ is some point on the line connecting $\mathbf{x}$ and $\mathbf{X}$.* Since f is C^2 in some neighborhood of $\mathbf{X}$, there is some neighborhood $N(\mathbf{X})$ throughout which the $f_{x_i x_j}(\xi)$ coefficients in (18) have the same sign as $f_{x_i x_j}(\mathbf{X})$. Thus, in place of the right side of (18) we can use the quadratic form

$$\sum_{i=1}^{N}\sum_{j=1}^{N} f_{x_i x_j}(\mathbf{X})(x_i - X_i)(x_j - X_j) \tag{19}$$

to determine if $f(\mathbf{x}) - f(\mathbf{X})$ is less than or equal to zero, greater than or equal to zero, or of mixed sign in $N(\mathbf{X})$ and, from the theory of quadratic forms, these circumstances correspond to $\mathbf{A}$ being negative definite, positive definite, or having eigenvalues of mixed sign, respectively. ■

EXAMPLE 3. Identify and classify any local extrema and saddles of the function

$$f(x, y) = \ln[2x(y-1) + 1]. \tag{20}$$

(Here, x_1 is x and x_2 is y.) Setting

$$f_x(x, y) = \frac{2(y-1)}{2x(y-1)+1} = 0 \quad \text{and} \quad f_y(x, y) = \frac{2x}{2x(y-1)+1} = 0 \tag{21}$$

gives the single critical point $x = 0$, $y = 1$. Next,

$$f_{xx}(x, y) = -\frac{4(y-1)^2}{[2x(y-1)+1]^2},$$

*That is, $\xi_1 = x_1 + (X_1 - x_1)\tau, \ldots, \xi_n = x_n + (X_n - x_n)\tau$ for some τ in $0 \le \tau \le 1$. These are parametric equations of a **straight line** from $\mathbf{x}$ to $\mathbf{X}$ in n-space.

$$f_{yy}(x,y) = -\frac{4x^2}{[2x(y-1)+1]^2}, \tag{22}$$
$$f_{xy}(x,y) = \frac{2}{[2x(y-1)+1]^2},$$

so we see from (20)–(22) that $f, f_x, f_y, f_{xx}, f_{yy}$, and f_{xy} are continuous everywhere in the x, y plane except along the curve $2x(y-1)+1=0$ (shown as C in Fig. 3), along which they are undefined. Thus, f is C^2 in any neighborhood, about (0,1), of radius R or less (Fig. 3). Next,

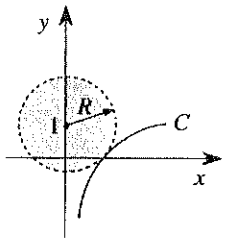


Figure 3. Neighborhood of (0,1).

$$\mathbf{A} = \begin{bmatrix} f_{xx}(0,1) & f_{xy}(0,1) \\ f_{xy}(0,1) & f_{yy}(0,1) \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix}, \tag{23}$$

with eigenvalues $\lambda = +2, -2$. The latter are of mixed sign so, according to Theorem 13.7.4, f has a saddle at (0,1).

COMMENT 1. We did not bother verifying that $\det\mathbf{A} \neq 0$ because that condition merely served to disallow zero as an eigenvalue and, indeed, we found that $\lambda = \pm 2$.

COMMENT 2. Even without reducing the quadratic form

$$f(x,y) - f(0,1) \sim \frac{1}{2!}\left[f_{xx}(0,1)x^2 + f_{yy}(0,1)(y-1)^2 + 2f_{xy}(0,1)x(y-1)\right], \tag{24}$$

or,

$$2\left[f(x,y) - f(0,1)\right] \sim \left[0x^2 + 0(y-1)^2 + 4x(y-1)\right] = 4x(y-1), \tag{25}$$

to canonical form, we can see from (25) that $f(x,y) - f(0,1)$ is of mixed sign in every neighborhood of (0,1) as indicated in Fig. 4. ∎

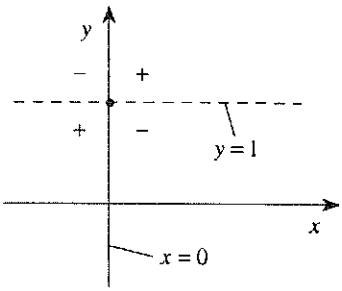


Figure 4. Mixed sign of $2x(y-1)$ near (0,1).

EXAMPLE 4. Consider

$$f(x,y,z) = \sin(x^2+y^2+z^2) + xy + xz + yz. \tag{26}$$

Working out f_x, f_y, and f_z, it is easily found that (0,0,0) is a critical point of f (although not the only one; see Exercise 9). Working out the second-order partial derivatives as well, we find that f is indeed C^2 in a neighborhood of (0,0,0); in fact, it is C^2 everywhere. Further, $f_{xx} = f_{yy} = f_{zz} = 2$ and $f_{xy} = f_{xz} = f_{yz} = 1$ at (0,0,0) so

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}, \tag{27}$$

the eigenvalues of which are $\lambda = 1, 1, 4$. Since all of the λ's are positive, f has a local minimum at the critical point (0,0,0). ∎

EXAMPLE 5. Consider

$$f(x,y) = x^3y - xy^3. \tag{28}$$

Then (Exercise 4) there is only one critical point, at (0,0), and $f_{xx} = f_{yy} = f_{xy} = 0$ there, so

$$\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}. \tag{29}$$

Since $\det\mathbf{A} = 0$ (and $\lambda = 0, 0$), Theorem 13.7.4 *does not apply*; (0,0) is said to be a **higher-order critical point**.

Of course, higher-order critical points occur for functions of a single variable as well, as in Example 1, where we had $f''(1) = f'''(1) = 0$ and we had to look to $f''''(1)$ for clarification. But whereas Theorem 13.7.2 covered such higher-order critical points, Theorem 13.7.4 does not.

In the present example inspection happens to suffice, for if we factor f as $xy(x - y)(x + y)$ we can see that f is alternately positive and negative in 45° wedges as shown in Fig. 5. Hence, the critical point of f at (0,0) is a saddle. ∎

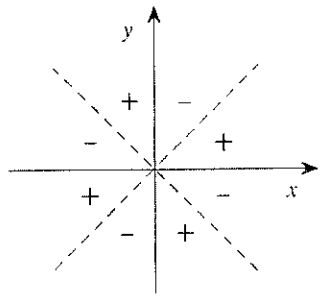


Figure 5. Sign of $f(x, y) = x^3y - xy^3$.

An important physical application of these ideas involves equilibrium states and their stability, because many force fields (such as gravitational force fields) have associated with them a **potential energy** $V(x, y, z)$, such that $\partial V/\partial x$, $\partial V/\partial y$, $\partial V/\partial z$ are the x, y, z force components, respectively, at any field point (x, y, z). Thus, the critical points of V are the **equilibrium points** of the force field, points at which the force is zero. If V has a local minimum at such a point, then the equilibrium at that point is classified as stable; if V has a local maximum or saddle there, then the equilibrium at that point is unstable.

For example, it is known from **Coulomb's law** of electrostatics that the **electric potential** $V(x, y, z)$ at any field point x, y, z, induced by a charge of strength Q coulombs (where Q can be positive or negative) at x_0, y_0, z_0 is proportional to Q and inversely proportional to the distance between x, y, z and x_0, y_0, z_0. Letting the constant of proportionality be k, then

$$V(x, y, z) = k\frac{Q}{\sqrt{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2}}, \tag{30}$$

and the force components induced at any point x, y, z in the coordinate directions, per unit charge at x, y, z, are given by $F_x = \partial V/\partial x$, $F_y = \partial V/\partial y$, $F_z = \partial V/\partial z$, respectively.

EXAMPLE 6. *Equilibrium Points of Charge Distribution.* Let three charges, each of strength Q coulombs, be placed in the x, y plane at the points (1,0), (0,0), and (0,1), respectively (Fig. 6). Find any equilibrium points and determine their stability. Superimposing the potentials induced by the individual charges gives

$$V(x, y) = kQ\left(\frac{1}{\sqrt{x^2 + y^2}} + \frac{1}{\sqrt{(x - 1)^2 + y^2}} + \frac{1}{\sqrt{x^2 + (y - 1)^2}}\right). \tag{31}$$

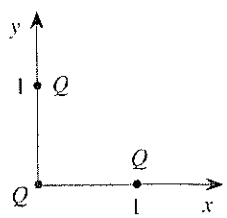


Figure 6. Three charges.

Using computer algebra we can evaluate $\partial V/\partial x$ and $\partial V/\partial y$ and solve the equations

$$\frac{\partial V}{\partial x}(x, y) = 0, \qquad \frac{\partial V}{\partial y}(x, y) = 0 \tag{32}$$

for x, y; the *Maple* commands are given at the end of this section. The result is

$$X = 0.0475, \qquad Y = 0.5130. \tag{33}$$

Next, evaluating $\partial^2 V/\partial x^2$, $\partial^2 V/\partial x \partial y$, $\partial^2 V/\partial y^2$ at that point gives, for the $\mathbf{A}$ matrix in (17),

$$\mathbf{A} = \begin{bmatrix} V_{xx}(X,Y) & V_{xy}(X,Y) \\ V_{xy}(X,Y) & V_{yy}(Y,Y) \end{bmatrix} = \begin{bmatrix} -14.375 & -1.448 \\ -1.448 & 31.014 \end{bmatrix}, \tag{34}$$

with eigenvalues

$$\lambda = -14.42, \quad 31.06. \tag{35}$$

Since the λ's are of mixed sign the equilibrium point (33) is unstable. Physically, that means that if a charge is placed at that point, then even the slightest displacement will cause it to move away.

COMMENT. Does it not seem peculiar that the equilibrium point (33) is not on the axis of symmetry, the line $y = x$? This point is discussed at the end of the section when we give the *Maple* commands used in this example. ∎

13.7.3. Constrained extrema and Lagrange multipliers. In Section 13.7.2 we studied the problem of finding one or more points, if any, that extremize a given function f:

$$f(x_1, \ldots, x_n) = \text{extremum}, \tag{36}$$

where, by an **extremum** we mean a local maximum or a local minimum. Now, however, we consider the problem of extremizing a given function subject to one or more **constraints**.

To illustrate, suppose we wish to find the point on the plane $2x - y + 4z = 3$ that is closest to the origin. Since the distance from the origin to (x, y, z) is $\sqrt{x^2 + y^2 + z^2}$, we wish to minimize the function $\sqrt{x^2 + y^2 + z^2}$ subject to the constraint that x, y, z are related according to $2x - y + 4z = 3$. Surely, if we minimize the distance we will also be minimizing the square of the distance, $x^2 + y^2 + z^2$, so let us state the problem in the more convenient form

$$f(x, y, z) = x^2 + y^2 + z^2 = \text{minimum}, \tag{37a}$$

subject to the constraint

$$g(x, y, z) = 2x - y + z = 3. \tag{37b}$$

There is only one function being extremized, but there could be more than one constraint:

$$f(x_1, \ldots, x_n) = \text{extremum}, \tag{38a}$$

subject to the constraints

$$\begin{aligned} g_1(x_1, \ldots, x_n) &= c_1, \\ &\vdots \\ g_k(x_1, \ldots, x_n) &= c_k. \end{aligned} \tag{38b}$$

In this discussion let $n = 3$, for definiteness, and let $k = 1$; other cases will be found in the exercises, and the choice $n = 3$ and $k = 1$ should suffice in explaining the main ideas. Thus, we consider the problem

$$f(x, y, z) = \text{extremum}, \tag{39a}$$
$$g(x, y, z) = c, \tag{39b}$$

where we are using the simpler notation x, y, z in place of x_1, x_2, x_3. We assume, throughout this Section 13.7.3 that f and g are C^1 in some neighborhood of the point under consideration, at which f has an extremum.

To begin, we write

$$df = \frac{\partial f}{\partial x}(\mathbf{X})dx + \frac{\partial f}{\partial y}(\mathbf{X})dy + \frac{\partial f}{\partial z}(\mathbf{X})dz = 0 \tag{40}$$

at $\mathbf{X}$. [That is, let us use the more convenient differential form of the chain rule, as in (9), rather than the derivative form, as in (8).] However, and this is a key point, we cannot use the arbitrariness of the dx, dy, dz increments to infer that $\partial f/\partial x = \partial f/\partial y = \partial f/\partial z = 0$ at X, because those increments are *not* arbitrary. Rather, x, y, z are related through the constraint equation (39b) so the dx, dy, dz increments are related as well.

One way out of this difficulty is to solve (39b) for z as a function of x and y, which will be possible if $g_z(\mathbf{X}) \neq 0$. Then, if (39b) gives us $z = Z(x, y)$, we can put that result into (39a) and write

$$f(x, y, Z(x, y)) \equiv F(x, y) = \text{extremum}. \tag{41}$$

Although it is true that x, y, z are not independent variables (for they are related through the constraint equation), x and y, by themselves, *are* independent.* Thus,

$$dF = \frac{\partial F}{\partial x}dx + \frac{\partial F}{\partial y}dy = 0 \tag{42}$$

does imply that

$$\frac{\partial F}{\partial x} = 0, \tag{43a}$$
$$\frac{\partial F}{\partial y} = 0, \tag{43b}$$

because dx and dy in (42) are arbitrary independent increments. The idea, then, is to solve (43a,b) for x and y; then, $z = Z(x, y)$ gives z.

EXAMPLE 7. Let us use this method to solve the problem stated in (37), above. Solving (37b) for z gives $z = 3 - 2x + y$ so

$$f(x, y, Z(x, y)) = x^2 + y^2 + (3 - 2x + y)^2 \equiv F(x, y). \tag{44}$$

*That is, we can assign values to x and y independently. Then, z is not arbitrary but follows from the constraint equation (39b).

Then,

$$\begin{aligned} \frac{\partial F}{\partial x} &= 2x - 4(3 - 2x + y) = 0, \\ \frac{\partial F}{\partial y} &= 2y + 2(3 - 2x + y) = 0 \end{aligned} \tag{45}$$

so $x = 1$ and $y = -1/2$. Finally, $z = 3 - 2x + y = 3 - 2 - 1/2 = 1/2$. To verify that $\mathbf{X} = (1, -\frac{1}{2}, \frac{1}{2})$ does give a minimum [as specified in (37a)], and not a maximum or a saddle, we can apply Theorem 13.7.4 to $F(x, y)$. We find that

$$\mathbf{A} = \begin{bmatrix} F_{xx} & F_{xy} \\ F_{xy} & F_{yy} \end{bmatrix} = \begin{bmatrix} 10 & -4 \\ -4 & 4 \end{bmatrix}$$

has eigenvalues $\lambda = 2$ and 12. Since both are positive, $\mathbf{X} = (1, -\frac{1}{2}, \frac{1}{2})$ does give a minimum, as desired. ∎

The essential idea behind the method described above is to use the constraint equations to eliminate variables so that those remaining are independent. Let us call that procedure the **method of elimination**, to distinguish it from the **method of Lagrange multipliers** due the the great French mathematician *Joseph Louis Lagrange* (1736–1813). Like Euler, Lagrange also worked on the applications of mathematics to mechanics.

To explain Lagrange's method, consider once again the representative problem (39). Besides df vanishing at an extremum, so does dg (indeed, $dg = 0$ everywhere because g is a constant). Thus,

$$df = f_x dx + f_y dy + f_z dz = 0, \tag{46a}$$

$$dg = g_x dx + g_y dy + g_z dz = 0, \tag{46b}$$

from which it follows that

$$df - \lambda dg = (f_x - \lambda g_x)dx + (f_y - \lambda g_y)dy + (f_z - \lambda g_z)dz = 0, \tag{47}$$

where λ is a yet to be determined parameter, the so-called **Lagrange multiplier**. Remember that x, y, z are not independent variables because they are related through (39b). Thus, we cannot argue that the dx, dy, dz increments in (47) are arbitrary and hence that each of the coefficients $f_x - \lambda g_x$, $f_y - \lambda g_y$, $f_z - \lambda g_z$ must vanish at the extremum $\mathbf{X}$. However, we can reach that same conclusion by a different logic. For suppose that $g_z \neq 0$ at $\mathbf{X}$. Then, by choosing λ to be f_z/g_z (which number exists because $g_z \neq 0$ at $\mathbf{X}$), (47) reduces to

$$(f_x - \lambda g_x)dx + (f_y - \lambda g_y)dy = 0. \tag{48}$$

[When we write $\lambda = f_z/g_z$, we do not fear that we are setting a constant equal to a function of x, y, z, for (46) and (47) hold *at* $\mathbf{X}$, so f_x, g_x, f_y, ..., really mean $f_x(\mathbf{X})$, $g_x(\mathbf{X})$, $f_y(\mathbf{X})$, ... ; that is, they are numbers, not functions.] Now, since x and y, by themselves, are independent variables, we can argue that dx and dy

in (48) *are* arbitrary independent increments so that the coefficients $f_x - \lambda g_x$ and $f_y - \lambda g_y$ must vanish at $\mathbf{X}$.

Consequently, we have the four equations

$$f_x - \lambda g_x = 0, \quad f_y - \lambda g_y = 0, \quad f_z - \lambda g_z = 0 \tag{49a}$$

and

$$g = c \tag{49b}$$

on the four unknowns x, y, z, and λ. In effect, the Lagrange multiplier method amounts to forming a new function,

$$\boxed{f^* \equiv f - \lambda g} \tag{50}$$

(or $f + \lambda g$; it doesn't matter), and extremizing f^* subject to *no* constraints so that $f_x^* = f_y^* = f_z^* = 0$. , which equations are identical to (49a). Observe that even though we already used the constraint equation $g = c$ in obtaining (46b), we still need to append $g = c$ to equations (49) because we lost information when we took the differential of $g = c$, in (46b), in that the constant c thereby dropped out.

EXAMPLE 7. Let us solve the problem stated in (37) once again, this time using the Lagrange multiplier method. With

$$f^* = f - \lambda g = x^2 + y^2 + z^2 - \lambda(2x - y + z),$$

we have

$$\begin{aligned} f_x^* &= 2x - 2\lambda = 0, \\ f_y^* &= 2y + \lambda = 0, \\ f_z^* &= 2z - \lambda = 0 \end{aligned} \tag{51a}$$

and

$$g = 2x - y + z = 3. \tag{51b}$$

Solving (51) gives $x = 1$, $y = -1/2$, $z = 1/2$, and $\lambda = 1$. Thus, $\mathbf{X} = (1, -\frac{1}{2}, \frac{1}{2})$ as found in Example 7 by the method of elimination.

COMMENT. In this problem we have no special interest in the ultimate value of λ; λ is simply an auxiliary variable. However, in physical applications λ often turns out to have physical significance. For instance, in studying the dynamics of the motion of a bead along a curved frictionless wire λ might turn out to be the normal force between the bead and the wire. ∎

EXAMPLE 8. Suppose that we wish to construct a tin can in the shape of a rectangular prism. Let the base dimensions be x by y and let the height be z, such that the volume xyz is a prescribed value V. The base is to be twice as thick as the other five sides so its cost per unit area is 2α, where α is the cost per unit area of the other sides. Seeking to minimize the material cost gives the problem:

$$\text{Material cost:} \quad f(x, y, z) = \alpha(3xy + 2xz + 2yz) = \text{minimum}, \tag{52a}$$

subject to the constraint

$$\text{Volume:} \quad g(x, y, z) = xyz = V. \tag{52b}$$

Defining $f^* = f - \lambda g$, set

$$\begin{aligned} f_x^* &= f_x - \lambda g_x = \alpha(3y + 2z) - \lambda yz = 0, \\ f_y^* &= f_y - \lambda g_y = \alpha(3x + 2z) - \lambda xz = 0, \\ f_z^* &= f_z - \lambda g_z = \alpha(2x + 2y) - \lambda xy = 0. \end{aligned} \tag{53}$$

Solving the four equations (52b) and (53) gives $x = y = (2V/3)^{1/3}$, $z = \frac{3}{2}(2V/3)^{1/3}$. ∎

Closure. In this section we review the theory of maxima and minima for functions of a single variable before turning to the unconstrained and constrained extremization of functions of more than one variable, considering only local extrema at interior points of the domain of the function f being extremized. If $f(x_1, \ldots, x_n)$ is C^1 in some neighborhood of the point of extremum, $\mathbf{X}$, then the extremum condition $df = 0$ implies that we must have

$$\frac{\partial f}{\partial x_1} = \cdots = \frac{\partial f}{\partial x_n} = 0 \tag{54}$$

at $\mathbf{X}$. Points at which (54) holds are called critical points of f. However, f need not have an extremum (local maximum or local minimum) at a critical point; it could have a saddle there. Assuming a bit more of f (namely, that it be C^2 rather than C^1 in some neighborhood of $\mathbf{X}$) and looking to the second-order terms in Taylor's formula, we show in Theorem 13.7.4 that everything hinges on the eigenvalues of the matrix $\mathbf{A} = \{f_{x_i x_j}(\mathbf{X})\}$ associated with f: if all of $\mathbf{A}$'s eigenvalues are negative then f has a local maximum at $\mathbf{X}$, if all are positive then f has a local minimum there, and if they are mixed in sign then f has a saddle there.

For the extremization of f in the presence of one or more constraints of the form $g = c$, the extremum condition $df = 0$ still holds but no longer implies (54). We study two methods of solving such problems, the method of elimination and the method of Lagrange multipliers. Working with three independent variables (x, y, z) for definiteness, we find that both methods work if $g_z(\mathbf{X}) \neq 0$ [or, failing that, if either $g_y(\mathbf{X}) \neq 0$ or $g_z(\mathbf{X}) \neq 0$].

In closing, let us mention an important type of extremization problem that does not fit the categories covered in this section. Namely, we seek $x_1, \ldots, x_n$ so that

$$f(x_1, \ldots, x_n) = b_1 x_1 + \cdots + b_n x_n = \text{extremum}, \tag{55a}$$

subject to the constraints

$$\begin{aligned} g_1(x_1, \ldots, x_n) &= a_{11} x_1 + \cdots + a_{1n} x_n = c_1, \\ &\vdots \\ g_k(x_1, \ldots, x_n) &= a_{k1} x_1 + \cdots + a_{kn} x_n = c_k, \end{aligned} \tag{55b}$$

where all the x_j's are required to be *nonnegative*. Or the constraints might be *inequalities* such as

$$a_{11}x_1 + \cdots + a_{1n}x_n \leq c_1, \tag{56}$$

but this case is equivalent to the one stated above since we can introduce an additional variable x_{n+1} and rewrite (56) as

$$a_{11}x_1 + \cdots + a_{1n}x_n + x_{n+1} = c_1$$

with the proviso that x_{n+1} is nonnegative too; x_{n+1} is called a **slack variable**.

Such problems are of great interest in the field known as **operations research**, often abbreviated as **OR**. Solutions are obtained not by setting various partial derivatives equal to zero but by special techniques such as the **simplex method**, which fall under the general heading of **linear programming**. This topic is beyond our present scope.*

Computer software. Let us go through the *Maple* commands used in Example 6. First, define the function V by entering

$$V := (x\hat{}2+y\hat{}2)\hat{}(-1/2)+((x-1)\hat{}2+y\hat{}2)\hat{}(-1/2)+(x\hat{}2+(y-1)\hat{}2)\hat{}(-1/2)) :$$

We can obtain $\partial V/\partial x$ by a diff(V, x) command, but since we will subsequently wish to call that quantity, let us both evaluate it and give it the name Vx at the same time. Similarly for $\partial V/\partial y$. Thus, enter

$$Vx := \text{diff}(V, x) :$$

and

$$Vy := \text{diff}(V, y) :$$

(Whereas we've been using colons you may prefer semicolons so that the results will be printed.) To find the solution of (32), use the fsolve command, and define the solution(s) as X, Y at the same time:

$$XY := \text{fsolve}(\{Vx = 0, Vy = 0\}, \{x, y\}, \{x = 0..1, y = 0..1\});$$

where we have used the optional $\{x = 0..1, y = 0..1\}$ to narrow the solution search because it can be seen by inspection (Exercise 11) that the equilibrium point(s), if any, must lie in that unit square. The output is

$$XY := \{y = .5129880394, x = .04746059840\}$$

Next,

$$Vxx := \text{diff}(V, x, x);$$
$$Vxy := \text{diff}(V, x, y);$$
$$Vyy := \text{diff}(V, y, y);$$

*See, for instance, G. B. Dantzig, *Linear Programming and Extensions*, 2nd ed. (Princeton, NJ: Princeton University Press, 1963) or R. Fletcher, *Practical Methods of Optimization*, 2nd ed. (New York: Wiley, 1988).

give $\partial^2 V/\partial x^2$, $\partial^2 V/\partial x \partial y$, $\partial^2 V/\partial y^2$. To evaluate these second-order derivatives at X, Y and define them as the elements of the **A** matrix, enter

$$a11 := \text{eval(subs}(XY, Vxx));$$
$$a12 := \text{eval(subs}(XY, Vxy));$$
$$a22 := \text{eval(subs}(XY, Vyy));$$

which gives

$$a11 := -14.37506987$$
$$a12 := -1.448243172$$
$$a22 := 31.01365215$$

respectively. Then enter

with(linalg):

to access the linear algebra package. Finally,

$$A := \text{matrix}(2, 2, [a11, a12, a12, a22]);$$

and

eigenvals(");

give the eigenvalues

$$-14.42123281, \quad 31.05981510$$

as stated in the example.

Note the use of the "**subs**" in the eval commands. In case that usage is not clear, let us give another example: to evaluate $3x^2$ at $x = 5$ use the command eval(subs(5,3 $*$ x^2)). Note also the strange outcome, that the equilibrium point is not on the line of symmetry $y = x$ even though both the potential $V(x, y)$ is a symmetric function of x and y or, said differently, even though the physical arrangement of charges is symmetric about that line. (You might want to think about that paradox before reading on.) However, we must realize that the fsolve command does not necessarily give all solutions; in fact, if we do not use the option to narrow the search to the unit square then the command gives no solutions at all. If there *is* symmetry about the line $y = x$ then there must be at least one more solution within the unit square, at the point $x = 0.5129880394, y = 0.04746059840$. To test that idea we can change the $\{x = 0..1, y = 0..1\}$ option to $\{x = 0.5..0.52, y = 0.045..0.050\}$, say. That change does indeed produce the suspected solution, so there are *two* equilibrium points and they are arranged symmetrically about the line $y = x$.

EXERCISES 13.7

1. The given function has a critical point at $x = 1$. Classify it as a local maximum, local minimum, or horizontal inflection point.

(a) $(x-1)^2 \ln x$ (b) $3(x-1)^4 + 5$
(c) $x(x-2)(x-1)^2$ (d) $(x+1)(x-3)(1-x)^3$
(e) $\sin[5(x-1)^4]$ (f) $\exp[8(x-1)^5]$
(g) $(1-x)\sin[(x^2-1)^3]$ (h) $\exp[-(\ln x)^3]$

2. Find all critical points of the function and classify them as local maxima, local minima, or horizontal inflection points.

(a) $\exp(-x^3)$, $-\infty < x < \infty$
(b) $1/(x^2 - 4x + 5)$, $-\infty < x < \infty$
(c) $\exp(12x - x^3)$, $-\infty < x < \infty$
(d) $-3(\ln x)^3$, $0 < x < \infty$
(e) $(\ln x)^4$, $0 < x < \infty$
(f) $\exp(-\sin x)$, $-1 < x < 5$
(g) $x^2 e^{-x}$, $-\infty < x < \infty$
(h) $4\sqrt{x} - x^2$, $0 < x < \infty$

3. Show that if $f(x)$ and $g(x)$ have local maxima at $x = X$, then the product $f(x)g(x)$ has a critical point there, which can be a local maximum, a local minimum, or a horizontal inflection point.

4. Show that the only critical point of $f(x,y) = x^3y - xy^3$ is (0,0) as claimed in Example 5.

5. Find, and classify, all local maxima, local minima, and saddles for the given function of x and y. Use Theorem 13.7.4 if possible.

(a) $2x^2 + xy + y^2 + 7y + 8$ (b) $x^2 + xy + y^2 + 6y$
(c) $x^2 + 4xy + y^2 + 6x + 8$ (d) $-6xy + x^5$
(e) $x(2x - y)$ (f) $\exp[-(x^2 + y^2 + 1)]$
(g) $\ln(x^2 + xy + y^2 + 4)$ (h) $\exp(x^2 - y^2)$
(i) $2xy - \sin xy$ (j) $x^3 - xy + x^5$
(k) $\ln(1 + x^2 + y^2)$ (l) $x^3 - xy + x^2y - 2$
(m) $1/(xy + y^2 + y + 1)$ (n) $2xy + x^8$

6. Same as Exercise 5.

(a) $f(x,y,z) = x^2 + 3xy + z^2$
(b) $f(x,y,z) = \exp(2x^2 + xz - 5z^2)$
(c) $f(x,y,z) = x^2 + y^2 + z^2 + xy + xz + yz$
(d) $f(x,y,z) = \ln(x^2 + y^2 + z^2 + 1)$
(e) $f(w,x,y,z) = \exp(w^2 - wx - yz + z^2)$
(f) $f(w,x,y,z) = 7 - 4(w^2 + x^2 + y^2 + z^2) + 4wx + 2wz$

7. Is it true that $f(x,y,z) = 1 + x^2 + 5y^2 + 2(z-1)^2 + 2\sin[x(z-1)]$ has a local minimum at (0,0,1)?

8. Let $f(x,y,z) = (1 + x^2 + y^2)^5 + \alpha \sin(xy) + 6z^2(1-z)$. For what values of α, if any, will f have a local minimum at (0,0,0)?

9. In Example 4 we considered the critical point at (0,0,0). Show that f has infinitely many other critical points as well. Find them and show that they comprise an infinite set of circles in 3-space. HINT: You will need to find nontrivial solutions of the equations

$$\begin{aligned} 2cx + y + z &= 0, \\ x + 2cy + z &= 0, \\ x + y + 2cz &= 0, \end{aligned}$$

where $c \equiv \cos(x^2 + y^2 + z^2)$.

10. For the case of a function of only two variables, $f(x,y)$, one can find the eigenvalues of the **A** matrix in (17) by the quadratic formula and examine their signs. Doing so, one can use Theorem 13.7.4 to obtain the following more specialized and frequently stated theorem, which we ask you to prove.

THEOREM 13.7.5 *Maximum, Minimum, Saddle*
Let

$$f_x(a,b) = 0 \quad \text{and} \quad f_y(a,b) = 0, \tag{10.1}$$

where (a,b) is an interior point in the domain of definition of f, and let f be C^2 in some neighborhood of (a,b). Define

$$D \equiv \begin{vmatrix} f_{xx}(a,b) & f_{xy}(a,b) \\ f_{xy}(a,b) & f_{yy}(a,b) \end{vmatrix}. \tag{10.2}$$

(i) If $f_{xx}(a,b) > 0$ and $D > 0$, then f has a local minimum at (a,b).
(ii) If $f_{xx}(a,b) < 0$ and $D > 0$, then f has a local maximum at (a,b).
(iii) If $D < 0$, then f has a saddle at (a,b).

---

11. (*More about Example 6*) (a) In Example 6 we state that it can be seen by inspection that the equilibrium point(s), if any, must lie within the unit square $0 < x < 1, 0 < y < 1$. Explain why that must be so.
(b) To understand the result obtained in Example 6, draw a

neat approximate sketch of the forces acting on a charge (of the same sign as Q, say) at the equilibrium point $x \approx 0.047, y \approx 0.513$, showing how the forces do indeed balance at that point. Also, explain, in physical terms, why the equilibrium is unstable.
(c) Repeat the calculation of Example 6, again using computer software, with the charge at the origin removed.
(d) Repeat the calculation of Example 6, again using computer software, with another charge Q added at the point $x = 1, y = 1$.
(e) Repeat the calculation of Example 6, again using computer software, but for a charge configuration supplied by your instructor.

12. (*Linear least-squares fit*) Suppose that the values $f(x_1) = f_1, f(x_2) = f_2, \ldots, f(x_N) = f_N$ are known; e.g., they may be experimental datum points. And suppose that we wish to find the best linear fit $f(x) \approx ax + b \equiv F(x)$ to these data, best in the sense that the total squared error E is minimized:

$$E(a,b) = \sum_{j=1}^{N} [F(x_j) - f_j]^2 = \text{minimum}. \tag{12.1}$$

It may be assumed that $x_1, \ldots, x_N$ are distinct points.
(a) Derive the results

$$a = \frac{N \sum x_j f_j - (\sum f_j)(\sum x_j)}{N \sum x_j^2 - (\sum x_j)^2}, \tag{12.2}$$

$$b = \frac{(\sum x_j^2)(\sum f_j) - (\sum x_j)(\sum x_j f_j)}{N \sum x_j^2 - (\sum x_j)^2}, \tag{12.3}$$

where $\sum$ denotes $\sum_{j=1}^{N}$.
(b) Verify that (12.2) and (12.3) give a local minimum of $E(a,b)$ and also that the denominators in (12.2) and (12.3) cannot vanish. HINT: The Schwarz inequality $|\mathbf{u} \cdot \mathbf{v}| \leq \|\mathbf{u}\| \, \|\mathbf{v}\|$, where $\mathbf{u} = (u_1, \ldots, u_N)$, $\mathbf{v} = (v_1, \ldots, v_N)$ should be useful if $\mathbf{u}$ and $\mathbf{v}$ are suitable chosen. Also, you may use Theorem 13.7.5 in Exercise 10 if you wish.
(c) Show that the local minimum of $E(a,b)$ is, in fact, an *absolute* minimum.

13. (a) Find the point on the line $x + 2y = 2$ that is closest to the origin, by the method of elimination and also by Lagrange's method. Prove that your result is indeed a minimum.
(b) Repeat part (a) for the curve $x = y^2 - 2$.

14. (*Optimum soup can*) (a) Suppose that we wish to design a cylindrical soup can, out of sheet metal having a cost of α cents per unit area, with a prescribed volume V, so as to minimize the material cost. Letting x be the radius and y be the height, we have the optimization problem

$$\text{Cost}: \quad f(x,y) = \alpha(2\pi x^2 + 2\pi xy) = \min \tag{14.1}$$

subject to the constraint

$$\text{Volume}: \quad \pi x^2 y = V. \tag{14.2}$$

Solve for x and y by the method of elimination and also by the Lagrange method, and show that $x = (V/2\pi)^{1/3}$, $y = 2(V/2\pi)^{1/3}$. Verify that this solution does give a minimum.
(b) Notice from the solution to part (a) that the resulting soup can has a height that is equal to its diameter, whereas actual soup cans are a bit taller than that. Perhaps the manufacturer would not only minimize the cost, for a given volume, but would also like a height-to-diameter ratio that makes the can *look* as large as possible. Suppose the proportions that make the can look as large as possible are defined by $y = kx$, where k is an empirically determined constant. Then the optimization problem (14.1) and (14.2) can be restated as

$$\begin{aligned} a[\alpha(2\pi x^2 + 2\pi xy)] + (1-a)(y - kx)^2 &= \min, \\ \pi x^2 y &= V, \end{aligned} \tag{14.3a,b}$$

where we prescribe a such that $0 \leq a \leq 1$. For example, if we care only about minimizing the cost we set $a = 1$, if we care only about making the volume look as large as possible we set $a = 0$, if we care equally about these considerations we set $a = 0.5$, and so on. [You need not solve (14.3).] Why was it important to square the $y - kx$ term in (14.3a)? Why did we *not* need to square the square-bracketed material cost term in (14.3a)?

15. (a) Show that the application of the Lagrange method to the problem

$$f = x_1^2 + x_2^2 = \text{extremum}, \tag{15.1}$$

$$g = a_{11}x_1^2 + a_{22}x_2^2 + 2a_{12}x_1x_2 = c \tag{15.2}$$

produces the eigenvalue problem

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{\lambda} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \tag{15.3}$$

NOTE: Observe the correspondence between a given eigenvalue problem (15.3) and the finding of maximum and mini-

mum distances from the origin of an x_1, x_2 plane to quadric curves $a_{11}x_1^2 + a_{22}x_2^2 + 2a_{12}x_1x_2 =$ constant. As discussed in Section 11.6, $a_{11}x_1^2 + a_{22}x_2^2 + 2a_{12}x_1x_2$ is the quadratic form associated with the **A** matrix in (15.3).
(b) Sketch a representative quadric curve (15.2) in an x_1, x_2 plane and indicate on that sketch the graphical significance of the eigenvectors of **A**. Explain.
(c) Find the points on the curve

$$x_1^2 + 2x_2^2 + x_1x_2 = 1$$

that are closest to the origin, and those that are farthest from the origin. Sketch the curve in the x_1, x_2 place, and show the eigenvectors of **A** and the points that you found, that are closest and farthest from the origin.
(d) Same as (c), for $x_1^2 + x_2^2 + x_1 = 1$.
(e) Same as (c), for $x_1^2 + x_2^2 - 4x_1 + 2x_2 = 2$.

16. Find the point (x, y), on the given line, that is closest to the point $(1, -2)$.

(a) $2x + y = 4$ (b) $x + y = -5$
(c) $x - 3y = 2$ (d) $x + 2y = 3$

17. Find the point (x, y, z), on the given plane, that is closest to the point $(2, 0, -1)$.

(a) $x + y + z = 4$ (b) $x - y + 2z = 4$
(c) $2x - y + z = 3$ (d) $2x + y + z = 0$

18. Find the point on the curve $y = x^{3/2}$ that is closest to the point(1,0).

19. Find the points at which $f = x + z$ is maximized and minimized, respectively, subject to the constraint $x^2 + y^2 + z^2 = 1$. Interpret the situation geometrically.

20. The constrained extremum problem

$$\begin{aligned} z &= \text{maximum}, \\ x + y + z &= 1, \\ \frac{z^2}{xy^3} &= 3 \end{aligned}$$

occurs in seeking to maximize the yield in an ammonia reactor as discussed in B. Noble, *Applications of Undergraduate Mathematics in Engineering* (New York: Macmillan, 1967), p. 75. Show that the desired maximum of z is $(17 + 4\sqrt{13})/9$.

21. (*More than one constraint*) (a) Extend the Lagrange method to cover cases that include more than one constraint. Specifically, explain how to solve the problem

$$\begin{aligned} f(x, y, z) &= \text{extremum}, \\ g_1(x, y, z) &= c_1, \\ g_2(x, y, z) &= c_2 \end{aligned}$$

by that method, and explain the logic behind each step. HINT: You should end up extremizing $f^* = f - \lambda_1 g_1 - \lambda_2 g_2$, where λ_1 and λ_2 are Lagrange multipliers.
(b) Apply that method to the problem

$$\begin{aligned} f &= x^2 + y^2 + z^2 = \text{extremum}, \\ g_1 &= x^2 + 4y^2 + 4z^2 = 4, \\ g_2 &= x + y + z = 0. \end{aligned}$$

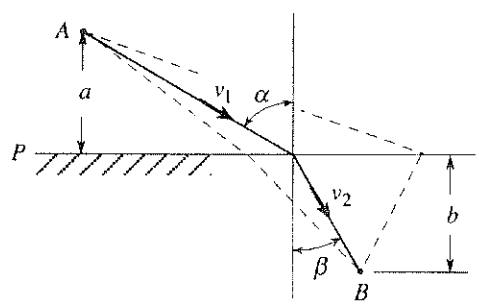
(c) Apply that method to the problem

$$\begin{aligned} f &= x + 2z = \text{extremum}, \\ g_1 &= x^2 + y^2 + z^2 = 1, \\ g_2 &= y - z = 0. \end{aligned}$$

(d) Apply that method to the problem

$$\begin{aligned} f &= x + y + z = \text{extremum}, \\ g_1 &= x^2 + y^2 + z^2 = 1, \\ g_2 &= x + y = 0. \end{aligned}$$

22. (*Fermat's principle*) (a) Let there be a light source at a given point A (shown in the first figure) in a medium in which the speed of light is v_1. Of the rays emitted from A, we are interested in the one that arrives at a given point B in a second medium in which the speed of light is v_2, the interface



between the two media being the plane P. Fermat's fundamental principle of optics* states that of all possible paths

*A lawyer by profession, *Pierre de Fermat* (1601–1665) was a mathematician only by avocation. Like most mathematicians of his time, Fermat studied problems of physics as well, and his **princi-**

(such as those shown as dashed lines), the one actually followed (shown as the solid line) is the one which is such that the travel time T from A to B is a minimum. Thus, show that

$$T(\alpha,\beta) = \frac{a}{v_1 \cos\alpha} + \frac{b}{v_2 \cos\beta} = \text{minimum}, \tag{22.1}$$

subject to the constraint that

$$g(\alpha,\beta) = a\tan\alpha + b\tan\beta = \text{constant}, \tag{22.2}$$

where α is the angle of incidence and β is the angle of refraction.

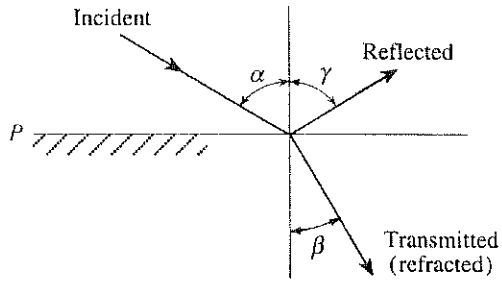
(b) Solving (22.1) and (22.2), derive the **law of refraction**

$$\frac{\sin\alpha}{\sin\beta} = \frac{v_1}{v_2}. \tag{22.3}$$

(c) Besides the incident and transmitted (refracted) rays, there will also be a reflected ray, as shown in the second figure. Using the same ideas as above, derive the **law of reflection**

$$\alpha = \gamma, \tag{22.4}$$

where γ is the angle of reflection.



NOTE: We have tacitly assumed that the paths within each medium are straight. The truth of this assumption can itself be deduced from Fermat's principle of least time.

13.8 Leibniz Rule

Recall, from the calculus, the **fundamental theorem of the integral calculus**: if f is a continuous function on the closed interval $[a, b]$, then

$$\frac{d}{dt}\int_a^t f(x)\,dx = f(t) \tag{1}$$

for $a \le t \le b$. We will see, in this section and in Chapter 16, that it is also useful to be able to differentiate integrals of the form

$$I(t) = \int_{a(t)}^{b(t)} f(x,t)\,dx, \tag{2}$$

which is more general than the integral $\int_a^t f(x)\,dx$ in (1) inasmuch as it allows for an arbitrary t dependence in both integration limits as well as in the integrand itself.

ple of least time was fundamental in the development of the science of optics. But Fermat is best remembered for his pioneering work in number theory. Of particular interest has been "Fermat's last theorem" (which, along with most of his other results, he wrote into the margin of his copy of the book *Arithmetica* by the great Alexandrian mathematician Diophantus), namely, that for $n > 2$ no integer solutions x, y, z exist for the equation $x^n + y^n = z^n$. He noted that he had "discovered a truly marvelous proof of this, which however the margin is not large enough to contain." Besides these marginal notes, and some results contained in letters to friends Fermat published but little. Thus it was, in 1665, that he perished.

To differentiate $I(t)$ it is convenient to put the sources of t dependence in a and b in evidence by expressing $I(t)$ as

$$I(t) = F(a(t), b(t), t). \tag{3}$$

That is, I is a function of the lower limit a (which, in turn, is a function of t), and of the upper limit b (which, in turn, is a function of t). The third argument in F indicates that even if a and b are fixed, I is still a function of t through the t dependence in the integrand $f(x, t)$. The form of (3) suggests chain differentiation, which step gives

$$I'(t) = \frac{\partial F}{\partial a} a'(t) + \frac{\partial F}{\partial b} b'(t) + \frac{\partial}{\partial t} \int_a^b f(x, t)\, dx, \tag{4}$$

where a, b are regarded as constants in the last term in (4), the $\partial/\partial t$ being with respect to the t dependence in the integrand. To evaluate $\partial F/\partial a$ and $\partial F/\partial b$ apply (1):

$$\frac{\partial F}{\partial a} = \frac{\partial}{\partial a} \int_a^b f(x, t)\, dx = -\frac{\partial}{\partial a} \int_b^a f(x, t)\, dx = -f(a, t) \tag{5}$$

and

$$\frac{\partial F}{\partial b} = \frac{\partial}{\partial b} \int_a^b f(x, t)\, dx = f(b, t). \tag{6}$$

Finally, it is reasonable to expect that

$$\frac{\partial}{\partial t} \int_a^b f(x, t)\, dx = \int_a^b \frac{\partial}{\partial t} f(x, t)\, dx, \tag{7}$$

that is, that we can interchange the order of the differentiation and the integration, in which case we have the result

$$\boxed{\begin{aligned} \frac{d}{dt} \int_{a(t)}^{b(t)} f(x, t)\, dx = \int_{a(t)}^{b(t)} \frac{\partial}{\partial t} f(x, t)\, dx \\ + b'(t) f(b(t), t) - a'(t) f(a(t), t), \end{aligned}} \tag{8}$$

which formula is known as the **Leibniz rule**.

Under what conditions is the Leibniz rule valid?* According to the fundamental theorem (1), (5) is valid if $f(b, t)$ is a continuous function of b, that is, if f is a continuous function of its first argument. For the chain differentiation (4) to be valid, we ask that $\partial f/\partial a$ and $\partial f/\partial b$ be continuous. Finally, it can be shown that (7) holds if f and $\partial f/\partial t$ are continuous functions of x and t. Putting these results together, we state the following theorem.†

*The issue is whether the limit interchange is possible, for the integrals are limits of Riemann sums and the derivatives are limits of difference quotients. Limit interchange came up in the closure part of Section 13.3, which we urge you to review. For the validity of the interchange in (7) (which is called *differentiation under the integral sign*) to be guaranteed, we need f to be sufficiently well behaved.

†While the discussion preceding the theorem outlines the main ideas, it is not a complete proof of the theorem. For a detailed proof, see Theorem 9-38 on p.220 of T. M. Apostol's *Mathematical Analysis* (Reading, MA: Addison-Wesley, 1957).

THEOREM 13.8.1 *Leibniz Rule*
Let $a(t)$ and $b(t)$ be differentiable for $t_1 \le t \le t_2$, and let A and B be finite constants such that $A \le a(t) \le B$ and $A \le b(t) \le B$ on $t_1 \le t \le t_2$. If $f(x,t)$ and $f_t(x,t)$ are continuous on the rectangle $A \le x \le B$, $t_1 \le t \le t_2$, then the Leibniz rule (8) holds for each t on $t_1 \le t \le t_2$.

EXAMPLE 1. Given

$$I(t) = \int_{-t}^{t^2} \cos(tx^2)\,dx, \tag{9}$$

evaluate $I'(t)$ by means of the Leibniz rule, if that rule applies. In this case $a(t) = -t$ and $b(t) = t^2$ are both differentiable on every interval $t_1 \le t \le t_2$, no matter how large, and $f(x,t) = \cos(tx^2)$ and $f_t(x,t) = -x^2 \sin(tx^2)$ are continuous on every rectangle in the x,t plane, so (8) applies and gives

$$\begin{aligned} I'(t) &= -\int_{-t}^{t^2} x^2 \sin(tx^2)\,dx + (2t)\cos[t(t^2)^2] - (-1)\cos[t(-t)^2] \\ &= -\int_{-t}^{t^2} x^2 \sin(tx^2)\,dx + 2t\cos(t^5) + \cos(t^3) \end{aligned} \tag{10}$$

at each point in the x,t plane. ∎

Thus far we have not explained the *purpose* of the Leibniz rule–how it is useful to us. The two examples to follow illustrate its use, and additional applications are included among the exercises.

EXAMPLE 2. To solve the forced harmonic oscillator initial-value problem

$$mx'' + kx = f(t); \qquad x(0) = 0,\ x'(0) = 0 \tag{11}$$

one can proceed by finding homogeneous and particular solutions by the methods explained in Chapter 3 (with the particular solution found by the method of variation of parameters) or by using the Laplace transform method explained in Chapter 5. Doing so, suppose we obtain the result

$$x(t) = \frac{1}{m\omega}\int_0^t \sin\omega(t-\tau)\,f(\tau)\,d\tau, \tag{12}$$

where $\omega = \sqrt{k/m}$ is the natural frequency of the oscillator. To check that result we can use the Leibniz rule, according to which

$$\begin{aligned} x'(t) &= \frac{1}{m\omega}\int_0^t \omega\cos\omega(t-\tau)\,f(\tau)\,d\tau + \frac{1}{m\omega}(1)(\sin 0)f(t) - 0 \\ &= \frac{1}{m}\int_0^t \cos\omega(t-\tau)\,f(\tau)\,d\tau, \end{aligned} \tag{13}$$

and

$$x''(t) = -\frac{\omega}{m}\int_0^t \sin\omega(t-\tau)\,f(\tau)\,d\tau + \frac{1}{m}(1)(\cos 0)f(t) - 0$$
$$= -\frac{\omega}{m}\int_0^t \sin\omega(t-\tau)\,f(\tau)\,d\tau + \frac{1}{m}f(t). \tag{14}$$

Putting (12) and (14) into (11) does give the identity $f(t) = f(t)$. Further, it follows from (12) and (13) that the initial conditions $x(0) = 0$ and $x'(0) = 0$ are safisfied as well. ∎

EXAMPLE 3. Sometimes we can use the Leibniz rule to evaluate integrals. For example, suppose we wish to evaluate

$$I = \int_0^\infty x^2 e^{-x^2}\,dx. \tag{15}$$

If we recall the known result*

$$\int_0^\infty e^{-x^2}\,dx = \frac{\sqrt{\pi}}{2}, \tag{16}$$

then we can evaluate I by considering instead the integral

$$J(a) = \int_0^\infty e^{-ax^2}\,dx, \tag{17}$$

for if we set $ax^2 = y^2$ (or $\sqrt{a}x = y$) in (16), then

$$J(a) = \int_0^\infty e^{-y^2}\,\frac{dy}{\sqrt{a}} = \frac{\sqrt{\pi}}{2}a^{-1/2}. \tag{18}$$

Differentiating (17) and the right-hand side of (18) with respect to a gives

$$\int_0^\infty -x^2 e^{-ax^2}\,dx = -\frac{1}{2}\frac{\sqrt{\pi}}{2}a^{-3/2} \tag{19}$$

and, finally, setting $a = 1$ in (19) gives the result

$$I = \int_0^\infty x^2 e^{-x^2}\,dx = \frac{\sqrt{\pi}}{4}. \tag{20}$$

COMMENT 1. Observe that there was no parameter in the original integral (15). To evaluate I we considered a slightly different integral J, with a suitably inserted parameter.

COMMENT 2. Actually, Theorem 13.8.1 does not suffice for the purposes of this example because of the infinite upper integration limit. That is, there does not exist the finite B that is called for in the theorem. This situation is more complicated because

$$\frac{d}{da}\int_0^\infty e^{-ax^2}\,dx = \frac{d}{da}\left(\lim_{C\to\infty}\int_0^C e^{-ax^2}\,dx\right). \tag{21}$$

*The result stated in (16) is well known and is the subject of Exercise 9 of Section 4.5.

Thus, besides the integral from 0 to C being the limit of a sequence of Riemann sums, and the derivative with respect to a being the limit of a difference quotient, we have the additional limit of C tending to ∞. The upshot is that if one or both limits in (8) are infinite ($+\infty$ or $-\infty$), then we need to impose conditions that are slightly more stringent than those in Theorem 13.8.1. Relevant theorems are to be found in books on advanced calculus, such as F. B. Hildebrand's *Advanced Calculus for Applications*, 2nd ed. (Englewood Cliffs, NJ: Prentice Hall, 1976), Sec. 7.9. We assert that (8) does hold in this example and in the exercises to follow. ∎

Closure. In summary, it is often useful to be able to differentiate a given integral with respect to a parameter without having to first carry out the integration. The formula for doing so is the Leibniz rule (8), and sufficient conditions for the validity of (8) are given in Theorem 13.8.1. Besides single integrals one is also interested in differentiating *multiple* integrals with respect to a parameter but, in the interest of brevity, we do not consider that case here.

EXERCISES 13.8

1. Apply the Leibniz rule:

(a) $\dfrac{d}{dt}\displaystyle\int_0^{t^2} \sin(tx^2)\,dx$

(b) $\dfrac{d}{dt}\displaystyle\int_3^{t} x^t \sin x\,dx$

(c) $\dfrac{d}{d\alpha}\displaystyle\int_\alpha^{2} e^{x^2}\,dx$

(d) $\dfrac{d}{d\alpha}\displaystyle\int_{-2\alpha^2}^{-\alpha} e^{\alpha x^3}\,dx$

(e) $\dfrac{d^2}{dx^2}\displaystyle\int_x^{2x} \ln(u^2+x^2)\,du$

(f) $\dfrac{d}{dy}\displaystyle\int_{y^2}^{1} x\,dx/(x^3+y^3)$

(g) $\dfrac{d^2}{da^2}\displaystyle\int_{5a}^{a^2} \cos(v^2+a^2)\,dv$

2. Derive the Taylor series of the given function $f(x)$ about $x=0$, up to and including terms of second order, using the Leibniz rule to obtain $f'(x)$ and $f''(x)$.

(a) $\int_0^x e^{-t^2}\,dt$ (b) $\int_{-x}^{\cos x} dt/(t^3+1)$

(c) $\int_0^{2\sin x} \ln(t^2+1)\,dt$ (d) $\int_0^{1+2x} e^{-xt^2}\,dt$

(e) $\int_x^{2x} \sin(xt^2)\,dt$ (f) $\int_{-x}^{3\sin x} \cos(xt^2)\,dt$

3. Show, by repeated differentiation of the formula

$$\int_0^\infty e^{-ax}\,dx = \frac{1}{a},$$

that $\int_0^\infty x^n e^{-x}\,dx = n!$ for $n = 0, 1, 2, \ldots$.

4. Evaluate $\int_0^\infty x^4 e^{-x^2}\,dx$, using any of the formulas found in Example 3.

5. (a) To evaluate

$$I = \int_0^\infty \frac{(\ln x)^2}{1+x^3}\,dx,$$

differentiate the formula

$$\int_0^\infty \frac{x^a}{1+x^3}\,dx = \frac{\pi}{3\sin\left(\frac{a+1}{3}\pi\right)} \tag{5.1}$$

a suitable number of times. Thus, show that $I = 10\pi^3/(81\sqrt{3})$. (b) Use (5.1) to evaluate the integral

$$I = \int_0^\infty \frac{x\ln x}{1+x^3}\,dx.$$

HINT: Recall that $d(x^a)/da = x^a \ln x$.

6. Show that

(a) $y(x) = \frac{1}{6}\int_a^x (x-t)^3 f(t)\,dt$ satisfies the initial-value problem $y''''(x) = f(x)$; $y(a) = y'(a) = y''(a) =$

$y'''(a) = 0$
(b) $y(x) = (1/x)\int_a^x (x-t)f(t)\,dt$ satisfies the initial-value problem $(xy)'' = f(x)$; $y(a) = y'(a) = 0$
(c) $y(x) = e^x + \int_0^x t^2 \cosh(x-t)\,dt$ satisfies the intial-value problem $y'' - y = 2x$; $y(0) = y'(0) = 1$

7. Evaluate $\int_0^1 x^{0.7}\ln x\,dx$ by differentiating the known formula

$$\int_0^1 x^a\,dx = \frac{1}{a+1}. \quad (a > -1)$$

8. Show that

$$\int_0^1 \frac{x^3-1}{\ln x}\,dx = \ln 4.$$

HINT: Considering

$$I(a) = \int_0^1 \frac{x^a - 1}{\ln x}\,dx, \quad (a \geq 0)$$

show that $I'(a) = 1/(a+1)$ with the "initial condition" $I(0) = 0$. Solve for $I(a)$ and set $a = 3$.

9. The conditions in Theorem 13.8.1 are sufficient, but not necessary. To illustrate this point, show that Leibniz differentiation of the integral

$$I(t) = \int_0^t \frac{e^{\sqrt{x}}}{\sqrt{x}}\,dx \qquad (t > 0)$$

does give the correct result even though the integrand $f(x,t) = e^{\sqrt{x}}/\sqrt{x}$ is not continuous at $x = 0$.

10. Consider a one-dimensional fluid flow, say flow in a channel of cross-sectional area A, with velocity $v(x,t)$ m/sec and mass density $\sigma(x,t)$ grams/m^3; x is measured positive downstream and t is the time. Consider a "control volume" $a(t) \leq x \leq b(t)$ that drifts with the fluid; i.e., $da/dt = v(a(t),t)$ and $db/dt = v(b(t),t)$. Then show that the principle of conservation of mass (i.e., that the amount of mass within the control volume remains constant) can be expressed as

$$\frac{d}{dt}\int_{a(t)}^{b(t)} \sigma(x,t)A\,dx = A\int_{a(t)}^{b(t)} \left[\frac{\partial\sigma}{\partial t} + \frac{\partial}{\partial x}(v\sigma)\right] dx = 0.$$

11. Show that

(a) $$u(x,t) = \int_0^1 f(\xi)\frac{e^{-(\xi-x)^2/(4\alpha^2 t)}}{2\alpha\sqrt{\pi t}}\,d\xi \tag{11.1}$$

$(-\infty < x < \infty, t > 0)$ satisfies the PDE (partial differential equation) $\alpha^2 u_{xx} = u_t$, known as the (one- dimensional) **diffusion equation**, where α^2 is a constant known as the diffusivity of the medium, and where f is a prescribed continuous function

(b) $$y(x,t) = \frac{F(x-ct) + F(x+ct)}{2} + \frac{1}{2c}\int_{x-ct}^{x+ct} G(\alpha)\,d\alpha \tag{11.2}$$

$(-\infty < x < \infty, t > 0)$ satisfies the PDE $c^2 y_{xx} = y_{tt}$, known as the (one-dimensional) **wave equation**, as well as the initial conditions $y(x,0) = F(x)$ and $y_t(x,0) = G(x)$, where $F(x)$ is differentiable and $G(x)$ is continuous

(c) $$u(x,y) = \frac{y}{\pi}\int_{-1}^{1} \frac{f(\xi)}{(\xi-x)^2 + y^2}\,d\xi \tag{11.3}$$

$(-\infty < x < \infty,\ y > 0)$ satisfies the PDE $u_{xx} + u_{yy} = 0$, known as the (two-dimensional) **Laplace equation**, where f is a prescribed continuous function.

12. (*Double and triple integrals*) The Leibniz rule (8) can be extended to multiple integrals. For our purposes it will suffice to limit that extension to the case of double and triple integrals over regions that are independent of the parameter. In the one-dimensional version (8), if $a(t)$ and $b(t)$ are constants then we have

$$\frac{d}{dt}\int_a^b f(x,t)\,dx = \int_a^b \frac{\partial}{\partial t} f(x,t)\,dx, \tag{12.1}$$

i.e., the derivative of the integral equals the integral of the derivative. In two and three dimensions the analogs of (12.1) are

$$\frac{d}{dt}\iint_{\mathcal{R}} f(x,y,t)\,dA = \iint_{\mathcal{R}} \frac{\partial}{\partial t} f(x,y,t)\,dA \tag{12.2}$$

and

$$\frac{d}{dt}\iiint_{\mathcal{R}} f(x,y,z,t)\,dV = \iiint_{\mathcal{R}} \frac{\partial}{\partial t} f(x,y,z,t)\,dV. \tag{12.3}$$

Sufficient conditions for the validity of (12.2) and (12.3) are that f and $\partial f/\partial t$ are continuous on the closed region $\mathcal{R}$ over the t interval of interest.

(a) Use (12.2) to show that

$$u(x,y) = \int_0^1\int_0^1 \ln\left[(x-\xi)^2 + (y-\eta)^2\right] f(\xi,\eta)\,d\xi\,d\eta \tag{12.4}$$

satisfies the PDE $u_{xx} + u_{yy} = 0$, known as the (two-dimensional) **Laplace equation**, at all points (x,y) outside

the square $0 \le x \le 1$, $0 \le y \le 1$, where f is any prescribed function that is continuous on that square.
(b) Use (12.3) to show that

$$u(x,y,z) = \int_0^1 \int_0^1 \int_0^1 \frac{f(\xi,\eta,\zeta)\,d\xi\,d\eta\,d\zeta}{\sqrt{(x-\xi)^2+(y-\eta)^2+(z-\zeta)^2}}$$

satisfies the (three-dimensional) **Laplace equation** $u_{xx} + u_{yy} + u_{zz} = 0$ at all points (x,y,z) outside the cube $0 \le x \le 1, 0 \le y \le 1, 0 \le z \le 1$, where f is any prescribed function that is continuous on that cube.

Chapter 13 Review

For the most part, Sections 13.1–13.5.1 are devoted to a review of the necessary preliminaries from the calculus, including point set theory, limits and continuity, composite functions and chain differentiation, Taylor's formula, and Taylor series.

In Section 13.5.2 we extend Taylor's formula and Taylor series to functions of more than one variable, such as $f(x,y)$, by parametrizing a line from the initial point (a,b) to the final point (x_0,y_0) by

$$x = a + (x_0 - a)t, \qquad y = b + (y_0 - b)t, \qquad (0 \le t \le 1)$$

because then $f(x,y) = f(x(t),y(t))$ reduces to a function $F(t)$ of the single variable t. The single-variable Taylor expansion of $F(t)$ about $t = 0$ leads us to the desired Taylor expansion of $f(x,y)$ about (a,b). An alternative and readily implemented approach is given in Exercise 11 of that section, namely, simply expanding in one variable at a time.

In Section 13.6 we consider relations such as $f(x,y) = 0$ on x and y, and the relations $f(x,y,u,v) = 0$ and $g(x,y,u,v) = 0$ on x, y, u, and v, and inquire as to whether such relations imply the existence of implicit functions $y(x)$, and $u(x,y)$ and $v(x,y)$, respectively. The chief result is the Implicit Function Theorem 13.6.2, which emphasizes the importance of the nonvanishing of the Jacobian determinant $\dfrac{\partial(f_1,\ldots,f_n)}{\partial(u_1,\ldots,u_n)}$ at the point in question, if the relations

$$\begin{aligned} f_1(x_1,\ldots,x_n,u_1,\ldots,u_n) &= 0, \\ &\vdots \\ f_n(x_1,\ldots,x_n,u_1,\ldots,u_n) &= 0 \end{aligned}$$

are to imply the existence of inverse functions $u_1(x_1,\ldots,x_n),\ldots,u_n(x_1,\ldots,x_n)$. We also find that the result

$$\frac{dy}{dx}\frac{dx}{dy} = 1,$$

for functions $y(x)$ and $x(y)$, generalizes to

$$\frac{\partial(x,y)}{\partial(u,v)}\frac{\partial(u,v)}{\partial(x,y)} = 1, \qquad \frac{\partial(x,y,z)}{\partial(u,v,w)}\frac{\partial(u,v,w)}{\partial(x,y,z)} = 1,$$

and so on, and that the chain rule

$$\frac{du}{dx}\frac{dx}{ds} = \frac{du}{ds}$$

for the function $u(x(s))$ generalizes to

$$\frac{\partial(u,v)}{\partial(x,y)}\frac{\partial(x,y)}{\partial(r,s)} = \frac{\partial(u,v)}{\partial(r,s)},$$
$$\frac{\partial(u,v,w)}{\partial(x,y,z)}\frac{\partial(x,y,z)}{\partial(r,s,t)} = \frac{\partial(u,v,w)}{\partial(r,s,t)},$$

and so on.

In Section 13.7 we review the theory of maxima and minima for functions of a single variable before turning to the unconstrained and constrained extremization of functions of more than one variable, considering only local extrema at interior points of the domain of the function being extremized. If $f(x_1,\ldots,x_n)$ is C^1 in some neighborhood of the point of extremum, $\mathbf{X}$, then the extremum condition $df = 0$ implies that we must have

$$\frac{\partial f}{\partial x_1} = \cdots = \frac{\partial f}{\partial x_n} = 0$$

at $\mathbf{X}$, which is called a critical point of f. To determine if f has a maximum, minimum, or saddle at $\mathbf{X}$, we use Taylor's formula about $\mathbf{X}$, through second-order terms because the first-order terms vanish by virtue of (7). The second-order terms constitute a quadratic form and everything hinges on the eigenvalues of the matrix $\mathbf{A} = \{f_{x_ix_j}(\mathbf{X})\}$ of that quadratic form: if all of $\mathbf{A}$'s eigenvalues are negative then f has a local maximum at $\mathbf{X}$, if all are positive then f has a local minimum there, and if they are mixed in sign then f has a saddle there.

For the extremization of $f(x_1,\ldots,x_n)$ in the presence of one or more constraints of the form $g(x_1,\ldots,x_n) =$ constant, the extremum condition $df = 0$ still holds, but no longer implies (7). We present two methods of solving such problems, elimination and the method of Lagrange multipliers.

Finally, in Section 13.8 we derive the Leibniz rule

$$\frac{d}{dt}\int_{a(t)}^{b(t)} f(x,t)\,dx = \int_{a(t)}^{b(t)} \frac{\partial}{\partial t}f(x,t)\,dx + b'(t)f(b(t),t) - a'(t)f(a(t),t)$$

for the differentiation of an integral that depends on a parameter t, which rule amounts to a generalization of the familiar fundamental theorem of the integral calculus,

$$\frac{d}{dt}\int_a^t f(x)\,dx = f(t).$$

Chapter 14

Vectors in 3-Space

14.1 Introduction

Vectors have already been studied in some detail in Chapter 9. However, except for Sections 9.1–9.3 that discussion focus mostly on n-tuple vectors in n-space. Here, however, we return to "arrow vectors" in physical 3-space and introduce a number of additional concepts, including the cross product of two vectors, combinations of dot and cross products, the differentiation of vector functions, and polar, cylindrical, and spherical coordinate systems.

Though each is of importance in its own right, Chapters 13–15 can be thought of as preparing the way for the especially important Chapter 16 on Scalar and Vector Field Theory.

14.2 Dot and Cross Product

It will be assumed here that you have studied Sections 9.1–9.3. By way of a brief review, recall that in Section 9.2 we introduced the notion of a vector in terms of a directed magnitude; we defined the magnitude or *norm* $\|\mathbf{u}\|$ of a vector $\mathbf{u}$, the *addition* of vectors $\mathbf{u}+\mathbf{v}$, the *scalar multiplication* $\alpha\mathbf{u}$, the *negative inverse* $-\mathbf{u}$, and the zero vector $\mathbf{0}$. And in Section 9.3 we defined the *dot product* $\mathbf{u}\cdot\mathbf{v}$ between any two vectors $\mathbf{u}$ and $\mathbf{v}$ as

$$\boxed{\mathbf{u}\cdot\mathbf{v} = \begin{cases} \|\mathbf{u}\|\,\|\mathbf{v}\|\cos\theta & \text{if } \mathbf{u},\ \mathbf{v} \neq \mathbf{0}, \\ 0 & \text{if } \mathbf{u} = \mathbf{0} \text{ or } \mathbf{v} = \mathbf{0}, \end{cases}} \tag{1}$$

where θ is the interior angle ($0 \le \theta \le \pi$) between $\mathbf{u}$ and $\mathbf{v}$ (Fig. 1). If $\mathbf{u} = \mathbf{v} \neq \mathbf{0}$, then $\theta = 0$, and (1) gives $\mathbf{u}\cdot\mathbf{u} = \|\mathbf{u}\|^2$, which result holds even if $\mathbf{u} = \mathbf{v} = \mathbf{0}$. Solving for $\|\mathbf{u}\|$, we obtain the relationship

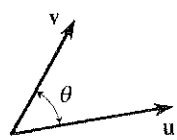


Figure 1. The angle θ in (1).

$$\|\mathbf{u}\| = \sqrt{\mathbf{u}\cdot\mathbf{u}}, \tag{2}$$

noted in Section 9.3, between the norm and the dot product. The dot product, defined by (1), admits the following properties:

$$\begin{aligned} \mathbf{u}\cdot\mathbf{v} &= \mathbf{v}\cdot\mathbf{u} && \text{(commutativity)} && (3a)\\ (\alpha\mathbf{u})\cdot\mathbf{v} &= \alpha(\mathbf{u}\cdot\mathbf{v}) && \text{(associtivity)} && (3b)\\ (\mathbf{u}+\mathbf{v})\cdot\mathbf{w} &= \mathbf{u}\cdot\mathbf{w}+\mathbf{v}\cdot\mathbf{w} && \text{(distributivity)} && (3c) \end{aligned}$$

for any vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$ and scalar α.* Of these properties, (3b) and (3c) are equivalent to the single statement that

$$(\alpha\mathbf{u}+\beta\mathbf{v})\cdot\mathbf{w} = \alpha(\mathbf{u}\cdot\mathbf{w})+\beta(\mathbf{v}\cdot\mathbf{w}) \qquad \text{(linearity)} \qquad (4)$$

for any vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$ and any scalars α, β, and the property (4) is known as *linearity*.

These properties permit us to manipulate dot products as if by ordinary scalar algebra. For example, just as

$$(2a+b)(4a-3b) = 8a^2 - 2ab - 3b^2,$$

where a, b are scalars, it is equally true that

$$\begin{aligned} (2\mathbf{u}+\mathbf{v})\cdot(4\mathbf{u}-3\mathbf{v}) &= 8\mathbf{u}\cdot\mathbf{u} - 2\mathbf{u}\cdot\mathbf{v} - 3\mathbf{v}\cdot\mathbf{v}\\ &= 8\,\|\mathbf{u}\|^2 - 2\mathbf{u}\cdot\mathbf{v} - 3\,\|\mathbf{v}\|^2 \end{aligned}$$

since

$$\begin{aligned} (2\mathbf{u}+\mathbf{v})\cdot(4\mathbf{u}-3\mathbf{v}) &= 2\mathbf{u}\cdot(4\mathbf{u}-3\mathbf{v}) + \mathbf{v}\cdot(4\mathbf{u}-3\mathbf{v}) && \text{by (4)}\\ &= 2(4\mathbf{u}-3\mathbf{v})\cdot\mathbf{u} + (4\mathbf{u}-3\mathbf{v})\cdot\mathbf{v} && \text{by (3a)}\\ &= 2(4\mathbf{u}\cdot\mathbf{u} - 3\mathbf{v}\cdot\mathbf{u} + 4\mathbf{u}\cdot\mathbf{v} - 3\mathbf{v}\cdot\mathbf{v} && \text{by (4)}\\ &= 8\mathbf{u}\cdot\mathbf{u} - 6\mathbf{u}\cdot\mathbf{v} + 4\mathbf{u}\cdot\mathbf{v} - 3\mathbf{v}\cdot\mathbf{v} && \text{by (3a)}\\ &= 8\,\|\mathbf{u}\|^2 - 2\mathbf{u}\cdot\mathbf{v} - 3\,\|\mathbf{v}\|^2 . \end{aligned}$$

Before leaving the dot product, observe from (1) that if $\mathbf{u}$ and $\mathbf{v}$ are nonzero, and $\theta = \pi/2$, then $\mathbf{u}$ and $\mathbf{v}$ are **perpendicular** to each other, and $\mathbf{u}\cdot\mathbf{v} = 0$. The converse, however, is not true for if $\mathbf{u}\cdot\mathbf{v} = 0$, then either $\mathbf{u}$ and $\mathbf{v}$ are perpendicular, *or* at least one of $\mathbf{u}$ and $\mathbf{v}$ is zero. Thus, we introduce a separate term, *orthogonal*: $\mathbf{u}$ and $\mathbf{v}$ are **orthogonal** if $\mathbf{u}\cdot\mathbf{v} = 0$. Perpendicularity implies orthogonality, but only if $\mathbf{u}$ and $\mathbf{v}$ are both nonzero does orthogonality imply perpendicularity.

Further, we say that a *set* of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is an **orthogonal set** if every vector in the set is orthogonal to every other vector in the set:

$$\mathbf{v}_i\cdot\mathbf{v}_j = 0 \quad \text{if } i \neq j. \qquad (5)$$

If, in addition, each $\mathbf{v}_j$ is scaled or normalized so as to have unit length, then

$$\mathbf{v}_i\cdot\mathbf{v}_j = \begin{cases} 1, & i = j\\ 0, & i \neq j \end{cases} \qquad (6)$$

*In Part III (Chapters 13–16), all scalars are *real* numbers.

since $\mathbf{v}_j \cdot \mathbf{v}_j = \|\mathbf{v}_j\|^2 = 1$, and we say that $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is an **orthonormal set**, namely, both orthogonal and normalized. Such sets are used extensively in subsequent sections.

Besides the dot product, we now introduce another kind of product between two vectors, the **cross product** of $\mathbf{u}$ and $\mathbf{v}$, denoted as $\mathbf{u} \times \mathbf{v}$, and defined as†

$$\boxed{\mathbf{u} \times \mathbf{v} = \begin{cases} \|\mathbf{u}\|\,\|\mathbf{v}\| \sin\theta\, \hat{\mathbf{e}} & \text{if } \mathbf{u}, \mathbf{v} \neq \mathbf{0},\ \theta \neq 0, \text{ and } \theta \neq \pi, \\ \mathbf{0} & \text{if } \mathbf{u} = \mathbf{0},\ \mathbf{v} = \mathbf{0},\ \theta = 0, \text{ or } \theta = \pi, \end{cases}} \tag{7}$$

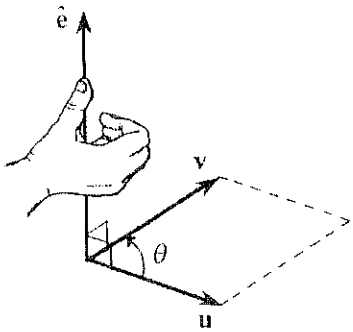


Figure 2. Cross product.

where θ is again the interior angle between $\mathbf{u}$ and $\mathbf{v}$, and $\hat{\mathbf{e}}$ is a unit vector normal to the plane of $\mathbf{u}$ and $\mathbf{v}$, directed such that the vectors $\mathbf{u}$, $\mathbf{v}$, $\hat{\mathbf{e}}$ form a *right-handed system*. That is, curling the other four fingers of our right hand from $\mathbf{u}$ (the first vector in $\mathbf{u} \times \mathbf{v}$) into $\mathbf{v}$ (the second vector in $\mathbf{u} \times \mathbf{v}$), our thumb points in the $\hat{\mathbf{e}}$ direction (Fig. 2).

Observe, first, that while the dot product $\mathbf{u} \cdot \mathbf{v}$ is a scalar, the cross product $\mathbf{u} \times \mathbf{v}$ is a vector. For that reason, $\mathbf{u} \cdot \mathbf{v}$ is also called the **scalar product** of $\mathbf{u}$ and $\mathbf{v}$, and $\mathbf{u} \times \mathbf{v}$ is also called the **vector product** of $\mathbf{u}$ and $\mathbf{v}$. (Some authors use the notation "$\mathbf{uv}$" in place of our $\mathbf{u} \cdot \mathbf{v}$, and "$\mathbf{u} \wedge \mathbf{v}$" in place of our $\mathbf{u} \times \mathbf{v}$.)

EXAMPLE 1. Let $\mathbf{u}$, $\mathbf{v}$ be as shown in Fig. 3: $\|\mathbf{u}\| = 5$, $\|\mathbf{v}\| = 4$, and $\theta = 60°$. Then $\mathbf{u} \times \mathbf{v} = (5)(4)(\sin 60°)\hat{\mathbf{e}} = (10\sqrt{3})\hat{\mathbf{e}}$ as sketched in the figure; $\mathbf{v} \times \mathbf{u}$ is of the same magnitude as $\mathbf{u} \times \mathbf{v}$ but points in the opposite direction. ∎

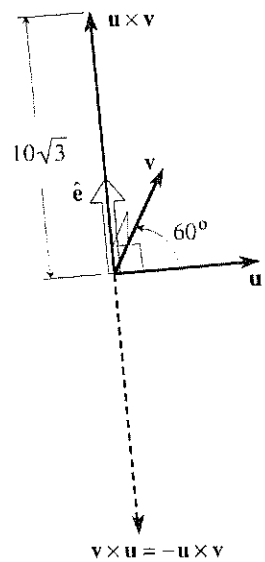


Figure 3. $\mathbf{u} \times \mathbf{v}$ and $\mathbf{v} \times \mathbf{u}$.

In contrast with the commutativity property of the dot product, namely $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$, we have $\mathbf{u} \times \mathbf{v} = -(\mathbf{v} \times \mathbf{u})$ for the cross product since

$$\mathbf{u} \times \mathbf{v} = \|\mathbf{u}\|\,\|\mathbf{v}\| \sin\theta\, \hat{\mathbf{e}}_1, \qquad \mathbf{v} \times \mathbf{u} = \|\mathbf{u}\|\,\|\mathbf{v}\| \sin\theta\, \hat{\mathbf{e}}_2,$$

where, according to the right-hand rule, $\hat{\mathbf{e}}_2 = -\hat{\mathbf{e}}_1$. The property $\mathbf{u} \times \mathbf{v} = -(\mathbf{v} \times \mathbf{u})$ is known as *anticommutativity* and is illustrated in Fig. 3. While commutativity fails to hold, associative and distributive properties do hold:

$$\mathbf{u} \times \mathbf{v} = -(\mathbf{v} \times \mathbf{u}) \qquad (\textit{anti}\text{commutativity}) \tag{8a}$$

$$(\alpha\mathbf{u}) \times \mathbf{v} = \alpha(\mathbf{u} \times \mathbf{v}) \qquad (\text{associativity}) \tag{8b}$$

$$(\mathbf{u} + \mathbf{v}) \times \mathbf{w} = (\mathbf{u} \times \mathbf{w}) + (\mathbf{v} \times \mathbf{w}) \qquad (\text{distributivity}) \tag{8c}$$

Of these properties, (8b) and (8c) are equivalent to the single statement

$$(\alpha\mathbf{u} + \beta\mathbf{v}) \times \mathbf{w} = \alpha(\mathbf{u} \times \mathbf{w}) + \beta(\mathbf{v} \times \mathbf{w}) \qquad (\text{linearity}) \tag{9}$$

† You may wonder why we need to spell out, in (7), that $\mathbf{u} \times \mathbf{v} = \mathbf{0}$ if $\theta = 0$ or $\theta = \pi$ since, after all, (7) *already* seems to say that $\mathbf{u} \times \mathbf{v} = \|\mathbf{u}\|\,\|\mathbf{v}\| \sin\theta\, \hat{\mathbf{e}} = 0\,\hat{\mathbf{e}} = \mathbf{0}$ (since $\sin 0 = 0$ and $\sin \pi = 0$). The catch is that $\hat{\mathbf{e}}$ is *undefined* if $\theta = 0$ or $\theta = \pi$ since then the $\mathbf{u}, \mathbf{v}$ vectors fail to define a unique plane.

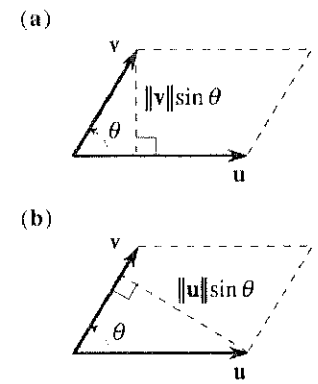


Figure 4. Area significance of $\|\mathbf{u} \times \mathbf{v}\|$.

for any vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ and scalars α, β, and the property (9) is known as *linearity*. Proof of (8b) and (8c) is left for the exercises.

Besides these properties there is a geometrical significance to be noted. Specifically, *the magnitude of* $\mathbf{u} \times \mathbf{v}$ *is equal to the area of the* $\mathbf{u}, \mathbf{v}$ *parallelogram* as shown in Fig. 4 (in which $\mathbf{u}$ and $\mathbf{v}$ lie in the plane of the paper). For the area (shaded) is the altitude $\|\mathbf{v}\|\sin\theta$ times the base $\|\mathbf{u}\|$ as labeled in Fig. 4a, or, equivalently, the altitude $\|\mathbf{u}\|\sin\theta$ times the base $\|\mathbf{v}\|$ as labeled in Fig. 4b. Let us state this result as an equation, for reference:

$$\boxed{\|\mathbf{u} \times \mathbf{v}\| = \text{area of } \mathbf{u}, \mathbf{v} \text{ parallelogram.}} \tag{10}$$

Recall from Example 1 in Section 9.3 that a simple physical example of the dot product involves the work W done when a body undergoes a linear displacement from an initial point A to a final point B under the action of a constant force $\mathbf{F}$. Specifically, if we denote the displacement vector as $\mathbf{AB}$, then $W = \mathbf{F} \cdot \mathbf{AB}$. Likewise there is, again from the subject of mechanics, a simple physical example of the cross product, namely, the moment of a force.

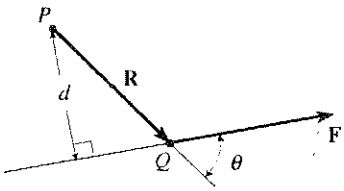


Figure 5. Moment of $\mathbf{F}$ about P.

EXAMPLE 2. *Moment of a Force.* In mechanics, one defines the *moment* M of a force $\mathbf{F}$, about a point P, as the product $\|\mathbf{F}\|\, d$, where d is the perpendicular distance from P to the line of action L of $\mathbf{F}$, as shown in Fig. 5. Besides the scalar quantity M, we can make the moment into a vector by using the right-hand rule. Specifically, one defines the vector moment $\mathbf{M}$ of $\mathbf{F}$, about P, as

$$\mathbf{M} \equiv \mathbf{R} \times \mathbf{F}, \tag{11}$$

where $\mathbf{R} = \mathbf{PQ}$ is a vector from P to the tail of $\mathbf{F}$ (Fig. 5) for then

$$\mathbf{M} = \|\mathbf{R}\|\,\|\mathbf{F}\|\sin\theta\,\hat{\mathbf{e}} = \|\mathbf{F}\|\,(\|\mathbf{R}\|\sin\theta)\,\hat{\mathbf{e}} = \|\mathbf{F}\|\,d\,\hat{\mathbf{e}} = M\hat{\mathbf{e}},$$

where $\hat{\mathbf{e}}$ is a unit vector perpendicular to the plane of $\mathbf{R}$ and $\mathbf{F}$ and in the direction dictated by the right-hand rule. For instance, if $\mathbf{R}$ and $\mathbf{F}$ in Fig. 5 are in the plane of the paper, then $\hat{\mathbf{e}}$ is directed perpendicular to the paper and toward the reader.

Notice that $\mathbf{M}$ is independent of the position of the point Q on the line of action L. For suppose that Q_1 and Q_2 are any two points on L (Fig. 6). Then

$$\begin{aligned} \mathbf{PQ}_1 \times \mathbf{F} - \mathbf{PQ}_2 \times \mathbf{F} &= (\mathbf{PQ}_1 - \mathbf{PQ}_2) \times \mathbf{F} && \text{by (8c),} \\ &= \mathbf{Q}_2\mathbf{Q}_1 \times \mathbf{F} && \text{by Fig. 6} \\ &= \mathbf{0} \end{aligned}$$

since the angle between $\mathbf{Q}_2\mathbf{Q}_1$ and $\mathbf{F}$ is either 0 or π. Thus, $\mathbf{PQ}_1 \times \mathbf{F} = \mathbf{PQ}_2 \times \mathbf{F}$, as claimed. ∎

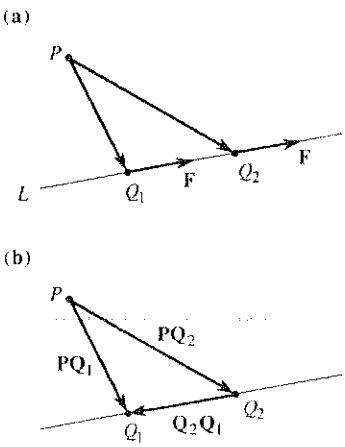


Figure 6. Location of Q.

Closure. Whereas a dot product can be defined for *any* vector space, the cross product is specific to 3-space. In this section we recalled the dot product (1), from Section 9.3, and introduce the cross product (7). Both operations are linear, but

whereas the dot product is commutative ($\mathbf{u}\cdot\mathbf{v} = \mathbf{v}\cdot\mathbf{u}$), the cross product is anti-commutative ($\mathbf{u}\times\mathbf{v} = -\mathbf{v}\times\mathbf{u}$).

EXERCISES 14.2

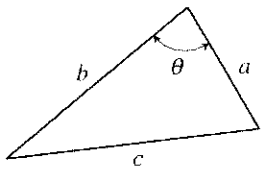
1. If $\mathbf{u}\times\mathbf{v} = 2\mathbf{u}$, what can one conclude about $\mathbf{u}$? About $\mathbf{v}$? Explain.

2. If $\mathbf{u}\times\mathbf{v} = \mathbf{0}$ and $\mathbf{u}\cdot\mathbf{v} = 0$, must $\mathbf{u} = \mathbf{0}$ and/or $\mathbf{v} = \mathbf{0}$? Explain.

3. Use the dot product to derive the *law of cosines*,

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

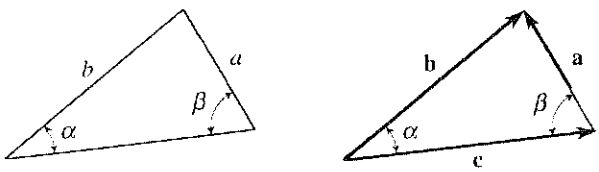
where c, a, b, θ are shown in the figure. HINT: Regard the sides of the triangle as vectors, oriented such that $\mathbf{c} = \mathbf{a} - \mathbf{b}$, and note that $\mathbf{c}\cdot\mathbf{c} = (\mathbf{a}-\mathbf{b})\cdot(\mathbf{a}-\mathbf{b})$.



4. Use the cross product to derive the *law of sines*,

$$\frac{a}{\sin\alpha} = \frac{b}{\sin\beta},$$

where a, b, α, β are as shown in part (a) of the accompanying figure. HINT: Regard the sides of the triangle as vectors, as in part (b), and cross a suitable vector into both sides of the equation $\mathbf{b} - \mathbf{c} = \mathbf{a}$.



(a) (b)

5. Derive the **Lagrange identity**

$$\|\mathbf{u}\times\mathbf{v}\|^2 = \|\mathbf{u}\|^2\,\|\mathbf{v}\|^2 - (\mathbf{u}\cdot\mathbf{v})^2. \tag{5.1}$$

6. Why did we bother to include the parentheses in the right-hand side of (8c) but not in the analogous equation (3c)?

7. If $\mathbf{u}\cdot\mathbf{v} = 0$ and $\mathbf{u}\times\mathbf{v} \equiv \mathbf{w}$ is $\{\mathbf{u}, \mathbf{v}, \mathbf{w}\}$ necessarily an orthogonal set? Explain.

8. Is $\mathbf{u}\times(\mathbf{v}\times\mathbf{w})$ necessarily equal to $(\mathbf{u}\times\mathbf{v})\times\mathbf{w}$? Prove or disprove.

9. Prove the associative law (8b). [Proof of the distributive law (8c) is left for the exercises in Section 14.3.]

10. (a) Prove that

$$(\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \cdots + \alpha_n\mathbf{u}_n)\cdot\mathbf{v} = \alpha_1(\mathbf{u}_1\cdot\mathbf{v}) + \alpha_2(\mathbf{u}_2\cdot\mathbf{v}) + \cdots + \alpha_n(\mathbf{u}_n\cdot\mathbf{v}),$$

namely, that (4) holds even if $n > 2$.
(b) Prove that

$$(\alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \cdots + \alpha_n\mathbf{u}_n)\times\mathbf{v} = \alpha_1(\mathbf{u}_1\times\mathbf{v}) + \alpha_2(\mathbf{u}_2\times\mathbf{v}) + \cdots + \alpha_n(\mathbf{u}_n\times\mathbf{v}),$$

namely, that (9) holds even if $n > 2$.

11. (*Linear dependence*) Prove that vectors $\mathbf{u}, \mathbf{v}$ are linearly dependent if and only if $\mathbf{u}\times\mathbf{v} = \mathbf{0}$.

12. Prove that if $\mathbf{u}$ and $\mathbf{v}$ are nonzero vectors, and $\mathbf{u}\times\mathbf{v} = \mathbf{0}$, then $\mathbf{u} = \alpha\mathbf{v}$ for some nonzero constant α.

14.3 Cartesian Coordinates

In Sections 9.2 and 14.2 we introduce vectors in 3-space and define the norm of a vector, the multiplication of a vector by a scalar, addition and subtraction of vec-

tors, and the dot and cross products of two vectors. These definitions are *intrinsic* inasmuch as they contain no reference whatsoever to any coordinate system.

In applications, however, our computations are often carried out more conveniently if we introduce a reference coordinate system and a corresponding set of base vectors. Since the operations listed above are defined intrinsically, it follows that the answers obtained will be the same whether we choose one coordinate system and set of base vectors or another. Thus, the choice is a matter of preference and convenience.

Although there are many coordinate systems from which to choose, we limit our attention here to those that we regard as the most important: *Cartesian*, *cylindrical*, and *spherical*. In the present section we consider the Cartesian case.

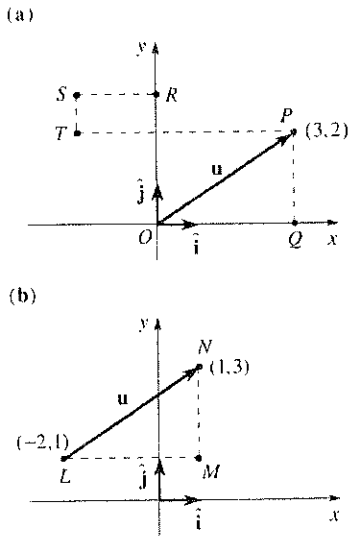


Figure 1. Cartesian x, y system.

Beginning with 2-space, for simplicity, we adopt the rectangular Cartesian x, y coordinate system shown in Fig. 1, together with the reference vectors $\hat{\mathbf{i}}, \hat{\mathbf{j}}$, where $\hat{\mathbf{i}}$ is a unit vector in the positive x direction, and $\hat{\mathbf{j}}$ is a unit vector in the positive y direction. We have placed $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ with their tails at the origin, but one can place them wherever one wishes, as long as their magnitudes and directions are preserved. Observe that $\left\{\hat{\mathbf{i}}, \hat{\mathbf{j}}\right\}$ constitutes an orthonormal set since

$$\hat{\mathbf{i}} \cdot \hat{\mathbf{i}} = \hat{\mathbf{j}} \cdot \hat{\mathbf{j}} = (1)(1)\cos 0^\circ = 1 \qquad \text{and} \qquad \hat{\mathbf{i}} \cdot \hat{\mathbf{j}} = (1)(1)\cos 90^\circ = 0. \tag{1}$$

Now, consider the vector $\mathbf{u}$ in Fig. 1a; whether $\mathbf{u}$ is being used as a *position vector* to locate the point P, namely the point $(3, 2)$ in the x, y plane or to designate a certain force vector * or a velocity vector or whatever, is not important here. We can express $\mathbf{u}$ as a linear combination of $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ as follows:

$$\mathbf{u} = \mathbf{OQ} + \mathbf{QP} = 3\hat{\mathbf{i}} + 2\hat{\mathbf{j}}.$$

The end result, $\mathbf{u} = 3\hat{\mathbf{i}} + 2\hat{\mathbf{j}}$, is independent of the fact that the $\mathbf{u}$ vector happens to spring from the origin; if we move its tail to $(-2, 1)$, say, as in Fig. 1b, then

$$\mathbf{u} = \mathbf{LM} + \mathbf{MN} = 3\hat{\mathbf{i}} + 2\hat{\mathbf{j}},$$

which result is the same as above.

More generally, let $\mathbf{u}$ be a vector from the origin to *any* point (a, b) [or, equivalently, from (x_0, y_0) to $(x_0 + a, y_0 + b)$]. By the same reasoning as above, we can express $\mathbf{u}$ in terms of $\hat{\mathbf{i}}, \hat{\mathbf{j}}$ as

$$\mathbf{u} = a\hat{\mathbf{i}} + b\hat{\mathbf{j}}. \tag{2}$$

Is (2) unique? For example, breaking OP (in Fig. 1a) into OQP gave $\mathbf{u} = 3\hat{\mathbf{i}} + 2\hat{\mathbf{j}}$; but if we had broken it into $ORSTP$, instead, might we have obtained a different result, such as $\mathbf{u} = 4\hat{\mathbf{i}} + 2\hat{\mathbf{j}}$? No. To prove that the expression (2) of a given vector $\mathbf{u}$, as a linear combination of $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$, is indeed unique, suppose that $\mathbf{u}$ also admits an expansion

$$\mathbf{u} = a'\hat{\mathbf{i}} + b'\hat{\mathbf{j}}. \tag{3}$$

*For example, if we choose a scale of 1 newton per unit length, then $\mathbf{u}$, in Fig. 1a, would represent a force of $\sqrt{13}$ newtons, in the direction OP.

Subtracting equals from equals, namely, (3) from (2) yields

$$(a - a')\hat{\mathbf{i}} + (b - b')\hat{\mathbf{j}} = \mathbf{0}. \tag{4}$$

Dotting both sides of (4) with $\hat{\mathbf{i}}$, then with $\hat{\mathbf{j}}$, and using (1), we find that $a - a' = 0$ and $b - b' = 0$ so that $a = a'$ and $b = b'$. Consequently, (3) is necessarily identical to (2) so (2) is unique, as claimed.

Similarly for 3-space: Suppose that we adopt a Cartesian x, y, z coordinate system, with reference vectors $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$ (Fig. 2), where

$$\hat{\mathbf{i}} \cdot \hat{\mathbf{i}} = \hat{\mathbf{j}} \cdot \hat{\mathbf{j}} = \hat{\mathbf{k}} \cdot \hat{\mathbf{k}} = 1, \tag{5a}$$

$$\hat{\mathbf{i}} \cdot \hat{\mathbf{j}} = \hat{\mathbf{i}} \cdot \hat{\mathbf{k}} = \hat{\mathbf{j}} \cdot \hat{\mathbf{k}} = 0. \tag{5b}$$

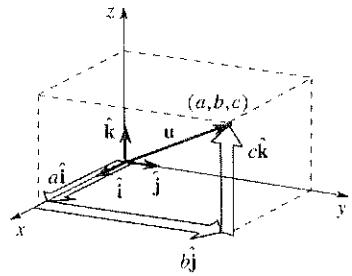


Figure 2. Cartesian coordinates and $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$.

It follows from (5) that $\left\{\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}\right\}$ is an orthonormal set. Furthermore,

$$\hat{\mathbf{i}} \times \hat{\mathbf{j}} = \left\|\hat{\mathbf{i}}\right\| \left\|\hat{\mathbf{j}}\right\| \sin 90^\circ \, \hat{\mathbf{k}} = \hat{\mathbf{k}}. \tag{6}$$

Working out all of the various cross products in this manner, we find that

$$\hat{\mathbf{i}} \times \hat{\mathbf{i}} = \hat{\mathbf{j}} \times \hat{\mathbf{j}} = \hat{\mathbf{k}} \times \hat{\mathbf{k}} = \mathbf{0}, \tag{7a}$$

$$\hat{\mathbf{i}} \times \hat{\mathbf{j}} = \hat{\mathbf{k}}, \qquad \hat{\mathbf{j}} \times \hat{\mathbf{k}} = \hat{\mathbf{i}}, \qquad \hat{\mathbf{k}} \times \hat{\mathbf{i}} = \hat{\mathbf{j}}, \tag{7b}$$

$$\hat{\mathbf{j}} \times \hat{\mathbf{i}} = -\hat{\mathbf{k}}, \qquad \hat{\mathbf{k}} \times \hat{\mathbf{j}} = -\hat{\mathbf{i}}, \qquad \hat{\mathbf{i}} \times \hat{\mathbf{k}} = -\hat{\mathbf{j}}. \tag{7c}$$

Figure 3. Mnemonic device for (7b) and (7c).

Equations (7b) and (7c) can be remembered easily if one keeps in mind the mnemonic device shown in Fig. 3. Note carefully, however, that the signs in (7b), (7c), and Fig. 3, are correct only if our x, y, z coordinate system is *right-handed*. If we were to choose a *left-handed* system, the signs in (7b), (7c), and Fig. 3 would be reversed; for example, we would have $\hat{\mathbf{i}} \times \hat{\mathbf{j}} = -\hat{\mathbf{k}}$, and $\hat{\mathbf{j}} \times \hat{\mathbf{k}} = -\hat{\mathbf{i}}$. *As long as we are consistent* it does not matter whether we choose a right-handed system or a left-handed one. But to minimize confusion, we shall *always* adopt right-handed systems, as is standard mathematical practice.

Analogous to the expression (2) for 2-space, any given vector $\mathbf{u}$ in 3-space, say from $(0, 0, 0)$ to (a, b, c) [or, equivalently, from (x_0, y_0, z_0) to $(x_0 + a, y_0 + b, z_0 + c)$], can be expressed as a unique linear combination of $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$, namely,

$$\mathbf{u} = a\hat{\mathbf{i}} + b\hat{\mathbf{j}} + c\hat{\mathbf{k}} \tag{8}$$

as illustrated in Fig. 2. Here a, b, c are called the x, y, z **components** of $\mathbf{u}$, respectively. Instead of the letters a, b, c, it is more common to use the letters u_1, u_2, u_3 for the components of $\mathbf{u}$ so that

$$\mathbf{u} = u_1\hat{\mathbf{i}} + u_2\hat{\mathbf{j}} + u_3\hat{\mathbf{k}}. \tag{9}$$

Since each vector in 3-space can be expressed, or *expanded*, as a unique linear combination of $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$, we say that the set $\left\{\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}\right\}$ is a **basis** for 3-space, an

orthonormal basis, in fact. (Bases and expansions are covered in more detail in Section 9.9, but that section is not a prerequisite for the present one.)

The form (9) is especially convenient for calculations. For example, if

$$\mathbf{u} = u_1\hat{\mathbf{i}} + u_2\hat{\mathbf{j}} + u_3\hat{\mathbf{k}} \qquad \text{and} \qquad \mathbf{v} = v_1\hat{\mathbf{i}} + v_2\hat{\mathbf{j}} + v_3\hat{\mathbf{k}}, \tag{10}$$

then

$$\begin{aligned}
\mathbf{u} + \mathbf{v} &= (u_1 + v_1)\hat{\mathbf{i}} + (u_2 + v_2)\hat{\mathbf{j}} + (u_3 + v_3)\hat{\mathbf{k}}, && \text{(11a)}\\
\mathbf{u} - \mathbf{v} &= (u_1 - v_1)\hat{\mathbf{i}} + (u_2 - v_2)\hat{\mathbf{j}} + (u_3 - v_3)\hat{\mathbf{k}}, && \text{(11b)}\\
\alpha\mathbf{u} &= (\alpha u_1)\hat{\mathbf{i}} + (\alpha u_2)\hat{\mathbf{j}} + (\alpha u_3)\hat{\mathbf{k}}, && \text{(11c)}\\
\mathbf{u}\cdot\mathbf{v} &= (u_1\hat{\mathbf{i}} + u_2\hat{\mathbf{j}} + u_3\hat{\mathbf{k}})\cdot(v_1\hat{\mathbf{i}} + v_2\hat{\mathbf{j}} + v_3\hat{\mathbf{k}})\\
&= u_1v_1\hat{\mathbf{i}}\cdot\hat{\mathbf{i}} + u_1v_2\hat{\mathbf{i}}\cdot\hat{\mathbf{j}} + \cdots + u_3v_2\hat{\mathbf{k}}\cdot\hat{\mathbf{j}} + u_3v_3\hat{\mathbf{k}}\cdot\hat{\mathbf{k}}\\
&= u_1v_1 + u_2v_2 + u_3v_3, && \text{(11d)}\\
\mathbf{u}\times\mathbf{v} &= (u_1\hat{\mathbf{i}} + u_2\hat{\mathbf{j}} + u_3\hat{\mathbf{k}})\times(v_1\hat{\mathbf{i}} + v_2\hat{\mathbf{j}} + v_3\hat{\mathbf{k}})\\
&= u_1v_1\hat{\mathbf{i}}\times\hat{\mathbf{i}} + u_1v_2\hat{\mathbf{i}}\times\hat{\mathbf{j}} + \cdots + u_3v_2\hat{\mathbf{k}}\times\hat{\mathbf{j}} + u_3v_3\hat{\mathbf{k}}\times\hat{\mathbf{k}}\\
&= (u_2v_3 - u_3v_2)\hat{\mathbf{i}} - (u_1v_3 - u_3v_1)\hat{\mathbf{j}} + (u_1v_2 - u_2v_1)\hat{\mathbf{k}}. && \text{(11e)}
\end{aligned}$$

The right-hand side of (11e) happens to be expressible as a third-order determinant so that

$$\boxed{\mathbf{u}\times\mathbf{v} = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix},} \tag{12}$$

this form being easier to remember than (11e).

EXAMPLE 1. Given

$$\mathbf{u} = \hat{\mathbf{i}} - 2\hat{\mathbf{k}}, \qquad \mathbf{v} = 3\hat{\mathbf{i}} + \hat{\mathbf{j}} + \hat{\mathbf{k}}, \qquad \mathbf{w} = \hat{\mathbf{j}} - \hat{\mathbf{k}},$$

compute $(\mathbf{u} + 2\mathbf{v}) \times \mathbf{w}$ and $\mathbf{u}\cdot\mathbf{w}$.

$$\begin{aligned}
(\mathbf{u} + 2\mathbf{v}) \times \mathbf{w} &= [(1+6)\hat{\mathbf{i}} + (0+2)\hat{\mathbf{j}} + (-2+2)\hat{\mathbf{k}}] \times (\hat{\mathbf{j}} - \hat{\mathbf{k}})\\
&= (7\hat{\mathbf{i}} + 2\hat{\mathbf{j}}) \times (\hat{\mathbf{j}} - \hat{\mathbf{k}}) = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ 7 & 2 & 0 \\ 0 & 1 & -1 \end{vmatrix} = -2\hat{\mathbf{i}} + 7\hat{\mathbf{j}} + 7\hat{\mathbf{k}}
\end{aligned}$$

and $\mathbf{u}\cdot\mathbf{w} = (1)(0) + (0)(1) + (-2)(-1) = 2$. ∎

Closure. In this section we introduce the (always right-handed) Cartesian coordinate system x, y, z, and its corresponding orthonormal base vectors $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$. Non-Cartesian systems are the subject of Section 14.6.

EXERCISES 14.3

1. Let $\mathbf{u} = 2\hat{\mathbf{i}} - \hat{\mathbf{j}} - 3\hat{\mathbf{k}}$, $\mathbf{v} = \hat{\mathbf{i}} + \hat{\mathbf{j}} - \hat{\mathbf{k}}$, $\mathbf{w} = 3\hat{\mathbf{i}} + 2\hat{\mathbf{k}}$, $\mathbf{x} = 8\hat{\mathbf{i}} - \hat{\mathbf{j}} - 11\hat{\mathbf{k}}$.

(a) Compute $2\mathbf{u} - \mathbf{v}$, $\mathbf{u}\cdot\mathbf{v}$, $\mathbf{v}\cdot\mathbf{u}$, $\mathbf{u}\times\mathbf{v}$, $\mathbf{v}\times\mathbf{u}$ and $\|\mathbf{u}\times\mathbf{v}\|$.
(b) Compute $(\mathbf{u}+3\mathbf{w})\times\mathbf{v}$, $\|(\mathbf{u}+3\mathbf{w})\times\mathbf{v}\|$, $\mathbf{u}\cdot\mathbf{w}$ and $|\mathbf{u}\cdot\mathbf{w}|$.
(c) Compute $(\mathbf{u}\cdot 2\mathbf{w})\mathbf{v}$, $\mathbf{x}\times(-2\mathbf{u}) + 3\mathbf{v}$, $(\mathbf{u}+\mathbf{w})\times\mathbf{v}$ and $\|\mathbf{x}\|$.
(d) Compute $\mathbf{u} + 2\mathbf{v} - \mathbf{w} + 3\mathbf{x}$, $(2\mathbf{u}+\mathbf{w})\cdot(\mathbf{v}-\mathbf{x})$ and $\mathbf{w}\times(\mathbf{v}-\mathbf{x})$.
(e) Compute $(3\mathbf{v}+\mathbf{w}-\mathbf{x})\cdot\mathbf{v}$, $(\mathbf{w}-\mathbf{x})\times(2\mathbf{v})$ and $\|\mathbf{u}\times(\mathbf{v}\times\mathbf{w})\|$.
(f) With $\mathbf{u}, \mathbf{v}, \mathbf{x}$ tail to tail, show that the three vectors lie in a plane. Do this two ways: first, by taking suitable cross and dot products and, second, by showing that one of the vectors can be expressed as a linear combination of the others.
(g) With $\mathbf{u}, \mathbf{v}, \mathbf{w}$ tail to tail, show — by taking suitable cross and dot products — that the three vectors do *not* lie in the same plane.
(h) Find a vector perpendicular to $\mathbf{u}$ and $\mathbf{v}$. Verify your answer.
(i) Find a vector perpendicular to $\mathbf{u}$ and $\mathbf{w}$. Verify your answer.
(j) Find a vector perpendicular to $\mathbf{v}$ and $\mathbf{w}$. Verify your answer.
(k) With $\mathbf{u}, \mathbf{v}, \mathbf{w}$ tail to tail, find the area of the triangle formed by their heads, and find a vector perpendicular to that triangle..
(l) Repeat part (k) for $\mathbf{u}, \mathbf{v}, \mathbf{x}$.
(m) Repeat part (k) for $\mathbf{u}, \mathbf{w}, \mathbf{x}$.
(n) With $\mathbf{u}, \mathbf{v}, \mathbf{w}, \mathbf{x}$ tail to tail, show whether or not their heads lie in a common plane.

2. Consider three forces: $\mathbf{F}_1 = 2\hat{\mathbf{i}}+\hat{\mathbf{k}}$ with its tail at $(2,1,-1)$, $\mathbf{F}_2 = \hat{\mathbf{i}} - \hat{\mathbf{j}} + \hat{\mathbf{k}}$ with its tail at $(0,0,0)$, and $\mathbf{F}_3 = \hat{\mathbf{i}} + 4\hat{\mathbf{j}}$ with its tail at $(1,1,1)$.

(a) Find the total moment $\mathbf{M}$ about $(0,0,0)$.
(b) Find the total moment $\mathbf{M}$ about $(1,1,0)$.
(c) Find the total moment $\mathbf{M}$ about $(8,6,-10)$.
(d) Find the total moment $\mathbf{M}$ about $(3,2,1)$.
(e) Determine, if possible, an additional force $\mathbf{F}_4$ with its tail at $(2,-1,-1)$, such that the total moment $\mathbf{M}$ about $(1,1,1)$ is zero. If it is *not* possible, explain that circumstance in geometrical terms.
(f) Same as part (e) with $(2,-1,-1)$ changed to $(3,6,-1)$.
(g) Same as part (e) with $(2,-1,-1)$ changed to $(-1,-4,3)$.

3. Use a suitable cross product to determine whether or not the following points lie on a straight line.

(a) $(2,3,1)$, $(1,-1,4)$, $(2,2,1)$
(b) $(1,3,0)$, $(2,-1,1)$, $(3,-5,2)$
(c) $(1,0,2)$, $(-4,3,0)$, $(-9,6,-2)$
(d) $(1,0,2)$, $(-4,3,0)$, $(12,-6,6)$
(e) $(1,0,2)$, $(-4,3,0)$, $(-3,3,2)$
(f) $(1,0,1)$, $(2,-1,4)$, $(5,7,-3)$

4. Use a suitable property of the cross product to find the area of the triangles with the following vertices.

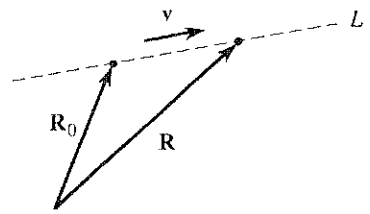
(a) $(1,4,3)$, $(2,0,-1)$, $(0,0,5)$
(b) $(2,-2,1)$, $(4,0,3)$, $(2,3,5)$
(c) $(0,0,2)$, $(0,4,0)$, $(3,1,1)$
(d) $(8,7,4)$, $(2,-5,0)$, $(1,6,8)$
(e) $(1,1,1)$, $(1,1,2)$, $(3,2,1)$
(f) $(5,4,1)$, $(3,2,3)$, $(-4,0,6)$
(g) $(2,-1,1)$, $(0,2,3)$, $(9,6,1)$
(h) $(14,1,-6)$, $(2,2,1)$, $(0,4,9)$

5. (a) Determine a unit normal vector to the plane $x+2y-z=5$ by finding three distinct (noncollinear) points A, B, C in the plane, crossing $\mathbf{AB}$ with $\mathbf{AC}$, and normalizing.
(b) Repeat part (a) for the plane $2x+y-z=0$.
(c) Repeat part (a) for the plane $x-3y+4z=-2$.
(d) Repeat part (a) for the plane $x+3y+z=4$.
(e) Repeat part (a) for the plane $ax+by+cz=d$.

6. (*Equation of a line*) (a) Let $\mathbf{R}_0$ be a given point in 3-space (i.e., the point at the head of the position vector $\mathbf{R}_0$ issuing from the origin). And let $\mathbf{v}$ be a given (nonzero) vector (see the figure). Show that the locus of points $\mathbf{R}$ which constitute a straight line L through $\mathbf{R}_0$, parallel to $\mathbf{v}$, is determined by either of the equations

$$(\mathbf{R}-\mathbf{R}_0)\times\mathbf{v} = \mathbf{0}, \qquad \text{or} \tag{6.1}$$

$$\mathbf{R} = \mathbf{R}_0 + \mathbf{v}t \qquad (-\infty < t < \infty). \tag{6.2}$$



(b) Using (6.2), show that the straight line through $(2,5,-1)$, parallel to the vector $4\hat{\mathbf{i}} - \hat{\mathbf{k}}$, may be defined by the parametric equations

$$x = 2+4t, \qquad y = 5, \qquad z = -1-t$$

for $-\infty < t < \infty$.
(c) Determine parametric equations for the straight line through $(1, 0, 5)$ which is w parallel to the vector $-3\hat{\mathbf{i}}+\hat{\mathbf{j}}+2\hat{\mathbf{k}}$.

7. (*Equation of a plane*) Let $\mathbf{R}_0$ be a position vector from the origin to a given plane, and let $\mathbf{R}_0$ be normal to the plane so that $\|\mathbf{R}_0\|$ is the shortest distance from the origin to the plane. If $\mathbf{R} = x\hat{\mathbf{i}}+y\hat{\mathbf{j}}+z\hat{\mathbf{k}}$ is the position vector to any point (x, y, z) in the plane, and $\hat{\mathbf{n}}$ is a unit normal vector to the plane, then surely $(\mathbf{R} - \mathbf{R}_0) \cdot \hat{\mathbf{n}} = 0$ or, since $\mathbf{R}_0 \cdot \hat{\mathbf{n}} = \|\mathbf{R}_0\|$,

$$\mathbf{R} \cdot \hat{\mathbf{n}} = \|\mathbf{R}_0\| . \tag{7.1}$$

Equation (7.1) is the general equation of a plane in vector form and is interesting because each quantity has a clear geometrical significance: $\mathbf{R} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$ is a position vector to any point (x, y, z) in the plane, $\hat{\mathbf{n}}$ is a unit vector to the plane (and is unique to within a factor of ± 1), and $\|\mathbf{R}_0\|$ is the shortest distance from the origin to the plane.

(a) Find the shortest distance from the origin to the plane, and a unit vector normal to the plane, for the plane defined by the equation $2x + y - 3z = 4$.
(b) Repeat (a) for $x - y - z = 5$.
(c) Repeat (a) for $3x + y - z = 0$.
(d) Repeat (a) for $x + 2y + 3z = 8$.
(e) Repeat (a) for $3y - z = 5$.
(f) Repeat (a) for $ax + by + cz = d$.
(g) Repeat (a) for $a(x - x_0) + b(y - y_0) + c(z - z_0) = 0$.

8. The vectors $\mathbf{u} = (1, 1, 2)$ and $\mathbf{v} = (3, 2, -1)$ determine a plane (by their span). Find a nonzero vector in that plane that is perpendicular to $\mathbf{w} = (2, 4, 3)$. As usual, explain your reasoning.

9. Same as Exercise 8, for

(a) $\mathbf{u} = (4, -1, 1)$, $\mathbf{v} = (1, 1, 2)$, $\mathbf{w} = (3, 0, 5)$
(b) $\mathbf{u} = (1, 2, 3)$, $\mathbf{v} = (3, 2, 1)$, $\mathbf{w} = (1, 2, 4)$
(c) $\mathbf{u} = (1, 2, 3)$, $\mathbf{v} = (1, 0, 2)$, $\mathbf{w} = (3, 0, 5)$
(d) $\mathbf{u} = (1, 0, 1)$, $\mathbf{v} = (0, 1, 1)$, $\mathbf{w} = (0, 0, 1)$

14.4 Multiple Products

We have discussed the dot and cross products of two vectors, $\mathbf{u} \cdot \mathbf{v}$ and $\mathbf{u} \times \mathbf{v}$, respectively. Products of more than two vectors are also encountered and are the subject of this section.

What products of *three* vectors $\mathbf{u}, \mathbf{v}, \mathbf{w}$ are possible? Writing down such combinations, consider

$$\mathbf{u} \cdot (\mathbf{v} \cdot \mathbf{w}), \quad \mathbf{u} \times (\mathbf{v} \cdot \mathbf{w}), \quad \mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}), \quad \mathbf{u} \times (\mathbf{v} \times \mathbf{w}). \tag{1}$$

Now, $\mathbf{v} \cdot \mathbf{w}$ is a scalar so the first two members of (1) amount to a vector dotted with a scalar and a vector crossed with a scalar, respectively. Such quantities are *not defined* so the first two items in (1) can be discarded. It remains to consider

$$\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w}), \tag{2}$$

called the **scalar triple product** because the result is a scalar, and

$$\mathbf{u} \times (\mathbf{v} \times \mathbf{w}), \tag{3}$$

called the **vector triple product** because the result is a vector.

14.4.1. Scalar triple product. First, the scalar triple product $\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w})$.* For notational simplicity, we can drop the parentheses and write $\mathbf{u} \cdot \mathbf{v} \times \mathbf{w}$ since $\mathbf{u} \cdot \mathbf{v} \times \mathbf{w}$

*The triple product $\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w})$ is sometimes denoted as "$(\mathbf{uvw})$." That notation is not used in this book.

can mean only $(\mathbf{u}\cdot\mathbf{v})\times\mathbf{w}$ or $\mathbf{u}\cdot(\mathbf{v}\times\mathbf{w})$, and of these two the former is not defined and is not a viable candidate.

To observe an important fact about $\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}$, let us sketch the "$\mathbf{u},\mathbf{v},\mathbf{w}$ parallelepiped" shown in Fig. 1 (for the case where $\mathbf{u},\mathbf{v},\mathbf{w}$ do not lie in the same plane). Now, $\mathbf{v}\times\mathbf{w}$ is the magnitude (norm) of $\mathbf{v}\times\mathbf{w}$ times the unit vector $\hat{\mathbf{e}}$ shown in Fig. 1. Thus,

$$\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}=\mathbf{u}\cdot\|\mathbf{v}\times\mathbf{w}\|\,\hat{\mathbf{e}}=\|\mathbf{v}\times\mathbf{w}\|\,(\mathbf{u}\cdot\hat{\mathbf{e}}). \tag{4}$$

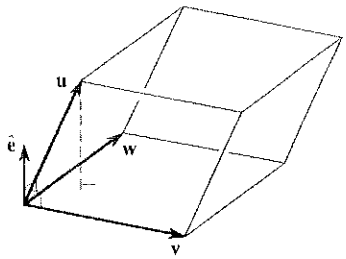


Figure 1. The $\mathbf{u},\mathbf{v},\mathbf{w}$ parallelepiped.

But $\|\mathbf{v}\times\mathbf{w}\|$ is, according to (10) in Section 14.2, the area of the $\mathbf{v},\mathbf{w}$ parallelogram, that is, the *base area* of the $\mathbf{u},\mathbf{v},\mathbf{w}$ parallelepiped. And $\mathbf{u}\cdot\hat{\mathbf{e}}$ is equal to the *altitude* (the dashed line). Hence it follows from (4) that $\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}$ is the volume of the $\mathbf{u},\mathbf{v},\mathbf{w}$ parallelepiped. However, if $\mathbf{u}$ were directed such that $\mathbf{u}\cdot\hat{\mathbf{e}}$ were negative, then $\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}$ would be the *negative* of the volume. In any case, if we introduce absolute values, then

$$\boxed{|\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}|=\text{volume of }\mathbf{u},\mathbf{v},\mathbf{w}\text{ parallelepiped.}} \tag{5}$$

Finally, if we recall from Section 14.3 that

$$\mathbf{v}\times\mathbf{w}=\begin{vmatrix}\hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}}\\ v_1 & v_2 & v_3\\ w_1 & w_2 & w_3\end{vmatrix},$$

where $\mathbf{v}=v_1\hat{\mathbf{i}}+v_2\hat{\mathbf{j}}+v_3\hat{\mathbf{k}}$ and $\mathbf{w}=w_1\hat{\mathbf{i}}+w_2\hat{\mathbf{j}}+w_3\hat{\mathbf{k}}$, it should not be hard to see that $\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}$ can be expressed neatly as

$$\boxed{\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}=\begin{vmatrix}u_1 & u_2 & u_3\\ v_1 & v_2 & v_3\\ w_1 & w_2 & w_3\end{vmatrix}.} \tag{6}$$

Now, interchanging two rows of a determinant changes the sign of the determinant. If, in (6), we interchange the first row with the second, and then the third with the first, the new determinant can be identified [according to the pattern in (6)] as $\mathbf{w}\cdot\mathbf{u}\times\mathbf{v}$. Since there were two sign changes, we see that $\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}=\mathbf{w}\cdot\mathbf{u}\times\mathbf{v}$, or

$$\boxed{\mathbf{u}\cdot\mathbf{v}\times\mathbf{w}=\mathbf{u}\times\mathbf{v}\cdot\mathbf{w},} \tag{7}$$

that is, interchanging the dot and the cross leaves the scalar triple product unchanged.

14.4.2. Vector triple product. Next, consider the vector triple product $\mathbf{u}\times(\mathbf{v}\times\mathbf{w})$. Here we *do* need to retain the parentheses since $\mathbf{u}\times\mathbf{v}\times\mathbf{w}$ could mean either $(\mathbf{u}\times\mathbf{v})\times\mathbf{w}$ or $\mathbf{u}\times(\mathbf{v}\times\mathbf{w})$. Both of these quantities are indeed defined, and they are not necessarily equal to each other; for example, $\hat{\mathbf{i}}\times(\hat{\mathbf{i}}\times\hat{\mathbf{j}})=\hat{\mathbf{i}}\times\hat{\mathbf{k}}=-\hat{\mathbf{j}}$, whereas $(\hat{\mathbf{i}}\times\hat{\mathbf{i}})\times\hat{\mathbf{j}}=\mathbf{0}\times\hat{\mathbf{j}}=\mathbf{0}$.

Thus, let us state, for reference, that in general

$$(\mathbf{u} \times \mathbf{v}) \times \mathbf{w} \neq \mathbf{u} \times (\mathbf{v} \times \mathbf{w}). \tag{8}$$

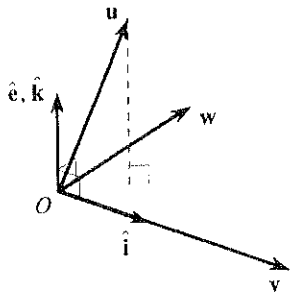


Figure 2. Derivation of (9).

Recall that $\mathbf{v} \times \mathbf{w}$ is perpendicular to the $\mathbf{v}, \mathbf{w}$ plane so it must be some multiple of $\hat{\mathbf{e}}$ (Fig. 2). Similarly, $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$ is perpendicular to both $\mathbf{u}$ and $\mathbf{v} \times \mathbf{w}$, hence, to both $\mathbf{u}$ and $\hat{\mathbf{e}}$. Being perpendicular to $\hat{\mathbf{e}}$, it must lie in the $\mathbf{v}, \mathbf{w}$ plane so it must be expressible in the form $\alpha\mathbf{v} + \beta\mathbf{w}$. To determine α and β, let us introduce a convenient reference coordinate system and set of base vectors. Specifically, let the point O in Fig. 2 be the origin of a Cartesian coordinate system with the usual base vectors $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$. And let the system be oriented so that $\hat{\mathbf{i}}$ is aligned with $\mathbf{v}$, and $\hat{\mathbf{k}}$ coincides with the $\hat{\mathbf{e}}$ shown in Fig. 2. Then $\hat{\mathbf{j}}$ must lie in the $\mathbf{v}, \mathbf{w}$ plane so we can express

$$\mathbf{v} = v_1\hat{\mathbf{i}}, \qquad \mathbf{w} = w_1\hat{\mathbf{i}} + w_2\hat{\mathbf{j}}, \qquad \mathbf{u} = u_1\hat{\mathbf{i}} + u_2\hat{\mathbf{j}} + u_3\hat{\mathbf{k}}.$$

Hence $\mathbf{v} \times \mathbf{w} = v_1 w_2 \hat{\mathbf{k}}$, and

$$\begin{aligned}
\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) &= (u_1\hat{\mathbf{i}} + u_2\hat{\mathbf{j}} + u_3\hat{\mathbf{k}}) \times (v_1 w_2 \hat{\mathbf{k}}) \\
&= -u_1 v_1 w_2 \hat{\mathbf{j}} + u_2 v_1 w_2 \hat{\mathbf{i}} \\
&= (u_1 w_1 + u_2 w_2) v_1 \hat{\mathbf{i}} - u_1 v_1 (w_1 \hat{\mathbf{i}} + w_2 \hat{\mathbf{j}}) \\
&= (u_1 w_1 + u_2 w_2)\mathbf{v} - u_1 v_1 \mathbf{w}.
\end{aligned}$$

But $u_1 w_1 + u_2 w_2 = \mathbf{u} \cdot \mathbf{w}$ and $u_1 v_1 = \mathbf{u} \cdot \mathbf{v}$ so we have the useful formula

$$\boxed{\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) = (\mathbf{u} \cdot \mathbf{w})\mathbf{v} - (\mathbf{u} \cdot \mathbf{v})\mathbf{w}.} \tag{9}$$

Closure. We could consider quadruple products such as $\mathbf{u} \times (\mathbf{v} \times (\mathbf{w} \times \mathbf{x}))$ and $(\mathbf{u} \times \mathbf{v}) \cdot (\mathbf{w} \times \mathbf{x})$, quintuple products, and so on, but the triple products discussed here will suffice for our purposes. In the subsequent sections, the scalar triple product will occur more than the vector triple product so keep in mind the geometrical property (5), the computational formula (6), and the property (7).

EXERCISES 14.4

1. If $\mathbf{u} = \hat{\mathbf{i}} - \hat{\mathbf{j}}$, $\mathbf{v} = 2\hat{\mathbf{i}} + \hat{\mathbf{j}} + 3\hat{\mathbf{k}}$, $\mathbf{w} = \hat{\mathbf{j}} - \hat{\mathbf{k}}$, evaluate

(a) $\mathbf{u} \cdot \mathbf{v} \times \mathbf{w}$ and $\mathbf{u} \times \mathbf{v} \cdot \mathbf{w}$
(b) $\mathbf{u} \times (\mathbf{v} \times \mathbf{w})$ and $(\mathbf{u} \times \mathbf{v}) \times \mathbf{w}$

2. If $\mathbf{A} = \hat{\mathbf{i}} + \hat{\mathbf{j}} - \hat{\mathbf{k}}$, $\mathbf{B} = 3\hat{\mathbf{j}} - \hat{\mathbf{k}}$, $\mathbf{C} = \hat{\mathbf{i}} + 2\hat{\mathbf{k}}$, evaluate

(a) $\mathbf{A} \cdot \mathbf{B} \times \mathbf{C}$ and $\mathbf{A} \times \mathbf{B} \cdot \mathbf{C}$
(b) $\mathbf{A} \times (\mathbf{B} \times \mathbf{C})$ and $(\mathbf{A} \times \mathbf{B}) \times \mathbf{C}$

3. Verify the identity (9) by working out the left- and right-hand sides separately, for each case.

(a) $\mathbf{u} = \hat{\mathbf{i}}$, $\mathbf{v} = \hat{\mathbf{j}}$, $\mathbf{w} = \hat{\mathbf{k}}$
(b) $\mathbf{u} = \hat{\mathbf{i}}$, $\mathbf{v} = \hat{\mathbf{i}}$, $\mathbf{w} = \hat{\mathbf{j}}$
(c) $\mathbf{u} = \hat{\mathbf{j}} - \hat{\mathbf{k}}$, $\mathbf{v} = \hat{\mathbf{i}} + 3\hat{\mathbf{j}}$, $\mathbf{w} = \hat{\mathbf{i}} + 2\hat{\mathbf{k}}$
(d) $\mathbf{u} = 8\hat{\mathbf{i}} + 6\hat{\mathbf{j}}$, $\mathbf{v} = \hat{\mathbf{i}} - 4\hat{\mathbf{k}}$, $\mathbf{w} = 2\hat{\mathbf{j}}$

4. Find the volume of the parallelepiped having the following vectors as adjacent edges.

(a) $\hat{\mathbf{i}} - 2\hat{\mathbf{j}}$, $3\hat{\mathbf{j}}$, $\hat{\mathbf{i}} + \hat{\mathbf{k}}$ (b) $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, $\hat{\mathbf{i}} + \hat{\mathbf{j}} + \hat{\mathbf{k}}$

(c) $\hat{\mathbf{i}} + 2\hat{\mathbf{j}}, 3\hat{\mathbf{i}}, \hat{\mathbf{j}} + \hat{\mathbf{k}}$ (d) $3\hat{\mathbf{j}}, 2\hat{\mathbf{i}} - \hat{\mathbf{j}}, \hat{\mathbf{i}} + 5\hat{\mathbf{k}}$
(e) $\hat{\mathbf{i}}, \hat{\mathbf{i}} + \hat{\mathbf{j}}, \hat{\mathbf{i}} + \hat{\mathbf{j}} + \hat{\mathbf{k}}$ (f) $\hat{\mathbf{i}} + \hat{\mathbf{j}}, \hat{\mathbf{i}}, \hat{\mathbf{i}} + \hat{\mathbf{j}} + \hat{\mathbf{k}}$

5. Show that $(\mathbf{u} \times \mathbf{v}) \times \mathbf{w} = (\mathbf{w} \cdot \mathbf{u})\mathbf{v} - (\mathbf{w} \cdot \mathbf{v})\mathbf{u}$. You may use (9).

6. Show that $\mathbf{u} \times (\mathbf{v} \times \mathbf{w}) + \mathbf{w} \times (\mathbf{u} \times \mathbf{v}) + \mathbf{v} \times (\mathbf{w} \times \mathbf{u}) = \mathbf{0}$.

7. If $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, are distinct vectors issuing from a common point, show that the vector $(\mathbf{A} \times \mathbf{B}) + (\mathbf{B} \times \mathbf{C}) + (\mathbf{C} \times \mathbf{A})$ is perpendicular to the plane containing their heads.

8. (*Orthogonal separation*) Let a given (nonzero) vector $\mathbf{u}$ be separated into the sum of two orthogonal vectors, one parallel to a given (nonzero) vector $\mathbf{v}$ and the other perpendicular to $\mathbf{v}$ (i.e., $\mathbf{u} = \mathbf{u}_{\text{par}} + \mathbf{u}_{\text{perp}}$). Show that

$$\mathbf{u}_{\text{par}} = (\mathbf{u} \cdot \hat{\mathbf{v}})\hat{\mathbf{v}}, \qquad \mathbf{u}_{\text{perp}} = \hat{\mathbf{v}} \times (\mathbf{u} \times \hat{\mathbf{v}}), \tag{8.1}$$

where $\hat{\mathbf{v}} = \mathbf{v}/\|\mathbf{v}\|$.

9. Prove the following identities involving quadruple products. HINT: Use (7) and (9).

(a) $(\mathbf{A} \times \mathbf{B}) \cdot (\mathbf{C} \times \mathbf{D}) = (\mathbf{A} \cdot \mathbf{C})(\mathbf{B} \cdot \mathbf{D}) - (\mathbf{A} \cdot \mathbf{D})(\mathbf{B} \cdot \mathbf{C})$
This is **Lagrange's identity**, of which equation (5.1) of Exercise 5 in Section 14.2 is a special case.
(b) $(\mathbf{A} \times \mathbf{B}) \times (\mathbf{C} \times \mathbf{D}) = (\mathbf{A} \cdot \mathbf{B} \times \mathbf{D})\mathbf{C} - (\mathbf{A} \cdot \mathbf{B} \times \mathbf{C})\mathbf{D}$

10. In deriving (9), we introduced the orthonormal basis $\left\{\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}\right\}$, oriented in such a way that $\mathbf{v} = v_1\hat{\mathbf{i}}$, $\mathbf{w} = w_1\hat{\mathbf{i}} + w_2\hat{\mathbf{j}}$, and $\mathbf{u} = u_1\hat{\mathbf{i}} + u_2\hat{\mathbf{j}} + u_3\hat{\mathbf{k}}$. Although this orientation was *convenient* (since then $\mathbf{v}$ had only one component, and $\mathbf{w}$ had only two, so that the calculations were rendered as short as possible), we claim that it was not essential. Verify this claim by rederiving (9) using the expansions $\mathbf{u} = u_1\hat{\mathbf{i}} + u_2\hat{\mathbf{j}} + u_3\hat{\mathbf{k}}$, $\mathbf{v} = v_1\hat{\mathbf{i}} + v_2\hat{\mathbf{j}} + v_3\hat{\mathbf{k}}$, $\mathbf{w} = w_1\hat{\mathbf{i}} + w_2\hat{\mathbf{j}} + w_3\hat{\mathbf{k}}$, which, of course, hold for *any* orientation.

11. (*Linear dependence of* $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$) Prove that vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ are linearly dependent if and only if $\mathbf{u} \cdot \mathbf{v} \times \mathbf{w} = 0$.

12. Do the following sets of four points lie in a plane? HINT: You may use the result stated in Exercise 11, without proving it, if you wish.

(a) $(1, 3, 0)$, $(2, 1, -1)$, $(0, 0, 4)$, $(5, 0, 8)$
(b) $(2, 1, -1)$, $(1, 3, 0)$, $(5, 0, 9)$, $(0, 0, 4)$
(c) $(4, 0, 0)$, $(0, 1, 0)$, $(1, 2, -4)$, $(0, 0, 1)$
(d) $(0, 0, 0)$, $(2, 6, -1)$, $(1, 0, 1)$, $(1, 2, 0)$
(e) $(1, 0, 1)$, $(2, 1, 3)$, $(1, -1, 0)$, $(3, -1, 2)$
(f) $(0, 0, 2)$, $(0, 1, 3)$, $(1, 2, 3)$, $(2, 3, 4)$

13. Give an example of three vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ (*not* the ones given in the text) such that $(\mathbf{u} \times \mathbf{v}) \times \mathbf{w} \neq \mathbf{u} \times (\mathbf{v} \times \mathbf{w})$, and an example of three vectors $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ such that $(\mathbf{u} \times \mathbf{v}) \times \mathbf{w} = \mathbf{u} \times (\mathbf{v} \times \mathbf{w})$. (The latter part of the exercise is simple if we let at least one of the three vectors be $\mathbf{0}$. To make the problem more interesting, let us insist that $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ all be nonzero.)

14.5 Differentiation of a Vector Function of a Single Variable

Suppose a particle moves about in 3-space so that its position vector $\mathbf{R}(t)$, from the origin to the particle, is a function of the time t. By definition, its velocity vector $\mathbf{v}(t)$ is the derivative $\mathbf{v} = d\mathbf{R}/dt$. This is but one example of many, where we need to differentiate a vector function. Although we discuss vectors at length in Chapter 9 and in the Sections 14.1–14.4, we have not yet introduced the concept of the differentiataion of a vector function. In fact, we have not yet introduced the concept of a vector function!

Suppose a vector $\mathbf{u}(\tau)$ is defined for each value of a real variable τ on some τ interval. Then we say that $\mathbf{u}$ is a **vector function** of τ on that interval. The physical nature of $\mathbf{u}$, and of τ, is in no way restricted. For instance, $\mathbf{u}$ might represent the position vector of a particle or an electric field vector, and its argument τ might be the time, arc length along a curve, or it might be a purely mathematical parameter.

To define the derivative $d\mathbf{u}/d\tau$, or $\mathbf{u}'(\tau)$, we stay as close as possible to the definition of the derivative given in the calculus, and define

$$\boxed{\frac{d\mathbf{u}}{d\tau} = \mathbf{u}'(\tau) \equiv \lim_{\Delta\tau \to 0} \frac{\mathbf{u}(\tau + \Delta\tau) - \mathbf{u}(\tau)}{\Delta\tau}.} \tag{1}$$

If the limit on the right-hand side exists, for a given value of τ, we say that $\mathbf{u}(\tau)$ is **differentiable** at that τ, and that $\mathbf{u}'(\tau)$ is its **derivative**.

Observe that the right-hand side of (1) is of the general form

$$\lim_{\tau \to \tau_0} \mathbf{F}(\tau). \tag{2}$$

Thus, our definition (1) is meaningful only if we also define the **limit of a vector sequence**, which we do as follows. By

$$\lim_{\tau \to \tau_0} \mathbf{F}(\tau) = \mathbf{L} \tag{3}$$

[i.e., $\mathbf{F}(\tau)$ tends to $\mathbf{L}$ as τ tends to τ_0] we shall mean that

$$\lim_{\tau \to \tau_0} \|\mathbf{F}(\tau) - \mathbf{L}\| = 0, \tag{4}$$

where the limit in (4) is the ordinary limit defined in the calculus since the norm $\|\mathbf{F}(\tau) - \mathbf{L}\|$ is a scalar function of τ. Further, we say that $\mathbf{F}(\tau)$ is **continuous** at τ_0 if $\lim_{\tau \to \tau_0} \mathbf{F}(\tau) = \mathbf{F}(\tau_0)$.

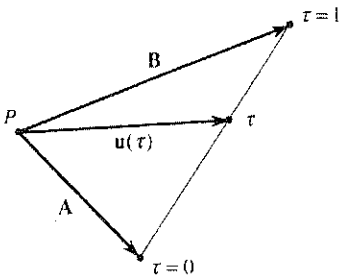


Figure 1. $\mathbf{u}(\tau)$ given by (5).

EXAMPLE 1. To illustrate (1), suppose that

$$\mathbf{u}(\tau) = \mathbf{A} + \tau^3(\mathbf{B} - \mathbf{A}), \qquad (0 \le \tau \le 1) \tag{5}$$

where $\mathbf{A}$ and $\mathbf{B}$ are given vectors that do not vary with τ. If the tail of $\mathbf{u}$ is fixed at P, then $\mathbf{u}(\tau)$ varies with τ as shown in Fig. 1. Then (1) gives

$$\begin{aligned} \mathbf{u}'(\tau) &= \lim_{\Delta\tau \to 0} \frac{\mathbf{A} + (\tau + \Delta\tau)^3(\mathbf{B} - \mathbf{A}) - \mathbf{A} - \tau^3(\mathbf{B} - \mathbf{A})}{\Delta\tau} \\ &= \lim_{\Delta\tau \to 0} \frac{3\tau^2\Delta\tau + 3\tau(\Delta\tau)^2 + (\Delta\tau)^3}{\Delta\tau}(\mathbf{B} - \mathbf{A}) \\ &= \lim_{\Delta\tau \to 0} \left[3\tau^2 + 3\tau\Delta\tau + (\Delta\tau)^2\right](\mathbf{B} - \mathbf{A}) = 3\tau^2(\mathbf{B} - \mathbf{A}). \end{aligned} \tag{6}$$

To illustrate, we have shown $\mathbf{u}'(\tau)$ at $\tau = 0.5$ in Fig. 2. ∎

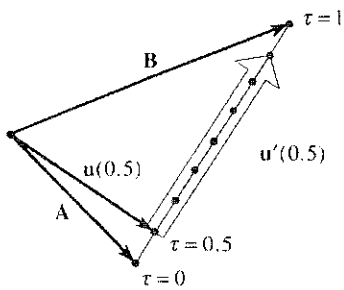


Figure 2. $\mathbf{u}'(\tau)$ at $\tau = 0.5$.

Just as we have rules such as $(uv)' = u'v + uv'$ for the differentiation of scalar functions, we have a number of such rules, which we now list, for the differentiation of vector functions:

$$(f\mathbf{u})' = f'\mathbf{u} + f\mathbf{u}', \qquad \text{where } f = f(\tau) \tag{7a}$$
$$(\alpha\mathbf{u} + \beta\mathbf{v})' = \alpha\mathbf{u}' + \beta\mathbf{v}', \qquad \alpha, \beta \text{ constants} \tag{7b}$$
$$(\mathbf{u}\cdot\mathbf{v})' = \mathbf{u}'\cdot\mathbf{v} + \mathbf{u}\cdot\mathbf{v}', \tag{7c}$$
$$(\mathbf{u}\times\mathbf{v})' = \mathbf{u}'\times\mathbf{v} + \mathbf{u}\times\mathbf{v}', \tag{7d}$$
$$(\mathbf{u}\cdot\mathbf{v}\times\mathbf{w})' = \mathbf{u}'\cdot\mathbf{v}\times\mathbf{w} + \mathbf{u}\cdot\mathbf{v}'\times\mathbf{w} + \mathbf{u}\cdot\mathbf{v}\times\mathbf{w}', \tag{7e}$$
$$[\mathbf{u}\times(\mathbf{v}\times\mathbf{w})]' = \mathbf{u}'\times(\mathbf{v}\times\mathbf{w}) + \mathbf{u}\times(\mathbf{v}'\times\mathbf{w}) + \mathbf{u}\times(\mathbf{v}\times\mathbf{w}'), \tag{7f}$$
$$\frac{d}{d\tau}\mathbf{u}(f(\tau)) = \frac{d\mathbf{u}}{df}f'(\tau), \qquad \text{(chain rule)} \tag{7g}$$

provided, of course, that the needed derivatives do exist.

As representative, let us prove (7c). Using the definition (1), we find that

$$\begin{aligned}
(\mathbf{u}\cdot\mathbf{v})' &= \lim_{\Delta\tau\to 0}\frac{\mathbf{u}(\tau+\Delta\tau)\cdot\mathbf{v}(\tau+\Delta\tau) - \mathbf{u}(\tau)\cdot\mathbf{v}(\tau)}{\Delta\tau}\\
&= \lim_{\Delta\tau\to 0}\frac{\mathbf{u}(\tau+\Delta\tau)\cdot\{\mathbf{v}(\tau) + [\mathbf{v}(\tau+\Delta\tau) - \mathbf{v}(\tau)]\} - \mathbf{u}(\tau)\cdot\mathbf{v}(\tau)}{\Delta\tau}\\
&= \lim_{\Delta\tau\to 0}\frac{[\mathbf{u}(\tau+\Delta\tau) - \mathbf{u}(\tau)]\cdot\mathbf{v}(\tau) + \mathbf{u}(\tau+\Delta\tau)\cdot[\mathbf{v}(\tau+\Delta\tau) - \mathbf{v}(\tau)]}{\Delta\tau}\\
&= \lim_{\Delta\tau\to 0}\left[\frac{\mathbf{u}(\tau+\Delta\tau) - \mathbf{u}(\tau)}{\Delta\tau}\right]\cdot\mathbf{v}(\tau) + \lim_{\Delta\tau\to 0}\mathbf{u}(\tau+\Delta\tau)\cdot\left[\frac{\mathbf{v}(\tau+\Delta\tau) - \mathbf{v}(\tau)}{\Delta\tau}\right]\\
&= \mathbf{u}'(\tau)\cdot\mathbf{v}(\tau) + \mathbf{u}(\tau)\cdot\mathbf{v}'(\tau).
\end{aligned} \tag{8}$$

Proofs of the other formulas are left for the exercises.

To illustrate the use of these rules, let us differentiate the $\mathbf{u}(\tau)$ given by (5) again, but instead of using the cumbersome difference quotient definition (1) let us use the relevant formulas in (7):

$$\begin{aligned}
\mathbf{u}'(\tau) &= \frac{d}{d\tau}[\mathbf{A} + \tau^3(\mathbf{B} - \mathbf{A})]\\
&= \frac{d}{d\tau}\mathbf{A} + \frac{d}{d\tau}[\tau^3(\mathbf{B} - \mathbf{A})] \qquad \text{per (7b) with } \alpha = \beta = 1\\
&= 0 + \frac{d}{d\tau}(\tau^3)(\mathbf{B} - \mathbf{A}) + \tau^3\frac{d}{d\tau}(\mathbf{B} - \mathbf{A}) \qquad \text{per (7a)}\\
&= 3\tau^2(\mathbf{B} - \mathbf{A}) + \mathbf{0}\\
&= 3\tau^2(\mathbf{B} - \mathbf{A}),
\end{aligned} \tag{9}$$

as found in Example 1.

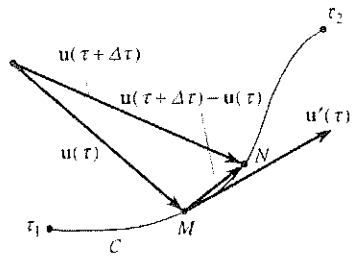


Figure 3. Graphical interpretation of $\mathbf{u}(\tau)$.

The definition (1) admits a graphical interpretation. Suppose that the tail of the $\mathbf{u}(\tau)$ vector is "bound" at a fixed point P, and that the head of the vector traces a space curve C as τ varies continuously over some interval $\tau_1 \le \tau \le \tau_2$, for example, as sketched in Fig. 3. The difference $\mathbf{u}(\tau + \Delta\tau) - \mathbf{u}(\tau)$, in (1) is the vector $\mathbf{MN}$. The endpoint M is fixed, for a given value of τ, and the other endpoint N moves along C and tends to M as $\Delta\tau \to 0$. The limiting direction of

$[\mathbf{u}(\tau+\Delta\tau)-\mathbf{u}(\tau)]/\Delta\tau$ as $\Delta\tau \to 0$, that is, the direction of $\mathbf{u}'(\tau)$ defines a straight line called the **tangent line** to $\mathcal{C}$ at τ [provided that $\mathbf{u}(\tau)$ is indeed differentiable there, and that $\mathbf{u}'(\tau) \neq \mathbf{0}$, for the zero vector would not imply any particular direction], and $\mathbf{u}'(\tau)$ is called a **tangent vector**. In Example 1, for instance, $\mathcal{C}$ is simply a straight line, and $\mathbf{u}'(\tau)$ was indeed found to be tangent to that line.

EXAMPLE 2. Suppose that a particle is located, with respect to a plane Cartesian coordinate system, by the position vector

$$\mathbf{R}(t) = x(t)\hat{\mathbf{i}} + y(t)\hat{\mathbf{j}} = t\hat{\mathbf{i}} + t^2\hat{\mathbf{j}}, \tag{10}$$

where t is the time. Then $\mathbf{R}'(t) = \mathbf{v}(t)$ is called the **velocity** of the particle, and $\mathbf{R}''(t) = \mathbf{v}'(t) = \mathbf{a}(t)$ is called its **acceleration**:

$$\begin{aligned}
\mathbf{v}(t) = \mathbf{R}'(t) &= \frac{d}{dt}(t\hat{\mathbf{i}} + t^2\hat{\mathbf{j}}) \\
&= \frac{d}{dt}(t\hat{\mathbf{i}}) + \frac{d}{dt}(t^2\hat{\mathbf{j}}) \qquad \text{per (7b) with } \alpha = \beta = 1 \\
&= \frac{d}{dt}(t)\hat{\mathbf{i}} + \frac{d}{dt}(t^2)\hat{\mathbf{j}} \qquad \text{per (7a)} \\
&= \hat{\mathbf{i}} + 2t\hat{\mathbf{j}},
\end{aligned} \tag{11}$$

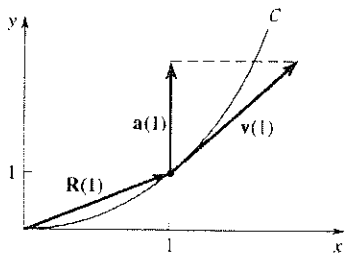


Figure 4. The motion (10).

and, similarly,

$$\mathbf{a}(t) = \mathbf{v}'(t) = 2\hat{\mathbf{j}}. \tag{12}$$

In this case the acceleration happens to be a constant vector. The vectors $\mathbf{R}$, $\mathbf{v}$, and $\mathbf{a}$ are shown in Fig. 4 at the instant $t = 1$, at which $\mathbf{R}(1) = \hat{\mathbf{i}} + \hat{\mathbf{j}}$, $\mathbf{v}(1) = \hat{\mathbf{i}} + 2\hat{\mathbf{j}}$, and $\mathbf{a}(1) = 2\hat{\mathbf{j}}$. Since $x = t$ and $y = t^2$, it follows that $y = x^2$ on $\mathcal{C}$. When $t = 1$ we have $x = 1$, and the slope of $\mathcal{C}$ is $dy/dx = 2x = 2$, which is identical to the slope of $v(1) = \hat{\mathbf{i}} + 2\hat{\mathbf{j}}$, so $\mathbf{v}(1)$ is indeed tangent to $\mathcal{C}$, as stated above this example. Of course, $\mathbf{v}(t)$ is tangent to $\mathcal{C}$ for every time t; we chose $t = 1$ just to illustrate. ∎

Closure. We introduce the idea of a vector function, $\mathbf{u}(\tau)$, and define its derivative $\mathbf{u}'(\tau)$ as the limit of the difference quotient given in (1). Because this definition echoes the definition of the derivative of a *scalar* function, familiar from the differential calculus, there are no surprises among the rules listed in (7). Thus, you should feel quite comfortable with the result that if $\mathbf{R}(t) = t\hat{\mathbf{i}} + t^2\hat{\mathbf{j}}$, then $\mathbf{R}'(t) = \hat{\mathbf{i}} + 2t\hat{\mathbf{j}}$, given in (11), for instance.

EXERCISES 14.5

1. For each of the following $\mathbf{u}(\tau)$ vectors, determine $\mathbf{u}'$, $\mathbf{u}''$, and $\|\mathbf{u}'\|$.

(a) $\mathbf{A} + \tau^2\mathbf{B}$ ($\mathbf{A}, \mathbf{B}$ constant vectors)

(b) $6\hat{\mathbf{i}} - e^{\tau}\hat{\mathbf{j}}$

(c) $\tau^2\hat{\mathbf{i}} - 4\hat{\mathbf{j}} + 3\cos 2\tau\hat{\mathbf{k}}$

(d) $\cos\tau\hat{\mathbf{i}} + \sin\tau\hat{\mathbf{j}} + \sin\tau\hat{\mathbf{k}}$

(e) $e^{-\tau}(\cos 2\tau\hat{\mathbf{i}} + \sin 2\tau\hat{\mathbf{j}})$
(f) $e^{\tau}(\cos 2\tau\hat{\mathbf{i}} + \sin 2\tau\hat{\mathbf{j}})$
(g) $\cos 3\tau\hat{\mathbf{i}} - 4\hat{\mathbf{j}} - \sin 3\tau\hat{\mathbf{k}}$
(h) $3\tau\hat{\mathbf{i}} - \hat{\mathbf{j}} + 2\tau^3\hat{\mathbf{k}}$

2. (a) Letting $\mathbf{u} = \tau^2\hat{\mathbf{i}} - \tau\hat{\mathbf{j}} - 3\tau\hat{\mathbf{k}}$, $\mathbf{v} = 2\cos\tau\hat{\mathbf{i}} - \tau\hat{\mathbf{k}}$, and $\mathbf{w} = \sin\tau\hat{\mathbf{j}}$, verify (7c) by working out the left- and right-hand sides.
(b) Repeat part (a), for (7d).
(c) Repeat part (a), for (7e).
(d) Repeat part (a), for (7f).

3. (a) Letting $\mathbf{u} = 3\tau\hat{\mathbf{i}} + \tau^4\hat{\mathbf{k}}$, $\mathbf{v} = \hat{\mathbf{i}} - 4\tau\hat{\mathbf{j}}$, and $\mathbf{w} = 5\tau^2\hat{\mathbf{i}} + \hat{\mathbf{j}} - \tau\hat{\mathbf{k}}$, verify (7c) by working out the left- and right-hand sides.
(b) Repeat part (a), for (7d).
(c) Repeat part (a), for (7e).
(d) Repeat part (a), for (7f).

4. Use the definition (1) to
(a) prove (7a)
(b) prove (7d)

5. (a) Show that (7e) follows from (7c) and (7d).
(b) Show that (7f) follows from (7d).

6. Derive (7g). HINT: Express $\mathbf{u}(f(\tau)) = u_1(f(\tau))\hat{\mathbf{i}} + u_2(f(\tau))\hat{\mathbf{j}} + u_3(f(\tau))\hat{\mathbf{k}}$.

7. (a) Obtain an expression for $(\mathbf{u}\cdot\mathbf{v})''$ analogous to the expression for $(\mathbf{u}\cdot\mathbf{v})'$ given by (7c).
(b) Obtain an expression for $(\mathbf{u}\times\mathbf{v})''$ analogous to the expression for $(\mathbf{u}\times\mathbf{v})'$ given by (7d).

8. (a) Derive the formula

$$\|\mathbf{u}\|' = \frac{\mathbf{u}\cdot\mathbf{u}'}{\|\mathbf{u}\|}. \tag{8.1}$$

(b) Show, from (8.1), that *if* $\|\mathbf{u}(\tau)\| = \textit{constant} \neq 0$, *then* $\mathbf{u}'(\tau)$ *is necessarily either perpendicular to* $\mathbf{u}(\tau)$, *or* $\mathbf{0}$.

9. In (8), we give an intrinsic proof of (7c), that is, one that does not rely on any coordinate system. Alternatively, prove (7c) by introducing Cartesian coordinates so that $\mathbf{u}(\tau) = u_1(\tau)\hat{\mathbf{i}} + u_2(\tau)\hat{\mathbf{j}} + u_3(\tau)\hat{\mathbf{k}}$ and $\mathbf{v}(\tau) = v_1(\tau)\hat{\mathbf{i}} + v_2(\tau)\hat{\mathbf{j}} + v_3(\tau)\hat{\mathbf{k}}$, differentiating $\mathbf{u}(\tau)\cdot\mathbf{v}(\tau)$, and then identifying the result as $\mathbf{u}'\cdot\mathbf{v} + \mathbf{u}\cdot\mathbf{v}'$.

14.6 Non-Cartesian Coordinates (Optional)

In preceding sections we worked either with no reference coordinate system at all, or with a Cartesian system. Often, it is convenient to work with a coordinate system other than Cartesian, a "non-Cartesian" system. Our purpose in the present section is to introduce three important non-Cartesian systems: plane polar, cylindrical, and spherical. We choose to do this within the context of a physical application, the kinematics of a point (or particle). Specifically, we will develop expressions for the position vector $\mathbf{R}$, the velocity vector $\mathbf{v}$, and the acceleration $\mathbf{a}$ of a point as it moves about in space, in terms of these three coordinate systems.

For reference, let us first note the corresponding results in a Cartesian system x, y, z, with base vectors $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$. If

$$\mathbf{R}(t) = x(t)\hat{\mathbf{i}} + y(t)\hat{\mathbf{j}} + z(t)\hat{\mathbf{k}} \tag{1}$$

denotes the **position vector** from the origin of a (stationary) Cartesian coordinate system to the point $(x(t), y(t), z(t))$, where t is the time, then the *velocity* $\mathbf{v}$ and *acceleration* $\mathbf{a}$ of the point are

$$\mathbf{v}(t) = \mathbf{R}'(t) = x'(t)\hat{\mathbf{i}} + y'(t)\hat{\mathbf{j}} + z'(t)\hat{\mathbf{k}}, \tag{2}$$

$$\mathbf{a}(t) = \mathbf{v}'(t) = \mathbf{R}''(t) = x''(t)\hat{\mathbf{i}} + y''(t)\hat{\mathbf{j}} + z''(t)\hat{\mathbf{k}}, \tag{3}$$

respectively, where primes denote differentiation with respect to the time t.

14.6.1. Plane polar coordinates. The coordinates r, θ and their respective unit base vectors $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta$ are shown in Fig. 1; r, θ are related to the reference x, y coordinates according to

$$\begin{aligned} x &= r\cos\theta, \\ y &= r\sin\theta, \end{aligned} \tag{4a}$$

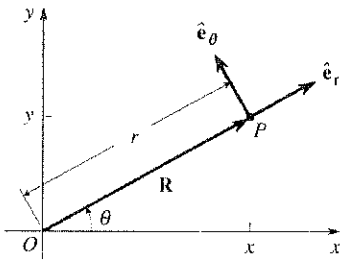


Figure 1. Plane polar coordinates.

or, solving (4a) for r and θ,

$$\begin{aligned} r &= \sqrt{x^2 + y^2}, \\ \theta &= \tan^{-1}\frac{y}{x}, \end{aligned} \tag{4b}$$

where θ is undefined at the origin. It will be understood that θ is measured in radians unless we specify that it is measured in degrees.

In general, both r and θ may be positive, negative, or zero, but in this book we choose always to have $r \geq 0$ so that r is the *distance* from the origin O to the point in question, say P. In addition, if we limit θ according to $0 \leq \theta < 2\pi$, each point in the x, y plane (except for the origin) corresponds to a unique value of θ.*

The base vectors $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta$ at any point P (other than the origin) are in the positive r, θ directions, respectively. Observe that while $\hat{\mathbf{i}}, \hat{\mathbf{j}}$ are *constant* vectors,† $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta$ are not! Specifically, $\hat{\mathbf{e}}_r$ and $\hat{\mathbf{e}}_\theta$ vary with θ, although not with r. For example, if we imagine sliding $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta$ to a new r location, say $r + \Delta r$, with θ held fixed, we see that both vectors remain unchanged because both their magnitude (unity) and their direction remain unchanged, even though their location is different. But if we vary θ, with r fixed, $\hat{\mathbf{e}}_r$ and $\hat{\mathbf{e}}_\theta$ both rotate. The upshot is that

$$\hat{\mathbf{e}}_r = \hat{\mathbf{e}}_r(\theta), \qquad \hat{\mathbf{e}}_\theta = \hat{\mathbf{e}}_\theta(\theta). \tag{5}$$

To obtain expressions for the velocity $\mathbf{v}$ and acceleration $\mathbf{a}$ of a point P moving in the x, y plane, we begin with the position vector from O to P (Fig. 1), namely,

$$\boxed{\mathbf{R} = r\hat{\mathbf{e}}_r.} \tag{6}$$

Before proceeding, two points should be made in connection with (6). First, it is tempting to write "$\mathbf{R} = r\hat{\mathbf{e}}_r + \theta\hat{\mathbf{e}}_\theta$," in place of (6), in an effort to follow the pattern

*Unfortunately, the relation $\theta = \tan^{-1} y/x$, given in (4b), does not completely suffice to determine the unique value of θ at a given point without taking into account the signs of x and y. For example, if $x = y = 1$, then surely $\theta = \pi/4$, whereas $\theta = \tan^{-1} 1$ gives both (the correct value) $\theta = \pi/4$ and the (incorrect value) $\theta = 5\pi/4$. Such ambiguity can always be resolved by the original relations $x = r\cos\theta$, $y = r\sin\theta$. For example, for the case $x = y = 1$ mentioned above, $\theta = \pi/4$ satisfies the relations $1 = \sqrt{2}\cos\theta$, $1 = \sqrt{2}\sin\theta$, whereas $\theta = 5\pi/4$ does not and is to be discarded.

†We should qualify that statement: Cartesian base vectors $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$ do not vary with the *space* coordinates x, y, z. They may, however, vary with *time* if the coordinate system itself is rotating with respect to the observer.

in the Cartesian formula $\mathbf{R} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}}$. However, the quoted expression is incorrect; the correct version is given by (6), since (from Fig. 1) the magnitude of $\mathbf{R}$ is r and its direction is $\hat{\mathbf{e}}_r$. Second, is it not strange that θ does not appear in (6)? The answer is that θ *does* appear in (6) since, according to (5), $\hat{\mathbf{e}}_r$ is a function of θ; that is, (6) is really $\mathbf{R} = r\hat{\mathbf{e}}_r(\theta)$.

Now, if the point P moves about according to $r = r(t)$, $\theta = \theta(t)$, then

$$\mathbf{R} = r(t)\hat{\mathbf{e}}_r(\theta(t)), \tag{7}$$

so its velocity is [recall (7a) in Section 14.5]

$$\mathbf{v}(t) = \dot{\mathbf{R}} = \dot{r}\hat{\mathbf{e}}_r + r\dot{\hat{\mathbf{e}}}_r, \tag{8}$$

where we choose to use overhead dots, rather than the more usual primes, to denote differentiation with respect to the time t. Since $\hat{\mathbf{e}}_r = \hat{\mathbf{e}}_r(\theta(t))$, we can use chain differentiation to express

$$\dot{\hat{\mathbf{e}}}_r = \frac{d}{dt}\hat{\mathbf{e}}_r(\theta(t)) = \frac{d\hat{\mathbf{e}}_r}{d\theta}\frac{d\theta}{dt} = \dot{\theta}\frac{d\hat{\mathbf{e}}_r}{d\theta}. \tag{9}$$

Fine, but what is $d\hat{\mathbf{e}}_r/d\theta$? Since we are working with the base vectors $\hat{\mathbf{e}}_r$ and $\hat{\mathbf{e}}_\theta$, we wish to expand *all* vectors in terms of these base vectors. Thus, we wish to expand $d\hat{\mathbf{e}}_r/d\theta$ in the form $(\ )\hat{\mathbf{e}}_r + (\ )\hat{\mathbf{e}}_\theta$. To accomplish that, let us fall back on the definition of the derivative,

$$\frac{d\hat{\mathbf{e}}_r}{d\theta} = \lim_{\Delta\theta\to 0}\frac{\hat{\mathbf{e}}_r(\theta + \Delta\theta) - \hat{\mathbf{e}}_r(\theta)}{\Delta\theta}. \tag{10}$$

The vector difference in the numerator is readily evaluated if we slide the two vectors back to O (Fig. 2), so that they are tail-to-tail. Then $\hat{\mathbf{e}}_r(\theta + \Delta\theta) - \hat{\mathbf{e}}_r(\theta)$ is the little vector from a to b; its length is $1\Delta\theta$ and its direction, in the limit as $\Delta\theta \to 0$, is $\hat{\mathbf{e}}_\theta$.* Thus,

$$\frac{d\hat{\mathbf{e}}_r}{d\theta} = \lim_{\Delta\theta\to 0}\frac{(1\Delta\theta)\hat{\mathbf{e}}_\theta}{\Delta\theta} = \hat{\mathbf{e}}_\theta \tag{11}$$

so (9) becomes $\dot{\hat{\mathbf{e}}}_r = \dot{\theta}\hat{\mathbf{e}}_\theta$, and hence (8) becomes

$$\boxed{\mathbf{v}(t) = \dot{r}\hat{\mathbf{e}}_r + r\dot{\theta}\hat{\mathbf{e}}_\theta.} \tag{12}$$

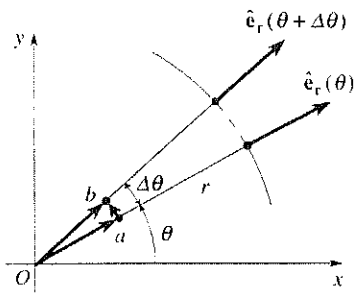


Figure 2. Calculation of $d\hat{\mathbf{e}}_r/d\theta$.

The expressions $\dot{r}, r\dot{\theta}$ for the r, θ velocity components, respectively, may well be familiar to you from a course in introductory physics.

Continuing, we differentiate once more to obtain the acceleration,

$$\mathbf{a}(t) = \dot{\mathbf{v}}(t) = \ddot{r}\hat{\mathbf{e}}_r + \dot{r}\dot{\hat{\mathbf{e}}}_r + \dot{r}\dot{\theta}\hat{\mathbf{e}}_\theta + r\ddot{\theta}\hat{\mathbf{e}}_\theta + r\dot{\theta}\dot{\hat{\mathbf{e}}}_\theta. \tag{13}$$

*As $\Delta\theta \to 0$, the length of the vector from a to b, say $\mathbf{ab}$, approaches the arc length of the circular arc through those points and centered at O. Thus, it follows from the formula $s = r\theta$ for the arc length s of a circular arc of radius r and angle θ, that $\|\mathbf{ab}\| \sim 1\Delta\theta$ as $\Delta\theta \to 0$.

We have already seen that $\dot{\hat{\mathbf{e}}}_r = \dot{\theta}\hat{\mathbf{e}}_\theta$, but we still need to evaluate the $\dot{\hat{\mathbf{e}}}_\theta$ in (13). We have

$$\dot{\hat{\mathbf{e}}}_\theta = \frac{d}{dt}\hat{\mathbf{e}}_\theta[\theta(t)] = \frac{d\hat{\mathbf{e}}_\theta}{d\theta}\frac{d\theta}{dt} = \dot{\theta}\frac{d\hat{\mathbf{e}}_\theta}{d\theta}, \tag{14}$$

where

$$\frac{d\hat{\mathbf{e}}_\theta}{d\theta} = \lim_{\Delta\theta\to 0}\frac{\hat{\mathbf{e}}_\theta(\theta+\Delta\theta) - \hat{\mathbf{e}}_\theta(\theta)}{\Delta\theta}. \tag{15}$$

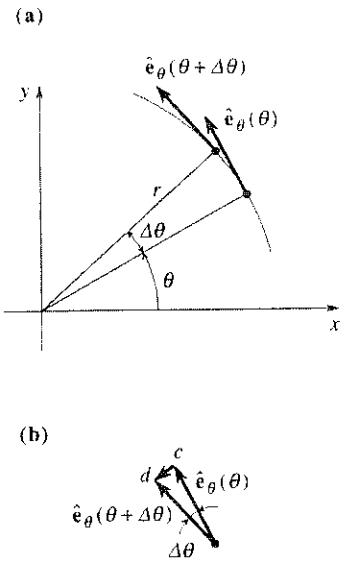


Figure 3. Calculation of $d\hat{\mathbf{e}}_\theta/d\theta$.

To evaluate the difference $\hat{\mathbf{e}}_\theta(\theta+\Delta\theta) - \hat{\mathbf{e}}_\theta(\theta)$, where these two vectors are shown in Fig. 3a, we pick them up and place them tail to tail (without altering their orientation), as shown in Fig. 3b. Their difference is the little vector from c to d, namely, $(1\,\Delta\theta)(-\hat{\mathbf{e}}_r)$ so that

$$\frac{d\hat{\mathbf{e}}_\theta}{d\theta} = \lim_{\Delta\theta\to 0}\frac{(1\,\Delta\theta)(-\hat{\mathbf{e}}_r)}{\Delta\theta} = -\hat{\mathbf{e}}_r. \tag{16}$$

Then (14) becomes $\dot{\hat{\mathbf{e}}}_\theta = -\dot{\theta}\hat{\mathbf{e}}_r$, and (13) gives

$$\mathbf{a}(t) = \ddot{r}\hat{\mathbf{e}}_r + \dot{r}\dot{\theta}\hat{\mathbf{e}}_\theta + \dot{r}\dot{\theta}\hat{\mathbf{e}}_\theta + r\ddot{\theta}\hat{\mathbf{e}}_\theta - r\dot{\theta}^2\hat{\mathbf{e}}_r$$

or,

$$\boxed{\mathbf{a}(t) = (\ddot{r} - r\dot{\theta}^2)\hat{\mathbf{e}}_r + (r\ddot{\theta} + 2\dot{r}\dot{\theta})\hat{\mathbf{e}}_\theta.} \tag{17}$$

Of the four terms constituting $\mathbf{a}$, two with well-known names are the *centripetal acceleration* $-r\dot{\theta}^2\hat{\mathbf{e}}_r$, and the *Coriolis acceleration* $2\dot{r}\dot{\theta}\hat{\mathbf{e}}_\theta$.*

Although the formulas (12) and (17) are of great importance, for our purposes (especially in Chapter 16 on scalar and vector field theory) the most important part of this discussion concerns the differentiation of the base vectors, for in non-Cartesian coordinate systems the base vectors generally depend on one or more of the coordinates so whenever we differentiate a vector we need to face up to differentiating the base vectors. In particular, for plane polar coordinates we found that the base vectors $\hat{\mathbf{e}}_r$ and $\hat{\mathbf{e}}_\theta$ vary with θ, but not with r, and that

$$\boxed{\frac{d\hat{\mathbf{e}}_r}{d\theta} = \hat{\mathbf{e}}_\theta \quad \text{and} \quad \frac{d\hat{\mathbf{e}}_\theta}{d\theta} = -\hat{\mathbf{e}}_r.} \tag{18a,b}$$

The method that we used to derive equations (18a,b) will be called the **difference quotient method**. That method is appealing pedagogically because it relies on fundamentals — the difference quotient definitions (10) and (15) together with the graphical representation of the various terms. However, it is heuristic rather than rigorous. For instance, expressing $\mathbf{cd} = (1\Delta\theta)(-\hat{\mathbf{e}}_r)$, in (16), seemed to be correct but was not proved.

*The latter is named after the French engineer *G. Coriolis* (1792–1843), but it might be mentioned that the expressions $a_r = \ddot{r} - r\dot{\theta}^2$ and $a_\theta = r\ddot{\theta} + 2\dot{r}\dot{\theta}$ were derived by Euler in his book *Theoria Motus Corporum Solidorum seu Rigidorum*, published in 1765.

Thus, we now present a different line of approach which *is* rigorous, and which we will call the **transform method**. The idea is to express $\hat{\mathbf{e}}_r$ and $\hat{\mathbf{e}}_\theta$ (temporarily) in terms of the alternative Cartesian basis $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$, the point being that $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ are constant vectors and hence readily differentiated. Specifically,

$$\hat{\mathbf{e}}_r = \cos\theta\hat{\mathbf{i}} + \sin\theta\hat{\mathbf{j}}, \tag{19a}$$

$$\hat{\mathbf{e}}_\theta = -\sin\theta\hat{\mathbf{i}} + \cos\theta\hat{\mathbf{j}}. \tag{19b}$$

[The fact that $\hat{\mathbf{e}}_r$ and $\hat{\mathbf{e}}_\theta$ depend only on θ, as stated in (5), is seen to follow from the absence of any r dependence in the right-hand sides of (19).] Taking $d/d\theta$, we have

$$\frac{d\hat{\mathbf{e}}_r}{d\theta} = -\sin\theta\hat{\mathbf{i}} + \cos\theta\hat{\mathbf{j}}, \tag{20a}$$

$$\frac{d\hat{\mathbf{e}}_\theta}{d\theta} = -\cos\theta\hat{\mathbf{i}} - \sin\theta\hat{\mathbf{j}}. \tag{20b}$$

Again, we emphasize that this step was simple because $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ are constant vectors. With the differentiation completed, we now return to the polar base vectors by the inverse relations

$$\hat{\mathbf{i}} = \cos\theta\hat{\mathbf{e}}_r - \sin\theta\hat{\mathbf{e}}_\theta, \tag{21a}$$

$$\hat{\mathbf{j}} = \sin\theta\hat{\mathbf{e}}_r + \cos\theta\hat{\mathbf{e}}_\theta. \tag{21b}$$

That is, putting (21) into (20),

$$\begin{aligned}\frac{d\hat{\mathbf{e}}_r}{d\theta} &= -\sin\theta(\cos\theta\hat{\mathbf{e}}_r - \sin\theta\hat{\mathbf{e}}_\theta) + \cos\theta(\sin\theta\hat{\mathbf{e}}_r + \cos\theta\hat{\mathbf{e}}_\theta)\\ &= (\sin^2\theta + \cos^2\theta)\hat{\mathbf{e}}_\theta = \hat{\mathbf{e}}_\theta,\end{aligned}$$

and

$$\begin{aligned}\frac{d\hat{\mathbf{e}}_\theta}{d\theta} &= -\cos\theta(\cos\theta\hat{\mathbf{e}}_r - \sin\theta\hat{\mathbf{e}}_\theta) - \sin\theta(\sin\theta\hat{\mathbf{e}}_r + \cos\theta\hat{\mathbf{e}}_\theta)\\ &= -(\cos^2\theta + \sin^2\theta)\hat{\mathbf{e}}_r = -\hat{\mathbf{e}}_r,\end{aligned}$$

as before.

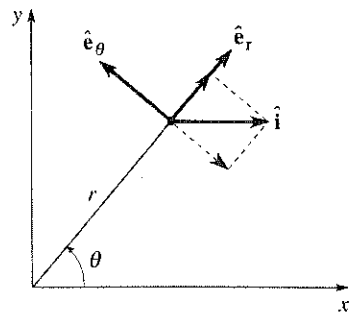


Figure 4. Expanding $\hat{\mathbf{i}}$.

In case the source of the relations (19) and (21) was not clear, let us elaborate. Consider (21a), for instance. Since $\{\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta\}$ is an orthonormal basis we can, according to the important equation (24) in Section 9.9, expand $\hat{\mathbf{i}}$ as $\hat{\mathbf{i}} = (\hat{\mathbf{i}}\cdot\hat{\mathbf{e}}_r)\hat{\mathbf{e}}_r + (\hat{\mathbf{i}}\cdot\hat{\mathbf{e}}_\theta)\hat{\mathbf{e}}_\theta = (1)(1)\cos\theta\hat{\mathbf{e}}_r + (1)(1)\cos(\theta + \frac{\pi}{2})\hat{\mathbf{e}}_\theta = \cos\theta\hat{\mathbf{e}}_r - \sin\theta\hat{\mathbf{e}}_\theta$ (see Fig. 4). Or, we could simply break $\hat{\mathbf{i}}$ into its $\hat{\mathbf{e}}_r$ and $\hat{\mathbf{e}}_\theta$ components graphically as indicated in Fig. 4. Similarly for (21b), (19a), and (19b).

We call the latter method the transform method because the "transform equations" (19) take us from $\hat{\mathbf{e}}_r$, $\hat{\mathbf{e}}_\theta$ to $\hat{\mathbf{i}}$, $\hat{\mathbf{j}}$. Then the $d/d\theta$ operation is readily carried out [in equations (20)] because $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ are constant vectors. With the results in hand,

we then use the "inverse equations" (21) to return to $\hat{\mathbf{e}}_r$, $\hat{\mathbf{e}}_\theta$. If this idea is unclear, we suggest that you review Section 5.1.

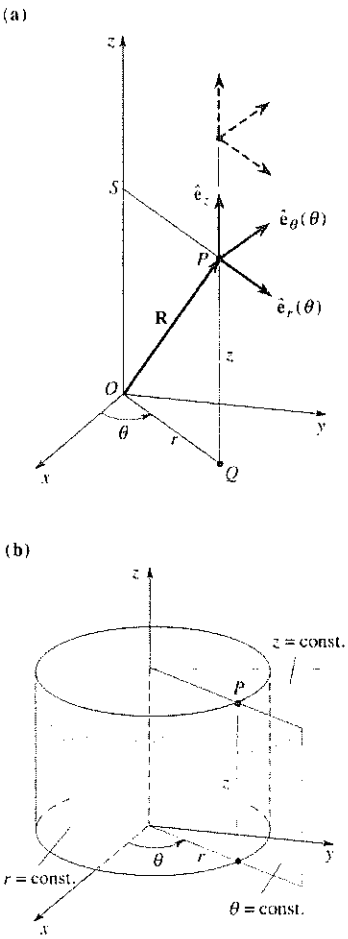


Figure 5. Cylindrical coordinates.

14.6.2. Cylindrical coordinates. The cylindrical coordinate system r, θ, z with orthonormal base vectors $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z$, shown in Fig. 5, is essentially the plane polar coordinate system that we have just discussed, with the coordinate z and the corresponding base vector $\hat{\mathbf{e}}_z$ added. Observe that the z in r, θ, z is the same as the Cartesian z variable. Similarly, $\hat{\mathbf{e}}_z$ is identical to $\hat{\mathbf{k}}$, and we would write $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{k}}$, except that $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z$ looks better.

As z is varied, with r and θ fixed, the point P (in Fig. 5a) moves along a vertical line. The $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z$ triad translates (see the dashed arrows in Fig. 5a) but does not rotate. Thus, since those vectors do not vary either in magnitude or in direction, they do not vary with z. Similarly, as r is varied, with θ and z fixed, P slides along the line SP. The triad translates but does not rotate so $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z$ do not vary with r. Finally, $\hat{\mathbf{e}}_z$ does not vary with θ but $\hat{\mathbf{e}}_r$ and $\hat{\mathbf{e}}_\theta$ do, as discussed in Section 14.6.1 above. The upshot is that $\hat{\mathbf{e}}_r = \hat{\mathbf{e}}_r(\theta)$, $\hat{\mathbf{e}}_\theta = \hat{\mathbf{e}}_\theta(\theta)$, and $\hat{\mathbf{e}}_z = \hat{\mathbf{k}}$ is a constant vector, and the derivatives $d\hat{\mathbf{e}}_r/d\theta$ and $d\hat{\mathbf{e}}_\theta/d\theta$ are as given in (18).

As we did for plane polar coordinates, let us develop expressions for the position vector, velocity, and acceleration. It might be tempting to think that the position vector is $\mathbf{R} = r\hat{\mathbf{e}}_r + \theta\hat{\mathbf{e}}_\theta + z\hat{\mathbf{e}}_z$, by analogy with $\mathbf{R} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$. No, an expression for $\mathbf{R}$ is obtained from Fig. 5a: $\mathbf{R} = \mathbf{OQ} + \mathbf{QP}$, where $\mathbf{OQ} = r\hat{\mathbf{e}}_r$ and $\mathbf{QP} = z\hat{\mathbf{e}}_z$. Thus,

$$\boxed{\mathbf{R} = r\hat{\mathbf{e}}_r + z\hat{\mathbf{e}}_z,} \tag{22}$$

which is the same as (6), but with $z\hat{\mathbf{e}}_z$ added. Remember, whereas $\hat{\mathbf{e}}_r$ and $\hat{\mathbf{e}}_\theta$ are functions of θ, $\hat{\mathbf{e}}_z$ is a *constant vector*; at every point P, $\hat{\mathbf{e}}_z$ has the same magnitude and the same direction. Thus, if $r = r(t)$, $\theta = \theta(t)$, and $z = z(t)$, where t is the time, then the velocity $\mathbf{v} = \dot{\mathbf{R}}$ and the acceleration $\mathbf{a} = \ddot{\mathbf{R}}$ are given by the formulas

$$\boxed{\mathbf{v}(t) = \dot{\mathbf{R}}(t) = \dot{r}\hat{\mathbf{e}}_r + r\dot{\theta}\hat{\mathbf{e}}_\theta + \dot{z}\hat{\mathbf{e}}_z} \tag{23}$$

and

$$\boxed{\mathbf{a}(t) = \ddot{\mathbf{R}}(t) = (\ddot{r} - r\dot{\theta}^2)\hat{\mathbf{e}}_r + (r\ddot{\theta} + 2\dot{r}\dot{\theta})\hat{\mathbf{e}}_\theta + \ddot{z}\hat{\mathbf{e}}_z.} \tag{24}$$

That is, we merely append $\dot{z}\hat{\mathbf{e}}_z$ to (12) and $\ddot{z}\hat{\mathbf{e}}_z$ to (17).

EXAMPLE 1. Let a particle move according to

$$r = \text{constant}, \qquad \theta = \omega t, \qquad z = vt,$$

where ω and v are constants, and $0 \le t < \infty$. Evidently, the path of motion is helical. From (23) and (24),

$$\mathbf{v}(t) = r\omega\hat{\mathbf{e}}_\theta + v\hat{\mathbf{e}}_z, \tag{25a}$$

$$\mathbf{a}(t) = -r\omega^2\hat{\mathbf{e}}_r. \tag{25b}$$

Observe that $\mathbf{v}(t)$ is *not* constant, even though its components $r\omega$ and v are constant, because $\hat{\mathbf{e}}_\theta = \hat{\mathbf{e}}_\theta(\theta(t))$ varies with t. [Indeed, if $\mathbf{v}(t)$ *were* constant, then $\mathbf{a}(t) = \dot{\mathbf{v}}(t)$ would be zero, not $-r\omega^2\hat{\mathbf{e}}_r$.] ∎

Under what circumstances is one coordinate system to be favored over another? One important guideline is the shape of the region involved. For instance, cylindrical coordinates are most convenient for studying fluid flow in a cylindrical pipe, and Cartesian coordinates are most convenient for studying the electric field in a rectangular prism. That is, *a particular coordinate system is convenient if the region is bounded by constant-coordinate surfaces.* Thus, Cartesian coordinates are convenient if the region is bounded by constant-x, constant-y, and constant-z planes, that is, if it is a rectangular prism. How about cylindrical coordinates? Constant-r, -θ, and -z surfaces are shown in Fig. 5b, where the θ = constant surface is known as a *meridional plane*. Thus, some representative regions for which cylindrical coordinates are especially convenient are as follows (sketch them).

$$\begin{aligned}
&\text{Soup can}: \quad 0 \le r \le r_1, \quad 0 \le \theta < 2\pi, \quad 0 \le z \le z_1.\\
&\text{Coaxial cable}: \quad r_1 \le r \le r_2, \quad 0 \le \theta < 2\pi, \quad 0 \le z \le z_1.\\
&\text{Wedge}: \quad 0 \le r \le r_1, \quad 0 \le \theta < \theta_1, \quad 0 \le z \le z_1.
\end{aligned}$$

14.6.3. Spherical coordinates. Finally, we present the spherical coordinates ρ, ϕ, θ with their respective orthonormal unit base vectors $\hat{\mathbf{e}}_\rho, \hat{\mathbf{e}}_\phi, \hat{\mathbf{e}}_\theta$ shown in Fig. 6. Some authors use the notation r, ϕ, θ, but we prefer to emphasize the difference between the spherical coordinate ρ and the cylindrical coordinate r by using different letters. What difference? Well, $\rho = \sqrt{x^2+y^2+z^2}$ is the distance from the origin to P, whereas $r = \sqrt{x^2+y^2}$ is not (unless $z = 0$); it is the perpendicular distance from P to the z axis. The angle θ, however, *is* the same in the two coordinate systems, cylindrical and spherical.

Observe from Fig. 6b that the ρ = constant surface through P is the *spherical* surface of radius ρ centered at the origin, the ϕ = constant surface is *conical*, and the θ = constant surface is a *meridional plane*.

In general, ρ, ϕ, and θ may be positive, negative, or zero, but we choose, in this book, always to have $\rho \ge 0$ so that ρ is the *distance* from the origin to P. In addition, we normally limit ϕ and θ so that $0 \le \phi \le \pi$ and $0 \le \theta < 2\pi$ since those ranges permit us to reach any point in 3-space in a unique manner.

From Fig. 6a we can see that ρ, ϕ, θ are related to the reference x, y, z coordinates according to

$$\begin{aligned}
x &= \rho \sin\phi \cos\theta\\
y &= \rho \sin\phi \sin\theta \qquad (26a)\\
z &= \rho \cos\phi,
\end{aligned}$$

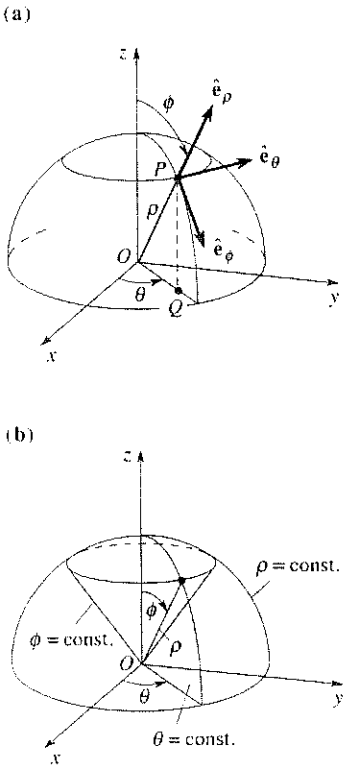


Figure 6. Spherical coordinates.

or, solving (26a) for ρ, ϕ, and θ,

$$\begin{aligned} \rho &= \sqrt{x^2+y^2+z^2}, \\ \phi &= \cos^{-1}\frac{z}{\sqrt{x^2+y^2+z^2}}, \\ \theta &= \tan^{-1}\frac{y}{x}, \end{aligned} \tag{26b}$$

where θ is undefined on the z axis (i.e., for $x = y = 0$) and ϕ is undefined at the origin. That is, on the positive z axis ϕ is zero and on the negative z axis ϕ is π so at the origin ϕ is discontinuous and undefined.

EXAMPLE 2. Given the point $(x, y, z) = (2, -1, 3)$, find ρ, ϕ, θ (such that $\rho \geq 0$, $0 \leq \phi \leq \pi$, and $0 \leq \theta < 2\pi$ as noted above). From (26b), $\rho = \sqrt{2^2 + (-1)^2 + 3^2} = \sqrt{14}$, $\phi = \cos^{-1}(3/\sqrt{14}) = 36.70°$, and $\theta = \tan^{-1}(-1/2) = 153.4°$ *and* $333.4°$ (within the interval $0 \leq \theta < 360°$). To choose between the latter two values, put the computed ρ, ϕ, θ values back into (26a) and see if they give the original x, y, z values: $\rho = \sqrt{14}$, $\phi = 36.70°$, and $\theta = 333.4°$ gives $(x, y, z) = (2, -1, 3)$ but $\rho = \sqrt{14}$, $\phi = 36.70°$, and $\theta = 153.4°$ does not. Thus, $\rho = \sqrt{14}$, $\phi = 36.70°$, and $\theta = 333.4°$. ∎

As emphasized earlier in this section, it is essential, in working with this coordinate system, to know the "space derivatives" of the base vectors, namely, $\partial/\partial\rho$, $\partial/\partial\phi$, $\partial/\partial\theta$ of each of the three base vectors. From Fig. 6a we see that if we hold ϕ and θ fixed and vary ρ, the $\hat{\mathbf{e}}_\rho, \hat{\mathbf{e}}_\phi, \hat{\mathbf{e}}_\theta$, triad translates parallel to itself as P moves along the radial line. Thus, $\hat{\mathbf{e}}_\rho, \hat{\mathbf{e}}_\phi, \hat{\mathbf{e}}_\theta$ do not vary with ρ so $\partial\hat{\mathbf{e}}_\rho/\partial\rho = \partial\hat{\mathbf{e}}_\phi/\partial\rho = \partial\hat{\mathbf{e}}_\theta/\partial\rho = \mathbf{0}$. Next, imagine varying ϕ, with ρ and θ fixed. Then (Fig. 6a) $\hat{\mathbf{e}}_\theta$ moves parallel to itself, and hence does not change, whereas $\hat{\mathbf{e}}_\rho$ and $\hat{\mathbf{e}}_\phi$ both rotate. Finally, if we vary θ, with ρ and ϕ fixed, we see that all three base vectors rotate. The upshot is that the dependence of the base vectors on the coordinates is as follows:

$$\hat{\mathbf{e}}_\rho = \hat{\mathbf{e}}_\rho(\phi, \theta), \qquad \hat{\mathbf{e}}_\phi = \hat{\mathbf{e}}_\phi(\phi, \theta), \qquad \hat{\mathbf{e}}_\theta = \hat{\mathbf{e}}_\theta(\theta). \tag{27}$$

Thus, there are five nonzero derivatives to work out. To do so, we could use either of the two methods presented above in Section 14.6.1, the difference quotient method or the transform method. However, the calculations are harder than for the case of plane polar coordinates so we merely state the results:

$$\boxed{\begin{aligned} &\frac{\partial\hat{\mathbf{e}}_\rho}{\partial\rho} = \mathbf{0}, && \frac{\partial\hat{\mathbf{e}}_\rho}{\partial\phi} = \hat{\mathbf{e}}_\phi, && \frac{\partial\hat{\mathbf{e}}_\rho}{\partial\theta} = \sin\phi\,\hat{\mathbf{e}}_\theta, \\ &\frac{\partial\hat{\mathbf{e}}_\phi}{\partial\rho} = \mathbf{0}, && \frac{\partial\hat{\mathbf{e}}_\phi}{\partial\phi} = -\hat{\mathbf{e}}_\rho, && \frac{\partial\hat{\mathbf{e}}_\phi}{\partial\theta} = \cos\phi\,\hat{\mathbf{e}}_\theta, \\ &\frac{\partial\hat{\mathbf{e}}_\theta}{\partial\rho} = \mathbf{0}, && \frac{\partial\hat{\mathbf{e}}_\theta}{\partial\phi} = \mathbf{0}, && \frac{\partial\hat{\mathbf{e}}_\theta}{\partial\theta} = -\sin\phi\,\hat{\mathbf{e}}_\rho - \cos\phi\,\hat{\mathbf{e}}_\phi, \end{aligned}} \tag{28}$$

and defer their derivation to Section 14.6.4, where we use a different, and simpler, approach.

Let us illustrate the use of (28) by deriving spherical-coordinate expressions for the velocity and acceleration, expressions analogous to those obtained for plane polar and cylindrical coordinates. We begin with the position vector $\mathbf{R} = \mathbf{OP} = \|\mathbf{OP}\|\,\hat{\mathbf{e}}_\rho$, that is,

$$\boxed{\mathbf{R} = \rho\hat{\mathbf{e}}_\rho.} \tag{29}$$

Since $\hat{\mathbf{e}}_\rho = \hat{\mathbf{e}}_\rho(\phi(t), \theta(t))$, differentiation with respect to t gives

$$\begin{aligned}\mathbf{v} = \dot{\mathbf{R}} &= \dot{\rho}\,\hat{\mathbf{e}}_\rho + \rho\left(\frac{\partial\hat{\mathbf{e}}_\rho}{\partial\phi}\frac{d\phi}{dt} + \frac{\partial\hat{\mathbf{e}}_\rho}{\partial\theta}\frac{d\theta}{dt}\right)\\ &= \dot{\rho}\,\hat{\mathbf{e}}_\rho + \rho(\dot{\phi}\,\hat{\mathbf{e}}_\phi + \dot{\theta}\sin\phi\,\hat{\mathbf{e}}_\theta)\end{aligned}$$

or

$$\boxed{\mathbf{v}(t) = \dot{\rho}\,\hat{\mathbf{e}}_\rho + \rho\dot{\phi}\,\hat{\mathbf{e}}_\phi + \rho\dot{\theta}\sin\phi\,\hat{\mathbf{e}}_\theta.} \tag{30}$$

One more differentiataion (Exercise 7) gives

$$\boxed{\begin{aligned}\mathbf{a}(t) = {} & (\ddot{\rho} - \rho\dot{\phi}^2 - \rho\dot{\theta}^2\sin^2\phi)\hat{\mathbf{e}}_\rho + (\rho\ddot{\phi} + 2\dot{\rho}\dot{\phi} - \rho\dot{\theta}^2\sin\phi\cos\phi)\hat{\mathbf{e}}_\phi\\ & + (\rho\ddot{\theta}\sin\phi + 2\dot{\rho}\dot{\phi}\sin\phi + 2\rho\dot{\theta}\dot{\phi}\cos\phi)\hat{\mathbf{e}}_\theta\end{aligned}} \tag{31}$$

for the acceleration.

14.6.4. Omega method. Here, we will develop an interesting alternative method for deriving the space derivatives of base vectors, and use that method to derive equations (28).

Consider a rigid (i.e., nondeformable) body $\mathcal{B}$ undergoing an arbitrary motion through 3-space, and let $\mathbf{A}$ be any fixed vector within $\mathcal{B}$ (Fig. 7). That is, $\mathbf{A}$ is a vector from one material point in $\mathcal{B}$ to another so $\|\mathbf{A}\|$ is constant with time because $\mathcal{B}$ is rigid. That is not to say that $\mathbf{A}$ is a constant vector, for a vector is comprised of a magnitude and a direction, and the direction of $\mathbf{A}$ varies with the time t, in general, as $\mathcal{B}$ tumbles through space. Thus, $\mathbf{A} = \mathbf{A}(t)$. It might help to think of $\mathcal{B}$ as a potato, say, and $\mathbf{A}$ as a needle pushed into the potato.

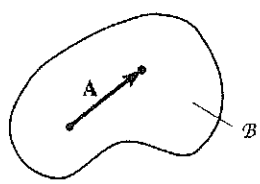


Figure 7. Fixed vector in $\mathcal{B}$.

Since $\|\mathbf{A}\|$ is constant ,

$$\|\mathbf{A}\|^2 = \mathbf{A}\cdot\mathbf{A} = \text{constant} \tag{32}$$

as well. Differentiating (32) with respect to t gives [recall (7c) in Section 14.5] $\dot{\mathbf{A}}\cdot\mathbf{A} + \mathbf{A}\cdot\dot{\mathbf{A}} = 2\mathbf{A}\cdot\dot{\mathbf{A}} = 0$ so

$$\mathbf{A}\cdot\dot{\mathbf{A}} = 0. \tag{33}$$

Equation (33) makes sense because if $\dot{\mathbf{A}}$ were *not* orthogonal to $\mathbf{A}$, then it would

have a nonzero component *along* $\mathbf{A}$, which component would correspond to a nonzero rate of stretching or shrinking of $\mathbf{A}$, in violation of the assumption that $\mathbf{A}$ is of fixed length.

It follows from (33) (see Exercise 12) that there exists a vector $\mathbf{\Omega}_1$ such that

$$\dot{\mathbf{A}} = \mathbf{\Omega}_1 \times \mathbf{A}. \tag{34}$$

Similarly, if $\mathbf{B}$ is any other fixed vector in $\mathcal{B}$ then there exists a vector $\mathbf{\Omega}_2$ such that

$$\dot{\mathbf{B}} = \mathbf{\Omega}_2 \times \mathbf{B}. \tag{35}$$

Since $\mathcal{B}$ is rigid it follows that

$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\| \, \|\mathbf{B}\| \cos \alpha = \text{constant} \tag{36}$$

because $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ are constant, as is the angle α between them. Differentiating (36) with respect to t gives

$$\dot{\mathbf{A}} \cdot \mathbf{B} + \mathbf{A} \cdot \dot{\mathbf{B}} = 0 \tag{37}$$

or, using (34) and (35),

$$\begin{aligned} (\mathbf{\Omega}_1 \times \mathbf{A}) \cdot \mathbf{B} + \mathbf{A} \cdot (\mathbf{\Omega}_2 \times \mathbf{B}) &= 0, \\ \mathbf{\Omega}_1 \times \mathbf{A} \cdot \mathbf{B} + \mathbf{A} \times \mathbf{\Omega}_2 \cdot \mathbf{B} &= 0, \qquad \text{[by (7) in Section 14.4]} \end{aligned}$$

so

$$(\mathbf{\Omega}_1 - \mathbf{\Omega}_2) \times \mathbf{A} \cdot \mathbf{B} = 0. \tag{38}$$

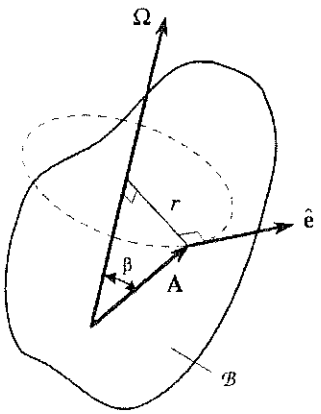


Figure 8. The angular velocity $\mathbf{\Omega}$.

Since $\mathbf{B}$ is arbitrary, it follows from (38) (Exercise 13) that $(\mathbf{\Omega}_1 - \mathbf{\Omega}_2) \times \mathbf{A} = \mathbf{0}$. Finally, since $\mathbf{A}$ is arbitrary it follows from the latter (Exercise 14) that $\mathbf{\Omega}_1 - \mathbf{\Omega}_2 = \mathbf{0}$. Thus, $\mathbf{\Omega}_1 = \mathbf{\Omega}_2$, and we can say that for any fixed vector $\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots$ in $\mathcal{B}$ we have $\dot{\mathbf{A}} = \mathbf{\Omega} \times \mathbf{A}$, $\dot{\mathbf{B}} = \mathbf{\Omega} \times \mathbf{B}$, $\dot{\mathbf{C}} = \mathbf{\Omega} \times \mathbf{C}$, and so on. Or, stated more concisely,

$$\boxed{\dot{\mathbf{A}} = \mathbf{\Omega} \times \mathbf{A},} \tag{39}$$

where $\mathbf{A}$ is *any* fixed vector in $\mathcal{B}$.

To understand what $\mathbf{\Omega}$ is, let us write out (39) as

$$\dot{\mathbf{A}} = (\|\mathbf{\Omega}\|)\,(\|\mathbf{A}\| \sin \beta)\,\hat{\mathbf{e}} = \Omega\, r\, \hat{\mathbf{e}}, \tag{40}$$

where β, r, and $\hat{\mathbf{e}}$ are shown in Fig. 8; $\hat{\mathbf{e}}$ is tangent to the circle of radius r (shown as dashed). But (40) is the familiar result, encountered in physics, that the velocity of a point moving in a circle of radius r with angular velocity Ω is Ωr in the tangential direction. That is, Ω is the instantaneous angular velocity of $\mathcal{B}$. Using the right-hand rule to give it a direction, Ω is therefore the **angular velocity vector** of $\mathcal{B}$; $\mathbf{\Omega}$ need not be constant with time but Fig. 8 and equation (39) hold at each instant.

Before using (39) to derive (28), let us consider the simpler case of cylindrical coordinates. Imagine the base vectors $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z$ (Fig. 5a) as welded steel rods

within a body $\mathcal{B}$ (which could be built up around the rods using modeling clay), and imagine that body to be in motion according to $r = r(t)$, $\theta = \theta(t)$, and $z = z(t)$, where $r(t)$, $\theta(t)$, and $z(t)$ are arbitrary. Then what is the angular velocity $\mathbf{\Omega}$ of the body? If we vary r only (holding θ and z fixed), then the body translates, with no angular velocity. Similarly if we vary z (holding r and θ fixed). But if we vary θ, then $\mathcal{B}$ undergoes an angular velocity $\dot{\theta}(t)$ or, using the right-hand rule to make it into a vector, $\dot{\theta}(t)\hat{\mathbf{e}}_z$. Since angular velocity is a vector, we can get the total angular velocity by adding these three contributions: $\mathbf{\Omega} = \mathbf{0} + \mathbf{0} + \dot{\theta}\,\hat{\mathbf{e}}_z$ so

$$\mathbf{\Omega} = \dot{\theta}\,\hat{\mathbf{e}}_z. \tag{41}$$

To determine the space derivatives of $\hat{\mathbf{e}}_r$, say, let $\mathbf{A}$ be $\hat{\mathbf{e}}_r$ (which does have fixed length) in (39). Then (39) gives

$$\frac{d\hat{\mathbf{e}}_r}{dt} = \mathbf{\Omega} \times \hat{\mathbf{e}}_r = \dot{\theta}\,\hat{\mathbf{e}}_z \times \hat{\mathbf{e}}_r = \dot{\theta}\,\hat{\mathbf{e}}_\theta. \tag{42}$$

On the other hand, we can use chain differentiation to write

$$\begin{aligned}\frac{d}{dt}\,\hat{\mathbf{e}}_r(r(t),\theta(t),z(t)) &= \frac{\partial\hat{\mathbf{e}}_r}{\partial r}\frac{dr}{dt} + \frac{\partial\hat{\mathbf{e}}_r}{\partial\theta}\frac{d\theta}{dt} + \frac{\partial\hat{\mathbf{e}}_r}{\partial z}\frac{dz}{dt}\\ &= \dot{r}\frac{\partial\hat{\mathbf{e}}_r}{\partial r} + \dot{\theta}\frac{\partial\hat{\mathbf{e}}_r}{\partial\theta} + \dot{z}\frac{\partial\hat{\mathbf{e}}_r}{\partial z}.\end{aligned} \tag{43}$$

Next, we equate the right-hand sides of (42) and (43) and use the fact that $r(t), \theta(t), z(t)$ (and hence $\dot{r}, \dot{\theta}, \dot{z}$) are arbitrary:

$$\dot{r}\mathbf{0} + \dot{\theta}\hat{\mathbf{e}}_\theta + \dot{z}\mathbf{0} = \dot{r}\frac{\partial\hat{\mathbf{e}}_r}{\partial r} + \dot{\theta}\frac{\partial\hat{\mathbf{e}}_r}{\partial\theta} + \dot{z}\frac{\partial\hat{\mathbf{e}}_r}{\partial z} \tag{44}$$

so

$$\frac{\partial\hat{\mathbf{e}}_r}{\partial r} = \mathbf{0}, \qquad \frac{\partial\hat{\mathbf{e}}_r}{\partial\theta} = \hat{\mathbf{e}}_\theta, \qquad \frac{\partial\hat{\mathbf{e}}_r}{\partial z} = \mathbf{0}. \tag{45}$$

That is, since $\dot{r}(t), \dot{\theta}(t), \dot{z}(t)$ are arbitrary we can set $\dot{r} = 1, \dot{\theta} = 0, \dot{z} = 0$ in (44), and learn that $\partial\hat{\mathbf{e}}_r/\partial r = \mathbf{0}$. Then we can set $\dot{r} = 0, \dot{\theta} = 1, \dot{z} = 0$, and so on. [Of course we didn't really need to include the $\partial\hat{\mathbf{e}}_r/\partial r$ and $\partial\hat{\mathbf{e}}_r/\partial z$ terms in (43) because we knew, from our discussion in Section 14.6.1 and 14.6.2 that $\hat{\mathbf{e}}_r$ is a function of θ alone, but we included those terms to show that we don't *need* to know that information in advance.]

Similarly, if we use $\mathbf{A} = \hat{\mathbf{e}}_\theta$ we obtain $\partial\hat{\mathbf{e}}_\theta/\partial r = \mathbf{0}$, $\partial\hat{\mathbf{e}}_\theta/\partial\theta = -\hat{\mathbf{e}}_r$, and $\partial\hat{\mathbf{e}}_\theta/\partial z = \mathbf{0}$, and if we use $\mathbf{A} = \hat{\mathbf{e}}_z$ we obtain $\partial\hat{\mathbf{e}}_z/\partial r = \partial\hat{\mathbf{e}}_z/\partial\theta = \partial\hat{\mathbf{e}}_z/\partial z = \mathbf{0}$.

Understand that we *imagine* the representative point P as undergoing a motion $r(t)$, $\theta(t)$, $z(t)$ simply in order to be able to use this method, which we call the **omega method** because of the prominent role played by $\mathbf{\Omega}$.

Turning to spherical coordinates, what is $\mathbf{\Omega}$ in this case? Referring to Fig. 6a, we see that varying ρ (holding ϕ and θ fixed) results in no angular velocity of the base vector triad, only translation; varying ϕ (holding ρ and θ fixed) results in an

angular velocity $\dot{\phi}(t)\hat{\mathbf{e}}_\theta$; and varying θ (holding ρ and ϕ fixed) results in an angular velocity about the z axis, $\dot{\theta}(t)\hat{\mathbf{k}}$. Thus,

$$\begin{aligned}\mathbf{\Omega} &= \dot{\phi}\hat{\mathbf{e}}_\theta + \dot{\theta}\hat{\mathbf{k}} \\ &= \dot{\phi}\hat{\mathbf{e}}_\theta + \dot{\theta}\left[(\hat{\mathbf{k}}\cdot\hat{\mathbf{e}}_\rho)\hat{\mathbf{e}}_\rho + (\hat{\mathbf{k}}\cdot\hat{\mathbf{e}}_\phi)\hat{\mathbf{e}}_\phi + (\hat{\mathbf{k}}\cdot\hat{\mathbf{e}}_\theta)\hat{\mathbf{e}}_\theta\right] \\ &= \dot{\phi}\hat{\mathbf{e}}_\theta + \dot{\theta}\left[\cos\phi\hat{\mathbf{e}}_\rho + \cos\left(\frac{\pi}{2}+\phi\right)\hat{\mathbf{e}}_\phi + 0\hat{\mathbf{e}}_\theta\right] \\ &= \dot{\phi}\hat{\mathbf{e}}_\theta + \dot{\theta}\left(\cos\phi\hat{\mathbf{e}}_\rho - \sin\phi\hat{\mathbf{e}}_\phi\right). \end{aligned} \tag{46}$$

Let us determine the space derivatives of $\hat{\mathbf{e}}_\rho$, say. Let $\mathbf{A} = \hat{\mathbf{e}}_\rho$ in (39). Then, on the one hand

$$\begin{aligned}\frac{d\hat{\mathbf{e}}_\rho}{dt} = \mathbf{\Omega}\times\hat{\mathbf{e}}_\rho &= \left[\dot{\phi}\hat{\mathbf{e}}_\theta + \dot{\theta}(\cos\phi\hat{\mathbf{e}}_\rho - \sin\phi\hat{\mathbf{e}}_\phi)\right]\times\hat{\mathbf{e}}_\rho \\ &= \dot{\phi}\hat{\mathbf{e}}_\phi + \dot{\theta}\sin\phi\hat{\mathbf{e}}_\theta, \end{aligned} \tag{47}$$

and on the other hand

$$\frac{d}{dt}\hat{\mathbf{e}}_\rho(\rho(t),\phi(t),\theta(t)) = \frac{\partial\hat{\mathbf{e}}_\rho}{\partial\rho}\dot{\rho} + \frac{\partial\hat{\mathbf{e}}_\rho}{\partial\phi}\dot{\phi} + \frac{\partial\hat{\mathbf{e}}_\rho}{\partial\theta}\dot{\theta}, \tag{48}$$

and comparison of (47) and (48) gives

$$\frac{\partial\hat{\mathbf{e}}_\rho}{\partial\rho} = \mathbf{0}, \qquad \frac{\partial\hat{\mathbf{e}}_\rho}{\partial\phi} = \hat{\mathbf{e}}_\phi, \qquad \frac{\partial\hat{\mathbf{e}}_\rho}{\partial\theta} = \sin\phi\hat{\mathbf{e}}_\theta,$$

in agreement with (28). Then, letting $\mathbf{A}$ be $\hat{\mathbf{e}}_\phi$ and $\hat{\mathbf{e}}_\theta$, in turn, we can obtain the remaining space derivatives, which steps we leave for the exercises.

Observe that $\mathbf{\Omega}$ is determined by the coordinate system: for plane polar and cylindrical coordinates $\mathbf{\Omega} = \dot{\theta}\hat{\mathbf{k}}$, and for spherical coordinates $\mathbf{\Omega} = \dot{\phi}\hat{\mathbf{e}}_\theta + \dot{\theta}\hat{\mathbf{k}}$. There is no need to use the omega method for Cartesian coordinates since $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$ are constant vectors so $\partial\hat{\mathbf{i}}/\partial x = \partial\hat{\mathbf{i}}/\partial y = \partial\hat{\mathbf{i}}/\partial z = \partial\hat{\mathbf{j}}/\partial x = \cdots = \mathbf{0}$, but it is worth mentioning that for that coordinate system $\mathbf{\Omega} = \mathbf{0}$.

Closure. In this section we derive expressions for the velocity and acceleration vectors in plane polar, cylindrical, and spherical coordinates, but that discussion merely serve as a common application thread and is not, in itself, the focal point. The focal point is the plane polar, cylindrical, and spherical coordinate systems: the definition of the coordinates by means of a sketch (e.g., Fig. 5), the relationship between those coordinates and the reference Cartesian coordinates (e.g., $x = r\cos\theta$, $y = r\sin\theta$, and $z = z$, the Cartesian coordinate z being identical to the cylindrical coordinate z), the expression for the position vector (e.g., $\mathbf{R} = r\hat{\mathbf{e}}_r + z\hat{\mathbf{e}}_z$ in cylindrical coordinates), and the space derivatives of the base vectors [e.g., equations (28) for spherical coordinates]. In fact, most of our attention is devoted to the derivation of the derivatives of the base vectors, and three distinct methods are

presented: the difference quotient method, the transform method, and the omega method. These derivatives will show up extensively in Chapter 16, but you don't need to derive them each time they arise; that has already been done here, and you can use the results, listed in (18) for cylindrical coordinates and (28) for spherical coordinates.

EXERCISES 14.6

1. Determine r, θ, z corresponding to each given point $P = (x, y, z)$. Give θ both in radians ($0 \le \theta < 2\pi$) and in degrees ($0° \le \theta < 360°$).

(a) $(2, 2, 3)$ (b) $(2, -1, 0)$
(c) $(-3, 2, 1)$ (d) $(0, 0, 5)$
(e) $(6, -1, 3)$ (f) $(-1, -2, 6)$
(g) $(-1, 5, 1)$ (h) $(6, 0, 4)$

2. Determine ρ, ϕ, θ corresponding to each given point $P = (x, y, z)$ such that $\rho \ge 0$, $0 \le \phi \le \pi$, $0 \le \theta < 2\pi$. Give ϕ and θ both in radians and in degrees.

(a) $(2, 0, 0)$ (b) $(0, 3, 0)$
(c) $(0, 5, 0)$ (d) $(1, 2, 3)$
(e) $(6, -3, 1)$ (f) $(0, -5, 0)$
(g) $(2, 3, 5)$ (h) $(2, 3, -5)$
(i) $(-2, 3, 5)$ (j) $(2, -3, 5)$
(k) $(-1, -2, 1)$ (l) $(4, -2, -3)$

3. (a) Derive the expressions

$$\begin{aligned} \hat{\mathbf{e}}_\rho &= \sin\phi(\cos\theta\hat{\mathbf{i}} + \sin\theta\hat{\mathbf{j}}) + \cos\phi\hat{\mathbf{k}}, \\ \hat{\mathbf{e}}_\phi &= \cos\phi(\cos\theta\hat{\mathbf{i}} + \sin\theta\hat{\mathbf{j}}) - \sin\phi\hat{\mathbf{k}}, \\ \hat{\mathbf{e}}_\theta &= -\sin\theta\hat{\mathbf{i}} + \cos\theta\hat{\mathbf{j}}. \end{aligned} \tag{3.1}$$

(b) Derive the "reverse" expressions

$$\begin{aligned} \hat{\mathbf{i}} &= \sin\phi\cos\theta\hat{\mathbf{e}}_\rho + \cos\phi\cos\theta\hat{\mathbf{e}}_\phi - \sin\theta\hat{\mathbf{e}}_\theta, \\ \hat{\mathbf{j}} &= \sin\phi\sin\theta\hat{\mathbf{e}}_\rho + \cos\phi\sin\theta\hat{\mathbf{e}}_\phi + \cos\theta\hat{\mathbf{e}}_\theta, \\ \hat{\mathbf{k}} &= \cos\phi\hat{\mathbf{e}}_\rho - \sin\phi\hat{\mathbf{e}}_\phi, \end{aligned} \tag{3.2}$$

either directly or by solving (3.1) for $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$. If you choose the latter, you may use computer software.

4. Compute $d\mathbf{A}/dt$ in each case.

(a) $\mathbf{A} = \sin\theta\hat{\mathbf{e}}_r - r\theta^2\hat{\mathbf{e}}_\theta + r\hat{\mathbf{e}}_z$; $r = t^2$, $\theta = 3t$
(b) $\mathbf{A} = t\hat{\mathbf{e}}_r$; $r = 4$, $\theta = 2t$
(c) $\mathbf{A} = r\hat{\mathbf{e}}_r$; $r = 6 + \sin t$, $\theta = \cos t$
(d) $\mathbf{A} = \rho^2\hat{\mathbf{e}}_\rho$; $\rho = t$, $\phi = 2t$, $\theta = 3t$
(e) $\mathbf{A} = \hat{\mathbf{e}}_\phi$; $\rho = 1 + t$, $\phi = t^2$, $\theta = \sin 2t$
(f) $\mathbf{A} = \hat{\mathbf{e}}_\rho + t\hat{\mathbf{e}}_\theta$; $\rho = 1$, $\phi = t$, $\theta = t^2$.

5. A particle moves in 3-space according to the given functions $x(t), y(t), z(t)$. Determine its velocity $\mathbf{v}(t)$ and acceleration $\mathbf{a}(t)$ in terms of $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z$, and t, for $t \ge 0$.

(a) $x = t^2$, $y = 2$, $z = 0$
(b) $x = 2t$, $y = 5t$, $z = 3t$
(c) $x = t^2 - t$, $y = t^2$, $z = -3t$
(d) $x = 3\cos t$, $y = 3\sin t$, $z = 6t$
(e) $x = 1$, $y = 4t$, $z = 2$
(f) $x = \sin t$, $y = \cos t$, $z = \sin t$

6. (a)–(f) Same as Exercise 5, but find $\mathbf{v}$ and $\mathbf{a}$ in terms of $\hat{\mathbf{e}}_\rho, \hat{\mathbf{e}}_\phi, \hat{\mathbf{e}}_\theta$, and t, for $t \ge 0$. You may use any formulas in Exercise 3.

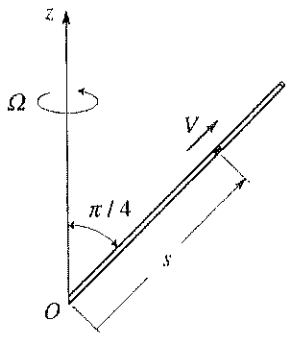
7. Evaluate each of the following by the difference quotient method, and show that your result agrees with that given in (28).

(a) $\partial\hat{\mathbf{e}}_\rho/\partial\phi$ (b) $\partial\hat{\mathbf{e}}_\rho/\partial\theta$
(c) $\partial\hat{\mathbf{e}}_\phi/\partial\phi$ (d) $\partial\hat{\mathbf{e}}_\phi/\partial\theta$
(e) $\partial\hat{\mathbf{e}}_\theta/\partial\theta$

8. Derive (31) from (30), with the help of (28).

9. Beginning at O at time $t = 0$, a bead moves at a constant speed V through a straight hollow tube, which rotates at constant angular velocity Ω about the z axis so as to sweep out a cone of half angle $\pi/4$, as shown. Compute the velocity $\mathbf{v}$, and the acceleration $\mathbf{a}$ of the particle, in terms of s, V, and Ω, using

(a) Cartesian coordinates
(b) cylindrical coordinates
(c) spherical coordinates



NOTE: Let us orient the x, y, z coordinate frame so that the tube lies in the x, z plane at $t = 0$.

10. Beginning at the "north pole" ($\phi = 0$), a particle spirals down the surface of a sphere of radius ρ such that the spherical coordinates ϕ and θ vary as $\phi = \Omega_1 t$, $\theta = \Omega_2 t$. Find the magnitude of the velocity and acceleration vectors at the instant when the particle crosses the "equator."

11. Observe, in (28), that none of the derivatives of $\hat{\mathbf{e}}_\rho$ contain an $\hat{\mathbf{e}}_\rho$ component, that none of the derivatives of $\hat{\mathbf{e}}_\phi$ contain an $\hat{\mathbf{e}}_\phi$ component, and that none of the derivatives of $\hat{\mathbf{e}}_\theta$ contain an $\hat{\mathbf{e}}_\theta$ component. The same pattern holds in (18). Explain why this pattern is more than coincidence.

12. Given vectors $\mathbf{u}$ and $\mathbf{v}$ in 3-space, such that $\mathbf{u} \cdot \mathbf{v} = 0$, where $\|\mathbf{u}\| \neq 0$, show that there exists a vector $\mathbf{\Omega}$ such that $\mathbf{v} = \mathbf{\Omega} \times \mathbf{u}$. HINT: Write $\mathbf{v} = \mathbf{\Omega} \times \mathbf{u}$ as three scalar equations on the components $\Omega_1, \Omega_2, \Omega_3$ of $\mathbf{\Omega}$, and solve.

13. Show that if $\mathbf{u} \cdot \mathbf{v} = 0$ for all vectors $\mathbf{v}$, then it must be true that $\mathbf{u} = \mathbf{0}$.

14. Show that if $\mathbf{u} \times \mathbf{v} = \mathbf{0}$ for all vectors $\mathbf{v}$, then it must be true that $\mathbf{u} = \mathbf{0}$.

15. Use the omega method to evaluate the three space derivatives of the spherical coordinate base vector

(a) $\hat{\mathbf{e}}_\phi$ (b) $\hat{\mathbf{e}}_\theta$

Chapter 14 Review

In this review let us limit attention to results that we urge you to know without needing to look them up.

Section 14.2

$$\|\mathbf{u} \times \mathbf{v}\| = \text{area of } \mathbf{u}, \mathbf{v} \text{ parallelogram}$$

Section 14.3 If $\mathbf{u} = u_1\hat{\mathbf{i}} + u_2\hat{\mathbf{j}} + u_3\hat{\mathbf{k}}$ and $\mathbf{v} = v_1\hat{\mathbf{i}} + v_2\hat{\mathbf{j}} + v_3\hat{\mathbf{k}}$, then

$$\mathbf{u} \times \mathbf{v} = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix}.$$

Section 14.4

$$|\mathbf{u} \cdot \mathbf{v} \times \mathbf{w}| = \text{volume of } \mathbf{u}, \mathbf{v}, \mathbf{w} \text{ parallelepiped}$$

$$\mathbf{u} \cdot \mathbf{v} \times \mathbf{w} = \begin{vmatrix} u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \\ w_1 & w_2 & w_3 \end{vmatrix}$$

$$\mathbf{u} \cdot \mathbf{v} \times \mathbf{w} = \mathbf{u} \times \mathbf{v} \cdot \mathbf{w}$$

Section 14.6

Polar coordinates: Know Fig. 1.

Coordinates r, θ and base vectors $\hat{\mathbf{e}}_r(\theta), \hat{\mathbf{e}}_\theta(\theta)$.

$$x = r\cos\theta$$
$$y = r\sin\theta$$

Position vector is $\mathbf{R} = r\hat{\mathbf{e}}_r$.

$$\frac{d\hat{\mathbf{e}}_r}{d\theta} = \hat{\mathbf{e}}_\theta, \qquad \frac{d\hat{\mathbf{e}}_\theta}{d\theta} = -\hat{\mathbf{e}}_r$$

Cylindrical coordinates: Know Fig. 5a.

Coordinates r, θ, z and base vectors $\hat{\mathbf{e}}_r(\theta), \hat{\mathbf{e}}_\theta(\theta), \hat{\mathbf{e}}_z$.

$$x = r\cos\theta$$
$$y = r\sin\theta$$
$$z = z$$

Position vector is $\mathbf{R} = r\hat{\mathbf{e}}_r + z\hat{\mathbf{e}}_z$.

$$\frac{d\hat{\mathbf{e}}_r}{d\theta} = \hat{\mathbf{e}}_\theta, \qquad \frac{d\hat{\mathbf{e}}_\theta}{d\theta} = -\hat{\mathbf{e}}_r$$

Spherical coordinates: Know Fig. 6a.

Coordinates ρ, ϕ, θ and base vectors $\hat{\mathbf{e}}_\rho(\phi, \theta), \hat{\mathbf{e}}_\phi(\phi, \theta), \hat{\mathbf{e}}_\theta(\theta)$.

$$x = \rho\sin\phi\cos\theta$$
$$y = \rho\sin\phi\sin\theta$$
$$z = \rho\cos\phi$$

Position vector is $\mathbf{R} = \rho\hat{\mathbf{e}}_\rho$.

Space derivatives of the base vectors given by (28); no need to memorize these.

Chapter 15

Curves, Surfaces, and Volumes

15.1 Introduction

In applications such as the calculation of moments of inertia, particle dynamics, and the gravitational force field induced by a given arrangement of mass, we need to know how to represent curves, surfaces, and volumes in 3-space, and how to obtain, from those representations, expressions for the differential arc length ds along a curve, the area element dA on a surface, and the volume element dV. We begin with space curves and integrals along those curves.

15.2 Curves and Line Integrals

Curves in space are, of course, of great importance. In mechanics, for example, the relationship between the forces acting on a point mass and the trajectory of the mass (i.e., the space curve developed by the motion) is of critical interest.

We met curves, indirectly, in Section 14.5, when we considered the variation of a vector function $\mathbf{u}(\tau)$ with τ. For, in general, if the tail of the vector $\mathbf{u}(\tau)$ is fixed, then the head of $\mathbf{u}(\tau)$ generates a space curve as the parameter τ is varied. Here, we shift our focus from the vector function $\mathbf{u}(\tau)$ to the space curve itself.

15.2.1. Curves. Let a Cartesian x, y, z coordinate system be specified, and let

$$x = x(\tau), \qquad y = y(\tau), \qquad z = z(\tau) \tag{1}$$

be continuous functions of a real parameter τ over a closed interval $[a, b]$, that is, over $a \leq \tau \leq b$. The points $P(\tau) \equiv (x(\tau), y(\tau), z(\tau))$ for $a \leq \tau \leq b$ are said to constitute a **curve** joining the endpoints $P(a)$ and $P(b)$, and we call (1) a **parametrization** of the curve.* If the endpoints coincide the curve is called a

*If each of $x(\tau)$, $y(\tau)$, and $z(\tau)$ is constant, the point set $P(\tau)$ is a single point and is called a *point curve*. Although we limit ourselves, in this section, to curves in 3-space, the definition given here may be extended to curves in n-space.

closed curve; if not, it is called an **arc**. If the tail of the position vector

$$\mathbf{R} = x(\tau)\hat{\mathbf{i}} + y(\tau)\hat{\mathbf{j}} + z(\tau)\hat{\mathbf{k}} \tag{2}$$

is fixed at the origin, then the head of $\mathbf{R}(\tau)$ generates the curve as τ varies from a to b.

As indicated in Section 14.5 (see the discussion associated with Fig. 3 therein), the derivative $\mathbf{R}'(\tau)$ (if it exists and is nonzero) is a **tangent vector** to the curve at the point $P(\tau)$.

EXAMPLE 1. Suppose that

$$x = \beta\tau + \cos\tau, \qquad y = \sin\tau, \qquad z = 0, \tag{3}$$

where β is a constant. Then $\mathbf{R}(\tau) = (\beta\tau + \cos\tau)\hat{\mathbf{i}} + \sin\tau\hat{\mathbf{j}}$. Since $z = 0$, the curve lies in a plane (the x, y plane in this case) and is therefore called a *plane curve*. Choosing different values of β and different τ intervals, we obtain different curves. Four such cases are illustrated in Fig. 1. ∎

We classify a curve $\mathcal{C}$ as **simple** if it does not intersect itself. If $\mathcal{C}$ is an arc it will be a *simple arc* if $\tau_1 \neq \tau_2$ implies that $P(\tau_1) \neq P(\tau_2)$ whenever τ_1 and τ_2 are in $[a, b]$; if $\mathcal{C}$ is a closed curve it will be a **simple closed curve** if $\tau_1 \neq \tau_2$ implies that $P(\tau_1) \neq P(\tau_2)$ whenever τ_1 and τ_2 are in $[a, b)$.* For instance, the curves in Fig. 1a, b, d are simple, and the one in Fig. 1c is not.

Further, we say that $\mathcal{C}$ is **smooth** if it possesses a tangent vector that varies continuously along the length of $\mathcal{C}$, and it is **piecewise smooth** if it is comprised of finitely many smooth segments, end to end. For instance, the curves in Fig. 1a, b, c are smooth, and the one in Fig. 1d is piecewise smooth.

In Example 1 we were given the parametric equations of the curve from which we were able to produce its graph. Sometimes we know the curve, but are not given parametric equations for it. In the next two examples we illustrate the determination of parametric equations for a given curve.

EXAMPLE 2. Find parametric equations for the curve $\mathcal{C}$ that is the intersection of the two planes

$$\begin{aligned} x - y + 2z &= 4, \\ 2x + y - z &= 2. \end{aligned} \tag{4}$$

$\mathcal{C}$ is the solution set of equations (4) so apply Gauss elimination and find that $z = \tau$ (arbitrary), $y = -2 + \frac{5}{3}\tau$, and $x = 2 - \frac{1}{3}\tau$. Indeed, these are the desired parametric equations of $\mathcal{C}$, where $-\infty < \tau < \infty$. Or, in 3-tuple vector form, $(x, y, z) =$

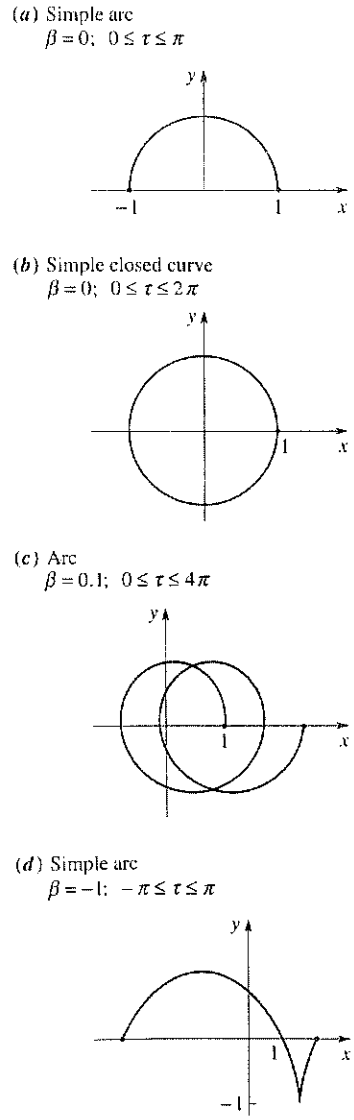


Figure 1. Several cases of the curve (3).

*$[a, b)$ means $a \leq \tau < b$. We write $[a, b)$ rather than $[a, b]$ because the endpoints $P(a)$ and $P(b)$ coincide: $P(a) = P(b)$.

$(2 - \frac{1}{3}\tau, -2 + \frac{5}{3}\tau, \tau) = (2, -2, 0) + \tau(-\frac{1}{3}, \frac{5}{3}, 1)$, where $(2, -2, 0)$ is from the origin to a particular point *on* C and $(-\frac{1}{3}, \frac{5}{3}, 1)$ is a vector *along* C. ∎

EXAMPLE 3. Find parametric equations for the curve C that is the ellipse

$$\left(\frac{x}{2}\right)^2 + \left(\frac{y}{3}\right)^2 = 1 \tag{5}$$

in the x, y plane. If we had a *circle* $x^2 + y^2 = 1$ we could use $x = \cos\tau$ and $y = \sin\tau$, that is, $x = r\cos\theta$ and $y = r\sin\theta$, where $r = 1$ and where we use θ as our parameter τ. Adapting that idea to (5), let $\frac{x}{2} = \cos\tau$ and $\frac{y}{3} = \sin\tau$. Thus, parametric equations for the ellipse (5) are

$$x = 2\cos\tau, \qquad y = 3\sin\tau, \tag{6}$$

where $0 \le \tau < 2\pi$, say. [Whether or not we add $z = 0$ to (6) depends on whether we are considering C as a curve in the x, y plane or in x, y, z space.] ∎

Observe that the parametrization of a given curve is not unique. In Example 3, for instance, the alternative parametrizations $x = 2\sin\tau$, $y = 3\cos\tau$ and $x = 2\cos\tau^2$, $y = 3\sin\tau^2$ work just as well.

One case worth emphasizing is the parametrization of a straight line from a given point $P_1(x_1, y_1, z_1)$ to another given point $P_2(x_2, y_2, z_2)$. We claim that

$$\boxed{\begin{aligned} x &= x_1 + (x_2 - x_1)\tau, \\ y &= y_1 + (y_2 - y_1)\tau, \\ z &= z_1 + (z_2 - z_1)\tau \end{aligned}} \qquad (0 \le \tau \le 1) \tag{7}$$

always works. Obviously, the right-hand sides of (7) give (x_1, y_1, z_1) when $\tau = 0$, and (x_2, y_2, z_2) when $\tau = 1$. Not quite as obvious is the fact that the curve thus parametrized is *straight*. To verify that it is straight, write the position vector from the origin to any point on the curve as

$$\mathbf{R}(\tau) = [x_1 + (x_2 - x_1)\tau]\hat{\mathbf{i}} + [y_1 + (y_2 - y_1)\tau]\hat{\mathbf{j}} + [z_1 + (z_2 - z_1)\tau]\hat{\mathbf{k}}.$$

Since the tangent vector to the curve,

$$\mathbf{R}'(\tau) = (x_2 - x_1)\hat{\mathbf{i}} + (y_2 - y_1)\hat{\mathbf{j}} + (z_2 - z_1)\hat{\mathbf{k}},$$

is a constant vector, it follows that the curve parametrized by (7) is, indeed, a straight line.

15.2.2. Arc length. Next, we consider the arc length $s(\tau)$ of a curve C, from some point $P(\tau_0)$ on C to any other point $P(\tau)$ on C. We know what we mean by the length of a straight line, but what is the length of a curved arc? As is done so often in mathematics, we use the limit concept to define this "new" quantity as the

limit of a sequence of "old" quantities.* Specifically, we approximate the curve by a system of linear segments, as in Fig. 2, and sum the lengths of those segments. Then we repeat the process using a finer system of linear segments, and so on. If the endpoints of any single linear segment (such as AB in Fig. 2) are located with respect to the origin of some coordinate system by position vectors $\mathbf{R}_A$ and $\mathbf{R}_B$, then the length of the segment is

$$\|\mathbf{R}_B - \mathbf{R}_A\| = \|\Delta\mathbf{R}\| = \sqrt{\Delta\mathbf{R}\cdot\Delta\mathbf{R}} = \sqrt{\frac{\Delta\mathbf{R}}{\Delta\tau}\cdot\frac{\Delta\mathbf{R}}{\Delta\tau}}\,\Delta\tau$$

so in differential form we have

$$\boxed{ds = \sqrt{\mathbf{R}'(\tau)\cdot\mathbf{R}'(\tau)}\,d\tau.} \tag{8}$$

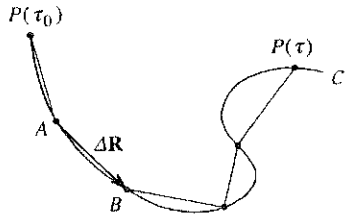


Figure 2. Approximation of C by linear segments.

Thus, the limiting process of summation described above evidently leads to the Riemann integral

$$\boxed{s(\tau) = \int_{\tau_0}^{\tau} \sqrt{\mathbf{R}'(t)\cdot\mathbf{R}'(t)}\,dt,} \tag{9}$$

which we adopt as our *definition* of the **arc length** of C, from $P(\tau_0)$ to $P(\tau)$, provided of course that the integral exists (i.e., is convergent). It can be shown that the arc length of a given arc is independent of the location and orientation of the reference Cartesian coordinate system and is also independent of the particular choice of parametrization $x(\tau), y(\tau), z(\tau)$. CAUTION: The integral in (9) gives the arc length only if $\tau \geq \tau_0$; if $\tau < \tau_0$, then the integral is negative and gives the negative of the arc length between $P(\tau_0)$ and $P(\tau)$.

EXAMPLE 4. As a simple illustration, use (9) to compute the arc length of the (semicircular) curve shown in Fig. 1a. Then

$$\begin{aligned}\mathbf{R}(\tau) &= x(\tau)\hat{\mathbf{i}} + y(\tau)\hat{\mathbf{j}} = \cos\tau\,\hat{\mathbf{i}} + \sin\tau\,\hat{\mathbf{j}},\\ \mathbf{R}'(\tau) &= -\sin\tau\,\hat{\mathbf{i}} + \cos\tau\,\hat{\mathbf{j}},\\ \mathbf{R}'(\tau)\cdot\mathbf{R}'(\tau) &= \sin^2\tau + \cos^2\tau = 1\end{aligned}$$

so (9) is simply

$$s = \int_0^{\pi} \sqrt{1}\,dt = \pi. \quad \blacksquare$$

*Recall, for instance, that the derivative is defined (in the calculus) as the limit of a difference quotient, the integral is defined as the limit of a sequence of Riemann sums, an infinite series is ordinarily defined as the limit of a sequence of partial sums, and so on. It is interesting that the limit is the fundamental and unifying concept of the calculus, yet it was not clarified until the nineteenth century, through the work of Cauchy (1789–1857), Weierstrass (1815–1897), and others, long after the birth of the calculus.

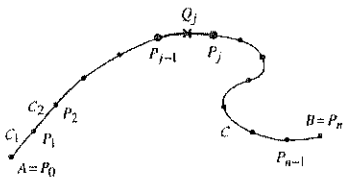


Figure 3. Partition of C.

15.2.3. Line integrals. Consider a given curve C and a function of $f(x, y, z)$ defined along C (perhaps off of C as well). Measuring arc length s from one endpoint, say A (Fig. 3), divide C into n arcs by specifying points P_0 $(= A)$, $P_1, P_2, \ldots, P_{n-1}, P_n$ $(= B)$ along C. Let the division be chosen arbitrarily, provided that the P_j points are spaced and numbered so that the arc length from A to P_j is less than arc length from A to P_k if $j < k$. Denote the arcs as $C_1, C_2, \ldots, C_n$, where the endpoints of C_j are P_{j-1} and P_j. On each arc C_j choose, arbitrarily, some point Q_j that is anywhere between the endpoints of C_j, or *at* one of the endpoints, and form the sum

$$J_n = \sum_{j=1}^{n} f(Q_j)\Delta s_j, \tag{10}$$

where Δs_j is the arc length of C_j. The choice of the P_j's and Q_j's defines a **partition** of C, and we call the largest Δs_j the **norm** of the partition. Actually, we introduce not just one partition but a sequence of them such that the norm of the nth partition tends to zero as $n \to \infty$. If the corresponding sequence of sums $J_1, J_2, \ldots$ converges to a limit, we call that limit the **line integral** of f over C, and write it as

$$\int_C f\, ds. \tag{11}$$

In this chapter we study integrals over curves, surfaces, and volumes so the terms curve integral, surface integral, and volume integral would appear to be natural choices. However, (11) is usually called a line integral, rather than a curve integral, so we will use that terminology.

The limit described above, and hence the integral (11), will indeed exist (i.e., converge) if f is continuous (or even piecewise continuous) along C and if C is piecewise smooth and simple.

From this definition of the line integral it can be shown that the following two important properties follow:

$$\int_C (\alpha f + \beta g) ds = \alpha \int_C f\, ds + \beta \int_C g\, ds \qquad \text{(linearity)} \tag{12a}$$

and

$$\int_C f\, ds = \int_{C_1} f\, ds + \int_{C_2} f\, ds, \tag{12b}$$

where α, β are any scalars, f, g are any two functions (continuous along C), and where "$C = C_1 + C_2$," that is, where C is divided into two parts, C_1 and C_2 (just as C was divided into n parts in Fig. 3).

In practice, however, we do not evaluate line integrals by seeking the limit of the sequence $J_1, J_2, \ldots$. Rather, we introduce a coordinate system and proceed as follows. If C is parametrized with respect to some parameter τ by $x = x(\tau)$, $y = y(\tau)$, $z = z(\tau)$, for $a \le \tau \le b$, then

$$\boxed{\int_C f(x, y, z)\, ds = \int_a^b f(x(\tau), y(\tau), z(\tau)) \sqrt{\mathbf{R}'(\tau) \cdot \mathbf{R}'(\tau)}\, d\tau.} \tag{13}$$

EXAMPLE 5. *Mass of a Helical Wire.* Determine the mass

$$M = \int_C \sigma \, ds \tag{14}$$

of a wire that is in the shape of a curve C and that has a mass density σ (mass per unit length) that varies along C. In particular, suppose that C is comprised of N turns of a circular helix of radius a, defined parametrically by

$$x(\tau) = a\cos\tau, \quad y(\tau) = a\sin\tau, \quad z(\tau) = b\tau, \qquad (0 \le \tau \le 2N\pi) \tag{15}$$

where τ is actually the cylindrical-coordinate angle θ. Observe that in one turn of the helix (i.e., as τ increases by 2π) z increases by $2\pi b$, which quantity is known as the *pitch* of the helix. As b is diminished the helix becomes more and more compressed, collapsing to a circle of radius a in the x, y plane as $b \to 0$; conversely, it becomes more and more stretched out as b is increased. And let the density vary linearly with τ as $\sigma(\tau) = c\tau$ for some positive constant c. Then

$$\begin{aligned} \mathbf{R}(\tau) &= a\cos\tau\hat{\mathbf{i}} + a\sin\tau\hat{\mathbf{j}} + b\tau\hat{\mathbf{k}}, \\ \mathbf{R}'(\tau) &= -a\sin\tau\hat{\mathbf{i}} + a\cos\tau\hat{\mathbf{j}} + b\hat{\mathbf{k}} \end{aligned}$$

so (13) gives

$$M = \int_0^{2N\pi} c\tau\sqrt{a^2+b^2}\, d\tau = c\sqrt{a^2+b^2}\,\frac{(2N\pi)^2}{2}. \quad \blacksquare$$

Closure. The key to space curves and curve integrals is in their parametric representation. We have considered space curves C defined parametrically by

$$x = x(\tau), \quad y = y(\tau), \quad z = z(\tau). \qquad (a \le \tau \le b)$$

The parameter τ might have a geometrical significance (as in Example 5, for instance, where τ amounted to the polar angle θ of an r, θ, z cylindrical coordinate system), it might represent time, or it might have no particular significance. Likewise, the differential arc length along C can be expressed in terms of τ as

$$ds = \sqrt{\mathbf{R}'(\tau)\cdot\mathbf{R}'(\tau)}\, d\tau,$$

where $\mathbf{R}(\tau) = x(\tau)\hat{\mathbf{i}} + y(\tau)\hat{\mathbf{j}} + z(\tau)\hat{\mathbf{k}}$ is a position vector from the origin of the reference Cartesian coordinate system to any point $P(\tau)$ on C. Finally, an integral $\int_C f\, ds$ along a curve C can, by parametrization, be reduced to an "ordinary integral" of a function of τ along a segment of a τ axis, as indicated in (13).

Computer software. Using *Maple*, you can obtain the graph of a curve C defined parametrically by

$$\mathbf{R}(\tau) = x(\tau)\hat{\mathbf{i}} + y(\tau)\hat{\mathbf{j}} + z(\tau)\hat{\mathbf{k}} \qquad (a \le \tau \le b)$$

using the **spacecurve** command. For instance, to plot the helix

$$\mathbf{R}(\tau) = 3\cos\tau\hat{\mathbf{i}} + 3\sin\tau\hat{\mathbf{j}} + 20\tau\hat{\mathbf{k}} \qquad (0 \le \tau \le 25)$$

enter

with(plots):

and return, to access the spacecurve command. Then enter

spacecurve([3 * cos (t), 3 * cos (t), 20 * t], t = 0..25);

and return. Two problems: first is that the graph is kinky because the default number of points used is only 50 and that is not enough for approximately four (25 divided by 2π) turns of the helix and, second, axes are not shown. To increase the number of points to 500, say, use the option numpoints = 500, and to show the labeled coordinate axes use the option axes = FRAMED or axes = BOXED. Thus, use

spacecurve([3 * cos (t), 3 * cos (t), 20 * t], t = 0..25,
numpoints = 500, axes = BOXED);

for instance.

EXERCISES 15.2

1. Determine a parametrization for each of the following curves: the intersection of

(a) the planes $x - 2y + z = 4,\ 2x + y - z = 0$
(b) the planes $4x - 3y + 5z = 1,\ x + y - 2z = 6$
(c) the planes $2x + y + z = 3,\ x + 4z = 5$
(d) the planes $x - y + z = 0,\ 5x - y + z = 2$
(e) the planes $2x + y - z = 1,\ 2y + z = 4$
(f) the plane $x + y + 2z = 5$ and the circular cylinder $x^2 + y^2 = 4$
(g) the plane $4x - y + z = 3$ and the circular cylinder $x^2 + z^2 = 1$
(h) the plane $x - y + 2z = 1$ and the elliptic cylinder $x^2 + 4y^2 = 4$
(i) surfaces $2x - y - 3z = 5$ and $z = x^2 + y + 1$, between $(0, -2, -1)$ and $(2, -4, 1)$
(j) surfaces $x - y^2 + z = 0$ and $x - y + 2z = 0$, between $(6, 2, -2)$ and $(1, 1, 0)$
(k) surfaces $x = y^2 + z^2$ and $z = xy$, between $(0, 0, 0)$ and $(1, -1/\sqrt{2}, -1/\sqrt{2})$

2. Give a parametrization of the straight line connecting

(a) $(5, -1, 2)$ and $(2, 0, 6)$ (b) $(1, 2, 3)$ and $(3, 7, -4)$
(c) $(8, 1, 7)$ and $(0, 28, 12)$ (d) $(1, 0, 0)$ and $(-6, 0, 0)$

3. Let each of the following $\mathbf{R}(\tau)$'s, for $0 \le \tau < \infty$, define a space curve. In each case, find the arc length s as a function of the parameter τ, namely, $s(\tau)$ if $\tau = 0$ is taken to correspond to $s = 0$. NOTE: Naturally, $\cos\tau\hat{\mathbf{i}}$, for instance, means $(\cos\tau)\hat{\mathbf{i}}$.

(a) $\hat{\mathbf{i}} + \tau\hat{\mathbf{j}} + \tau^2\hat{\mathbf{k}}$ (b) $\cos\tau\hat{\mathbf{i}} + 4\tau\hat{\mathbf{j}} + \sin\tau\hat{\mathbf{k}}$
(c) $\tau\hat{\mathbf{i}} - 3\tau\hat{\mathbf{j}} + 5(\tau + 4)\hat{\mathbf{k}}$ (d) $\cos\tau(\hat{\mathbf{i}} + \hat{\mathbf{j}}) - \sqrt{2}\sin\tau\hat{\mathbf{k}}$
(e) $\hat{\mathbf{i}} - 7\tau^2\hat{\mathbf{j}} + 3\hat{\mathbf{k}}$ (f) $\sin\tau\hat{\mathbf{i}} + \cos\tau\hat{\mathbf{j}} - \cos\tau\hat{\mathbf{k}}$

4. (*Serret–Frenet formulas*) The arc length s along a space curve $\mathcal{C}$ is itself a convenient parameter for the parametrization of $\mathcal{C}$. Letting τ be s, we have

$$\mathbf{R}(s) = x(s)\hat{\mathbf{i}} + y(s)\hat{\mathbf{j}} + z(s)\hat{\mathbf{k}}. \tag{4.1}$$

In what follows there are problems and statements. We ask you to follow along with the statements and to respond only to problems introduced by the italicized word *show*.

(a) Since $d\mathbf{R}$ is tangent to $\mathcal{C}$, $d\mathbf{R}/ds$ must be a tangent vector to $\mathcal{C}$. *Show* why the **tangent vector**

$$\frac{d\mathbf{R}}{ds} = \hat{\mathbf{T}} \tag{4.2}$$

is a unit vector.

(b) Since $\hat{\mathbf{T}}(s)$ is a unit vector, $\hat{\mathbf{T}} \cdot \hat{\mathbf{T}} = 1$. Differentiating that equation with respect to s, *show* that $\hat{\mathbf{T}} \cdot \hat{\mathbf{T}}' = 0$. Hence, $\hat{\mathbf{T}}'$ is either perpendicular to $\hat{\mathbf{T}}'$ or else it is $\mathbf{0}$ so we can express

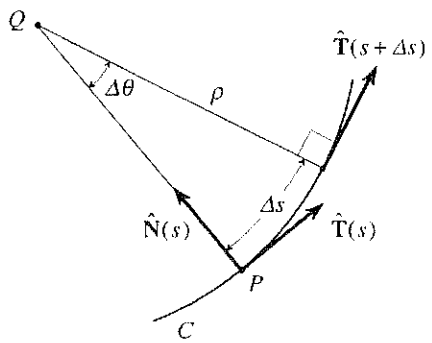
$$\frac{d\hat{\mathbf{T}}}{ds} = \kappa\hat{\mathbf{N}} \qquad (\kappa \geq 0) \tag{4.3}$$

where $\hat{\mathbf{N}}$ is a unit normal vector perpendicular to $\hat{\mathbf{T}}$ and κ is a scalar multiplier, which can be considered as nonnegative without loss of generality. $\hat{\mathbf{N}}(s)$ is called the **principal normal** and $\kappa(s)$ is called the **curvature** of C. If the curve happens to be straight (in some neighborhood of the point in question), then $\hat{\mathbf{T}}(s)$ is a constant vector and $\hat{\mathbf{T}}'(s) = \mathbf{0}$ so that, from (4.3), $\kappa = 0$, as is reasonable since a straight line has no curvature. In that case $\hat{\mathbf{N}} = (1/\kappa)\hat{\mathbf{T}}'$ is undefined. The plane containing $\hat{\mathbf{T}}$ and $\hat{\mathbf{N}}$ is called the **osculating plane**.

(c) In (b), we stated that κ is the curvature, which means the numerical inverse of the **radius of curvature** ρ,

$$\kappa = \frac{1}{\rho}. \tag{4.4}$$

Show that it follows from (4.3) that κ is indeed the curvature. HINT: Use the accompanying figure (where Q is the



local center of curvature corresponding to the point P on C) and a difference quotient approximation of $\hat{\mathbf{T}}'(s)$. Note from (4.4) that if the radius of curvature ρ is large then the curvature κ is small, and if ρ is small then κ is large. Finally, we introduce a third unit vector $\hat{\mathbf{B}}(s)$, the **binormal**, according to

$$\hat{\mathbf{B}} = \hat{\mathbf{T}} \times \hat{\mathbf{N}} \tag{4.5}$$

so that at each point P on C $\left\{\hat{\mathbf{T}}, \hat{\mathbf{N}}, \hat{\mathbf{B}}\right\}$ is a right-handed orthonormal set. The spatial orientation of the $\hat{\mathbf{T}}, \hat{\mathbf{N}}, \hat{\mathbf{B}}$ triad varies, in general, as the representative point P [at the tip of $\mathbf{R}(s)$] moves along the curve C. Thus, in general, $\hat{\mathbf{T}}(s), \hat{\mathbf{N}}(s), \hat{\mathbf{B}}(s)$ will all be functions of s. Indeed, the variation of $\hat{\mathbf{T}}$ with s has already been given in (4.3). Let us obtain analogous expressions for $\hat{\mathbf{N}}'(s)$ and $\hat{\mathbf{B}}'(s)$.

(d) First $\hat{\mathbf{B}}'(s)$: From (4.5), *show* that $\hat{\mathbf{B}}' = \hat{\mathbf{T}} \times \hat{\mathbf{N}}'$. It follows from the latter that $\hat{\mathbf{B}}'$ is perpendicular to $\hat{\mathbf{T}}$ (also to $\hat{\mathbf{N}}'$, but that fact will not be used here). But $\hat{\mathbf{B}}$ is a unit vector so, by the same logic as given in (b) for $\hat{\mathbf{T}}$, $\hat{\mathbf{B}}'$ must be perpendicular to $\hat{\mathbf{B}}$ or be $\mathbf{0}$. Thus, $\hat{\mathbf{B}}'$ is perpendicular to the $\hat{\mathbf{B}}$, $\hat{\mathbf{T}}$ plane, or else it is $\mathbf{0}$ so it must be expressible in the form

$$\frac{d\hat{\mathbf{B}}}{ds} = \nu\hat{\mathbf{N}}, \qquad (\nu \geq 0) \tag{4.6}$$

where the scalar multiple ν in (4.6) is called the **torsion of** C at P because it is a measure of the rate at which the curve twists out of the osculating plane; if the curve is straight we define $\nu = 0$. NOTE: The torsion is generally denoted as τ, not ν, in the literature, but we have already used τ extensively as a parameter and prefer to use a different letter for the torsion.

(e) Finally, differentiating $\hat{\mathbf{N}} = \hat{\mathbf{B}} \times \hat{\mathbf{T}}$ (for recall that $\left\{\hat{\mathbf{T}}, \hat{\mathbf{N}}, \hat{\mathbf{B}}\right\}$ is a right-handed orthonormal set) *show* that

$$\frac{d\hat{\mathbf{N}}}{ds} = -\kappa\hat{\mathbf{T}} - \nu\hat{\mathbf{B}}. \tag{4.7}$$

Equations (4.3), (4.6), and (4.7), namely,

$$\boxed{\begin{aligned} \frac{d\hat{\mathbf{T}}}{ds} &= \kappa\hat{\mathbf{N}}, \\ \frac{d\hat{\mathbf{N}}}{ds} &= -\kappa\hat{\mathbf{T}} - \nu\hat{\mathbf{B}}, \\ \frac{d\hat{\mathbf{B}}}{ds} &= \nu\hat{\mathbf{N}}, \end{aligned}} \tag{4.8a,b,c}$$

are known as the **Serret–Frenet formulas.*** The latter are fundamental to the study of space curves for the following reason. If $\hat{\mathbf{T}}, \hat{\mathbf{N}}, \hat{\mathbf{B}}$ are expressed in terms of their scalar components then (4.8) is a system of nine linear first-order differential

*These formulas were obtained by *Frederic–Jean Frenet* (1816–1900) in his dissertation (1847) and, independently, by *Joseph Alfred Serret* (1819-1885) in 1851.

equations. If $\kappa(s)$ and $\nu(s)$ are prescribed continuous functions, then that system admits a solution for $\hat{\mathbf{T}}(s), \hat{\mathbf{N}}(s), \hat{\mathbf{B}}(s)$ that is unique to within the specification of $\hat{\mathbf{T}}(0), \hat{\mathbf{N}}(0), \hat{\mathbf{B}}(0)$. With $\hat{\mathbf{T}}(s)$ thus determined, we can determine $\mathbf{R}(s)$ and hence the space curve C, by integrating (4.2). For details, we refer the interested reader to J. J. Stoker, *Differential Geometry* (New York: Wiley–Interscience, 1969) or D. J. Struik, *Differential Geometry*, 2nd ed. (Reading, MA: Addison–Wesley, 1961).

5. First, read Exercise 4. Given the circular helix

$$\mathbf{R}(\tau) = a\cos\tau\hat{\mathbf{i}} + a\sin\tau\hat{\mathbf{j}} + b\tau\hat{\mathbf{k}} \qquad (a, b > 0,\ 0 \le \tau < \infty) \tag{5.1}$$

derive the results

$$\begin{aligned} s(\tau) &= \sqrt{a^2+b^2}\,\tau, \\ \hat{\mathbf{T}}(\tau) &= c(-a\sin\tau\hat{\mathbf{i}} + a\cos\tau\hat{\mathbf{j}} + b\hat{\mathbf{k}}), \\ \hat{\mathbf{N}}(\tau) &= -\cos\tau\hat{\mathbf{i}} - \sin\tau\hat{\mathbf{j}}, \\ \hat{\mathbf{B}}(\tau) &= c(b\sin\tau\hat{\mathbf{i}} - b\cos\tau\hat{\mathbf{j}} + a\hat{\mathbf{k}}), \\ \kappa(\tau) &= a/(a^2+b^2), \\ \rho(\tau) &= (a^2+b^2)/a, \\ \nu(\tau) &= -b/(a^2+b^2), \end{aligned} \tag{5.2}$$

where $c \equiv 1/\sqrt{a^2+b^2}$. HINT: Don't forget to use chain differentiation. For example, $\hat{\mathbf{T}} = \dfrac{d\mathbf{R}}{ds} = \dfrac{d\mathbf{R}}{d\tau}\dfrac{d\tau}{ds}$.

6. In Exercise 5 we see that for a circular helix the curvature $\kappa(\tau)$ and torsion $\nu(\tau)$ are constant along the helix. In this exercise let us work in the opposite direction. Specifically, let κ and ν be constants in the Frenet–Serret equations (4.8) and show, by deriving it, that if κ and ν are not both zero then the solution of (4.8) is

$$\begin{aligned} \hat{\mathbf{N}} &= \mathbf{C}_1\sin\omega s + \mathbf{C}_2\cos\omega s, \\ \hat{\mathbf{T}} &= \frac{\kappa}{\omega}\mathbf{C}_2\sin\omega s - \frac{\kappa}{\omega}\mathbf{C}_1\cos\omega s + \nu\mathbf{C}_3, \\ \hat{\mathbf{B}} &= \frac{\nu}{\omega}\mathbf{C}_2\sin\omega s - \frac{\nu}{\omega}\mathbf{C}_1\cos\omega s - \kappa\mathbf{C}_3, \end{aligned} \tag{6.1}$$

where $\omega \equiv \sqrt{\kappa^2+\nu^2}$, and $\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3$ are the (vector) constants of integration. Then show that integration of (4.2) gives

$$\mathbf{R} = \mathbf{C}_4\cos\omega s + \mathbf{C}_5\sin\omega s + \mathbf{C}_6 s + \mathbf{C}_7. \tag{6.2}$$

Finally, verify that (5.1), in Exercise 5, is indeed of the form (6.2).

7. (*Velocity and acceleration of a particle*) First, read Exercise 4. Let the velocity of a moving particle be given by a position vector $\mathbf{R} = \mathbf{R}(t)$, where t is the time. We wish to obtain expressions for the velocity $\mathbf{v} = d\mathbf{R}/dt$ and the acceleration $\mathbf{a} = d\mathbf{v}/dt = d^2\mathbf{R}/dt^2$ in terms of $\hat{\mathbf{T}}, \hat{\mathbf{N}}, \hat{\mathbf{B}}, \kappa$, and ν.

(a) With the help of chain differentiation (namely, $\dfrac{d(\,)}{dt} = \dfrac{d(\,)}{ds}\dfrac{ds}{dt} = v\dfrac{d(\,)}{ds}$, where $\dfrac{ds}{dt} = v = \|\mathbf{v}\|$ is the *speed*), derive the formulas

$$\boxed{\begin{aligned} \mathbf{v} &= v\hat{\mathbf{T}}, \\ \mathbf{a} &= \frac{dv}{dt}\hat{\mathbf{T}} + \kappa v^2\hat{\mathbf{N}}. \end{aligned}} \tag{7.1a,b}$$

NOTE: Equation (7.1b) is simple and informative. Since $\mathbf{a}$ is a linear combination of $\hat{\mathbf{T}}$ and $\hat{\mathbf{N}}$ it lies in the osculating plane. The first term on the right is simply the linear acceleration, the acceleration that would occur if the particle moved in a straight line with variable speed $v(t)$. The second term is directed toward the instantaneous center of curvature (the point Q in the figure in Exercise 4) and its magnitude is κv^2, or v^2/ρ. This is the familiar *centripetal acceleration* that we meet in physics when studying circular motion. Note further that we can now appreciate that the Serret–Frenet formulas (4.8) play the same role as the formulas for the spatial derivatives of the non-Cartesian base vectors (such as $\partial\hat{\mathbf{e}}_\rho/\partial\phi$, $\partial\hat{\mathbf{e}}_\phi/\partial\theta$, and so on) in Section 14.6. For in deriving (7.1b) from (7.1a) we needed to evaluate $d\hat{\mathbf{T}}/dt$. Chain differentiation gave $(d\hat{\mathbf{T}}/ds)(ds/dt)$, and then the Serret–Frenet formula (4.8a) gave $d\hat{\mathbf{T}}/ds$ as a linear combination of the base vectors $\hat{\mathbf{T}}$, $\hat{\mathbf{N}}$ and $\hat{\mathbf{B}}$.

(b) Actually, the *third* derivative $d^3\mathbf{R}/dt^3$ is also of interest in engineering applications (e.g., in cam design and in the analysis of mechanisms) and, quaintly, is called the **jerk**, $\mathbf{j}(t)$. Show, from (7.1b), that

$$\mathbf{j} = \left(\frac{\mathbf{d^2v}}{\mathbf{dt^2}} - \kappa^2\mathbf{v}^3\right)\hat{\mathbf{T}} + 3\kappa\mathbf{v}\frac{\mathbf{dv}}{\mathbf{dt}}\hat{\mathbf{N}} - \kappa\nu\mathbf{v}^3\hat{\mathbf{B}}. \tag{7.2}$$

8. Compute the normal $\hat{\mathbf{N}}(x)$ for the plane curve $y = x^3$. Sketch the curve over $-2 \le x \le 2$, and show $\hat{\mathbf{N}}$ at a number of points, say, $x = \pm 0.01$ and $x = \pm 1.0$.

9. Let a curve in the x, y plane be defined by $y = y(x)$. Show that

$$\boxed{\kappa = \frac{1}{\rho} = \frac{|y''|}{(1+y'^2)^{3/2}},} \tag{9.1}$$

provided, of course, that $y(x)$ is twice differentiable.

10. Prove that $P(\tau) = (x(\tau), y(\tau), z(\tau))$ is a plane curve if and only if

$$\begin{vmatrix} x' & y' & z' \\ x'' & y'' & z'' \\ x''' & y''' & z''' \end{vmatrix} = 0, \tag{10.1}$$

where primes denote $d/d\tau$. HINT: Recall the general equation of a plane, $ax + by + cz = d$, where a, b, c, d are constants and a, b, c are not all zero. If $P(\tau)$ lies in a plane, then $ax(\tau) + by(\tau) + c(\tau) = d$.

11. Let each of the following $\mathbf{R}(\tau)$'s, for $0 \le \tau < \infty$, define a space curve. Use the results in Exercise 10 to show whether or not the curve is a plane curve.

(a) $\tau\hat{\mathbf{i}} - 2\tau\hat{\mathbf{j}} + \tau^3\hat{\mathbf{k}}$
(b) $\tau^2\hat{\mathbf{i}} + \tau^2\hat{\mathbf{j}} - 3\tau\hat{\mathbf{k}}$
(c) $2\hat{\mathbf{i}} - \hat{\mathbf{j}} + 5\tau\hat{\mathbf{k}}$
(d) $3\hat{\mathbf{i}} + \sin\tau\hat{\mathbf{j}} + \cos\tau\hat{\mathbf{k}}$

12. (a)–(f) Obtain a computer plot of the curve defined in the corresponding part of Exercise 3, over $0 \le \tau \le 5$.

15.3 Double and Triple Integrals

Having studied curve integrals, we can turn to surface and volume integrals. Since these are multiple integrals, let us begin with a brief review of double and triple integrals.

15.3.1. Double integrals. Consider the double integral

$$I = \iint_{\mathcal{R}} f(x, y)\, dA, \tag{1}$$

where $\mathcal{R}$ is a region in the x, y plane. We assume that $\mathcal{R}$ is a closed, bounded region in the x, y plane,* that its boundary $\mathcal{C}$ is a closed piecewise smooth curve, and that $f(x, y)$ is defined on $\mathcal{R}$.

Let us lay down a rectangular grid on $\mathcal{R}$ consisting of a finite number of lines parallel to the coordinate axes, for example, as shown in Fig. 1. The N rectangles lying entirely within $\mathcal{R}$ (the shaded ones in Fig. 1) constitute a **partition** of $\mathcal{R}$, and the greatest of the dimensions $\Delta x_j, \Delta y_j$ for $j = 1, \ldots, N$ will be called the **norm** of the partition and will be denoted by the symbol $|p|$. Let (x_j, y_j) be an arbitrarily selected point in the jth partition rectangle for each $j = 1, 2, \ldots, N$. Then, denoting the area $\Delta x_j \Delta y_j$ as ΔA_j,

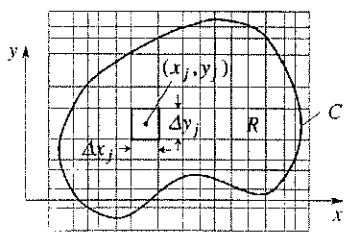


Figure 1. Partition of $\mathcal{R}$.

$$\sum_{j=1}^{N} f(x_j, y_j)\Delta A_j \tag{2}$$

is called the **Riemann sum** corresponding to the chosen partition and (x_j, y_j) points. The idea, in principle at least, is to compute the Riemann sum (2), then

*The term *closed region* is defined in Section 13.2.2. $\mathcal{R}$ being *bounded* means that it can be enclosed within a sufficiently large circle (or sphere in the three-dimensional case).

to introduce a finer partition (i.e., one with a smaller norm) and compute the new Riemann sum, and so on, such that the norm of the partitions tends to zero. Two things happen: the partitioning rectangles become arbitrarily small, and the unshaded area within $\mathcal{R}$ (Fig. 1) evidently tends to zero.

If the sequence of values of the Riemann sum thus generated converges to a unique limit, independent of the choice of the partition sequence and (x_j, y_j) points, that limit is, by definition, the **double integral** (1). That is,

$$\iint_{\mathcal{R}} f(x,y)\,dA \equiv \lim_{|p|\to 0} \sum f(x_j, y_j)\Delta A_j. \tag{3}$$

It can be shown that the limit will indeed exist if f is continuous in the closed region $\mathcal{R}$, that condition being sufficient, not necessary. If the limit does exist, we say that the integral *converges*, or *exists*, and that f is *integrable* on $\mathcal{R}$, Riemann-integrable to be more specific since the Riemann definition (3) is not the only definition of the double integral.

Double integrals admit the following properties, each of which corresponds to an analogous property for line integrals:

$$\iint_{\mathcal{R}} [\alpha f(x,y) + \beta g(x,y)]dA = \alpha \iint_{\mathcal{R}} f(x,y)\,dA + \beta \iint_{\mathcal{R}} g(x,y)\,dA, \tag{4a}$$

$$\iint_{\mathcal{R}} f(x,y)\,dA = \iint_{\mathcal{R}_1} f(x,y)\,dA + \iint_{\mathcal{R}_2} f(x,y)\,dA, \tag{4b}$$

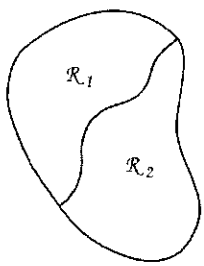


Figure 2. Subdivision of $\mathcal{R}$.

where α is a constant and all integrals are assumed to exist. The property (4a) is known as linearity. In (4b) $\mathcal{R}$ is divided (by a piecewise smooth curve) into two subregions, $\mathcal{R}_1$ and $\mathcal{R}_2$, as illustrated in Fig. 2. In addition to (4a) and (4b), we call attention to the **mean value theorem** for double integrals, which states that if $f(x,y)$ is continuous throughout the closed region $\mathcal{R}$, then there exists at least one point (x_0, y_0) in $\mathcal{R}$ such that

$$\boxed{\iint_{\mathcal{R}} f(x,y)\,dA = f(x_0, y_0)A,} \tag{4c}$$

where

$$A = \iint_{\mathcal{R}} dA \tag{5}$$

is the **area** of $\mathcal{R}$.

Besides serving to define the double integral on the left, (3) is useful from a computational standpoint since the Riemann sum (2), for a given partition, will provide an accurate approximation to the integral if the partition is sufficiently fine. [In practice, however, one normally uses approximations which are more efficient and more sophisticated than the simplest approximation provided by (2).]

Of course, we resort to numerical integration only when necessary, only when we are unable to carry out the evaluation analytically. How do we evaluate a double integral analytically? By dealing with it as an **iterated integral**, that is, by integrating on one variable at a time. The procedure is similar to that employed in differentiation. For example, in computing $u_{xy} = \frac{\partial}{\partial y}\left(\frac{\partial}{\partial x} u\right)$, first we differentiate u with respect to x, holding y fixed, and then we differentiate the resulting function $\partial u/\partial x$ with respect to y. Analogously, if we recall that $\Delta A_j = \Delta x_j \Delta y_j$ in (2), this fact suggests that we write $\iint_{\mathcal{R}} f(x,y)\,dA$ as $\iint_{\mathcal{R}} f(x,y)\,dx\,dy$, and that we contemplate two *successive* integrations, first on x and then on y. Thus, if $\mathcal{R}$ is the region shown in Fig. 3a, let us tentatively express

$$\iint_{\mathcal{R}} f(x,y)\,dA = \int_{y_1}^{y_2} \left\{ \int_{x_1(y)}^{x_2(y)} f(x,y)\,dx \right\} dy$$

or, omitting the parentheses for brevitiy,

$$\iint_{\mathcal{R}} f(x,y)\,dA = \int_{y_1}^{y_2} \int_{x_1(y)}^{x_2(y)} f(x,y)\,dx\,dy. \tag{6}$$

In words, we integrate across the (bold) horizontal sliver in Fig. 3a, from $x_1(y)$ on the left to $x_2(y)$ on the right, and then we sweep the sliver from bottom ($y = y_1$) to top ($y = y_2$).

Observe, in (6), that the "inner" integral, $\int_{x_1(y)}^{x_2(y)} f(x,y)\,dx$, is a function of y, say $F(y)$; the subsequently carried out "outer" integration is of the form $\int_{y_1}^{y_2} F(y)\,dy$, in which it is to be noted that the limits y_1 and y_2 are constants.

Alternatively, if we integrate first on y and then on x, as in Fig. 3b, we have

$$\int_{x_1}^{x_2} \left\{ \int_{y_1(x)}^{y_2(x)} f(x,y)\,dy \right\} dx \qquad \text{or} \qquad \int_{x_1}^{x_2} \int_{y_1(x)}^{y_2(x)} f(x,y)\,dy\,dx. \tag{7}$$

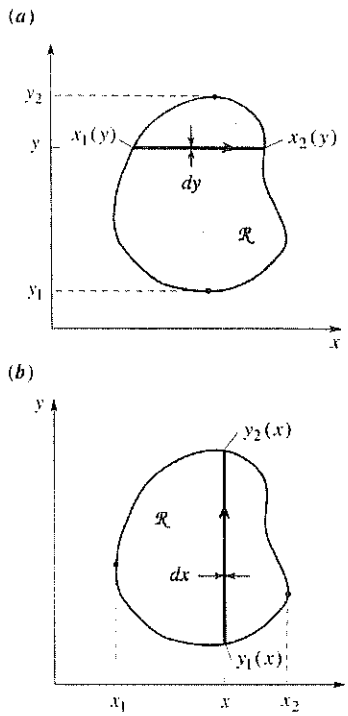


Figure 3. Iterated integrals.

Are the iterated integrals in (6) and (7) necessarily equal? This question is analogous to the one regarding *differentiation*: for example, are the mixed partial derivatives f_{xy} and f_{yx} equal? (We suggest that before continuing you review Theorem 13.3.1 and the Closure in Section 13.3, about interchanging the order of two limit processes.)

Returning to the question at hand, we state without proof that if $f(x,y)$ is continuous on $\mathcal{R}$, then all three quantities are equal: the double integral (1), the iterated integral (6), and the "inverted" iterated integral (7),*

*See J. E. Marsden and A. J. Tromba, *Vector Calculus* (San Fransisco: W. H. Freeman, 1976), Chap.5, especially *Fubini's theorem* for rectangular regions on p.224 and the rectangularization idea on p.230. Note (above Example 1) that we have stated that all three integrals are equal if f is continuous on $\mathcal{R}$. If f is not continuous, it is possible for the iterated integrals in (6) and (7) to be unequal (or for one or both to fail to exist), and it is even possible for them to be equal, but not equal to the double integral $\iint_{\mathcal{R}} f(x,y)\,dA$. For such an example, see T. M. Apostol, *Mathematical Analysis* (Reading, MA: Addison–Wesley, 1957), Exer. 10-9.

$$\iint_{\mathcal{R}} f(x,y)\,dA = \int_{y_1}^{y_2}\int_{x_1(y)}^{x_2(y)} f(x,y)\,dx\,dy = \int_{x_1}^{x_2}\int_{y_1(x)}^{y_2(x)} f(x,y)\,dy\,dx. \quad (8)$$

EXAMPLE 1. To illustrate the truth of (8), let us evaluate

$$I = \iint_{\mathcal{R}} xy^2\,dA, \quad (9)$$

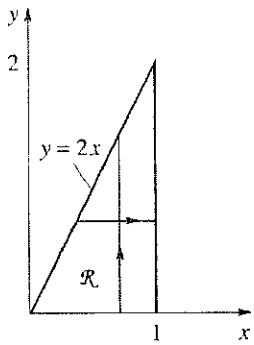


Figure 4. The region $\mathcal{R}$ in (9).

where $\mathcal{R}$ is the triangular region depicted in Fig. 4. First, observe that $f(x,y) = xy^2$ is indeed continuous on $\mathcal{R}$. If we integrate first on x, then $x_1(y) = y/2$, $x_2(y) = 1$, $y_1 = 0$, and $y_2 = 2$ so

$$I = \int_0^2 \left\{ \int_{y/2}^1 xy^2\,dx \right\} dy = \int_0^2 \left\{ \frac{x^2y^2}{2}\bigg|_{y/2}^1 \right\} dy$$
$$= \int_0^2 \left(\frac{y^2}{2} - \frac{y^4}{8} \right) dy = \frac{8}{6} - \frac{32}{40} = \frac{8}{15}. \quad (10)$$

Alternatively, integrating first on y, as in (7), we have

$$I = \int_0^1 \left\{ \int_0^{2x} xy^2\,dy \right\} dx = \int_0^1 \left\{ \frac{xy^3}{3}\bigg|_0^{2x} \right\} dx = \frac{8}{15}, \quad (11)$$

which does agree with (10), in accordance with (8). ∎

EXAMPLE 2. Evaluate the iterated integral

$$I = \int_2^4 \int_{y/2}^{\sqrt{y}} e^{y/x}\,dx\,dy. \quad (12)$$

In Example 1 we were given a picture of $\mathcal{R}$, from which we deduced the integration limits $x_1(y) = y/2$, $x_2(y) = 1$, $y_1 = 0$, $y_2 = 0$ in (10). In the present example we are given the limits. Fine, but it is difficult to get started because $\exp(y/x)$ is a formidable function of x to integrate. It is more promising to invert the order of integration so we can integrate first with respect to y because $\exp(y/x)$ is a simple function of y. Thus, consider instead

$$I = \int_?^? \int_?^? e^{y/x}\,dy\,dx. \quad (13)$$

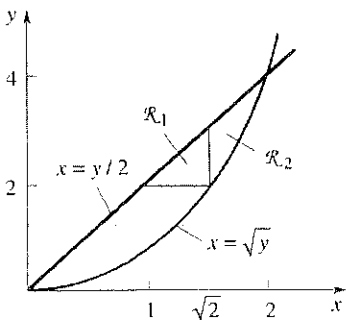


Figure 5. Region of integration.

But what are the new limits of integration? [They are *not* merely $\int_{y/2}^{\sqrt{y}}\int_2^4$, i.e., the reverse of the limits in (12).] To obtain the limits in (13) we need a picture of the region $\mathcal{R}$. That picture can be inferred from the limits in (12) and is the shaded region shown in Fig. 5. To integrate first on y and then on x we must break $\mathcal{R}$ into the subregions because the bottom boundaries of $\mathcal{R}_1$ and $\mathcal{R}_2$ are different, $y = 2$ for $\mathcal{R}_1$ and $y = x^2$ for $\mathcal{R}_2$. Thus,

$$I = \iint_{\mathcal{R}_1} e^{y/x}\,dy\,dx + \iint_{\mathcal{R}_2} e^{y/x}\,dy\,dx$$

$$= \int_1^{\sqrt{2}} \int_2^{2x} e^{y/x}\, dy\, dx + \int_{\sqrt{2}}^{2} \int_{x^2}^{2x} e^{y/x}\, dy\, dx$$
$$= \int_1^{\sqrt{2}} \left\{ xe^{y/x}\Big|_2^{2x} \right\} dx + \int_{\sqrt{2}}^{2} \left\{ xe^{y/x}\Big|_{x^2}^{2x} \right\} dx.$$

Thus,

$$I = \frac{e^2}{2} + (\sqrt{2} - 1)e^{\sqrt{2}} - \int_1^{\sqrt{2}} xe^{2/x}\, dx. \tag{14}$$

The remaining integral in (14) can be evaluated in terms of the tabulated exponential integral function, but let us not pursue that point. At the very least we have succeeded, by inverting the order of integration, in reducing I from a double integral to a single integral.

COMMENT. Be sure to see that the functional nature of the integrand was not relevant in our search for the new limits needed in (13). Rather, the original limits [in (12)] implied the region $\mathcal{R}$ (shaded in Fig. 6), and the region $\mathcal{R}$, in turn, implied the new limits. ∎

15.3.2. Triple integrals. Next, consider the triple integral

$$\iiint_{\mathcal{R}} f(x, y, z)\, dV. \tag{15}$$

We assume that $\mathcal{R}$ is a closed, bounded region in 3-space, that $f(x, y, z)$ is defined in $\mathcal{R}$, and that the boundary $\mathcal{S}$ of $\mathcal{R}$ is a piecewise smooth surface. By a **smooth surface** we mean one that has a continuously turning normal,* and by a **piecewise smooth surface** we mean one that consists of at most a finite number of smooth surfaces arranged edge to edge.

To define the triple integral (15), we partition $\mathcal{R}$ into a cellular structure analogous to that shown in Fig. 1. This time, however, the cells are rectangular prisms rather than rectangles. Corresponding to this N-cell partition we form the Riemann sum

$$\sum_{j=1}^{N} f(x_j, y_j, z_j)\Delta V_j, \tag{16}$$

where $\Delta V_j = \Delta x_j \Delta y_j \Delta z_j$ is the volume of the jth cell. Then we define a finer partition, compute its Riemann sum, and repeat the procedure over and over (in principle; one does not actually carry out this procedure in practice). If the sequence of values of the Riemann sum thus generated converges to a limit that limit is, by definition, the **triple integral** (15). That is,

$$\iiint_{\mathcal{R}} f(x, y, z)\, dV \equiv \lim_{|p| \to 0} \sum f(x_j, y_j, z_j)\Delta V_j \tag{17}$$

*For present purposes, one's intuitive notion of the normal to a surface should suffice; the term is defined in Section 15.4.

if the limit on the right exists, where $|p|$ is the *norm* of the partition, namely, the greatest of the dimensions $\Delta x_j, \Delta y_j, \Delta z_j$ for $j = 1, \ldots, N$. It can be shown that the limit (17) will indeed exist if f is continuous in the closed region $\mathcal{R}$, that condition being sufficient but not necessary.

We state, without writing them out, that properties (4a) to (4c) carry over in the obvious way to the case of triple integrals. For example, the **mean value theorem** for triple integrals states that if $f(x, y, z)$ is continuous throughout the closed region $\mathcal{R}$, then there exists at least one point (x_0, y_0, z_0) in $\mathcal{R}$ such that

$$\boxed{\iiint\limits_{\mathcal{R}} f(x, y, z)\, dV = f(x_0, y_0, z_0)V,} \tag{18}$$

where

$$V = \iiint\limits_{\mathcal{R}} dV \tag{19}$$

is the **volume** of $\mathcal{R}$.

To evaluate a triple integral analytically, we treat it as an iterated integral and integrate on one variable at a time.

EXAMPLE 3. Evaluate

$$I = \iiint\limits_{\mathcal{R}} yz^2\, dV, \tag{20}$$

where $\mathcal{R}$ is the closed region bounded by the planes $x = 0$, $y = 0$, $z = 0$, and $3x + 2y + 6z = 6$. With $dV = dx\, dy\, dz$, we write

$$I = \int_?^? \int_?^? \int_?^? yz^2\, dx\, dy\, dz, \tag{21}$$

but how do we obtain the integration limits? It is helpful to re-express (21) as

$$I = \int_?^? \left\{ \int_?^? \int_?^? yz^2\, dx\, dy \right\} dz \tag{22}$$

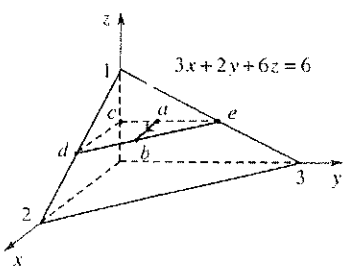


Figure 6. The region $\mathcal{R}$.

and to attend first to the bracketed iterated integral on x and y. In that iterated integral z is held fixed, at some value between $z = 0$ and $z = 1$ because we are not yet integrating on z. Thus, the x, y integration is being carried out on a horizontal slice such as the triangle cde (Fig. 6). In turn, within that iterated integral we are integrating on x first, with y held fixed at some point between c and e. In other words, first we integrate on x from a ($x = 0$) to b ($x = 2 - \frac{2}{3}y - 2z$), then we sweep the line ab from cd ($y = 0$) to e ($y = 3 - 3z$). Finally, we sweep the triangle cde from $z = 0$ to $z = 1$. Thus, we have

$$I = \int_0^1 \int_0^{3-3z} \int_0^{2-\frac{2}{3}y-2z} yz^2\, dx\, dy\, dz$$

$$
\begin{aligned}
&= \int_0^1 \int_0^{3-3z} yz^2 x \Big|_{x=0}^{x=2-\frac{2}{3}y-2z} dy\, dz \\
&= \int_0^1 \int_0^{3-3z} yz^2 \left(2 - \frac{2}{3}y - 2z\right) dy\, dz \\
&= \int_0^1 z^2 \left(y^2 - \frac{2}{9}y^3 - y^2 z\right) \Big|_{y=0}^{y=3-3z} dz \\
&= \int_0^1 z^2 \left[9(1-z)^2 - 6(1-z)^3 - 9(1-z)^2 z\right] dz = \frac{1}{20}. \qquad (23)
\end{aligned}
$$

In case the upper y limit ($y = 3 - 3z$) is not clear, observe that the point e is on the plane $3x + 2y + 6z = 6$ at $x = 0$, hence $y + 3z = 3$ and $y = 3 - 3z$.

COMMENT 1. The limits are *not* $\int_0^1 \int_0^3 \int_0^2$. The latter would be correct if $\mathcal{R}$ were a rectangular prism, but it is not. The key to determining the correct limits lies in carrying out the sketch (Fig. 6), including the triangle cde and the line ab.

COMMENT 2. How do we know that x varies from 0 (at a) to $2 - \frac{2}{3}y - 2z$ (at b) and not vice versa, and similarly for the y and z limits? The key is that $dV = dx\, dy\, dz$ must be positive because it is a volume element. We can ensure that each dV is positive by choosing the limits as we have, so that each dx, dy, and dz is positive; for instance, if the z limits had been chosen as 1 to 0, instead of 0 to 1, then each dz would be *negative*.

COMMENT 3. As for double integrals, if the integrand is continuous on the region $\mathcal{R}$, then the triple integral is equal to the iterated integral, with the iteration carried out in any order. That is,

$$
\begin{aligned}
\iiint_{\mathcal{R}} f\, dV &= \int_{z_1}^{z_2} \int_{y_1(z)}^{y_2(z)} \int_{x_1(y,z)}^{x_2(y,z)} f\, dx\, dy\, dz \\
&= \int_{z_1}^{z_2} \int_{x_1(z)}^{x_2(z)} \int_{y_1(x,z)}^{y_2(x,z)} f\, dy\, dx\, dz = \cdots, \qquad (24)
\end{aligned}
$$

and so on, where the three dots are included because there are actually six possible integration sequences: xyz, yxz, zxy, xzy, zyx, and yzx. Integrating in the sequence zyx, for instance, we would have

$$
I = \int_0^2 \int_0^{3-\frac{3}{2}x} \int_0^{1-\frac{1}{2}x-\frac{1}{3}y} yz^2\, dz\, dy\, dx, \qquad (25)
$$

where the limits are obtained from Fig. 6. Evaluation of the iterated integral in (25) once again gives $I = 1/20$, as in (23).

COMMENT 4. Observe the pattern in (24); for instance, consider the final iterated integral, in which we integrate first on y, then on x, and then on z. The limits can depend only on the variables that have not yet been integrated. Thus, the y limits can, in general, depend on x and z, the x limits can depend only on z, and the z limits must be constants. ∎

Closure. This section is included as a brief review of double and triple integrals. Of special importance are the properties (4a)–(4c) and the corresponding ones for

triple integrals, the distinction between double and triple integrals and their iterated integral counterparts, and the determination of integration limits.

Computer software. Besides single integrals, the *Maple* int command can be used to evaluate multiple integrals. For instance, to evaluate the integral (9), enter

$$\text{int}(\text{int}(x * y\hat{}2,\ x = y/2..1),\ y = 0..2);$$

and return. The result is $8/15$, as obtained analytically in Example 1.

EXERCISES 15.3

Exercises 1–9 correspond to Section 15.3.1, and the remainder correspond to Section 15.3.2.

1. Evaluate each of the following iterated integrals. Then integrate again, with the order of integration reversed, and show that the same result is obtained.

(a) $\int_0^1 \int_0^y y^2\,dx\,dy$ (b) $\int_0^1 \int_x^{2x} x^3 y\,dy\,dx$

(c) $\int_0^2 \int_y^2 xe^{x+y}\,dx\,dy$ (d) $\int_1^2 \int_0^y \sin(x-y)\,dx\,dy$

(e) $\int_0^1 \int_y^{(2y+1)/4} (x+y)^8\,dx\,dy$ (f) $\int_0^1 \int_{-y}^{2-y} x\,dx\,dy$

(g) $\int_{-1}^3 \int_{-x}^x x^2\,dy\,dx$ (h) $\int_1^4 \int_2^{\sqrt{x}} x^2\,dy\,dx$

2. Show that $\int_0^T \int_0^t x(\tau)\,d\tau\,dt$ can be reduced to the single integral $\int_0^T (T-\tau)x(\tau)\,d\tau$ (T = constant). HINT: See the Comment at the end of Example 2.

3. Show that

(a) $\int_0^6 \int_{y/2}^3 \frac{1}{x} e^{y/x}\,dx\,dy = 3(e^2-1)$

(b) $\int_4^{16} \int_{y/2}^{\sqrt{y}} \frac{1}{x} e^{y/x}\,dx\,dy = 7e^2 - e^4 - \int_4^8 e^{16/x}\,dx$

(c) $\int_0^1 \int_{y/2}^y \sin(x^2 y)\,dx\,dy = \int_0^{1/2} \frac{\cos(x^3) - \cos(2x^3)}{x^2}\,dx + \int_{1/2}^1 \frac{\cos(x^3) - \cos(x^2)}{x^2}\,dx$

(d) $\int_0^{\pi/2} \int_x^{\pi/2} \frac{\sin y}{y}\,dy\,dx = 1$

4. (*Mass and center of gravity*) Let $\sigma(x,y)$ be the density of a distribution of mass over a region $\mathcal{R}$ in the x,y plane [i.e., $\sigma(x,y)$ is the mass per unit area at (x,y)]. Then the x,y coordinates of the **center of gravity** are defined as

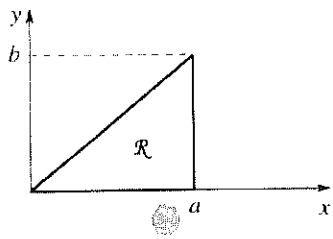
$$x_c = \frac{1}{M}\iint_{\mathcal{R}} x\sigma(x,y)\,dA, \qquad y_c = \frac{1}{M}\iint_{\mathcal{R}} y\sigma(x,y)\,dA, \tag{4.1}$$

where

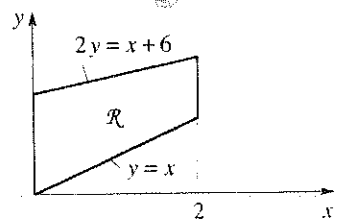
$$M = \iint_{\mathcal{R}} \sigma(x,y)\,dA \tag{4.2}$$

is the total mass. Evaluate x_c and y_c in each case. In parts (a) to (d), $\sigma(x,y)$ is a constant, say σ.

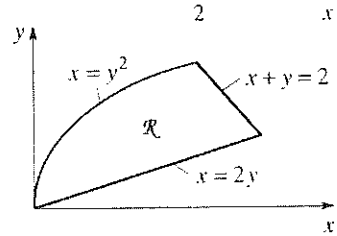
(a)



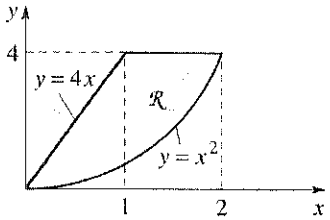
(b)



(c)



(d)



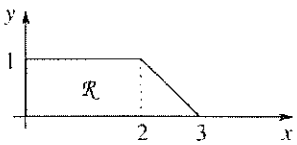
(e) Same $\mathcal{R}$ as in part (a), but $\sigma(x, y) = 1 + x$
(f) Same $\mathcal{R}$ as in part (a), but $\sigma(x, y) = y$
(g) Same $\mathcal{R}$ as in part (c), but $\sigma(x, y) = x$
(h) Same $\mathcal{R}$ as in part (c), but $\sigma(x, y) = y$

5. (*Moments of inertia*) Let $\sigma(x, y)$ be the density of a distribution of mass over a region $\mathcal{R}$ in the x, y plane; i.e., $\sigma(x, y)$ is the mass per unit area at (x, y). Then the **moments of inertia** I_x and I_y, about the x, y axes, respectively, are defined as

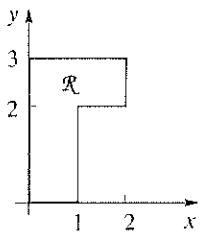
$$I_x = \iint_{\mathcal{R}} y^2 \sigma(x, y)\, dA, \qquad I_y = \iint_{\mathcal{R}} x^2 \sigma(x, y)\, dA. \tag{5.1}$$

Evaluate I_x and I_y in each case. In parts (a) and (b), $\sigma(x, y)$ is a constant, say σ.

(a)



(b)



(c) $\mathcal{R}$, σ as in Exercise 4(a)
(d) $\mathcal{R}$, σ as in Exercise 4(b)
(e) $\mathcal{R}$, σ as in Exercise 4(c)
(f) $\mathcal{R}$, σ as in Exercise 4(d)
(g) $\mathcal{R}$, σ as in Exercise 4(e)
(h) $\mathcal{R}$, σ as in Exercise 4(f)

6. The **exponential integral** function $E_1(t)$ is defined by the formula

$$\boxed{E_1(t) = \int_t^{\infty} \frac{e^{-\tau}}{\tau}\, d\tau, \quad (t > 0)} \tag{6.1}$$

and is tabulated in the literature, just as are $\sin t, e^t$, and so on. Evaluate the integral $\int_0^x E_1(t)\, dt$, and show that it equals $1 - e^{-x} + xE_1(x)$. (See the index for applications of the exponential integral function within this text.)

7. If

$$f(x, y) = \begin{cases} 1, & 0 \le y \le 2x, \quad 0 \le x \le 1 \\ 0, & \text{elsewhere,} \end{cases}$$

show that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{u} f(x, y)\, dx\, dy = \begin{cases} 0, & u \le 0 \\ u^2, & 0 < u < 1 \\ 1, & u \ge 1. \end{cases}$$

8. Use computer software to evaluate the integral in (14), and thus obtain a numerical value for I.

9. (a)–(h) Use computer software to evaluate the iterated integral given in the corresponding part of Exercise 1.

10. Evaluate each of the following iterated integrals.

(a) $\int_1^2 \int_0^{3x} \int_{2y}^{x} dz\, dy\, dx$
(b) $\int_{-1}^1 \int_0^z \int_2^3 z^2 \sin(yz)\, dx\, dy\, dz$
(c) $\int_0^{\pi} \int_0^{2y} \int_0^{y+z} \cos(x + y)\, dx\, dz\, dy$
(d) $\int_{-1}^3 \int_0^{z^2} \int_{\pi y}^0 \sin \frac{x}{y}\, dx\, dy\, dz$
(e) $\int_0^2 \int_0^{4z} \int_0^{2-z} e^{x+y}\, dx\, dy\, dz$
(f) $\int_0^1 \int_0^1 \int_0^y z^3 y^2 e^{xyz}\, dx\, dy\, dz$

11. Evaluate

$$I = \int_0^1 \int_0^{\pi z} \int_{y/\pi}^{z} \sin \frac{y}{x}\, dx\, dy\, dz.$$

HINT: The x integration looks quite difficult, so try inverting the order of integration to the form

$$I = \int_0^1 \int_?^? \int_?^? \sin \frac{y}{x}\, dy\, dx\, dz.$$

To determine the "?" integration limits we do *not* need to draw the full three-dimensional region of integration. Rather, a look at the x, y plane (with z regarded as a *fixed* value, somewhere between its limits of 0 and 1) will suffice since it is only the x and y integrations that we are interchanging.

12. Evaluate the following integrals. The idea, in each case, is the same as discussed in Exercise 4.

(a) $\mathcal{R}$: the tetrahedron with vertices at $(1, 0, 0)$, $(0, 2, 0)$, $(0, 0, 1)$, $(0, 0, 0)$; $\sigma =$ constant. Evaluate x_c.

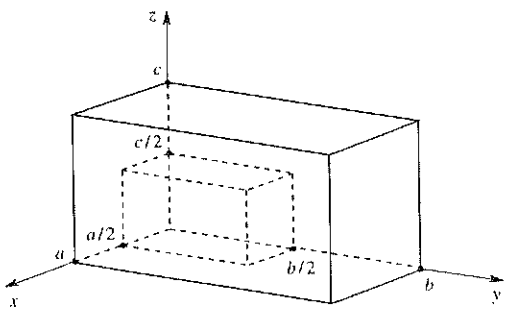
(b) $\mathcal{R}$: the tetrahedron with vertices at $(1,0,1)$, $(0,0,1)$, $(0,1,1)$, $(0,0,0)$; $\sigma = x$. Evaluate z_c.
(c) $\mathcal{R}$: the finite region bounded by $z = 4 - x^2 - y^2$, and the x, y plane; $\sigma =$ constant. Evaluate z_c.
(d) $\mathcal{R}$: the finite region bounded by $z = 1 - x^2 - y^2$, the x, z plane, and the x, y plane; $\sigma =$ constant. Evaluate z_c.
(e) Repeat part (d), but with $\sigma = z$.

13. (*Moments of inertia*) Let $\sigma(x,y,z)$ be the density of a distribution of a mass over a region $\mathcal{R}$ in x, y, z space [i.e., $\sigma(x,y,z)$ is the mass per unit volume at (x,y,z)]. Then the **moments of inertia** I_x, I_y, and I_z, about the x, y, z axes, respectively, are defined as

$$I_x = \iiint_{\mathcal{R}} (y^2 + z^2)\,\sigma(x,y,z)\,dV$$
$$I_y = \iiint_{\mathcal{R}} (x^2 + z^2)\,\sigma(x,y,z)\,dV$$
$$I_z = \iiint_{\mathcal{R}} (x^2 + y^2)\,\sigma(x,y,z)\,dV.$$

In each case evaluate I_x.

(a) $\mathcal{R}$, σ as given in Exercise 12(a).
(b) $\mathcal{R}$ the rectangular prism $0 < x < a, 0 < y < b, 0 < z < c$; $\sigma =$ constant. Show that $I_x = \sigma abc(b^2 + c^2)/3$.
(c) $\mathcal{R}$ the rectangular prism $0 < x < a, 0 < y < b, 0 < z < c$, but with the corner $0 < x < a/2, 0 < y < b/2, 0 < z < c/2$ cut out, as shown in the figure; $\sigma =$ constant. HINT: You may use the result given in part (b).



14. Determine the "?" integration limits. HINT: You can use the given limits to infer the region $\mathcal{R}$. Sketching $\mathcal{R}$, you can then determine the new limits. However, that method might involve a challenging three-dimensional sketch and would not work at all for a quadruple integral, say, because $\mathcal{R}$ would then be four-dimensional. Thus, we suggest that you use one or more two-dimensional pictures. In part (a), for instance, only the y and z integrations are being interchanged so it suffices to consider the y, z plane. In part (b) you can accomplish the desired change in the order of integration by using two successive interchanges, as follows: $xyz \to yxz \to yzx$. The first interchange requires a look at the x, y plane only, and the second interchange requires a look at the x, z plane only.

(a) $$\int_0^1 \int_0^z \int_0^z f(x,y,z)\,dx\,dy\,dz = \int_?^? \int_?^? \int_?^? f(x,y,z)\,dx\,dz\,dy$$

(b) $$\int_0^1 \int_0^z \int_0^{2z} f(x,y,z)\,dx\,dy\,dz = \int_?^? \int_?^? \int_?^? f(x,y,z)\,dy\,dz\,dx$$

(c) $$\int_0^1 \int_0^z \int_0^z f(x,y,z)\,dx\,dy\,dz = \int_?^? \int_?^? \int_?^? f(x,y,z)\,dz\,dy\,dx$$

(d) $$\int_0^2 \int_0^z \int_z^2 f(x,y,z)\,dx\,dy\,dz = \int_?^? \int_?^? \int_?^? f(x,y,z)\,dy\,dz\,dx$$

(e) $$\int_0^4 \int_{z/2}^{\sqrt{z}} \int_0^{y+z} f(x,y,z)\,dx\,dy\,dz = \int_?^? \int_?^? \int_?^? f(x,y,z)\,dx\,dz\,dy$$

(f) $$\int_4^6 \int_1^3 \int_0^2 f(x,y,z)\,dx\,dy\,dz = \int_?^? \int_?^? \int_?^? f(x,y,z)\,dz\,dy\,dx$$

15. (a)–(f) Use computer software to evaluate the integral in the corresponding part of Exercise 10.

15.4 Surfaces

Aiming at a consideration of surface integrals in Section 15.5, we begin by discussing the parametric representation of surfaces, the tangent plane, and the normal.

15.4.1. Parametric representation of surfaces. To begin, recall that a space curve may be represented parametrically by equations

$$x = x(u), \quad y = y(u), \quad z = z(u) \tag{1}$$

over some interval of the parameter u. (Previously we used τ as our parameter; here we prefer u.) Or, in vector form, we express the position vector $\mathbf{R}(u)$, from the origin of the reference x, y, z coordinate system to a point on the curve, as

$$\mathbf{R}(u) = x(u)\hat{\mathbf{i}} + y(u)\hat{\mathbf{j}} + z(u)\hat{\mathbf{k}}. \tag{2}$$

Similarly, it is natural to expect a *two*-parameter family of curves

$$x = x(u, v), \quad y = y(u, v), \quad z = z(u, v) \tag{3}$$

or

$$\mathbf{R}(u, v) = x(u, v)\hat{\mathbf{i}} + y(u, v)\hat{\mathbf{j}} + z(u, v)\hat{\mathbf{k}}, \tag{4}$$

over some intervals of the real parameters u and v, to define a surface. That is, for each fixed v the parametrization (3) [or (4)] defines a curve (in general). Thus, as we vary v we produce a family of such curves which, in general, will generate a surface.

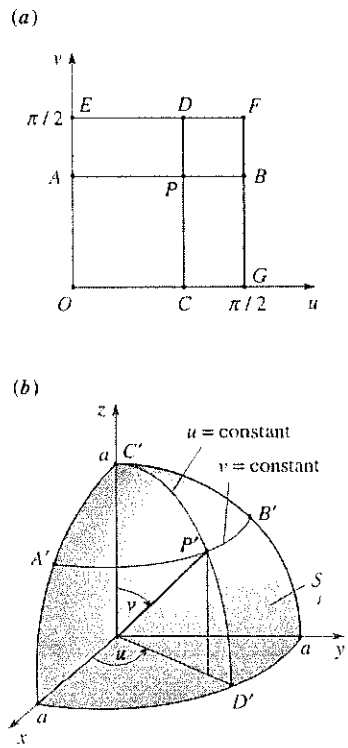


Figure 1. The surface S generated by (5).

EXAMPLE 1. Let

$$\begin{aligned} x &= a \sin v \cos u, \\ y &= a \sin v \sin u, \\ z &= a \cos v, \end{aligned} \tag{5}$$

where $0 \le u \le \pi/2$, $0 \le v \le \pi/2$, and a is a positive constant. Squaring and adding these three equations gives $x^2 + y^2 + z^2 = a^2$, which represents a spherical surface of radius a, centered at the origin. In fact, a comparison of (5) with (26a) in Section 14.6 reveals that a, v, u are actually the spherical coordinates ρ, ϕ, θ, respectively, in disguise. Since $\rho = a =$ constant, the surface is spherical. More precisely, it is one-eighth of a spherical surface as shown in Fig. 1b, where it is denoted as S. Thus, in this example the parameters u and v happen to bear a geometrical significance; they are the spherical polar angles shown in Fig. 1b.

If v is fixed and u is varied over the line AB in Fig. 1a, then (5) generates the curve $A'B'$ shown in Fig. 1b. Similarly, if u is fixed and v is varied over CD, then (5) generates the curve $C'D'$. If we think of S as part of the earth's globe, with C' as the north pole, then the $v =$ constant curves are the lines of latitude and the $u =$ constant curves are the lines of longitude.

COMMENT 1. It is common to refer to u, v as **curvilinear coordinates** of the generated surface S. In this example the network of lines of latitude (the $v =$ constant curves) and

longitude (the u = constant curves) forms a coordinate system on the surface of the sphere just as a rectilinear network of x = constant and y = constant lines forms a coordinate system on the x, y plane. We call the u = constant and v = constant curves on $\mathcal{S}$ the **coordinate curves**. In particular, along the v = constant curve only u is changing so we call the v = constant curves the **u coordinate curves**. Similarly, we call the u = constant curves the **v coordinate curves**. For instance, in Fig. 1b $A'B'$ is a u coordinate curve, with u increasing from 0 at A' to $\pi/2$ at B', and $C'D'$ is a v coordinate curve, with v increasing from 0 at C' to $\pi/2$ at D'. In Fig. 2 we show many such u and v coordinate curves so as to more clearly see the u, v coordinate "mesh" on the spherical surface.

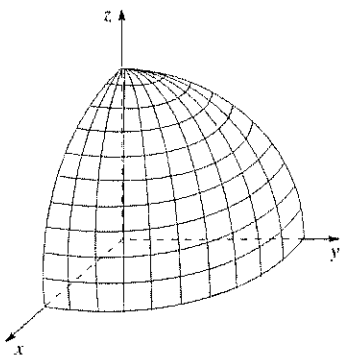


Figure 2. u, v coordinate mesh on $\mathcal{S}$.

COMMENT 2. If the surface $\mathcal{S}$ can be expressed in the form $z = f(x, y)$, then we can parametrize $\mathcal{S}$ by $x = u$, $y = v$, and $z = f(u, v)$. In the present case,

$$z = \sqrt{a^2 - x^2 - y^2} \tag{6}$$

so an alternative parametrization, to that given by (5), is

$$x = u, \qquad y = v, \qquad z = \sqrt{a^2 - u^2 - v^2}. \tag{7}$$

In that case the u, v coordinate curves through a given point P on $\mathcal{S}$ are as shown in Fig. 3a, and the overall coordinate mesh is as shown in Fig. 3b. ∎

15.4.2. Tangent plane and normal. Let a given vector function $\mathbf{R}(u, v)$, for some u and v intervals, define a surface $\mathcal{S}$. Since v is held constant in the partial derivative $\mathbf{R}_u$, it follows that $\mathbf{R}_u$ (if it exists and is nonzero) is a tangent vector to $\mathcal{S}$, along the u coordinate curve (and in the increasing u direction) as shown in Fig. 4. Similarly, $\mathbf{R}_v$ (if it exists and is nonzero) is a tangent vector to $\mathcal{S}$ along the v coordinate curve (and in the increasing v direction).

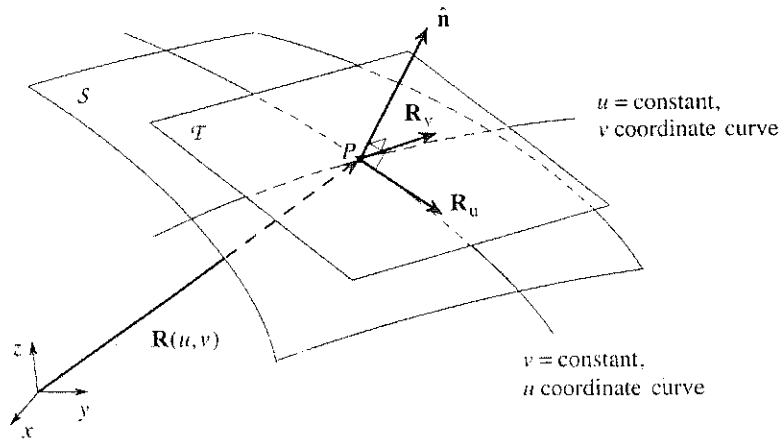


Figure 4. Tangent plane and normal.

From Fig. 4 we see that a **normal** vector to $\mathcal{S}$, at P, can be obtained from $\mathbf{R}(u, v)$ as

$$\mathbf{n} = \mathbf{R}_u \times \mathbf{R}_v \tag{8}$$

or

$$\boxed{\hat{\mathbf{n}} = \frac{\mathbf{R}_u \times \mathbf{R}_v}{\|\mathbf{R}_u \times \mathbf{R}_v\|},} \tag{9}$$

provided that

$$\mathbf{R}_u \times \mathbf{R}_v \neq 0 \tag{10}$$

at P, that is, provided that $\mathbf{R}_u$ and $\mathbf{R}_v$ are nonzero and noncollinear (i.e., provided that $\mathbf{R}_u$ and $\mathbf{R}_v$ are linearly independent). In Example 1, $\mathbf{R}_u$ and $\mathbf{R}_v$ happen to be not only linearly independent but even orthogonal at each point on $\mathcal{S}$, but in general the u and v coordinate curves need not be orthogonal.

If (10) is satisfied, then $\mathbf{R}_u$ and $\mathbf{R}_v$ determine a plane, the **tangent plane** $\mathcal{T}$ to $\mathcal{S}$ at P. For if $u = u(\tau)$, $v = v(\tau)$ are parametric equations of any curve $\mathcal{C}$ through P, in $\mathcal{S}$, then $d\mathbf{R}/d\tau$ is tangent to $\mathcal{C}$ at P. But

$$\frac{d\mathbf{R}}{d\tau} = \frac{d}{d\tau}\mathbf{R}(u(\tau), v(\tau)) = \mathbf{R}_u u' + \mathbf{R}_v v' \tag{11}$$

is in $\mathcal{T}$ because it is a linear combination of $\mathbf{R}_u$ and $\mathbf{R}_v$. (For the second equality, which is chain differentiation, to hold, let us assume that $\mathbf{R}_u$ and $\mathbf{R}_v$ are continuous functions of u and v at P and, of course, that u and v are differentiable functions of τ at that point.)

Given $\mathbf{R}(u, v)$, and a point P on $\mathcal{S}$, how can we obtain the equation of the tangent plane $\mathcal{T}$ to $\mathcal{S}$ at P? Let $\mathbf{R}_p = x_p\hat{\mathbf{i}} + y_p\hat{\mathbf{j}} + z_p\hat{\mathbf{k}}$ and $\mathbf{R} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$ be position vectors to P, and to any point on $\mathcal{T}$, respectively (Fig. 5). Then

$$(\mathbf{R} - \mathbf{R}_p) \cdot \hat{\mathbf{n}} = 0, \tag{12a}$$

where $\hat{\mathbf{n}}$ can be computed from (9) and is a known vector, as is $\mathbf{R}_p$. If we denote $\hat{\mathbf{n}} = a\hat{\mathbf{i}} + b\hat{\mathbf{j}} + c\hat{\mathbf{k}}$, say, then (12a) becomes

$$\boxed{a(x - x_p) + b(y - y_p) + c(z - z_p) = 0} \tag{12b}$$

or

$$\boxed{ax + by + cz = d,} \tag{12c}$$

where $d = ax_p + by_p + cz_p$. Equations (12) are equivalent, each one being the equation of the tangent plane at P; (12b) and (12c) may well look familiar from the calculus.

Naturally, the normal $\hat{\mathbf{n}}$ can be determined only to within a factor of ± 1. For instance, if $\hat{\mathbf{n}} = (\hat{\mathbf{i}} + \hat{\mathbf{j}} - 2\hat{\mathbf{k}})/\sqrt{6}$ is a normal to $\mathcal{S}$ at some particular point P, then so is $\hat{\mathbf{n}} = -(\hat{\mathbf{i}} + \hat{\mathbf{j}} - 2\hat{\mathbf{k}})/\sqrt{6}$. The two possible $\hat{\mathbf{n}}$ vectors at P determine a line which we call the **normal line** at P. A unique tangent plane exists at P if and only if a unique normal line exists there, and the condition (10) is *sufficient, but not necessary* (see Exercise 10), for the existence of a unique normal line.

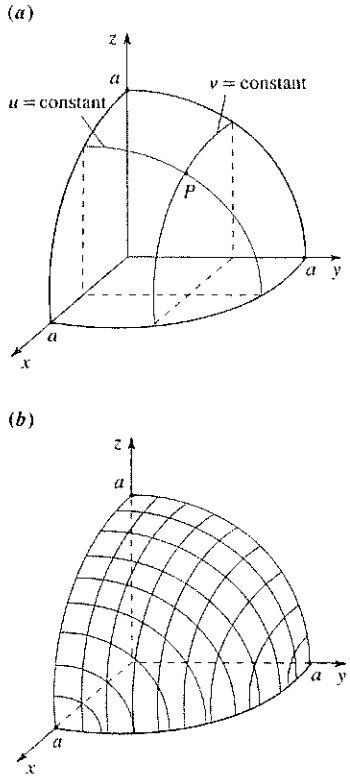


Figure 3. Using (7), instead.

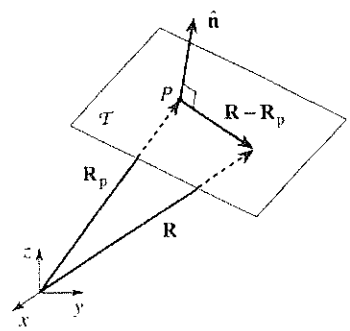


Figure 5. Obtaining equation of tangent plane.

EXAMPLE 2. For the surface $\mathcal{S}$ defined parametrically by

$$x = u + v,$$

$$y = u - v^2, \qquad (13)$$
$$z = u^2 + 3v^4,$$

find the tangent plane to $\mathcal{S}$ at $u = 0$, $v = 1$, that is, at the point $(x, y, z) = (1, -1, 3)$. Since $\mathbf{R} = (u + v)\hat{\mathbf{i}} + (u - v^2)\hat{\mathbf{j}} + (u^2 + 3v^4)\hat{\mathbf{k}}$, we have

$$\mathbf{R}_u = \hat{\mathbf{i}} + \hat{\mathbf{j}} + 2u\hat{\mathbf{k}} = \hat{\mathbf{i}} + \hat{\mathbf{j}},$$
$$\mathbf{R}_v = \hat{\mathbf{i}} - 2v\hat{\mathbf{j}} + 12v^3\hat{\mathbf{k}} = \hat{\mathbf{i}} - 2\hat{\mathbf{j}} + 12\hat{\mathbf{k}}$$

at $u = 0$ and $v = 1$ so (9) gives

$$\hat{\mathbf{n}} = \frac{4\hat{\mathbf{i}} - 4\hat{\mathbf{j}} - \hat{\mathbf{k}}}{\sqrt{33}}. \qquad (14)$$

Thus, (12b) gives

$$\frac{4}{\sqrt{33}}(x - 1) - \frac{4}{\sqrt{33}}(y + 1) - \frac{1}{\sqrt{33}}(z - 3) = 0 \qquad (15)$$

or

$$4x - 4y - z = 5 \qquad (16)$$

as the equation of the desired tangent plane. ∎

EXAMPLE 3. The relation

$$x^2 + y^2 + z^2 + z^4 = 1 \qquad (17)$$

defines a surface that is somewhat of a sphere compressed at its north and south poles; its intercepts along the coordinate axes are ± 1 along each of the x and y axes, and ± 0.618 along the z axis. Find the equation of the tangent plane at the point $(x, y, z) = \left(\frac{1}{2}, \frac{1}{2}, \sqrt{\frac{\sqrt{3}-1}{2}}\right)$ $= (0.5, 0.5, 0.605)$ on $\mathcal{S}$.

To use (12b), say, we need to know a, b, c; that is, we need to know $\hat{\mathbf{n}}$. Since $\hat{\mathbf{n}}$ is given, by (9), in terms of $\mathbf{R}(u, v)$, our first step is to parametrize $\mathcal{S}$ in order to find $\mathbf{R}(u, v)$. We can solve (17) (with the help of the quadratic formula) for z, in the form $z = f(x, y)$, and then parametrize $\mathcal{S}$ by $x = u$, $y = v$, $z = f(u, v)$, but it is easier to solve (17) for x as

$$x = \sqrt{1 - y^2 - z^2 - z^4}, \qquad (18)$$

and hence to parametrize $\mathcal{S}$ by $y = u$, $z = v$, $x = \sqrt{1 - u^2 - v^2 - v^4}$. Then

$$\mathbf{R}(u, v) = \sqrt{1 - u^2 - v^2 - v^4}\,\hat{\mathbf{i}} + u\hat{\mathbf{j}} + v\hat{\mathbf{k}} \qquad (19)$$

so, with $u = 0.5$ and $v = 0.605$, (19) gives $\mathbf{R}_u = -\hat{\mathbf{i}} + \hat{\mathbf{j}}$ and $\mathbf{R}_v = -2.096\hat{\mathbf{i}} + \hat{\mathbf{k}}$. Next, (9) gives

$$\hat{\mathbf{n}} = 0.396\hat{\mathbf{i}} + 0.396\hat{\mathbf{j}} + 0.829\hat{\mathbf{k}}. \qquad (20)$$

Hence, $a = 0.396$, $b = 0.396$, $c = 0.829$, $x_p = 0.5$, $y_p = 0.5$, and $z_p = 0.605$ so (12b) gives

$$0.396x + 0.396y + 0.829z = 0.898 \qquad (21)$$

as the equation of the desired tangent plane.

COMMENT. Alternatively, if $\mathcal{S}$ is given in the form of a relation

$$f(x, y, z) = 0, \tag{22}$$

then the Taylor series expansion about (x_p, y_p, z_p) gives

$$\begin{aligned} f(x_p, y_p, z_p) + f_x(x_p, y_p, z_p)(x - x_p) + f_y(x_p, y_p, z_p)(y - y_p) \\ + f_z(x_p, y_p, z_p)(z - z_p) + \cdots = 0, \end{aligned} \tag{23}$$

where the three dots denote the terms of second order and higher. If we linearize by dropping those higher-order terms, and note that $f(x_p, y_p, z_p) = 0$ [because $f(x, y, z) = 0$ is the equation of $\mathcal{S}$, and (x_p, y_p, z_p) is on $\mathcal{S}$], then (23) becomes

$$\boxed{f_x(x_p, y_p, z_p)(x - x_p) + f_y(x_p, y_p, z_p)(y - y_p) + f_z(x_p, y_p, z_p)(z - z_p) = 0,} \tag{24}$$

which is the same as (12b), but expressed in terms of $f(x, y, z)$ rather than in terms of $\mathbf{R}(u, v)$. We leave it as an exercise to show that, with $f(x, y, z) = x^2 + y^2 + z^2 + z^4 - 1$, (24) agrees with (21). ∎

Closure. A surface $\mathcal{S}$ may be represented by the parametric equations $x = x(u, v)$, $y = y(u, v)$, $z = z(u, v)$ or, equivalently, by the position vector

$$\mathbf{R}(u, v) = x(u, v)\hat{\mathbf{i}} + y(u, v)\hat{\mathbf{j}} + z(u, v)\hat{\mathbf{k}}.$$

Accordingly, $\mathcal{S}$ is covered by a mesh of u and v coordinate curves (namely, the $v =$ constant and $u =$ constant curves, respectively), as shown in Figs. 2 and 3b for two different parametrizations. From $\mathbf{R}(u, v)$ we can obtain the normal to $\mathcal{S}$ from (9) (except at points on $\mathcal{S}$ at which $\mathbf{R}_u \times \mathbf{R}_v = \mathbf{0}$) and, knowing $\hat{\mathbf{n}}$, the condition (12a) gives us the equation of the tangent plane in the forms (12b) or (12c). Alternatively, if in place of $\mathbf{R}(u, v)$ we know $\mathcal{S}$ through a relation $f(x, y, z) = 0$, then Taylor-expanding f and linearizing gives the equation of the tangent plane in the form (24). Naturally, from a comparison of (12) and (24) we can see that the normal vector to $\mathcal{S}$ can be expressed, in terms of f, as

$$\hat{\mathbf{n}} = \frac{f_x\hat{\mathbf{i}} + f_y\hat{\mathbf{j}} + f_z\hat{\mathbf{k}}}{\sqrt{f_x^2 + f_y^2 + f_z^2}}, \tag{25}$$

where f_x, f_y, f_z are evaluated at the point (x_p, y_p, z_p) on $\mathcal{S}$.

EXERCISES 15.4

1. In Fig. 1 we show the curves on $\mathcal{S}$ that are the images of the lines AB and CD in the u, v plane, namely, $A'B'$ and $C'D'$, respectively. Sketch, trace, or photocopy the surface $\mathcal{S}$ in Fig. 1b and show the curve on $\mathcal{S}$ that is the image of the following line in the u, v plane.

(a) OE (b) EF (c) FG (d) OG
(e) the line $v = u$ from O to F
(f) the line $v = \pi/2 - u$ from E to G

2. Consider the parametrization

$$x = u, \quad y = u + v, \quad z = 0,$$

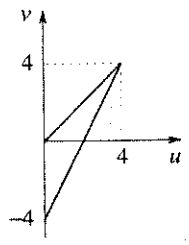
over $0 < u < 4, \quad 0 < v < 4$.

(a) Show $\mathcal{S}$ in a labeled sketch.
(b) Show the $u = 1, 2, 3$ and $v = 1, 2, 3$ coordinate curves on $\mathcal{S}$.
(c) Evaluate $\hat{\mathbf{n}}$ using (9). Are there any points on $\mathcal{S}$ at which (10) is not satisfied?

3. Consider the parametrization

$$x = u, \quad y = u - v, \quad z = 4 - 2u + v, \tag{3.1}$$

over the open region shown (shaded) below.



(a) Eliminating u and v from (3.1), obtain a nonparametric representation of the surface $\mathcal{S}$ [i.e., in the form $f(x, y, z) = 0$], and show $\mathcal{S}$ in a neat, labeled sketch.
(b) Show the $u = 1, 2, 3$ and $v = 1, 2, 3$ coordinate curves on $\mathcal{S}$.
(c) Evaluate $\hat{\mathbf{n}}$ using (9).

4. Consider the parametrization

$$x = u\cos v, \quad y = u\sin v, \quad z = 0,$$

over $0 \le u < 3, \quad 0 \le v < 2\pi$.

(a) Show $\mathcal{S}$ in a labeled sketch.
(b) Show the $u = 1, 2$ and $v = 0, \pi/4, \pi/2, 3\pi/4, \pi, 5\pi/4, 3\pi/2, 7\pi/4$ coordinate curves on $\mathcal{S}$.
(c) Evaluate $\hat{\mathbf{n}}$ using (9). Show that (10) is *not* satisfied at the origin, but that a unit normal $\hat{\mathbf{n}}$ does, nevertheless, exist there.

5. Consider the parametrization

$$x = au\cos v, \quad y = bu\sin v, \quad z = 0,$$

over $0 \le u < 1, \quad 0 \le v < 2\pi$.

(a) Show $\mathcal{S}$ in a labeled sketch.
(b) Show the $u = \frac{1}{2}$ and $v = \pi/4, 3\pi/4, 5\pi/4, 7\pi/4$ coordinate curves on $\mathcal{S}$.
(c) Evaluate $\hat{\mathbf{n}}$ using (9). Show that (10) is *not* satisfied at the origin, but that a unit normal $\hat{\mathbf{n}}$ does, nevertheless, exist there.

6. Consider the parametrization

$$x = (1 - u)\cos v, \quad y = (1 - u)\sin v, \quad z = u, \tag{6.1}$$

over $0 < u \le 1, \quad 0 \le v < 2\pi$.

(a) Eliminating u and v from (6.1), obtain a nonparametric representation of the surface $\mathcal{S}$, and show $\mathcal{S}$ in a neat, labeled sketch. Identify $\mathcal{S}$. (For example, is it a plane? A sphere? ...)
(b) Show the $u = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$ and $v = 0, \pi/4$ coordinate curves on $\mathcal{S}$.
(c) Evaluate $\hat{\mathbf{n}}$ using (9). Show that there is one point on $\mathcal{S}$ at which (10) is not satisfied. Does $\mathcal{S}$ admit a unique normal line at that point?
(d) Find the equation of the tangent plane $\mathcal{T}$ at $u = \frac{1}{2}, v = 0$. Sketch $\mathcal{T}$ and $\mathcal{S}$.
(e) Repeat part (d), for $u = \frac{1}{4}, v = 0$.
(f) Repeat part (d), for $u = \frac{1}{2}, v = \pi/4$.
(g) Repeat part (d), for $u = \frac{1}{2}, v = \pi/2$.

7. Consider the parametrization

$$x = a\cos u, \quad y = b\sin u, \quad z = v^2, \tag{7.1}$$

over $0 \le u < 2\pi, \quad 0 < v < 3$.

(a) Sketch $\mathcal{S}$, and verify that it is an elliptic cylinder.
(b) Show the $u = 0, \pi/4$, and $v = 2, 3$ coordinate curves on $\mathcal{S}$.

8. Give two different parametrizations of each surface:

(a) the x, y plane
(b) the x, z plane

(d) the plane $3x - 2y + z = 6$
(e) the hyperboloid $x^2 + y^2 - z^2 = 1$ over $0 \le z < \infty$
(f) the elliptic paraboloid $x^2 + y^2 = z$
(g) the quadratic cone $x^2 + y^2 = z^2$

9. Find any parametrization $x = x(u,v)$, $y = y(u,v)$, $z = z(u,v)$ of the given plane such that the u, v coordinate curves are orthogonal.

(a) $x + 2y - 4z = 5$ (b) $2x - y + z = 9$
(c) $2x + y + 3z = -6$ (d) $x - y - z = 1$
(e) $x - y - 5z = 0$ (f) $x + y - 4z = 0$

10. Surely, if condition (10) is satisfied at some point P on a surface $\mathcal{S}$, then there exists a unique normal line and a unique tangent plane at P. Prove that the condition (10) is sufficient but not necessary. HINT: Logically, a counterexample will suffice. For such an example, consider the point $x = y = z = 0$ on the surface $z = 0$ (i.e., the x, y plane), with the parametrization $x = u^3$, $y = v^3$, $z = 0$.

11. Find a normal $\hat{\mathbf{n}}$ and the equation of the tangent plane for the given surface $\mathcal{S}$ and a point P; if a unique normal line and tangent plane do not exist at that point then state that.

(a) $\mathcal{S}: \ z = x^2 + y^2, \quad P = (2,1,5)$
(b) $\mathcal{S}: \ z = x^2 - 4y^2, \quad P = (1,2,-3)$
(c) $\mathcal{S}: \ x^4 + y^4 + z^4 = 18, \quad P = (1,-1,2)$
(d) $\mathcal{S}: \ x = yz, \quad P = (3,-1,-3)$
(e) $\mathcal{S}: \ y = x^2 + z^4, \quad P = (1,2,1)$
(f) $\mathcal{S}: \ x^2 + y^2 = 16\sin^2 z, \quad P = (2,2,\pi/4)$
(g) $\mathcal{S}: \ x = u,\ y = u + v^2,\ z = v + 1, \quad P: \ u = 3,\ v = 2$
(h) $\mathcal{S}: \ x = u\cos v,\ y = u\sin v,\ z = u^2, \quad P: \ u = 2,\ v = \pi/6$
(i) $\mathcal{S}: \ x = u\sin v,\ y = u\cos v,\ z = u, \quad P: \ u = 0,\ v = \pi$
(j) $\mathcal{S}: \ x = ve^u,\ y = u - v,\ z = v^2, \quad P: \ u = 2,\ v = 1$
(k) $\mathcal{S}: \ x = u,\ y = u^2 + v,\ z = u - v, \quad P: \ u = 0,\ v = 1$
(l) $\mathcal{S}: \ x = 4u^2,\ y = u^2 - v^2,\ z = 2v^2 + 3, \quad P: \ u = 1,\ v = -1$

15.5 Surface Integrals

In Section 15.3.1 we reviewed double integrals $\iint f\,dA$ over regions in the x, y plane. Having developed the concept of the parametric representation of general curved surfaces in Section 15.4, we are now ready to generalize the concept of double integrals by considering double integrals on curved surfaces in 3-space. For example, suppose that we have a distribution of electric charges over a surface, and we wish to know the total charge or the electric field induced by that distribution. Such quantities are expressible as integrals of the sort that we are about to consider.

15.5.1. Area element dA. Consider a surface $\mathcal{S}$ given parametrically by $\mathbf{R}(u,v) = x(u,v)\hat{\mathbf{i}} + y(u,v)\hat{\mathbf{j}} + z(u,v)\hat{\mathbf{k}}$, where $\mathbf{R}(u,v)$ is C^1 (i.e., $\mathbf{R}$ and its first-order partial derivatives $\mathbf{R}_u$ and $\mathbf{R}_v$ are continuous) and $\mathbf{R}_u \times \mathbf{R}_v \neq \mathbf{0}$ on $\mathcal{S}$. These conditions ensure that $\mathcal{S}$ is "nice." Specifically, they ensure that a unique normal line exists at each point on $\mathcal{S}$ and varies continuously on $\mathcal{S}$ (i.e., is a continuous function of u and v). Such a surface is said to be **smooth**.

To obtain an expression for the area element dA on $\mathcal{S}$, in terms of our u, v coordinate system, we begin by taking a differential of the position vector $\mathbf{R}(u,v)$ from the origin to any point P on $\mathcal{S}$:

$$d\mathbf{R} = \mathbf{R}_u du + \mathbf{R}_v dv. \tag{1}$$

Thus, along the u = constant and v = constant curves through P we have $d\mathbf{R} = \mathbf{R}_v dv$ (since u = constant, so that $du = 0$) and $d\mathbf{R} = \mathbf{R}_u du$, respectively, as

shown in Fig. 1. These vectors define a parallelogram (Fig. 1) lying in the tangent plane to S at P. The area of this parallelogram is [recall (10) in Section 14.2]

$$dA = \|\mathbf{R}_u du \times \mathbf{R}_v dv\| = \|\mathbf{R}_u \times \mathbf{R}_v\| \, du \, dv. \tag{2}$$

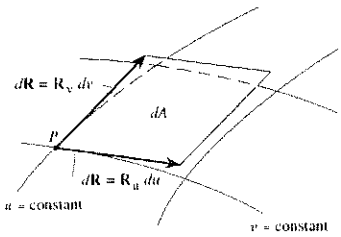


Figure 1. Area element dA.

As du and dv tend to zero this plane area element lies closer and closer to S so that it seems reasonable to define the **area** of the curved surface S by the double integral

$$\boxed{A = \iint_S dA = \iint_{\mathcal{R}} \|\mathbf{R}_u \times \mathbf{R}_v\| \, du \, dv,} \tag{3}$$

where $\mathcal{R}$ is the region in the u, v plane that corresponds to the surface S in 3-space. Notice that we have not proved (3), it is a definition.* Further, we call

$$\boxed{dA = \|\mathbf{R}_u \times \mathbf{R}_v\| \, du \, dv} \tag{4}$$

the **area element** on S.

To obtain a computational version of (4), cross $\mathbf{R}_u = x_u\hat{\mathbf{i}} + y_u\hat{\mathbf{j}} + z_u\hat{\mathbf{k}}$ with $\mathbf{R}_v = x_v\hat{\mathbf{i}} + y_v\hat{\mathbf{j}} + z_v\hat{\mathbf{k}}$. The norm of the resulting vector is the square root of the sum of the squares of its components. Carrying out these steps, we obtain

$$dA = \sqrt{EG - F^2} \, du \, dv, \tag{5a}$$

$$\begin{aligned} E &= x_u^2 + y_u^2 + z_u^2, \\ F &= x_u x_v + y_u y_v + z_u z_v, \\ G &= x_v^2 + y_v^2 + z_v^2. \end{aligned} \tag{5b}$$

Notice that we may attach a geometrical significance to F for, by inspection, we see that $F = \mathbf{R}_u \cdot \mathbf{R}_v$. Thus, if $\mathbf{R}(u, v)$ is such that F is identically zero, then that condition implies that $\mathbf{R}_u$ and $\mathbf{R}_v$ are perpendicular to each other at each point on S. In that event, we say that the curvilinear coordinates u and v are **orthogonal**; that is, they form an "orthogonal net" on S.

EXAMPLE 1. *Surface Area of a Sphere.* To illustrate the use of (5), let us compute the surface area of the spherical surface S defined by

$$x = a \sin v \cos u, \quad y = a \sin v \sin u, \quad z = a \cos v, \tag{6}$$

over $0 \le u \le \pi/2$ and $0 \le v \le \pi/2$ (which surface was the subject of Example 1 in Section 15.4). S constitutes one-eighth of a complete spherical surface and is shown in Fig. 2. Putting (6) into (5b) gives

$$E = a^2 \sin^2 v, \quad F = 0, \quad G = a^2. \tag{7}$$

*For instance, our use of the differential version of the chain rule, in (1), was heuristic, not rigorous.

Thus,

$$dA = a^2 |\sin v|\, du\, dv = a^2 \sin v\, du\, dv \tag{8}$$

since $\sin v \geq 0$ over $0 \leq v \leq \pi/2$, so

$$A = a^2 \int_0^{\pi/2} \int_0^{\pi/2} \sin v\, du\, dv = \frac{\pi a^2}{2}. \tag{9}$$

Of course, (9) is correct since we know that the surface area of the sphere is $4\pi a^2$, and $\mathcal{S}$ is only one-eighth of that.

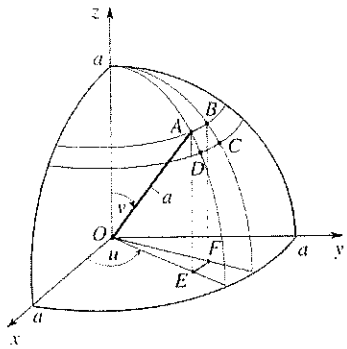


Figure 2. The surface $\mathcal{S}$.

COMMENT 1. The expression (8) could, in this case, have been obtained directly from Fig. 2, without recourse to (5). For dA is the shaded area $ABCD$ so

$$\begin{aligned} dA \sim (AB)(AD) &= (EF)(AD) = (OE\, du)(a\, dv) \\ &= (a \sin v\, du)(a\, dv) = a^2 \sin v\, du\, dv. \end{aligned} \tag{10}$$

Here AB, for example, denotes the length of the arc connecting the points A and B, du is the angle FOE, and dv is the angle DOA. But while the connection between (8) and the geometry contained in Fig. 2 is interesting and supportive, we should emphasize that the "method of pictures" used in (10) may not be feasible in other cases, whereas (5) is simple and automatic.

COMMENT 2. In this case $F = 0$ everywhere on $\mathcal{S}$ so the u, v coordinate curves must be *orthogonal*. This result is no great surprise since it should be evident that the curves of longitude and latitude intersect, at each point on $\mathcal{S}$, at right angles.

COMMENT 3. A striking feature of this example is the beautiful simplicity of the calculation in (9). This simplicity emphasizes that the u, v curvilinear coordinates defined in (6) are rather "natural" for the representation of this particular surface. ∎

There are two special cases of (5) that should be singled out:

Case 1: z=0. If $z = 0$, so that $\mathcal{S}$ is flat and lies in the x, y plane, then $x = x(u, v)$, $y = y(u, v)$, $z = 0$, and (5b) simplifies to

$$E = x_u^2 + y_u^2, \quad F = x_u x_v + y_u y_v, \quad G = x_v^2 + y_v^2. \tag{11}$$

It follows from (11) that

$$\begin{aligned} EG - F^2 &= x_u^2 y_v^2 - 2x_u y_u x_v y_v + x_v^2 y_u^2 \\ &= (x_u y_v - x_v y_u)^2, \end{aligned} \tag{12}$$

which is none other than the *Jacobian* $\partial(x, y)/\partial(u, v)$ squared! Thus, for the case where $z = 0$, (5) simplifies to the form

$$\boxed{dA = \left| \frac{\partial(x, y)}{\partial(u, v)} \right| du\, dv,} \tag{13}$$

the absolute value signs being included to ensure that we have the *positive* square root in (5a) since area must surely be positive.

EXAMPLE 2. As an illustration of Case 1, note that the parametrization choice

$$x = u\cos v, \qquad y = u\sin v, \qquad z = 0 \tag{14}$$

amounts to the familiar case of *plane polar coordinates*; that is, u is r and v is θ. Then the Jacobian is

$$\frac{\partial(x,y)}{\partial(u,v)} = \begin{vmatrix} x_u & x_v \\ y_u & y_v \end{vmatrix} = \begin{vmatrix} \cos v & -u\sin v \\ \sin v & u\cos v \end{vmatrix} = u,$$

so (13) gives

$$dA = |u|\,du\,dv$$

or, in terms of the more familiar letters r and θ,

$$\boxed{dA = r\,dr\,d\theta,} \tag{15}$$

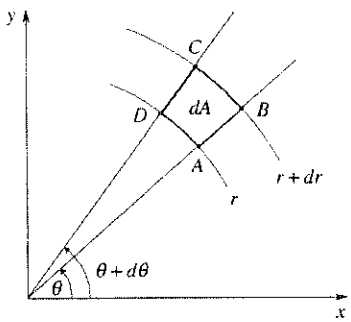


Figure 3. dA in polar coordinates.

where we omit the absolute values since the polar variable r is understood (in this book) to be nonnegative.

Again, seeking a geometrical interpretation, note that the length of AB in Fig. 3 is dr, while the length of AD is $r\,d\theta$ so that $dA \sim (dr)(r\,d\theta) = r\,dr\,d\theta$, as in (15).

COMMENT. Since $dA = dx\,dy$ in Cartesian coordinates and $dA = r\,dr\,d\theta$ in polar coordinates, it is tempting to write $dx\,dy = r\,dr\,d\theta$. However, the latter is **not** correct. Heuristically, we can think of it this way. In the United States the smallest monetary unit is the penny so we can write $d(\text{money}) = \text{penny}$. In the country of Rumanova the smallest monetary unit is the yink so we can write $d(\text{money}) = \text{yink}$. Any amount of money can be formed by a suitable aggregation of pennies or of yinks. But it does not follow that a penny equals a yink. ∎

Case 2: $z = f(x, y)$. If S is given in the form

$$z = f(x,y), \tag{16}$$

over some region $\mathcal{R}$ in the x, y plane, then as noted in Section 15.4 (see Comment 2 in Example 1), we can parametrize S by

$$x = u, \quad y = v, \quad z = f(x,y). \tag{17}$$

Putting (17) into (5) yields $E = 1 + f_u^2$, $F = f_u f_v$, $G = 1 + f_v^2$, and hence the form

$$dA = \sqrt{1 + f_u^2 + f_v^2}\,du\,dv.$$

But since $x = u$ and $y = v$ it seems silly to retain the new variables u and v. Thus, let us write, instead,

$$\boxed{dA = \sqrt{1 + f_x^2 + f_y^2}\,dx\,dy.} \tag{18}$$

Correspondingly,

$$\boxed{A = \iint_{\mathcal{R}} \sqrt{1 + f_x^2 + f_y^2}\, dx\, dy} \tag{19}$$

is the area of $\mathcal{S}$, where $\mathcal{R}$ is the region in the x, y plane lying directly beneath $\mathcal{S}$.

EXAMPLE 3. *Surface Area of a Sphere, Revisited.* Let us compute the surface area of the surface $\mathcal{S}$ shown in Fig. 2 (Example 1) again, this time using the parametrization

$$x = u, \qquad y = v, \qquad z = \sqrt{a^2 - u^2 - v^2}. \tag{20}$$

Since $x = u$ and $y = v$, the coordinate curves on $\mathcal{S}$ are simply the projections onto $\mathcal{S}$ of the $x =$ constant and $y =$ constant lines in the x, y plane. To illustrate, the coordinate curves through a given point P on $\mathcal{S}$ are sketched in Fig. 4. To visualize the coordinate curves more easily, imagine the $x =$ constant and $y =$ constant lines in the x, y plane (such as AB and CD, respectively) as thin wires, and imagine shining a flashlight upward, as shown in the figure. Then the shadow on $\mathcal{S}$ of the $x =$ constant lines will be the $u =$ constant curves (e.g., $A'B$), and the shadow on $\mathcal{S}$ of the $y =$ constant lines will be the $v =$ constant curves (e.g., $C'D$). Observe that the u, v curves are *not* orthogonal, in this case, since $F = f_u f_v$ is not zero, nor do they *look* orthogonal in Fig. 4.

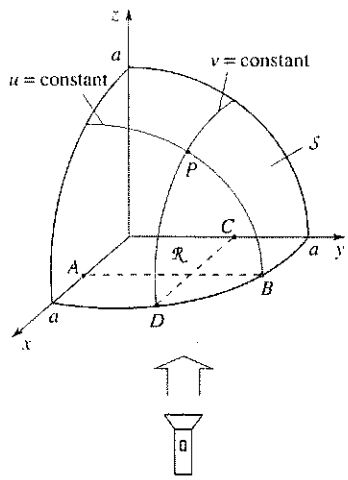


Figure 4. Visualizing the u, v coordinate curves on $\mathcal{S}$.

Returning to our calculation of the surface area of $\mathcal{S}$, we use (19), where $\mathcal{R}$ is the shaded quarter disk in the x, y plane (Fig. 4). Since $f_x = (\frac{1}{2})(-2x)(a^2 - x^2 - y^2)^{-1/2}$ and $f_y = (\frac{1}{2})(-2y)(a^2 - x^2 - y^2)^{-1/2}$, (19) gives

$$A = \int_0^a \int_0^{\sqrt{a^2 - y^2}} \frac{a}{\sqrt{a^2 - x^2 - y^2}}\, dx\, dy. \tag{21}$$

Evaluation of the iterated integral in (21) is left for the exercises; the answer, of course, is $\pi a^2/2$, as before. ∎

15.5.2. Surface integrals. We are ready to study **surface integrals**, integrals of the form

$$\iint_{\mathcal{S}} f\, dA, \tag{22}$$

where the function f is known on the surface $\mathcal{S}$.

Our definition of the surface integral (22) will be essentially the same as that given in Section 15.3.1 for the double integral $\iint_{\mathcal{R}} f\, dA$ on some region $\mathcal{R}$ in the x, y plane. That is, we partition $\mathcal{S}$ into N parts $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_N$, and define the *norm* of the partition $|p|$ as the "size" of the largest one of these parts. (As a measure of the "size" of an element $\mathcal{S}_j$, we can, for example, use the least upper bound of the linear distances between all possible pairs of points on $\mathcal{S}_j$.) Next, let (x_j, y_j, z_j) be an arbitrarily selected point on $\mathcal{S}_j$ for each $j = 1, 2, \ldots, N$. Then, denoting the

area of S_j as ΔA_j,

$$\sum_{j=1}^{N} f(x_j, y_j, z_j)\Delta A_j \tag{23}$$

is called the **Riemann sum** corresponding to the chosen partition and (x_j, y_j, z_j) points. The idea, in principle at least, is to compute the Riemann sum (23), then to introduce a finer partition (i.e., one with a smaller norm) and compute the new Riemann sum, and so on, such that the norm of the partitions tends to zero. If the sequence of values of the Riemann sum thus generated converges to a unique limit, independent of the choice of the partition sequence and (x_j, y_j, z_j) points, that limit is, by definition, the **surface integral** $\iint_S f\,dA$; that is,

$$\iint_S f(x, y, z)\,dA \equiv \lim_{|p|\to 0} \sum f(x_j, y_j, z_j)\Delta A_j. \tag{24}$$

To *evaluate* a surface integral we do *not* use the definition (24) (although we *could* carry out that exercise in sufficiently simple cases). Rather, we convert the integral to an iterated integral by parametrizing S in the form $\mathbf{R}(u, v)$ and using the expression $dA = \|\mathbf{R}_u \times \mathbf{R}_v\|\,du\,dv = \sqrt{EG - F^2}\,du\,dv$. Thus,

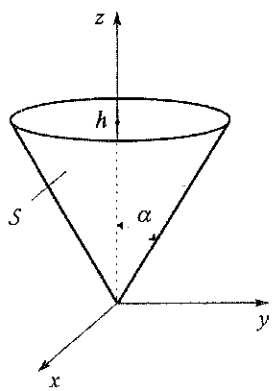
$$\iint_S f\,dA = \iint_{\mathcal{R}} f(x(u, v), y(u, v), z(u, v))\sqrt{EG - F^2}\,du\,dv, \tag{25a}$$

where $\mathcal{R}$ is the region in the u, v plane that corresponds to the surface S in 3-space. Or if S is given in the form $z = z(x, y)$, then

$$\iint_S f\,dA = \iint_{\mathcal{R}} f(x, y, z(x, y))\sqrt{1 + z_x^2 + z_y^2}\,dx\,dy, \tag{25b}$$

where $\mathcal{R}$ is the "shadow" of S on the x, y plane.

(a)



(b)

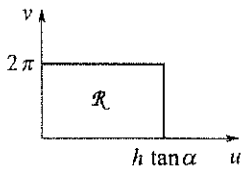


Figure 5. Conical shell.

EXAMPLE 4. *Mass of a Conical Shell.* Consider a thin shell in the shape of a right circular cone S of height h as shown in Fig. 5a. Assuming that the mass density distribution σ (mass per unit surface area) is known, we wish to compute the total mass, which may be expressed as the surface integral

$$M = \iint_S \sigma\,dA. \tag{26}$$

First, let us parametrize S. (Before reading on, you should try to make up a suitable parametrization, remembering that the parametrization of a given surface is not unique, and that the rule of thumb is to construct a parametrization which is as simple and natural as possible.) Adopting the parametrization

$$\begin{aligned} x &= u\cos v, \\ y &= u\sin v, \\ z &= u\cot\alpha, \end{aligned} \tag{27}$$

over the region $\mathcal{R}$ shown in Fig. 5b, we compute $E = x_u^2 + y_u^2 + z_u^2 = 1 + \cot^2\alpha = 1/\sin^2\alpha$, $F = x_u x_v + y_u y_v + z_u z_v = 0$, and $G = x_v^2 + y_v^2 + z_v^2 = u^2$, so that the form (25b) gives

$$M = \iint_{\mathcal{R}} \sigma(u,v)\,\frac{u}{\sin\alpha}\,du\,dv = \frac{1}{\sin\alpha}\int_0^{2\pi}\int_0^{h\tan\alpha} \sigma(u,v)u\,du\,dv. \tag{28}$$

For example, if $\sigma(u,v) = \text{constant} = \sigma_0$, in which case we say that the shell is homogeneous, then

$$M = \frac{\sigma_0}{\sin\alpha}\int_0^{2\pi}\int_0^{h\tan\alpha} u\,du\,dv = \pi\sigma_0 h^2 \sin\alpha\sec^2\alpha. \tag{29}$$

COMMENT 1. Where did we get (27)? In cylindrical coordinates the equation of the cone is $z = r\cot\alpha$. Also, $x = r\cos\theta$ and $y = r\sin\theta$. These three equations are a suitable parametrization of $\mathcal{S}$, where the parameters are r and θ. Finally, we took r to be u and θ to be v, to be consistent with our usual u, v notation. If we chose $r = v$ and $\theta = u$, instead, the same final result (29) would be obtained.

COMMENT 2. Alternatively, spherical polars are an attractive choice because the cone is a constant ϕ surface, namely, $\phi = \alpha$. Thus, $x = \rho\sin\phi\cos\theta = \rho\sin\alpha\cos\theta$, $y = \rho\sin\phi\sin\theta = \rho\sin\alpha\sin\theta$, and $z = \rho\cos\phi = \rho\cos\alpha$ so if we let ρ be u and θ be v, say, then we have the alternative parametrization

$$\begin{aligned} x &= u\sin\alpha\cos v,\\ y &= u\sin\alpha\sin v,\\ z &= u\cos\alpha. \end{aligned} \tag{30}$$

From (30) we obtain $E = 1$, $F = 0$, and $G = u^2\sin^2\alpha$ so (with $\sigma = \sigma_0$ again)

$$M = \int_0^{2\pi}\int_0^{h\sec\alpha} \sigma_0 u\sin\alpha\,du\,dv, \tag{31}$$

which gives the same final result as obtained in (29).

COMMENT 3. As a partial check of (29), observe that $M \to 0$ as $\alpha \to 0$, as it should, and $M \to \infty$ as $\alpha \to \pi/2$, as it should. ∎

Closure. The key to this section is the definition, by (4), of the area element dA of a smooth surface $\mathcal{S}$ that is defined parametrically in terms of u and v. In turn, (4) springs from Fig. 1, which should be understood and remembered. Working out $\|\mathbf{R}_u \times \mathbf{R}_v\|$ in terms of $x(u,v)$, $y(u,v)$, $z(u,v)$ then produce the computationally convenient version (5). In addition, we call attention to two important special cases. In the first, $\mathcal{S}$ is flat and lies in the x, y plane and dA reduces to the absolute magnitude of the Jacobian times $du\,dv$; in the second, the shape of $\mathcal{S}$ is known in the form $z = f(x,y)$, and dA then simplifies to the form given in (18).

EXERCISES 15.5

1. In each case evaluate the surface area A of the surface $\mathcal{S}$ using (5).

(a) Let $\mathcal{S}$ be the *circular cylinder* $x^2 + y^2 = 1$, between $z = 0$ and $z = 1 + y$. HINT: Use $x = \cos v$, $y = \sin v$, $z = u$.
(b) Let $\mathcal{S}$ be the *circular cylinder* $x^2 + y^2 = 1$, between $z = 0$ and $z = 1 - y^2$. HINT: Same hint as in part (a).
(c) Let $\mathcal{S}$ be the *quadric cone* $x^2 + y^2 = z^2$, between $z = 0$ and $z = h$. Show that $A = \sqrt{2}\pi h^2$. HINT: Use $x = u\cos v$, $y = u\sin v$, $z = u$.

2. Let $\mathcal{S}$ be the *elliptic paraboloid* $x^2 + y^2 = z$, between $z = 0$ and $z = h$.

(a) Using (5), show that $A = \left(\frac{\pi}{6}\right)[(1 + 4h)^{3/2} - 1]$.
(b) Applying Taylor's series to the answer in part (a), show that $A \sim \pi h$ as $h \to 0$. Explain why this result looks correct [and hence provides us with a partial check on our answer to part (a)].

3. Let $\mathcal{S}$ be the one-sheeted *hyperboloid* $x^2 + y^2 - z^2 = 1$, between $z = 0$ and $z = h$.

(a) Show that

$$A = \pi\left[h\sqrt{1 + 2h^2} + \frac{1}{\sqrt{2}}\ln\left|\sqrt{2}\,h + \sqrt{1 + 2h^2}\right|\right]. \tag{3.1}$$

(b) Applying Taylor's series to (3.1), show that $A \sim 2\pi h$ as $h \to 0$. Explain why this result looks correct [and hence provides us with a partial check on (a)].

4. Let $\mathcal{S}$ be the *hyperboloid* $z^2 - x^2 - y^2 = 1$, between $z = 1$ and $z = h$.

(a) Using (5), with $x = u\cos v$, $y = u\sin v$, $z = \sqrt{u^2 + 1}$, show that

$$A = \pi\left[h\sqrt{2h^2 - 1} - \tfrac{1}{\sqrt{2}}\ln\left|\sqrt{2}\,h + \sqrt{2h^2 - 1}\right| - 1 + \tfrac{1}{\sqrt{2}}\ln\left|1 + \sqrt{2}\right|\right]. \tag{4.1}$$

HINT: Same hint as in Exercise 3(a).

(b) To examine the case where $\mathcal{S}$ is very "shallow," set $h = 1 + \epsilon$ in (4.1), and consider the behavior of A as $\epsilon \to 0$. Explain why the result looks correct [and hence provides us with a partial check on (4.1)].

(c) In the same spirit as part (b), show that $A \sim \sqrt{2}\pi h^2$ as $h \to \infty$. Explain why this result looks correct [and therefore provides us with another partial check on (4.1)].

5. Parametrizing the circular cylinder $x^2 + y^2 = a^2$ by $x = a\cos v$, $y = a\sin v$, $z = u$, show that the area element is $dA = a\,du\,dv$. Interpret this result geometrically, with a labeled sketch.

6. Find the area A of the following regions in the x, y plane:

(a) the region enclosed by $r = \sin\theta$ $(0 \le \theta < \pi)$
(b) the region enclosed by the limaçon $r = 2 + \cos\theta$ $(0 \le \theta < 2\pi)$
(c) the region enclosed by one leaf of the "daisy" $r = 2\sin 3\theta$
(d) the region (in the second quadrant) between the circle $r = \sin\theta$ $(0 \le \theta < \pi)$ and the cardioid $r = 1 + \cos\theta$ $(0 \le \theta < 2\pi)$.

7. Consider a portion $\mathcal{S}$ of a plane $ax + by + cz = d$ with a "shadow" on the x, y plane designated as the region $\mathcal{R}$. We assume that the plane is not perpendicular to the x, y plane, so that $c \neq 0$ and $z = (d - ax - by)/c$. Using (19), show that the area of $\mathcal{S}$ is

$$A = (\sec\alpha)(\text{area of } \mathcal{R}), \tag{7.1}$$

where α is the acute angle between the z axis and the normal line to $\mathcal{S}$.

8. Use (19) to show that the surface area of the paraboloid $z = h(1 - x^2 - y^2)$, between $z = 0$ and $z = h$, is

$$A = \iint_{\mathcal{R}} \sqrt{1 + 4h^2(x^2 + y^2)}\,dx\,dy, \tag{8.1}$$

where $\mathcal{R}$ is the unit disk $x^2 + y^2 \le 1$. To integrate (8.1), change to polar coordinates r, θ, and show that

$$A = \frac{\pi}{6h^2}\left[(1 + 4h^2)^{3/2} - 1\right]. \tag{8.2}$$

As a partial check on (8.2), show that the right-hand side of (8.2) tends to π (namely, the area of the unit disk) as $h \to 0$.

9. Sketch the surface $\mathcal{S}$ defined by $z = \epsilon(1 - x^2)(1 - y^2)$ over the square $0 \le x \le 1$, $0 \le y \le 1$, where $0 \le \epsilon \ll 1$. Show that (19) gives the area of $\mathcal{S}$ as

$$A = \int_0^1\int_0^1 \sqrt{1 + 4\epsilon^2\left[x^2(1 - y^2)^2 + (1 - x^2)^2y^2\right]}\,dx\,dy.$$

(9.1)

Evidently, this integral is difficult to evaluate. However, recalling that ϵ is small, suppose that we set $t \equiv 4\epsilon^2\left[x^2(1-y^2)^2+(1-x^2)^2y^2\right]$ and expand the integrand as a Maclaurin series in t:

$$\begin{aligned}\sqrt{1+4\epsilon^2\left[x^2(1-y^2)^2+(1-x^2)^2y^2\right]}\\ &= (1+t)^{1/2}\\ &= 1+\frac{1}{2}t-\frac{1}{8}t^2+\cdots \qquad (|t|<1)\\ &= 1+2\epsilon^2[x^2(1-y^2)^2+(1-x^2)^2y^2]\\ &\quad -2\epsilon^4[x^2(1-y^2)^2+(1-x^2)^2y^2]^2+\cdots.\end{aligned}$$

(9.2)

Replacing the integrand in (9.1) by the infinite series on the right-hand side of (9.2), we succeed in trading in one function that is very difficult to integrate for many that are quite simple to integrate (since they are of the form $x^m y^n$). In fact, we do not even need many terms, for good accuracy, if ϵ is sufficiently small, for then $t \ll 1$ so that $(1+t)^{1/2} \approx 1+\frac{1}{2}t$ or $1+\frac{1}{2}t-\frac{1}{8}t^2$, say, may suffice. Using $(1+t)^{1/2} \approx 1+\frac{1}{2}t$, show that $A \approx 1+\frac{32}{45}\epsilon^2$. NOTE: The preceding discussion does not pretend to be any more than heuristic. A rigorous basis for integrating the infinite series in a term-by-term manner is provided by the following *theorem*: If $|a_n(x,y)| < M_n$ over the closed region $\mathcal{R}$, where the M_n's are constants and $\sum^\infty M_n$ is convergent, then

$$\iint_{\mathcal{R}} \sum^{\infty} a_n(x,y)\,dx\,dy = \sum^{\infty} \iint_{\mathcal{R}} a_n(x,y)\,dx\,dy \tag{9.3}$$

(i.e., we may integrate the series term by term).

10. Evaluate $\iint_{\mathcal{S}}(1+x)dA$, where the surface $\mathcal{S}$ is

(a) the plane $z = 1+y$ with vertices at $(0,0,1)$, $(1,0,1)$, $(1,1,2)$, and $(0,1,2)$
(b) the plane $z = x+y$ with vertices at $(0,0,0)$, $(1,0,1)$, $(0,1,1)$
(c) the cylinder $x^2+y^2=1$, between $z=0$ and $z=h$
(d) the cylinder $x^2+y^2=1$, between $z=0$ and $z=1+x$
(e) the hemisphere $x^2+y^2+z^2=9$, between $x=0$ and $x=3$
(f) the sphere $x^2+y^2+z^2=4$
(g) the quadric cone $x^2+y^2=z^2$, between $z=0$ and $z=h$
HINT: Use $x=u\cos v$, $y=u\sin v$, $z=u$.
(h) elliptic paraboloid $x^2+y^2=z$, between $z=0$ and $z=h$

11. (*Mass and center of gravity*) Let σ be the mass density of a (negligibly thick) distribution of mass over a surface $\mathcal{S}$. That is, σ is the mass per unit area at each point on $\mathcal{S}$; it may vary over $\mathcal{S}$. Then the x, y, z coordinates of the **center of gravity** are defined as

$$\begin{aligned} x_c &= \frac{1}{M}\iint_{\mathcal{S}} x\sigma\,dA,\\ y_c &= \frac{1}{M}\iint_{\mathcal{S}} y\sigma\,dA,\\ z_c &= \frac{1}{M}\iint_{\mathcal{S}} z\sigma\,dA,\end{aligned} \tag{11.1}$$

where

$$M = \iint_{\mathcal{S}} \sigma\,dA \tag{11.2}$$

is the total mass. Evaluate x_c in each case. (You need not evaluate y_c, z_c.)

(a) $\mathcal{S}$ is the plane surface $z = x+2y$ with vertices at $(0,0,0)$, $(1,0,1)$, $(0,1,2)$; $\sigma = \text{constant} = \sigma_0$.
(b) $\mathcal{S}$ is the plane surface $z = x+y$ with vertices at $(0,0,0)$, $(1,0,1)$, $(0,1,1)$; $\sigma = 1+y$.
(c) $\mathcal{S}$ is the plane surface $z = 2-x$ with vertices at $(0,0,2)$, $(0,1,2)$, $(2,0,0)$, $(2,1,0)$; $\sigma = 4-x$.
(d) $\mathcal{S}$ is the plane surface $z = 2-x$ with vertices at $(0,0,2)$, $(0,1,2)$, $(2,0,0)$; $\sigma = 1+x$.
(e) $\mathcal{S}$ is the cylindrical surface $x^2+y^2=4$ between $z=0$ and $z=h$; $\sigma = 4+x$.
(f) $\mathcal{S}$ is the cylindrical surface $x^2+y^2=4$ between $z=0$ and $z=2+x$; $\sigma = \text{constant} = \sigma_0$.
(g) $\mathcal{S}$ is the hemispherical surface $x^2+y^2+z^2=1$ between $x=0$ and $x=1$; $\sigma = \text{constant} = \sigma_0$.
(h) $\mathcal{S}$ is the spherical surface $x^2+y^2+z^2=9$; $\sigma = 4+x$.

12. (*Moments of inertia*) Let σ be the mass density of a (negligibly thick) distribution of mass over a surface $\mathcal{S}$, as in Exercise 11. Then the **moments of inertia** I_x, I_y, I_z about the x, y, z axes, respectively, are defined as

$$\begin{aligned} I_x &= \iint_{\mathcal{S}} (y^2+z^2)\sigma\,dA,\\ I_y &= \iint_{\mathcal{S}} (x^2+z^2)\sigma\,dA,\\ I_z &= \iint_{\mathcal{S}} (x^2+y^2)\sigma\,dA.\end{aligned} \tag{12.1}$$

Evaluate I_x in each case:

(a) same as in Exercise 11(a)

(b) same as in Exercise 11(b)
(c) same as in Exercise 11(c)
(d) $\mathcal{S}$ is the spherical surface $x^2+y^2+z^2=c^2$; $\sigma=$ constant $=\sigma_0$
(e) $\mathcal{S}$ is the cylindrical surface $y^2+z^2=4$ between $x=0$ and $x=3$; $\sigma=$ constant $=\sigma_0$
(f) $\mathcal{S}$ is a flat disk of radius c, lying in the y,z plane, with its center at the origin; $\sigma=$ constant $=\sigma_0$
(g) $\mathcal{S}$ is a flat disk of radius c, lying in the x,z plane, with its center at the origin; $\sigma=$ constant $=\sigma_0$.

13. In each case $\mathcal{S}$ is flat and lies in the x,y plane. Evaluate each integral by introducing polar coordinates r,θ.

(a) $\iint_{\mathcal{S}} \frac{y}{x}\,dx\,dy$, $\mathcal{S}$: $1\le r\le 2$, $0\le\theta\le\pi/4$
(b) $\iint_{\mathcal{S}} \sin(x^2+y^2)\,dx\,dy$, $\mathcal{S}$: $r\le 2$, $0\le\theta\le\pi/3$
(c) $\iint_{\mathcal{S}} \frac{y(x^2+y^2)^2}{x}\,dx\,dy$, $\mathcal{S}$: $1\le r\le 2$, $-\pi/3\le\theta\le\pi/3$
(d) $\iint_{\mathcal{S}} (x^2+y^2)^{-1/2}\,dx\,dy$, $\mathcal{S}$: $1\le r\le 2$, $0\le\theta\le\pi/2$
(e) $\iint_{\mathcal{S}} x^2(x^2+y^2)^{-3/2}\,dx\,dy$, $\mathcal{S}$: triangle with vertices at $(0,0)$, $(1,0)$, and $(1,1)$
(f) $\iint_{\mathcal{S}} e^{-(x^2+y^2)}\,dx\,dy$, $\mathcal{S}$: $r\le 1$, $0\le\theta\le\pi$

14. To evaluate

$$I=\int_0^1\int_0^{1-y} e^{x/(x+2y)}\,dx\,dy, \tag{14.1}$$

make the change of variables

$$u=x,\qquad v=x+2y. \tag{14.2}$$

The resulting integrand will be an easy function of u and a hard function of v so integrate on u first. To determine the u,v integration limits first infer the region of integration in the x,y plane, from (14.1), then use the transformation (14.2) to determine the region of integration in the u,v plane. Thus, show that

$$I=\frac{e}{4}-1+\frac{1}{2e}\int_1^2 ve^{2/v}\,dv. \tag{14.3}$$

Finally, use computer software to evaluate the v integral in (14.3), and thus show that $I=0.76624517$. HINT: For numerical integration using *Maple*, see ?int[numerical].

15. (a) To evaluate

$$I=\int_0^1\int_0^1 \sin\frac{xy}{xy+1}\,dx\,dy, \tag{15.1}$$

make the change of variables $u=x$, $v=xy$ and, using the ideas outlined in Exercise 14, show that $I=0.174971978$.
(b) In part (a), the choice $u=xy$, $v=xy+1$ might have seemed more natural. Show why that choice does *not* work.

15.6 Volumes and Volume Integrals

Having discussed the generation of curves by one-parameter families of the form $x=x(u)$, $y=y(u)$, $z=z(u)$, and the generation of surfaces by two-parameter families of the form $x=x(u,v)$, $y=y(u,v)$, $z=z(u,v)$, it should be no surprise that we now consider the generation of *volumes* by three-parameter families of the form

$$x=x(u,v,w),\qquad y=y(u,v,w),\qquad z=z(u,v,w), \tag{1}$$

or in terms of the position vector $\mathbf{R}$,

$$\mathbf{R}(u,v,w)=x(u,v,w)\hat{\mathbf{i}}+y(u,v,w)\hat{\mathbf{j}}+z(u,v,w)\hat{\mathbf{k}}. \tag{2}$$

That is, for each fixed w the parametrization (1) [or (2)] defines a surface (in general). Thus, as we vary w we produce a family of such surfaces which, in general, will generate a volume.

15.6.1. Volume element dV. We assume that $\mathbf{R}(u, v, w)$ is C^1 in the region $\mathcal{V}$ of interest (i.e., $\mathbf{R}$, $\mathbf{R}_u$, $\mathbf{R}_v$, and $\mathbf{R}_w$ are continuous in $\mathcal{V}$), and that $\mathbf{R}_u$, $\mathbf{R}_v$, $\mathbf{R}_w$ are linearly independent (or, equivalently, that $\mathbf{R}_u \cdot \mathbf{R}_v \times \mathbf{R}_w \neq 0$) at each point in $\mathcal{V}$. Mimicing our discussion of surface area in Section 15.5.1, consider the $u =$ constant, $v =$ constant, and $w =$ constant surfaces through a given point P as sketched in Fig. 1. If the curve PQ is the intersection of the $w =$ constant, and $v =$ constant surfaces through P, then only u varies along PQ so PQ is the u coordinate curve through P. Similarly, PR is the v coordinate curve through P, and PS is the w coordinate curve through P.

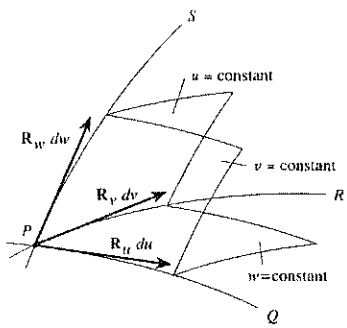


Figure 1. Edges of the parallelepiped for dV.

Since $\mathbf{R}_u$, $\mathbf{R}_v$, $\mathbf{R}_w$ are linearly independent, by assumption, the vectors $\mathbf{R}_u\,du$, $\mathbf{R}_v\,dv$, $\mathbf{R}_w\,dw$ determine a parallelepiped of nonzero volume dV. According to (5) in Section 14.4,

$$\begin{aligned} dV &= |\mathbf{R}_u\,du \cdot \mathbf{R}_v\,dv \times \mathbf{R}_w\,dw| = |\mathbf{R}_u \cdot \mathbf{R}_v \times \mathbf{R}_w|\,du\,dv\,dw \\ &= \left\| \begin{matrix} x_u & y_u & z_u \\ x_v & y_v & z_v \\ x_w & y_w & z_w \end{matrix} \right\| du\,dv\,dw, \end{aligned} \tag{3}$$

where the inner vertical rules on the right-hand side of (3) denote determinant, and the outer ones denote absolute value. But the determinant in (3) is none other than the Jacobian $\partial(x, y, z)/\partial(u, v, w)$* so that

$$\boxed{dV = \left| \frac{\partial(x, y, z)}{\partial(u, v, w)} \right| du\,dv\,dw,} \tag{4}$$

and this quantity is hereby *defined* to be the **volume element** at P. Then the volume V of a given region $\mathcal{V}$ in x, y, z space may be expressed as

$$\boxed{V = \iiint_{\mathcal{V}} dV = \iiint_{\mathcal{R}} \left| \frac{\partial(x, y, z)}{\partial(u, v, w)} \right| du\,dv\,dw,} \tag{5}$$

where $\mathcal{R}$ is the region in u, v, w space corresponding to the region $\mathcal{V}$ in x, y, z space.

EXAMPLE 1. *Cylindrical Coordinates.* If we let the cylindrical coordinates r, θ, z be u, v, w, respectively, then

$$x = u \cos v, \qquad y = u \sin v, \qquad z = w \tag{6}$$

*Recall that $\det \mathbf{A}^{\mathrm{T}} = \det \mathbf{A}$ for any square matrix $\mathbf{A}$. Therefore, the Jacobian $J(u, v, w) = \dfrac{\partial(x, y, z)}{\partial(u, v, w)}$ can be expressed either as

$$\begin{vmatrix} x_u & x_v & x_w \\ y_u & y_v & y_w \\ z_u & z_v & z_w \end{vmatrix} \quad \text{or as} \quad \begin{vmatrix} x_u & y_u & z_u \\ x_v & y_v & z_v \\ x_w & y_w & z_w \end{vmatrix}.$$

is our parametrization. Then

$$\frac{\partial(x,y,z)}{\partial(u,v,w)} = \begin{vmatrix} x_u & x_v & x_w \\ y_u & y_v & y_w \\ z_u & z_v & z_w \end{vmatrix} = \begin{vmatrix} \cos v & -u\sin v & 0 \\ \sin v & u\cos v & 0 \\ 0 & 0 & 1 \end{vmatrix} = u \tag{7}$$

so (4) gives $dV = |u|\,du\,dv\,dw$ or, returning to the more familiar r, θ, z notation, $dV = |r|\,dr\,d\theta\,dz$. Since $r \geq 0$, we have the result

$$\boxed{dV = r\,dr\,d\theta\,dz,} \tag{8}$$

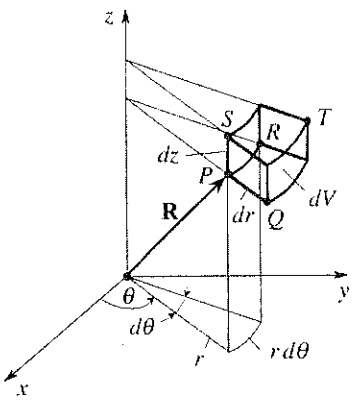


Figure 2. Cylindrical coordinates.

which result admits the simple geometrical interpretation indicated in Fig. 2. To explain, let us write down the position vector [recall (22) in Section 14.6]

$$\mathbf{R} = r\hat{\mathbf{e}}_r(\theta) + z\hat{\mathbf{e}}_z. \tag{9}$$

Then

$$\begin{aligned} d\mathbf{R} &= dr\hat{\mathbf{e}}_r + r\frac{d\hat{\mathbf{e}}_r}{d\theta}\,d\theta + dz\hat{\mathbf{e}}_z \\ &= dr\hat{\mathbf{e}}_r + r\,d\theta\hat{\mathbf{e}}_\theta + dz\hat{\mathbf{e}}_z \end{aligned} \tag{10}$$

or, in the notation of Fig. 2,

$$\mathbf{PT} = \mathbf{PQ} + \mathbf{PR} + \mathbf{PS}. \tag{11}$$

Since $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z$ are orthogonal, it follows that $dV = \|\mathbf{PQ}\|\,\|\mathbf{PR}\|\,\|\mathbf{PS}\| = (dr)(r\,d\theta)(dz)$, as given by (8). ∎

EXAMPLE 2. *Spherical Coordinates.* If we let the spherical coordinates ρ, ϕ, θ be u, v, w, respectively, then

$$\begin{aligned} x &= u\sin v\cos w, \\ y &= u\sin v\sin w, \\ z &= u\cos v, \end{aligned} \tag{12}$$

so

$$\begin{aligned} \frac{\partial(x,y,z)}{\partial(u,v,w)} &= \begin{vmatrix} x_u & x_v & x_w \\ y_u & y_v & y_w \\ z_u & z_v & z_w \end{vmatrix} \\ &= \begin{vmatrix} \sin v\cos w & u\cos v\cos w & -u\sin v\sin w \\ \sin v\sin w & u\cos v\sin w & u\sin v\cos w \\ \cos v & -u\sin v & 0 \end{vmatrix} \\ &= u^2\sin v. \end{aligned} \tag{13}$$

Thus, (4) gives $dV = |u^2 \sin v|\, du\, dv\, dw = \rho^2 |\sin\phi|\, d\rho\, d\phi\, d\theta$. As noted in Section 14.6.3, we normally limit ϕ and θ so that $0 \le \phi \le \pi$ and $0 \le \theta < 2\pi$. Since $0 \le \sin\phi \le 1$ on $0 \le \phi \le \pi$ we can drop the absolute value signs and write

$$\boxed{dV = \rho^2 \sin\phi\, d\rho\, d\phi\, d\theta.} \tag{14}$$

Alternatively,

$$\mathbf{R} = \rho \hat{\mathbf{e}}_\rho(\phi, \theta) \tag{15}$$

so

$$\begin{aligned} d\mathbf{R} &= d\rho \hat{\mathbf{e}}_\rho + \rho \left(\frac{\partial \hat{\mathbf{e}}_\rho}{\partial \phi}\, d\phi + \frac{\partial \hat{\mathbf{e}}_\rho}{\partial \theta}\, d\theta \right) \\ &= d\rho \hat{\mathbf{e}}_\rho + \rho d\phi \hat{\mathbf{e}}_\phi + \rho \sin\phi\, d\theta \hat{\mathbf{e}}_\theta, \end{aligned} \tag{16}$$

where we have used (28) in Section 14.6. And since $\hat{\mathbf{e}}_\rho, \hat{\mathbf{e}}_\phi, \hat{\mathbf{e}}_\theta$ are orthogonal, it follows from (16) that $dV = (d\rho)(\rho\, d\phi)(\rho \sin\phi\, d\theta)$ as given by (14). For geometrical interpretation see Fig. 3. ∎

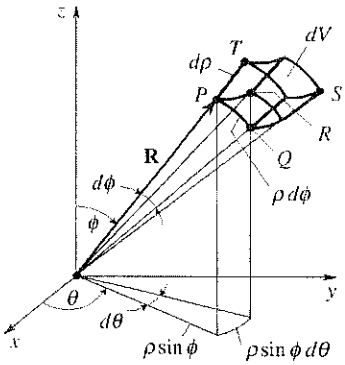


Figure 3. Spherical coordinates.

To tie the present section together with the material in Section 14.6 on cylindrical and spherical coordinates, observe that for those coordinate systems we can derive the *volume element* not only from (4), but also from the product of the three orthogonal components of the $d\mathbf{R}$ vector. Furthermore, the products of those components taken two at a time give the *area elements* on the constant-coordinate surfaces.

For instance, on a conical ϕ = constant surface $d\phi = 0$ so (16) becomes $d\mathbf{R} = d\rho \hat{\mathbf{e}}_\rho + \rho \sin\phi\, d\theta\, \hat{\mathbf{e}}_\theta$, and $dA = (d\rho)(\rho \sin\phi\, d\theta) = \rho \sin\phi\, d\rho\, d\theta$ is the relevant area element, namely, the shaded area in Fig. 3. Similarly, on a spherical ρ = constant surface $d\rho = 0$ so (16) becomes $d\mathbf{R} = \rho\, d\phi \hat{\mathbf{e}}_\phi + \rho \sin\phi\, d\theta\, \hat{\mathbf{e}}_\theta$, and $dA = (\rho\, d\phi)(\rho \sin\phi\, d\theta) = \rho^2 \sin\phi\, d\phi\, d\theta$. Of course, these results are the same as would be obtained using the $dA = \sqrt{EG - F^2}\, du\, dv$ formula in Section 15.5.

Thus, since it contains so much information, the expression for $d\mathbf{R}$ for any given orthogonal coordinate system, is extremely important. Let us summarize these results, for reference, for cylindrical and spherical coordinates.

Cylindrical coordinates:

$$\boxed{\begin{aligned} \mathbf{R} &= r\hat{\mathbf{e}}_r + z\hat{\mathbf{e}}_z \\ d\mathbf{R} &= dr\hat{\mathbf{e}}_r + r\,d\theta\hat{\mathbf{e}}_\theta + dz\hat{\mathbf{e}}_z \\ dA &= \begin{cases} r\, d\theta\, dz & \text{(constant-}r\text{ surface)} \\ dr\, dz & \text{(constant-}\theta\text{ surface)} \\ r\, dr\, d\theta & \text{(constant-}z\text{ surface)} \end{cases} \\ dV &= r\, dr\, d\theta\, dz. \end{aligned}} \tag{17}$$

Spherical coordinates:

$$\boxed{\begin{aligned} \mathbf{R} &= \rho\hat{\mathbf{e}}_\rho \\ d\mathbf{R} &= d\rho\hat{\mathbf{e}}_\rho + \rho\, d\phi\hat{\mathbf{e}}_\phi + \rho\sin\phi\, d\theta\hat{\mathbf{e}}_\theta \\ dA &= \begin{cases} \rho^2\,|\sin\phi|\, d\phi\, d\theta & \text{(constant-}\rho\text{ surface)} \\ \rho\,|\sin\phi|\, d\rho\, d\theta & \text{(constant-}\phi\text{ surface)} \\ \rho d\rho\, d\phi & \text{(constant-}\theta\text{ surface)} \end{cases} \\ dV &= \rho^2\,|\sin\phi|\, d\rho\, d\phi\, d\theta, \end{aligned}} \tag{18}$$

where the absolute value signs are to be enforce that the area and volume elements always be positive.

15.6.2. Volume integrals. With dV in hand, as given by (4), we can deal with the **volume integral** $\iiint_{\mathcal{V}} f\, dV$ of a given function f over a given region $\mathcal{V}$ in 3-space for if x, y, z are parametrized by u, v, w, then

$$\boxed{\int_{\mathcal{V}} f\, dV = \int_{\mathcal{R}} f(x(u,v,w),\ y(u,v,w),\ z(u,v,w)) \left|\frac{\partial(x,y,z)}{\partial(u,v,w)}\right| du\, dv\, dw,} \tag{19}$$

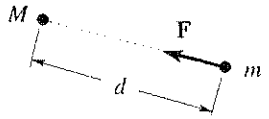


Figure 4. Gravitational force $\mathbf{F}$ exerted by M on m.

where $\mathcal{R}$ is the region in u, v, w space corresponding to the region $\mathcal{V}$ in x, y, z space. We use single integral signs in (19) to denote triple integrals, for compactness, and use that notation routinely in Chapter 16. There should be no confusion, in (19), because of the $\mathcal{V}$ and $\mathcal{R}$, which are three-dimensional regions, and because the dV and $du\, dv\, dv$ clearly denote triple integrals.

EXAMPLE 3. *Gravitational Attraction of a Cone.* **Newton's law of gravitation** states that the force of attraction $\mathbf{F}$ exerted by any one point mass M on any other point mass m is given by

$$\mathbf{F} = G\,\frac{Mm}{d^2}\,\hat{\mathbf{e}}, \tag{20}$$

where (Fig. 4) d is the distance of separation, $\hat{\mathbf{e}}$ is a unit vector directed from m toward M, and G ($= 6.67 \times 10^{-8}$ cm^3/g sec^2) is the universal gravitational constant; (20) is said to be an *inverse square law* since the force varies as the inverse square of the distance. (By M and m being *point masses*, we mean that their sizes are negligible compared with d.) If M is at (x, y, z) and m is at (x_0, y_0, z_0), we can express $\hat{\mathbf{e}}$ as

$$\hat{\mathbf{e}} = \frac{(x-x_0)\hat{\mathbf{i}} + (y-y_0)\hat{\mathbf{j}} + (z-z_0)\hat{\mathbf{k}}}{\sqrt{(x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2}}. \tag{21}$$

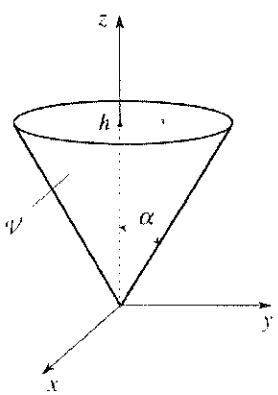


Figure 5. The cone.

The problem that we pose here is to calculate the force of attraction exerted at the origin, per unit mass at the origin (i.e., $m = 1$), by the solid right circular cone of uniform mass density σ (g/cm^3) shown in Fig. 5. Since the cone is not a point mass, we need to

find the forces induced by the individual mass elements that constitute the cone and to sum them by integration. With M changed to $\sigma\, dV$ and $x_0 = y_0 = z_0 = 0$, (20) and (21) give

$$\mathbf{F}(0,0,0) = G \iiint\limits_{\mathcal{V}} \frac{x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}}{(x^2+y^2+z^2)^{3/2}}\, \sigma\, dV. \tag{22}$$

Surely, the symmetry about the z axis implies that the x and y components of $\mathbf{F}$ will turn out to be zero so that (22) can be simplified to

$$\mathbf{F}(0,0,0) = G\sigma\hat{\mathbf{k}} \iiint\limits_{\mathcal{V}} \frac{z}{(x^2+y^2+z^2)^{3/2}}\, dV. \tag{23}$$

In what coordinate system shall we work out the integral? That is, how shall we parametrize the conical region? Spherical coordinates seem like a good choice inasmuch as the conical surface is a constant-ϕ coordinate surface. However, the top surface, $z = h$, is *not* a constant-coordinate surface. That is, $\rho \neq$ constant, $\phi \neq$ constant, and $\theta \neq$ constant on $z = h$. Using cylindrical coordinates the top surface *is* a constant-coordinate surface (namely, $z = h$) but the conical surface is not. Thus, the choice seems to be "six of one, half a dozen of the other." Let us work the problem both ways.

Using cylindrical coordinates: That is, choose u, v, w to be the familiar cylindrical coordinates r, θ, z, respectively. Then $dV = r\, dr\, d\theta\, dz$ and $x^2 + y^2 = r^2$ so (23) becomes

$$\begin{aligned}
\mathbf{F} &= \sigma G\hat{\mathbf{k}} \int_0^h \int_0^{2\pi} \int_0^{z\tan\alpha} \frac{zr\, dr\, d\theta\, dz}{(r^2+z^2)^{3/2}} \\
&= 2\pi\sigma G\hat{\mathbf{k}} \int_0^h \int_{u=z^2}^{u=z^2(1+\tan^2\alpha)} \frac{z\, du\, dz}{2u^{3/2}} \qquad (u = r^2 + z^2) \\
&= 2\pi\sigma G\hat{\mathbf{k}} \int_0^h z\left(-\frac{1}{\sqrt{u}}\right)\Bigg|_{z^2}^{z^2/\cos^2\alpha} dz \\
&= 2\pi\sigma G\hat{\mathbf{k}} \int_0^h z\left(-\frac{\cos\alpha}{z} + \frac{1}{z}\right) dz,
\end{aligned} \tag{24}$$

so that

$$\mathbf{F} = 2\pi\sigma G h(1 - \cos\alpha)\hat{\mathbf{k}}. \tag{25}$$

In case the $z\tan\alpha$ limit was not clear, note that $r/z = \tan\alpha$ on the conical surface.

Using spherical coordinates: This time choose u, v, w to be the spherical coordinates ρ, ϕ, θ, respectively. Then $dV = \rho^2 \sin\phi\, d\rho\, d\phi\, d\theta$, $z = \rho\cos\phi$, and $x^2 + y^2 + z^2 = \rho^2$ so (23) becomes

$$\begin{aligned}
\mathbf{F} &= \sigma G\hat{\mathbf{k}} \int_0^{2\pi} \int_0^{\alpha} \int_0^{h/\cos\phi} \frac{\rho\cos\phi}{(\rho^2)^{3/2}}\, \rho^2 \sin\phi\, d\rho\, d\phi\, d\theta \\
&= 2\pi\sigma G\hat{\mathbf{k}} \int_0^{2\pi} \cos\phi \sin\phi \frac{h}{\cos\phi}\, d\phi = 2\pi\sigma G h(1 - \cos\alpha)\hat{\mathbf{k}},
\end{aligned} \tag{26}$$

which result agrees with (25).

COMMENT. As a partial check, observe from (26) that $\mathbf{F} \to \mathbf{0}$ as $\alpha \to 0$, as it should. ∎

Closure. With a three-dimensional region generated by a three-parameter family $x = x(u, v, w)$, $y = y(u, v, w)$, $z = z(u, v, w)$, and the position vector given by $\mathbf{R} = x = x(u, v, w)\hat{\mathbf{i}} + y(u, v, w)\hat{\mathbf{j}} + z(u, v, w)\hat{\mathbf{k}}$, the volume element dV is defined as the volume of the parallelipiped with edges $\mathbf{R}_u\, du$, $\mathbf{R}_v\, dv$, $\mathbf{R}_w\, dw$ as shown in Fig. 1. Thus,

$$\begin{aligned} dV &= |(\mathbf{R}_u\, du) \cdot (\mathbf{R}_v\, dv) \times (\mathbf{R}_w\, dw)| \\ &= |\mathbf{R}_u \cdot \mathbf{R}_v \times \mathbf{R}_w|\, du\, dv\, dw \\ &= \left\| \begin{matrix} x_u & y_u & z_u \\ x_v & y_v & z_v \\ x_w & y_w & z_w \end{matrix} \right\| du\, dv\, dw \\ &= \left| \frac{\partial(x, y, z)}{\partial(u, v, w)} \right| du\, dv\, dw. \end{aligned}$$

That is, dV is the absolute magnitude of the Jacobian of x, y, z with respect to u, v, w times $du\, dv\, dw$.

EXERCISES 15.6

1. Compute the volume of a sphere of radius R by triple integration:

(a) using cylindrical coordinates
(b) using spherical coordinates.

2. Show by triple integration that the volume of the right circular cone shown in Fig. 5 is $V = (\pi h^3/3)\tan^2\alpha$:

(a) using cylindrical coordinates
(b) using spherical coordinates.

3. Evaluate the gravitational force $\mathbf{F}$ given by (23), this time letting $\mathcal{V}$ be the hemisphere $x^2 + y^2 + z^2 \le R^2$, for $0 \le z \le R$

(a) using cylindrical coordinates
(b) using spherical coordinates.

4. (*Mass and center of gravity*) First, read Exercise 4 of Section 15.3. There, the mass was a two-dimensional distribution over a region $\mathcal{R}$ in the x, y plane. Here, we consider a distribution over a three-dimensional region $\mathcal{R}$ in x, y, z space. If the mass density is $\sigma(x, y, z)$ [i.e., $\sigma(x, y, z)$ is the mass per unit volume at x, y, z], then the x, y, z coordinates of the **center of gravity** are defined as

$$\begin{aligned} x_c &= \frac{1}{M} \iiint_{\mathcal{R}} x\sigma(x, y, z)\, dV, \\ y_c &= \frac{1}{M} \iiint_{\mathcal{R}} y\sigma(x, y, z)\, dV, \\ z_c &= \frac{1}{M} \iiint_{\mathcal{R}} z\sigma(x, y, z)\, dV, \end{aligned} \tag{4.1}$$

where

$$M = \iiint_{\mathcal{R}} \sigma(x, y, z)\, dV$$

is the total mass. In each of the following, take the density σ to be a constant.

(a) Let $\mathcal{R}$ be the solid cone considered in Example 3. Compute z_c.
(b) Let $\mathcal{R}$ be one eighth of the sphere $x^2 + y^2 + z^2 \le R^2$, namely, that part which lies within $x \ge 0, y \ge 0, z \ge 0$. Compute x_c, y_c, z_c.
(c) Let $\mathcal{R}$ be that part of the sphere $x^2 + y^2 + z^2 \le R^2$

which lies between $z = a$ and $z = R$ (where $0 < a < R$). Compute z_c.
(d) Let $\mathcal{R}$ be the spherical sector $0 \le \rho \le R$, $0 \le \phi \le \alpha$, $0 \le \theta < 2\pi$. Compute z_c.

5. (*Moments of inertia*) First, read Exercise 13 of Section 15.3. In each of the following, take the density σ to be a constant. Let M be the total mass.

(a) Consider a hollow sphere $a^2 \le x^2 + y^2 + z^2 \le b^2$. Show that

$$I_x = \frac{2}{5} M \frac{b^5 - a^5}{b^3 - a^3}.$$

(b) In the limit $a \to b$, the hollow sphere in part (a) is called a "thin shell." For a thin shell of radius b, show that the answer to part (a) reduces to the form $I_x = \frac{2}{3}Mb^2$.
(c) Consider a hollow right circular cylinder $a^2 \le x^2 + y^2 \le b^2$, $0 \le z \le h$. Show that $I_x = (M/12)(3b^2 + 3a^2 + 4h^2)$.
(d) For the cylinder in part (c), show that $I_z = (M/2)(a^2+b^2)$.
(e) For the right circular cone shown in Fig. 5, show that $I_z = (3M/10)(h \tan\alpha)^2$.
(f) For the right circular cone shown in Fig. 5, show that $I_x = (3Mh^2/20)(\tan^2\alpha + 4)$.

6. In the case of cylindrical coordinates, we set $u = r$, $v = \theta$, $w = z$. Show that the result (8) would obtain if, instead, we set $u = \theta$, $v = r$, $w = z$, say.

7. (*Gravitational force induced by hollow sphere*) Consider a hollow sphere, $a^2 \le x^2 + y^2 + z^2 \le b^2$, of uniform mass density σ.

(a) Show that the gravitational force of attraction induced by the hollow sphere, per unit mass at $(0, 0, z_0)$, is

$$\mathbf{F}(0,0,z_0) = \sigma G\hat{\mathbf{k}} \iiint_{\mathcal{V}} \frac{(z - z_0)dV}{[x^2 + y^2 + (z - z_0)^2]^{3/2}}. \tag{7.1}$$

(b) Carrying out the integration in (7.1), show that

$$\mathbf{F}(0,0,z_0) = \begin{cases} -\dfrac{MG}{z_0^2}\hat{\mathbf{k}} & \text{if } |z_0| > b \\ \mathbf{0} & \text{if } |z_0| < a. \end{cases} \tag{7.2}$$

NOTE: The result (7.2) is remarkable for it says that the force induced by the hollow sphere at any point outside the sphere is the same as if the entire mass of the sphere were compressed into a *point* mass M at the origin, and that there is *no* force within the cavity.
(c) Evaluate $\mathbf{F}(0, 0, z_0)$ for the case where $a < z_0 < b$. HINT: The result obtained in part (b) should be all that you need.

8. Using the results obtained in Example 3:

(a) What is the gravitational force induced by an infinite slab ($-\infty < x < \infty$, $-\infty < y < \infty$, $0 < z < h$) of uniform mass density σ, per unit mass, at any point on either face of the slab ($z = 0$ or $z = h$)?
(b) What is the gravitational force induced by an infinite half-space ($-\infty < x < \infty$, $-\infty < y < \infty$, $0 \le z < \infty$) of uniform mass density σ, per unit mass, at any point on the face $z = 0$?

Chapter 15 Review

The single most important idea about curves, surfaces, and volumes is their parametrization by

$$\mathbf{R}(\tau) = x(\tau)\hat{\mathbf{i}} + y(\tau)\hat{\mathbf{j}} + z(\tau)\hat{\mathbf{k}}, \tag{1a}$$
$$\mathbf{R}(u,v) = x(u,v)\hat{\mathbf{i}} + y(u,v)\hat{\mathbf{j}} + z(u,v)\hat{\mathbf{k}}, \tag{1b}$$
$$\mathbf{R}(u,v,w) = x(u,v,w)\hat{\mathbf{i}} + y(u,v,w)\hat{\mathbf{j}} + z(u,v,w)\hat{\mathbf{k}}, \tag{1c}$$

respectively, where $\mathbf{R}$ is the position vector. We could have used u instead of τ in (1a), to promote the (u), (u, v), (u, v, w) pattern in (1a)–(1c), but τ is more traditional for curves. As should not be surprising, all quantities of interest, for

curves, surfaces, and volumes can be expressed in terms of $\mathbf{R}$. The differential arc length, area element, and volume element are

$$ds = \sqrt{\mathbf{R}'(\tau)\cdot\mathbf{R}'(\tau)}\,d\tau, \tag{2}$$
$$dA = \|\mathbf{R}_u \times \mathbf{R}_v\|\,du\,dv, \tag{3}$$
$$dV = |\mathbf{R}_u \cdot \mathbf{R}_v \times \mathbf{R}_w|\,du\,dv\,dw, \tag{4}$$

respectively, and the normal to a surface is, to within a factor of ± 1,

$$\hat{\mathbf{n}} = \frac{\mathbf{R}_u \times \mathbf{R}_v}{\|\mathbf{R}_u \times \mathbf{R}_v\|}. \tag{5}$$

Computationally, it is important to express these results in terms of the components x, y, z of $\mathbf{R}$. Accordingly, (2)–(4) become

$$ds = \sqrt{x'^2 + y'^2 + z'^2}\,d\tau, \tag{6}$$

$$dA = \sqrt{EG - F^2}\,du\,dv,$$
$$\begin{aligned} E &= x_u^2 + y_u^2 + z_u^2 \\ F &= x_u x_v + y_u y_v + z_u z_v \\ G &= x_v^2 + y_v^2 + z_v^2, \end{aligned} \tag{7}$$

$$dV = \left\| \begin{matrix} x_u & y_u & z_u \\ x_v & y_v & z_v \\ x_w & y_w & z_w \end{matrix} \right\| du\,dv\,dw = \left| \frac{\partial(x,y,z)}{\partial(u,v,w)} \right| du\,dv\,dw. \tag{8}$$

We note two special cases of (7), one when the surface is flat and lies in the x, y plane ($z = 0$), and one when the surface is known in the form $z = f(x, y)$.

$$z = 0: \qquad dA = \left| \frac{\partial(x,y)}{\partial(u,v)} \right| du\,dv, \tag{9a}$$

$$z = f(x,y): \qquad dA = \sqrt{1 + f_x^2 + f_y^2}\,dx\,dy. \tag{9b}$$

Observe the appearance in (8) and (9a) of the Jacobian determinants.

Chapter 16

Scalar and Vector Field Theory

16.1 Introduction

A great many phenomena are governed by ordinary differential equations (ODE's). That is, there is only one independent variable, such as a space variable or time, and one or more dependent variables. If there are two or more independent variables, then the dependent variable is generally called a *field*, and the governing differential equation will be a partial differential equation (PDE), known as a *field equation*. For instance, the concentration $c(x, y, z, t)$ of a pollutant that spreads by diffusion in a lake is governed by the PDE

$$\alpha^2 \left(\frac{\partial^2 c}{\partial x^2} + \frac{\partial^2 c}{\partial y^2} + \frac{\partial^2 c}{\partial z^2} \right) = \frac{\partial c}{\partial t}, \tag{1}$$

where α^2 is a constant known as the diffusivity. As another example, let $\sigma(x, y, z, t)$ and $\mathbf{v}(x, y, z, t)$ be the mass density and velocity of a certain fluid flow, respectively. Then the principle of conservation of mass implies that σ and $\mathbf{v}$ are related through the PDE

$$\frac{\partial \sigma}{\partial t} + \frac{\partial}{\partial x}(\sigma v_x) + \frac{\partial}{\partial y}(\sigma v_y) + \frac{\partial}{\partial z}(\sigma v_z) = 0, \tag{2}$$

where v_x, v_y, v_z are the x, y, z components of $\mathbf{v}$, respectively. In this chapter we study the calculus of such scalar and vector fields.

Our plan is as follows. In Sections 16.3–16.6 we introduce the so-called divergence, gradient, and curl, which are specific differential operators that can act on scalar and vector fields. Those results are extended to non-Cartesian coordinates (cylindrical and spherical) in Section 16.7, which is optional so it can be omitted in a shorter course.

The "payoff" comes mostly in Sections 16.8–16.10, on the Gauss divergence theorem, Stokes's theorem, and irrotational fields. For instance, in Section 16.8 we derive (1) and (2) above, using the divergence theorem. Thus, one of the major outcomes of this chapter is the derivation of some of the important field equations

of mathematical physics. After pausing in Chapter 17 to develop the Fourier series and Fourier transform, we will then come back to those field equations and learn how to solve them – in Chapters 18–20 on PDE's. We continue to treat the slightly more difficult non-Cartesian cases as optional, in Sections 16.8–16.10, by including them only in optional subsections.

16.2 Preliminaries

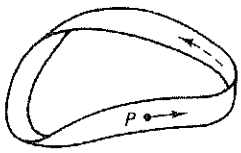


Figure 1. Möbius band.

16.2.1. Topological considerations. Throughout this chapter we ask that the curves, surfaces, and volumes under consideration be sufficiently "decent" for our purposes. Specifically, we ask the curves to be piecewise smooth (Section 15.2.1). We ask the surfaces to be piecewise smooth too (Section 15.3.2) but, in addition, we ask them to be orientable. Since nonorientable surfaces are rarely encountered in applications, and since a rigorous definition would be a long story, we will define the term only in an intuitive way. Namely, we say that a surface is **orientable** if it is two-sided, that is, if we can paint one side green and the other side blue. The classic example of a *non*orientable surface is the *Möbius band*, which can be constructed by taking a rectangular strip of paper, twisting one end through 180°, then taping that end to the other (Fig. 1). If an insect starts at any point P on one side of the band and follow the arrows, it eventually visits the entire surface, so that the Möbius band has only one side. Our objection to such a surface is that we are going to "orient" a given surface $\mathcal{S}$ by defining a normal vector $\hat{\mathbf{n}}$ at some point on $\mathcal{S}$ and then extending the field of normals continuously over the rest of $\mathcal{S}$. But when we do that with the Möbius band, say, we end up with *two* vectors, $\hat{\mathbf{n}}$ and $-\hat{\mathbf{n}}$, at each point on $\mathcal{S}$.

Finally, we ask that the volumes under consideration be regions (Section 13.2.2) that are bounded by piecewise smooth orientable surfaces.

16.2.2. Scalar and vector fields. The objects of interest in this chapter are scalar fields and vector fields. By a **scalar field** we shall mean a *scalar valued function*, say f, defined over a region $\mathcal{R}$, which is a one-, two-, or three-dimensional connected subset of 3-space. Besides the space variables, f may depend on the time t as well. Normally, f will be real valued.

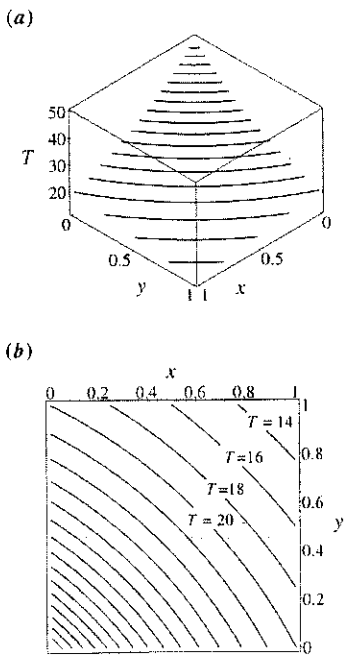


Figure 2. Plots of $T(x, y)$.

EXAMPLE 1. *Temperature Field.* Let $T(x, y)$ be the temperature at any given point (x, y) in a square plate which extends over $0 \le x \le 1$, $0 \le y \le 1$, and let

$$T(x, y) = \frac{100}{(x+1)^2 + (y+1)^2}. \tag{1}$$

In this case the region $\mathcal{R}$ is the two-dimensional set $0 \le x \le 1$, $0 \le y \le 1$, and the scalar function is $T(x, y)$, given by (1). By way of graphical display, we have used *Maple* to obtain the three-dimensional plot of T, above the x, y plane, that is given in Fig. 2a. Plotted are a number of **level curves** of T, curves along which $T(x, y)$ is a constant, like level roads on a mountain. The projection of those curves down onto the x, y plane give the

level curves shown in Fig. 2b. We can see from (1) that the level curves of T, in the x, y plane, are the concentric circles $(x+1)^2 + (y+1)^2 =$ constant, centered at $(-1, -1)$. ∎

In Example 1 we encountered level curves. More generally, we say that the **level set** of value c, of a function $f(x_1, \ldots, x_n)$, is the set of points $(x_1, \ldots, x_n)$ for which $f(x_1, \ldots, x_n) = c$. If $n = 2$, the level set is, in general, a level *curve*, and if $n = 3$ the level set is, in general, a level *surface*. To see why we say "in general," consider a few examples for the case where $n = 2$. Let

$$f(x, y) = \ln(x^2 + y^2), \qquad g(x, y) = \sin(x + 3y), \qquad h(x, y) = 6,$$

each defined over the whole plane. It is not hard to see that the level sets of f are concentric circles centered at the origin. Specifically, for any given value c $(-\infty < c < \infty)$ the level set of f is the circle $x^2 + y^2 = e^c$. Turning to g, the level set of value $\frac{1}{2}$, say, is not just a single curve, it is the *set* of straight lines $x + 3y = \pi/6 + 2n\pi$ $(n = 0, \pm 1, \pm 2, \ldots)$, and the level set of value 4, say, is *empty*. Finally, turning to h, the level set of value 2, say, is empty, whereas the level set of value 6 is the whole plane!

Just as a scalar field is a scalar valued function defined over a given region, a *vector valued function* defined over a given region is called a **vector field**.*

EXAMPLE 2. *Fluid Velocity Field.* Imagine a steady, uniform flow of fluid (such as air or water), at a speed U meters/second, parallel to a wall as shown in Fig. 3. That is, at every instant and at every point (x, y) in the region (namely, $-\infty < x < \infty, 0 < y < \infty$), the fluid velocity $\mathbf{v}$ is equal to $U\hat{\mathbf{i}}$, where U is a constant. Then the vector function

$$\mathbf{v} = U\hat{\mathbf{i}}, \tag{2}$$

in the stated region, constitutes a vector field – a simple one, but a vector field nonetheless.

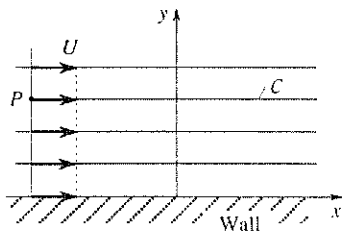


Figure 3. Steady, uniform flow field.

Since the flow is steady, every fluid particle that passes through the point P (Fig. 3), at one time or another, must follow the same path, namely, the curve C. Fluid mechanicists call such curves *streamlines*. (In general, streamlines do not exist if the flow is *un*steady.) Thus we speak of the family of horizontal lines in Fig. 3 as the *streamline pattern*.

As a more complicated case, let the wall have a semicircular bump of radius a, centered at the origin, and let the fluid velocity field be

$$\mathbf{v} = U\hat{\mathbf{i}} + \frac{Ua^2}{(x^2 + y^2)^2}\left[(y^2 - x^2)\hat{\mathbf{i}} - 2xy\hat{\mathbf{j}}\right], \qquad (a^2 \le x^2 + y^2 < \infty) \tag{3}$$

as depicted in Fig. 4. At the top of the bump, for example, where $x = 0$ and $y = a$,

$$\mathbf{v}(0, a) = U\hat{\mathbf{i}} + \frac{Ua^2}{a^4}(a^2\hat{\mathbf{i}} - 0\hat{\mathbf{j}}) = 2U\hat{\mathbf{i}}.$$

Furthermore, it follows from (3) that $\mathbf{v} = 0$ at $(\pm a, 0)$, which points are therefore called *stagnation points* of the flow.

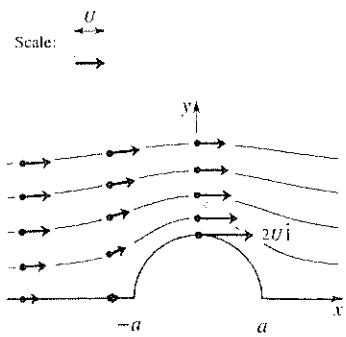


Figure 4. The flow field (3).

*That is, a vector field $\mathbf{v}$ on a region $\mathcal{R}$, which is a subset of $\mathbb{R}^n$, is a transformation or mapping $\mathbf{v} : \mathcal{R} \subset \mathbb{R}^n \to \mathbb{R}^n$. In this chapter n will be 2 or 3, the general case is considered in J. E. Marsden and A. J. Tromba, *Vector Calculus* (San Francisco: W. H. Freeman, 1976). Similarly, a scalar field f on a region $\mathcal{R}$, which is a subset of $\mathbb{R}^n$, is a mapping $f : \mathcal{R} \subset \mathbb{R}^n \to \mathbb{R}^1$.

COMMENT. In Section 16.10 we derive the PDE governing the flow field shown in Fig. 4. Later, when we study PDE's, we will solve that PDE and obtain the result stated in (3). In fact, we will solve the same problem again, in Section 23.6, using the method of conformal mapping. ∎

Closure. Scalar and vector fields are scalar and vector valued *functions*, respectively, defined on one-, two-, or three-dimensional regions of space. They may be constant or nonconstant, and they may or may not vary with the time t as well as the space variables. Examples of scalar fields are the temperature fields $T(x, y, z) = 100(x^2 + y^2 + 5z^2)$ and $T(x, t) = 50(\sin x)e^{-2t}$, and examples of vector fields are the fluid velocity fields $\mathbf{v}(x, y) = 4xy\hat{\mathbf{i}} - (x + 3y^2)\hat{\mathbf{j}}$ and $\mathbf{v}(x, y, z, t) = 10\hat{\mathbf{i}} + x^3yz \sin t\hat{\mathbf{k}}$.

Computer software. The computer plots in Fig. 2 were obtained using the *Maple* **plot3d** and **contourplot** commands, as follows. For Fig. 2a we used

```
with(plots):
plot3d(100/((x + 1)^2 + (y + 1)^2), x = 0..1, y = 0..1, style = contour);
```

and for Fig. 2b we used

```
with(plots):
contourplot(100/((x + 1)^2 + (y + 1)^2), x = 0..1, y = 0..1);
```

If we omit the style = contour option in the first, then instead of the plot being comprised of level curves it would, as the default condition, be comprised of constant-x and constant-y curves. That is, the display would be the u, v coordinate curves under the parametrization $x = u$ and $y = v$.

EXERCISES 16.2

1. Let $T(x, y)$ be a temperature field in $|x| \le 10$, $0 \le y \le 10$. Determine, and sketch the $T = 0, 20, 40, 60$ level sets. If the level set is empty, state that.

(a) $T = x^2 + y^2$ (b) $T = 10(x - y)$
(c) $T = 10(x + y)$ (d) $T = x^2 - y^2$
(e) $T = 2x^2$ (f) $T = 100/(x^2 + y^2 + 1)$
(g) $T = x + 20y$ (h) $T = x^2y^2/20$
(i) $T = 60 \sin(\pi xy/128)$ (j) $T = 53$
(k) $T = \sin[(x^2 + y^2)/24]$ (l) $T = 20$

2. The position vector $\mathbf{R} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}}$ to any point (x, y) in the x, y plane constitutes a vector field. Draw the $\mathbf{R}$ vector (to scale) at eight points that are equally spaced around the circle $r = 1$, and at eight points that are equally spaced around the circle $r = 3$.

3. Consider the vector field $\mathbf{v} = x\hat{\mathbf{i}} - y\hat{\mathbf{j}}$ in $0 \le x < \infty$, $0 \le y < \infty$.

(a) On a single graph, sketch the $\mathbf{v}$ vectors at the 16 points (m, n), where m and n are integers such that $0 \le m \le 3$ and $0 \le n \le 3$. (Here it will help to assign a convenient scale, as we have done in Fig. 4, so that the arrow representations of the vectors will not lie on top of each other.)

(b) Determine the curves along which $\mathbf{v}$ has constant magnitude and those along which $\mathbf{v}$ has constant direction. Are these results in agreement with your sketch of the field in part (a)?

4. Given the vector field $\mathbf{w}(x, y, z)$, determine parametric equations $x = x(\tau)$, $y = y(\tau)$, $z = z(\tau)$ for the curve(s) through the point $(2, 3, -1)$ along which $\mathbf{w}$ has constant direction.

(a) $\mathbf{w} = y\hat{\mathbf{i}} - x\hat{\mathbf{j}} + z\hat{\mathbf{k}}$
(b) $\mathbf{w} = (y + 2z)\hat{\mathbf{i}} - x\hat{\mathbf{j}} + (x + y)\hat{\mathbf{k}}$
(c) $\mathbf{w} = x^2\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$
(d) $\mathbf{w} = yz\hat{\mathbf{i}} - \hat{\mathbf{j}} - (x + 2)\hat{\mathbf{k}}$
(e) $\mathbf{w} = z^2\hat{\mathbf{i}} - 3\hat{\mathbf{j}} - (y - x)\hat{\mathbf{k}}$
(f) $\mathbf{w} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$

5. (a) For the velocity field defined by (3), determine the *speed* $\|\mathbf{v}\|$ as a function of x, along the wall (namely, $y = 0$ for $|x| \geq a$ and $x^2 + y^2 = a^2$ for $|x| \leq a$).
(b) Show that as $r = \sqrt{x^2 + y^2} \to \infty$ along any ray $y = mx$, the velocity $\mathbf{v}$ tends to the uniform flow $U\hat{\mathbf{i}}$, which flow is called the *free stream*.

6. Recall the plane polar coordinate base vectors $\hat{\mathbf{e}}_r(\theta)$ and $\hat{\mathbf{e}}_\theta(\theta)$ from Section 14.6. Each of these is a vector field because there is an $\hat{\mathbf{e}}_r$ vector and an $\hat{\mathbf{e}}_\theta$ vector defined at each point in the plane (except the origin).

(a) Sketch the $\hat{\mathbf{e}}_r(\theta)$ field.
(b) Sketch the $\hat{\mathbf{e}}_\theta(\theta)$ field.

7. Use computer software to obtain plots of the four streamlines shown in Fig. 4, taking $U = a = 1$ and taking their initial points as $(-3.5, 0.5)$, $(-3.5, 1)$, $(-3.5, 1.5)$, and $(-3.5, 2)$. HINT: Equation (3) gives us the ODEs

$$\begin{aligned} x'(t) &= 1 + \frac{y^2 - x^2}{(x^2 + y^2)^2}, \\ y'(t) &= -\frac{2xy}{(x^2 + y^2)^2} \end{aligned} \tag{7.1}$$

for the motion $x(t), y(t)$ of any given fluid particle.

16.3 Divergence

Fundamental in scalar and vector field theory are certain differential operators, known as the divergence, gradient, and curl, that operate on those fields. We will introduce the divergence in this section and the gradient and curl in the next two.

Let $\mathbf{v}$ be a vector field defined in a region $\mathcal{R}$. The physical nature of $\mathbf{v}$ is immaterial here, but for definiteness let us think of $\mathbf{v}$ as a fluid velocity field. Focusing our attention on a particular point $P = (x, y, z)$ in the flow, let us introduce a **control volume** around P, as shown in Fig. 1, and denote it as $\mathcal{B}$. For example, $\mathcal{B}$ might be chosen to be a prism, a sphere, or an arbitrary "potato" shape as in the figure. A control volume is only a mathematical region rather than a physical presence, and it is normally introduced so that we can keep track of the flux of some quantity of interest such as mass, electric charge, or heat.

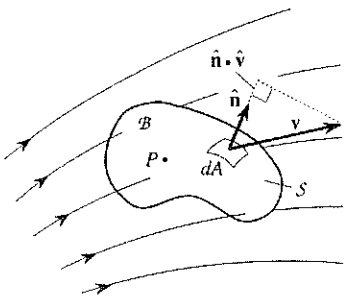


Figure 1. Control volume.

For instance, consider the integral*

$$\boxed{I = \int_S \hat{\mathbf{n}} \cdot \mathbf{v}\, dA,} \tag{1}$$

*In Chapter 15 we denote double and triple integrals by $\iint$ and $\iiint$, respectively. In this section we switch to the more compact $\int$ notation used in (1) unless, of course, integration limits are to be spelled out. Another point of notation is that some authors prefer to combine $\hat{\mathbf{n}}$ and dA in (1) as $\hat{\mathbf{n}}\, dA = d\mathbf{A}$: then (1) becomes $\int_S \mathbf{v} \cdot d\mathbf{A}$, which is a bit more compact than (1), but we will not use that notation.

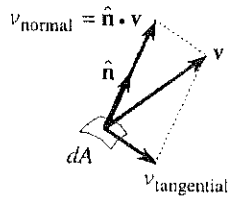


Figure 2. Flux across dA.

where $\hat{\mathbf{n}}$ is a unit outward normal vector at each point on $\mathcal{S}$, as illustrated, at a representative point in Fig. 1. The set of all such $\hat{\mathbf{n}}$ vectors on $\mathcal{S}$ would look something like the quills on a frightened porcupine. What is the physical significance of I? If, at a given point on the surface $\mathcal{S}$ of $\mathcal{B}$, we break the fluid velocity $\mathbf{v}$ into normal and tangential components (Fig. 2), then the flux across dA is due entirely to the normal component $\hat{\mathbf{n}} \cdot \mathbf{v}$ since the tangential flow is along the surface and therefore does not cross it. Then the outflow through the surface element dA is the normal velocity component $\hat{\mathbf{n}} \cdot \mathbf{v}$ (meters per second, say) times dA (square meters), that is, $\hat{\mathbf{n}} \cdot \mathbf{v}\, dA$ $\mathrm{m}^3/\mathrm{sec}$. Thus, integrating over the entire surface $\mathcal{S}$, I is the *net outflow*, in $\mathrm{m}^3/\mathrm{sec}$, say, across $\mathcal{S}$ – outflow because we took $\hat{\mathbf{n}}$ to be the outward unit normal. This outflow can be positive, zero, or negative. For example, if heat is extracted from the fluid and the fluid contracts, one can imagine a net *in*flow into the control volume so that the outflow will be negative.

Thus, I is an example of a flux integral. We will call it a **volume flux** because it is volume per unit time. If, besides denoting the vector velocity field as $\mathbf{v}(x, y, z, t)$, we also denote the scalar density field as $\sigma(x, y, z, t)$ (mass per unit volume), then

$$J = \int_{\mathcal{S}} \sigma \hat{\mathbf{n}} \cdot \mathbf{v}\, dA \tag{2}$$

would also be a flux, not a volume flux but a mass flux because the dimensions of $\sigma \hat{\mathbf{n}} \cdot \mathbf{v}\, dA$ are mass per unit volume times volume per unit time, hence mass per unit time.

Next, let us divide the volume flux integral in (1) by the volume V of $\mathcal{B}$ to obtain the outflow *per unit volume*. Finally, we shrink $\mathcal{B}$ down to point P and obtain the outflow per unit volume at the point P. This result is called the **divergence** of $\mathbf{v}$ at P and is defined as

$$\boxed{\operatorname{div} \mathbf{v}(P) \equiv \lim_{\mathcal{B} \to 0} \left\{ \frac{\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA}{V} \right\},} \tag{3}$$

where $\mathcal{B} \to 0$ means that $\mathcal{B}$ shrinks to the point P in such a way that the maximum linear dimension (the "diameter") of $\mathcal{B}$ tends to zero.* If we assume that $\mathbf{v}$ is C^1 and that $\mathcal{B}$ has a piecewise smooth orientable surface $\mathcal{S}$, then it can be shown that the limit in (3) does indeed exist at each point P in the field.†

Observe that $\operatorname{div} \mathbf{v}(P)$ is a scalar at each point P since $\hat{\mathbf{n}} \cdot \mathbf{v}$, dA, and V are scalars. Thus, $\operatorname{div} \mathbf{v}$ is itself a scalar field associated with the given vector field $\mathbf{v}$.

Observe, further, that (3) provides an intrinsic, or invariant, definition of $\operatorname{div} \mathbf{v}$. That is, it contains no reference to any particular coordinate system so the value of

*Observe that the limit as $\mathcal{B} \to 0$ is not the same as the limit as $V \to 0$. For instance, if we make a coin thinner and thinner, then its volume V tends to zero even though it is not true that $\mathcal{B} \to 0$.

†Let $\mathbf{v} = v_x(x, y, z, t)\hat{\mathbf{i}} + v_y(x, y, z, t)\hat{\mathbf{j}} + v_z(x, y, z, t)\hat{\mathbf{k}}$, where t is the time. By $\mathbf{v}$ being C^1 in $\mathcal{R}$, we mean that v_x, v_y and v_z are all C^1 in $\mathcal{R}$; i.e., $\partial/\partial x$, $\partial/\partial y$, and $\partial/\partial z$ of v_x, v_y, and v_z all exist and are continuous in $\mathcal{R}$. Since all the first-order partial derivatives of $\mathbf{v}$ exist and are continuous in $\mathcal{R}$, we also say that $\mathbf{v}$ is **continuously differentiable** in $\mathcal{R}$.

div $\mathbf{v}(P)$ at any given point P is uniquely determined, independent of the choice of the reference coordinate system.

Thus, the definition of div $\mathbf{v}$ given by (3) has two advantages: it readily admits a clear and simple physical interpretation of div $\mathbf{v}(P)$ as the outflow per unit volume at P, and it is invariant with respect to coordinate system. However, the limit definition (3) is of little use computationally. To obtain a computationally convenient expression, we need to introduce a reference coordinate system. Introducing a Cartesian system, since that is evidently the simplest choice, let us carry out the limit indicated in (3). Since the limit in (3) is independent of the shape of $\mathcal{B}$, let us choose the simplest shape, namely, one bounded by constant coordinate surfaces. Thus, for $\mathcal{B}$ let us choose the rectangular prism shown in Fig. 3, with $P = (x, y, z)$ at its center.

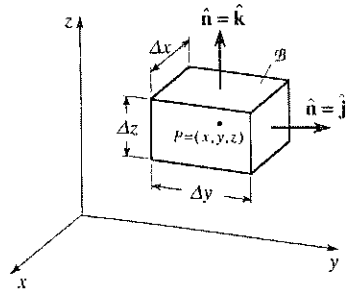


Figure 3. Cartesian coordinates.

Consider first the contribution from the front and back faces, on which the outward unit normal is $\hat{\mathbf{n}} = +\hat{\mathbf{i}}$ and $\hat{\mathbf{n}} = -\hat{\mathbf{i}}$, respectively. Let us express $\mathbf{v} = v_x(x, y, z, t)\hat{\mathbf{i}} + v_y(x, y, z, t)\hat{\mathbf{j}} + v_z(x, y, z, t)\hat{\mathbf{k}}$, where the x, y, z subscripts specify the x, y, z components of $\mathbf{v}$ rather than partial derivatives,* and where we have not included any time dependence in $\mathbf{v}$ because it will not play an active role in the present discussion. Then

$$\begin{aligned}\int_{\text{front face}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA &= \int_{\text{front face}} \hat{\mathbf{i}} \cdot \left(v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}\right) dA \\ &= \int_{\text{front face}} v_x\left(x + \frac{\Delta x}{2}, y', z'\right) dy'\, dz' \\ &= v_x\left(x + \frac{\Delta x}{2}, y_1, z_1\right) \Delta y\, \Delta z, \end{aligned} \tag{4}$$

for some point $(x + \Delta x/2, y_1, z_1)$ on the front face, where the last equality in (4) follows from the mean value theorem (4c) in Section 15.3. [To distinguish the x, y, z coordinates of the fixed point $P = (x, y, z)$ from the dummy variables of integration, we use x', y', z' for the latter.] Similarly,

$$\begin{aligned}\int_{\text{back face}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA &= \int_{\text{back face}} -\hat{\mathbf{i}} \cdot \left(v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}\right) dA \\ &= -v_x\left(x - \frac{\Delta x}{2}, y_2, z_2\right) \Delta y\, \Delta z \end{aligned} \tag{5}$$

for some point $(x - \Delta x/2, y_2, z_2)$ on the back face. Since the volume of $\mathcal{B}$ is $V = \Delta x\, \Delta y\, \Delta z$, it follows that

$$\begin{aligned}&\lim_{\mathcal{B}\to 0}\left\{\frac{\int_{\text{front face+back face}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA}{V}\right\} \\ &\quad = \lim_{\Delta x, \Delta y, \Delta z \to 0}\left\{\frac{\left[v_x\left(x + \Delta x/2, y, z\right) - v_x\left(x - \Delta x/2, y, z\right)\right] \Delta y \Delta z}{\Delta x\, \Delta y\, \Delta z}\right\} \\ &\quad = \lim_{\Delta x \to 0} \frac{v_x\left(x + \Delta x/2, y, z\right) - v_x\left(x - \Delta x/2, y, z\right)}{\Delta x} = \frac{\partial v_x}{\partial x}. \end{aligned} \tag{6}$$

*Thus, partial derivatives will be expressed as $\partial(\ )/\partial x$, $\partial(\ )/\partial y$, and so on.

Similarly, the top and bottom faces contribute $\partial v_z/\partial z$, and the right and left faces contribute $\partial v_y/\partial y$, so that

$$\boxed{\operatorname{div} \mathbf{v} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}.} \tag{7}$$

Analogous expressions are given in Section 16.7, for cylindrical and spherical coordinates.

EXAMPLE 1. Evaluate the divergence of the field $\mathbf{v} = (x^2/z)\hat{\mathbf{i}} - 3\hat{\mathbf{j}} + yz\hat{\mathbf{k}}$ at the point $P = (2, 9, -1)$. First, let us check to see if $\mathbf{v}$ is C^1 in some region $\mathcal{R}$ containing P. The partial derivatives of $v_x = x^2/z$, $v_y = -3$ and $v_z = yz$ with respect to x, y, and z exist and are continuous functions of x, y, z everywhere except on the plane $z = 0$, where $\partial v_x/\partial x = 2x/z$ and $\partial v_x/\partial z = -x^2/z^2$ are undefined ("blow up"). But P is not in the plane $z = 0$ so we can say that $\mathbf{v}$ is indeed C^1 in a region $\mathcal{R}$ containing P, where $\mathcal{R}$ is the half-space $-\infty < x < \infty$, $-\infty < y < \infty$, $-\infty < z < 0$. Then (7) gives

$$\operatorname{div} \mathbf{v} = \frac{\partial}{\partial x}\left(\frac{x^2}{z}\right) + \frac{\partial}{\partial y}(-3) + \frac{\partial}{\partial z}(yz) = \frac{2x}{z} + y,$$

so at P we have $\operatorname{div} \mathbf{v} = -4 + 9 = 5$.

COMMENT. For brevity, in subsequent examples we will not verify explicitly that $\mathbf{v}$ is C^1. ∎

EXAMPLE 2. If $\mathbf{v} = y^3 zt\hat{\mathbf{j}}$, where t is the time, then

$$\operatorname{div} \mathbf{v} = \frac{\partial}{\partial x}(0) + \frac{\partial}{\partial y}(y^3 zt) + \frac{\partial}{\partial z}(0) = 3y^2 zt.$$

That is, t is treated as a constant because only space derivatives are involved. ∎

Rewriting (7) as

$$\operatorname{div} \mathbf{v} = \left(\hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}\right) \cdot (v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}) \tag{8}$$

suggests that we define a vector differential operator

$$\boxed{\nabla \equiv \hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z},} \tag{9}$$

in which case we can write $\operatorname{div} \mathbf{v}$ in the operator form

$$\boxed{\operatorname{div} \mathbf{v} = \nabla \cdot \mathbf{v}.} \tag{10}$$

We will have more to say about the $\nabla \cdot \mathbf{v}$ operation in subsequent sections, but one point that must be mentioned immediately is that $\nabla \cdot \mathbf{v}$ is not a dot product in the usual sense. Specifically, it is *not* true that $\nabla \cdot \mathbf{v}$ equals $\|\nabla\|$ times $\|\mathbf{v}\|$ times the cosine of the angle between ∇ and $\mathbf{v}$ because $\nabla = \hat{\mathbf{i}}(\partial/\partial x) + \hat{\mathbf{j}}(\partial/\partial y) + \hat{\mathbf{k}}(\partial/\partial z)$ is a vector *differential operator*, not a vector with a length and a direction. Thus, $\|\nabla\|$ is not defined nor is the "angle between ∇ and $\mathbf{v}$." Furtheremore, $\nabla \cdot \mathbf{v}$ is *not* the same as $\mathbf{v} \cdot \nabla$ since

$$\begin{aligned}\nabla \cdot \mathbf{v} &= \left(\hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}\right) \cdot (v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{k}} + v_z\hat{\mathbf{k}}) \\ &= \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}, \end{aligned} \tag{11}$$

whereas

$$\begin{aligned}\mathbf{v} \cdot \nabla &= (v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{k}} + v_z\hat{\mathbf{k}}) \cdot \left(\hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}\right) \\ &= v_x\frac{\partial}{\partial x} + v_y\frac{\partial}{\partial y} + v_z\frac{\partial}{\partial z}. \end{aligned} \tag{12}$$

The symbol ∇, introduced by Hamilton,* is read as *del*, sometimes as *nabla* because it looks like an ancient Hebrew instrument of that name.

Closure. Given any vector field $\mathbf{v}$ (that is C^1 in the region $\mathcal{R}$ of interest), the scalar valued divergence of $\mathbf{v}$ at any point P is defined by (3), and has the physical significance of being the outflow per unit volume at P (hence the name "divergence"). Even if $\mathbf{v}$ is not a fluid velocity field (for instance, it might be an electric field, a gravitational force field, etc.) we can always *think* of it as a fluid velocity field and of div $\mathbf{v}$ as the outflow per unit volume. For computational purposes, we rely on the form (7) for Cartesian coordinates or on an analogous form for a cylindrical or spherical coordinate system. Finally, we see in (10) that the divergence of $\mathbf{v}$ amounts to a $\nabla \cdot$ operation on $\mathbf{v}$, where ∇ is the vector differential operator defined in (9); ∇ is our link to the gradient and curl in Sections 16.4 and 16.5.

EXERCISES 16.3

1. Work out div $\mathbf{v}$ for the following vector fields, each of which is defined over all of 3-space, say. Further, evaluate div $\mathbf{v}$ at $P = (3, -1, 4)$, and verify that $\mathbf{v}$ is C^1 in some region $\mathcal{R}$ containing P. (t is the time.)

(a) $\mathbf{v} = a\hat{\mathbf{i}} + b\hat{\mathbf{j}} + c\hat{\mathbf{k}}$ (a, b, c constants)
(b) $\mathbf{v} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$
(c) $\mathbf{v} = x\hat{\mathbf{i}} - y\hat{\mathbf{j}} + z^2t\hat{\mathbf{k}}$
(d) $\mathbf{v} = xyz(\hat{\mathbf{i}} + \hat{\mathbf{j}} - 3\hat{\mathbf{k}})$

**William R. Hamilton* (1805–1865) was both a great mathematician and a great physicist. While still an undergraduate, he was named Professor of Astronomy at Trinity College and Royal Astronomer of Ireland.

(e) $\mathbf{v} = xy\hat{\mathbf{i}} - 2(x^2 + z^2)\hat{\mathbf{k}}$
(f) $\mathbf{v} = \hat{\mathbf{i}} + \hat{\mathbf{j}} + z\hat{\mathbf{k}}$
(g) $\mathbf{v} = z\hat{\mathbf{i}} + x^2t^2\hat{\mathbf{j}}$
(h) $\mathbf{v} = \sin(x^2 + y^2 + z^2)\hat{\mathbf{j}}$

2. Determine the divergence $\operatorname{div}\mathbf{v}$ of the fluid flow field $\mathbf{v}$ given by (3) in Section 16.2.

3. Make up a vector field $\mathbf{v}(x, y, z)$ such that

(a) $\nabla \cdot \mathbf{v} = 0$ everywhere
(b) $\nabla \cdot \mathbf{v} > 0$ everywhere
(c) $\nabla \cdot \mathbf{v} < 0$ everywhere
(d) $\nabla \cdot \mathbf{v} > 0$ in $x^2 + y^2 + z^2 > 1$ and $\nabla \cdot \mathbf{v} < 0$ in $x^2 + y^2 + z^2 < 1$
(e) $\nabla \cdot \mathbf{v} > 0$ in $|x| > 1$ and $\nabla \cdot \mathbf{v} < 0$ in $|x| < 1$

4. This exercise is to promote understanding of the limit definition (3). Specifically, we ask you to evaluate $\operatorname{div}\mathbf{v}$ at $P = (0, 0, 0)$, say, by actually evaluating $(\int_S \hat{\mathbf{n}} \cdot \mathbf{v}\, dA)/V$ and taking the limit as $\mathcal{B} \to 0$. Take $\mathcal{B}$ to be the cube $|x| \le \epsilon$, $|y| \le \epsilon$, $|z| \le \epsilon$. Show that your result agrees with that obtained (much more readily) from (7). NOTE: Do not merely mimic our steps (4)–(6). Rather, actually compute $\int \hat{\mathbf{n}} \cdot \mathbf{v}\, dA$ on each of the six faces, add the results, divide by $V = 8\epsilon^3$, and take the limit as $\epsilon \to 0$.

(a) $\mathbf{v} = 2\hat{\mathbf{i}} - \hat{\mathbf{j}} + 4\hat{\mathbf{k}}$
(b) $\mathbf{v} = 3\hat{\mathbf{i}} + 4\hat{\mathbf{j}} - 2\hat{\mathbf{k}}$
(c) $\mathbf{v} = 5xe^z\hat{\mathbf{i}}$
(d) $\mathbf{v} = (x + 1)\sin y\hat{\mathbf{j}}$
(e) $\mathbf{v} = x\hat{\mathbf{i}} + 2y\hat{\mathbf{j}} - 4z^3\hat{\mathbf{k}}$
(f) $\mathbf{v} = (x^2 - 4x + yz^2)\hat{\mathbf{i}} + \hat{\mathbf{j}}$

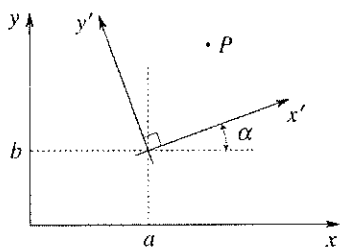
5. (*Invariance property of* $\operatorname{div}\mathbf{v}$) Let a C^1 vector field $\mathbf{v}$ be represented in terms of some particular Cartesian x, y frame as $\mathbf{v} = v_x(x, y)\hat{\mathbf{i}} + v_y(x, y)\hat{\mathbf{j}}$. (We limit ourselves to the two-dimensional case merely to reduce the algebra; the story is essentially identical for three-dimensional fields.) Then, as derived in this section,

$$\operatorname{div}\mathbf{v} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y}. \tag{5.1}$$

As emphasized in the text, this number will be the same, at a given point $P = (x, y)$, independent of the choice of the location and orientation of the reference coordinate system. In other words, it should be equally true that

$$\operatorname{div}\mathbf{v} = \frac{\partial v_{x'}}{\partial x'} + \frac{\partial v_{y'}}{\partial y'} \tag{5.2}$$

for any values of a, b and α (see the accompanying figure).



Here we ask you to *show* that this is true, namely, that

$$\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} = \frac{\partial v_{x'}}{\partial x'} + \frac{\partial v_{y'}}{\partial y'}. \tag{5.3}$$

HINT: First, show that

$$\begin{aligned} v_x &= (\cos\alpha)v_{x'} - (\sin\alpha)v_{y'}, \\ v_y &= (\text{etc.})v_{x'} - (\text{etc.})v_{y'}, \\ x' &= (\cos\alpha)(x - a) + (\sin\alpha)(y - b), \\ y' &= (\text{etc.})(x - a) + (\text{etc.})(y - b). \end{aligned}$$

(We leave the etceteras for you to determine.) Then use chain differentiation to express $\partial/\partial x$ and $\partial/\partial y$ as linear combinations of $\partial/\partial x'$ and $\partial/\partial y'$.

16.4 Gradient

We found in Section 16.3 that the divergence of a vector field $\mathbf{v}$ can be expressed in the form $\operatorname{div}\mathbf{v} = \nabla \cdot \mathbf{v}$, where ∇ is a vector differential operator which, in Cartesian coordinates, is given by

$$\nabla = \hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}. \tag{1}$$

Thus, dotting ∇ into a vector field $\mathbf{v}$ produces a scalar field associated with $\mathbf{v}$, namely, the divergence of $\mathbf{v}$. With ∇ in hand, we wonder if it might not be profitable to consider other possible actions of ∇, specifically, ∇u (where u is a scalar field) and $\nabla \times \mathbf{v}$ (where $\mathbf{v}$ is a vector field). In fact, these combinations will be of great importance. Just as $\nabla \cdot \mathbf{v}$ has a name ("the divergence of $\mathbf{v}$," or div $\mathbf{v}$, for short) which suggests its physical significance, we call ∇u the **gradient** of u, or grad u, and we call $\nabla \times \mathbf{v}$ the **curl** of $\mathbf{v}$, or curl $\mathbf{v}$.

The pattern is as follows:*

	Input		*Output*
$\operatorname{div}\mathbf{v} = \nabla \cdot \mathbf{v}$:	vector field $\mathbf{v}$	$\rightarrow$	scalar field $\nabla \cdot \mathbf{v}$
$\operatorname{grad} u = \nabla u$:	scalar field u	$\rightarrow$	vector field ∇u
$\operatorname{curl}\mathbf{v} = \nabla \times \mathbf{v}$:	vector field $\mathbf{v}$	$\rightarrow$	vector field $\nabla \times \mathbf{v}$

In this section we consider the gradient. In the next we turn to the curl.

As stated, we define

$$\boxed{\operatorname{grad} u \equiv \nabla u,} \tag{2}$$

where u is a C^1 scalar field. In Cartesian coordinates,

$$\operatorname{grad} u = \left(\hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}\right) u$$

or

$$\boxed{\operatorname{grad} u = \frac{\partial u}{\partial x}\hat{\mathbf{i}} + \frac{\partial u}{\partial y}\hat{\mathbf{j}} + \frac{\partial u}{\partial z}\hat{\mathbf{k}}.} \tag{3}$$

EXAMPLE 1. If $u(x, y, z) = x \sin y - z^3$ and $v(x, y, z, t) = 5x^2 zt$, where t is the time, then $\nabla u = \sin y\hat{\mathbf{i}} + x \cos y\hat{\mathbf{j}} - 3z^2\hat{\mathbf{k}}$ and $\nabla v = 10xzt\hat{\mathbf{i}} + 5x^2 t\hat{\mathbf{k}}$. ∎

Directional derivative. At this point we introduce the so-called directional derivative of a scalar field $u(x, y, z)$ because it will help us to understand the gradient, and because it is important in its own right. We consider a space curve C, parametrized

*A fourth combination, $\nabla\mathbf{v}$, is possible, but is not discussed in this book. For if we write out $\nabla\mathbf{v}$, in Cartesian coordinates, we have (in two dimensions, for brevity)

$$\nabla\mathbf{v} = \left(\hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y}\right)(v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}}) = \frac{\partial v_x}{\partial x}\hat{\mathbf{i}}\hat{\mathbf{i}} + \frac{\partial v_y}{\partial y}\hat{\mathbf{j}}\hat{\mathbf{j}} + \frac{\partial v_y}{\partial x}\hat{\mathbf{i}}\hat{\mathbf{j}} + \frac{\partial v_x}{\partial y}\hat{\mathbf{j}}\hat{\mathbf{i}}. \tag{a}$$

The objects $\hat{\mathbf{i}}\hat{\mathbf{i}}, \hat{\mathbf{j}}\hat{\mathbf{j}}, \hat{\mathbf{i}}\hat{\mathbf{j}}, \hat{\mathbf{j}}\hat{\mathbf{i}}$ are neither scalars nor vectors; they are called **dyads**, and the linear combination of dyads (a) is called a **second-order tensor**. Tensors are important in such applications as continuum mechanics, differential geometry, and the theory of relativity but are not discussed here. We refer the interested reader to the 68-page book *Tensor Analysis* by H. D. Block (Columbus, Ohio: Charles E. Merrill, 1962).

by $x = x(s)$, $y = y(s)$, and $z = z(s)$, where s is the arc length along C from some reference point on C, and we wish to compute the rate of change du/ds along C. By chain differentiation, we have

$$\frac{d}{ds}u(x(s), y(s), z(s)) = \frac{\partial u}{\partial x}\frac{dx}{ds} + \frac{\partial u}{\partial y}\frac{dy}{ds} + \frac{\partial u}{\partial z}\frac{dz}{ds}, \tag{4}$$

which formula holds because we have assumed $u(x, y, z)$ to be C^1. That is, chain differentiation is essentially an interpolation formula, whereby du/ds is computed as a linear combination of the rates of change $\partial u/\partial x$, $\partial u/\partial y$, and $\partial u/\partial z$ in the orthogonal coordinate directions. For such interpolation to be valid, we surely need those three partial derivatives to be continuous at the point in question (Theorem 13.4.1), and that is why we have assumed that $u(x, y, z)$ is C^1. In fact, typically, the scalar fields that arise in applications are indeed C^1, perhaps with breakdowns at one or more isolated points.

Continuing, observe that the right-hand side of (4) is a dot product:

$$\frac{du}{ds} = \left(\frac{\partial u}{\partial x}\hat{\mathbf{i}} + \frac{\partial u}{\partial y}\hat{\mathbf{j}} + \frac{\partial u}{\partial z}\hat{\mathbf{k}}\right) \cdot \left(\frac{dx}{ds}\hat{\mathbf{i}} + \frac{dy}{ds}\hat{\mathbf{j}} + \frac{dz}{ds}\hat{\mathbf{k}}\right). \tag{5}$$

The first vector on the right is ∇u, and the second is $d\mathbf{R}/ds$, where $\mathbf{R}(s) = x(s)\hat{\mathbf{i}} + y(s)\hat{\mathbf{j}} + z(s)\hat{\mathbf{k}}$ is the position vector from the origin to the point $P = (x(s), y(s), z(s))$ on C. Not only is $d\mathbf{R}/ds$ a tangent vector to C at P, it is a *unit* vector because

$$\left\|\frac{d\mathbf{R}}{ds}\right\| = \left\|\lim_{\Delta\to 0}\frac{\Delta\mathbf{R}}{\Delta s}\right\| = \lim_{\Delta\to 0}\frac{\|\Delta\mathbf{R}\|}{\Delta s} = 1, \tag{6}$$

by the definition of arc length (Section 15.2.2). If we denote that unit tangent vector $d\mathbf{R}/ds$ as $\hat{\mathbf{s}}$, then (5) becomes

$$\boxed{\frac{du}{ds} = \nabla u \cdot \hat{\mathbf{s}}.} \tag{7}$$

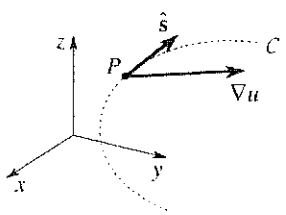


Figure 1. Computing du/ds at P.

If we wish, we can dispense with the curve C altogether. If we want the derivative du/ds, at any point P, in any desired direction defined by a unit vector $\hat{\mathbf{s}}$, then it is given by the gradient of u, at P, dotted into $\hat{\mathbf{s}}$ (Fig. 1). For instance, if $\hat{\mathbf{s}} = \hat{\mathbf{i}}$ then (7) gives $du/ds = (u_x\hat{\mathbf{i}} + u_y\hat{\mathbf{j}} + u_z\hat{\mathbf{k}}) \cdot \hat{\mathbf{i}} = \partial u/\partial x$. Similarly, if $\hat{\mathbf{s}} = \hat{\mathbf{j}}$ then $du/ds = \partial u/\partial y$, and if $\hat{\mathbf{s}} = \hat{\mathbf{k}}$ then $du/ds = \partial u/\partial z$. More generally, however, (7) gives the rate of change of u in *any* $\hat{\mathbf{s}}$ direction.

EXAMPLE 2. Compute the directional derivative of the field $u(x, y, z) = x^2 - 3yz$ in the direction of the vector $\hat{\mathbf{i}} + \hat{\mathbf{j}} - 2\hat{\mathbf{k}}$, at the point $(2, -1, 4)$. Using (7),

$$\frac{du}{ds}(2, -1, 4) = (2x\hat{\mathbf{i}} - 3z\hat{\mathbf{j}} - 3y\hat{\mathbf{k}})\Big|_{(2,-1,4)} \cdot \frac{\hat{\mathbf{i}} + \hat{\mathbf{j}} - 2\hat{\mathbf{k}}}{\sqrt{6}} = -\frac{14}{\sqrt{6}}.$$

COMMENT. Remember that the gradient is a vector, but the directional derivative is a scalar. ∎

Interpretation of ∇u. We were able to interpret the scalar div $\mathbf{v}(P)$ simply, as the outflow per unit volume at P. We do not have an analogous physical interpretation of the vector ∇u, but we can use (7) to draw geometrical conclusions as to its direction and magnitude.

Consider any point P in a region throughout which a C^1 scalar field u is defined. Suppose that $\nabla u \neq \mathbf{0}$ at P and that there is a $u =$ constant surface $\mathcal{S}$ through P and a tangent plane $\mathcal{T}$ (Fig. 2); for instance, if u is a temperature field, then $\mathcal{S}$ is an isothermal surface. If $\hat{\mathbf{s}}$, at P, is chosen as any vector in the tangent plane $\mathcal{T}$, then surely du/ds must be zero. Since $du/ds = \nabla u \cdot \hat{\mathbf{s}} = 0$ for *every* $\hat{\mathbf{s}}$ at P in the tangent plane, and both ∇u and $\hat{\mathbf{s}}$ are nonzero, it follows that ∇u is normal to the tangent plane $\mathcal{T}$ and hence to the surface $\mathcal{S}$, at P.

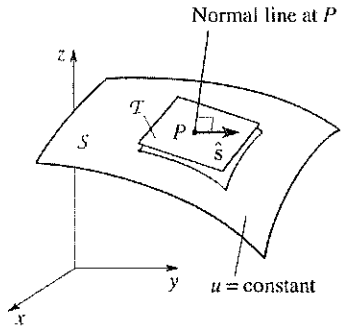


Figure 2. Geometrical interpretation of ∇u.

If, letting $\hat{\mathbf{s}}$ be in the tangent plane, we learn that ∇u is normal to $\mathcal{S}$, then to seek additional information about ∇u it seems logical to let $\hat{\mathbf{s}}$ be along the normal line at P, say in the direction of increasing u, for definiteness. Then, writing du/dn and $\hat{\mathbf{n}}$ for du/ds and $\hat{\mathbf{s}}$, respectively, (7) gives

$$\begin{aligned} \frac{du}{dn} &= \nabla u \cdot \hat{\mathbf{n}} \\ &= \|\nabla u\| \, (1) \cos 0 = \|\nabla u\| \,, \end{aligned} \tag{8}$$

so that the magnitude of ∇u is the directional derivative of u along the normal line to $\mathcal{S}$, in the direction of increasing u.

In summary, we can say this about the gradient ∇u of a scalar field $u(x, y, z)$ at a point P: *its direction is normal to the $u =$ constant surface through P, in the direction of increasing u, and its magnitude is equal to the directional derivative du/dn in that direction.*

EXAMPLE 3. Let our scalar field be a two-dimensional temperature field $T(x, y)$, defined on the rectangle $0 \leq x \leq 2.2, 0 \leq y \leq 1$. Suppose that T is not known analytically, but is measured experimentally, and that when the data are organized into level curves (isotherms in this case) the resulting graph is as shown in Fig. 3. The problem that we pose is to determine ∇T at $P = (1, 0.4)$. Drawing the lines PA and PB, we compute

$$\frac{\partial T}{\partial x}(P) \approx \frac{T(A) - T(P)}{\|\mathbf{PA}\|} \approx \frac{60 - 40}{0.51} \approx 39 \tag{9a}$$

and

$$\frac{\partial T}{\partial y}(P) \approx \frac{T(B) - T(P)}{\|\mathbf{PB}\|} \approx \frac{60 - 40}{0.2} \approx 100 \tag{9b}$$

so that

$$\nabla T(P) = \frac{\partial T}{\partial x}(P)\hat{\mathbf{i}} + \frac{\partial T}{\partial y}(P)\hat{\mathbf{j}} \approx 39\hat{\mathbf{i}} + 100\hat{\mathbf{j}}. \tag{10}$$

In (9a) we appealed to the definition

$$\frac{\partial T}{\partial x} = \lim_{\Delta x \to 0} \frac{T(x + \Delta x, y) - T(x, y)}{\Delta x} \tag{11}$$

but were unable to pass to the limit as $\Delta x \to 0$ because of the discrete spacing of the level curves provided in the plot. Thus, we took the smallest possible Δx, namely $\|\mathbf{PA}\|$, and "lived" with the approximation (9a). Similarly for (9b).

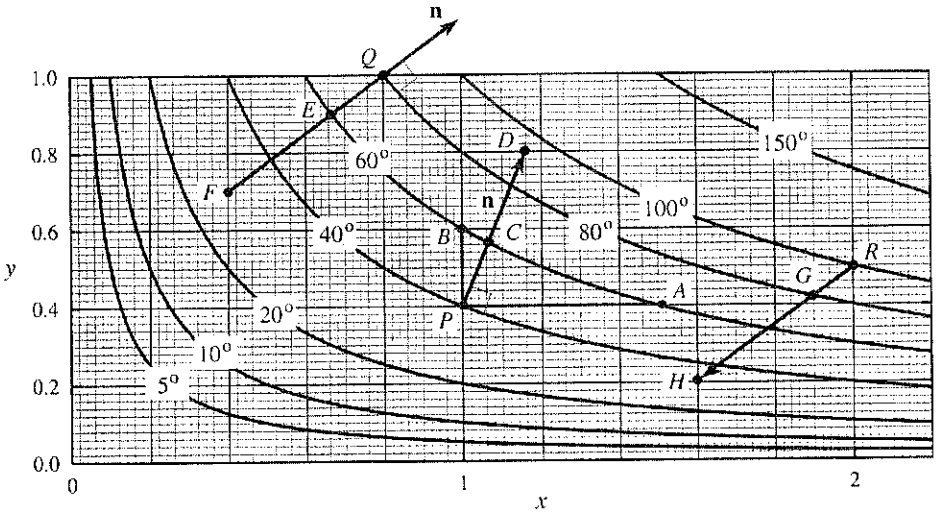


Figure 3. Level curves of $T(x, y)$.

In place of the Cartesian formula (10), we can compute ∇T from the intrinsic formula

$$\nabla T = \frac{dT}{dn}\hat{\mathbf{n}} \tag{12}$$

that is implied by the italicized summary preceding this example. Here, $\hat{\mathbf{n}}$ is the unit normal vector to the T = constant curve in the direction of increasing T, and dT/dn is the directional derivative of T in the direction of $\hat{\mathbf{n}}$. Drawing the normal to the $T = 40$ curve, from P to some other convenient point, say D, we have

$$\hat{\mathbf{n}} = \frac{\mathbf{PD}}{\|\mathbf{PD}\|} \approx \frac{0.16\hat{\mathbf{i}} + 0.4\hat{\mathbf{j}}}{\sqrt{(0.16)^2 + (0.4)^2}} \approx 0.37\hat{\mathbf{i}} + 0.93\hat{\mathbf{j}} \tag{13}$$

and

$$\frac{dT}{dn} \approx \frac{T(C) - T(P)}{\|\mathbf{PC}\|} \approx \frac{60 - 40}{0.18} \approx 111 \tag{14}$$

so (12) gives

$$\nabla T \approx (111)(0.37\hat{\mathbf{i}} + 0.93\hat{\mathbf{j}}) \approx 41\hat{\mathbf{i}} + 103\hat{\mathbf{j}}. \tag{15}$$

In fact, the temperature field that we have plotted in Fig. 3 is actually $T(x, y) = 100xy$. Thus, the exact value of ∇T at $(1, 0.4)$ is

$$\nabla T = \frac{\partial T}{\partial x}\hat{\mathbf{i}} + \frac{\partial T}{\partial y}\hat{\mathbf{j}} = 100y\hat{\mathbf{i}} + 100x\hat{\mathbf{j}} = 40\hat{\mathbf{i}} + 100\hat{\mathbf{j}}, \tag{16}$$

in view of which our estimates (10) and (15) are seen to be respectable, considering that our Δx, Δy, and Δn increments were not very small. More sophisticated difference quotients could have been used, to squeeze out more accuracy, but the thrust of this example is to illustrate concepts, not to maximize numerical accuracy.

As a second calculation, using Fig. 3, let us compute ∇T at $Q = (0.8, 1)$. Let us use (12). In this case we cannot use a forward difference quotient for dT/dn because Q is at a boundary point of the region and a forward step would carry us outside the region, where there is no data.* Thus, use a backward difference quotient instead. First, draw a normal line FQ. Then,

$$\hat{\mathbf{n}} = \frac{\mathbf{FQ}}{\|\mathbf{FQ}\|} \approx \frac{0.4\hat{\mathbf{i}} + 0.3\hat{\mathbf{j}}}{\sqrt{(0.4)^2 + (0.3)^2}} \approx 0.8\hat{\mathbf{i}} + 0.6\hat{\mathbf{j}}, \tag{17}$$

$$\frac{dT}{dn} \approx \frac{T(Q) - T(F)}{\|\mathbf{QF}\|} \approx \frac{80 - 60}{0.17} \approx 118 \tag{18}$$

so

$$\nabla T(0.8, 1) \approx (118)(0.8\hat{\mathbf{i}} + 0.6\hat{\mathbf{j}}) \approx 94\hat{\mathbf{i}} + 71\hat{\mathbf{j}} \tag{19}$$

compared with the exact value $100\hat{\mathbf{i}} + 80\hat{\mathbf{j}}$.

Finally, let us compute dT/ds at $R = (2, 0.5)$, in the direction $\mathbf{RH}$, where H is at $(1.6, 0.2)$. We have

$$\frac{dT}{ds}(R) \approx \frac{T(G) - T(R)}{\|\mathbf{GR}\|} \approx \frac{80 - 100}{0.13} \approx -154 \tag{20}$$

compared with the exact value -160. ∎

Closure. Having met the vector differential operator ∇, in Section 16.3, we define the gradient of a scalar field u as $\operatorname{grad} u = \nabla u$. In Cartesian coordinates,

$$\operatorname{grad} u = \frac{\partial u}{\partial x}\hat{\mathbf{i}} + \frac{\partial u}{\partial y}\hat{\mathbf{j}} + \frac{\partial u}{\partial z}\hat{\mathbf{k}}. \tag{21}$$

Next, we derived the formula

$$\frac{du}{ds} = \nabla u \cdot \hat{\mathbf{s}} \tag{22}$$

for the directional derivative du/ds in any given $\hat{\mathbf{s}}$ direction. Besides being of importance in its own right, (22) help us to understand the gradient ∇u. We learn

*In the difference quotient definition of the derivative,

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}, \tag{a}$$

Δx can tend to zero through positive values, in which case we call the difference quotient in (a) a **forward difference quotient**, or through negative values. In the latter case we can re-express (a) as

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x) - f(x - \Delta x)}{\Delta x}, \tag{b}$$

and call it a **backward difference quotient**.

that ∇u at any point P is normal to the u = constant surface through P, in the direction of increasing u, and its magnitude is equal to the directional derivative du/dn in that direction.

EXERCISES 16.4

1. Work out grad u for the following scalar fields, each of which is defined over all of 3-space. Further, evaluate grad u at $(9, 4, -1)$.

(a) $u = 6x$ (b) $u = x^2$
(c) $u = z\sin(x^2 + y^2)$ (d) $u = x^3$
(e) $u = xyz$ (f) $u = x^2 + y^2 + z^2$
(g) $u = x^2 + z^2$ (h) $u = x - y$
(i) $u = x + y + z$ (j) $u = z - x^2 - y^2$

2. Evaluate the directional derivative du/ds, at the designated point P, in the direction of the given vector $\mathbf{v}$.

(a) $u = x^2 + y^2 + z^2,\ P = (2, 1, 5),\ \mathbf{v} = \hat{\mathbf{i}}$
(b) $u = x + y + 3z,\ P = (1, 0, 3),\ \mathbf{v} = 2\hat{\mathbf{i}} - \hat{\mathbf{j}} + 5\hat{\mathbf{k}}$
(c) $u = xyz,\ P = (1, -1, 2),\ \mathbf{v} = 3\hat{\mathbf{i}} - \hat{\mathbf{k}}$
(d) $u = xyz,\ P = (3, 2, 1),\ \mathbf{v} = \hat{\mathbf{i}} + \hat{\mathbf{j}} + \hat{\mathbf{k}}$
(e) $u = \sqrt{x(y+z)},\ P = (1, 2, 3),\ \mathbf{v} = \hat{\mathbf{i}} - \hat{\mathbf{j}} + \hat{\mathbf{k}}$
(f) $u = 3x^2,\ P = (2, 6, -1),\ \mathbf{v} = 2\hat{\mathbf{j}} + 5\hat{\mathbf{k}}$

3. For the temperature field $T(x, y)$ given in Fig. 3, evaluate the following quantities directly from the figure, with the help of suitable difference quotients, and compare your results with the exact values obtained from the expression $T = 100xy$.

(a) ∇T at $(0.4, 1)$ (b) ∇T at $(1, 1)$
(c) ∇T at $(1, 0.2)$ (d) ∇T at $(0.5, 0.2)$
(e) ∇T at $(1.6, 0.5)$ (f) ∇T at $(1.5, 1)$
(g) dT/ds at $(0.4, 1)$ in the direction of $-\hat{\mathbf{i}} - \hat{\mathbf{j}}$
(h) dT/ds at $(1, 1)$ in the direction of $-\hat{\mathbf{i}} - 2\hat{\mathbf{j}}$
(i) dT/ds at $(1, 0.2)$ in the direction of $\hat{\mathbf{i}} + \hat{\mathbf{j}}$
(j) dT/ds at $(0.5, 0.2)$ in the direction of $-\hat{\mathbf{j}}$
(k) dT/ds at $(1.6, 0.5)$ in the direction of $\hat{\mathbf{i}} - \hat{\mathbf{j}}$
(l) dT/ds at $(1.5, 1)$ in the direction of $-\hat{\mathbf{i}} - 2\hat{\mathbf{j}}$

4. Let the electric potential (i.e., the voltage) be given by $V(x, y, z) = 3x^2y - xz$. If a positive charge is placed at $P = (x, y, z)$, in what direction will the charge begin to move? NOTE: It is known, from electric field theory, that such a charge will begin to move in the direction of maximum rate of voltage drop.

(a) $P = (2, 3, -1)$ (b) $P = (4, 0, -1)$
(c) $P = (1, 2, 5)$ (d) $P = (0, 0, 4)$
(e) $P = (0, 2, 1)$ (f) $P = (3, -2, 0)$

5. (*Convective derivative*) Let $\mathbf{v}(x, y, z, t) = v_x(x, y, z, t)\hat{\mathbf{i}} + v_y(x, y, z, t)\hat{\mathbf{j}} + v_z(x, y, z, t)\hat{\mathbf{k}}$ be a fluid velocity field, and consider some scalar property of the flow, such as the temperature field $T(x, y, z, t)$. If we swim along any desired path according to $x = x(t)$, $y = y(t)$, $z = z(t)$, then the timewise rate of change of T that we observe is, by chain differentiation,

$$\frac{d}{dt}T(x(t), y(t), z(t)) = \frac{\partial T}{\partial t} + \frac{\partial T}{\partial x}\frac{dx}{dt} + \frac{\partial T}{\partial y}\frac{dy}{dt} + \frac{\partial T}{\partial z}\frac{dz}{dt}. \tag{5.1}$$

(a) If we choose to *drift* with the fluid, then $dx/dt = v_x$, $dy/dt = v_y$, and $dz/dt = v_z$. In this case, show that (a) becomes

$$\frac{dT}{dt} = \frac{\partial T}{\partial t} + (\mathbf{v}\cdot\nabla)T. \tag{5.2}$$

This is often called the **convective derivative** because it is the derivative obtained when we drift, or convect, with the fluid. The special notation D/Dt, suggested by *Sir G. G. Stokes* (1819–1903), is often used:

$$\boxed{\frac{D(\ )}{Dt} = \frac{\partial(\ )}{\partial t} + (\mathbf{v}\cdot\nabla)(\ ).} \tag{5.3}$$

(b) To understand the nature of the contributions of each of the two terms on the right-hand side of (5.2) apply (5.2) to the three simple cases: $T = 2t$, $\mathbf{v} = 0$; $T = 3x$, $\mathbf{v} = U\hat{\mathbf{i}}$; $T = 2t + 3x$, $\mathbf{v} = U\hat{\mathbf{i}}$. Include whatever words of explanation (and sketches) seem appropriate.

(c) Does it matter if we rewrite $\mathbf{v}\cdot\nabla$ in equation (5.3) as $\nabla\cdot\mathbf{v}$? That is, is it true that

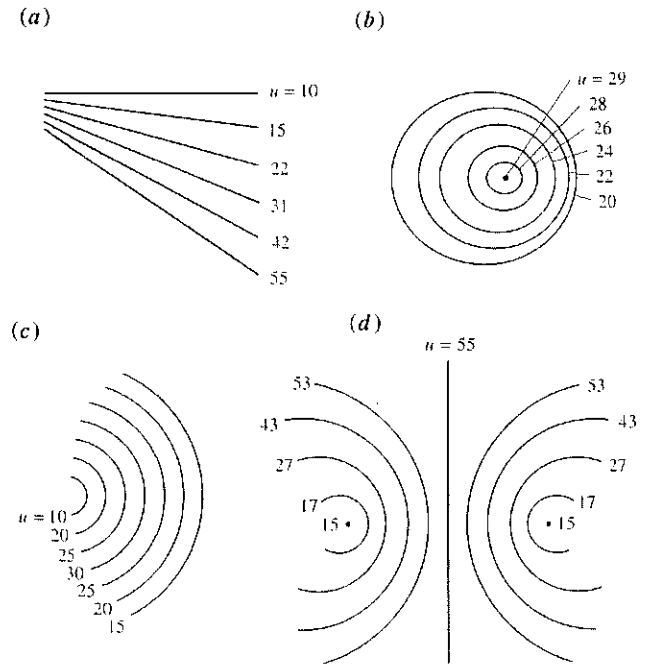
$$\frac{D(\ )}{Dt} = \frac{\partial(\ )}{\partial t} + (\operatorname{div}\mathbf{v})(\ )\,?$$

6. Evaluate the rate of change of temperature, dT/dt, measured by a thermometer that drifts along with the fluid velocity $\mathbf{v}$. Evaluate dT/dt at $x = 2$, $y = -1$, $z = 3$, and $t = 4$. HINT: Read Exercise 5.

(a) $T = 100t^2 + 2(x^2 + y^2)$, $\mathbf{v} = 20xt\hat{\mathbf{i}} - 10y\hat{\mathbf{j}}$
(b) $T = 25xt^2 - 2xyz$, $\mathbf{v} = 20\hat{\mathbf{i}} + xy\hat{\mathbf{j}} - t\hat{\mathbf{k}}$
(c) $T = (x + y + z)t$, $\mathbf{v} = xz\hat{\mathbf{i}} - xt\hat{\mathbf{j}}$
(d) $T = y^2 - zt$, $\mathbf{v} = \hat{\mathbf{i}} + 2x^2t\hat{\mathbf{j}} - \hat{\mathbf{k}}$

7. Define $u(x, y)$ to be 1 along the curve $y = x^2$ except for the origin $x = y = 0$, and define it to be zero at all other points in the plane, including the origin. Show that du/ds exists and equals 0 at the origin, for *all* directions $\hat{\mathbf{s}}$, but that $u(x, y)$ is, nevertheless, discontinuous at the origin. NOTE: Recall from the calculus that if $y'(x)$ exists at x_0, then $y(x)$ is necessarily continuous at x_0. For functions of more than one variable, we see, from this example, that the existence of directional derivatives in *every direction*, at a given point, does *not* suffice to imply that the function is continuous at that point.

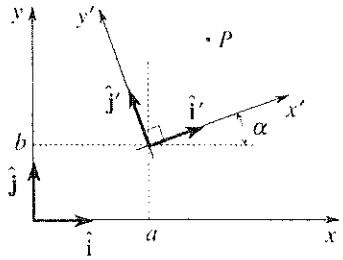
8. Some level curves of a scalar field $u(x, y)$ are sketched below. Trace or photocopy that sketch, and indicate the point(s), if any, at which $\nabla u \approx \mathbf{0}$, as well as the point(s) at which ∇u takes on its greatest magnitude.



9. (*Invariance property of* ∇u) In Section 16.3 we stress the invariance of the divergence of $\mathbf{v}$: that is, the value of $\nabla \cdot \mathbf{v}$ at a given field point is the same, whether we use one Cartesian coordinate system or another. The same invariance holds for the gradient. Prove that claim – in two dimensions, for simplicity. That is, show that

$$\frac{\partial u}{\partial x}\hat{\mathbf{i}} + \frac{\partial u}{\partial y}\hat{\mathbf{j}} = \frac{\partial u}{\partial x'}\hat{\mathbf{i}}' + \frac{\partial u}{\partial y'}\hat{\mathbf{j}}', \tag{9.1}$$

where the x, y and x', y' systems differ by an arbitrary displacement and an arbitrary rotation as shown in the figure.



10. (*Limit definition of gradient*) recall that we began with the limit definition of the divergence,

$$\operatorname{div} \mathbf{v}(P) \equiv \lim_{\mathcal{B}\to 0} \left\{ \frac{\int_S \hat{\mathbf{n}} \cdot \mathbf{v}\, dA}{V} \right\}. \tag{10.1}$$

Working out the right-hand side, for the case of Cartesian coordinates, gave

$$\operatorname{div} \mathbf{v} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z},$$

and expressing the latter as

$$\begin{aligned} \operatorname{div} \mathbf{v} &= \left(\hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z} \right) \cdot (v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}) \\ &= \nabla \cdot \mathbf{v} \end{aligned}$$

led us to the del operator,

$$\nabla = \hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}.$$

With the ∇ operator in hand, we then used ∇ to introduce the gradient, as $\operatorname{grad} u = \nabla u$. Alternatively, the gradient can be introduced intrinsically, by a limit definition analogous to the definition (10.1) of the divergence, namely,

$$\boxed{\operatorname{grad} u(P) \equiv \lim_{\mathcal{B}\to 0} \left\{ \frac{\int_S \hat{\mathbf{n}}\, u\, dA}{V} \right\}}. \tag{10.2}$$

Following steps analogous to those in Section 16.3, derive from (10.2) the Cartesian coordinate expression of $\operatorname{grad} u$ as

$$\operatorname{grad} u = \frac{\partial u}{\partial x}\hat{\mathbf{i}} + \frac{\partial u}{\partial y}\hat{\mathbf{j}} + \frac{\partial u}{\partial z}\hat{\mathbf{k}}$$

which, of course, is the same as ∇u.

16.5 Curl

Having considered the divergence and gradient, $\operatorname{div}\mathbf{v} = \nabla\cdot\mathbf{v}$ and $\operatorname{grad} u = \nabla u$, we now turn to the **curl of** $\mathbf{v}$, defined here as

$$\boxed{\operatorname{curl}\mathbf{v} \equiv \nabla\times\mathbf{v},} \tag{1}$$

where $\mathbf{v}$ is a C^1 vector field. Working out the right-hand side of (1) in Cartesian coordinates yields

$$\begin{aligned}\operatorname{curl}\mathbf{v} &= \left(\hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}\right)\times(v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}})\\ &= \begin{vmatrix}\hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}}\\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z}\\ v_x & v_y & v_z\end{vmatrix}\end{aligned} \tag{2}$$

or*

$$\boxed{\operatorname{curl}\mathbf{v} = \left(\frac{\partial v_z}{\partial y} - \frac{\partial v_y}{\partial z}\right)\hat{\mathbf{i}} - \left(\frac{\partial v_z}{\partial x} - \frac{\partial v_x}{\partial z}\right)\hat{\mathbf{j}} + \left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y}\right)\hat{\mathbf{k}}.} \tag{3}$$

EXAMPLE 1. If $\mathbf{v}(x,y,z,t) = xyz\hat{\mathbf{i}} - 2y^2t\hat{\mathbf{k}}$, then

$$\begin{aligned}\operatorname{curl}\mathbf{v} &= (-4yt - 0)\hat{\mathbf{i}} - (0 - xy)\hat{\mathbf{j}} + (0 - xz)\hat{\mathbf{k}}\\ &= -4yt\hat{\mathbf{i}} + xy\hat{\mathbf{j}} - xz\hat{\mathbf{k}}. \quad \blacksquare\end{aligned}$$

To interpret $\operatorname{curl}\mathbf{v}$ physically, let $\mathbf{v} = v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}$ be a fluid velocity field. Let us examine $\operatorname{curl}\mathbf{v}$ which, in the study of fluid mechanics, is called the **vorticity**

*Evidently, if we expand the determinant about the *first row*, we do obtain the terms on the right-hand side of (3). Yet if we expand about the third row, say, we obtain

$$v_x\left(\hat{\mathbf{j}}\frac{\partial}{\partial z} - \frac{\partial}{\partial y}\hat{\mathbf{k}}\right) - v_y\left(\hat{\mathbf{i}}\frac{\partial}{\partial z} - \frac{\partial}{\partial x}\hat{\mathbf{k}}\right) + v_z\left(\hat{\mathbf{i}}\frac{\partial}{\partial y} - \frac{\partial}{\partial x}\hat{\mathbf{j}}\right), \tag{a}$$

which is not the same. The point to bear in mind is that ∇ is located to the left of $\mathbf{v}$ in (1) so that $\partial/\partial x$, $\partial/\partial y$, and $\partial/\partial z$ act on v_x, v_y, and v_z. Thus, the terms in (a) need to be rearranged: the first term, $v_x\hat{\mathbf{j}}\,\partial/\partial z$, should be rearranged as $(\partial v_x/\partial z)\hat{\mathbf{j}}$, and so on, for each of the other terms. If we remember to do that, the correct expression is obtained no matter which row or column is expanded about.

vector. For simplicity, let $\mathbf{v}$ be a *plane flow*, where $v_z = 0$ and v_x, v_y do not vary with z. Thus, $\mathbf{v} = v_x(x, y)\hat{\mathbf{i}} + v_y(x, y)\hat{\mathbf{j}}$ and (3) reduces to

$$\operatorname{curl} \mathbf{v} = \left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y} \right) \hat{\mathbf{k}}. \tag{4}$$

Consider, heuristically, the motion of the little rectangular element of fluid shown in Fig. 1. At time t it is in position 1. If the element were *rigid*, its motion would consist of a translation plus a rotation. If so, then at time $t + \Delta t$ it might be in configuration 2, say, due to a translation OO' plus a rotation $\Delta\alpha$, and its angular velocity (taken as positive counterclockwise, with the right-hand rule used to assign a vector direction) would be $\boldsymbol{\omega} \sim (\Delta\alpha/\Delta t)\hat{\mathbf{k}}$. However, a fluid element is *deformable*, not rigid, so it may also suffer a shear deformation as indicated in the configuration 3. What, then, is the fluid's "angular velocity" to mean? It is *defined*, in fluid mechanics, as the average angular velocity of initially perpendicular edges such as OA and OB. Then

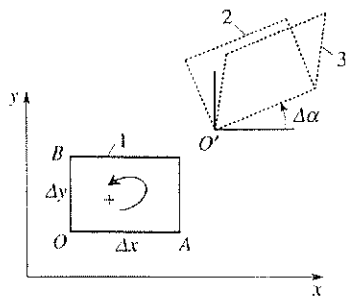


Figure 1. Plane motion of fluid element.

$$\text{angular velocity of } OA = \frac{v_y(A) - v_y(O)}{\Delta x}\hat{\mathbf{k}} \to \frac{\partial v_y}{\partial x}\hat{\mathbf{k}} \quad \text{as } \Delta x \to 0, \tag{5a}$$

$$\text{angular velocity of } OB = \frac{v_x(O) - v_x(B)}{\Delta y}\hat{\mathbf{k}} \to -\frac{\partial v_x}{\partial y}\hat{\mathbf{k}} \quad \text{as } \Delta y \to 0, \tag{5b}$$

so that the

$$\text{average of the two} \equiv \boldsymbol{\omega} = \frac{1}{2}\left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y} \right) \hat{\mathbf{k}}. \tag{6}$$

In case (5a) is not clear, consider Fig. 2. The average of the x velocity components $v_x(O)$ and $v_x(A)$ produce the x-wise translation of OA, their difference produces the stretching or contracting of OA, and the average of $v_y(O)$ and $v_y(A)$ produce the y-wise translation of OA. But the *angular velocity* of OA is due to the difference of $v(O)$ and $v(A)$. It is the vertical velocity of A relative to O, namely, $v_y(A) - v_y(O)$ divided by the radius OA of that motion. Similarly, the angular velocity of OB is $v_x(O) - v_x(B)$ divided by OB.

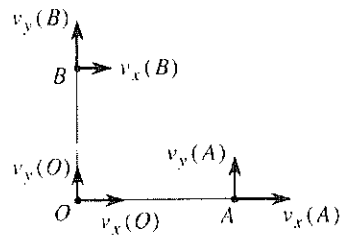


Figure 2. Angular velocities of OA and OB.

Comparing (4) and (6), we see that $\operatorname{curl} \mathbf{v} = 2\boldsymbol{\omega}$, from which we draw the following simple and important physical interpretation of the curl of a vector field: $\operatorname{curl} \mathbf{v}(P)$ *is twice the angular velocity of the fluid at* P.

Although our (heuristic) proof of the italicized claim was for plane flows, it is not hard to show that that result holds for nonplanar flows as well. And if our vector field is not a fluid velocity field (e.g., it might be an electric field, magnetic field, or gravitational force field) we can at least think of it as a velocity field in order to have access to the physical interpretation stated above in italics.

EXAMPLE 2. *Shear Flow.* To illustrate, consider the simple flow

$$\mathbf{v} = \kappa y \hat{\mathbf{i}}, \tag{7}$$

where κ is a positive constant. Shown in Fig. 3 the flow is known as a *shear flow* and is similar to the flow in the *boundary layer* on an aircraft wing. Considering a typical

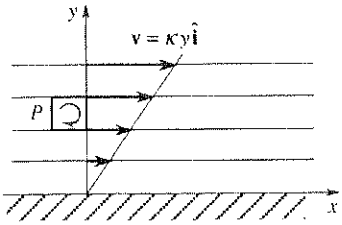


Figure 3. Shear flow.

fluid particle such as P, observe that the top of the particle is translating faster than the bottom so the particle undergoes a clockwise angular velocity as it translates to the right. Furthermore, it appears that all particles have the same angular velocity (though different translational velocities) since, for a given particle size, the velocity differential from top to bottom of the particle is the same for each particle. Thus, since curl $\mathbf{v}$ is twice the angular velocity of the fluid particle, and the latter is evidently a constant and in the $-\hat{\mathbf{k}}$ direction (by the right-hand rule), we expect curl $\mathbf{v}$ to be a constant and in the $-\hat{\mathbf{k}}$ direction. Let us see:

$$\operatorname{curl}\mathbf{v} = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ \kappa y & 0 & 0 \end{vmatrix} = -\kappa\hat{\mathbf{k}}, \tag{8}$$

which is a constant and in the $-\hat{\mathbf{k}}$ direction, as expected. ∎

It is interesting to observe that, in itself, the streamline pattern tells us nothing about the vorticity, curl $\mathbf{v}$. For instance, the three flows $\mathbf{v} = 3y\hat{\mathbf{i}}$, $-3y\hat{\mathbf{i}}$, and $3\hat{\mathbf{i}}$ all have the same streamline pattern: the set of horizontal lines, as in Fig. 3. But in these cases the fluid particles are rotating clockwise, counterclockwise, and not at all, respectively. To carry this point a bit further, consider the flow over a semicircular bump, shown in Fig. 4 of Section 16.2. It seems reasonable to imagine the fluid particles moving and rotating as sketched here in Fig. 4. Yet,

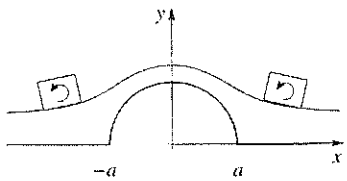


Figure 4. Particles rotating like this?

$$\mathbf{v} = U\hat{\mathbf{i}} + \frac{Ua^2}{(x^2+y^2)^2}\left[(y^2-x^2)\hat{\mathbf{i}} - 2xy\hat{\mathbf{j}}\right]$$

and, working out curl $\mathbf{v}$, we find that terms cancel so that curl $\mathbf{v} = \mathbf{0}$ everywhere in the field. Thus, the particles are not rotating at all even as they move along curved streamlines, in the same way that seats on a Ferris wheel remain horizontal even as they move in a vertical circle.

EXAMPLE 3. *Maxwell's Equations.* To illustrate the role of the curl in applications, we note that the equations

$$\nabla \times \mathbf{E} = 0, \tag{9a}$$

$$\nabla \times \mathbf{H} = \mathbf{J}, \tag{9b}$$

are two of the four classical **Maxwell's equations** governing steady (i.e., not varying with time) electromagnetic fields. $\mathbf{E}$, $\mathbf{H}$, and $\mathbf{J}$ are the electric field intensity, magnetic field intensity, and the current density, respectively. ∎

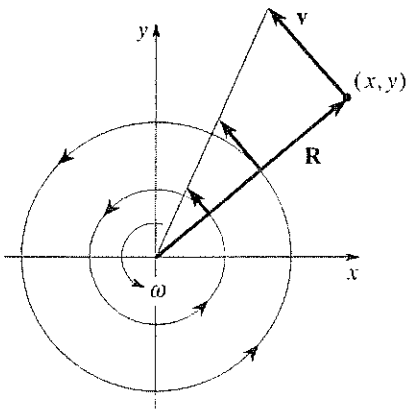
Closure. We define the curl of a vector field $\mathbf{v}$ as $\nabla \times \mathbf{v}$. In Cartesian coordinates the latter can be evaluated by (3), but it is easier to remember the determinant form given in (2). To attach a physical significance to $\nabla \times \mathbf{v}$, think of $\mathbf{v}$ as a fluid velocity field (whether it is or not). Then $\nabla \times \mathbf{v}$ at any given point P is twice the angular velocity of the fluid at P.

EXERCISES 16.5

1. Work out curl $\mathbf{v}$ for the following vector fields. Further, evaluate curl $\mathbf{v}$ at $(3, 4, -1)$.

(a) $\mathbf{v} = a\hat{\mathbf{i}} + b\hat{\mathbf{j}} + c\hat{\mathbf{k}}$ $(a, b, c$ constants$)$
(b) $\mathbf{v} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$
(c) $\mathbf{v} = x\hat{\mathbf{i}} - y^2\hat{\mathbf{j}} + z^3\hat{\mathbf{k}}$
(d) $\mathbf{v} = xyz(\hat{\mathbf{i}} + \hat{\mathbf{j}} - 3\hat{\mathbf{k}})$
(e) $\mathbf{v} = xy\hat{\mathbf{i}} - 2(x^2 + z^2)\hat{\mathbf{k}}$
(f) $\mathbf{v} = \hat{\mathbf{i}} + \hat{\mathbf{j}} + z\hat{\mathbf{k}}$
(g) $\mathbf{v} = z\hat{\mathbf{i}} + x^2\hat{\mathbf{j}}$
(h) $\mathbf{v} = \sin(x + y + z)\hat{\mathbf{j}}$
(i) $\mathbf{v} = f(x)\hat{\mathbf{i}} + g(y)\hat{\mathbf{j}} + h(z)\hat{\mathbf{k}}$
(j) $\mathbf{v} = f(y)\hat{\mathbf{i}} + g(x)\hat{\mathbf{j}}$

2. (*Solid-body rotation*) Consider a plane fluid flow that is a counterclockwise "solid-body rotation" about the z axis with angular velocity $\boldsymbol{\omega} = \omega\hat{\mathbf{k}}$, as sketched in the figure. That is,



the fluid might just as well be frozen solid, and spun about the z axis. Introducing polar coordinates r, θ, the position vector is

$$\mathbf{R} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} = r\cos\theta\hat{\mathbf{i}} + r\sin\theta\hat{\mathbf{j}}. \tag{2.1}$$

Since $\dot{r} = 0$ and $\dot{\theta} = \omega$ (where dots denote time derivatives), we have for the fluid velocity field

$$\begin{aligned}\mathbf{v} = \dot{\mathbf{R}} &= r(-\dot{\theta}\sin\theta\hat{\mathbf{i}} + \dot{\theta}\cos\theta\hat{\mathbf{j}}) \\ &= \omega(-r\sin\theta\hat{\mathbf{i}} + r\cos\theta\hat{\mathbf{j}}) = \omega(-y\hat{\mathbf{i}} + x\hat{\mathbf{j}}).\end{aligned} \tag{2.2}$$

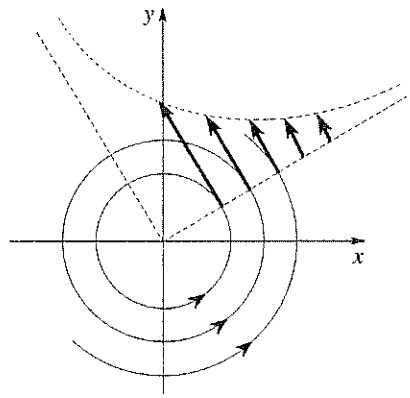
(a) Work out curl $\mathbf{v}$ for the $\mathbf{v}$ field given by (2.2) and explain, in physical terms, why the result makes sense.

(b) Derive (2.2) differently, by differentiating $\mathbf{R} = r\hat{\mathbf{e}}_r$ with respect to time.

3. (*Vortex flow*) The fluid velocity field

$$\mathbf{v} = \frac{\Gamma}{2\pi r}\hat{\mathbf{e}}_\theta \qquad (\Gamma = \text{constant}) \tag{3.1}$$

is purely tangential, i.e., there is no radial $(\hat{\mathbf{e}}_r)$ component so that the streamlines are concentric circles centered at the origin, as sketched in the accompanying figure. Furthermore,



the magnitude of the tangential velocity tends to zero as $r \to \infty$, and to ∞ as $r \to 0$. [Equation (3.1) is said to give the flow induced by a *vortex* of strength Γ located at the origin.] The problem that we pose is to re-express (3.1) in terms of $x, y, \hat{\mathbf{i}}, \hat{\mathbf{j}}$ and show that $\nabla \times \mathbf{v} = \mathbf{0}$ everywhere in the field (except at the origin, where $\nabla \times \mathbf{v}$ is not defined) so that every fluid particle (not at the origin) has no spin. [This result may seem strange in view of the fact that the particles are in circular motion! Evidently, the particles must maintain their spatial orientation (much as the seats on a Ferris wheel do) as they carry out their circular motion.]

4. (*Invariance property of curl* $\mathbf{v}$) Let a C^1 vector field $\mathbf{v}$ be represented in terms of some particular x, y frame as $\mathbf{v} = v_x(x, y)\hat{\mathbf{i}} + v_y(x, y)\hat{\mathbf{j}}$. (We limit ourselves to the two-dimensional case merely to reduce the algebra.) Then, according to (3),

$$\text{curl}\,\mathbf{v} = \left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y}\right)\hat{\mathbf{k}}. \tag{4.1}$$

Show that this vector is the same even if a different coordinate system is used, say, the x', y' system shown in the figure. That

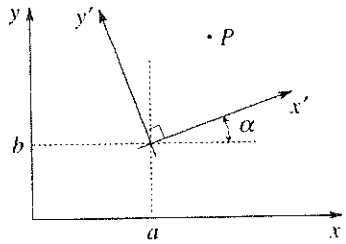
is, show that

$$\left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y}\right)\hat{\mathbf{k}} = \left(\frac{\partial v_{y'}}{\partial x'} - \frac{\partial v_{x'}}{\partial y'}\right)\hat{\mathbf{k}}'$$

or, since $\hat{\mathbf{k}}' = \hat{\mathbf{k}}$, that

$$\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y} = \frac{\partial v_{y'}}{\partial x'} - \frac{\partial v_{x'}}{\partial y'} \tag{4.2}$$

for any values of a, b and α.



5. Show whether or not $\nabla \times \mathbf{v}$ is necessarily orthogonal to $\mathbf{v}$. HINT: That is, is $\mathbf{v} \cdot \nabla \times \mathbf{v}$ necessarily zero? Write

$$\mathbf{v} \cdot \nabla \times \mathbf{v} = \begin{vmatrix} v_x & v_y & v_z \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ v_x & v_y & v_z \end{vmatrix},$$

and see if the determinant is necessarily zero. It is true that two rows are identical, which seems to imply that the determinant must be zero, but be careful, and recall the footnote associated with equation (3).

6. (*Limit definition of curl*) First, read Exercise 10 in Section 16.4. Similarly, we can introduce the curl by the limit definition

$$\boxed{\operatorname{curl} \mathbf{v}(P) \equiv \lim_{\mathcal{B}\to 0}\left\{\frac{\int_S \hat{\mathbf{n}} \times \mathbf{v}\, dA}{V}\right\},} \tag{6.1}$$

and deduce from (6.1) that curl $\mathbf{v}$ is indeed $\nabla \times \mathbf{v}$. That is what we ask you to do: using Cartesian coordinates, derive from (6.1) the result

$$\begin{aligned}\operatorname{curl} \mathbf{v} = &\left(\frac{\partial v_z}{\partial y} - \frac{\partial v_y}{\partial z}\right)\hat{\mathbf{i}} - \left(\frac{\partial v_z}{\partial x} - \frac{\partial v_x}{\partial z}\right)\hat{\mathbf{j}} \\ &+ \left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y}\right)\hat{\mathbf{k}},\end{aligned} \tag{6.2}$$

which, as claimed above, is the same as $\nabla \times \mathbf{v}$. NOTE: Let us write the three results together, so the pattern among them can be seen:

$$\operatorname{div} \mathbf{v} = \lim_{\mathcal{B}\to 0}\left\{\frac{\int_S \hat{\mathbf{n}} \cdot \mathbf{v}\, dA}{V}\right\} = \nabla \cdot \mathbf{v}, \tag{6.3a}$$

$$\operatorname{grad} u = \lim_{\mathcal{B}\to 0}\left\{\frac{\int_S \hat{\mathbf{n}}\, u\, dA}{V}\right\} = \nabla u, \tag{6.3b}$$

and

$$\operatorname{curl} \mathbf{v} = \lim_{\mathcal{B}\to 0}\left\{\frac{\int_S \hat{\mathbf{n}} \times \mathbf{v}\, dA}{V}\right\}. \tag{6.3c}$$

16.6 Combinations; Laplacian

We have introduced the div, grad, and curl operators, and found that all three can be expressed in terms of ∇. Specifically, div $= \nabla\cdot$, grad $= \nabla$, and curl $= \nabla\times$.

In this section, we consider combinations: one of the operators above acting on a combination of fields such as $\nabla(\alpha u + \beta \tilde{v})$ and $\nabla \cdot (u\mathbf{v})$, or a combination of the operators above acting on a single field, such as $\nabla \cdot \nabla u$ and $\nabla \times (\nabla u)$. Beginning with the former, the action of one such operator on a combination of fields, we note the following results:

If α, β are scalars, then

$$\nabla \cdot (\alpha \mathbf{u} + \beta \mathbf{v}) = \alpha \nabla \cdot \mathbf{u} + \beta \nabla \cdot \mathbf{v}, \tag{1}$$

$$\nabla(\alpha u + \beta v) = \alpha \nabla u + \beta \nabla v, \tag{2}$$
$$\nabla \times (\alpha \mathbf{u} + \beta \mathbf{v}) = \alpha \nabla \times \mathbf{u} + \beta \nabla \times \mathbf{v} \tag{3}$$

so that div, grad, and curl are *linear* operators. Furthermore,

$$\nabla \cdot (u\mathbf{v}) = \nabla u \cdot \mathbf{v} + u \nabla \cdot \mathbf{v}, \tag{4}$$
$$\nabla \times (u\mathbf{v}) = \nabla u \times \mathbf{v} + u \nabla \times \mathbf{v}, \tag{5}$$
$$\nabla \cdot (\mathbf{u} \times \mathbf{v}) = \mathbf{v} \cdot \nabla \times \mathbf{u} - \mathbf{u} \cdot \nabla \times \mathbf{v}, \tag{6}$$
$$\nabla \times (\mathbf{u} \times \mathbf{v}) = \mathbf{u} \nabla \cdot \mathbf{v} - \mathbf{v} \nabla \cdot \mathbf{u} + (\mathbf{v} \cdot \nabla)\mathbf{u} - (\mathbf{u} \cdot \nabla)\mathbf{v}, \tag{7}$$
$$\nabla(\mathbf{u} \cdot \mathbf{v}) = (\mathbf{u} \cdot \nabla)\mathbf{v} + (\mathbf{v} \cdot \nabla)\mathbf{u} + \mathbf{u} \times (\nabla \times \mathbf{v}) + \mathbf{v} \times (\nabla \times \mathbf{u}), \tag{8}$$

where $(\mathbf{u} \cdot \nabla)\mathbf{v}$ in (7) and (8) means

$$\begin{aligned}(\mathbf{u} \cdot \nabla)\mathbf{v} &= (\mathbf{u} \cdot \nabla)(v_x \hat{\mathbf{i}} + v_y \hat{\mathbf{j}} + v_z \hat{\mathbf{k}}) \\ &= (\mathbf{u} \cdot \nabla v_x)\hat{\mathbf{i}} + (\mathbf{u} \cdot \nabla v_y)\hat{\mathbf{j}} + (\mathbf{u} \cdot \nabla v_z)\hat{\mathbf{k}} \\ &= \left(u_x \frac{\partial v_x}{\partial x} + u_y \frac{\partial v_x}{\partial y} + u_z \frac{\partial v_x}{\partial z}\right)\hat{\mathbf{i}} + (\text{etc.})\hat{\mathbf{j}} + (\text{etc.})\hat{\mathbf{k}},\end{aligned}$$

and similarly for $(\mathbf{v} \cdot \nabla)\mathbf{u}$.

Derivation of these results is simple, and we will limit our discussion to the verification of one representative case, (4):

$$\begin{aligned}\nabla \cdot (u\mathbf{v}) &= \left(\hat{\mathbf{i}} \frac{\partial}{\partial x} + \hat{\mathbf{j}} \frac{\partial}{\partial y} + \hat{\mathbf{k}} \frac{\partial}{\partial z}\right) \cdot (uv_x \hat{\mathbf{i}} + uv_y \hat{\mathbf{j}} + uv_z \hat{\mathbf{k}}) \\ &= \frac{\partial}{\partial x}(uv_x) + \frac{\partial}{\partial y}(uv_y) + \frac{\partial}{\partial z}(uv_z) \\ &= \frac{\partial u}{\partial x} v_x + \frac{\partial u}{\partial y} v_y + \frac{\partial u}{\partial z} v_z + u\left(\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}\right) \\ &= \nabla u \cdot \mathbf{v} + u \nabla \cdot \mathbf{v}.\end{aligned}$$

The form of this result is not surprising and is reminiscent of the (scalar) statement $(d/dx)(uv) = (du/dx)v + u(dv/dx)$ from the calculus. That is, ∇ is a differential operator, and $u\mathbf{v}$ is a product so we get two terms: one in which ∇ acts on u, and one in which it acts on $\mathbf{v}$.

Next, look at the action of a combination of div, grad, and curl operators on a single field. One such case that will be of great importance to us is the operator $\nabla \cdot \nabla$; that is, div grad. This operator is called the **Laplacian** operator,* or just the Laplacian, and is denoted by the symbol ∇^2. That is,

$$\nabla^2 \equiv \nabla \cdot \nabla = \left(\hat{\mathbf{i}} \frac{\partial}{\partial x} + \hat{\mathbf{j}} \frac{\partial}{\partial y} + \hat{\mathbf{k}} \frac{\partial}{\partial z}\right) \cdot \left(\hat{\mathbf{i}} \frac{\partial}{\partial x} + \hat{\mathbf{j}} \frac{\partial}{\partial y} + \hat{\mathbf{k}} \frac{\partial}{\partial z}\right)$$

*After *Pierre Simon de Laplace* (1749–1827). Some authors use the symbol Δ in place of ∇^2.

or,

$$\boxed{\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}.} \tag{9}$$

Read as *del square*, ∇^2 does *not* mean ∇ times ∇. It is a mathematical symbol that means ∇ *dot* ∇, in the same way that the matrix symbol $\mathbf{A}^{-1}$ means the inverse matrix of $\mathbf{A}$, not $1/\mathbf{A}$. The Laplacian can act on scalar fields *or* on vector fields:

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}, \tag{10}$$

$$\begin{aligned}\nabla^2 \mathbf{v} &= \nabla^2(v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}) \\ &= \hat{\mathbf{i}}\nabla^2 v_x + \hat{\mathbf{j}}\nabla^2 v_y + \hat{\mathbf{k}}\nabla^2 v_z \\ &= \left(\frac{\partial^2 v_x}{\partial x^2} + \frac{\partial^2 v_x}{\partial y^2} + \frac{\partial^2 v_x}{\partial z^2}\right)\hat{\mathbf{i}} + (\text{etc.})\hat{\mathbf{j}} + (\text{etc.})\hat{\mathbf{k}}.\end{aligned} \tag{11}$$

Besides the div grad combination, just mentioned, some others to be noted are

$$\boxed{\begin{aligned}\text{div curl}\,\mathbf{v} &= \nabla\cdot\nabla\times\mathbf{v} = 0, \\ \text{curl grad}\,u &= \nabla\times\nabla u = \mathbf{0},\end{aligned}} \tag{12,13}$$

$$\text{curl curl}\,\mathbf{v} = \text{curl}^2\mathbf{v} = \nabla\times(\nabla\times\mathbf{v}) = \nabla(\nabla\cdot\mathbf{v}) - \nabla^2\mathbf{v}, \tag{14}$$

which hold if $\mathbf{v}$ and u are C^2.

We have drawn attention to (12) and (13), by placing them in a box because they will be important to us later, and also because they are striking results, striking in the sense that the divergence of the curl of *every* C^2 vector field is found to be zero, and the curl of the gradient of *every* C^2 scalar field is found to be zero.

As a mnemonic device to remember (12) and (13), think of the letters d, g, and c (for divergence, gradient, and curl). The only combinations that come readily to mind (for the author, at least) are **dc** (for direct current), and **cg** (for center of gravity). This is, the divergence of a curl is zero, and the curl of a gradient is zero.

EXAMPLE 1. If $u = x^2y$ and $\mathbf{v} = x^2y\hat{\mathbf{i}} - z^3\hat{\mathbf{j}} + 2\hat{\mathbf{k}}$, then

$$\begin{aligned}\nabla^2 u &= 2y \qquad \text{by(10)}, \\ \nabla^2\mathbf{v} &= (2y+0+0)\hat{\mathbf{i}} + (0+0-6z)\hat{\mathbf{j}} + (0+0+0)\hat{\mathbf{k}} \qquad \text{by(11)} \\ &= 2y\hat{\mathbf{i}} - 6z\hat{\mathbf{j}}, \\ \nabla\cdot\nabla\times\mathbf{v} &= 0 \qquad \text{by(12)}, \\ \nabla\times\nabla u &= \mathbf{0} \qquad \text{by(13)}, \\ \nabla\times(\nabla\times\mathbf{v}) &= \nabla(2xy) - (2y\hat{\mathbf{i}} - 6z\hat{\mathbf{j}}) \qquad \text{by(14)}, \\ &= (2x+6z)\hat{\mathbf{j}}. \quad \blacksquare\end{aligned}$$

Closure. The identities (1)–(8) and (12)–(14) are of general utility in vector analysis somewhat as the various trigonometric identities are of utility in the calculus, and we will use most of them in the remainder of this chapter. The Laplace operator ∇^2 will be central in our study of partial differential equations in Chapters 18–20.

EXERCISES 16.6

1. Evaluate $\nabla^2 u$ and $\nabla \times \nabla u$ in each case.

(a) $u = x$
(b) $u = x^2 + y^2$
(c) $u = x^2 + y^2 + z^2$
(d) $u = x^2y^3 - 6$
(e) $u = xe^y$
(f) $u = \cos(x - 2y)$
(g) $u = xyz$
(h) $u = \sin(xyz)$
(i) $u = ax + by + cz + d$ $(a, b, c, d$ constants$)$

2. Evaluate $\nabla \cdot \nabla \times \mathbf{v}$ and $\nabla \times (\nabla \times \mathbf{v})$ in each case.

(a) $\mathbf{v} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$
(b) $\mathbf{v} = xe^y\hat{\mathbf{i}} - 2\hat{\mathbf{j}} + z^2\hat{\mathbf{k}}$
(c) $\mathbf{v} = z^3\hat{\mathbf{i}}$
(d) $\mathbf{v} = yz\hat{\mathbf{i}} - x^2\hat{\mathbf{j}}$
(e) $\mathbf{v} = x^2y\hat{\mathbf{j}}$
(f) $\mathbf{v} = f(y)\hat{\mathbf{i}} + g(x)\hat{\mathbf{j}}$

3. Are the parentheses needed in (14), i.e., in $\nabla \times (\nabla \times \mathbf{v})$? Explain.

4. $\nabla^2 u = 0$ is called **Laplace's equation**. Show that the following are solutions of Laplace's equation.

(a) $xy + 3z$
(b) $x^2 - y^2 - 2xz$
(c) $x^3 - 3xy^2$
(d) $y^3 - 3x^2y$
(e) $\ln(x^2 + y^2)$ for $x^2 + y^2 \neq 0$
(f) $\dfrac{1}{x^2 + y^2 + z^2}$ for $x^2 + y^2 + z^2 \neq 0$

5. If $\mathbf{v}$ is any vector, and $\mathbf{R} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$, show that $(\mathbf{v} \cdot \nabla)\mathbf{R} = \mathbf{v}$.

6. Derive the following equations [as we have done in the text for (4)].

(a) equation (5)
(b) equation (6)
(c) equation (7)
(d) equation (12)
(e) equation (13)
(f) equation (14)

7. The following scalar equations occur in fluid mechanics, solid mechanics, and electromagnetics. Re-express them more concisely in terms of vector and vector differential operator notation. Here x, y, z subscripts denote the respective components of a given vector, *not* partial derivatives. To illustrate what we are after, consider the equations

$$\frac{\partial u}{\partial x} = F_x, \qquad \frac{\partial u}{\partial y} = F_y, \qquad \frac{\partial u}{\partial z} = F_z \tag{7.1}$$

If we multiply these equations by $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$, respectively, and add them, we obtain

$$\frac{\partial u}{\partial x}\hat{\mathbf{i}} + \frac{\partial u}{\partial y}\hat{\mathbf{j}} + \frac{\partial u}{\partial z}\hat{\mathbf{k}} = F_x\hat{\mathbf{i}} + F_y\hat{\mathbf{j}} + F_z\hat{\mathbf{k}}$$

or, in more concise vector notation, $\nabla u = \mathbf{F}$.

(a) $$\frac{\partial \sigma}{\partial t} + \frac{\partial}{\partial x}(\sigma v_x) + \frac{\partial}{\partial y}(\sigma v_y) + \frac{\partial}{\partial z}(\sigma v_z) = 0$$

(b) $$\sigma\frac{\partial^2 u_x}{\partial t^2} = \mu\left(\frac{\partial^2 u_x}{\partial x^2} + \frac{\partial^2 u_x}{\partial y^2} + \frac{\partial^2 u_x}{\partial z^2}\right) + (\lambda + \mu)\frac{\partial \Theta}{\partial x} + F_x$$

$$\sigma\frac{\partial^2 u_y}{\partial t^2} = \mu\left(\frac{\partial^2 u_y}{\partial x^2} + \frac{\partial^2 u_y}{\partial y^2} + \frac{\partial^2 u_y}{\partial z^2}\right) + (\lambda + \mu)\frac{\partial \Theta}{\partial y} + F_y$$

$$\sigma\frac{\partial^2 u_z}{\partial t^2} = \mu\left(\frac{\partial^2 u_z}{\partial x^2} + \frac{\partial^2 u_z}{\partial y^2} + \frac{\partial^2 u_z}{\partial z^2}\right) + (\lambda + \mu)\frac{\partial \Theta}{\partial z} + F_z,$$

where c, μ, λ are constants

(c) $$\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} + \frac{\partial B_x}{\partial t} = 0$$

$$\frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} + \frac{\partial B_y}{\partial t} = 0$$

$$\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} + \frac{\partial B_z}{\partial t} = 0$$

(d) $\dfrac{\partial H_z}{\partial y} - \dfrac{\partial H_y}{\partial z} = J_x$

$\dfrac{\partial H_x}{\partial z} - \dfrac{\partial H_z}{\partial x} = J_y$

$\dfrac{\partial H_y}{\partial x} - \dfrac{\partial H_x}{\partial y} = J_z$

8. The differential operator $\nabla^4 \equiv \nabla^2\nabla^2$ is known as the **biharmonic** (or **bi-Laplacian**) operator, and the partial differential equation $\nabla^4 u = 0$ is known as the **biharmonic equation**.

(a) If $u = u(x, y)$, show that

$$\nabla^4 u = u_{xxxx} + 2u_{xxyy} + u_{yyyy},$$

where the subscripts denote partial derivatives.
(b) Write out $\nabla^4 u$, in Cartesian coordinates, for the three-dimensional case $u = u(x, y, z)$.

9. (*Caution*) Familiar properties such as $\mathbf{A}\cdot\mathbf{B} = \mathbf{B}\cdot\mathbf{A}$, $\mathbf{B}\times\mathbf{A} = -\mathbf{A}\times\mathbf{B}$, and $\mathbf{A}\cdot(\mathbf{B}\times\mathbf{C}) = (\mathbf{A}\times\mathbf{B})\cdot\mathbf{C}$ are, in general, *not* true if one or more of the vectors is the vector differential operator $\boldsymbol{\nabla}$. Specifically, show that

$$\boldsymbol{\nabla}\cdot\mathbf{v} \neq \mathbf{v}\cdot\boldsymbol{\nabla}, \qquad \boldsymbol{\nabla}\times\mathbf{v} \neq -\mathbf{v}\times\boldsymbol{\nabla}, \qquad \text{and}$$

$$\boldsymbol{\nabla}\cdot(\mathbf{v}\times\mathbf{w}) \neq (\boldsymbol{\nabla}\times\mathbf{v})\cdot\mathbf{w}.$$

10. Given that

$$\boldsymbol{\nabla}\times\mathbf{v} = -\frac{1}{c}\frac{\partial\mathbf{w}}{\partial t}, \tag{10.1}$$

$$\boldsymbol{\nabla}\times\mathbf{w} = \frac{1}{c}\frac{\partial\mathbf{v}}{\partial t}, \tag{10.2}$$

$$\boldsymbol{\nabla}\cdot\mathbf{v} = 0, \tag{10.3}$$

$$\boldsymbol{\nabla}\cdot\mathbf{w} = 0, \tag{10.4}$$

show that $\mathbf{v}$ and $\mathbf{w}$ satisfy the *wave equations*

$$c^2\nabla^2\mathbf{v} = \frac{\partial^2\mathbf{v}}{\partial t^2} \qquad \text{and} \qquad c^2\nabla^2\mathbf{w} = \frac{\partial^2\mathbf{w}}{\partial t^2}.$$

HINT: One of equations (1) to (14) will be useful. Also, note that $\dfrac{\partial}{\partial t}\boldsymbol{\nabla}\cdot\mathbf{u} = \boldsymbol{\nabla}\cdot\dfrac{\partial\mathbf{u}}{\partial t}$ and $\dfrac{\partial}{\partial t}\boldsymbol{\nabla}\times\mathbf{u} = \boldsymbol{\nabla}\times\dfrac{\partial\mathbf{u}}{\partial t}$.

11. Evaluate $\nabla^2\mathbf{v}$ if

(a) $\mathbf{v} = xz^2\hat{\mathbf{i}} + y\sin z\hat{\mathbf{k}}$
(b) $\mathbf{v} = x^2\hat{\mathbf{i}} + y^2\hat{\mathbf{j}} + z^2\hat{\mathbf{k}}$
(c) $\mathbf{v} = x\hat{\mathbf{i}} - 2y\hat{\mathbf{j}} + 3z\hat{\mathbf{k}}$
(d) $\mathbf{v} = \sin(xz)\hat{\mathbf{j}}$

16.7 Non-Cartesian Systems; Div, Grad, Curl, and Laplacian (Optional)

Recall the pattern of the last several sections. Our starting point was the definition

$$\operatorname{div}\mathbf{v} \equiv \lim_{\mathcal{B}\to 0}\left\{\frac{\int_S \hat{\mathbf{n}}\cdot\mathbf{v}\,dA}{V}\right\}. \tag{1}$$

To obtain a more tractable form, we introduced a reference Cartesian coordinate system and, accordingly, specified $\mathcal{B}$ to be a prism, of dimension Δx by Δy by Δz, the faces of $\mathcal{B}$ being made up of constant-coordinate surfaces: $x =$ constant, $y =$ constant, $z =$ constant. Simplification of the right-hand side of (1) then led to the form

$$\operatorname{div} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z}, \tag{2}$$

which we then expressed as

$$\left(\hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}\right)\cdot(v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}),$$

or $\nabla \cdot \mathbf{v}$, where

$$\nabla = \hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}. \tag{3}$$

With ∇ in hand we then introduced the gradient and curl as $\operatorname{grad} u \equiv \nabla u$ and $\operatorname{curl} \mathbf{v} \equiv \nabla \times \mathbf{v}$.

But Cartesian coordinates are not necessarily the most convenient ones in a given application. For example, in studying the temperature field in a spherical domain, one would certainly be better advised to use spherical coordinates. We begin by considering the cases of plane polar, cylindrical, and spherical coordinates. Or, since the plane polar system is really a special case of the cylindrical system, it will suffice to consider cylindrical and spherical coordinates. In the final subsection, 16.7.3, we give the extension to *any* orthogonal curvilinear coordinate system.

16.7.1. Cylindrical coordinates. Our plan is to derive the ∇ operator, for cylindrical coordinates, from its Cartesian expression (3) by expanding

$$\begin{aligned} \hat{\mathbf{i}} &= \cos\theta\hat{\mathbf{e}}_r - \sin\theta\hat{\mathbf{e}}_\theta, \\ \hat{\mathbf{j}} &= \sin\theta\hat{\mathbf{e}}_r + \cos\theta\hat{\mathbf{e}}_\theta, \\ \hat{\mathbf{k}} &= \hat{\mathbf{e}}_z, \end{aligned} \tag{4}$$

recalling the relations $x = r\cos\theta$, $y = r\sin\theta$, $z = z$, and using chain differentiation. Thus (with $c \equiv \cos\theta$ and $s \equiv \sin\theta$, for brevity),

$$\begin{aligned} \nabla &= \hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z} \\ &= (c\hat{\mathbf{e}}_r - s\hat{\mathbf{e}}_\theta)\left(\frac{\partial}{\partial r}\frac{\partial r}{\partial x} + \frac{\partial}{\partial \theta}\frac{\partial \theta}{\partial x} + \frac{\partial}{\partial z}\frac{\partial z}{\partial x}\right) \\ &\quad + (s\hat{\mathbf{e}}_r + c\hat{\mathbf{e}}_\theta)\left(\frac{\partial}{\partial r}\frac{\partial r}{\partial y} + \frac{\partial}{\partial \theta}\frac{\partial \theta}{\partial y} + \frac{\partial}{\partial z}\frac{\partial z}{\partial y}\right) \\ &\quad + (\hat{\mathbf{e}}_z)\left(\frac{\partial}{\partial r}\frac{\partial r}{\partial z} + \frac{\partial}{\partial \theta}\frac{\partial \theta}{\partial z} + \frac{\partial}{\partial z}\frac{\partial z}{\partial z}\right). \end{aligned} \tag{5}$$

Recalling from (41) and (42) in Section 13.6 that $r_x = \cos\theta$, $r_y = \sin\theta$, $\theta_x = -\sin\theta/r$, $\theta_y = \cos\theta/r$, noting that $z_x = z_y = 0$ and $z_z = 1$, and using the identity $\cos^2\theta + \sin^2\theta = 1$, (5) reduces to

$$\boxed{\nabla = \hat{\mathbf{e}}_r\frac{\partial}{\partial r} + \hat{\mathbf{e}}_\theta\frac{1}{r}\frac{\partial}{\partial \theta} + \hat{\mathbf{e}}_z\frac{\partial}{\partial z}.} \tag{6}$$

The ∇ operator is the key. It immediately gives, for the gradient of a scalar field u,

$$\boxed{\nabla u = \frac{\partial u}{\partial r}\hat{\mathbf{e}}_r + \frac{1}{r}\frac{\partial u}{\partial \theta}\hat{\mathbf{e}}_\theta + \frac{\partial u}{\partial z}\hat{\mathbf{e}}_z,} \tag{7}$$

and with it we can work out the divergence ($\nabla\cdot$), curl ($\nabla\times$), and Laplacian (∇^2) of a given field. In doing so, it will be crucial to remember that the cylindrical coordinate base vectors are not constants. Specifically, recall from Section 14.6 that $\hat{\mathbf{e}}_r$ and $\hat{\mathbf{e}}_\theta$ are functions of θ and $\hat{\mathbf{e}}_z$ is a constant, with the only nonzero derivatives being

$$\frac{d\hat{\mathbf{e}}_r}{d\theta} = \hat{\mathbf{e}}_\theta \qquad \text{and} \qquad \frac{d\hat{\mathbf{e}}_\theta}{d\theta} = -\hat{\mathbf{e}}_r. \tag{8}$$

For example, given the vector field $\mathbf{v} = v_r\hat{\mathbf{e}}_r + v_\theta\hat{\mathbf{e}}_\theta + v_z\hat{\mathbf{e}}_z$, consider its divergence

$$\nabla\cdot\mathbf{v} = \left(\hat{\mathbf{e}}_r\frac{\partial}{\partial r} + \hat{\mathbf{e}}_\theta\frac{1}{r}\frac{\partial}{\partial\theta} + \hat{\mathbf{e}}_z\frac{\partial}{\partial z}\right)\cdot\left(v_r\hat{\mathbf{e}}_r + v_\theta\hat{\mathbf{e}}_\theta + v_z\hat{\mathbf{e}}_z\right). \tag{9}$$

Carrying out the dot product on the right-hand side yields nine terms: the first term in ∇ dotted with the first term in $\mathbf{v}$, the first with the second, the first with the third, the second with the first, and so on. As representative, consider the first term dotted with the first,

$$\left(\hat{\mathbf{e}}_r\frac{\partial}{\partial r}\right)\cdot\left(v_r\hat{\mathbf{e}}_r\right). \tag{10}$$

Notice carefully that the latter is ambiguous because there are two operations to carry out, a dot product and a derivative, and we are not told which to do first and which to do second. In the case of (10) we get the same result either way, but that is not true for all of the terms in (9). We state, without proof, that the correct answer in such cases, as could be verified by working the problem in a different way, is always obtained if we *do the differentation first and then the dot product (or cross product if we are computing the curl rather than the divergence)*. Thus, to work out (10) we write

$$\begin{aligned}\left(\hat{\mathbf{e}}_r\frac{\partial}{\partial r}\right)\cdot\left(v_r\hat{\mathbf{e}}_r\right) &= \hat{\mathbf{e}}_r\cdot\frac{\partial}{\partial r}\left(v_r\hat{\mathbf{e}}_r\right) = \hat{\mathbf{e}}_r\cdot\left(\frac{\partial v_r}{\partial r}\hat{\mathbf{e}}_r + v_r\frac{\partial\hat{\mathbf{e}}_r}{\partial r}\right)\\ &= \frac{\partial v_r}{\partial r}\hat{\mathbf{e}}_r\cdot\hat{\mathbf{e}}_r + v_r\hat{\mathbf{e}}_r\cdot\mathbf{0} = \frac{\partial v_r}{\partial r}\end{aligned} \tag{11}$$

since $\partial\hat{\mathbf{e}}_r/\partial r = \mathbf{0}$. Observe that in the first equality, in (11), we have slid the $\partial/\partial r$ past the dot. That step gives $\hat{\mathbf{e}}_r\cdot\partial(v_r\hat{\mathbf{e}}_r)/\partial r$, which is unambiguous; it instructs us to *first* work out the derivative $\partial(v_r\hat{\mathbf{e}}_r)/\partial r$ and *then* the dot product, as explained above in italicized type.

Really, the only terms requiring such special care are the second term dotted with the first (because $\partial\hat{\mathbf{e}}_r/\partial\theta \neq \mathbf{0}$) and the second term dotted with the second (because $\partial\hat{\mathbf{e}}_\theta/\partial\theta \neq \mathbf{0}$):

$$\begin{aligned}\left(\hat{\mathbf{e}}_\theta\frac{1}{r}\frac{\partial}{\partial\theta}\right)\cdot\left(v_r\hat{\mathbf{e}}_r\right) &= \hat{\mathbf{e}}_\theta\frac{1}{r}\cdot\frac{\partial}{\partial\theta}\left(v_r\hat{\mathbf{e}}_r\right)\\ &= \hat{\mathbf{e}}_\theta\frac{1}{r}\cdot\left(\frac{\partial v_r}{\partial\theta}\hat{\mathbf{e}}_r + v_r\frac{\partial\hat{\mathbf{e}}_r}{\partial\theta}\right)\\ &= \hat{\mathbf{e}}_\theta\frac{1}{r}\cdot\left(\frac{\partial v_r}{\partial\theta}\hat{\mathbf{e}}_r + v_r\hat{\mathbf{e}}_\theta\right) = \frac{v_r}{r}\end{aligned} \tag{12}$$

and

$$\begin{aligned}\left(\hat{\mathbf{e}}_\theta \frac{1}{r}\frac{\partial}{\partial\theta}\right)\cdot(v_\theta\hat{\mathbf{e}}_\theta) &= \hat{\mathbf{e}}_\theta\frac{1}{r}\cdot\frac{\partial}{\partial\theta}(v_\theta\hat{\mathbf{e}}_\theta)\\ &= \hat{\mathbf{e}}_\theta\frac{1}{r}\cdot\left(\frac{\partial v_\theta}{\partial\theta}\hat{\mathbf{e}}_\theta + v_\theta\frac{\partial\hat{\mathbf{e}}_\theta}{\partial\theta}\right)\\ &= \hat{\mathbf{e}}_\theta\frac{1}{r}\cdot\left(\frac{\partial v_\theta}{\partial\theta}\hat{\mathbf{e}}_\theta - v_\theta\hat{\mathbf{e}}_r\right) = \frac{1}{r}\frac{\partial v_\theta}{\partial\theta}.\end{aligned} \tag{13}$$

Of the nine terms that result from (9) we have worked out three, in (11)–(13). Let us work out just one more, say the first term in ∇ dotted with the third in $\mathbf{v}$, and then state the final result:

$$\begin{aligned}\left(\hat{\mathbf{e}}_r\frac{\partial}{\partial r}\right)\cdot(v_z\hat{\mathbf{e}}_z) &= \hat{\mathbf{e}}_r\cdot\frac{\partial}{\partial r}(v_z\hat{\mathbf{e}}_z) = \hat{\mathbf{e}}_r\cdot\left(\frac{\partial v_z}{\partial r}\hat{\mathbf{e}}_z + v_z\frac{\partial\hat{\mathbf{e}}_z}{\partial z}\right)\\ &= \hat{\mathbf{e}}_r\cdot\left(\frac{\partial v_z}{\partial r}\hat{\mathbf{e}}_z + \mathbf{0}\right) = 0+0=0.\end{aligned} \tag{14}$$

In fact, five of the nine terms give 0, and the final result is

$$\nabla\cdot\mathbf{v} = \frac{\partial v_r}{\partial r} + \frac{1}{r}v_r + \frac{1}{r}\frac{\partial v_\theta}{\partial\theta} + \frac{\partial v_z}{\partial z}, \tag{15}$$

which is often expressed, more compactly, as

$$\boxed{\nabla\cdot\mathbf{v} = \frac{1}{r}\frac{\partial}{\partial r}(rv_r) + \frac{1}{r}\frac{\partial}{\partial\theta}v_\theta + \frac{\partial}{\partial z}v_z.} \tag{16}$$

Proceeding in similar fashion, one obtains

$$\boxed{\nabla\times\mathbf{v} = \left(\frac{1}{r}\frac{\partial v_z}{\partial\theta} - \frac{\partial v_\theta}{\partial z}\right)\hat{\mathbf{e}}_r + \left(\frac{\partial v_r}{\partial z} - \frac{\partial v_z}{\partial r}\right)\hat{\mathbf{e}}_\theta + \frac{1}{r}\left(\frac{\partial(rv_\theta)}{\partial r} - \frac{\partial v_r}{\partial\theta}\right)\hat{\mathbf{e}}_z} \tag{17}$$

and

$$\boxed{\nabla^2 u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial\theta^2} + \frac{\partial^2 u}{\partial z^2}.} \tag{18}$$

Thus, the gradient and the Laplacian of any scalar field $u(r,\theta,z)$ are given by (7) and (18), respectively, and the divergence and curl of any vector field $\mathbf{v}(r,\theta,z)$ are given by (16) and (17), respectively.

EXAMPLE 1. Consider the scalar field $u = rz^2\sin\theta$, and the vector field $\mathbf{v} = z^2\hat{\mathbf{e}}_r - 2\hat{\mathbf{e}}_\theta + r\hat{\mathbf{e}}_z$. Then (7) and (18) give

$$\nabla u = z^2\sin\theta\hat{\mathbf{e}}_r + z^2\cos\theta\hat{\mathbf{e}}_\theta + 2rz\sin\theta\hat{\mathbf{e}}_z \tag{19}$$

and

$$\nabla^2 u = 0 + \frac{z^2}{r}\sin\theta - \frac{rz^2}{r^2}\sin\theta + 2r\sin\theta = 2r\sin\theta, \tag{20}$$

respectively. To use (16) and (17) to work out $\nabla\cdot\mathbf{v}$ and $\nabla\times\mathbf{v}$ we need, first, to identify the components $v_r(r,\theta,z)$, $v_\theta(r,\theta,z)$, and $v_z(r,\theta,z)$ of $\mathbf{v}$. Since $\mathbf{v} = v_r\hat{\mathbf{e}}_r + v_\theta\hat{\mathbf{e}}_\theta + v_z\hat{\mathbf{e}}_z = z^2\hat{\mathbf{e}}_r - 2\hat{\mathbf{e}}_\theta + r\hat{\mathbf{e}}_z$, we see that $v_r = z^2$, $v_\theta = -2$, and $v_z = r$. Thus,

$$\nabla\cdot\mathbf{v} = \frac{1}{r}\frac{\partial}{\partial r}(rz^2) + \frac{1}{r}\frac{\partial}{\partial\theta}(-2) + \frac{\partial}{\partial z}(r) = \frac{z^2}{r} \tag{21}$$

from (16), and

$$\begin{aligned}\nabla\times\mathbf{v} &= (0-0)\hat{\mathbf{e}}_r + (2z-1)\hat{\mathbf{e}}_\theta + \frac{1}{r}(-2-0)\hat{\mathbf{e}}_z\\ &= (2z-1)\hat{\mathbf{e}}_\theta - \frac{2}{r}\hat{\mathbf{e}}_z\end{aligned} \tag{22}$$

from (17).

COMMENT. In this case $u = rz^2\sin\theta$ is readily re-expressed in terms of x, y, z since $r\sin\theta$ is y. Thus, $u = yz^2$ so

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0 + 0 + 2y = 2y,$$

which does agree with (20) because $y = r\sin\theta$. ∎

As noted above, we need not derive these operators for the case of plane polar coordinates because they merely follow from (7), (16), (17), and (18). That is, if

$$u = u(r,\theta) \qquad \text{and} \qquad \mathbf{v} = v_r(r,\theta)\hat{\mathbf{e}}_r + v_\theta(r,\theta)\hat{\mathbf{e}}_\theta,$$

then all z derivatives are zero and $v_z = 0$ too, so those formulas merely reduce to

$$\nabla u = \frac{\partial u}{\partial r}\hat{\mathbf{e}}_r + \frac{1}{r}\frac{\partial u}{\partial\theta}\hat{\mathbf{e}}_\theta, \tag{23}$$

$$\nabla^2 u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial\theta^2}, \tag{24}$$

$$\nabla\cdot\mathbf{v} = \frac{1}{r}\frac{\partial}{\partial r}(rv_r) + \frac{1}{r}\frac{\partial}{\partial\theta}v_\theta, \tag{25}$$

$$\nabla\times\mathbf{v} = \frac{1}{r}\left(\frac{\partial(rv_\theta)}{\partial r} - \frac{\partial v_r}{\partial\theta}\right)\hat{\mathbf{e}}_z. \tag{26}$$

16.7.2. Spherical coordinates. Here we consider scalar fields u and vector fields

$$\mathbf{v} = v_\rho\hat{\mathbf{e}}_\rho + v_\phi\hat{\mathbf{e}}_\phi + v_\theta\hat{\mathbf{e}}_\theta, \tag{27}$$

where u, v_ρ, v_ϕ, and v_θ are functions of the spherical coordinates ρ, ϕ, θ that were introduced in Section 14.6, the latter being related to x, y, z according to

$$\begin{aligned} x &= \rho \sin\phi \cos\theta, \\ y &= \rho \sin\phi \sin\theta, \\ z &= \rho \cos\phi. \end{aligned} \tag{28}$$

Derivation of ∇, $\nabla \cdot \mathbf{v}$, $\nabla^2 u$, and so on, follows exactly the same lines as indicated above for cylindrical coordinates. For brevity we merely state the main results and leave derivations for the exercises.

$$\boxed{\nabla = \hat{\mathbf{e}}_\rho \frac{\partial}{\partial \rho} + \hat{\mathbf{e}}_\phi \frac{1}{\rho} \frac{\partial}{\partial \phi} + \hat{\mathbf{e}}_\theta \frac{1}{\rho \sin\phi} \frac{\partial}{\partial \theta},} \tag{29}$$

$$\boxed{\nabla u = \frac{\partial u}{\partial \rho} \hat{\mathbf{e}}_\rho + \frac{1}{\rho} \frac{\partial u}{\partial \phi} \hat{\mathbf{e}}_\phi + \frac{1}{\rho \sin\phi} \frac{\partial u}{\partial \theta} \hat{\mathbf{e}}_\theta,} \tag{30}$$

$$\boxed{\nabla^2 u = \frac{1}{\rho^2} \left[\frac{\partial}{\partial \rho} \left(\rho^2 \frac{\partial u}{\partial \rho} \right) + \frac{1}{\sin\phi} \frac{\partial}{\partial \phi} \left(\sin\phi \frac{\partial u}{\partial \phi} \right) + \frac{1}{\sin^2\phi} \frac{\partial^2 u}{\partial \theta^2} \right],} \tag{31}$$

$$\boxed{\nabla \cdot \mathbf{v} = \frac{1}{\rho^2} \frac{\partial}{\partial \rho} (\rho^2 v_\rho) + \frac{1}{\rho \sin\phi} \frac{\partial}{\partial \phi} (v_\phi \sin\phi) + \frac{1}{\rho \sin\phi} \frac{\partial v_\theta}{\partial \theta},} \tag{32}$$

$$\boxed{\begin{aligned} \nabla \times \mathbf{v} = {} & \frac{1}{\rho \sin\phi} \left(\frac{\partial}{\partial \phi} (v_\theta \sin\phi) - \frac{\partial v_\phi}{\partial \theta} \right) \hat{\mathbf{e}}_\rho \\ & + \frac{1}{\rho} \left(\frac{1}{\sin\phi} \frac{\partial v_\rho}{\partial \theta} - \frac{\partial (\rho v_\theta)}{\partial \rho} \right) \hat{\mathbf{e}}_\phi + \frac{1}{\rho} \left(\frac{\partial (\rho v_\phi)}{\partial \rho} - \frac{\partial v_\rho}{\partial \phi} \right) \hat{\mathbf{e}}_\theta. \end{aligned}} \tag{33}$$

EXAMPLE 2. Given $u = \rho^2 \sin^2\phi \sin^2\theta$ and $\mathbf{v} = \rho \hat{\mathbf{e}}_\rho$, use the given formulas to determine $\nabla^2 u$ and $\nabla \cdot \mathbf{v}$. From (31),

$$\begin{aligned} \nabla^2 u = \frac{1}{\rho^2} \Bigg[& \frac{\partial}{\partial \rho} \left(2\rho^3 \sin^2\phi \sin^2\theta \right) + \frac{1}{\sin\phi} \frac{\partial}{\partial \phi} \left(2\rho^2 \sin^2\phi \cos\phi \sin^2\theta \right) \\ & + \frac{1}{\sin^2\phi} 2\rho^2 \sin^2\phi (\cos^2\theta - \sin^2\theta) \Bigg] = 2, \end{aligned}$$

where we omit the final steps, for brevity, since they are straightforward. And from (32), with $v_\rho = \rho$, $v_\phi = 0$, and $v_\theta = 0$,

$$\nabla \cdot \mathbf{v} = \frac{1}{\rho^2} \frac{\partial}{\partial \rho} \rho^3 = 3.$$

COMMENT. We see from (28) that $u = \rho^2 \sin^2\phi \sin^2\theta$ is readily expressed in terms of Cartesian coordinates as $u = y^2$, and $\mathbf{v} = \rho\hat{\mathbf{e}}_\rho$ is simply the position vector $\mathbf{v} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$. Thus,

$$\nabla^2 u = \frac{\partial^2}{\partial x^2}(y^2) + \frac{\partial^2}{\partial y^2}(y^2) + \frac{\partial^2}{\partial z^2}(y^2) = 0 + 2 + 0 = 2$$

and

$$\nabla \cdot \mathbf{v} = \frac{\partial}{\partial x}(x) + \frac{\partial}{\partial y}(y) + \frac{\partial}{\partial z}(z) = 1 + 1 + 1 = 3,$$

both of which results agree with the results found above using (31) and (32). ∎

A word of CAUTION regarding the use of determinants for the representation of $\nabla \times \mathbf{v}$ in non-Cartesian coordinate systems: Recall that it turns out that the cross product $\mathbf{u} \times \mathbf{v}$ of two vectors $\mathbf{u} = u_x\hat{\mathbf{i}} + u_y\hat{\mathbf{j}} + u_z\hat{\mathbf{k}}$ and $\mathbf{v} = v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}$ can be expressed as the determinant

$$\mathbf{u} \times \mathbf{v} = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ u_x & u_y & u_z \\ v_x & v_y & v_z \end{vmatrix}, \tag{34}$$

as can be verified by working out $(u_x\hat{\mathbf{i}} + u_y\hat{\mathbf{j}} + u_z\hat{\mathbf{k}}) \times (v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}})$ term by term, expanding the determinant in (34), and verifying that the two results are identical.

Similarly for *any* right-handed set of orthonormal base vectors $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3$, where by right-handed we mean that $\hat{\mathbf{e}}_1 \times \hat{\mathbf{e}}_2 = \hat{\mathbf{e}}_3$, $\hat{\mathbf{e}}_2 \times \hat{\mathbf{e}}_3 = \hat{\mathbf{e}}_1$, and $\hat{\mathbf{e}}_3 \times \hat{\mathbf{e}}_1 = \hat{\mathbf{e}}_2$. That is, if $\mathbf{u} = u_1\hat{\mathbf{e}}_1 + u_2\hat{\mathbf{e}}_2 + u_3\hat{\mathbf{e}}_3$ and $\mathbf{v} = v_1\hat{\mathbf{e}}_1 + v_2\hat{\mathbf{e}}_2 + v_3\hat{\mathbf{e}}_3$, then

$$\mathbf{u} \times \mathbf{v} = \begin{vmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix}, \tag{35}$$

as can be verified by working out $(u_1\hat{\mathbf{e}}_1 + u_2\hat{\mathbf{e}}_2 + u_3\hat{\mathbf{e}}_3) \times (v_1\hat{\mathbf{e}}_1 + v_2\hat{\mathbf{e}}_2 + v_3\hat{\mathbf{e}}_3)$ term by term, expanding the determinant in (35), and verifying that the two results are identical. For instance, $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3$ might be the right-handed cylindrical coordinate base vectors $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z$, or the right-handed spherical coordinate base vectors $\hat{\mathbf{e}}_\rho, \hat{\mathbf{e}}_\phi, \hat{\mathbf{e}}_\theta$.

However, if the $\mathbf{u}$ vector is not an ordinary vector but is the vector differential operator ∇, then the determinant form for $\nabla \times \mathbf{v}$ gives INCORRECT results for non-Cartesian systems. For example, the cylindrical coordinate formula

$$\nabla \times \mathbf{v} = \begin{vmatrix} \hat{\mathbf{e}}_r & \hat{\mathbf{e}}_\theta & \hat{\mathbf{e}}_z \\ \partial/\partial r & (1/r)\partial/\partial\theta & \partial/\partial z \\ v_r & v_\theta & v_z \end{vmatrix}$$

is INCORRECT, as can be seen by expanding the determinant and comparing the result with the correct result given by (17). A generalized determinant formula for $\nabla \times \mathbf{v}$ that *is* correct will be given below by (43).

16.7.3. General orthogonal coordinates. Let us give the extension of these results to any right-handed orthogonal curvilinear coordinate system with coordinates q_1, q_2, q_3 and corresponding base vectors $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3$. Let q_1, q_2, q_3 be related to the reference Cartesian coordinates according to

$$\begin{aligned} x &= x(q_1, q_2, q_3), \\ y &= y(q_1, q_2, q_3), \\ z &= z(q_1, q_2, q_3). \end{aligned} \tag{36}$$

For instance, q_1, q_2, q_3 might be the cylindrical coordinates r, θ, z, respectively.

The key is to identify the quantities h_1, h_2, h_3 in

$$d\mathbf{R} = h_1 dq_1\, \hat{\mathbf{e}}_1 + h_2 dq_2\, \hat{\mathbf{e}}_2 + h_3 dq_3\, \hat{\mathbf{e}}_3, \tag{37}$$

which are given by

$$\boxed{\begin{aligned} h_1 &= \sqrt{\left(\frac{\partial x}{\partial q_1}\right)^2 + \left(\frac{\partial y}{\partial q_1}\right)^2 + \left(\frac{\partial z}{\partial q_1}\right)^2}, \\ h_2 &= \sqrt{\left(\frac{\partial x}{\partial q_2}\right)^2 + \left(\frac{\partial y}{\partial q_2}\right)^2 + \left(\frac{\partial z}{\partial q_2}\right)^2}, \\ h_3 &= \sqrt{\left(\frac{\partial x}{\partial q_3}\right)^2 + \left(\frac{\partial y}{\partial q_3}\right)^2 + \left(\frac{\partial z}{\partial q_3}\right)^2}. \end{aligned}} \tag{38}$$

With the h_j's determined from (38), we have

$$\boxed{\nabla = \hat{\mathbf{e}}_1 \frac{1}{h_1}\frac{\partial}{\partial q_1} + \hat{\mathbf{e}}_2 \frac{1}{h_2}\frac{\partial}{\partial q_2} + \hat{\mathbf{e}}_3 \frac{1}{h_3}\frac{\partial}{\partial q_3},} \tag{39}$$

$$\boxed{\nabla u = \frac{1}{h_1}\frac{\partial u}{\partial q_1}\hat{\mathbf{e}}_1 + \frac{1}{h_2}\frac{\partial u}{\partial q_2}\hat{\mathbf{e}}_2 + \frac{1}{h_3}\frac{\partial u}{\partial q_3}\hat{\mathbf{e}}_3,} \tag{40}$$

$$\boxed{\nabla^2 u = \frac{1}{h_1 h_2 h_3}\left[\frac{\partial}{\partial q_1}\left(\frac{h_2 h_3}{h_1}\frac{\partial u}{\partial q_1}\right) + \frac{\partial}{\partial q_2}\left(\frac{h_1 h_3}{h_2}\frac{\partial u}{\partial q_2}\right) + \frac{\partial}{\partial q_3}\left(\frac{h_1 h_2}{h_3}\frac{\partial u}{\partial q_3}\right)\right],} \tag{41}$$

$$\boxed{\nabla \cdot \mathbf{v} = \frac{1}{h_1 h_2 h_3}\left[\frac{\partial}{\partial q_1}(h_2 h_3 v_1) + \frac{\partial}{\partial q_2}(h_1 h_3 v_2) + \frac{\partial}{\partial q_3}(h_1 h_2 v_3)\right],} \tag{42}$$

and

$$\boxed{\nabla \times \mathbf{v} = \frac{1}{h_1 h_2 h_3}\begin{vmatrix} h_1\hat{\mathbf{e}}_1 & h_2\hat{\mathbf{e}}_2 & h_3\hat{\mathbf{e}}_3 \\ \dfrac{\partial}{\partial q_1} & \dfrac{\partial}{\partial q_2} & \dfrac{\partial}{\partial q_3} \\ h_1 v_1 & h_2 v_2 & h_3 v_3 \end{vmatrix},} \tag{43}$$

where v_1, v_2, v_3 denote the $\hat{\mathbf{e}}_1$, $\hat{\mathbf{e}}_2$, $\hat{\mathbf{e}}_3$ components of $\mathbf{v}$, respectively; that is, $\mathbf{v} = v_1\hat{\mathbf{e}}_1 + v_2\hat{\mathbf{e}}_2 + v_3\hat{\mathbf{e}}_3$.

EXAMPLE 3. In cylindrical coordinates, let us identify q_1, q_2, q_3 as r, θ, z, respectively, although the order will not matter provided that q_1, q_2, q_3 is a right-handed system. The relations (36) are

$$\begin{aligned} x &= r\cos\theta = q_1 \cos q_2, \\ y &= r\sin\theta = q_1 \sin q_2, \\ z &= z \qquad\;\; = q_3 \end{aligned}$$

and putting these into (38) gives $h_1 = 1$, $h_2 = q_1$, $h_3 = 1$. With these h_j's, (39)–(43) do give results that are the same as (6), (7), (18), (16), and (17), respectively (Exercise 10). ∎

Closure. The key, in obtaining the divergence, gradient, curl, and Laplacian operators for a non-Cartesian coordinate system is to obtain the ∇ operator. That can be done from its Cartesian form, (3), by expressing $x, y, z, \hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$ in terms of the desired coordinates and base vectors. With ∇ in hand, the gradient of a scalar field u is given immediately as ∇u. Knowing the derivatives of the non-Cartesian base vectors, we can also work out expressions for $\nabla \cdot \mathbf{v}$, $\nabla \times \mathbf{v}$, and $\nabla^2 u$. The results, for cylindrical and spherical coordinates, are given in this section and framed for emphasis. The key point to keep in mind, in deriving formulas for $\nabla \cdot \mathbf{v}$ and $\nabla \times \mathbf{v}$, is that the derivatives within the ∇ act not only upon the scalar components within $\mathbf{v}$ but also upon the *base vectors* within $\mathbf{v}$, and the derivatives of the base vectors are not all zero for a non-Cartesian system, in general. Similarly, the derivatives within the first ∇, in $\nabla^2 u = \nabla \cdot \nabla u$, act not only upon u but also upon the base vectors within the second ∇. Of course it should be appreciated that once the results (6), (7), (16)–(18), and (30)–(33) are in hand we can simply use those formulas without dealing with the above-mentioned details.

EXERCISES 16.7

Cylindrical coordinates:

1. Evaluate ∇u, $\nabla^2 u$, $\nabla \cdot \mathbf{v}$, and $\nabla \times \mathbf{v}$, using the relevant formulas (7), (16), (17), and (18).

(a) $u = r, \quad \mathbf{v} = r\hat{\mathbf{e}}_r + z\hat{\mathbf{e}}_z$
(b) $u = r\sin 2\theta, \quad \mathbf{v} = \cos\theta\hat{\mathbf{e}}_r - \sin\theta\hat{\mathbf{e}}_\theta$
(c) $u = r^2\sin\theta, \quad \mathbf{v} = z^2\hat{\mathbf{e}}_r$
(d) $u = z^3 r, \quad \mathbf{v} = \hat{\mathbf{e}}_\theta$
(e) $u = 1/r, \quad \mathbf{v} = r\sin\theta\hat{\mathbf{e}}_z$
(f) $u = 6, \quad \mathbf{v} = z^2\hat{\mathbf{e}}_r + (r^2 z + \cos^2\theta)\hat{\mathbf{e}}_z$
(g) $u = r^3\cos\theta, \quad \mathbf{v} = 6\hat{\mathbf{e}}_r - 3\hat{\mathbf{e}}_\theta - 2\hat{\mathbf{e}}_z$
(h) $u = r^2 + z^2, \quad \mathbf{v} = rz\hat{\mathbf{e}}_r + r^2 z^2\cos\theta\hat{\mathbf{e}}_\theta$

2. Evaluate $\nabla \cdot \mathbf{v}$ two different ways, and show that the results are in agreement with each other: first, use (16); second, re-express $\mathbf{v}$ in terms of $x, y, z, \hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$ and use the formula $\nabla \cdot \mathbf{v} = \partial v_x/\partial x + \partial v_y/\partial y + \partial v_z/\partial z$.

(a) $\mathbf{v} = \hat{\mathbf{e}}_r$
(b) $\mathbf{v} = \hat{\mathbf{e}}_\theta$
(c) $\mathbf{v} = r\hat{\mathbf{e}}_r + z\hat{\mathbf{e}}_z$
(d) $\mathbf{v} = \cos^2\theta\hat{\mathbf{e}}_\theta$
(e) $\mathbf{v} = r^2 z\sin\theta\hat{\mathbf{e}}_z$
(f) $\mathbf{v} = \cos\theta\hat{\mathbf{e}}_r - \sin\theta\hat{\mathbf{e}}_\theta$
(g) $\mathbf{v} = r\hat{\mathbf{e}}_\theta$
(h) $\mathbf{v} = r^3\hat{\mathbf{e}}_r$
(i) $\mathbf{v} = \cos\theta\hat{\mathbf{e}}_z$

3. Carry out steps analogous to those followed in (9)–(16), in deriving the formula (16) for $\nabla \cdot \mathbf{v}$, to derive

(a) the formula (26) for $\nabla \times \mathbf{v}$ in plane polar coordinates
(b) the formula (17) for $\nabla \times \mathbf{v}$ in cylindrical coordinates

4. Derive the formula (18) for $\nabla^2 u$

(a) by writing $\nabla^2 u = \nabla \cdot \nabla u$ and using (7) and (16)
(b) by beginning with the Cartesian version

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2},$$

introducing the relations $x = r\cos\theta$, $y = r\sin\theta$, $z = z$, and using chain differentiation. HINT: $\partial^2/\partial x^2$, for example, is a double application of the operator

$$\frac{\partial}{\partial x} = \frac{\partial}{\partial r}\frac{\partial r}{\partial x} + \frac{\partial}{\partial \theta}\frac{\partial \theta}{\partial x} + \frac{\partial}{\partial z}\frac{\partial z}{\partial x} = \cos\theta\frac{\partial}{\partial r} - \frac{\sin\theta}{r}\frac{\partial}{\partial \theta}.$$

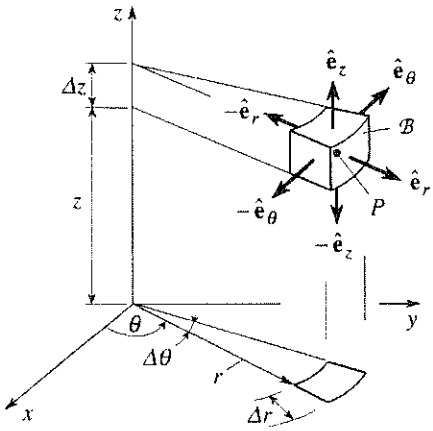
Similarly,

$$\frac{\partial}{\partial y} = \sin\theta\frac{\partial}{\partial r} + \frac{\cos\theta}{r}\frac{\partial}{\partial \theta} \quad \text{and} \quad \frac{\partial}{\partial z} = \frac{\partial}{\partial z}.$$

5. (a) Derive the formula (16) for $\nabla \cdot \mathbf{v}$ using the limit definition

$$\operatorname{div}\mathbf{v} = \lim_{\mathcal{B}\to 0}\left\{\frac{\int_S \hat{\mathbf{n}}\cdot\mathbf{v}\,dA}{V}\right\}, \tag{5.1}$$

and choosing $\mathcal{B}$ as shown in the accompanying figure, bounded by constant-r, constant-θ, constant-z surfaces. HINT: The analogous derivation, for Cartesian coordinates, is given in Section 16.3.



(b) Derive the formula (9) for ∇u using the limit definition

$$\operatorname{grad} u = \lim_{\mathcal{B}\to 0}\left\{\frac{\int_S \hat{\mathbf{n}}\,u\,dA}{V}\right\}, \tag{5.2}$$

and choosing $\mathcal{B}$ as shown in the figure.
(c) Derive the formula (17) for $\nabla \times \mathbf{v}$ using the limit definition

$$\operatorname{curl}\mathbf{v} = \lim_{\mathcal{B}\to 0}\left\{\frac{\int_S \hat{\mathbf{n}}\times\mathbf{v}\,dA}{V}\right\}, \tag{5.3}$$

and choosing $\mathcal{B}$ as shown in the figure.

6. Show that the given u is a solution of the *Laplace equation* $\nabla^2 u = 0$.

(a) $u = r^n\sin n\theta$ for any integer n (for $r \neq 0$ if n is negative)
(b) $u = r^n\cos n\theta$ for any integer n (for $r \neq 0$ if n is negative)
(c) $u = \ln r \quad (r \neq 0)$
(d) $u = \dfrac{1}{\sqrt{r^2 + z^2}} \quad (r^2 + z^2 \neq 0)$

7. Besides $\nabla^2 u$, we can also have $\nabla^2 \mathbf{v}$, the Laplacian operating on a vector field. Rather than ask you to work out $\nabla^2 \mathbf{v}$ in

cylindrical coordinates, because it is a bit messy, we ask you to derive the plane polar coordinate result

$$\nabla^2\mathbf{v} = \left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v_r}{\partial r}\right) + \frac{1}{r^2}\left(\frac{\partial^2 v_r}{\partial\theta^2} - 2\frac{\partial v_\theta}{\partial\theta} - v_r\right)\right]\hat{\mathbf{e}}_r + \left[\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial v_\theta}{\partial r}\right) + \frac{1}{r^2}\left(\frac{\partial^2 v_\theta}{\partial\theta^2} + 2\frac{\partial v_r}{\partial\theta} - v_\theta\right)\right]\hat{\mathbf{e}}_\theta \tag{7.1}$$

by working out

$$\left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial\theta^2}\right)(v_r(r,\theta)\hat{\mathbf{e}}_r + v_\theta(r,\theta)\hat{\mathbf{e}}_\theta)$$

term by term and remembering that $\partial\hat{\mathbf{e}}_r/\partial r = \partial\hat{\mathbf{e}}_\theta/\partial r = 0$, $\partial\hat{\mathbf{e}}_r/\partial\theta = \hat{\mathbf{e}}_\theta$, and $\partial\hat{\mathbf{e}}_\theta/\partial\theta = -\hat{\mathbf{e}}_r$.

8. Use (7.1), above, to evaluate $\nabla^2\mathbf{v}$.

(a) $\mathbf{v} = r^2\hat{\mathbf{e}}_r$ (b) $\mathbf{v} = \cos\theta\hat{\mathbf{e}}_r - \sin\theta\hat{\mathbf{e}}_\theta$
(c) $\mathbf{v} = \hat{\mathbf{e}}_\theta$ (d) $\mathbf{v} = r\cos\theta\hat{\mathbf{e}}_r - r^2\sin 3\theta\hat{\mathbf{e}}_\theta$

9. Write out the *biharmonic equation* $\nabla^4 u = 0$, where $\nabla^4 \equiv \nabla^2\nabla^2$, in both Cartesian coordinates and cylindrical coordinates.

10. Verify that for the cylindrical coordinates $q_1 = r$, $q_2 = \theta$, $q_3 = z$, equations (39)–(43) do give the same results as (6), (7), (18), (16), and (17) as claimed in Example 3.

Spherical coordinates:

11. Evaluate ∇u, $\nabla^2 u$, $\nabla\cdot\mathbf{v}$, and $\nabla\times\mathbf{v}$, using the relevant formulas (30)–(33).

(a) $u = \rho$, $\mathbf{v} = \rho\hat{\mathbf{e}}_\rho$
(b) $u = \rho^2$, $\mathbf{v} = 3\hat{\mathbf{e}}_\phi$
(c) $u = \sin\phi$, $\mathbf{v} = \rho\hat{\mathbf{e}}_\phi$
(d) $u = \rho^2\sin\theta$, $\mathbf{v} = \hat{\mathbf{e}}_\rho + \hat{\mathbf{e}}_\phi + \hat{\mathbf{e}}_\theta$
(e) $u = \rho\cos\theta$, $\mathbf{v} = \rho^3\hat{\mathbf{e}}_\theta$
(f) $u = \cos\phi$, $\mathbf{v} = 6\hat{\mathbf{e}}_\theta$

12. Using the expression

$$\nabla = \hat{\mathbf{e}}_\rho\frac{\partial}{\partial\rho} + \hat{\mathbf{e}}_\phi\frac{1}{\rho}\frac{\partial}{\partial\phi} + \hat{\mathbf{e}}_\theta\frac{1}{\rho\sin\phi}\frac{\partial}{\partial\theta},$$

and, as needed, the expressions for the space derivatives of $\hat{\mathbf{e}}_\rho, \hat{\mathbf{e}}_\phi, \hat{\mathbf{e}}_\theta$ given in Section 14.6.3, derive

(a) the formula (32) for $\nabla\cdot\mathbf{v}$
(b) the formula (33) for $\nabla\times\mathbf{v}$
(c) the formula (31) for $\nabla^2 u$

13. Find values of α, if any, such that $u = \rho^\alpha$ is a solution of the *Laplace equation* $\nabla^2 u = 0$.

14. For spherical polars, x, y, z are related to ρ, ϕ, θ according to (28). Letting q_1, q_2, q_3 be ρ, ϕ, θ, respectively, show that (39)–(43) give expressions for ∇, ∇u, $\nabla^2 u$, $\nabla\cdot\mathbf{v}$, and $\nabla\times\mathbf{v}$ that agree with (29)–(33).

15. The vector equation

$$\frac{\partial\mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} = -\frac{1}{\sigma}\nabla p, \tag{15.1}$$

where $\mathbf{v}$ is the vector velocity field, p is the scalar pressure field, t is the time, and σ is the fluid mass density, is known in fluid mechanics as the *equation of motion* (for a fluid having no viscosity) since it is the fluid mechanics version of Newton's second law of motion $\mathbf{F} = m\mathbf{a}$ for a particle of constant mass m. Write the three scalar equations that follow from (15.1), in

(a) Cartesian coordinates x, y, z
(b) cylindrical coordinates r, θ, z
(c) spherical coordinates ρ, ϕ, θ

16.8 Divergence Theorem

16.8.1. Divergence theorem. In Sections 16.1–16.7 we introduce the quantities that we will need – div, grad, curl, Laplacian, and so on. In this section we begin to use those quantities to obtain important results. In particular, we derive the Gauss divergence theorem and use it to derive well known equations in fluid mechanics and heat transfer.*

*Generally attributed to *Gauss*, the divergence theorem was obtained independently by the Rus-

THEOREM 16.8.1 *Divergence Theorem*
Let $\mathcal{V}$ be a closed region in 3-space, the boundary of which is a piecewise smooth orientable surface $\mathcal{S}$, and let $\mathbf{v}$ be a vector field that is defined and C^1 in $\mathcal{V}$. If $\hat{\mathbf{n}}$ denotes the outward unit normal on $\mathcal{S}$, then

$$\boxed{\int_{\mathcal{V}} \operatorname{div} \mathbf{v}\, dV = \int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA.} \tag{1}$$

Outline of Proof: We begin by proving (1) for the simple case where $\mathcal{V}$ is a rectangular prism $x_1 \le x \le x_2$, $y_1 \le y \le y_2$, $z_1 \le z \le z_2$. Write out

$$\int_{\mathcal{V}} \operatorname{div} \mathbf{v}\, dV = \int_{\mathcal{V}} \left(\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z} \right) dV. \tag{2}$$

Consider the first term on the right:

$$\begin{aligned} \int_{\mathcal{V}} \frac{\partial v_x}{\partial x}\, dV &= \int_{z_1}^{z_2} \int_{y_1}^{y_2} \left(\int_{x_1}^{x_2} \frac{\partial v_x}{\partial x}\, dx \right) dy\, dz \\ &= \int_{z_1}^{z_2} \int_{y_1}^{y_2} v_x(x_2, y, z)\, dy\, dz - \int_{z_1}^{z_2} \int_{y_1}^{y_2} v_x(x_1, y, z)\, dy\, dz \\ &= \int\!\!\int_{x=x_2 \text{ face}} \mathbf{v} \cdot \hat{\mathbf{i}}\, dy\, dz + \int\!\!\int_{x=x_1 \text{ face}} \mathbf{v} \cdot (-\hat{\mathbf{i}})\, dy\, dz. \end{aligned} \tag{3}$$

Since $\hat{\mathbf{n}} = +\hat{\mathbf{i}}$ on the $x = x_2$ face and $\hat{\mathbf{n}} = -\hat{\mathbf{i}}$ on the $x = x_1$ face (Fig. 1), (3) amounts to

$$\int_{\mathcal{V}} \frac{\partial v_x}{\partial x}\, dV = \int\!\!\int_{x_1, x_2 \text{ faces}} \mathbf{v} \cdot \hat{\mathbf{n}}\, dA. \tag{4}$$

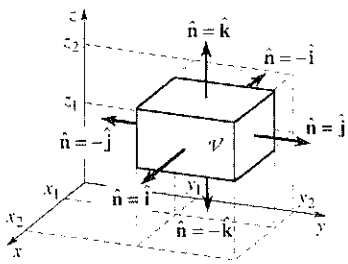


Figure 1. The prism $\mathcal{V}$.

Similarly,

$$\begin{aligned} \int_{\mathcal{V}} \frac{\partial v_y}{\partial y}\, dV &= \int_{z_1}^{z_2} \int_{x_1}^{x_2} \left(\int_{y_1}^{y_2} \frac{\partial v_y}{\partial y}\, dy \right) dx\, dz \\ &= \text{etc.} = \int\!\!\int_{y_1, y_2 \text{ faces}} \mathbf{v} \cdot \hat{\mathbf{n}}\, dA \end{aligned} \tag{5}$$

and

$$\int_{\mathcal{V}} \frac{\partial v_z}{\partial z}\, dV = \int\!\!\int_{z_1, z_2 \text{ faces}} \mathbf{v} \cdot \hat{\mathbf{n}}\, dA. \tag{6}$$

sian mathematician *Michel Ostrogradsky* (1801–1861), and was published in the memoirs of the St. Petersburg Academy of Sciences in 1831.

Then, adding (4)–(6) gives (1). We need to justify using different orders of integration in the three volume integrals; that is, to integrate $\partial v_x/\partial x$, $\partial v_y/\partial y$, $\partial v_z/\partial z$ we integrated first with respect to x, y, z, respectively. Since the theorem assumes $\mathbf{v}$ to be C^1, it follows that the integrand [written out in the right-hand side of (2)] is continuous, so the integral has a unique value, independent of the order of integration. [See equations (8) and (24), and the associated discussion, in Section 15.3.]

The preceding proof can be generalized to any closed region $\mathcal{V}$ in 3-space, the boundary of which is a piecewise smooth orientable surface. A partial generalization is covered in the exercises. Let us use this space instead to outline a different approach which has the advantage of flowing directly from the limit definition of the divergence,

$$\operatorname{div} \mathbf{v}(P) = \lim_{\mathcal{B}\to 0} \left\{ \frac{\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA}{V} \right\}, \tag{7}$$

but understand that our presentation will be heuristic, not rigorous. To begin, let us break $\mathcal{V}$ into N subregions, or cells, where N is very large; one might think of $\mathcal{V}$ as being filled with millions of fish eggs. Let us number them, from 1 to N, call the jth cell $\mathcal{B}_j$, with surface $\mathcal{S}_j$, volume dV_j, and with some point in $\mathcal{B}_j$ designated as P_j. Some such cells are sketched in Fig. 2a, at an exaggerated scale so we can label them. Let us write down (7) for each cell. Dropping the limit, but reminding ourselves of it by writing dV in place of V, we have the N equations

$$\begin{aligned} \operatorname{div} \mathbf{v}(P_1)dV_1 &= \int_{\mathcal{S}_1} \hat{\mathbf{n}}_1 \cdot \mathbf{v}\, dA \\ &\vdots \\ \operatorname{div} \mathbf{v}(P_N)dV_N &= \int_{\mathcal{S}_N} \hat{\mathbf{n}}_N \cdot \mathbf{v}\, dA. \end{aligned} \tag{8}$$

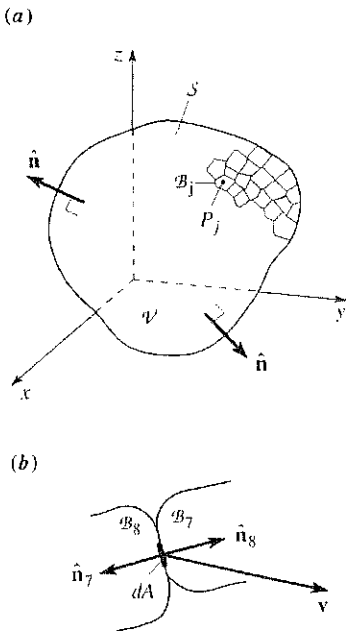


Figure 2. Partition of $\mathcal{V}$.

Adding these, the sum of the left-hand sides is $\int_{\mathcal{V}} \operatorname{div} \mathbf{v}\, dV$, namely, the left-hand member of (1). Interpreting the sum of the right-hand sides of (8) is a bit trickier. Consider two adjacent cells, say $\mathcal{B}_7$ and $\mathcal{B}_8$ (Fig. 2b), and observe that the contributions $\hat{\mathbf{n}}_7 \cdot \mathbf{v}\, dA$ to $\int_{\mathcal{S}_7} \hat{\mathbf{n}}_7 \cdot \mathbf{v}\, dA$ and $\hat{\mathbf{n}}_8 \cdot \mathbf{v}\, dA$ to $\int_{\mathcal{S}_8} \hat{\mathbf{n}}_8 \cdot \mathbf{v}\, dA$ exactly cancel because $\hat{\mathbf{n}}_8 = -\hat{\mathbf{n}}_7$. Such cancellation occurs along all of the internal boundaries. Only the contributions from the dA elements lying on the outer surface $\mathcal{S}$ survive because these have no neighbors and hence suffer no such cancellation. Thus, the sum of the right-hand sides of (8) is $\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA$, namely, the right-hand member of (1). ■

From the preceding heuristic proof we can see that the divergence theorem (1) [more precisely, the divergence theorem *formula* (1) since (1) is a formula, not a theorem] is actually a macro version of the micro statement (7). Also important, in understanding (1), is its relationship with the fundamental theorem of the integral calculus, which we state [for comparison with (1)] in this form:

$$\int_a^b F'(x)\, dx = F(x)\Big|_{x=a}^{x=b}, \tag{9}$$

if $F'(x)$ is continuous on $a \le x \le b$. That is, if the integrand is a derivative and is continuous then the integral can be reduced to an evaluation at the boundary of the integration region, namely, the endpoints a and b. Analogously, the volume integral in (1) can be reduced to an evaluation at the boundary of the region $\mathcal{V}$ (namely, on its surface $\mathcal{S}$) if the integrand is a kind of derivative, namely, a divergence. In fact, (1) reduces to (9) in the one-dimensional case, where $\mathbf{v} = v_x(x,y,z)\hat{\mathbf{i}} + v_y(x,y,z)\hat{\mathbf{j}} + v_z(x,y,z)\hat{\mathbf{k}}$ is merely $v_x(x)\hat{\mathbf{i}}$. Thus, the three-dimensional generalization of the derivative, $F'(x)$ in (9), is the divergence div $\mathbf{v}$ in (1).

To evaluate the integrals in (1), we need expressions for dV and dA. Of these, dV is easy since it depends only on the coordinate system being used and we know, once and for all, that

$$dV = \begin{cases} dx\,dy\,dz & \text{(Cartesian)} \\ r\,dr\,d\theta\,dz & \text{(cylindrical)} \\ \rho^2\,|\sin\phi|\,d\rho\,d\phi\,d\theta. & \text{(spherical)} \end{cases} \tag{10}$$

However, dA is harder inasmuch as it is specific to the particular surface under consideration. The general formula for dA was given in Section 15.5 by (4) and (5), and special cases were covered by (13) and (18) therein. Since you may not have studied Section 15.5, let us try to get by, in this section and the next, relying only on equation (18) from Section 15.5, namely,

$$dA = \sqrt{1 + f_x^2 + f_y^2}\,dx\,dy \tag{11}$$

if the surface $\mathcal{S}$ is known in the form $z = f(x,y)$. The subscripts in (11) denote partial derivatives. Here, dA is the desired area element on $\mathcal{S}$ and $dx\,dy$ is the projection, or shadow, of that area onto the x, y plane (Fig. 3).

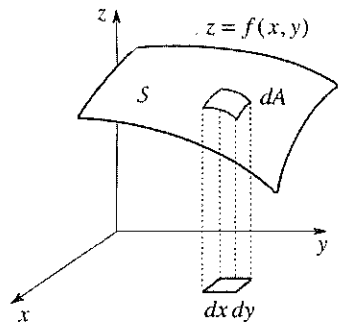


Figure 3. Interpretation of (11).

In the example to follow, we illustrate the "mechanics" of the calculations in (1), and then we turn to physical applications.

EXAMPLE 1. *Verification of (1).* Let $\mathbf{v} = x(y+1)z^3\hat{\mathbf{j}}$, and let $\mathcal{V}$ be the pentahedron shown in Fig. 4, with faces given by the planes $x = 0$, $x = 2$, $y = 0$, $z = 0$, and the slanted plane $EFGH$. The problem that we pose is merely to evaluate both sides of (1), for that case, and to verify that they are equal. Naturally, such verification by no means proves the truth of (1), but if the left- and right-hand sides of (1) are unequal, for this single example, that result would suffice to *dis*prove (1).

We will need the equation of the plane $EFGH$. We know it is of the form

$$ax + by + cz = d, \tag{12}$$

and if we force the plane to go through three noncollinear point, such as E, F, and G, that should uniquely determine its equation. Since $E = (2,0,1)$, $F = (0,0,1)$, and $G = (0,1,0)$, we obtain, from (12), $2a + c = d$, $c = d$, and $b = d$. Thus, $a = 0$, $b = d$, $c = d$ so (12) becomes $dy + dz = d$, or

$$y + z = 1. \tag{13}$$

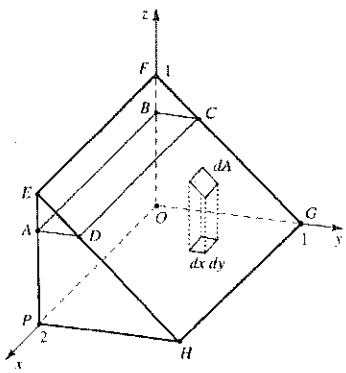


Figure 4. The pentahedron $\mathcal{V}$.

We begin with the left-hand side of (1):

$$\begin{aligned}\int_{\mathcal{V}} \operatorname{div} \mathbf{v}\, dV &= \int_{\mathcal{V}} \left\{ \frac{\partial}{\partial x}(0) + \frac{\partial}{\partial y}[x(y+1)z^3] + \frac{\partial}{\partial z}(0) \right\} dV \\ &= \int_0^1 \left\{ \int_0^{1-z} \int_0^2 x z^3 \, dx\, dy \right\} dz \\ &= 2\int_0^1 z^3(1-z)\, dz = \frac{1}{10}, \end{aligned} \tag{14}$$

where the limits $\int_0^{1-z} \int_0^2$ correspond to the shaded rectangle $ABCD$.

Turning to the surface integral in (1), we consider the five faces separately. Since those faces are planes, $\hat{\mathbf{n}}$ is simply a constant over each one. On the $x = 0$ face $\hat{\mathbf{n}} = -\hat{\mathbf{i}}$, and $\hat{\mathbf{n}} \cdot \mathbf{v} = -\hat{\mathbf{i}} \cdot x(y+1)z^3\hat{\mathbf{j}} = 0$, so $\int \hat{\mathbf{n}} \cdot \mathbf{v}\, dA = \int 0\, dA = 0$ over that face. Similarly, for the $z = 0$ face, since $\hat{\mathbf{n}} \cdot \mathbf{v} = (-\hat{\mathbf{k}}) \cdot x(y+1)z^3\hat{\mathbf{j}} = 0$ there. On $y = 0$, dA is $dx\, dz$, and

$$\begin{aligned}\int_{y=0 \text{ face}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA &= \int_{y=0 \text{ face}} -\hat{\mathbf{j}} \cdot [x(y+1)z^3\hat{\mathbf{j}}]\Big|_{y=0} dA \\ &= -\int_0^1 \int_0^2 x z^3 \, dx\, dz = -\frac{1}{2}. \end{aligned} \tag{15}$$

Finally, consider the slanted face. We need expressions for $\hat{\mathbf{n}}$ and dA. One way to compute $\hat{\mathbf{n}}$ is to cross the vector $\mathbf{HG}$ (from H to G) into the vector $\mathbf{HE}$ (from H to E):

$$\mathbf{n} = \mathbf{HG} \times \mathbf{HE} = -2\hat{\mathbf{i}} \times (-\hat{\mathbf{j}} + \hat{\mathbf{k}}) = 2\hat{\mathbf{k}} + 2\hat{\mathbf{j}}.$$

Normalizing the latter (i.e., to unit length),

$$\hat{\mathbf{n}} = \frac{\mathbf{n}}{\|\mathbf{n}\|} = \frac{2\hat{\mathbf{j}} + 2\hat{\mathbf{k}}}{\sqrt{8}} = \frac{1}{\sqrt{2}}(\hat{\mathbf{j}} + \hat{\mathbf{k}}). \tag{16}$$

Next, (13) gives $z = 1 - y$ so $f(x, y)$ is $1 - y$, $f_x = 0$, $f_y = -1$, and (11) gives

$$dA = \sqrt{1 + (0)^2 + (-1)^2}\, dx\, dy = \sqrt{2}\, dx\, dy. \tag{17}$$

Thus,

$$\begin{aligned}\int_{y+z=1 \text{ face}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA &= \int_0^1 \int_0^2 \left(\frac{\hat{\mathbf{j}} + \hat{\mathbf{k}}}{\sqrt{2}} \right) \cdot [x(y+1)z^3\hat{\mathbf{j}}]\Big|_{z=1-y} (\sqrt{2}\, dx\, dy) \\ &= \int_0^1 \int_0^2 x(y+1)(1-y)^3 \, dx\, dy = \frac{3}{5}. \end{aligned} \tag{18}$$

Adding the results for the five parts of $\mathcal{S}$,

$$\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA = 0 + 0 + 0 - \frac{1}{2} + \frac{3}{5} = \frac{1}{10},$$

which does indeed agree with (14).

COMMENT 1. Surely $\mathbf{v}$ is C^1 in the closed region $\mathcal{V}$, as required by the divergence theorem since the partial derivatives of $v_x = 0$, $v_y = x(y+1)z^3$, $v_z = 0$ with respect to x, y, z are continuous functions of x, y, z in $\mathcal{V}$.

COMMENT 2. Observe that the formula (17) converts the surface integration on the slanted plane to an equivalent integration over its "shadow" (the rectangle $PHGO$) in the x, y plane; hence the limits $\int_0^1 \int_0^2$ in (18). Also, remember to express any z's in the integrand in terms of x, y according to $z = f(x, y) = 1 - y$. ∎

Our cross product method of deriving $\hat{\mathbf{n}}$, in Example 1, works only if the surface is planar. A more general method, for deriving normals to surfaces, is as follows. Suppose the surface is given by the relation

$$g(x, y, z) = c, \tag{19}$$

where c is a constant. If we *imagine* $g(x, y, z)$ to be a scalar field defined in 3-space, say a temperature field for definiteness, then the particular surface $\mathcal{S}$ defined by (19) is an isothermal surface, and ∇g will be normal to that surface. Thus, if we normalize ∇g, then we have the unit normal vector

$$\boxed{\hat{\mathbf{n}} = \pm \frac{\nabla g}{\|\nabla g\|},} \tag{20}$$

where the correct sign is to be chosen so that $\hat{\mathbf{n}}$ is outward, not inward. For instance, in Example 1 the equation of the slanted face is $y + z = 1$, so $g(x, y, z) = y + z$ and

$$\hat{\mathbf{n}} = \pm \frac{\nabla(y+z)}{\|\nabla(y+z)\|} = \pm \frac{\hat{\mathbf{j}} + \hat{\mathbf{k}}}{\sqrt{2}}. \tag{21}$$

Clearly, the plus sign gives the *out*ward unit normal, as desired, and then (21) agrees with (16), which was found by other means. We will call (20) the **gradient method** for obtaining the field of normals to a surface defined by (19).

Consider two important physical applications.

EXAMPLE 2. *Continuity Equation of Fluid Mechanics.* Consider a fluid flow within some region $\mathcal{R}$ in 3-space; let $\mathbf{v}(x, y, z, t)$ and $\sigma(x, y, z, t)$ be the velocity and density (i.e., the mass per unit volume) fields, respectively.* Let $\mathcal{V}$ be a stationary control volume of arbitrary shape (although we ask that its surface be piecewise smooth and orientable), located within $\mathcal{R}$. The mass within $\mathcal{V}$ at any time t is $M = \int_{\mathcal{V}} \sigma \, dV$, so the rate of increase of M is†

$$\frac{dM}{dt} = \frac{d}{dt} \int_{\mathcal{V}} \sigma(x, y, z, t) \, dV = \int_{\mathcal{V}} \frac{\partial \sigma}{\partial t} \, dV. \tag{22}$$

*The letter ρ is usually used for the mass density, but we will reserve ρ for the ρ in ρ, ϕ, θ spherical coordinate systems.

†For the last equality in (22), see equation (12.3) in Exercise 12 of Section 13.8.

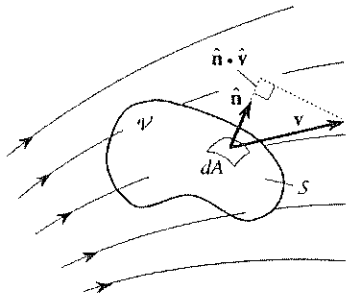


Figure 5. Control volume $\mathcal{V}$.

Next, observe that the rate at which mass enters $\mathcal{V}$ through its boundary $\mathcal{S}$ is the flux integral (Fig. 5)

$$-\int_{\mathcal{S}} \sigma\, \hat{\mathbf{n}} \cdot \mathbf{v}\, dA, \tag{23}$$

where $\hat{\mathbf{n}}$ is the unit outward normal to $\mathcal{S}$. [The latter expression differs in sign from (2) in Section 16.3 because there we considered outflow, whereas here we consider inflow. We could avoid the minus sign by taking $\hat{\mathbf{n}}$ to be the *in*ward normal, but it is tradition, in vector field theory, to work with outward normals.]

Finally, suppose that there is a creation of mass within the field at a rate $f(x, y, z, t)$ mass per unit volume per unit time; where f is positive mass is being created, and where f is negative mass is being destroyed. Then the net rate of creation of mass within $\mathcal{V}$ is

$$\int_{\mathcal{V}} f(x, y, z, t)\, dV \tag{24}$$

mass per unit time.

Since the rate of increase of mass within $\mathcal{V}$ is certainly equal to the rate at which mass enters $\mathcal{V}$ through $\mathcal{S}$, plus the rate at which mass is created, it follows from (22)–(24) that

$$\int_{\mathcal{V}} \frac{\partial \sigma}{\partial t}\, dV = -\int_{\mathcal{S}} \sigma\, \hat{\mathbf{n}} \cdot \mathbf{v}\, dA + \int_{\mathcal{V}} f\, dV. \tag{25}$$

To combine these integrals we convert the surface integral to a volume integral using the divergence theorem (1), with "$\mathbf{v}$" in (1) taken to be $\sigma\mathbf{v}$ in (25):

$$\int_{\mathcal{S}} \sigma\, \hat{\mathbf{n}} \cdot \mathbf{v}\, dA = \int_{\mathcal{S}} \hat{\mathbf{n}} \cdot (\sigma\mathbf{v})\, dA = \int_{\mathcal{V}} \nabla \cdot (\sigma\mathbf{v})\, dV. \tag{26}$$

Thus (25) becomes

$$\int_{\mathcal{V}} \left[\frac{\partial \sigma}{\partial t} + \nabla \cdot (\sigma\mathbf{v}) - f \right] dV = 0. \tag{27}$$

We could accept (27) as our final result, but its integral form is inconvenient. To achieve the final simplification, observe that the control volume $\mathcal{V}$ is arbitrary. If $\int_{\mathcal{V}} [\ ]\, dV = 0$ for *arbitrary* $\mathcal{V}$, within $\mathcal{R}$, and we are willing to assume that the [ ] integrand is continuous, then it can be concluded that $[\ ] = 0$ throughout $\mathcal{R}$. For suppose that [ ] were *non*zero, say positive, at some point P in $\mathcal{R}$. Then by the assumed continuity of [ ] there must be some neighborhood $\mathcal{N}$ of P such that $[\ ] > 0$ throughout $\mathcal{N}$. Since $\mathcal{V}$ is arbitrary, we can take $\mathcal{V}$ to be $\mathcal{N}$. Then $\int_{\mathcal{V}} [\ ]\, dV = \int_{\mathcal{N}} [\ ]\, dV > 0$, which contradicts (27). Thus, [ ] must be identically zero through $\mathcal{R}$: $\partial\sigma/\partial t + \nabla \cdot (\sigma\mathbf{v}) - f = 0$, or

$$\frac{\partial \sigma}{\partial t} + \nabla \cdot (\sigma\mathbf{v}) = f. \tag{28}$$

Thus far, (28) involves logic and bookkeeping, no physics, for who can argue with the statement that the rate of increase of mass within $\mathcal{V}$ equals the rate at which it enters through $\mathcal{S}$ plus the rate at which it is generated within $\mathcal{V}$? Finally, the *physics*: if matter can be neither created nor destroyed, then the mass generation term, f, must be zero, and (28) reduces to the form

$$\boxed{\frac{\partial \sigma}{\partial t} + \nabla \cdot (\sigma\mathbf{v}) = 0,} \tag{29}$$

which is the well known **continuity equation** of fluid mechanics and is a statement of the physical principle of conservation of mass. It is a partial differential equation, or *field equation*, governing the scalar density field σ and the vector velocity field $\mathbf{v}$. If $\mathbf{v} = v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}$, then (29) becomes

$$\frac{\partial \sigma}{\partial t} + \frac{\partial}{\partial x}(\sigma v_x) + \frac{\partial}{\partial y}(\sigma v_y) + \frac{\partial}{\partial z}(\sigma v_z) = 0. \tag{30}$$

The latter is only one equation on the four unknowns σ, v_x, v_y, v_z, but (30) is not the only relevant field equation. For instance, Newton's vector second law of motion leads to another field equation relating σ and $\mathbf{v}$, which is a vector equation or, equivalently, three scalar equations. That step would appear to "close" the system since, together with (30), it would give four equations on the four unknowns. However, it also introduces another unknown field, the pressure field $p(x, y, z, t)$ so we need to add to our system of equations an *equation of state*, relating the pressure, density, and temperature fields. Having thus included the temperature field, say $T(x, y, z, t)$, we need to use the law(s) of thermodynamics to obtain the field equation(s) needed to close the system. The upshot is that the various unknown fields, $\sigma, v_x, v_y, v_z, p, T$, are governed by a system of coupled field equations, most of them being PDE's (partial differential equations).

COMMENT. It is often possible to simplify this complex state of affairs. For instance, if the fluid is water then we might very well be able to neglect its compressibility and, to a good approximation, assume that $\sigma =$ constant. In that case $\partial\sigma/\partial t = 0$, so (29) becomes $0 + \nabla \cdot (\sigma \mathbf{v}) = \sigma \nabla \cdot \mathbf{v} = 0$ or, equivalently,

$$\nabla \cdot \mathbf{v} = 0. \tag{31}$$

Then, even though Newton's vector equation of motion introduces the unknown pressure field p, those three equations plus (31) comprise a closed system of four PDE's governing the four unknowns v_x, v_y, v_z, and p. We can still solve the relevant field equation for the temperature field T if we wish, but the point is that if $\sigma =$ constant then that equation stands alone, and is uncoupled from the PDE's governing the velocity and pressure fields.

How do we know whether $\sigma =$ constant is a sufficiently accurate approximation? There is always a relevant nondimensional parameter that tells us whether or not one effect is negligible compared to others. The nondimensional parameter that provides a measure of the effects of compressibility is the *Mach number* M, defined as the ratio of the fluid velocity to the local speed of sound. If $M \ll 1$ we can, with good accuracy, make the simplifying assumption that $\sigma =$ constant, that is, that the flow is incompressible. For instance, since the speed of sound in water is around 1500 m/s, the Mach number within a water wave is, no doubt, tiny compared to unity, so in developing the field equations governing water waves one would surely take the water to be incompressible. ∎

EXAMPLE 3. *Unsteady Heat Conduction.* Consider the unsteady conduction of heat within some region $\mathcal{R}$ in 3-space. Our derivation of a partial differential equation governing the temperature field $T(x, y, z, t)$ will closely parallel our derivation of the continuity equation in the preceding example.

First, we choose a fixed arbitrary control volume $\mathcal{V}$ within $\mathcal{R}$. Instead of a mass balance, this time we carry out a heat balance. That is, we equate the rate at which heat

accumulates in $\mathcal{V}$ to the rate at which it enters through the surface $\mathcal{S}$ of $\mathcal{V}$ plus the rate at which it is generated within $\mathcal{V}$.

Since the heat (in calories, say) contained in a mass m at (absolute) temperature T is mcT, where c is the *specific heat* of the material, the rate of accumulation of heat within $\mathcal{V}$ is

$$\frac{d}{dt}\int_{\mathcal{V}} cT\sigma\, dV \qquad \text{or} \qquad \int_{\mathcal{V}} c\sigma \frac{\partial T}{\partial t}\, dV. \tag{32}$$

As in Example 2, σ is the mass density (mass per unit volume) of the material and dV is an infinitesimal volume element. Thus, $\sigma\, dV$ in (32) is $d(\text{mass})$.

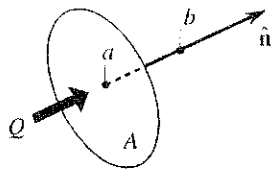


Figure 6. Fourier's law.

To compute the heat flux into $\mathcal{V}$ through $\mathcal{S}$ we need the **Fourier law of heat conduction**, which states that the heat flux Q (calories per second, say) through an area element A is proportional to A and to the derivative of the temperature field normal to the element A (Fig. 6):

$$Q = -kA\frac{\partial T}{\partial n}, \tag{33}$$

where the constant of proportionality, k, is called the thermal *conductivity* of the given material: for instance, k for copper is much larger than k for wood. Why the minus sign? If Q really is positive (i.e., in the $+\hat{\mathbf{n}}$ direction), then $T(a) > T(b)$ (since heat flows from hot to cold) and $\partial T/\partial n$ will be negative.

According to the Fourier law (33), the heat flux into $\mathcal{V}$ through an area element dA on the surface $\mathcal{S}$ of $\mathcal{V}$ is $k\,\frac{\partial T}{\partial n}\,dA$, where there is no minus sign this time because $\hat{\mathbf{n}}$ is the outward unit normal whereas we seek the inward heat flux. Thus, the heat flux into $\mathcal{V}$, through $\mathcal{S}$, is

$$\int_{\mathcal{S}} k\frac{\partial T}{\partial n}\, dA. \tag{34}$$

Further, if heat is being generated within $\mathcal{V}$ at the rate $f(x, y, z, t)$ calories per unit volume per unit time, then the net rate of generation of heat within $\mathcal{V}$ is

$$\int_{\mathcal{V}} f(x, y, z, t)\, dV. \tag{35}$$

And since the rate of increase of heat within $\mathcal{V}$ equals the rate of heat entering through $\mathcal{S}$ plus the rate of generation within, we have

$$\int_{\mathcal{V}} c\sigma\frac{\partial T}{\partial t}\, dV = \int_{\mathcal{S}} k\frac{\partial T}{\partial n}\, dA + \int_{\mathcal{V}} f\, dV. \tag{36}$$

To convert the surface integral in (36) to a volume integral using the divergence theorem $\int_{\mathcal{S}} \hat{\mathbf{n}}\cdot\mathbf{v}\, dV = \int_{\mathcal{V}} \nabla\cdot\mathbf{v}\, dV$, we need to identify the vector field "$\mathbf{v}$" such that $k\,\partial T/\partial n = \hat{\mathbf{n}}\cdot\mathbf{v}$. From the directional derivative formula (7) in Section 16.4, with $\hat{\mathbf{s}} = \hat{\mathbf{n}}$, $\partial T/\partial n = \nabla T\cdot\hat{\mathbf{n}}$, so $k\,\partial T/\partial n = k\,\nabla T\cdot\hat{\mathbf{n}}$. Thus "$\mathbf{v}$" $= k\,\nabla T$. Then the divergence theorem gives $\int_{\mathcal{S}} k\,(\partial T/\partial n)\, dA = \int_{\mathcal{S}} k\,\nabla T\cdot\hat{\mathbf{n}}\, dA = \int_{\mathcal{V}} \nabla\cdot(k\,\nabla T)\, dV$ so (36) becomes

$$\int_{\mathcal{V}} \left[c\sigma\frac{\partial T}{\partial t} - \nabla\cdot(k\,\nabla T) - f\right] dV = 0. \tag{37}$$

From the arbitrariness of $\mathcal{V}$ there follows (as discussed in Example 2) the PDE, or field equation,

$$c\sigma\frac{\partial T}{\partial t} - \nabla\cdot(k\,\nabla T) - f = 0 \tag{38}$$

governing the temperature field $T(x, y, z, t)$. If the conductivity k is a constant, then $\nabla \cdot (k\,\nabla T) = k\nabla \cdot \nabla T = k\nabla^2 T$, and (38) may be written as

$$\alpha^2 \nabla^2 T - \frac{\partial T}{\partial t} = -F, \tag{39}$$

where $F = f(x, y, z, t)/(c\sigma)$, and $\alpha^2 = k/(c\sigma)$ is called the *diffusivity* of the material. The $-F$ term renders the PDE (39) nonhomogeneous and serves as a forcing function just as the $F_0 \sin \Omega t$ term in the ODE $mx'' + kx = F_0 \sin \Omega t$ makes that equation nonhomogeneous and serves as a forcing function. If $F(x, y, z, t) \equiv 0$, one usually writes (39) in the form

$$\boxed{\alpha^2 \nabla^2 T = \frac{\partial T}{\partial t},} \tag{40}$$

which is called the **heat equation** governing the unsteady diffusion of heat by conduction, in a material having diffusivity α^2. If a steady state is achieved, then $\partial T/\partial t = 0$ and (40) reduces to the **Laplace equation**

$$\boxed{\nabla^2 T = 0.} \tag{41}$$

COMMENT 1. In applications, the source term F in (39) may be nonzero. For example, there might be taking place within the region a chemical reaction that releases heat (exothermic) or absorbs heat (endothermic). Or, a mechanical member (such as a beam) subjected to a cyclic loading might be heated by mechanical hysteresis losses proportional to the local stress level.

COMMENT 2. The conduction of heat, in a medium, is an example of the physical process of diffusion, so (40) is also known as the **diffusion equation**. Another example is the diffusion of material, such as the spreading out by diffusion of a drop of dye place in a pot of water. Let $c(x, y, z, t)$ denote the concentration of the material (i.e., the mass of material per unit volume of medium). Analogous to the Fourier law (33) is **Fick's law**, that the material mass flux through an area element A is proportional to A and to the derivative $\partial c/\partial n$ normal to A. Consequently, just like the temperature field T, the concentration field c satisfies a diffusion equation

$$\beta^2 \nabla^2 c = \frac{\partial c}{\partial t}, \tag{42}$$

where β^2 is a mass diffisivity constant. Arguably, the diffusion equation and the Laplace equation are two of the three most prominent PDE's of mathematical physics, the third being the wave equation, which we will study in Chapter 20. ∎

EXAMPLE 4. *Green's Identities.* If u and v are scalar fields, then the combination $u\nabla v$ is a vector field. Applying the divergence theorem to that field gives

$$\int_{\mathcal{V}} \nabla \cdot (u\nabla v)\, dV = \int_{\mathcal{S}} \hat{\mathbf{n}} \cdot (u\nabla v)\, dA. \tag{43}$$

But from equation (4) in Section 16.6, with ∇v as "$\mathbf{v}$," we have $\nabla \cdot (u\nabla v) = \nabla u \cdot \nabla v + u\nabla \cdot \nabla v = \nabla u \cdot \nabla v + u\nabla^2 v$, and $\hat{\mathbf{n}} \cdot (u\nabla v) = u\,\hat{\mathbf{n}} \cdot \nabla v = u\,\partial v/\partial n$, so

$$\boxed{\int_{\mathcal{V}} (\nabla u \cdot \nabla v + u\nabla^2 v)\, dV = \int_{\mathcal{S}} u\,\frac{\partial v}{\partial n}\, dA.} \tag{44}$$

This formula, known as **Green's first identity**,* is important in the study of partial differential equations.

In some cases, evaluation of the $\partial v/\partial n$ term in (44) is simple. For example, if V is the cube $|x| < 1$, $|y| < 1$, $|z| < 1$, then $\partial v/\partial n$ on the $x = +1$ face is merely $\partial v/\partial x$, and on the $x = -1$ face it is $-\partial v/\partial x$. For more complicated surface shapes we suggest that you use the expression $\partial v/\partial n = \nabla v \cdot \hat{\mathbf{n}}$ to calculate $\partial v/\partial n$ on $\mathcal{S}$.

Now, interchanging the letters u and v in (44), it should be equally true that

$$\int_{\mathcal{V}} (\nabla v \cdot \nabla u + v\nabla^2 u)\, dV = \int_{\mathcal{S}} v \frac{\partial u}{\partial n}\, dA,$$

and subtracting these last two equations yields **Green's second identity**

$$\boxed{\int_{\mathcal{V}} (u\nabla^2 v - v\nabla^2 u)\, dV = \int_{\mathcal{S}} \left(u\frac{\partial v}{\partial n} - v\frac{\partial u}{\partial n} \right) dA.} \tag{45}$$

What conditions on u, v assure the validity of Green's identities? Recall that the three components of $\mathbf{v}$ in the divergence theorem (1) were to have continuous first-order partials with respect to x, y, z. Since in deriving (44) and (45) we first set "$\mathbf{v}$" $= u\nabla v = uv_x\hat{\mathbf{i}} + uv_y\hat{\mathbf{j}} + uv_z\hat{\mathbf{k}}$, and then "$\mathbf{v}$" $= v\nabla u = vu_x\hat{\mathbf{i}} + vu_y\hat{\mathbf{j}} + vu_z\hat{\mathbf{k}}$, we must ask uv_x, uv_y, uv_z, vu_x, vu_y, and vu_z to have continuous first-order partial derivatives in the closed region $\mathcal{V}$. Equivalently, we ask u and v to be C^2 in the closed region $\mathcal{V}$. Similarly, $\mathcal{S}$ is to be a piecewise smooth orientable surface, as in the divergence theorem. ∎

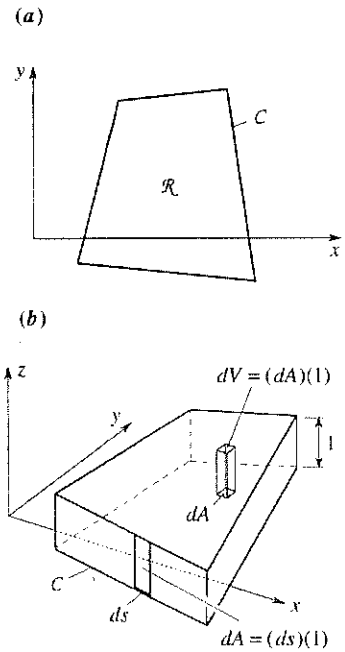


Figure 7. Plane case.

16.8.2. Two-dimensional case. Suppose $\mathbf{v}$ is but a *two*-dimensional field, $\mathbf{v} = v_x(x,y)\hat{\mathbf{i}} + v_y(x,y)\hat{\mathbf{j}}$, defined over a closed region $\mathcal{R}$ in the x, y plane (Fig. 7a). To have a three-dimensional region, so we can apply the divergence theorem (1), let us build up a volume $\mathcal{V}$ with $\mathcal{R}$ as its base, and of unit thickness in the z direction, as in Fig. 7b. In the divergence theorem

$$\begin{aligned}\int_{\mathcal{V}} \nabla \cdot \mathbf{v}\, dV &= \int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA \\ &= \int_{\text{top}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA + \int_{\text{bottom}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA + \int_{\text{side}} \hat{\mathbf{n}} \cdot \mathbf{v}\, dA,\end{aligned} \tag{46}$$

$\mathcal{S}$ is comprised of a top ($z = 1$) on which $\hat{\mathbf{n}} = +\hat{\mathbf{k}}$, a bottom ($z = 0$) on which $\hat{\mathbf{n}} = -\hat{\mathbf{k}}$, and a side on which $\hat{\mathbf{n}}$ is parallel to the x, y plane. Since $\hat{\mathbf{n}} \cdot \mathbf{v} = 0$ on the top and bottom, the "top" and "bottom" integrals in (46) are zero. Further, div $\mathbf{v}$ in (46) does not vary with z, so we can let dV extend from bottom to top; then $dV = (dA)(1)$, where dA is the base area of dV (Fig. 7b). Similarly, $\hat{\mathbf{n}} \cdot \mathbf{v}$ on the side face does not vary with z, so we can let dA in the side integral extend from

*George Green (1793–1841) was a self-taught English mathematical physicist. His work on electricity and magnetism helped to place field theory on a firm mathematical foundation. His "Green's identities," "Green's theorem," and "Green's functions" are important tools in the study of field theory and partial differential equations.

bottom to top (shaded in Fig. 7b); then $dA = (ds)(1)$, where ds is the arc length along $\mathcal{C}$. With these replacements, (46) becomes

$$\boxed{\int_{\mathcal{R}} \nabla \cdot \mathbf{v}\, dA = \int_{\mathcal{C}} \hat{\mathbf{n}} \cdot \mathbf{v}\, ds,} \tag{47}$$

which we call the **two-dimensional divergence theorem**. That is, we simply drop back by one dimension: $\mathcal{V} \to \mathcal{R}$, $dV \to dA$, $\mathcal{S} \to \mathcal{C}$, and $dA \to ds$, The integral on the right-hand side of (47) is a line integral. Line integrals were defined in Section 15.2.3.

EXAMPLE 5. Let us illustrate (47) with a simple example. Namely, verify (47) for the case where $\mathbf{v} = x^2 y\hat{\mathbf{i}} - 3\hat{\mathbf{j}}$ and $\mathcal{R}$ is the rectangle shown in Fig. 8. Then

$$\int_{\mathcal{R}} \nabla \cdot \mathbf{v}\, dA = \int_0^2 \int_0^3 2xy\, dx\, dy = 18,$$

and

$$\begin{aligned}\int_{\mathcal{C}} \hat{\mathbf{n}} \cdot \mathbf{v}\, ds &= \int_0^3 (-\hat{\mathbf{j}}) \cdot \mathbf{v}\Big|_{y=0} dx + \int_0^2 \hat{\mathbf{i}} \cdot \mathbf{v}\Big|_{x=3} dy \\ &\quad + \int_0^3 \hat{\mathbf{j}} \cdot \mathbf{v}\Big|_{y=2} dx + \int_0^2 (-\hat{\mathbf{i}}) \cdot \mathbf{v}\Big|_{x=0} dy \\ &= \int_0^3 3\, dx + \int_0^2 9y\, dy - \int_0^3 3\, dx + \int_0^2 0\, dy = 18,\end{aligned}$$

which results are indeed the same.

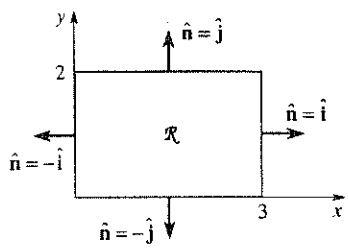


Figure 8. The region $\mathcal{R}$.

COMMENT. In the integral $\int_0^2 (-\hat{\mathbf{i}}) \cdot \mathbf{v}\Big|_{x=0} dy$, say, how do we know that the limits are 0 to 2, not 2 to 0? The idea is that the ds in (47) came from "$(ds)(1)$," namely, the area element dA, and since area elements are necessarily positive, we need to choose integration limits so that the ds increments are all positive. Hence we let y go from 0 to 2, not from 2 to 0. Similarly for the integrals on the edges $x = 3$, $y = 0$, and $y = 2$. ∎

16.8.3. Non-Cartesian coordinates. (Optional) Let us work two examples involving cylindrical and spherical coordinates.

EXAMPLE 6. *Cylindrical Coordinates.* Let $\mathbf{v} = r^2 z\hat{\mathbf{e}}_z$, and let $\mathcal{V}$ be the cone shown in Fig. 9. The problem that we pose is to verify the divergence theorem, for that case, by working out the left- and right-hand sides of (1) and verifying that they are equal. Insofar as $\mathcal{V}$ is concerned, cylindrical and spherical coordinates are equally convenient: in cylindrical coordinates the flat top is a constant-coordinate surface ($z = h$) but the conical part is not, and in spherical coordinates the conical part is a constant-coordinate surface ($\phi = \alpha$) but the flat top is not. However, $\mathbf{v}$ is given in cylindrical coordinates, so let us use cylindrical coordinates r, θ, z.

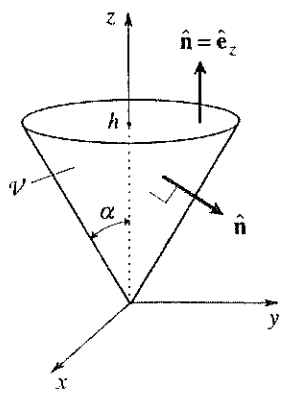


Figure 9. The cone $\mathcal{V}$.

Equation (16) in Section 16.7 gives

$$\nabla \cdot \mathbf{v} = \frac{1}{r}\frac{\partial}{\partial r}[r(0)] + \frac{1}{r}\frac{\partial}{\partial \theta}(0) + \frac{\partial}{\partial z}(r^2 z) = r^2,$$

and the volume element is $dV = r\,dr\,d\theta\,dz$, so

$$\int_{\mathcal{V}} \nabla \cdot \mathbf{v}\,dV = \int_0^h \int_0^{2\pi} \int_0^{z\tan\alpha} (r^2)\,r\,dr\,d\theta\,dz = \frac{\pi}{10}h^5 \tan^4\alpha, \tag{48}$$

where the r limits follow from the fact that the equation of the conical surface is $r = (\tan\alpha)z$.

To evaluate the right-hand side of (1) we break the integral into an integral over the flat top plus an integral on the conical surface. On the top we have $\hat{\mathbf{n}} = \hat{\mathbf{e}}_z$ and $dA = r\,dr\,d\theta$, so*

$$\begin{aligned}\int_{\text{top}} \hat{\mathbf{n}} \cdot \mathbf{v}\,dA &= \int_0^{2\pi} \int_0^{h\tan\alpha} (r^2 z)\Big|_{z=h} r\,dr\,d\theta \\ &= h\int_0^{2\pi} \int_0^{h\tan\alpha} r^3\,dr\,d\theta = \frac{\pi}{2}h^5 \tan^4\alpha.\end{aligned} \tag{49}$$

To determine $\hat{\mathbf{n}}$ on the conical surface we use the gradient method explained in connection with equations (19) and (20). That is, since the surface is defined by $r = (\tan\alpha)z$, we have

$$\begin{aligned} g(r,\theta,z) &= r - (\tan\alpha)z = 0, \\ \mathbf{n} &= \pm\nabla g = \pm\left(\hat{\mathbf{e}}_r\frac{\partial}{\partial r} + \hat{\mathbf{e}}_\theta\frac{1}{r}\frac{\partial}{\partial\theta} + \hat{\mathbf{e}}_z\frac{\partial}{\partial z}\right)[r - (\tan\alpha)z] \\ &= \pm\left[\hat{\mathbf{e}}_r - (\tan\alpha)\hat{\mathbf{e}}_z\right], \\ \|\mathbf{n}\| &= \sqrt{1+\tan^2\alpha} = \frac{1}{\cos\alpha}, \\ \hat{\mathbf{n}} &= +\frac{\hat{\mathbf{e}}_r - (\tan\alpha)\hat{\mathbf{e}}_z}{1/\cos\alpha} = (\cos\alpha)\hat{\mathbf{e}}_r - (\sin\alpha)\hat{\mathbf{e}}_z, \end{aligned} \tag{50}$$

where the plus sign was selected since it gives the *out*ward normal. Next, we need the area element dA on the conical surface. For that we rely on the general formula (5) in Section 15.5, $dA = \sqrt{EG - F^2}\,du\,dv$. To use that formula we need parametric equations $x = x(u,v)$, $y = y(u,v)$, $z = z(u,v)$ for the surface, which can be obtained by beginning with the equations $x = r\cos\theta$, $y = r\sin\theta$, $z = z$ relating x, y, z to r, θ, z, and adding the fact that the cone is given by $z = (\cot\alpha)r$:

$$\begin{aligned} x &= r\cos\theta = u\cos v, \\ y &= r\sin\theta = u\sin v, \\ z &= (\cot\alpha)r = (\cot\alpha)u. \end{aligned} \tag{51}$$

That is, let r and θ be u and v, respectively.

$$\begin{aligned} E &= x_u^2 + y_u^2 + z_u^2 = \cos^2 v + \sin^2 v + \cot^2\alpha = \sec^2\alpha, \\ F &= x_u x_v + y_u y_v + z_u z_v = (\cos v)(-u\sin v) + (\sin v)(u\cos v) = 0, \\ G &= x_v^2 + y_v^2 + z_v^2 = u^2\sin^2 v + u^2\cos^2 v = u^2 \end{aligned}$$

*If you don't see why $dA = r\,dr\,d\theta$, see (17) in Section 15.6.

so

$$dA = \sqrt{EG - F^2}\, du\, dv = (\operatorname{cosec}\alpha)\, u\, du\, dv. \tag{52}$$

Thus, remembering that $r = u$ and $z = (\cot\alpha)\, u$ on the cone,

$$\begin{aligned}\int_{\text{cone}} \hat{\mathbf{n}}\cdot\mathbf{v}\, dA &= \int_{\text{cone}} [(\cos\alpha)\hat{\mathbf{e}}_r - (\sin\alpha)\hat{\mathbf{e}}_z]\cdot(r^2 z\hat{\mathbf{e}}_z)(\operatorname{cosec}\alpha)u\, du\, dv \\ &= -\int_{\text{cone}} r^2 z u\, du\, dv \\ &= -\int_0^{2\pi}\int_0^{h\tan\alpha} (\cot\alpha)u^4\, du\, dv \\ &= -\frac{2\pi}{5}\, h^5 \tan^4\alpha. \end{aligned} \tag{53}$$

Adding (49) and (53) we obtain

$$\int_{\mathcal{S}} \hat{\mathbf{n}}\cdot\mathbf{v}\, dA = \frac{\pi}{10}\, h^5 \tan^4\alpha,$$

which does agree with (48).

COMMENT 1. Though it is clear by inspection that $\hat{\mathbf{n}} = \hat{\mathbf{e}}_z$ on the flat top, observe that the gradient method could have been used instead, for the equation of that surface is $z = h$, so $g(r,\theta,z) = z$ and

$$\hat{\mathbf{n}} = \pm\frac{\nabla g}{\|\nabla g\|} = \pm\hat{\mathbf{e}}_z \to +\hat{\mathbf{e}}_z.$$

COMMENT 2. To check results, we urge you to consider special cases. For instance, to check the result $\hat{\mathbf{n}} = (\cos\alpha)\hat{\mathbf{e}}_r - (\sin\alpha)\hat{\mathbf{e}}_z$, derived in (50), observe that for the special cases $\alpha = 0$ and $\alpha = \pi/2$ we have $\hat{\mathbf{n}} = \hat{\mathbf{e}}_r$ and $\hat{\mathbf{n}} = -\hat{\mathbf{e}}_z$, respectively, which results are obviously correct.

COMMENT 3. To see that $\mathbf{v}$ is C^1 in $\mathcal{V}$, as called for by the divergence theorem, express $\mathbf{v} = r^2 z\hat{\mathbf{e}}_z = (x^2+y^2)z\hat{\mathbf{k}}$, and observe that the partial derivatives of $v_x = 0$, $v_y = 0$, and $v_z = (x^2+y^2)z$ with respect to x, y, z are continuous functions of x, y, z in $\mathcal{V}$. ∎

EXAMPLE 7. *Gauss's Law.* An electrical charge q (which could be positive or negative) at the origin gives rise to an electric field $\mathbf{E}$ that is expressed most conveniently in terms of spherical coordinates as

$$\mathbf{E} = \frac{q}{4\pi\epsilon}\,\frac{1}{\rho^2}\,\hat{\mathbf{e}}_\rho, \tag{54}$$

where ϵ is a physical constant known as the electric permittivity of the medium. That field is radially directed and spherically symmetric, with a magnitude that tends to zero as $\rho \to \infty$ and to infinity as $\rho \to 0$ (since it is proportional to $1/\rho^2$), as hinted at in Fig. 10. A fundamental result in electrostatics, known as **Gauss's law**, states that if $\mathcal{S}$ is a smooth closed surface within the medium, then

$$\int_{\mathcal{S}} \hat{\mathbf{n}}\cdot\mathbf{E}\, dA = \begin{cases} 0 & \text{if } \mathcal{S} \text{ does not enclose the charge} \\ \dfrac{q}{\epsilon} & \text{if } \mathcal{S} \text{ does enclose the charge.} \end{cases} \tag{55}$$

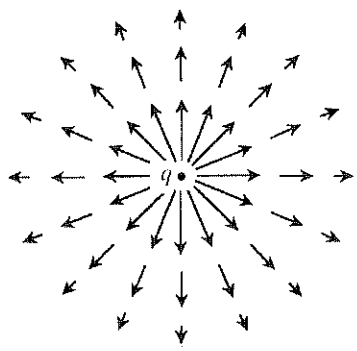


Figure 10. Charge-induced $\mathbf{E}$ field.

Our purpose, in this example, is to prove Gauss's law.

First, suppose that $\mathcal{S}$ does not enclose the charge at the origin. Then the divergence theorem gives

$$\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{E}\, dA = \int_{\mathcal{V}} \nabla \cdot \mathbf{E}\, dV = \int_{\mathcal{V}} 0\, dV = 0, \tag{56}$$

where $\mathcal{V}$ is the volume enclosed by $\mathcal{S}$, because [according to (32) in Section 16.7, with "$\mathbf{v}$" $= \mathbf{E}$, "v_ρ" $= q/(4\pi\epsilon\rho^2)$, and "v_ϕ" $=$ "v_θ" $= 0$]

$$\nabla \cdot \mathbf{E} = \frac{1}{\rho^2}\frac{\partial}{\partial \rho}\left(\frac{q}{4\pi\epsilon}\right) + 0 + 0 = 0. \tag{57}$$

However, we cannot claim that (56) holds if $\mathcal{S}$ encloses the origin because the $\mathbf{E}$ field fails to satisfy the conditions of the divergence theorem. Specifically, $\mathbf{E}$ is undefined at the origin because the $1/\rho^2$ is $1/0$ there. Similarly for the first-order partial derivatives of the x, y, z components of

$$\mathbf{E} = \frac{q}{4\pi\epsilon}\frac{1}{\rho^2}\hat{\mathbf{e}}_\rho = \frac{q}{4\pi\epsilon}\frac{1}{x^2+y^2+z^2}\frac{x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}}{\sqrt{x^2+y^2+z^2}},$$

with respect to x, y, z, so $\mathbf{E}$ is not C^1 in $\mathcal{V}$, as required by the divergence theorem.

To overcome this difficulty let us cut the singular point out, so we can stay away from it. That is, let us apply the divergence theorem not to $\mathcal{V}$, but to $\mathcal{V}'$, say, which is the same as $\mathcal{V}$ but with a sphere of radius ρ_0 cut out at the origin (Fig. 11). Then the surface of $\mathcal{V}'$ consists of the two parts $\mathcal{S}$ and $\mathcal{S}_0$; note that the outward (i.e., out of $\mathcal{V}'$) unit normal $\hat{\mathbf{n}}$ on $\mathcal{S}_0$ is $-\hat{\mathbf{e}}_\rho$. Now, the $\mathbf{E}$ field *is* C^1 within $\mathcal{V}'$ (because the origin has been cut out), so the divergence theorem gives

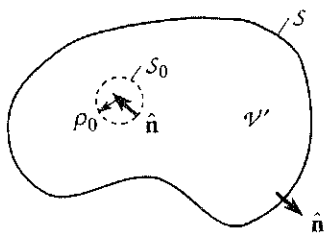


Figure 11. The region $\mathcal{V}'$.

$$\int_{\mathcal{V}'} \nabla \cdot \mathbf{E}\, dV = \int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{E}\, dA + \int_{\mathcal{S}_0} \hat{\mathbf{n}} \cdot \mathbf{E}\, dA \tag{58}$$

or, since $\nabla \cdot \mathbf{E} = 0$ within $\mathcal{V}'$,

$$\begin{aligned}\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{E}\, dA &= -\int_{\mathcal{S}_0} \hat{\mathbf{n}} \cdot \mathbf{E}\, dA \\ &= -\int_{\mathcal{S}_0} -\hat{\mathbf{e}}_\rho \cdot \frac{q}{4\pi\epsilon}\frac{1}{\rho^2}\hat{\mathbf{e}}_\rho\bigg|_{\rho=\rho_0} dA \\ &= \frac{q}{4\pi\epsilon\rho_0^2}\int_{\mathcal{S}_0} dA = \frac{q}{4\pi\epsilon\rho_0^2}4\pi\rho_0^2 = \frac{q}{\epsilon},\end{aligned}$$

as was to be shown. ∎

Closure. The divergence theorem, Theorem 16.8.1, tells us that an integral over a volume $\mathcal{V}$ can be reduced to an integration over the boundary of $\mathcal{V}$, its surface $\mathcal{S}$, provided that its integrand is a divergence (and provided that the surface $\mathcal{S}$ and the vector field $\mathbf{v}$ are sufficiently "decent"). We note that (1) is actually a generalization of the familiar result, from the integral calculus, that the integral $\int_a^b F(x)\, dx$ can

be reduced to an evaluation at the boundary of the x interval, its endpoints $x = a$ and $x = b$, provided that $f(x)$ is a derivative, for then

$$\int_a^b F'(x)\,dx = F(x)\Big|_{x=a}^{x=b}. \tag{59}$$

Whereas (59) reduces the calculation from one dimension (the line segment on the x axis) to none (the two endpoints), the divergence theorem (1) reduces the calculation from three dimensions (the region $\mathcal{V}$) to two (its surface $\mathcal{S}$). Of course we don't have to use (1) in that direction. In Example 2 and 3, for example, we used it to convert surface integrals to volume integrals.

Our examples are of two types. One type involves the verification of (1) for a particular case, that is, for a particular field $\mathbf{v}$ and region $\mathcal{V}$, the purpose being to foster understanding of the (1) through having to face up to obtaining expressions for $\nabla \cdot \mathbf{v}$, dV, $\hat{\mathbf{n}}$, and dA, and having to evaluate the two integrals. The other type involves the derivation of important results in fluid mechanics, heat conduction, and electrostatics. We will follow that same pattern in the next section as well.

In Examples 2 and 3 we derive two of the most important partial differential equations of mathematical physics, the heat equation and the Laplace equation. In Chapters 18–20 we return to these PDE's and learn how to solve them. Thus, Chapters 13–16 lay much of the groundwork for our study of PDE's in Chapters 18–20.

EXERCISES 16.8

1. In each case, verify the divergence theorem by working out $\int_{\mathcal{V}} \nabla \cdot \mathbf{v}\,dV$ and $\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{v}\,dA$ and showing that the results are equal.

(a) $\mathbf{v} = 2\hat{\mathbf{i}} - \hat{\mathbf{j}} + 4\hat{\mathbf{k}}$, $\mathcal{V}$: the rectangular prism $0 \le x \le 1$, $0 \le y \le 3$, $0 \le z \le 2$
(b) $\mathbf{v} = x\hat{\mathbf{i}} + 2y\hat{\mathbf{j}}$, $\mathcal{V}$: the cube $|x| \le 1$, $|y| \le 1$, $|z| \le 1$
(c) $\mathbf{v} = y\hat{\mathbf{i}} - 2\hat{\mathbf{j}} + x\hat{\mathbf{k}}$, $\mathcal{V}$: the rectangular prism $|x| \le 2$, $0 \le y \le 1$, $0 \le z \le 1$
(d) $\mathbf{v} = \hat{\mathbf{j}} + x^2 z\hat{\mathbf{k}}$, $\mathcal{V}$: the cube $0 \le x \le 1$, $0 \le y \le 1$, $0 \le z \le 1$
(e) $\mathbf{v} = 3xy^2\hat{\mathbf{j}}$, $\mathcal{V}$: the rectangular prism $0 \le x \le 1$, $0 \le y \le 1$, $0 \le z \le 1$
(f) $\mathbf{v} = x^2\hat{\mathbf{i}} - 2z\hat{\mathbf{j}}$, $\mathcal{V}$: the rectangular prism $|x| \le 2$, $|y| \le 2$, $|z| \le 3$, with the cubical cavity $0 < x < 1$, $0 < y < 1$, $0 < z < 1$
(g) $\mathbf{v} = xy\hat{\mathbf{j}}$, $\mathcal{V}$: the pentahedron with vertices at $(0,0,0)$, $(1,0,0)$, $(0,2,0)$, $(0,0,1)$, $(1,0,1)$, $(1,2,0)$
(h) $\mathbf{v} = y^2 z\hat{\mathbf{k}}$, $\mathcal{V}$: the pentahedron with vertices at $(0,0,0)$, $(2,0,0)$, $(0,0,3)$, $(2,0,3)$, $(0,4,3)$, $(2,4,3)$
(i) $\mathbf{v} = x^2 y \sin z\hat{\mathbf{i}}$, $\mathcal{V}$: same as in part (h)
(j) $\mathbf{v} = (3x^2 - 2yz)\hat{\mathbf{j}}$, $\mathcal{V}$: same as in part (h)
(k) $\mathbf{v} = 6\hat{\mathbf{i}} + 2x\hat{\mathbf{j}} + x^2 yz\hat{\mathbf{k}}$, $\mathcal{V}$: the pentahedron with vertices at $(0,0,0)$, $(1,0,0)$, $(0,2,0)$, $(1,2,0)$, $(0,0,5)$, $(0,2,5)$
(l) $\mathbf{v} = z^2\hat{\mathbf{k}}$, $\mathcal{V}$: the tetrahedron bounded by the planes $x = 0$, $y = 0$, $z = 0$, $2x + y + 2z = 2$
(m) $\mathbf{v} = x^2 z\hat{\mathbf{i}}$, $\mathcal{V}$: same as in part (l)
(n) $\mathbf{v} = x^2 z\hat{\mathbf{i}} - 2z(x^2 + 1)\hat{\mathbf{j}} + z\hat{\mathbf{k}}$, $\mathcal{V}$: same as in part (l)
(o) $\mathbf{v} = z^2\hat{\mathbf{k}}$, $\mathcal{V}$: the tetrahedron bounded by the planes $x = 0$, $y = 0$, $z = 1$, $z = x + y$

2. Let $\mathcal{S}$ be a piecewise smooth orientable closed surface enclosing a region of volume V. Show that

(a) $\int_{\mathcal{S}} \hat{\mathbf{n}}\,dA = \mathbf{0}$ HINT: Show that $\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{a}\,dA = 0$ for every constant vector $\mathbf{a}$.
(b) $\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot (x\hat{\mathbf{i}})\,dA = V$
(c) $\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot (x\hat{\mathbf{i}} + y\hat{\mathbf{j}})\,dA = 2V$
(d) $\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot (x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}})\,dA = 3V$

3. Verify Green's first identity (44) in each case.

(a) $u = 2x^3$, $v = xy^2$, $\mathcal{V}$: the rectangular prism $0 \le x \le 2$, $0 \le y \le 3$, $0 \le z \le 1$
(b) $u = 1$, $v = x^2 + y^2 + z^2$, $\mathcal{V}$: the cube $0 \le x \le 1$, $0 \le y \le 1$, $0 \le z \le 1$
(c) $u = 1$, $v = x^2 + y^2 + z^2$, $\mathcal{V}$: the pentahedron with vertices at $(0,0,0)$, $(1,0,0)$, $(0,1,0)$, $(0,0,1)$, $(1,0,1)$, $(1,1,0)$

4. Verify the two-dimensional divergence theorem (47) in each case.

(a) $\mathbf{v} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}}$, $\mathcal{R}$: the rectangle $0 \le x \le a$, $0 \le y \le b$
(b) $\mathbf{v} = x^2 y\hat{\mathbf{i}}$, $\mathcal{R}$: the rectangle $0 \le x \le 2$, $0 \le y \le 1$
(c) $\mathbf{v} = xy^2\hat{\mathbf{i}}$, $\mathcal{R}$: the triangle with vertices at $(0,0)$, $(2,0)$, $(2,1)$
(d) $\mathbf{v} = xy\hat{\mathbf{j}}$, $\mathcal{R}$: the triangle with vertices at $(0,0)$, $(2,0)$, $(0,1)$

5. In each case evaluate $I = \int_0^1 \int_0^1 \int_0^1 f\,dx\,dy\,dz$ directly. Then determine a vector field $\mathbf{v}$ such that $\nabla \cdot \mathbf{v} = f$, convert I to a surface integral by using the divergence theorem, and evaluate the surface integral.

(a) $f = x^2 yz$ (b) $f = xz^2 - 2y$
(c) $f = 4$ (d) $f = x + y^2 z$

6. Derive the following results from results in this section.

(a) $\displaystyle\int_{\mathcal{V}} \nabla^2 u\,dV = \int_{\mathcal{S}} \frac{\partial u}{\partial n}\,dA$

(b) $\displaystyle\int_{\mathcal{V}} \left[\|\nabla u\|^2 + u\nabla^2 u\right] dV = \int_{\mathcal{S}} u\,\frac{\partial u}{\partial n}\,dA$

CAUTION: Most authors write $(\nabla u)^2$ for $\nabla u \cdot \nabla u$, in place of our $\|\nabla u\|^2$.

7. Just as (47) is the two-dimensional version of (1), work out the two-dimensional versions of Green's

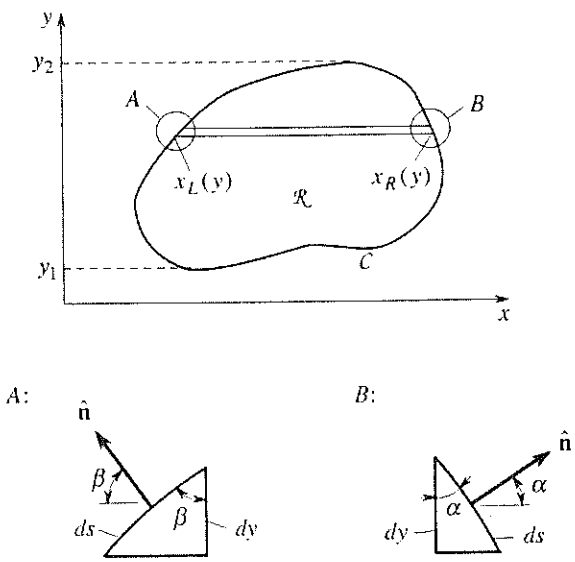
(a) first identity (b) second identity

8. (*Alternative proof of divergence theorem*) An alternative approach to the proof of the divergence theorem (1) is to express the volume integral as an iterated integral, and to carry out one integration. To illustrate the procedure, consider the simpler *two*-dimensional case (47), and suppose that $\mathcal{R}$ is *convex* in the x direction, i.e., each crosshatched sliver running from $x_L(y)$ to $x_R(y)$ (see sketch) lies entirely within $\mathcal{R}$. Now

$$\int_{\mathcal{R}} \nabla \cdot \mathbf{v}\,dA = \int_{\mathcal{R}} \left(\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y}\right) dx\,dy$$
$$= \int_{\mathcal{R}} \frac{\partial v_x}{\partial x}\,dx\,dy + \int_{\mathcal{R}} \frac{\partial v_y}{\partial y}\,dx\,dy.$$

The first term becomes

$$\int_{\mathcal{R}} \frac{\partial v_x}{\partial x}\,dx\,dy = \int_{y_1}^{y_2} \int_{x_L(y)}^{x_R(y)} \frac{\partial v_x}{\partial x}\,dx\,dy$$
$$= \int_{\text{right}} v_x(ds\,\cos\alpha) - \int_{\text{left}} v_x(ds\,\cos\beta)$$
$$= \int_{\text{right}} v_x(\hat{\mathbf{i}} \cdot \hat{\mathbf{n}})\,ds - \int_{\text{left}} v_x(-\hat{\mathbf{i}} \cdot \hat{\mathbf{n}})\,ds$$
$$= \int_{\text{right}} (v_x\hat{\mathbf{i}}) \cdot \hat{\mathbf{n}}\,ds + \int_{\text{left}} (v_x\hat{\mathbf{i}}) \cdot \hat{\mathbf{n}}\,ds$$
$$= \int_{\mathcal{C}} (v_x\hat{\mathbf{i}}) \cdot \hat{\mathbf{n}}\,ds. \tag{8.1}$$



We now state the problem: assuming that $\mathcal{R}$ is convex in the y direction as well, show that

$$\int_{\mathcal{R}} \frac{\partial v_y}{\partial y}\,dx\,dy = \int_{\mathcal{C}} (v_y\hat{\mathbf{j}}) \cdot \hat{\mathbf{n}}\,ds, \tag{8.2}$$

and hence infer (47) from (8.1) and (8.2). NOTE: Extension to nonconvex regions is not difficult but will not be considered.

9. Besides the divergence theorem, $\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \hat{\mathbf{v}}\,dA = \int_{\mathcal{V}} \nabla \cdot \hat{\mathbf{v}}\,dV$, one may derive the "companion" results

$$\boxed{\int_{\mathcal{S}} \hat{\mathbf{n}}\,u\,dA = \int_{\mathcal{V}} \nabla u\,dV} \tag{9.1}$$

and

$$\boxed{\int_S \hat{\mathbf{n}} \times \hat{\mathbf{v}}\, dA = \int_V \nabla \times \hat{\mathbf{v}}\, dV.} \tag{9.2}$$

(a) Derive (9.1). HINT: Write

$$\begin{aligned}\int_S \hat{\mathbf{n}}\, u\, dA &= \int_S [(\hat{\mathbf{n}}\cdot\hat{\mathbf{i}})\hat{\mathbf{i}} + (\hat{\mathbf{n}}\cdot\hat{\mathbf{j}})\hat{\mathbf{j}} + (\hat{\mathbf{n}}\cdot\hat{\mathbf{k}})\hat{\mathbf{k}}]u\, dA \\ &= \hat{\mathbf{i}}\int_S (\hat{\mathbf{n}}\cdot\hat{\mathbf{i}})\, u\, dA + \hat{\mathbf{j}}\int_S (\hat{\mathbf{n}}\cdot\hat{\mathbf{j}})\, u\, dA \\ &\quad + \hat{\mathbf{k}}\int_S (\hat{\mathbf{n}}\cdot\hat{\mathbf{k}})\, u\, dA \\ &= \hat{\mathbf{i}}\int_S \hat{\mathbf{n}}\cdot(u\hat{\mathbf{i}})\, dA + \hat{\mathbf{j}}\int_S \hat{\mathbf{n}}\cdot(u\hat{\mathbf{j}})\, dA \\ &\quad + \hat{\mathbf{k}}\int_S \hat{\mathbf{n}}\cdot(u\hat{\mathbf{k}})\, dA,\end{aligned}$$

apply the divergence theorem to each integral, and combine the results into one integral.

(b) Derive (9.2). HINT: Essentially the same hint as in part (a). Write

$$\begin{aligned}\int_S \hat{\mathbf{n}} \times \hat{\mathbf{v}}\, dA &= \int_S \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ \hat{\mathbf{n}}\cdot\hat{\mathbf{i}} & \hat{\mathbf{n}}\cdot\hat{\mathbf{j}} & \hat{\mathbf{n}}\cdot\hat{\mathbf{k}} \\ v_x & v_y & v_z \end{vmatrix} dA \\ &= \hat{\mathbf{i}}\int_S \hat{\mathbf{n}}\cdot(v_z\hat{\mathbf{j}} - v_y\hat{\mathbf{k}})\, dA + \text{etc.},\end{aligned}$$

apply the divergence theorem to each integral, and so on.

(c) Deduce, from equation (9.1), that $\int_S \hat{\mathbf{n}}\, dA = \mathbf{0}$ for every S. Verify this result, by direct integration, for the case where S is the surface of the rectangular prism $0 \le x \le a$, $0 \le y \le b$, $0 \le z \le c$. NOTE: Just as the global statement $\int_S \hat{\mathbf{n}}\cdot\mathbf{v}\, dA = \int_V \nabla\cdot\mathbf{v}\, dV$ corresponds to the local statement

$$\nabla\cdot\mathbf{v} = \lim_{V\to 0} \frac{\int_S \hat{\mathbf{n}}\cdot\mathbf{v}\, dA}{V}$$

(V = the volume enclosed by S), (9.1) and (9.2)correspond to the local statements

$$\nabla u = \lim_{V\to 0} \frac{\int_S \hat{\mathbf{n}}\, u\, dA}{V} \tag{9.3}$$

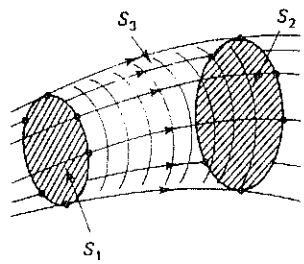
and

$$\nabla\times\mathbf{v} = \lim_{V\to 0} \frac{\int_S \hat{\mathbf{n}}\times\mathbf{v}\, dA,}{V} \tag{9.4}$$

respectively.

10. (*Continuity equation*) Let S_1 be a completely permeable plane surface within a steady fluid flow field. Let the velocity field be $\mathbf{v}(x, y, z)$, and let the mass density field be $\sigma(x, y, z)$. Then the set of streamlines that pass through the set points on the edge of S_1 form a "streamtube," as sketched in the accompanying figure. Let S_2 be another plane cross section of the streamtube an arbitrary distance downstream of S_1, and let the areas of S_1, S_2 be A_1, A_2, respectively. Recalling the continuity equation (29), namely $\nabla\cdot(\sigma\mathbf{v}) = 0$ (since the field is steady, so that $\partial\sigma/\partial t = 0$), integrate this equation over a control volume V which is bounded by S_1, S_2, and the streamtube surface S_3 between S_1 and S_2, apply the divergence theorem, and thus show that

$$\int_{S_1} \hat{\mathbf{n}}\cdot(\sigma\mathbf{v})\, dA + \int_{S_2} \hat{\mathbf{n}}\cdot(\sigma\mathbf{v})\, dA = 0. \tag{10.1}$$



If σ and $\mathbf{v}$ are constant over S_1 and S_2, and S_1, S_2 are normal to the flow, show that (10.1) reduces to the simpler form

$$\sigma_1 A_1 V_1 = \sigma_2 A_2 V_2, \tag{10.2}$$

often given in fluid mechanics texts, where $\|\mathbf{v}\| = V_1$ on S_1 and $\|\mathbf{v}\| = V_2$ on S_2.

11. (*Archimedes's principle*) Consider a solid body immersed in a fluid of uniform mass density σ. Archimedes's principle states that the net pressure force on the body is upward (i.e., it is a buoyant force) and equal to the weight of the fluid displaced by the body. Derive Archimedes's principle using the methods of this section. HINT: If we measure z upwards from the surface of the fluid, then the static pressure distribution in the fluid is $p(x, y, z) = -\sigma g z$. Write the net force $\mathbf{F}$ on the submerged body as an integral over its surface S, and then transform that to a volume integral using (9.1) in Exercise 9.

12. Verify the divergence theorem.

(a) $\mathbf{v} = 3r^2\hat{\mathbf{e}}_r - r\hat{\mathbf{e}}_\theta + 2\hat{\mathbf{e}}_z$, V: the cylinder $r \le 4$, $0 \le \theta < 2\pi$, $0 \le z \le 5$

(b) $\mathbf{v} = z\hat{\mathbf{e}}_z$, V: the cylinder $r \le 2$, $0 \le \theta < 2\pi$, $-3 \le z \le 6$

(c) $\mathbf{v} = rz^2\hat{\mathbf{e}}_\theta$, V: the half-cylinder $r \le 4$, $\pi/2 \le \theta \le 3\pi/2$,

$0 \le z \le 3$

(d) $\mathbf{v} = z^2\hat{\mathbf{e}}_r + 6\hat{\mathbf{e}}_z$, $\mathcal{V}$: the hollow cylinder $2 \le r \le 3$, $0 \le \theta < 2\pi$, $1 \le z \le 5$

(e) $\mathbf{v} = z^2\hat{\mathbf{e}}_z$, $\mathcal{V}$: the cone $r \le 2z$, $0 \le \theta < 2\pi$, $0 \le z \le 3$

(f) $\mathbf{v} = \rho^2\hat{\mathbf{e}}_\rho$, $\mathcal{V}$: the sphere $\rho \le a$

(g) $\mathbf{v} = \rho^3\hat{\mathbf{e}}_\rho$, $\mathcal{V}$: the hemisphere $\rho \le a$, $0 \le \phi < \pi/2$, $0 \le \theta < 2\pi$

(h) $\mathbf{v} = \rho^2 \sin\phi \sin\theta \hat{\mathbf{e}}_\theta$, $\mathcal{V}$: the region $2 \le \rho \le 3$, $0 \le \phi < \pi/2$, $0 \le \theta \le \pi/2$

(i) $\mathbf{v} = \dfrac{1}{\rho}\hat{\mathbf{e}}_\rho + \rho^3\hat{\mathbf{e}}_\phi - \rho^2 \sin^2\phi\hat{\mathbf{e}}_\theta$, $\mathcal{V}$: the hollow sphere $2 \le \rho \le 5$

16.9 Stokes's Theorem

Following Newton's death in 1727 the mathematical scene in the British Isles was relatively quiet for almost 100 years. There then appeared a number of exceptional mathematical physicists, beginning with *George Green* (1793–1841) and *William R. Hamilton* (1805–1865), and followed by the "Cambridge school," which included *Sir G. Gabriel Stokes* (1819–1903), *Lord Kelvin* (Sir William Thomson; 1824–1907), *James Clerk Maxwell* (1831–1879), and *Lord Rayleigh* (John William Strutt; 1842–1919).

In this section we examine two important and closely related theorems, one due to Stokes and one due to Green. Both of these theorems involve line integrals so we begin by reviewing and extending our study of line integrals, which began in Section 15.2.

16.9.1. Line integrals. We met the line integral $\int_C f\,ds$ and showed how to evaluate it by parametrizing the curve C according to $x = x(\tau)$, $y = y(\tau)$, $z = z(\tau)$ for $a \le \tau \le b$. Specifically,

$$\int_C f\,ds = \int_a^b f(x(\tau), y(\tau), z(\tau))\sqrt{\mathbf{R}'(\tau)\cdot\mathbf{R}'(\tau)}\,d\tau, \tag{1}$$

where $\mathbf{R}(\tau) = x(\tau)\hat{\mathbf{i}} + y(\tau)\hat{\mathbf{j}} + z(\tau)\hat{\mathbf{k}}$. That is, by parametrizing we reduce the line integral $\int_C f\,ds$ on a curve C to an "ordinary" integral of the form $\int_a^b F(\tau)\,d\tau$ on a segment of a τ axis. Actually, in engineering science applications it is more common to meet line integrals in a different form, namely, in the form

$$\int_C \mathbf{v}\cdot d\mathbf{R}, \tag{2}$$

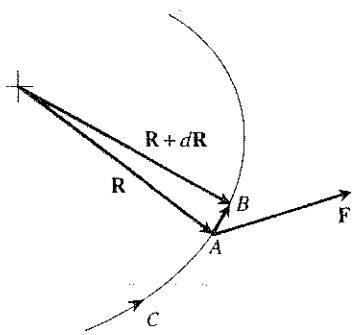


Figure 1. Work.

where $\mathbf{v}$ is a given vector field defined in some region containing C, and $\mathbf{R}$ is the position vector from some reference point, or origin, to points on the curve C. For example, if a particle is subjected to a force $\mathbf{F}$ as it moves along a curve C, then the work dW done by $\mathbf{F}$, as the particle moves an infinitesimal distance along C, say from A to B (Fig. 1) is

$$dW = \mathbf{F}\cdot\mathbf{AB} = \mathbf{F}\cdot d\mathbf{R}. \tag{3}$$

Evidently, then, the net work done in traversing the entire curve C is

$$W = \int_C \mathbf{F} \cdot d\mathbf{R}. \tag{4}$$

For instance, $\mathbf{F}(x, y, z)$ might be a gravitational force field induced by point or distributed systems of mass. Since the curve C, in Fig. 1, is actually traversed by the particle, we sometimes call it a *path* instead of a curve.

That the two forms $\int_C f\,ds$ and $\int_C \mathbf{v} \cdot d\mathbf{R}$ are equivalent can be seen by expressing

$$\int_C \mathbf{v} \cdot d\mathbf{R} = \int_C \mathbf{v} \cdot \frac{d\mathbf{R}}{ds}\,ds = \int_C f\,ds, \tag{5}$$

where f is $\mathbf{v} \cdot d\mathbf{R}/ds$.* However, there is one difference to keep in mind. In $\int_C \mathbf{v} \cdot d\mathbf{R}$ the curve C is said to be **oriented**; that is, we have a specific direction of traversal in mind, and we denote it graphically by an arrowhead, as we did in Fig. 1. In that illustration, observe that the work done in moving from A to B, $dW = \mathbf{F} \cdot \mathbf{AB}$, is the negative of the work done if instead we moved from B to A, $dW = \mathbf{F} \cdot \mathbf{BA}$, because $\mathbf{BA} = -\mathbf{AB}$. Thus, if we denote an oriented curve as C, and the same curve but oppositely oriented as "$-C$," then

$$\int_{-C} \mathbf{v} \cdot d\mathbf{R} = -\int_C \mathbf{v} \cdot d\mathbf{R}. \tag{6}$$

In contrast, the curve C in the form $\int_C f\,ds$ is *not* oriented because ds is arc length, which is always positive. That claim is consistent with (1) because we see in (1) that $ds = \sqrt{\mathbf{R}'(\tau) \cdot \mathbf{R}'(\tau)}\,d\tau$, where the positive square root is understood, and where each $d\tau$ is positive because $b \geq a$. Thus, in $\int_C f\,ds$ the curve C is not oriented, and in $\int_C \mathbf{v} \cdot d\mathbf{R}$ it is.

Just as we evaluate $\int_C f\,ds$ by parametrizing C as in (1), we evaluate $\int_C \mathbf{v} \cdot d\mathbf{R}$ in the same manner. We begin by parametrizing C by

$$\left.\begin{aligned} x &= x(\tau) \\ y &= y(\tau) \\ z &= z(\tau) \end{aligned}\right\} \tau : a \to b, \tag{7}$$

where we write $\tau : a \to b$ rather than $a \leq \tau \leq b$ because the value a (at the initial point on C) can be less than *or* greater than the value b (at the terminal point on C) because of the orientation of C. Then,

$$\begin{aligned} \int_C \mathbf{v} \cdot d\mathbf{R} &= \int_C (v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}) \cdot (dx\hat{\mathbf{i}} + dy\hat{\mathbf{j}} + dz\hat{\mathbf{k}}) \\ &= \int_C v_x(x, y, z)\,dx + v_y(x, y, z)\,dy + v_z(x, y, z)\,dz \end{aligned}$$

*We can regard the position vector $\mathbf{R}$ as a function of the arc length s along C by parametrizing C by $x = x(s)$, $y = y(s)$, $z = z(s)$ since then $\mathbf{R}(x(s), y(s), z(s))$ is a function of s.

$$= \int_a^b \left[v_x(x(\tau), y(\tau), z(\tau)) \frac{dx}{d\tau} + v_y(x(\tau), y(\tau), z(\tau)) \frac{dy}{d\tau} + v_z(x(\tau), y(\tau), z(\tau)) \frac{dz}{d\tau} \right] d\tau. \tag{8}$$

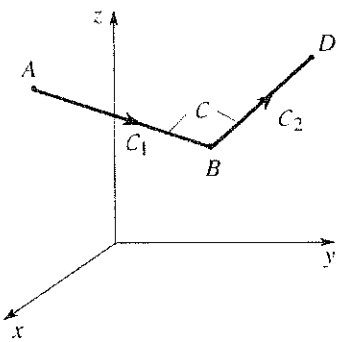


Figure 2. The curve C.

EXAMPLE 1. Evaluate $\int_C \mathbf{v} \cdot d\mathbf{R}$, where $\mathbf{v} = xz^2\hat{\mathbf{i}} - 3\hat{\mathbf{j}} + 2y\hat{\mathbf{k}}$, and where C is the bent line from $A = (2, 1, 3)$ to $B = (1, 2, 1)$ to $D = (3, 4, 5)$, shown schematically in Fig. 2. First, break C into the two parts C_1 and C_2, and write

$$\int_C \mathbf{v} \cdot d\mathbf{R} = \int_{C_1} \mathbf{v} \cdot d\mathbf{R} + \int_{C_2} \mathbf{v} \cdot d\mathbf{R},$$
$$= \int_{C_1} xz^2\, dx - 3\, dy + 2y\, dz + \int_{C_2} xz^2\, dx - 3\, dy + 2y\, dz, \tag{9}$$

which step is permissible for $\int_C \mathbf{v} \cdot d\mathbf{R}$ just as it is for $\int_C f\, ds$ [recall equations (12) in Section 15.2]. Next, use (7) in Section 15.2 to parametrize C_1 and C_2:

$$C_1: \quad \begin{aligned} x &= 2 + (1-2)\tau = 2 - \tau, & x'(\tau) &= -1 \\ y &= 1 + (2-1)\tau = 1 + \tau, & y'(\tau) &= 1 \\ z &= 3 + (1-3)\tau = 3 - 2\tau, & z'(\tau) &= -2 \end{aligned} \tag{10a}$$

and

$$C_2: \quad \begin{aligned} x &= 1 + (3-1)\tau = 1 + 2\tau, & x'(\tau) &= 2 \\ y &= 2 + (4-2)\tau = 2 + 2\tau, & y'(\tau) &= 2 \\ z &= 1 + (5-1)\tau = 1 + 4\tau, & z'(\tau) &= 4 \end{aligned} \tag{10b}$$

as τ goes from 0 to 1 (i.e., $\tau : 0 \to 1$). Finally, expressing dx, dy, dz in (9) as $x'(\tau)\, d\tau$, $y'(\tau)\, d\tau$, $z'(\tau)\, d\tau$, respectively, and using (10), we have

$$\int_C \mathbf{v} \cdot d\mathbf{R} = \int_0^1 [(2-\tau)(3-2\tau)^2(-1) - 3(1) + 2(1+\tau)(-2)]\, d\tau + \int_0^1 [(1+2\tau)(1+4\tau)^2(2) - 3(2) + 2(2+2\tau)(4)]\, d\tau$$
$$= -\frac{97}{6} + \frac{202}{3} = \frac{307}{6}. \quad \blacksquare \tag{11}$$

Recall from Section 15.2 that a curve is said to be *closed* if its endpoints coincide. If C is closed, then we usually write

$$\oint_C \mathbf{v} \cdot d\mathbf{R} \tag{12}$$

in place of $\int_C \mathbf{v} \cdot d\mathbf{R}$. One might think that we will always obtain $\oint_C \mathbf{v} \cdot d\mathbf{R} = 0$ because the endpoints of C coincide, just as it is true that $\int_a^a f(x)\, dx = 0$ for an

integration on the x axis, but that is not necessarily the case. For instance, suppose we close the contour C in Example 1 by adding a straight line segment from D to A (Fig. 2). Call that segment C_3. We find (Exercise 1) that $\int_{C_3} \mathbf{v} \cdot d\mathbf{R} = -\frac{259}{6}$ so

$$\oint_C \mathbf{v} \cdot d\mathbf{R} = \int_{ABDA} \mathbf{v} \cdot d\mathbf{R} = -\frac{97}{6} + \frac{202}{3} - \frac{259}{6} = 8, \tag{13}$$

which is not zero. Of course, a line integral around a closed path *may* be zero, but it need not.

Two examples of line integrals around closed contours are as follows. First, *Ampère's law*,

$$\oint_C \mathbf{H} \cdot d\mathbf{R} = I, \tag{14}$$

states that the line integral of the magnetic field intensity $\mathbf{H}$ (amperes/meter) around a closed curve C equals the current I (amperes) flowing through C. That is, think of C as made of a stiff wire, dipped in soap water so that there is a soap film S with C as its perimeter. Then I is the current crossing the surface S. Second, in fluid mechanics the line integral

$$\oint_C \mathbf{v} \cdot d\mathbf{R} = \Gamma \tag{15}$$

defines the *circulation* Γ (meters^2/second) of the fluid velocity field $\mathbf{v}$ around the contour C; Γ is important in aerodynamics because of its prominent role in the *Kutta–Joukowski formula*,

$$L = \sigma U \Gamma, \tag{16}$$

for the lift L per unit span of an airfoil (i.e., a wing) due to a fluid flow of velocity U and density σ. In (16), Γ is the circulation around any closed clockwise contour that encloses the airfoil such as the ellipse shown in Fig. 3. That is, consider a simple airfoil consisting of a flat plate at incidence α, normal to the paper, a cross section of which is shown in Fig. 3. Rather than the plate moving leftward with flight speed U, it is equivalent to consider the plate as stationary in a rightward flow of speed U, as in a wind tunnel. If $\alpha = 0$ the streamlines are simply horizontal straight lines, but as α is increased the streamlines distort more and more (somewhat as the semi-circular bump causes a distortion of the otherwise-horizontal streamlines in Fig. 4 of Section 16.2). Although the streamline pattern is not shown here, in Fig. 3, it could be found by the method of conformal mapping.* In any case, the result of such calculation is that the velocities above the plate are larger than those below the plate, as sketched at points A and B in the figure. Thus, for $d\mathbf{R}$'s of equal length, the $\mathbf{v} \cdot d\mathbf{R}$ contribution to Γ from point A is positive and larger than the negative contribution from point B so that a positive circulation Γ is established around C by virtue of the inclination α of the plate. According to (16), that circulation gives rise to a lift force. Even if you haven't studied fluid mechanics or aerodynamics you may have studied the Bernoulli equation (relating pressure and velocity) in a

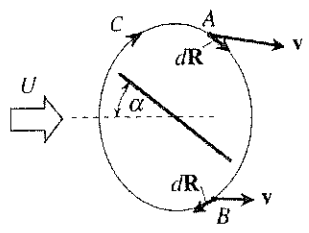


Figure 3. Lift on an airfoil.

*In fact, that solution is outlined, for the case where $\alpha = 90°$, in Exercise 6 in Section 23.6, and it could be modified for $\alpha \neq 90°$.

physics course. Since Bernoulli's equation states that high velocity gives low pressure and low velocity gives high pressure, it follows – from the fact that the velocity is greater above the plate than below it – that the pressure is lower above the plate than below it, and that pressure difference establishes the lift L.

16.9.2. Stokes's theorem. Recall from the Closure in Section 16.8 that the fundamental theorem of the integral calculus,

$$\int_a^b F'(x)\,dx = F(x)\Big|_{x=a}^{x=b}, \tag{17}$$

gives the integral along a (one-dimensional) line segment as an evaluation at the (zero-dimensional) boundary of that line, namely, its endpoints, and that the Gauss divergence theorem gives the integral over a (three-dimensional) region as an integration over the (two-dimensional) boundary of that region, namely, its surface. Filling the gap between those two results, Stokes's theorem gives the integral over a two-dimensional open surface $\mathcal{S}$ in terms of a line integral around its one-dimensional edge curve $\mathcal{C}$. Here we distinguish a **closed surface**, which encloses a volume (as does the surface $\mathcal{S}$ in the divergence theorem), from an **open surface**, which does not. For instance, the surface of a basketball is closed and the surface defined by a potato chip is open.

THEOREM 16.9.1 *Stokes's Theorem*
Let $\mathbf{v}$ be a C^1 vector field defined in a region $\mathcal{R}$ in 3-space. Let $\mathcal{S}$ be an open piecewise smooth orientable surface within $\mathcal{R}$, and let the edge of $\mathcal{S}$ be a piecewise smooth simple closed curve $\mathcal{C}$. Then

$$\boxed{\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \nabla \times \mathbf{v}\,dA = \oint_{\mathcal{C}} \mathbf{v} \cdot d\mathbf{R},} \tag{18}$$

where $\hat{\mathbf{n}}$ is a unit normal to $\mathcal{S}$, the orientation of $\mathcal{C}$ and the direction of $\hat{\mathbf{n}}$ being in accordance with the right-hand rule (Fig. 4).

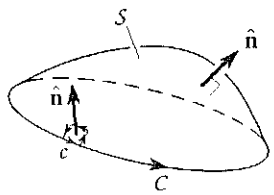


Figure 4. $\mathcal{S}$, $\mathcal{C}$, and $\hat{\mathbf{n}}$ in (18).

Before discussing the proof of Stokes's theorem we need to explain our statement in the theorem about the right-hand rule. First, observe that we refrain from calling $\hat{\mathbf{n}}$ the "outward" or "inward" normal to $\mathcal{S}$. If $\mathcal{S}$ were a closed surface, such as the surface of a sphere or cube, it would make sense to speak of an outward (or inward) normal to $\mathcal{S}$ since $\mathcal{S}$ would "inherit" an orientation from the volume $\mathcal{V}$ that it enclosed: normal vectors on $\mathcal{S}$ directed into $\mathcal{V}$ would be called inward, and those directed out of $\mathcal{V}$ would be called outward. But in Stokes's theorem $\mathcal{S}$ is not a closed surface so the terms "outward" and "inward" do not apply.

The idea, then, is that $\hat{\mathbf{n}}$, in (18), can be either of the two (oppositely directed) fields of normals on $\mathcal{S}$. Similarly, $\mathcal{C}$ can be oriented in either of two possible directions. Selecting one of those two orientations for $\mathcal{C}$, arbitrarily, we then choose

between the two possible normals so that $\hat{\mathbf{n}}$ and C are in accordance with the right-hand rule. That is, if we consider a small contour C', part of which coincides with (and is oriented in the same way as) the boundary curve C, then applying the right-hand rule to C' yields a normal $\hat{\mathbf{n}}$ (Fig. 4). We then continue this normal over the rest of S. For example, if the surface S is a flat surface in the x, y plane, then we can have $\hat{\mathbf{n}} = +\hat{\mathbf{k}}$ and C counterclockwise, *or* $\hat{\mathbf{n}} = -\hat{\mathbf{k}}$ and C clockwise.

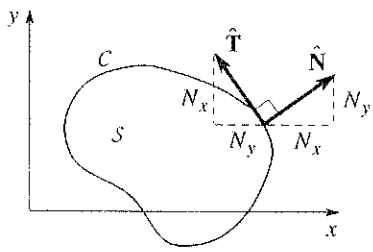


Figure 5. Plane region.

Proof: Let us limit our proof of (18) to the case where S is flat, in which case S can be assumed to lie in the plane of the paper as in Fig. 5; extension from the plane case to the general case will be left for the exercises.

Recall the two-dimensional divergence theorem given by (47) in Section 16.8.2. In the present case the region "$\mathcal{R}$" in (47) is S, and in place of "$\mathbf{v}$" and "$\hat{\mathbf{n}}$" it will be convenient to use the letters $\mathbf{V}$ and $\hat{\mathbf{N}}$. Thus,

$$\int_S \nabla \cdot \mathbf{V}\, dA = \int_C \hat{\mathbf{N}} \cdot \mathbf{V}\, ds. \tag{19}$$

Writing out the integrands gives

$$\int_S \left(\frac{\partial V_x}{\partial x} + \frac{\partial V_y}{\partial y} \right) dA = \int_C (N_x V_x + N_y V_y)\, ds, \tag{20}$$

where $\mathbf{V} = V_x \hat{\mathbf{i}} + V_y \hat{\mathbf{j}}$ and $\hat{\mathbf{N}} = N_x \hat{\mathbf{i}} + N_y \hat{\mathbf{j}}$ (Fig. 5). If we define a new vector $\mathbf{v} = v_x \hat{\mathbf{i}} + v_y \hat{\mathbf{j}} \equiv -V_y \hat{\mathbf{i}} + V_x \hat{\mathbf{j}}$, then (20) becomes

$$\begin{aligned} \int_S \left(\frac{\partial v_y}{\partial x} - \frac{\partial v_x}{\partial y} \right) dA &= \int_C (N_x v_y - N_y v_x)\, ds \\ &= \int_C (v_x \hat{\mathbf{i}} + v_y \hat{\mathbf{j}}) \cdot (-N_y \hat{\mathbf{i}} + N_x \hat{\mathbf{j}})\, ds \\ &= \int_C \mathbf{v} \cdot \hat{\mathbf{T}}\, ds, \end{aligned} \tag{21}$$

where $\hat{\mathbf{T}} = -N_y \hat{\mathbf{i}} + N_x \hat{\mathbf{j}}$ is the unit counterclockwise tangent vector to C. Observe that the integrand on the left is the z component of $\nabla \times \mathbf{v}$, namely, $\hat{\mathbf{k}} \cdot \nabla \times \mathbf{v}$, and that $\hat{\mathbf{T}}\, ds = d\mathbf{R}$ along C. Finally, $\hat{\mathbf{k}}$ is identical to the $\hat{\mathbf{n}}$ in Stokes's theorem so (21) becomes

$$\int_S \hat{\mathbf{n}} \cdot \nabla \times \mathbf{v}\, dA = \int_C \mathbf{v} \cdot d\mathbf{R},$$

which was to be proved.

Here S was flat and $\mathbf{v}$ was only a two-dimensional field. A proof for the general case, where S need not be flat and $\mathbf{v}$ is three-dimensional, is outlined in the exercises. ■

How can we understand Stokes's theorem from a physical point of view? For definiteness, think of $\mathbf{v}$ as a fluid velocity field (which we can do even if it is not).

Recall, from our discussion of fluid mechanics and the curl, that $\nabla \times \mathbf{v}$ is twice the fluid particle angular velocity and is known as the vorticity and denoted as $\mathbf{\Omega}$. Thus, $\int_S \hat{\mathbf{n}} \cdot \nabla \times \mathbf{v}\, dA = \int_S \hat{\mathbf{n}} \cdot \mathbf{\Omega}\, dA$ is the vorticity flux across S. Furthermore, $\oint_C \mathbf{v} \cdot d\mathbf{R}$ is the circulation around C. Thus, in fluid mechanics terminology, Stokes's theorem (18) states that *the circulation* Γ *of the flow field* $\mathbf{v}$, *around* C, *equals the vorticity flux through* S.*

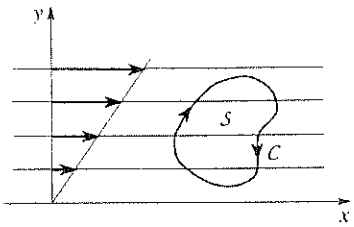


Figure 6. Shear flow.

EXAMPLE 2. *Vorticity Flux.* Consider the plane flow field $\mathbf{v} = \kappa y \hat{\mathbf{i}}$, where κ is some positive constant. The flow, called a *shear flow*, is shown in Fig. 6. Let S be a generic region in the x, y plane, and let the edge curve be clockwise, say, so that $\hat{\mathbf{n}} = -\hat{\mathbf{k}}$. Then $\mathbf{\Omega} = \nabla \times \mathbf{v} = -\kappa \hat{\mathbf{k}}$ is itself a "flow," a uniform flow, of magnitude κ, directed into the paper. Thus, the vorticity flux through S is simply κ times the area A of S, and that is exactly what the left-hand side of (18) gives,

$$\int_S \hat{\mathbf{n}} \cdot \nabla \times \mathbf{v}\, dA = \int_S \hat{\mathbf{n}} \cdot \mathbf{\Omega}\, dA = \int_S -\hat{\mathbf{k}} \cdot (-\kappa \hat{\mathbf{k}})\, dA = \kappa A. \quad \blacksquare$$

EXAMPLE 3. *Verification of Stokes's Theorem.* Verify Stokes's theorem for the case where $\mathbf{v} = xz\hat{\mathbf{j}}$, and where S is the surface $z = 4 - y^2$, cut off by the planes $x = 0$, $z = 0$, and $y = x$, with C oriented as shown in Fig. 7; we have broken C into the three parts, C_1, C_2, C_3. Let us work out the left-hand side of (18) first. Using the gradient method to find $\hat{\mathbf{n}}$ [equation (20) in Section 16.8], and the formula

$$dA = \sqrt{1 + f_x^2 + f_y^2}\, dx\, dy \tag{22}$$

for dA [equation (11) in Section 16.8, where S is the surface $z = f(x, y) = 4 - y^2$ in this case], we have

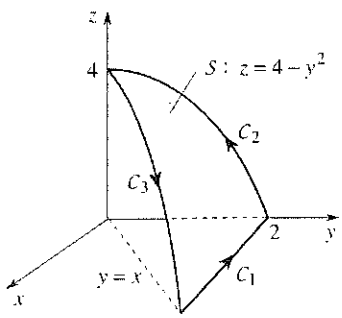


Figure 7. Example 3.

$$\nabla \times \mathbf{v} = \begin{vmatrix} \hat{\mathbf{i}} & \hat{\mathbf{j}} & \hat{\mathbf{k}} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ 0 & xz & 0 \end{vmatrix} = -x\hat{\mathbf{i}} + z\hat{\mathbf{k}},$$

$$\hat{\mathbf{n}} = \frac{\nabla(z + y^2)}{\|\nabla(z + y^2)\|} = \frac{2y\hat{\mathbf{j}} + \hat{\mathbf{k}}}{\sqrt{4y^2 + 1}},$$

$$\hat{\mathbf{n}} \cdot \nabla \times \mathbf{v} = \frac{z}{\sqrt{4y^2 + 1}},$$

$$dA = \sqrt{1 + (0)^2 + (-2y)^2}\, dx\, dy = \sqrt{1 + 4y^2}\, dx\, dy \tag{23}$$

so

$$\int_S \hat{\mathbf{n}} \cdot \nabla \times \mathbf{v}\, dA = \int_0^2 \int_0^y \left. \frac{z}{\sqrt{4y^2 + 1}} \right|_{z=4-y^2} \sqrt{1 + 4y^2}\, dx\, dy$$

$$= \int_0^2 \int_0^y (4 - y^2)\, dx\, dy = 4, \tag{24}$$

*Actually, (18) is not "Stokes's theorem" it is only the formula within Stokes's theorem, but for brevity it is convenient to refer to it as Stokes's theorem.

where the x, y integration limits correspond to the triangular "shadow" of $\mathcal{S}$ down onto the x, y plane. Of course, the integration is actually on $\mathcal{S}$ nonethless; it has simply been "referred" to the x, y plane by virtue of the relation (22) between dA on $\mathcal{S}$ and $dx\,dy$ in the x, y plane.

Turning to the right-hand side of (18), we have

$$\oint_{\mathcal{C}} \mathbf{v} \cdot d\mathbf{R} = \oint_{\mathcal{C}} xz\hat{\mathbf{j}} \cdot (dx\hat{\mathbf{i}} + dy\hat{\mathbf{j}} + dz\hat{\mathbf{k}}) = \oint_{\mathcal{C}} xz\,dy$$
$$= \oint_{\mathcal{C}_1} xz\,dy + \oint_{\mathcal{C}_2} xz\,dy + \oint_{\mathcal{C}_3} xz\,dy. \tag{25}$$

The first integral is zero because $z = 0$ on $\mathcal{C}_1$, and the second is zero because $x = 0$ on $\mathcal{C}_2$. To evaluate the $\mathcal{C}_3$ integral, we can parametrize $\mathcal{C}_3$ as follows. Let $x = \tau$, say. Then $y = \tau$ because $y = x$ on $\mathcal{C}_3$, and $z = 4 - \tau^2$ because $z = 4 - y^2$ on $\mathcal{C}_3$. Finally, τ goes from 0 to 2 because $x = \tau$ and x goes from 0 to 2 as we move along $\mathcal{C}_3$ from the initial point $(0, 0, 4)$ to the final point $(2, 2, 0)$. Thus, (25) gives

$$\oint_{\mathcal{C}} \mathbf{v} \cdot d\mathbf{R} = 0 + 0 + \int_0^2 \tau(4 - \tau^2)\,d\tau = 4, \tag{26}$$

in agreement with (24). ∎

EXAMPLE 4. *Ampère's Law.* As a physical application of Stokes's theorem, let us derive one of the celebrated Maxwell's equations of electromagnetic field theory. We consider a region of 3-space in which there is a magnetic field with magnetic field intensity $\mathbf{H}$ (amperes/meter), a charge density field σ (coulombs/meter^3), and a velocity field $\mathbf{v}$ (meters/second). In general, the scalar field σ and the vector fields $\mathbf{H}$ and $\mathbf{v}$ will vary with x, y, z, and the time t, but any dependence on t will not be relevant here.

We begin with Ampère's law,

$$\oint_{\mathcal{C}} \mathbf{H} \cdot d\mathbf{R} = I, \tag{27}$$

which states that the line integral of the magnetic field intensity $\mathbf{H}$ around any closed curve $\mathcal{C}$ equals the current I (amperes) passing through any control surface $\mathcal{S}$ (i.e., a mathematical surface, not a physical surface) having $\mathcal{C}$ as its boundary. (Note that when we say "any" surface or "any" curve $\mathcal{C}$ it is understood that they be suitably decent, namely, that they satisfy the conditions on $\mathcal{S}$ and $\mathcal{C}$ in Stokes's theorem.)

Since I is the flux of charge across $\mathcal{S}$ we can express it as

$$I = \int_{\mathcal{S}} \sigma\,\hat{\mathbf{n}} \cdot \mathbf{v}\,dA = \int_{\mathcal{S}} \hat{\mathbf{n}} \cdot (\sigma\mathbf{v})\,dA, \tag{28}$$

where $\hat{\mathbf{n}}$ is the unit normal to $\mathcal{S}$ in whichever of the two directions is chosen as the direction in which I is positive. As a partial check on (28), note that its units, (coulombs/m^3)(m/sec)(m^2), do indeed give coulombs/sec.

Using Stokes's theorem to express

$$\oint_C \mathbf{H} \cdot d\mathbf{R} = \int_S \hat{\mathbf{n}} \cdot \nabla \times \mathbf{H}\, dA, \tag{29}$$

(27)–(29) give

$$\int_S \hat{\mathbf{n}} \cdot (\nabla \times \mathbf{H} - \mathbf{J})\, dA = 0, \tag{30}$$

where it is traditional to combine the product $\sigma\mathbf{v}$ as a single field $\mathbf{J} = \sigma\mathbf{v}$, known as the current density. Finally, since (30) holds for *any* surface S, it follows that $\nabla \times \mathbf{H} - \mathbf{J}$ must be identically zero in the region so

$$\nabla \times \mathbf{H} = \mathbf{J}, \tag{31}$$

which is one of Maxwell's equations.

Why does it follow from (30) that $\nabla \times \mathbf{H} - \mathbf{J} = 0$? Can't the integral be zero by virtue of cancellation, without the integrand being *identically* zero? For a specific S, yes, but not for all possible choices of S. For suppose that $\nabla \times \mathbf{H} - \mathbf{J}$ is nonzero at some point P; let it be nonzero vector $\mathbf{a}$ at P. Then let us choose S to be a circular disk with its center at P, and let its normal $\hat{\mathbf{n}}$ be aligned with $\mathbf{a}$ (Fig. 8). Of course $\nabla \times \mathbf{H} - \mathbf{J} = \mathbf{a}$ only at P; it varies on S. But if it is continuous, which we assume, then it must be possible to choose the radius ϵ of S small enough so that $(\nabla \times \mathbf{H} - \mathbf{J}) \cdot \hat{\mathbf{n}} > 0$ everywhere on S, in which case the integral in (30) is positive rather than zero. This contradiction of (30) shows that it must not be possible for $\nabla \times \mathbf{H} - \mathbf{J}$ to be nonzero anywhere in the field.

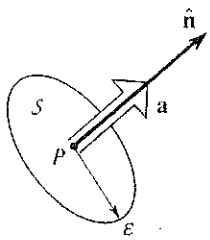


Figure 8. Choosing S.

COMMENT 1. The logic used to infer (31) from (30) is very similar to that used to infer (28) from (27) and also (38) from (37) in Section 16.8. We suggest that you review those two examples in Section 16.8

COMMENT 2. Understand that (31) is a *field equation*, specifically a partial differential equation relating the vector fields $\mathbf{H}$ and $\mathbf{J}$ at each point in the field. It happens to be a vector equation or, equivalently, the three scalar PDE's

$$\begin{aligned} \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} &= J_x, \\ \frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} &= J_y, \\ \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} &= J_z, \end{aligned} \tag{32}$$

whereas the field equations (28) and (38) in Section 16.8 were single scalar PDE's. ∎

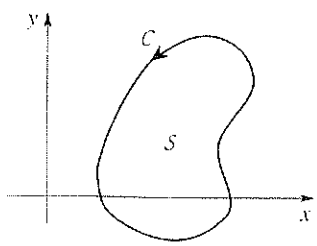


Figure 9. Stokes's theorem for plane case.

16.9.3. Green's theorem. Let us apply Stokes's theorem to the plane case where $\mathbf{v} = P(x, y)\hat{\mathbf{i}} + Q(x, y)\hat{\mathbf{j}}$, and where S is a flat surface in the x, y plane, and let C be counterclockwise (Fig. 9). Then $\hat{\mathbf{n}} = +\hat{\mathbf{k}}$ and $\nabla \times \mathbf{v} = \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)\hat{\mathbf{k}}$ so we have the following result, which is known as **Green's theorem**.

THEOREM 16.9.2 *Green's Theorem*
Let a C^1 vector field $\mathbf{v} = P(x,y)\hat{\mathbf{i}} + Q(x,y)\hat{\mathbf{j}}$ be defined in a region $\mathcal{R}$ in 2-space. Let $\mathcal{S}$ be a region within $\mathcal{R}$, and let the edge of $\mathcal{S}$ be a piecewise smooth simple closed curve $\mathcal{C}$, oriented counterclockwise. Then

$$\boxed{\int_{\mathcal{S}} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA = \oint_{\mathcal{C}} P\,dx + Q\,dy.} \tag{33}$$

Since Green's theorem is really but a special case of Stokes's theorem, let us limit our discussion of it to an illustrative example, with additional material reserved for the exercises.

EXAMPLE 5. Verify (33) for the case where $P(x,y) = xy^3$, and $Q(x,y) = x^2 - y^2$, and where $\mathcal{S}$ is as shown in Fig. 10. Then

$$\begin{aligned}
\int_{\mathcal{S}} \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dA &= \int_0^2 \int_{y/2}^1 (2x - 3xy^2)\,dx\,dy \\
&= \int_0^2 \left(x^2 - \frac{3}{2}x^2y^2 \right)\Big|_{x=y/2}^{x=1} dy \\
&= \int_0^2 \left(1 - \frac{3}{2}y^2 - \frac{y^2}{4} + \frac{3}{2}\frac{y^2}{4}y^2 \right) dy \\
&= \left(y - \frac{1}{2}y^3 - \frac{1}{12}y^3 + \frac{3}{40}y^5 \right)\Big|_0^2 = -\frac{4}{15}.
\end{aligned} \tag{34}$$

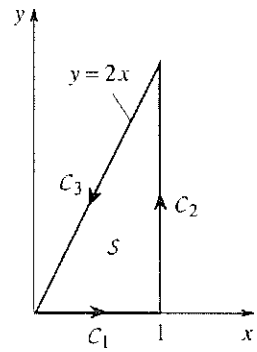


Figure 10. The region $\mathcal{S}$.

Breaking $\mathcal{C}$ into $\mathcal{C}_1 + \mathcal{C}_2 + \mathcal{C}_3$, we have

$$\begin{aligned}
\oint_{\mathcal{C}} P\,dx + Q\,dy &= \int_{\mathcal{C}_1} xy^3\,dx + (x^2 - y^2)\,dy + \int_{\mathcal{C}_2} xy^3\,dx + (x^2 - y^2)\,dy \\
&\quad + \int_{\mathcal{C}_3} xy^3\,dx + (x^2 - y^2)\,dy \\
&= \int_{\mathcal{C}_2} (x^2 - y^2)\,dy + \int_{\mathcal{C}_3} xy^3\,dx + (x^2 - y^2)\,dy
\end{aligned} \tag{35}$$

since $y = 0$ on $\mathcal{C}_1$, $dy = 0$ on $\mathcal{C}_1$, and $dx = 0$ on $\mathcal{C}_2$. Finally, since $x = 1$ on $\mathcal{C}_2$, and $y = 2x$ on $\mathcal{C}_3$, (35) gives

$$\oint_{\mathcal{C}} P\,dx + Q\,dy = \int_0^2 (1 - y^2)\,dy + \int_1^0 [8x^4 + (x^2 - 4x^2)(2)]\,dx = -\frac{4}{15}, \tag{36}$$

in agreement with (34).

COMMENT. Why didn't we parametrize $\mathcal{C}_2$ and $\mathcal{C}_3$ to evaluate the final integrals in (35)? Actually, we did; in $\mathcal{C}_2$ we used y as the parameter for $\mathcal{C}_2$ (i.e., we used $y = \tau$ but didn't

bother changing the name from y to τ), and we used x as the parameter for C_3. If you prefer, use τ explicitly. ∎

16.9.4. Non-Cartesian coordinates. (Optional) Let us illustrate the occurrence of non-Cartesian coordinates, in Stokes's theorem, with two examples.

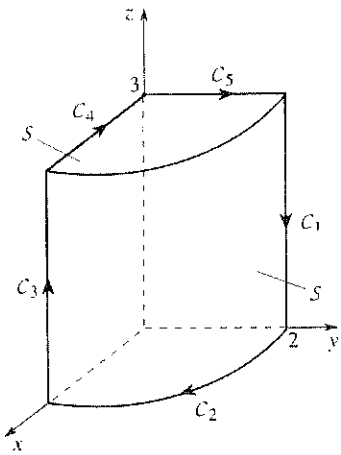


Figure 11. S and C for Example 6.

EXAMPLE 6. *Cylindrical Coordinates.* Verify Stokes's theorem for the case where

$$\mathbf{v} = -r^2 \cos\theta \hat{\mathbf{e}}_r + r^2 \hat{\mathbf{e}}_\theta + r \sin\theta \hat{\mathbf{e}}_z, \tag{37}$$

where S is as shown in Fig. 11 (essentially, a quarter of a soup can but with no bottom), and where C is oriented as shown in the figure.

Let us do the surface integral first. From the cylindrical coordinate expression for $\nabla \times \mathbf{v}$ in Section 16.7,

$$\nabla \times \mathbf{v} = \cos\theta \hat{\mathbf{e}}_r - \sin\theta \hat{\mathbf{e}}_\theta + (3r - r\sin\theta)\hat{\mathbf{e}}_z. \tag{38}$$

On the curved part of the surface $\hat{\mathbf{n}} = -\hat{\mathbf{e}}_r$ (in accordance with the orientation of C) and $dA = r\,d\theta\,dz$ (for a constant-r surface), and on the flat top $\hat{\mathbf{n}} = -\hat{\mathbf{e}}_z$ and $dA = r\,dr\,d\theta$ (for a constant-z surface) so

$$\begin{aligned}
\int_S \hat{\mathbf{n}} \cdot \nabla \times \mathbf{v}\, dA &= \int_0^3 \int_0^{\pi/2} (-\hat{\mathbf{e}}_r) \cdot \nabla \times \mathbf{v}\, r\, d\theta\, dz \Big|_{r=2} \\
&\quad + \int_0^{\pi/2} \int_0^2 (-\hat{\mathbf{e}}_z) \cdot \nabla \times \mathbf{v}\, r\, dr\, d\theta \Big|_{z=3} \\
&= -2 \int_0^3 \int_0^{\pi/2} \cos\theta\, d\theta\, dz - \int_0^{\pi/2} \int_0^2 (3r^2 - r^2 \sin\theta)\, dr\, d\theta \\
&= -\frac{10}{3} - 4\pi.
\end{aligned} \tag{39}$$

Turning to the line integral, we obtain

$$\begin{aligned}
\oint_C \mathbf{v} \cdot d\mathbf{R} &= \oint_C (-r^2 \cos\theta \hat{\mathbf{e}}_r + r^2 \hat{\mathbf{e}}_\theta + r \sin\theta \hat{\mathbf{e}}_z) \cdot (dr\, \hat{\mathbf{e}}_r + r\, d\theta\, \hat{\mathbf{e}}_\theta + dz \hat{\mathbf{e}}_z) \\
&= \oint_C -r^2 \cos\theta\, dr + r^3\, d\theta + r \sin\theta\, dz \\
&= \int_3^0 r \sin\theta\, dz \Big|_{r=2, \theta=\pi/2} + \int_{\pi/2}^0 r^3\, d\theta \Big|_{r=2, z=0} + \int_0^3 r \sin\theta\, dz \Big|_{r=2, \theta=0} \\
&\quad + \int_2^0 (-r^2 \cos\theta)\, dr \Big|_{\theta=0, z=3} + \int_0^2 (-r^2 \cos\theta)\, dr \Big|_{\theta=\pi/2, z=3} \\
&= -6 - 4\pi + 0 + \frac{8}{3} + 0 = -\frac{10}{3} - 4\pi,
\end{aligned} \tag{40}$$

in agreement with (39). Following the third equality in (40), the five integrals correspond to $C_1, \ldots, C_5$, respectively. On C_1, for instance, dr and $d\theta$ are zero because r and θ are

constant on C_1 so we omitted the $-r\cos\theta\,dr + r^3\,d\theta$ part of the integrand. ∎

EXAMPLE 7. *Spherical Coordinates.* Verify Stokes's theorem for the case where

$$\mathbf{v} = \rho\hat{\mathbf{e}}_\rho - \rho^2\hat{\mathbf{e}}_\theta, \tag{41}$$

where $\mathcal{S}$ is as shown in Fig. 12 (a quarter of a cone, of semi-angle $\pi/6$), and where $\mathcal{C}$ is oriented as shown in the figure.

From the spherical coordinate expression for $\nabla \times \mathbf{v}$ in Section 16.7,

$$\nabla \times \mathbf{v} = -\rho\cot\phi\hat{\mathbf{e}}_\rho + 3\rho\hat{\mathbf{e}}_\phi. \tag{42}$$

Further, $\hat{\mathbf{n}} = \hat{\mathbf{e}}_\phi$ and $dA = \rho|\sin\phi|\,d\rho\,d\theta$ since $\mathcal{S}$ is a constant-ϕ surface so

$$\begin{aligned}\int_{\mathcal{S}} \hat{\mathbf{n}}\cdot\nabla\times\mathbf{v}\,dA &= \int_0^{\pi/2}\int_0^2 \hat{\mathbf{e}}_\phi\cdot(-\rho\cot\phi\hat{\mathbf{e}}_\rho + 3\rho\hat{\mathbf{e}}_\phi)\,(\rho|\sin\phi|\,d\rho\,d\theta)\Big|_{\phi=\pi/6}\\ &= \frac{3}{2}\int_0^{\pi/2}\int_0^2 \rho^2\,d\rho\,d\theta = 2\pi.\end{aligned} \tag{43}$$

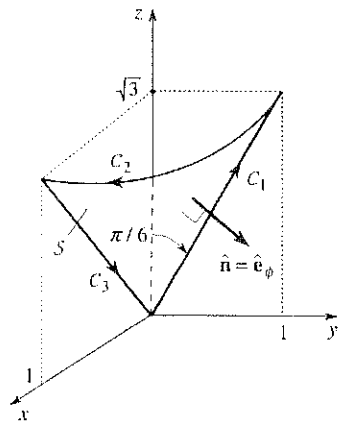


Figure 12. $\mathcal{S}$ and $\mathcal{C}$ for Example 7.

Next,

$$\begin{aligned}\oint_{\mathcal{C}} \mathbf{v}\cdot d\mathbf{R} &= \oint_{\mathcal{C}} (\rho\hat{\mathbf{e}}_\rho - \rho^2\hat{\mathbf{e}}_\theta)\cdot(d\rho\,\hat{\mathbf{e}}_\rho + \rho\,d\phi\,\hat{\mathbf{e}}_\phi + \rho|\sin\phi|\,d\theta\hat{\mathbf{e}}_\theta)\\ &= \oint_{\mathcal{C}} \rho\,d\rho - \rho^3\sin\phi\,d\theta\\ &= \int_0^2 \rho\,d\rho\Big|_{\phi=\pi/6,\theta=\pi/2} + \int_{\pi/2}^0 (-\rho^3\sin\phi)\,d\theta\Big|_{\rho=2,\phi=\pi/6} + \int_2^0 \rho\,d\rho\Big|_{\phi=\pi/6,\theta=0}\\ &= 2 + 2\pi - 2 = 2\pi,\end{aligned} \tag{44}$$

in agreement with (43).

COMMENT. If the expressions used for dA and $d\mathbf{R}$ are not clear to you, see equation (18) in Section 15.6. If $\mathcal{S}$ were not a constant-coordinate surface we could not use (18) for dA, but we can always fall back on the $dA = \sqrt{EG - F^2}\,du\,dv$ formula in Section 15.5. Similarly, in this example $\hat{\mathbf{n}} = \hat{\mathbf{e}}_\phi$ is clear by inspection, but in more difficult cases we can always fall back on the gradient method [see equation (20) in section 16.8]. ∎

Closure. Having already studied line integrals in the form $\int_{\mathcal{C}} f\,ds$, we first discuss line integrals that occur in the form $\int_{\mathcal{C}} \mathbf{v}\cdot d\mathbf{R}$ and note their equivalance, except for the fact that $\mathcal{C}$ is not oriented in the former but it is in the latter. The latter form is prominent in Stokes's theorem,

$$\int_{\mathcal{S}} \hat{\mathbf{n}}\cdot\nabla\times\mathbf{v}\,dA = \oint_{\mathcal{C}} \mathbf{v}\cdot d\mathbf{R}, \tag{45}$$

which evaluates the surface integral on $\mathcal{S}$ in terms of an integral along its boundary curve $\mathcal{C}$. Remember that $\mathcal{S}$ in (45) is an open surface, not a closed surface, and

that $\hat{\mathbf{n}}$ is not the "outward" normal because inward and outward are not meaningful for an open surface. Rather, the field of $\hat{\mathbf{n}}$'s on $\mathcal{S}$ are related to the orientation of $\mathcal{C}$ according to the right-hand rule. Finally, we derive Green's theorem (33), and observe that it is actually just a special case of Stokes's theorem for the case where $\mathbf{v} = P(x,y)\hat{\mathbf{i}} + Q(x,y)\hat{\mathbf{j}}$ and where $\mathcal{S}$ is a flat surface in the x, y plane.

Not only are the fundamental theorem of the integral calculus, Stokes's theorem, and the divergence theorem similar in expressing an integral over an n-dimensional region in terms of an evaluation over its $(n-1)$-dimensional boundary, they can all be generalized to a single statement in n dimensions using an elegant theory known as the theory of **differential forms**.*

EXERCISES 16.9

1. Derive the result $\int_{\mathcal{C}_3} \mathbf{v} \cdot d\mathbf{R} = -\frac{259}{6}$, claimed in the paragraph following Example 1.

2. Verify Stokes's theorem. $\mathcal{S}$ is a plane surface with straight edges. The vertices and orientation of $\mathcal{C}$ are given.

(a) $\mathbf{v} = xy\hat{\mathbf{i}} - (2x - y)\hat{\mathbf{k}}$, $\mathcal{C} : (0,0,0)$ to $(1,1,0)$ to $(1,0,0)$ to $(0,0,0)$

(b) $\mathbf{v} = xy^2\hat{\mathbf{i}} - y^3\hat{\mathbf{j}} + 4xyz\hat{\mathbf{k}}$, $\mathcal{C} : (0,1,0)$ to $(1,0,0)$ to $(1,1,0)$ to $(0,1,0)$

(c) $\mathbf{v} = y^2\hat{\mathbf{i}} - x^2\hat{\mathbf{j}} - \sin(x^2y^2z^2)\hat{\mathbf{k}}$, $\mathcal{C} : (0,0,0)$ to $(2,0,0)$ to $(1,-1,0)$ to $(0,0,0)$

(d) $\mathbf{v} = xe^y\hat{\mathbf{i}} + (x+z)\hat{\mathbf{j}} - \hat{\mathbf{k}}$, $\mathcal{C} : (-1,0,0)$ to $(0,1,0)$ to $(0,-1,0)$ to $(-1,0,0)$

(e) $\mathbf{v} = x^2yz\hat{\mathbf{j}}$, $\mathcal{C} : (0,1,0)$ to $(1,1,0)$ to $(1,0,1)$ to $(0,0,1)$ to $(0,1,0)$

(f) $\mathbf{v} = z\hat{\mathbf{i}} + y\hat{\mathbf{j}} + x\hat{\mathbf{k}}$, $\mathcal{C} : (1,0,0)$ to $(0,1,0)$ to $(-1,0,0)$ to $(0,-1,0)$ to $(1,0,0)$

(g) $\mathbf{v} = xy^2z\hat{\mathbf{j}}$, $\mathcal{C} : (0,0,1)$ to $(0,1,2)$ to $(1,2,0)$ to $(0,0,1)$

(h) $\mathbf{v} = xyz\hat{\mathbf{j}}$, $\mathcal{C} : (0,-1,0)$ to $(0,0,2)$ to $(1,1,0)$ to $(0,-1,0)$

(i) $\mathbf{v} = x^2z\hat{\mathbf{k}}$, $\mathcal{C} : (1,-1,0)$ to $(1,1,0)$ to $(0,0,1)$ to $(1,-1,0)$

3. Verify Stokes's theorem.

(a) $\mathbf{v} = x^2\hat{\mathbf{i}} + xz\hat{\mathbf{j}}$. $\mathcal{S}$ is a plane surface with edge curve $\mathcal{C}$ as follows: straight line from $(0,0,1)$ to $(0,1,1)$, then a straight line from $(0,1,1)$ to $(1,1,0)$, then back to $(0,0,1)$ along a curve parametrized by $x = \tau^2$, $y = \tau$, $z = 1 - \tau^2$, where $\tau : 1 \to 0$.

(b) $\mathbf{v} = x^2\hat{\mathbf{i}} - xz\hat{\mathbf{j}}$. $\mathcal{S}$ is a plane surface with edge curve $\mathcal{C}$ as follows: straight line from $(1,2,0)$ to $(1,1,0)$, then along a curve parametrized by $x = \tau^2$, $y = \tau$, $z = 1 - \tau^2$, where $\tau : 1 \to 0$, then along a curve parametrized by $x = \tau^2$, $y = 2\tau$, $z = 1 - \tau^2$, where $\tau : 0 \to 1$.

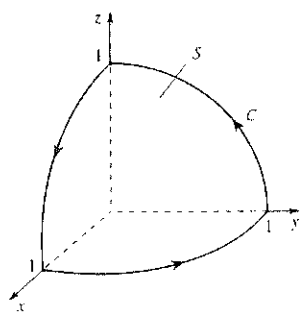
(c) $\mathbf{v} = y^2\hat{\mathbf{j}} - xy^2z\hat{\mathbf{k}}$. $\mathcal{S}$ is the surface of a unit cube with corners at $(0,0,0)$, $(1,0,0)$, $(0,1,0)$, $(0,0,1)$, $(1,0,1)$, $(1,1,0)$, $(0,1,1)$, $(1,1,1)$, excluding the face $x = 1$. Take $\hat{\mathbf{n}} = \hat{\mathbf{k}}$ on the $z = 1$ face.

(d) $\mathbf{v} = -y\hat{\mathbf{i}} + x\hat{\mathbf{j}} + 3\hat{\mathbf{k}}$. $\mathcal{S}$ is the surface $z = 4 - x^2 - y^2$ between $z = 0$ and $z = 4$, and $\hat{\mathbf{n}} = \hat{\mathbf{k}}$ at the point $(0,0,4)$ on $\mathcal{S}$.

(e) $\mathbf{v} = x^2y\hat{\mathbf{j}}$. $\mathcal{S}$ lies in the x, y plane and has edge curve $\mathcal{C}$ as follows: straight line from $(0,0,0)$ to $(0,-1,0)$, then a straight line from $(0,-1,0)$ to $(1,1,0)$, then along $y = x^2$ back to $(0,0,0)$.

(f) $\mathbf{v} = x\hat{\mathbf{i}} - xz\hat{\mathbf{k}}$. $\mathcal{S}$ lies in the x, z plane and has edge curve $\mathcal{C}$ as follows: $z = x$ from $(0,0,0)$ to $(1,0,1)$, then $x = z^2$ from $(1,0,1)$ to $(0,0,0)$.

(g) $\mathbf{v} = yz\hat{\mathbf{k}}$. $\mathcal{S}$ is the surface $z = 1 - x^2 - y^2$ cut off by the planes $x = 0$, $y = 0$, and $z = 0$, and $\mathcal{C}$ oriented as shown.

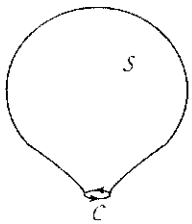


*For a readable introduction to differential forms, see Section 7.6 of J. E. Marsden and A. J. Tromba, *Vector Calculus* (San Fransisco: W. H. Freeman, 1976).

(h) $\mathbf{v} = x^2 z\hat{\mathbf{i}} - 3xy\hat{\mathbf{k}}$. $\mathcal{S}$ is the surface $x = 1 - z^2$ for $-1 \le x \le 1, 0 \le y \le 3$, with $\hat{\mathbf{n}} = -\hat{\mathbf{i}}$ at the point $(1, 1, 0)$ on $\mathcal{S}$.

(i) $\mathbf{v} = yz\hat{\mathbf{i}} + xy\hat{\mathbf{k}}$. $\mathcal{S}$ is the surface $x = z^2$ for $0 \le x \le 1$, $0 \le y \le 2$, with $\hat{\mathbf{n}} = -\hat{\mathbf{i}}$ at the point $(0, 1, 0)$ on $\mathcal{S}$.

4. Remember that $\mathcal{S}$ in Stokes's theorem is an open surface. Suppose we let $\mathcal{S}$ tend to a closed surface by letting $\mathcal{C}$ get very small (see the figure), until it shrinks to a point. Since $\mathcal{C}$



shrinks to a point, $\oint_{\mathcal{C}} \mathbf{v} \cdot d\mathbf{R} \to 0$ and, by Stokes's theorem,

$$\int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \nabla \times \mathbf{v}\, dA \to 0 \tag{4.1}$$

as well. Show that the result (4.1) also follows from the divergence theorem.

5. The following claim was put forth in an examination paper: "The line integral $\oint_{\mathcal{C}} f(x, y)\, dx + g(x, y)\, dy$, where $\mathcal{C}$ is the clockwise unit circle, is necessarily zero because on $\mathcal{C}$ we can express y as a function of x so that [taking $(1, 0)$ as the initial point of $\mathcal{C}$ and also the terminal point of $\mathcal{C}$]

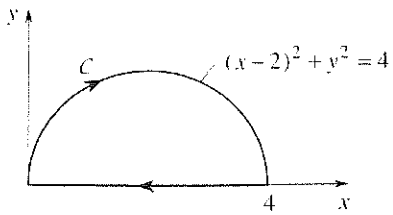
$$\oint_{\mathcal{C}} f(x, y)\, dx = \int_1^1 f(x, y(x))\, dx = \int_1^1 F(x)\, dx = 0$$

and, similarly,

$$\oint_{\mathcal{C}} g(x, y)\, dy = \int_0^0 g(x(y), y)\, dy = \int_0^0 G(y)\, dy = 0."$$

Give a critical assessment of that claim. Is it true? False? Explain your reasoning.

6. Given the field $\mathbf{F} = y^2\hat{\mathbf{i}}$ and the contour shown, the following



conflicting calculations were put forward in an exam paper:

$$\oint_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{R} = \oint_{\mathcal{C}} y^2\, dx$$
$$= \int_0^4 [4 - (x-2)^2]\, dx + \int_4^0 0\, dx = \frac{32}{3},$$

but

$$\oint_{\mathcal{C}} \mathbf{F} \cdot d\mathbf{R} = \int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \nabla \times \mathbf{F}\, dA$$
$$= \int_{\mathcal{S}} 2y\, dA = 2\int_0^4 y(y\, dx)$$
$$= 2\int_0^4 [4 - (x-2)^2]\, dx = \frac{64}{3}.$$

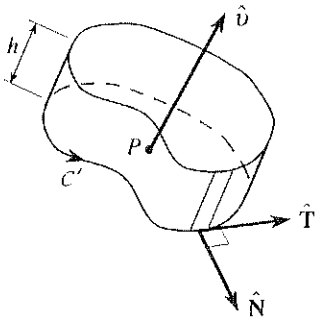
Find, and correct, the error.

7. (*An observation about Stokes's theorem*) Notice that the value of the line integral in Stokes's theorem (18) is independent of the shape of the surface $\mathcal{S}$. Thus, the surface integral in (18) must, similarly, be independent of the shape $\mathcal{S}$. That is, it must be true that $\int_{\mathcal{S}_1} \hat{\mathbf{n}} \cdot \nabla \times \mathbf{v}\, dA = \int_{\mathcal{S}_2} \hat{\mathbf{n}} \cdot \nabla \times \mathbf{v}\, dA$ for any two surfaces $\mathcal{S}_1$ and $\mathcal{S}_2$ that share the same edge curve $\mathcal{C}$. Provide an alternative argument as to why the two integrals (in the preceding sentence) must be equal. HINT: Note that $\mathcal{S}_1 + \mathcal{S}_2$ is a closed surface. Apply the divergence theorem to the surface integral of $\hat{\mathbf{n}} \cdot \nabla \times \mathbf{v}$ over that closed surface.

8. (*Heuristic proof of Stokes's theorem*) Our proof of Stokes's theorem was limited to the case where $\mathcal{S}$ is flat. This exercise is to indicate how to prove the theorem for a general surface $\mathcal{S}$ that is not necessarily flat. We use the limit definition of the curl,

$$\operatorname{curl} \mathbf{v}(P) \equiv \lim_{\mathcal{B} \to 0} \left\{ \frac{\int_{\mathcal{S}'} \hat{\mathbf{n}} \times \mathbf{v}\, dA}{V} \right\}, \tag{8.1}$$

given in Exercise 6 of Section 16.5. Let $\mathcal{B}$ be a right circular cylinder of a small height h, and let the base of $\mathcal{B}$ be a plane region of area A, bounded by a simple closed curve $\mathcal{C}'$, oriented so that the unit normal $\hat{\nu}$ to the base, and $\mathcal{C}'$, are in accordance with the right-hand rule as seen in the figure.



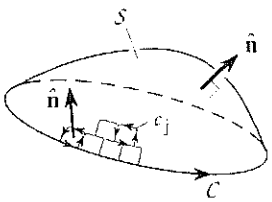
Dotting both sides of (8.1) with $\hat{\nu}$, and recalling that $\mathbf{A}\cdot\mathbf{B}\times\mathbf{C} = \mathbf{A}\times\mathbf{B}\cdot\mathbf{C}$, show that

$$\hat{\nu}\cdot\operatorname{curl}\mathbf{v} = \lim_{A\to 0}\left\{\frac{\int_{C'}\mathbf{v}\cdot d\mathbf{R}}{A}\right\}, \tag{8.2}$$

or, in differential form,

$$\hat{\nu}\cdot\operatorname{curl}\mathbf{v}\,dA \sim \int_{C'}\mathbf{v}\cdot d\mathbf{R} \tag{8.3}$$

as $A \to 0$. Now take the surface $\mathcal{S}$ in Stokes's theorem and partition it with a large number of such curves as C', as suggested in the figure, and number them as $C'_1, C'_2, \ldots$. Write



down (8.3) for each curve $(C'_1, C'_2, \ldots)$, realizing that $\hat{\nu}$ is the normal $\hat{\mathbf{n}}$ to $\mathcal{S}$. Add these equations and, noting internal cancellation analogous to that which occured in our proof of the divergence theorem, show that Stokes's theorem follows. [Note that the $\mathcal{S}'$ in (8.1) is the closed surface of the infinitesimal body $\mathcal{B}$, whereas the $\mathcal{S}$ in the figure above, is the open cap-like surface in Stokes's theorem.]

9. (*Faraday's law*) **Faraday's law** (that the emf around a closed curve equals the negative of the time rate of change of the magnetic flux through the curve) may be expressed as

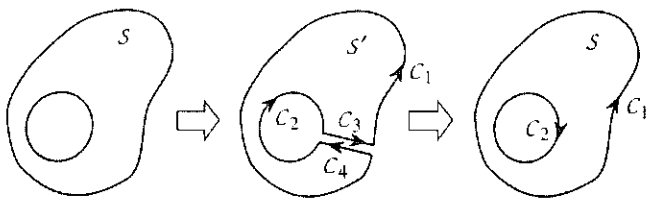
$$\begin{aligned}\oint_C \mathbf{E}\cdot d\mathbf{R} &= -\frac{d}{dt}\int_{\mathcal{S}}\mathbf{B}\cdot\hat{\mathbf{n}}\,dA\\ &= -\int_{\mathcal{S}}\frac{\partial\mathbf{B}}{\partial t}\cdot\hat{\mathbf{n}}\,dA\end{aligned} \tag{9.1}$$

for every (fixed) surface $\mathcal{S}$ with closed boundary curve C in the field, where $\mathbf{E}$ is the electric field intensity, $\mathbf{B}$ is the magnetic flux density, t is the time, and where the relative orientations of C and $\hat{\mathbf{n}}$ are the same as in Stokes's theorem. Applying Stokes's theorem to the line integral, use the arbitrariness of $\mathcal{S}$ and C to deduce (heuristically) that the relation

$$\nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t} \tag{9.2}$$

holds at all points in the field. Equation (9.2) is one of the *Maxwell's equations* for time-varying fields; for steady fields it reduces to $\nabla\times\mathbf{E} = 0$.

10. (*Surfaces with holes*) Here we indicate a modest extension of Stokes's theorem to the case where $\mathcal{S}$ has one or more holes in it. It should suffice to consider the surface $\mathcal{S}$ shown in the left-hand figure which, for simplicity, has only one hole in it and lies in the plane of the paper. Thus, the boundary C of $\mathcal{S}$ is not a simple closed curve, as is called for in Stokes's theorem.



(a) Slit the region as shown in the middle figure, and define the contours C_1, C_2, C_3, C_4 enclosing the region $\mathcal{S}'$. Since $\mathcal{S}'$ and $C = C_1 + C_2 + C_3 + C_4$ do satisfy the requirements given in Stokes's theorem, that theorem can now be applied. Letting the gap between C_3 and C_4 tend to zero, show that we obtain

$$\boxed{\int_{\mathcal{S}}\hat{\mathbf{n}}\cdot\nabla\times\mathbf{v}\,dA = \oint_{C_1}\mathbf{v}\cdot d\mathbf{R} + \oint_{C_2}\mathbf{v}\cdot d\mathbf{R},} \tag{10.1}$$

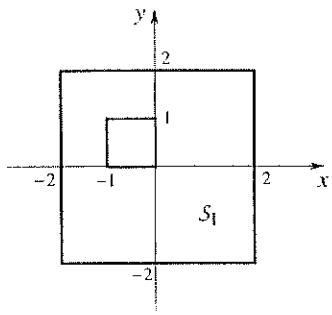
in accordance with the right-hand figure, where $\hat{\mathbf{n}}$ is directed out of the paper, toward the reader. NOTE: Since Green's theorem is but a special case of Stokes's theorem, a result analogous to (10.1) follows immediately for Green's theorem:

$$\boxed{\begin{aligned}\int_{\mathcal{S}}\left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right)dA &= \oint_{C_1} P\,dx + Q\,dy\\ &\quad + \oint_{C_2} P\,dx + Q\,dy.\end{aligned}} \tag{10.2}$$

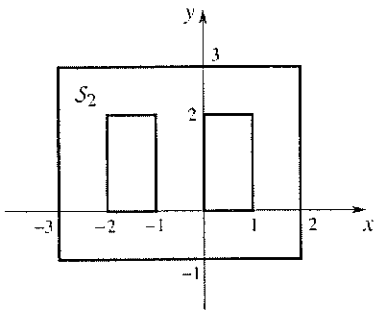
(b) Using pictures and reasoning, derive a formula analogous to (10.1) for the case where $\mathcal{S}$ has two holes rather than one.

11. Verify Stokes's theorem (10.1) above.

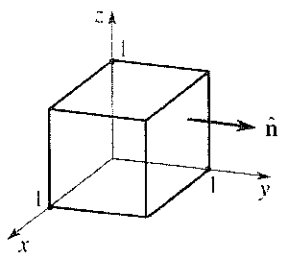
(a) $\mathbf{v} = y^3\hat{\mathbf{i}}$, $\mathcal{S}$ is the region $\mathcal{S}_1$ shown, and $\hat{\mathbf{n}} = \hat{\mathbf{k}}$.



(b) $\mathbf{v} = xy\hat{\mathbf{i}}$, $\mathcal{S}$ is the region $\mathcal{S}_2$ shown, and $\hat{\mathbf{n}} = -\hat{\mathbf{k}}$.



(c) $\mathbf{v} = x^3y(1+z)\hat{\mathbf{i}}+3\cos(x^2+z^2)\hat{\mathbf{k}}$, and $\mathcal{S}$ is the four-sided surface shown, open at the top and bottom.

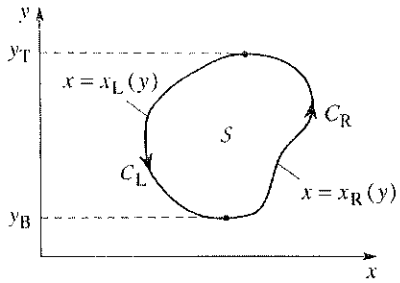


12. (*Direct proof of Green's theorem*) Green's theorem (33) can also be proved by treating the left-hand side as an iterated integral and integrating directly. Suppose that $\mathcal{S}$ is convex in both x and y. By **convex in** x, for instance, we mean that $\mathcal{S}$ can be described by the statement

$$x_L(y) \le x \le x_R(y) \quad \text{for} \quad y_B \le y \le y_T. \tag{12.1}$$

That is, every horizontal line from $\mathcal{C}_L$ to $\mathcal{C}_R$ (see figure) lies entirely within $\mathcal{S}$. Then

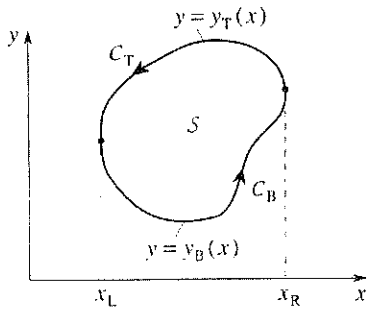
$$\int_{\mathcal{S}} \frac{\partial Q}{\partial x}\, dA = \int_{y_B}^{y_T} \int_{x_L(y)}^{x_R(y)} \frac{\partial Q}{\partial x}\, dx\, dy = \cdots = \oint_{\mathcal{C}} Q\, dy. \tag{12.2}$$



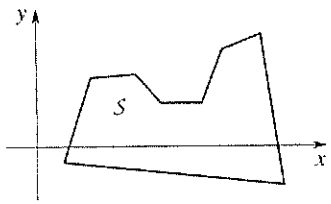
(a) Fill in the missing steps in (12.2).
(b) Next, show that

$$\int_{\mathcal{S}} \frac{\partial P}{\partial y}\, dA = \int_{x_L}^{x_R} \int_{y_B(x)}^{y_T(x)} \frac{\partial P}{\partial y}\, dy\, dx = \cdots = -\oint_{\mathcal{C}} P\, dx, \tag{12.3}$$

where $x_L, x_R, y_B(x)$ and $y_T(x)$ are shown below; (12.2) and (12.3) then give (33).



(c) Indicate how you could extend the convex-$\mathcal{S}$ result, proved above, to cover nonconvex regions such as the one shown below.



13. If $\mathcal{C}$ is a piecewise smooth simple closed curve in the x, y plane, oriented counterclockwise, show that the area A enclosed by $\mathcal{C}$ is given by the line integral

$$A = \frac{1}{2} \oint_{\mathcal{C}} (x\, dy - y\, dx). \tag{13.1}$$

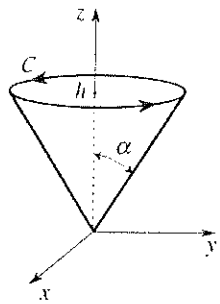
14. Use (13.1) in Exercise 13 to compute A if $\mathcal{C}$ is the boundary of the

(a) rectangle with vertices at $(0,0)$, $(a,0)$, (a,b), $(0,b)$
(b) triangle with vertices at $(a,0)$, (a,b), $(0,b)$
(c) circle $x^2+y^2=9$

15. Show that (33) does *not* hold for the case where $P = -y/(x^2+y^2)$, $Q = x/(x^2+y^2)$, and $\mathcal{S}$ is the unit disk $x^2+y^2 \le 1$. Explain why this failure does not violate Green's theorem.

16. Verify Stokes's theorem.

(a) $\mathbf{v} = \omega r\hat{\mathbf{e}}_\theta$ (ω = constant), $\mathcal{S}$: the waterglass surface consisting of the cylinder $r = a$ over $0 \le z \le h$, with flat base $z = 0$, $\hat{\mathbf{n}} = \hat{\mathbf{e}}_z$ on the base.
(b) $\mathbf{v} = 3r\hat{\mathbf{e}}_r - rz^2\hat{\mathbf{e}}_\theta + 5r^2\hat{\mathbf{e}}_z$, $\mathcal{S}$: the upside-down waterglass surface consisting of the cylinder $r = a$ over $0 \le z \le h$, with flat top $z = h$, $\hat{\mathbf{n}} = \hat{\mathbf{e}}_z$ on the face $z = h$.
(c) $\mathbf{v} = (50-2y)\hat{\mathbf{i}}$, $\mathcal{S}$: the flat disk $r \le 1$, $z = 0$, $\hat{\mathbf{n}} = \hat{\mathbf{e}}_z$.
(d) $\mathbf{v} = \omega r\hat{\mathbf{e}}_\theta$ (ω = constant), $\mathcal{S}$: the conical surface, shown below, open at the top, $\mathcal{C}$ oriented as shown.



(e) Repeat part (d), but with $\mathbf{v} = 3r\hat{\mathbf{e}}_r - r^{-1}\hat{\mathbf{e}}_\theta$.
(f) Repeat part (d), but with $\mathbf{v} = \rho^2 \sin\phi\hat{\mathbf{e}}_\theta$.

16.10 Irrotational Fields

16.10.1. Irrotational fields. Let $\mathbf{v}$ be a C^1 vector field in a domain $\mathcal{D}$.* If

$$\boxed{\nabla \times \mathbf{v} = \mathbf{0}} \tag{1}$$

at each point in $\mathcal{D}$, then $\mathbf{v}$ is said to be **irrotational** in $\mathcal{D}$. In Section 16.5 we saw that $\nabla \times \mathbf{v}$ can be interpreted physically as twice the fluid particle angular velocity at the point in question. Thus, if $\mathbf{v}$ is irrotational that means that the particles have no angular velocity, or "spin." For instance, the plane flow over a semicircular bump given by equation (3) in Section 16.2, and shown there in Fig. 4, was irrotational because $\nabla \times \mathbf{v} = \mathbf{0}$. In fact, irrotational flows form an important case in applications such as aerodynamics, water wave mechanics, and gravitational fields.

To study irrotational fields we need to define two more terms. First, we say that the line integral $\int_{\mathcal{C}} \mathbf{v} \cdot d\mathbf{R}$ between endpoints P and Q, is **path independent** if the value of the integral is the same for *every* piecewise smooth path $\mathcal{C}$ lying within $\mathcal{D}$; if it is path independent for *every* pair of endpoints P and Q in $\mathcal{D}$ then we say that it is path independent in $\mathcal{D}$.

Second, we need to add to our topological concept of connectedness, which was introduced in Section 13.2.2. A domain $\mathcal{D}$ is said to be **simply connected** if every closed curve in $\mathcal{D}$ can be shrunk, by a continuous deformation, to any point in $\mathcal{D}$. If it is not simply connected it is **multiply connected**. For instance, all of 3-space, the interior of a sphere or cube and the region between two concentric

*Recall from Section 13.2.2 that a domain is an open region, i.e., a connected set containing none of its boundary points.

spherical surfaces are simply connected, but the interior of a torus (i.e., a doughnut, bagel, or wedding band) is not; it is multiply connected. That is, whereas a closed curve such as $\mathcal{C}_1$ (Fig. 1) can be shrunk to any desired point within $\mathcal{D}$, a closed curve such as $\mathcal{C}_2$ cannot. In two dimensions, all of 2-space and the interior of a circle or square are simply connected, but the region between two concentric circles is not.

The focus of this section is the following theorem and its applications.

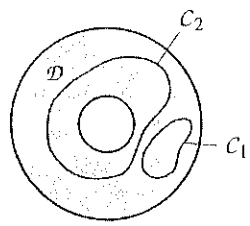


Figure 1. Torus.

THEOREM 16.10.1 *Irrotational Field*
If $\mathbf{v}$ is C^1 in a simply connected domain $\mathcal{D}$, then the following statements are equivalent (i.e., if one holds, then the other three do also).

(a) There exists a C^2 scalar function Φ in $\mathcal{D}$ such that $\mathbf{v} = \nabla\Phi$ throughout $\mathcal{D}$.

(b) $\nabla \times \mathbf{v} = \mathbf{0}$ throughout $\mathcal{D}$.

(c) $\int_{\mathcal{C}} \mathbf{v} \cdot d\mathbf{R} = 0$ for every piecewise smooth simple closed path $\mathcal{C}$ in $\mathcal{D}$.

(d) $\int_{\mathcal{C}} \mathbf{v} \cdot d\mathbf{R}$ is independent of path in $\mathcal{D}$.

Proof: Because of the claimed equivalence one might believe that we need to prove that (a) $\Rightarrow$ (b), (a) $\Rightarrow$ (c), (a) $\Rightarrow$ (d), (b) $\Rightarrow$ (a), (b) $\Rightarrow$ (c), (b) $\Rightarrow$ (d), and so on, twelve items altogether. No, it suffices to prove any closed logical loop, such as (a) $\Rightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (d) $\Rightarrow$ (a).

(a) $\Rightarrow$ (b): If $\mathbf{v} = \nabla\Phi = \Phi_x\hat{\mathbf{i}} + \Phi_y\hat{\mathbf{j}} + \Phi_z\hat{\mathbf{k}}$ (with subscripts denoting partial derivatives, as can be understood from the context), then $\nabla \times \mathbf{v} = (\Phi_{zy} - \Phi_{yz})\hat{\mathbf{i}} - (\Phi_{zx} - \Phi_{xz})\hat{\mathbf{j}} + (\Phi_{yx} - \Phi_{xy})\hat{\mathbf{k}} = \mathbf{0}$ because it follows from the assumption that Φ is C^2 that $\Phi_{yz} = \Phi_{zy}$, $\Phi_{zx} = \Phi_{xz}$, and $\Phi_{yx} = \Phi_{xy}$ (Theorem 13.3.1). Or, remember from (13) in Section 16.6 that the curl of a gradient is zero.

(b) $\Rightarrow$ (c): Let $\mathcal{C}$ be any piecewise smooth simple closed path within $\mathcal{D}$. Suppose that we can introduce a piecewise smooth surface $\mathcal{S}$ with $\mathcal{C}$ as its boundary. (Note that if $\mathcal{D}$ were not simply connected, $\mathcal{S}$ might have one or more holes in it, and then $\mathcal{C}$ would be only part of the boundary of $\mathcal{S}$. See Exercise 9 of Section 16.9.) Then, by Stokes's theorem,

$$\oint_{\mathcal{C}} \mathbf{v} \cdot d\mathbf{R} = \int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \nabla \times \mathbf{v}\, dA = \int_{\mathcal{S}} \hat{\mathbf{n}} \cdot \mathbf{0}\, dA = 0, \tag{2}$$

as claimed. However, given a piecewise smooth simple closed path $\mathcal{C}$, can we necessarily find a piecewise smooth surface $\mathcal{S}$ having $\mathcal{C}$ as its boundary, as assumed above? For the $\mathcal{C}$ shown in Fig. 2a, for example, such an $\mathcal{S}$ is readily constructed as sketched in the figure. But what about more complicated $\mathcal{C}$'s – for example, the overhand knot shown in Fig. 2a. In that case the idea is to add cancelling line segments, shown as dashed lines in Fig. 3, so that the curve $\mathcal{C}$ can be split into

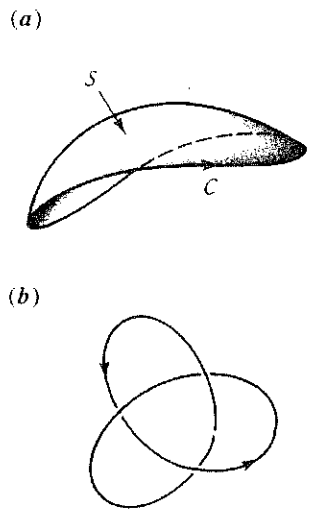


Figure 2. Is there an $\mathcal{S}$?

two noninterlocking curves $\mathcal{C}_1$ and $\mathcal{C}_2$ which *do* admit suitable surfaces $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively. Then, analogous to (1) we have

$$\oint_{\mathcal{C}} \mathbf{v}\cdot d\mathbf{R} = \oint_{\mathcal{C}_1} \mathbf{v}\cdot d\mathbf{R} + \oint_{\mathcal{C}_2} \mathbf{v}\cdot d\mathbf{R} = \int_{\mathcal{S}_1} \hat{\mathbf{n}}\cdot\nabla\times\mathbf{v}\,dA + \int_{\mathcal{S}_2} \hat{\mathbf{n}}\cdot\nabla\times\mathbf{v}\,dA$$
$$= \int_{\mathcal{S}_1} \hat{\mathbf{n}}\cdot\mathbf{0}\,dA + \int_{\mathcal{S}_2} \hat{\mathbf{n}}\cdot\mathbf{0}\,dA = 0+0=0.$$

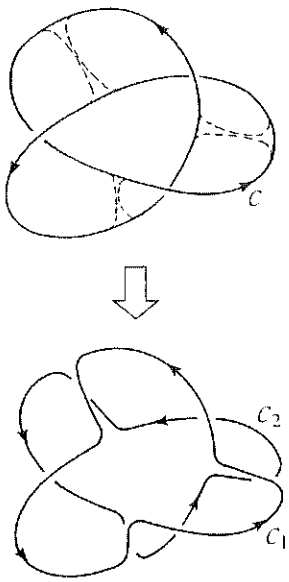


Figure 3. Undoing the knot.

Rigorous proof that we can handle *any* piecewise smooth simple closed path $\mathcal{C}$, however, is beyond our present scope.

(c) $\Rightarrow$ (d): Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be any two piecewise smooth paths from any point $P_0 = (x_0, y_0, z_0)$ in $\mathcal{D}$ to any point $P = (x, y, z)$ in $\mathcal{D}$ (Fig. 4a). Suppose that $\mathcal{C}_1$ and $\mathcal{C}_2$ are simple curves and that they intersect each other only at the endpoints P_0 and P. (The argument needed for the case where additional intersections occur will be omitted here.) Then $\mathcal{C}_1 + (-\mathcal{C}_2)$ is a piecewise smooth simple closed path (Fig. 4b) and, according to item (c),

$$\oint_{\mathcal{C}_1+(-\mathcal{C}_2)} \mathbf{v}\cdot d\mathbf{R} = 0. \tag{3}$$

But

$$\oint_{\mathcal{C}_1+(-\mathcal{C}_2)} \mathbf{v}\cdot d\mathbf{R} = \oint_{\mathcal{C}_1} \mathbf{v}\cdot d\mathbf{R} + \int_{-\mathcal{C}_2} \mathbf{v}\cdot d\mathbf{R}$$
$$= \int_{\mathcal{C}_1} \mathbf{v}\cdot d\mathbf{R} - \int_{\mathcal{C}_2} \mathbf{v}\cdot d\mathbf{R}, \tag{4}$$

and it follows from (3) and (4) that

$$\int_{\mathcal{C}_1} \mathbf{v}\cdot d\mathbf{R} = \int_{\mathcal{C}_2} \mathbf{v}\cdot d\mathbf{R}, \tag{5}$$

as claimed in (d).

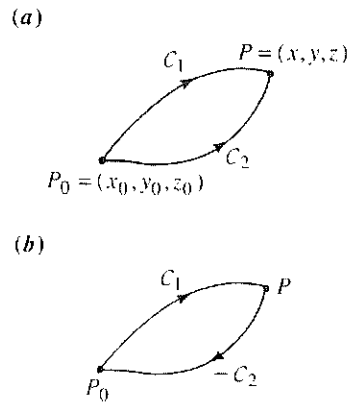


Figure 4. You take the high road, I'll take the low.

(d) $\Rightarrow$ (a): Regarding $P_0 = (x_0, y_0, z_0)$ as fixed and $P = (x, y, z)$ as a variable endpoint, it follows from the path independence [item (d)] that

$$f(x,y,z) \equiv \int_{(x_0,y_0,z_0)}^{(x,y,z)} \mathbf{v}\cdot d\mathbf{R} \tag{6}$$

is indeed a function of x, y, z. For if different paths to (x, y, z) were to give different values of the integral, then the integral would not be uniquely determined by x, y, z. That is, it would not be a single-valued function of x, y, z and, since functions are to be single-valued, it would not be a function of x, y, z.

Consider first the x dependence of the integral. Regarding y and z as fixed, take the path from (x_0, y_0, z_0) to (x, y, z) in (6) to be the one shown in Fig. 5. That

is, the path is fixed, except that the endpoint $P = (x, y, z)$ moves parallel to the x axis as x is varied, with y and z fixed. Then (6) becomes

$$\begin{aligned} f(x,y,z) &= \int_{(x_0,y_0,z_0)}^{(x_0,y,z)} \mathbf{v}\cdot d\mathbf{R} + \int_{(x_0,y,z)}^{(x,y,z)} \mathbf{v}\cdot d\mathbf{R} \\ &= \text{constant} + \int_{(x_0,y,z)}^{(x,y,z)} (v_x\,dx + v_y\,dy + v_z\,dz) \\ &= \text{constant} + \int_{x_0}^{x} v_x(x,y,z)\,dx, \end{aligned} \tag{7}$$

where $\mathbf{v} = v_x\hat{\mathbf{i}} + v_y\hat{\mathbf{j}} + v_z\hat{\mathbf{k}}$, and where the last step follows from the fact that y and z are constant along the path from (x_0, y, z) to (x, y, z). Or, introducing a dummy variable of integration to avoid confusion, we have

$$f(x,y,z) = \text{constant} + \int_{x_0}^{x} v_x(\xi,y,z)\,d\xi, \tag{8}$$

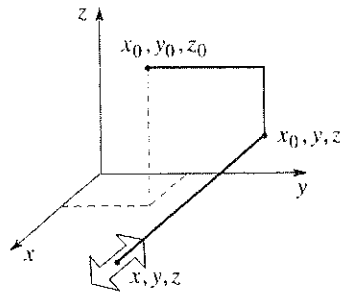


Figure 5. The x dependence of f.

and since v_x is assumed to be continuous it follows from the fundamental theorem of the integral calculus that $\partial f/\partial x = v_x(x, y, z)$. In the same manner, we may show that $\partial f/\partial y = v_y(x, y, z)$, and that $\partial f/\partial z = v_z(x, y, z)$. Thus, $\mathbf{v} = \nabla f$ and f is the scalar function that we've been looking for. ■

With Theorem 16.10.1 in hand, suppose we wish to evaluate a given line integral

$$I = \int_{\mathcal{C}} \mathbf{v}\cdot d\mathbf{R}. \tag{9}$$

First, check to see if $\mathbf{v}$ is irrotational (i.e., if $\nabla \times \mathbf{v} = \mathbf{0}$) in a simply connected domain $\mathcal{D}$ containing the path $\mathcal{C}$. If it is, and $\mathcal{C}$ is a closed path (piecewise smooth simple closed path, to be precise), then it follows immediately from item (c) that $I = 0$. If $\mathbf{v}$ is irrotational but $\mathcal{C}$ is not closed, then I is not necessarily zero, yet Theorem 16.10.1 still provides two alternative simplifications. First, we can use (d) to justify changing the path to a simpler one, keeping the endpoints fixed. Or we can use (a), which guarantees the existence of a scalar function (i.e., a scalar field) $\Phi(x, y, z)$ such that $\nabla\Phi = \mathbf{v}$. For if we can solve the relation $\nabla\Phi = \mathbf{v}$ for Φ, which function is known as the **scalar potential** or simply the **potential**, corresponding to the vector field $\mathbf{v}$, then

$$\begin{aligned} I = \int_{P_i}^{P_f} \mathbf{v}\cdot d\mathbf{R} &= \int_{P_i}^{P_f} \nabla\Phi\cdot d\mathbf{R} \\ &= \int_{P_i}^{P_f} \left(\frac{\partial\Phi}{\partial x}\hat{\mathbf{i}} + \frac{\partial\Phi}{\partial y}\hat{\mathbf{j}} + \frac{\partial\Phi}{\partial z}\hat{\mathbf{k}}\right)\cdot\left(dx\,\hat{\mathbf{i}} + dy\,\hat{\mathbf{j}} + dz\,\hat{\mathbf{k}}\right) \\ &= \int_{P_i}^{P_f} \left(\frac{\partial\Phi}{\partial x}dx + \frac{\partial\Phi}{\partial y}dy + \frac{\partial\Phi}{\partial z}dz\right) = \int_{P_i}^{P_f} d\Phi = \Phi\Big|_{P_i}^{P_f}, \end{aligned} \tag{10}$$

where P_i is the initial point and P_f is the final point. That is,

$$\boxed{\int_{P_i}^{P_f} \mathbf{v} \cdot d\mathbf{R} = \Phi\Big|_{P_i}^{P_f},} \tag{11}$$

the line integral is equal to the change in the potential.

Both of these lines of approach are illustrated in the example to follow.

EXAMPLE 1. Evaluate the line integral $I = \int_C \mathbf{v} \cdot d\mathbf{R}$, where

$$\mathbf{v} = 3x^2y^2\hat{\mathbf{i}} + (2x^3y - e^z)\hat{\mathbf{j}} + (2z - ye^z)\hat{\mathbf{k}}, \tag{12}$$

and where C is given parametrically by

$$\begin{aligned} x &= \left(1 - \frac{\tau}{2}\right) e^{\sqrt{\tau}} - 2\frac{\sin \tau^3}{\sin 8} \sqrt[3]{\frac{\tau}{2}}, \\ y &= -\frac{6}{\tau + 3} + \frac{21}{10}\tau, \\ z &= \frac{1}{16}\tau^5 - 1 \end{aligned} \tag{13}$$

for $\tau : 0 \to 2$. We could, of course, put the parametrization (13) into

$$\int_C \mathbf{v} \cdot d\mathbf{R} = \int_C 3x^2y^2\,dx + (2x^3y - e^z)\,dy + (2z - ye^z)\,dz. \tag{14}$$

That step would give an "ordinary" integral of the form $\int_0^2 F(\tau)\,d\tau$. However, the integrand $F(\tau)$ would, evidently, be quite unwieldy. Thus, it is best to check, first, to see if $\mathbf{v}$ is irrotational. In fact, we do find that $\nabla \times \mathbf{v} = \mathbf{0}$ so we can use either of the methods outlined above. Thus, we can discard the unwieldy path (13); all we need to extract from (13) are the initial and final points. Setting $\tau = 0$ and 2 in (13) gives $P_i = (1, -2, -1)$ and $P_f = (-2, 3, 1)$.

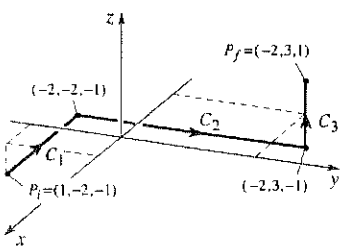


Figure 6. A simple path.

Let us evaluate I first by path simplification. What *is* a simple path? Surely, a straight line from P_i to P_f comes to mind. No, the simplest path will be one along which only one coordinate varies. Specifically, let us use the path C_1, C_2, C_3 shown in Fig. 6. Then

$$\begin{aligned} I &= \int_{C_1} \mathbf{v} \cdot d\mathbf{R} + \int_{C_2} \mathbf{v} \cdot d\mathbf{R} + \int_{C_3} \mathbf{v} \cdot d\mathbf{R} \\ &= \int_{C_1} 3x^2y^2\,dx\Big|_{y=-2,z=-1} + \int_{C_2} (2x^3y - e^z)\,dy\Big|_{x=-2,z=-1} \\ &\quad + \int_{C_3} (2z - ye^z)\,dz\Big|_{x=-2,y=3} \\ &= 12\int_1^{-2} x^2\,dx + \int_{-2}^{3} (-16y - e^{-1})\,dy + \int_{-1}^{1} (2z - 3e^z)\,dz \\ &= -76 - 3e - 2e^{-1}. \end{aligned} \tag{15}$$

To appreciate the simplification achieved, observe that in the $\mathcal{C}_1$ integral we kept only the dx part because y and z are constant on $\mathcal{C}_1$ (hence the dy's and dz's are zero). Further, the integrand of the dx integral simplifies because any y's or z's contained therein are constant on $\mathcal{C}_1$. Similarly for the $\mathcal{C}_2$ and $\mathcal{C}_3$ integrals.

Alternatively, let us evaluate I using the potential Φ, the existence of which is assured by the fact that $\nabla \times \mathbf{v} = \mathbf{0}$. To find $\Phi(x, y, z)$ we use the connection $\mathbf{v} = \nabla\Phi$ between $\mathbf{v}$ and Φ:

$$\frac{\partial \Phi}{\partial x} = v_x = 3x^2y^2, \tag{16a}$$

$$\frac{\partial \Phi}{\partial y} = v_y = 2x^3y - e^z, \tag{16b}$$

$$\frac{\partial \Phi}{\partial z} = v_z = 2z - ye^z. \tag{16c}$$

We need to integrate (16a,b,c) to solve for Φ. First, integrate (16a) with respect to x, holding y and z fixed (since y and z were held fixed in the derivative $\partial\Phi/\partial x$) :

$$\Phi(x, y, z) = \int 3x^2y^2\,\partial x = x^3y^2 + A(y, z), \tag{17}$$

where the ∂x notation is not standard in the literature but may be helpful in reminding us that y and z are fixed. We need to allow the integration constant A to depend on y and z; observe that $\partial/\partial x$ of (17) does give us back (16a).

Next, put (17) into the left-hand side of (16b):

$$2x^3y + \frac{\partial A}{\partial y} = 2x^3y - e^z. \tag{18}$$

Cancelling the $2x^3y$ terms and integrating partially on y gives

$$A(y, z) = -\int e^z\,\partial y = -ye^z + B(z), \tag{19}$$

where we let the integration constant B depend on z because z was held fixed in the partial integration on y. Putting (19) into (17) gives the updated expression for Φ,

$$\Phi(x, y, z) = x^3y^2 - ye^z + B(z). \tag{20}$$

Finally, putting (20) into the left-hand side of (16c) gives

$$0 - ye^z + B'(z) = 2z - ye^z. \tag{21}$$

Cancelling the $-ye^z$ terms and integrating on z gives

$$B(z) = \int 2z\,dz = z^2 + C, \tag{22}$$

where we use dz rather than ∂z because it is an ordinary integral on z, and the integration constant C really is just a constant. Thus,

$$\Phi(x, y, z) = x^3y^2 - ye^z + z^2 + C. \tag{23}$$

Then (11) gives

$$\int_C \mathbf{v} \cdot d\mathbf{R} = (x^3y^2 - ye^z + z^2 + C)\Big|_{(1,-2,-1)}^{(-2,3,1)} = -76 - 3e - 2e^{-1}, \tag{24}$$

in agreement with (15).

COMMENT 1. There will always occur an arbitrary additive constant in Φ, just as C occurs in (23), for an additive constant drops out when we ask $\nabla\Phi$ to equal $\mathbf{v}$. However, when Φ is evaluated between the endpoints C inevitably cancels out so one can simply set $C = 0$, say, when it first appears.

COMMENT 2. One might wonder how the procedure of integrating $\nabla\Phi = \mathbf{v}$ can *fail* to produce a scalar potential $\Phi(x, y, z)$, whether or not $\mathbf{v}$ is irrotational. To understand this point, recall the cancellation of the $2x^3y$ terms. If those terms did not cancel, then (18) would have presented a logical contradiction for $\partial A/\partial y$ would therefore have depended on x, whereas it is a function only of y and z. Similarly, in (21) the cancellation of the $-ye^z$ terms was essential because otherwise $B'(z)$ would have depended on y. These cancellations were not accidents; they occurred because $\mathbf{v}$ was irrotational.

COMMENT 3. An advantage of the potential method over path simplification is that with $\Phi(x, y, z)$ in hand we can now compute $\int_{P_i}^{P_f} \mathbf{v} \cdot d\mathbf{R}$ between *any* two points P_i and P_f, from (11). ∎

EXAMPLE 2. *Conservative Force Fields.* A force field $\mathbf{F}$ defined in a domain $\mathcal{D}$ is said to be **conservative** *if the work*

$$W = \oint_C \mathbf{F} \cdot d\mathbf{R} \tag{25}$$

is zero for every piecewise smooth simple closed path C in $\mathcal{D}$. If $\mathbf{F}$ is C^1 throughout $\mathcal{D}$, then it follows from Theorem 16.10.1 that $\mathbf{F}$ is conservative if and only if it is irrotational. If so, there exists a scalar potential Φ such that

$$\mathbf{F} = -\nabla\Phi, \tag{26}$$

where the minus sign is customary but not essential.

To illustrate, consider the uniform gravitational field $\mathbf{F} = -g\hat{\mathbf{k}}$ per unit mass, where g is the acceleration due to gravity. Surely $\nabla \times \mathbf{F} = \mathbf{0}$ since $\mathbf{F}$ is a constant. Then $\mathbf{F} = -g\hat{\mathbf{k}} = -\nabla\Phi$ so that

$$\frac{\partial\Phi}{\partial x} = 0, \qquad \frac{\partial\Phi}{\partial y} = 0, \qquad \frac{\partial\Phi}{\partial z} = g.$$

Integrating gives us

$$\Phi = gz + C, \tag{27}$$

which is the familiar "gravitational potential energy" from elementary physics.

Returning to the general case, suppose that $\mathbf{F}$ is conservative, and consider the work $\int \mathbf{F} \cdot d\mathbf{R}$ over any path from an initial point P_i to a final point P_f. With $\mathbf{F} = m\ddot{\mathbf{R}}$ from

Newton's second law, where m is the mass (assumed constant), we have

$$\begin{aligned}\int_{P_i}^{P_f} \mathbf{F}\cdot d\mathbf{R} &= \int_{P_i}^{P_f} m\ddot{\mathbf{R}}\cdot d\mathbf{R} = m\int \ddot{\mathbf{R}}\cdot\dot{\mathbf{R}}\,dt \\ &= \frac{1}{2}m\int \frac{d}{dt}(\dot{\mathbf{R}}\cdot\dot{\mathbf{R}})\,dt = \frac{1}{2}m\int d(\dot{\mathbf{R}}\cdot\dot{\mathbf{R}}) \\ &= \frac{1}{2}m(\dot{\mathbf{R}}\cdot\dot{\mathbf{R}})\Big|_{P_i}^{P_f} = \frac{1}{2}m\left\|\dot{\mathbf{R}}\right\|^2\Big|_{P_i}^{P_f}. \end{aligned} \tag{28}$$

But since $\mathbf{F}$ is conservative, we may also express

$$\int_{P_i}^{P_f} \mathbf{F}\cdot d\mathbf{R} = \int_{P_i}^{P_f} -\nabla\Phi\cdot d\mathbf{R} = -\int_{P_i}^{P_f} d\Phi = -\Phi\Big|_{P_i}^{P_f}. \tag{29}$$

It follows from (28) and (29) that $\frac{1}{2}m\left\|\dot{\mathbf{R}}\right\|^2\Big|_{P_i}^{P_f} = -\Phi\Big|_{P_i}^{P_f}$, or

$$\left(\frac{1}{2}m\left\|\dot{\mathbf{R}}\right\|^2 + \Phi\right)\Big|_{P_i}^{P_f} = 0. \tag{30}$$

That is, there is no change in $\frac{1}{2}m\left\|\dot{\mathbf{R}}\right\|^2 + \Phi$:

$$\frac{1}{2}m\left\|\dot{\mathbf{R}}\right\|^2 + \Phi = \text{constant}, \tag{31}$$

or

$$\text{kinetic energy} + \text{"potential energy"} = \text{constant}, \tag{32}$$

and (32) explains why such a force field is called conservative because the total energy (kinetic plus potential) is conserved. Now we see the motivation for including the minus sign in (26), namely, so we end up with plus signs in (30)–(32). ∎

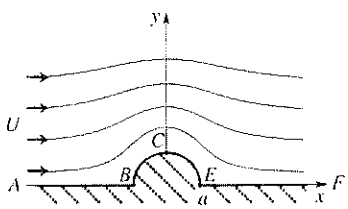


Figure 7. Flow over a building.

EXAMPLE 3. *Irrotational Fluid Flow.* Consider a building with a semicircular cross section. A cross wind, say of speed U (Fig. 7), will cause a lift force on the building (i.e., in the positive y direction) which can be quite large and which must be taken into account in the structural design. If we know the air velocity field $\mathbf{v}$ then we can use the Bernoulli equation of fluid mechanics (Exercise 12) to compute the pressure field, in particular the pressure distribution on the roof, integration of which gives the lift force. Here, we will limit our discussion to showing how to set up the field equation governing the velocity field $\mathbf{v}$.

Upstream (i.e., as $x \to -\infty$) the flow is simply the undisturbed free stream $U\hat{\mathbf{i}}$. Since $\nabla \times U\hat{\mathbf{i}} = 0$, we see that the flow is irrotational upstream. Since $\nabla \times \mathbf{v}$ is (Section 16.5) twice the fluid angular velocity, it follows that the fluid particles initially (i.e., upstream) have no spin; they undergo pure translation. Will they acquire angular velocity as they move downstream and pass over the semicircle? Consider the stresses acting on a typical fluid particle (Fig. 8). Just as an imbalance of horizontal or vertical normal stresses will, according to Newton's second law, cause a horizontal or vertical acceleration, respectively, an imbalance of shearing stresses results in a nonzero torque and hence an angular acceleration. But shearing stresses are possible only by virtue of the fluid property known as

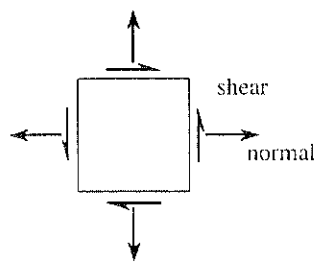


Figure 8. The stresses on a fluid particle.

viscosity. We propose that air has sufficiently small viscosity for us, to a good approximation, to consider it as inviscid, that is, having no viscosity at all. With no viscosity there are no shearing stresses, hence no torque, hence no angular acceleration. Thus, having started out as irrotational (i.e., upstream) the flow will remain irrotational throughout: not only does $\nabla \times \mathbf{v} \to \mathbf{0}$ as $x \to -\infty$, $\nabla \times \mathbf{v} = \mathbf{0}$ everwhere in the field. Consequently, there exists a scalar potential Φ, called the *velocity potential* in this application, such that

$$\mathbf{v} = \nabla\Phi. \tag{33}$$

What physics has thus far been left out? We have used Newton's second law and have also made an assumption on a relevant material property, the viscosity. Also relevant is the physical law of conservation of mass, which (see Example 2 in Section 16.8) can be expressed as the "continuity equation"

$$\frac{\partial\sigma}{\partial t} + \nabla\cdot(\sigma\mathbf{v}) = 0, \tag{34}$$

where σ is the fluid mass density field. If, besides assuming the fluid to be inviscid we also consider it to be incompressible, then the density σ is a constant. In that case (34) reduces to $0 + \sigma\nabla\cdot\mathbf{v} = 0$, or

$$\nabla\cdot\mathbf{v} = 0. \tag{35}$$

Finally, putting (33) into (35) gives

$$\boxed{\nabla^2\Phi = 0} \tag{36}$$

so the field equation governing Φ is the famous *Laplace equation* (36).

The remainder of this example is optional since it draws upon material contained in the optional Section 16.7.

In what coordinate system should we express (36)? The uniform free stream $U\hat{\mathbf{i}}$ is expressed most readily in terms of Cartesian coordinates, but the semicircular shape suggests using polar coordinates. When we return to this problem in a later chapter, and solve it by a method of separation of variables, we will see that we need to work in polar coordinates so let us make that choice here. We consider the region to be bounded by the radial lines $\theta = 0$ (EF in Fig. 9) and $\theta = \pi$ (AB), and the semicircles $r = a$ (BCE) and $r = R$ (AGF). Since the portion $ABCEF$ is rigid, we assume that the flow is *along* that part of the boundary, neither penetrating it nor separating from it; that is, $\hat{\mathbf{n}}\cdot\mathbf{v} = \hat{\mathbf{n}}\cdot\nabla\Phi = \partial\Phi/\partial n = 0$, where the second equality is simply the directional derivative formula (7) in Section 16.4, and $\hat{\mathbf{n}}$ is the unit outward normal to $\mathcal{D}$. On BCE and on EF we have

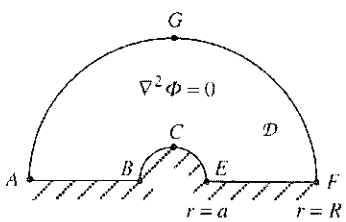


Figure 9. The domain $\mathcal{D}$.

$$\begin{aligned} BCE: &\quad \hat{\mathbf{n}}\cdot\nabla\Phi = -\hat{\mathbf{e}}_r\cdot\left(\frac{\partial\Phi}{\partial r}\hat{\mathbf{e}}_r + \frac{1}{r}\frac{\partial\Phi}{\partial\theta}\hat{\mathbf{e}}_\theta\right) = -\frac{\partial\Phi}{\partial r} = 0, \\ EF: &\quad \hat{\mathbf{n}}\cdot\nabla\Phi = -\hat{\mathbf{e}}_\theta\cdot\left(\frac{\partial\Phi}{\partial r}\hat{\mathbf{e}}_r + \frac{1}{r}\frac{\partial\Phi}{\partial\theta}\hat{\mathbf{e}}_\theta\right) = -\frac{1}{r}\frac{\partial\Phi}{\partial\theta} = 0 \end{aligned} \tag{37}$$

so $\partial\Phi/\partial r = 0$ on BCE, and $\partial\Phi/\partial\theta = 0$ on EF. Similarly, $\partial\Phi/\partial\theta = 0$ on AB.

On the semicircle AGF the boundary condition is that $\mathbf{v} \sim U\hat{\mathbf{i}}$ as $R \to \infty$. That is,

$$\nabla\Phi = \frac{\partial\Phi}{\partial x}\hat{\mathbf{i}} + \frac{\partial\Phi}{\partial y}\hat{\mathbf{j}} \sim U\hat{\mathbf{i}}$$

so $\partial\Phi/\partial x \sim U$ and $\partial\Phi/\partial y \to 0$ as $R \to \infty$.* Thus, $\Phi \sim Ux$, or expressing this result in terms of polar coordinates, $\Phi \sim Ur\cos\theta$ as $r \to \infty$. The resulting problem governing $\Phi(r,\theta)$ is as follows,

$$\begin{aligned}
&\nabla^2\Phi = \frac{\partial^2\Phi}{\partial r^2} + \frac{1}{r}\frac{\partial\Phi}{\partial r} + \frac{1}{r^2}\frac{\partial^2\Phi}{\partial\theta^2} = 0 \qquad (a < r < \infty,\ \ 0 < \theta < \pi)\\
&\frac{\partial\Phi}{\partial\theta}(r,0) = \frac{\partial\Phi}{\partial\theta}(r,\pi) = 0 \qquad (a < r < \infty)\\
&\frac{\partial\Phi}{\partial r}(a,\theta) = 0 \qquad (0 < \theta < \pi)\\
&\Phi(r,\theta) \sim Ur\cos\theta \qquad \text{as } r \to \infty, \qquad (0 < \theta < \pi)
\end{aligned} \tag{38}$$

and is solved within Chapter 20.

COMMENT 1. More generally, any irrotational incompressible flow is governed by a Laplace equation $\nabla^2\Phi = 0$ on the velocity potential Φ, and is known as a potential flow.

COMMENT 2. How can we quantify the accuracy of our two assumptions – that the fluid is inviscid and incompressible? One learns, in a course on fluid mechanics, that the accuracy of these assumptions rests on the size of two well known nondimensional parameters: the effects of viscosity will be negligible if the *Reynolds number* is much greater than unity, and the effects of compressibility will be negligible if the square of the *Mach number* is much less than unity.

COMMENT 3. It would be natural to wonder if the velocity potential Φ admits a simple physical significance. Such significance can in fact be attached to Φ in terms of so-called impulsive pressures, discussion of which can be found in the book by Wilson.* ∎

16.10.2. Non-Cartesian coordinates. (Optional) Let us limit our coverage of the non-Cartesian case to one example.

EXAMPLE 4. *Cylindrical Coordinates.* Evaluate the line integral $\int_C \mathbf{v}\cdot d\mathbf{R}$, where

$$\mathbf{v} = 3r^2 z\sin\theta\,\hat{\mathbf{e}}_r + r^2 z\cos\theta\,\hat{\mathbf{e}}_\theta + (r^3\sin\theta - 3z^2)\hat{\mathbf{e}}_z \tag{39}$$

and where C is a straight line from $P_i = (x_i, y_i, z_i) = (2,0,1)$ to $P_f = (x_f, y_f, z_f) = (1,1,0)$. Recalling that in cylindrical coordinates $d\mathbf{R} = dr\,\hat{\mathbf{e}}_r + r\,d\theta\,\hat{\mathbf{e}}_\theta + z\,\hat{\mathbf{e}}_z$, we have

$$\int_C \mathbf{v}\cdot d\mathbf{R} = \int_C 3r^2 z\sin\theta\,dr + r^3 z\cos\theta\,d\theta + (r^3\sin\theta - 3z^2)\,dz. \tag{40}$$

Using the cylindrical coordinate expression for $\nabla\times\mathbf{v}$, we find that $\nabla\times\mathbf{v} = \mathbf{0}$ so we need not evaluate the integral directly; we can use either path simplification or the potential method.

*$f(x) \sim g(x)$ as $x \to x_0$ means that $f(x)/g(x) \to 1$ as $x \to x_0$. We never write $f(x) \sim 0$ as $x \to x_0$ because $f(x)/0$ cannot tend to 1; it is better notation to write $f(x) \to 0$ rather than $f(x) \sim 0$. That is why we wrote $\partial\Phi/\partial y \to 0$ rather than $\partial\Phi/\partial y \sim 0$.

*D. H. Wilson, *Hydrodynamics* (London: Edward Arnold, 1959), Sec.10.

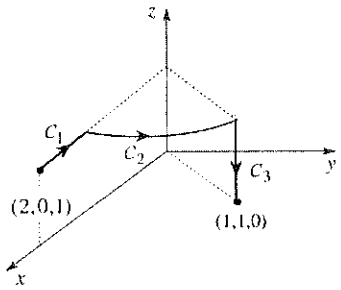


Figure 10. The modified path.

Since C is already a straight line, isn't it already simple? No, we can do better for the simplest path is made up of one or more segments on which only one coordinate varies. Specifically, let us use the path C_1, C_2, C_3 shown in Fig. 10. On C_1 only r varies (from 2 to $\sqrt{2}$) while θ and z are fixed ($\theta = 0$ and $z = 1$); on C_2 only θ varies (from 0 to $\pi/4$) while r and z are fixed ($r = \sqrt{2}$ and $z = 1$); and on C_3 only z varies (from 1 to 0) while r and θ are fixed ($r = \sqrt{2}$ and $\theta = \pi/4$) so

$$\begin{aligned}\int_C \mathbf{v}\cdot d\mathbf{R} &= \int_{C_1} 3r^2 z\sin\theta\, dr\Big|_{\theta=0,z=1} + \int_{C_2} r^3 z\cos\theta\, d\theta\Big|_{r=\sqrt{2},z=1} \\ &\quad + \int_{C_3} (r^3\sin\theta - 3z^2)\, dz\Big|_{r=\sqrt{2},\theta=\pi/4} \\ &= \int_2^{\sqrt{2}} 0\, dr + \int_0^{\pi/4} 2\sqrt{2}\cos\theta\, d\theta + \int_1^0 (2-3z^2)\, dz \\ &= 1. \end{aligned} \tag{41}$$

Alternatively, let us use the potential method. Write out the r, θ, z components of $\mathbf{v} = \nabla\Phi$:

$$\frac{\partial\Phi}{\partial r} = 3r^2 z\sin\theta, \tag{42a}$$

$$\frac{1}{r}\frac{\partial\Phi}{\partial\theta} = r^2 z\cos\theta, \tag{42b}$$

$$\frac{\partial\Phi}{\partial z} = r^3\sin\theta - 3z^2. \tag{42c}$$

Integrating (42a) gives

$$\Phi(r,\theta,z) = \int 3r^2 z\sin\theta\, \partial r = r^3 z\sin\theta + A(\theta,z), \tag{43}$$

and putting (43) into (42b) gives

$$\frac{1}{r}\left(r^3 z\cos\theta + \frac{\partial A}{\partial\theta}\right) = r^2 z\cos\theta$$

so that $\partial A/\partial\theta = 0$. Thus $A(\theta, z)$ does not vary with θ so it can be expressed as a function of z only, say $B(z)$. Then (43) becomes $\Phi(r,\theta,z) = r^3 z\sin\theta + B(z)$, and putting this expression into (42c) gives $r^3\sin\theta + B'(z) = r^3\sin\theta - 3z^2$. Thus, $B'(z) = -3z^2$ so $B(z) = -z^3 + C$, where C is an arbitrary constant. The result is that

$$\Phi(r,\theta,z) = r^3 z\sin\theta - z^3 + C \tag{44}$$

where we can, without loss, set $C = 0$, say. Finally,

$$\int_C \mathbf{v}\cdot d\mathbf{R} = \Phi\Big|_{P_i}^{P_f} = (r^3 z\sin\theta - z^3)\Big|_{r=2,\theta=0,z=1}^{r=\sqrt{2},\theta=\pi/4,z=0} = 1, \tag{45}$$

in agreement with (41). ∎

EXAMPLE 5. *Role of Simple Connectedness.* Our purpose in this final example is to indicate the significance of the requirement, in Theorem 16.10.1, that the domain $\mathcal{D}$ be

simply connected. Consider the vector field

$$\mathbf{H} = \frac{I}{2\pi r}\,\hat{\mathbf{e}}_\theta, \tag{46}$$

which is the magnetic field intensity induced in 3-space by a current I flowing in a wire which extends from $z = -\infty$ to $z = +\infty$ along the z axis (Fig. 11). Since $H_r = 0$, $H_z = 0$, and $H_\theta = I/2\pi r$ is a function of r only, (17) in Section 16.7 gives

$$\begin{aligned}\nabla \times \mathbf{H} &= \frac{1}{r}\frac{\partial}{\partial r}(rH_\theta)\,\hat{\mathbf{e}}_z \\ &= \frac{1}{r}\frac{\partial}{\partial r}\left(\frac{I}{2\pi}\right)\hat{\mathbf{e}}_z = \frac{1}{r}(0)\,\hat{\mathbf{e}}_z = \mathbf{0}\end{aligned} \tag{47}$$

so the field is irrotational. Yet,

$$\begin{aligned}\oint_{\mathcal{C}_1} \mathbf{H}\cdot d\mathbf{R} &= \oint_{\mathcal{C}_1} \frac{I}{2\pi r}\,\hat{\mathbf{e}}_\theta \cdot (dr\,\hat{\mathbf{e}}_r + r\,d\theta\,\hat{\mathbf{e}}_\theta + dz\,\hat{\mathbf{e}}_z) \\ &= \oint_{\mathcal{C}_1} \frac{I}{2\pi}\,d\theta = \frac{I}{2\pi}\,2\pi = I\end{aligned} \tag{48}$$

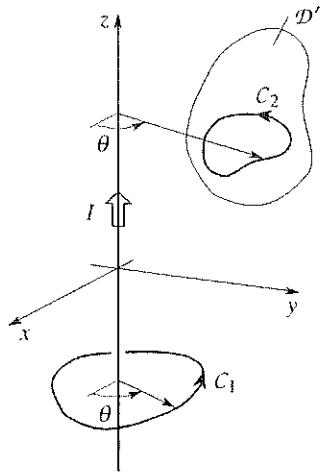


Figure 11. Current-carrying wire.

is nonzero, in apparent contradiction of Theorem 16.10.1. However, observe that $\mathcal{C}_1$ does not lie within a *simply connected* domain $\mathcal{D}$ throughout which $\mathbf{H}$ is C^1 and $\nabla \times \mathbf{H} = \mathbf{0}$. That is, $\mathbf{H} = (I/2\pi r)\,\hat{\mathbf{e}}_\theta$ is not even defined along the z axis ($r = 0$), let alone C^1 there. Similarly, the last equality in (47) holds only for $r \neq 0$. Thus, $\mathbf{H}$ is C^1 and $\nabla \times \mathbf{H} = \mathbf{0}$ in a region $\mathcal{D}$ consisting of all 3-space but with the z axis excluded, that is, the annulus $0 < r < \infty$. That domain is not simply connected so the result (48) does not violate the theorem.

Understand, in (48), that $\oint_{\mathcal{C}_1} d\theta = 2\pi$ because θ increases by 2π as $\mathcal{C}_1$ is traversed: $\theta_f = \theta_i + 2\pi$. However, for any curve that does not encircle the z axis, such as $\mathcal{C}_2$ (Fig. 11), $\theta_f = \theta_i$ so

$$\oint_{\mathcal{C}_2} \mathbf{H}\cdot d\mathbf{R} = \oint_{\mathcal{C}_2} \frac{I}{2\pi}\,d\theta = 0. \tag{49}$$

The result (49) follows from Theorem 16.10.1 because there *is* a simply connected domain $\mathcal{D}'$ (Fig. 11) containing $\mathcal{C}_2$, throughout which $\mathbf{H}$ is C^1 and irrotational. That is, $\mathcal{D}'$ is not pierced by the z axis. ∎

Closure. The focal point of this section is the irrotational field theorem, Theorem 16.10.1, in which each of the four stated items implies the other three. However, in applications the direction of inference is usually as follows: if $\mathbf{v}$ is irrotational, then there exists a scalar potential Φ such that $\mathbf{v} = \nabla\Phi$, the line integral $\oint_{\mathcal{C}} \mathbf{v}\cdot d\mathbf{R}$ around every closed path is zero, and the line integral $\int_{\mathcal{C}} \mathbf{v}\cdot d\mathbf{R}$ along every open path $\mathcal{C}$ is path independent and depends only upon the endpoints of $\mathcal{C}$. In particular, we show that if $\mathcal{C}$ is open, then the integral is simply the change in the potential between the endpoints P_i and P_f:

$$\int_{\mathcal{C}} \mathbf{v}\cdot d\mathbf{R} = \Phi\Big|_{P_i}^{P_f}. \tag{50}$$

Thus, if $\mathbf{v}$ is irrotational (and C^1 in a simply connected domain $\mathcal{D}$) then we can compute the line integral $\int_{\mathcal{C}} \mathbf{v} \cdot d\mathbf{R}$ along an open path $\mathcal{C}$ in $\mathcal{D}$ either by simplifying the path or by finding the potential function Φ and using (50). To find Φ, we integrate the three scalar components of $\mathbf{v} = \nabla\Phi$.

An important advantage of working with Φ, rather than $\mathbf{v}$, is that it is a single scalar function, whereas $\mathbf{v}$ is comprised of three scalar functions, namely, its three components. That scalar potentials are of great importance is witnessed by the fact that many of them are so well known, for instance, the gravitational potential induced by a distribution of mass, the electric potential (i.e., the voltage) induced by a distribution of charge, the entropy in thermodynamics (introduced in the exercises), and the velocity potential of fluid mechanics. In terms of these potentials we have such important results as the statement of conservation of energy (32) for conservative force fields and the Laplace equation (36) governing irrotational incompressible fluid flow.

EXERCISES 16.10

1. In what domain(s) are $\mathbf{v}$ both C^1 and irrotational?

(a) $\mathbf{v} = x^2\hat{\mathbf{i}} - 3y\hat{\mathbf{j}} + z^3\hat{\mathbf{k}}$
(b) $\mathbf{v} = 2\sin x\hat{\mathbf{i}} + y^3\hat{\mathbf{j}} + x\hat{\mathbf{k}}$
(c) $\mathbf{v} = e^x\hat{\mathbf{i}} - (2/y)\hat{\mathbf{j}} + z^2\hat{\mathbf{k}}$
(d) $\mathbf{v} = (x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}})/(x^2 + y^2 + z^2)$
(e) $\mathbf{v} = (-y\hat{\mathbf{i}} + x\hat{\mathbf{j}})/(x^2 + y^2)$
(f) $\mathbf{v} = x^3\hat{\mathbf{i}} - (z\hat{\mathbf{j}} - y\hat{\mathbf{k}})/(y^2 + z^2)$
(g) $\mathbf{v} = z\hat{\mathbf{i}} + y\hat{\mathbf{j}} - x\hat{\mathbf{k}}$

2. Show whether or not $\mathbf{v}$ is irrotational; if it is, find its potential Φ such that $\mathbf{v} = \nabla\Phi$. In what domain $\mathcal{D}$ is your result valid?

(a) $\mathbf{v} = \hat{\mathbf{i}} - 2\hat{\mathbf{j}} - 8\hat{\mathbf{k}}$
(b) $\mathbf{v} = x^3\hat{\mathbf{i}} - y\hat{\mathbf{j}} + \sin z\hat{\mathbf{k}}$
(c) $\mathbf{v} = z\hat{\mathbf{i}} + y^3\hat{\mathbf{j}} + x\hat{\mathbf{k}}$
(d) $\mathbf{v} = x(2z + y)\hat{\mathbf{i}} + \dfrac{x^2}{2}\hat{\mathbf{j}} + x^2\hat{\mathbf{k}}$
(e) $\mathbf{v} = ye^{-x}\hat{\mathbf{i}} + e^{-|x|}\hat{\mathbf{j}} + z^3\hat{\mathbf{k}}$
(f) $\mathbf{v} = 2ye^x\hat{\mathbf{i}} + 2e^{-|x|}\hat{\mathbf{j}} - 3\hat{\mathbf{k}}$
(g) $\mathbf{v} = 2xz\hat{\mathbf{i}} + 3y\hat{\mathbf{j}} + x^2\hat{\mathbf{k}}$
(h) $\mathbf{v} = z\hat{\mathbf{i}} + z^2\hat{\mathbf{j}} + (x + 2yz)\hat{\mathbf{k}}$
(i) $\mathbf{v} = 2ze^{2y}\hat{\mathbf{j}} + e^{2y}\hat{\mathbf{k}}$
(j) $\mathbf{v} = 2xy\hat{\mathbf{i}} + (2ze^{2y} + x^2)\hat{\mathbf{j}} + e^{2y}\hat{\mathbf{k}}$
(k) $\mathbf{v} = xy\hat{\mathbf{i}} - y\hat{\mathbf{k}}$
(l) $\mathbf{v} = 6\hat{\mathbf{i}} - 2\hat{\mathbf{j}} + xyz\hat{\mathbf{k}}$
(m) $\mathbf{v} = 2xz\hat{\mathbf{i}} + x^2\hat{\mathbf{k}}$
(n) $\mathbf{v} = xz\hat{\mathbf{i}} + 2(y + 1)z\hat{\mathbf{j}} + \left(\dfrac{x^2}{2} + y^2 + 2y\right)\hat{\mathbf{k}}$
(o) $\mathbf{v} = yz\hat{\mathbf{i}} + xz\hat{\mathbf{j}} + xy\hat{\mathbf{k}}$

3. Can functions $f(x, y, z)$, $g(x, y, z)$ be found such that $\mathbf{v}$ is irrotational? If so, find one such f and g. If not, why not?

(a) $\mathbf{v} = xy\hat{\mathbf{i}} - z\hat{\mathbf{j}} + f\hat{\mathbf{k}}$
(b) $\mathbf{v} = 2xy\hat{\mathbf{i}} + x^2\hat{\mathbf{j}} + f\hat{\mathbf{k}}$
(c) $\mathbf{v} = f\hat{\mathbf{i}} + 2xyz\hat{\mathbf{j}} + g\hat{\mathbf{k}}$
(d) $\mathbf{v} = z^2\hat{\mathbf{i}} + f\hat{\mathbf{j}} + g\hat{\mathbf{k}}$

4. Assuming that the conditions of Theorem 16.10.1 are met, are the following correct? Explain.

(a) $$\int_{P_i}^{P_f} \mathbf{v} \cdot d\mathbf{R} = \int_{x_i}^{x_f} v_x(x, y_i, z_i)\,dx + \int_{y_i}^{y_f} v_y(x_f, y, z_i)\,dy + \int_{z_i}^{z_f} v_z(x_f, y_f, z)\,dz$$

(b) $$\int_{P_i}^{P_f} \mathbf{v} \cdot d\mathbf{R} = \int_{x_i}^{x_f} v_x(x, y_f, z_f)\,dx + \int_{y_i}^{y_f} v_y(x_i, y, z_i)\,dy + \int_{z_i}^{z_f} v_z(x_i, y_f, z)\,dz$$

(c) $$\int_{P_i}^{P_f} \mathbf{v} \cdot d\mathbf{R} = \int_{x_i}^{x_f} v_x(x, y_i, z_i)\,dx + \int_{y_i}^{y_f} v_y(x_f, y, z_f)\,dy + \int_{z_i}^{z_f} v_z(x_i, y_f, z)\,dz$$

5. Consider $\int_{\mathcal{C}} \mathbf{v} \cdot d\mathbf{R}$, where $\mathcal{C}$ is the path $x = \sin\tau$, $y = \cos 3\tau$, $z = 2\tau$ from $\tau = 0$ to $\tau = \pi$. Show that $\mathbf{v}$ is irrotational, and use this fact to evaluate the integral two ways:

by path simplification and by the potential method. Sketch the simplified path.

(a) $\mathbf{v} = y\hat{\mathbf{i}} + x\hat{\mathbf{j}}$
(b) $\mathbf{v} = 3x^2 \cos 2y\hat{\mathbf{i}} - 2x^3 \sin 2y\hat{\mathbf{j}}$
(c) $\mathbf{v} = 5x^{3/2}e^{2y}\hat{\mathbf{i}} + (4x^{5/2}e^{2y} + 5y^2)\hat{\mathbf{j}}$
(d) $\mathbf{v} = 3\hat{\mathbf{i}} + 6y^2z^{5/2}\hat{\mathbf{j}} + (5y^3z^{3/2} + 2)\hat{\mathbf{k}}$
(e) $\mathbf{v} = (y^3 - \cos z)\hat{\mathbf{j}} + (y \sin z + 2)\hat{\mathbf{k}}$
(f) $\mathbf{v} = 2x^2\hat{\mathbf{i}} - 2yz\hat{\mathbf{j}} - (y^2 + 3)\hat{\mathbf{k}}$

6. Repeat Exercise 5, but where C is the path $x = \cos^3 \tau$, $y = \tau^2$, $z = 3\tau$ from $\tau = 0$ to $\tau = \pi$.

(a) $\mathbf{v} = x\hat{\mathbf{i}} + y\hat{\mathbf{j}} + z\hat{\mathbf{k}}$
(b) $\mathbf{v} = 8\hat{\mathbf{i}} + (z^2 - 3y)\hat{\mathbf{j}} + (2zy - \sqrt{z})\hat{\mathbf{k}}$
(c) $\mathbf{v} = (x^2 - 2y^{3/2})\hat{\mathbf{i}} - 3x\sqrt{y}\hat{\mathbf{j}} + e^z\hat{\mathbf{k}}$
(d) $\mathbf{v} = y^2z^2\hat{\mathbf{i}} + 2xyz^2\hat{\mathbf{j}} + 2xy^2z\hat{\mathbf{k}}$

7. (*Exact differentials*) An expression $P\,dx + Q\,dy + R\,dz$, where P, Q, and R are functions of x, y, z defined in some domain $\mathcal{D}$, is called a *first-order differential form* in three variables. If there exists a function f such that $P\,dx + Q\,dy + R\,dz = df$ in $\mathcal{D}$, then the form is said to be an **exact differential**. Assuming that P, Q, R are C^1 in $\mathcal{D}$, show that for $P\,dx + Q\,dy + R\,dz$ to be an exact differential it is necessary and sufficient that $\nabla \times (P\hat{\mathbf{i}} + Q\hat{\mathbf{j}} + R\hat{\mathbf{k}}) = 0$ in $\mathcal{D}$. Further, show that f is the scalar potential of $P\hat{\mathbf{i}} + Q\hat{\mathbf{j}} + R\hat{\mathbf{k}}$.

8. (*Stream function*) Consider the plane flow $\mathbf{v} = v_x(x, y)\hat{\mathbf{i}} + v_y(x, y)\hat{\mathbf{j}}$ of an incompressible fluid, where $\mathbf{v}$ is C^1. The volume flow rate (in meters^3/second, say) crossing the curve OP from left to right is

$$Q = \int_C v_x\,dy - v_y\,dx \equiv \int_C \mathbf{f} \cdot d\mathbf{R}, \tag{8.1}$$

where $\mathbf{f} = -v_y(x, y)\hat{\mathbf{i}} + v_x(x, y)\hat{\mathbf{j}}$. Since $\nabla \times \mathbf{f} = (\partial v_x/\partial x + \partial v_y/\partial y)\hat{\mathbf{k}}$ is zero by virtue of the continuity equation [(29) in Section 16.8, with $v_z = 0$ in this case], it follows that

$$Q = \int_{(0,0)}^{(x,y)} v_x\,dy - v_y\,dx \tag{8.2}$$

is a (single-valued) function of x and y, say $\Psi(x, y)$, where

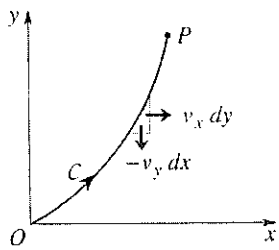
$$\frac{\partial \Psi}{\partial x} = -v_y, \qquad \frac{\partial \Psi}{\partial y} = v_x. \tag{8.3}$$

Ψ is called the **stream function**.

(a) Show that the $\Psi(x, y) =$ constant curves are the streamlines, as were defined in Example 2 of Section 16.2.

(b) If $\mathbf{v} = 2y\hat{\mathbf{i}}$, verify that $\nabla \cdot \mathbf{v} = 0$ (so that Ψ exists), and determine $\Psi(x, y)$. NOTE: Like the velocity potential Φ, the stream function Ψ can be determined only to within an arbitrary additive constant.

(c) Repeat part (b), for $\mathbf{v} = 2xe^{2y}\hat{\mathbf{i}} - e^{2y}\hat{\mathbf{j}}$. Sketch the streamline through $(1, 0)$.



(d) If $\nabla \times \mathbf{v} \equiv \Omega(x, y)\hat{\mathbf{k}}$, show that

$$\nabla^2\Psi = -\Omega(x, y). \tag{8.4}$$

NOTE: If the flow is incompressible *and* irrotational (so that $\Omega = 0$), then there exist both a velocity potential Φ satisfying $\nabla^2\Phi = 0$, *and* a streamfunction Ψ satisfying $\nabla^2\Psi = 0$. If the flow is incompressible but *not* irrotational, Φ does not exist – yet the stream function does exist and satisfies (8.4).

(e) If the flow is incompressible and irrotational, so that both Ψ and Φ exist, show that the $\Psi =$ constant curves and the $\Phi =$ constant curves are everywhere *orthogonal*; i.e., at each point $P = (x, y)$ the $\Psi =$ constant curve through P, and the $\Phi =$ constant curve through P intersect at a right angle. HINT: Consider $\nabla\Psi \cdot \nabla\Phi$.

9. (*Entropy of an ideal gas*) Any gas satisfying the equation of state $pv = RT$, where p is the pressure, v is the volume per mole, T is the absolute temperature, and R is the universal gas constant, is said to be an *ideal gas*. The first law of thermodynamics for one mole of an ideal gas can be expressed as

$$dq = p\,dv + c_v\,dT, \tag{9.1}$$

where dq is the heat input and $c_v = c_v(T)$ is the specific heat at constant volume. Show that the right-hand side of (9.1) is *not* an exact differential (see Exercise 7), but that if we multiply through by $1/T$, then

$$\frac{dq}{T} = \frac{p}{T}\,dv + \frac{c_v}{T}\,dT \tag{9.2}$$

is an exact differential, say $dq/T \equiv ds$, where $s(v, T)$ is called the *entropy* of the (one mole of) gas. [Since ds is an exact differential, s is indeed a *function* of v, T. That is, it is uniquely determined, for the given gas, by the point (v, T)

in the v, T plane (i.e., by the state), independent of the path history.]

10. (*Solenoidal fields*) Let $\mathbf{v}$ be a vector field defined in a domain $\mathcal{D}$. If $\nabla \cdot \mathbf{v} = 0$ at each point in $\mathcal{D}$, then $\mathbf{v}$ is said to be **solenoidal** in $\mathcal{D}$. We have the following companion to Theorem 16.10.1:

THEOREM 16.10.2 *Solenoidal Field*
If $\mathbf{v}$ is C^1 and solenoidal in a simply connected domain $\mathcal{D}$, then there exists a vector field $\mathbf{w}$ in $\mathcal{D}$ such that $\mathbf{v} = \nabla \times \mathbf{w}$.

Prove this theorem, for the case where $\mathcal{D}$ is the prism $x_1 < x < x_2$, $y_1 < y < y_2$, $z_1 < z < z_2$. HINT: We need to show that the equation $\mathbf{v} = \nabla \times \mathbf{w}$ does admit a solution $\mathbf{w}$. To do this, start with the three scalar equations

$$\frac{\partial w_z}{\partial y} - \frac{\partial w_y}{\partial z} = v_x, \qquad \frac{\partial w_x}{\partial z} - \frac{\partial w_z}{\partial x} = v_y, \qquad \frac{\partial w_y}{\partial x} - \frac{\partial w_x}{\partial y} = v_z. \tag{10.1}$$

There is enough leeway in (10.1) to set one of the components of $\mathbf{w}$, say w_x, equal to zero. Then the latter two equations in (10.1) become

$$\frac{\partial w_z}{\partial x} = -v_y \qquad \text{and} \qquad \frac{\partial w_y}{\partial x} = v_z, \tag{10.2}$$

which can be integrated to give w_y to within an arbitrary additive function $A(y, z)$ and w_z to within an arbitrary additive function $B(y, z)$. Putting these results into the first equation in (10.1), show that it is possible to choose $A(y, z) = 0$ so that

$$\mathbf{w} = 0\hat{\mathbf{i}} + \left[\int_{x_0}^{x} v_z(\xi, y, z)\, \partial\xi\right] \hat{\mathbf{j}} + \left[\int_{y_0}^{y} v_x(x_0, \eta, z)\, \partial\eta - \int_{x_0}^{x} v_y(\xi, y, z)\, \partial\xi\right] \hat{\mathbf{k}}, \tag{10.3}$$

where x_0, y_0 are any constants such that $x_1 \le x_0 \le x_2$ and $y_1 \le y_0 \le y_2$. Show that to the right-hand side of (10.3) we can add ∇f, the gradient of an arbitrary scalar function f that is C^2 in $\mathcal{D}$. NOTE: Observe the pattern that emerges in this section: If $\mathbf{v}$ is irrotational ($\nabla \times \mathbf{v} = 0$), $\mathbf{v}$ is expressible as the gradient of a scalar potential Φ ($\mathbf{v} = \nabla\Phi$). Analogously, if $\mathbf{v}$ is solenoidal ($\nabla \cdot \mathbf{v} = 0$), $\mathbf{v}$ is expressible as the curl of a vector potential $\mathbf{w}$ ($\mathbf{v} = \nabla \times \mathbf{w}$). The notions of irrotational and solenoidal fields are complementary in the sense that *every C^1 vector field* $\mathbf{v}$, *defined in a bounded simply connected domain $\mathcal{D}$, can be split as* $\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2$, *where* $\mathbf{v}_1$ *is irrotational in $\mathcal{D}$ and* $\mathbf{v}_2$ *is solenoidal in $\mathcal{D}$*, a fact that we state without proof.

11. Show that $\mathbf{v}$ is solenoidal (see Exercise 10), and determine a vector potential $\mathbf{w}$. (As noted in Exercise 10, $\mathbf{w}$ is not uniquely determined.)

(a) $\mathbf{v} = a\hat{\mathbf{i}} + b\hat{\mathbf{j}} + c\hat{\mathbf{k}}$ $\quad (a, b, c \text{ constants})$
(b) $\mathbf{v} = y\hat{\mathbf{i}} + x\hat{\mathbf{j}}$
(c) $\mathbf{v} = 3\hat{\mathbf{i}} + x^2\hat{\mathbf{j}} + 4y\hat{\mathbf{k}}$
(d) $\mathbf{v} = xy\hat{\mathbf{i}} - yz\hat{\mathbf{k}}$
(e) $\mathbf{v} = 2y\hat{\mathbf{i}} + x^2y^2\hat{\mathbf{j}} - 2x^2yz\hat{\mathbf{k}}$
(f) $\mathbf{v} = x^3y\hat{\mathbf{i}} - z\hat{\mathbf{j}} - 3x^2y(z+1)\hat{\mathbf{k}}$
(g) $\mathbf{v} = x\sin y\hat{\mathbf{i}} + (\cos y + 2z)\hat{\mathbf{j}} + x^2y\hat{\mathbf{k}}$

12. (*Integrating equations of motion*) In Example 3 we considered a single particle. Here we consider a continuum of particles. Consider the motion of a fluid, assumed irrotational and incompressible, in a domain $\mathcal{D}$, under the action of a uniform gravitational force field $-g\hat{\mathbf{k}}$ (per unit mass). If σ is the mass density, $\mathbf{v}$ is the velocity field, and p is the pressure field, then we state, without derivation, that Newton's second law leads to the *Euler equation*

$$\sigma \frac{d\mathbf{v}}{dt} = -\nabla p - \sigma g\hat{\mathbf{k}}, \tag{12.1}$$

or

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} + \nabla\left(\frac{p}{\sigma}\right) + g\hat{\mathbf{k}} = \mathbf{0}, \tag{12.2}$$

where the d/dt in (12.1) is the convective derivative defined in Exercise 5 of Section 16.4. Use equation (8) in Section 16.6 to show that $(\mathbf{v} \cdot \nabla)\mathbf{v} = \nabla(\frac{1}{2}v^2)$, where v is the speed, i.e., the norm of the velocity $\mathbf{v}$, and show that it follows from (12.2) that

$$\frac{\partial \Phi}{\partial t} + \frac{1}{2}v^2 + \frac{p}{\sigma} + gz = \text{"constant"} = F(t) \tag{12.3}$$

everywhere in $\mathcal{D}$ for arbitrary $F(t)$. Equation (12.3) is one form of the well known *Bernoulli equation*.

13. Given $\int_C \mathbf{v} \cdot d\mathbf{R}$, where C is the open path

$$r = \tau, \qquad \theta = \pi\tau/2, \qquad z = \sin\frac{\pi(\tau - 1)}{2}$$

for $\tau : 1 \to 2$, evaluate I both by path simplification and by the potential method, and justify the use of those methods.

(a) $\mathbf{v} = z\hat{\mathbf{e}}_r + r\hat{\mathbf{e}}_z$
(b) $\mathbf{v} = z\sin\theta\hat{\mathbf{e}}_r + z\cos\theta\hat{\mathbf{e}}_\theta + r\sin\theta\hat{\mathbf{e}}_z$
(c) $\mathbf{v} = \cos 5\theta\hat{\mathbf{e}}_r - 5\sin 5\theta\hat{\mathbf{e}}_\theta + z^2\hat{\mathbf{e}}_z$
(d) $\mathbf{v} = 2(rz - \cos\theta)\hat{\mathbf{e}}_r + 2\sin\theta\hat{\mathbf{e}}_\theta + r^2\hat{\mathbf{e}}_z$
(e) $\mathbf{v} = \rho^5\hat{\mathbf{e}}_\rho$
(f) $\mathbf{v} = \sin\phi\hat{\mathbf{e}}_\rho + \cos\phi\hat{\mathbf{e}}_\phi$
(g) $\mathbf{v} = 3\rho^2\cos\theta\hat{\mathbf{e}}_\rho - \rho^2\dfrac{\cos\theta}{\sin\phi}\hat{\mathbf{e}}_\theta$
(h) $\mathbf{v} = \dfrac{\cos\phi}{\rho}\hat{\mathbf{e}}_\phi - \dfrac{\sin\theta}{\rho\sin\phi}\hat{\mathbf{e}}_\theta$

14. In (9), we showed that

$$\int_{P_i}^{P_f} \mathbf{v}\cdot d\mathbf{R} = \int_{P_i}^{P_f} \nabla\Phi\cdot d\mathbf{R} = \Phi\Big|_{P_i}^{P_f}$$

by expressing $\nabla\Phi$ and $d\mathbf{R}$ in Cartesian coordinates. Show that we obtain the same result, namely $\Phi|_{P_i}^{P_f}$, if we use cylindrical coordinates or spherical coordinates.

Chapter 16 Review

We begin by defining the divergence of a vector field $\mathbf{v}$ as

$$\operatorname{div}\mathbf{v} = \lim_{\mathcal{B}\to 0}\left\{\frac{\int_S \hat{\mathbf{n}}\cdot\mathbf{v}\,dA}{V}\right\}.$$

Carrying out that limit in Cartesian coordinates gives

$$\operatorname{div}\mathbf{v} = \frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z},$$

which we identify as $\nabla\cdot\mathbf{v}$, where ∇ is the vector differential operator

$$\nabla = \hat{\mathbf{i}}\frac{\partial}{\partial x} + \hat{\mathbf{j}}\frac{\partial}{\partial y} + \hat{\mathbf{k}}\frac{\partial}{\partial z}.$$

With ∇ in hand, we use it to define

$$\operatorname{grad} u = \nabla u \qquad \text{and} \qquad \operatorname{curl}\mathbf{v} = \nabla\times\mathbf{v}.$$

Thinking of $\mathbf{v}$ as a fluid velocity field (even if it is not), we can interpret $\operatorname{div}\mathbf{v}$ and $\operatorname{curl}\mathbf{v}$ as the outflow per unit volume and twice the fluid angular velocity, respectively. With the help of the directional derivative formula,

$$\frac{du}{ds} = \nabla u\cdot\hat{\mathbf{s}},$$

we learn that ∇u at any point P is normal to the u = constant surface through P, in the direction of increasing u, and its magnitude is equal to the directional derivative du/dn in that direction.

Studying combinations (i.e., more than one operator and/or more than one field), we develop a number of identities such as

$$\nabla\cdot(u\mathbf{v}) = \nabla u\cdot\mathbf{v} + u\nabla\cdot\mathbf{v},$$

and meet the Laplace operator

$$\nabla^2 = \nabla \cdot \nabla = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2},$$

which is known as the Laplacian. In particular, we find that the divergence of every curl is zero (mnemonically, "dc"), and the curl of every gradient is zero ("cg"):

$$\begin{aligned} \text{div curl } \mathbf{v} &= \nabla \cdot \nabla \times \mathbf{v} = 0, \\ \text{curl grad } u &= \nabla \times \nabla u = \mathbf{0}. \end{aligned}$$

In the optional Section 16.7 we derive formulas for the gradient, divergence, curl, and Laplacian in cylindrical and spherical coordinates. Three crucial points, in that section, are as follows. First, the partial derivatives in the ∇, within $\nabla \cdot \mathbf{v}$ say, act not only on the scalar components of $\mathbf{v}$ but also on the base vectors in $\mathbf{v}$. Second, in terms such as

$$\hat{\mathbf{e}}_r \frac{\partial}{\partial r} \cdot (v_\theta \hat{\mathbf{e}}_\theta)$$

we need to do the derivative (of $v_\theta \hat{\mathbf{e}}_\theta$) before the dot or cross product. Third, whereas

$$\begin{aligned} \mathbf{u} \times \mathbf{v} &= (u_1\hat{\mathbf{e}}_1 + u_2\hat{\mathbf{e}}_2 + u_3\hat{\mathbf{e}}_3) \times (v_1\hat{\mathbf{e}}_1 + v_2\hat{\mathbf{e}}_2 + v_3\hat{\mathbf{e}}_3) \\ &= \begin{vmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \hat{\mathbf{e}}_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix} \end{aligned}$$

holds for any vectors $\mathbf{u}$ and $\mathbf{v}$ (where $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3$ are the right-handed orthonormal base vectors $\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}$, or $\hat{\mathbf{e}}_r, \hat{\mathbf{e}}_\theta, \hat{\mathbf{e}}_z$ or $\hat{\mathbf{e}}_\rho, \hat{\mathbf{e}}_\phi, \hat{\mathbf{e}}_\theta$), it does *not* hold, for cylindrical or spherical coordinates, if $\mathbf{u}$ is the ∇ operator.

In Sections 16.8 and 16.9 we study the Gauss divergence theorem and Stokes's theorem (as well as Green's theorem and Green's identities) and observe the pattern of being able to express an integral over an n-dimensional integral as an integral (or evaluation) over its $(n-1)$-dimensional boundary, provided that the integrand is a sort of derivative:

$$\begin{aligned} &1 \to 0: && \int_a^b F'(x)\,dx = F(x)\Big|_{x=a}^{x=b} && \text{(Fundamental Theorem)} \\ &2 \to 1: && \int_S \hat{\mathbf{n}} \cdot \nabla \times \mathbf{v}\,dA = \oint_C \mathbf{v} \cdot d\mathbf{R} && \text{(Stokes's Theorem)} \\ &3 \to 2: && \int_V \nabla \cdot \mathbf{v}\,dV = \int_S \hat{\mathbf{n}} \cdot \mathbf{v}\,dA && \text{(Divergence Theorem)} \end{aligned}$$

where the numbers at the left denote the dimensions n and $n-1$. [In fact, the Fundamental Theorem is just a one-dimensional version of the Divergence Theorem, where $\mathbf{v} = F(x)\hat{\mathbf{i}}$ and where the region V is a rectangular prism ex-

tending from $x = a$ to $x = b$ in the x-direction, for then $\int_{\mathcal{V}} \nabla \cdot \mathbf{v}\, dV$ reduces to $\int_a^b F'(x)\, dx$ times the area A of the end faces at $x = a$ and $x = b$, and $\int_S \hat{\mathbf{n}} \cdot \mathbf{v}\, dA$ reduces to $F(b) - F(a)$ times A; then A cancels from the left- and right-hand sides.] These theorems are especially useful in deriving various field equations – such as the continuity equation of fluid mechanics, the diffusion equation, and various Maxwell's equations.

In the final section, on irrotational fields, the key result is Theorem 16.10.1, which (under the conditions stated therein) expressed the equivalance between $\mathbf{v}$ being irrotational, the existence of a scalar potential Φ, the line integral around every closed loop being zero, and the line integral on an open path being independent of the path.

Chapter 17

Fourier Series, Fourier Integral, Fourier Transform

17.1 Introduction

This chapter is about a number of methods associated with the French applied mathematician *Joseph Fourier* (1768–1830), methods which will be central when we turn next to the subject of partial differential equations in Chapters 18–20. In fact, series of trigonometric functions, which we now call **Fourier series**, were already being actively studied – by Euler, Lagrange, d'Alembert, Daniel Bernoulli, and others – even when Fourier was born. For instance, d'Alembert had already given integral formulas for computing the coefficients in those series, and various specific "Fourier series" had been put forward. Yet, there was considerable debate regarding the class of functions that could be successfully expanded in such series, and various sets of sufficient conditions were slow to appear, the first being given by Peter Gustav Lejeune–Dirichlet around 1829. Part of the difficulty was that even the meaning of the term "function" was not yet clear or agreed upon. But it is also true that mathematical issues were deep and elusive. For instance, in 1873, Paul Du Bois–Reymond put forward an example of a function that is continuous on $(-\pi, \pi)$, yet has a Fourier series that fails to converge at any point in that interval! Thus, the subject of Fourier series has been one of the most fertile in the development of modern pure and applied mathematics.

Although Fourier neither invented "Fourier" series nor settled the outstanding fundamental questions, he did use them fruitfully, especially for problems regarding the conduction of heat in solids, governed by the PDE

$$\alpha^2 \nabla^2 T = \frac{\partial T}{\partial t}$$

that is derived in Section 16.8 and to which we return in Chapter 18. In claiming that an "arbitrary" function could be represented by a Fourier series Fourier overstated the case, and his work was faulted for lack of rigor. Yet he had the insight to

see the power of these new methods. His work advanced the use of Fourier methods and solidified techniques that would be needed to solve problems in field theory and that continue to be developed even today.

To explain what this chapter is about we ask you to recall, first, the importance of being able to expand a given vector in terms of a set of orthogonal base vectors. Likewise, we will find in this chapter that a given (sufficiently well-behaved) function can be expanded in terms of a set of "orthogonal functions." Fourier and his contemporaries did not have vector space concepts available to them but, essentially, they were seeking sets of orthogonal functions to be used as base vectors in an infinite-dimensional function space. That task is much more difficult than for finite-dimensional spaces. For instance, if we have n orthogonal vectors in an n-dimensional space (where n is finite), then they provide a basis. However, in an infinite-dimensional space an infinite set of orthogonal functions may, but need not, be a basis. For suppose we have an infinite set of orthogonal functions that *is* a basis. If we remove one of them, then we have an infinite number of them left, but that diminished set will not be a basis. Further, expansions in an infinite-dimensional space are infinite series, so subtle matters arise regarding their convergence and manipulation.

Fourier series are introduced and developed in Section 17.3, though not from a vector space point of view. The alternative, and more modern, vector space approach is given subsequently in Section 17.6. By then, the nagging question arises: Where does the set of orthogonal functions, which comprise the individual terms in a Fourier series, "come from"? Are there other such sets? The answer is provided, in Sections 17.7 and 17.8, by the Sturm–Liouville theory. There, it is revealed that such sets of orthogonal base vectors are generated as the eigenfunctions of second-order differential equation eigenvalue problems known as Sturm–Liouville problems. Thus, Sections 17.7 and 17.8 are the function space analog of Section 11.3 on symmetric matrices, wherein we found that the eigenvectors of a real symmetric $n \times n$ matrix provide an orthogonal basis for n-space.

In Section 17.9 we will let the period of the periodic functions under consideration tend to infinity, and will find that the Fourier series representation gives way to a Fourier integral representation. The Fourier integral gives us, by a mere rearrangement, the Fourier transform, in Section 17.10, which is very much analogous to the Laplace transform that we studied in Chapter 5. In fact, in the final section, 17.11, we show how to derive the Laplace transform from the Fourier transform.

For interesting historical (and mathematical) accounts, we suggest the little book by R. L. Jeffery [*Trigonometric Series* (Toronto: University of Toronto Press, (1956)] as well as the historical treatise by Morris Kline [*Mathematical Thought from Ancient to Modern Times* (New York: Oxford, 1972)].*

*Other standard sources include H. S. Carslaw, *Introduction to the Theory of Fourier's Series and Integrals*, 3rd ed. (New York: Dover, 1930), and I. Grattan-Guinness, *Joseph Fourier, 1768–1830* (Cambridge, MA: MIT Press, 1972).

17.2 Even, Odd, and Periodic Functions

Before taking up our study of Fourier series, in the next section, we need to define even, odd, and periodic functions.

Let f be defined on an x interval, finite or infinite, that is centered at $x = 0$. We say that f is an **even** function if

$$\boxed{f(-x) = f(x),} \tag{1}$$

and an **odd** function if

$$\boxed{f(-x) = -f(x),} \tag{2}$$

for all x in that interval. That is, the graph of f is *symmetric* about $x = 0$ if f is even, and *antisymmetric* about $x = 0$ if f is odd. Examples are shown in Fig. 1. For example, $5, x^2, 3x^4, \cos x, \sin|x|$, and e^{-x^2} are even, and $x, 3x^3, 2x^5$, $\sin x$, and $x\cos x$ are odd.

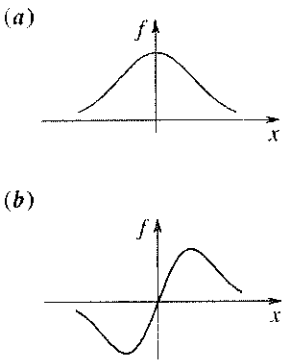


Figure 1. Even and odd.

There are several useful algebraic properties of even and odd functions, such as the following:

$$\text{even} + \text{even} = \text{even}, \tag{3a}$$
$$\text{even} \times \text{even} = \text{even}, \tag{3b}$$
$$\text{odd} + \text{odd} = \text{odd}, \tag{3c}$$
$$\text{odd} \times \text{odd} = \text{even}, \tag{3d}$$
$$\text{even} \times \text{odd} = \text{odd}. \tag{3e}$$

To prove (3e), for example, let $F(x)$ be even and let $G(x)$ be odd. Then $F(-x)G(-x) = F(x)[-G(x)] = -F(x)G(x)$, in accord with (2).

In addition, two useful integral properties are as follows. If f is even, then

$$\boxed{\int_{-A}^{A} f(x)\,dx = 2\int_{0}^{A} f(x)\,dx, \qquad (f \text{ even})} \tag{4a}$$

and if f is odd, then

$$\boxed{\int_{-A}^{A} f(x)\,dx = 0, \qquad (f \text{ odd})} \tag{4b}$$

for if we interpret the integrals in (4a) as areas (positive above the x axis, negative below it), then the area $\int_{-A}^{0} f(x)\,dx$ is equal to the $\int_{0}^{A} f(x)\,dx$ due to the symmetry of the graph of f. And in the case of (4b) the areas $\int_{-A}^{0} f(x)\,dx$ and $\int_{0}^{A} f(x)\,dx$ are negatives of each other, due to the antisymmetry of the graph of f, and hence cancel.

Alternatively, (4a) and (4b) follow directly from (1) and (2), respectively. For example, if f is odd, then

$$\begin{aligned}
\int_{-A}^{A} f(x)\,dx &= \int_{-A}^{0} f(x)\,dx + \int_{0}^{A} f(x)\,dx \\
&= \int_{A}^{0} f(-t)\,(-dt) + \int_{0}^{A} f(x)\,dx \qquad (x=-t) \\
&= \int_{0}^{A} f(-t)\,dt + \int_{0}^{A} f(x)\,dx \\
&= \int_{0}^{A} -f(t)\,dt + \int_{0}^{A} f(x)\,dx \qquad \text{(oddness of } f\text{)} \\
&= -\int_{0}^{A} f(x)\,dx + \int_{0}^{A} f(x)\,dx \qquad (t=x) \\
&= 0, \qquad (5)
\end{aligned}$$

as stated in (4b).

Note carefully that a given function is not necessarily even or odd; it may be *both* even and odd, or it may be *neither*. Every function can be uniquely decomposed into the sum of an even function, say f_e, and an odd function, say f_o, as demonstrated by the simple identity

$$\boxed{\begin{aligned} f(x) &= \frac{f(x)+f(-x)}{2} + \frac{f(x)-f(-x)}{2} \\ &\equiv \quad f_e(x) \quad + \quad f_o(x), \end{aligned}} \qquad (6)$$

for observe that

$$f_o(-x) = \frac{f(-x)-f(x)}{2} = -\frac{f(x)-f(-x)}{2} = -f_o(x)$$

and, similarly, that $f_e(-x) = f_e(x)$.

EXAMPLE 1. Surely $f(x) = e^x$ is neither even nor odd, since (Fig. 2) it is neither symmetric nor antisymmetric about $x = 0$. Putting $f(x) = e^x$ and $f(-x) = e^{-x}$ into (6) gives

$$f_e(x) = \frac{e^x + e^{-x}}{2} \qquad \text{and} \qquad f_o(x) = \frac{e^x - e^{-x}}{2}$$

as the even and odd parts of e^x, respectively. In fact, we recognize these functions as $\cosh x$ and $\sinh x$, so it is interesting that we can think of $\cosh x$ and $\sinh x$ as the even and odd parts of e^x, respectively. ∎

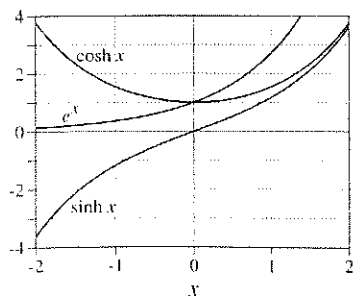


Figure 2. Even and odd parts of e^x.

Notice that (6) is reminiscent of other decompositions that occur in mathematics, whereby a mathematical object (such as a function, matrix, or vector) is broken

into the sum of two complementary parts. For instance, every square matrix can be expressed as the sum of a symmetric matrix and a skew-symmetric matrix (Exercise 6, Section 10.3), every vector in 3-space can be expressed as the sum of a vector along a given line and a vector perpendicular to that line, and every C^1 vector field can be expressed as the sum of an irrotational field and a solenoidal field (Exercise 10, Section 16.10).

Next, suppose that for a given function f there exists a positive constant T such that

$$\boxed{f(x+T) = f(x)} \tag{7}$$

for every x in the domain of f. Then we say that f is a **periodic** function of x, with period T. Sometimes we say that f is **T-periodic**.

EXAMPLE 2. For example, $\sin x$ is periodic with period 2π because $\sin(x+2\pi) = \sin x \cos 2\pi + \sin 2\pi \cos x = \sin x$ for all x. ∎

In graphical terms, one can think of the graph of a periodic function f as generated by stamping it out one period at a time, as with an inked woodblock.

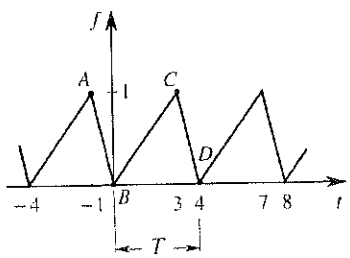


Figure 3. Periodic function f.

EXAMPLE 3. The function f shown in Fig. 3 is seen to be periodic with period $T = 4$, for if the segment BCD, for instance, is "stamped out" indefinitely to the right and left we generate the graph of f. There is nothing special about choosing the segment BCD for this purpose; ABC, or any other segment of length 4, would do as well. ∎

Notice that if f is periodic with period T, it is necessarily periodic with period $2T, 3T, 4T, \ldots$ as well. For example, $f(x+2T) = f((x+T)+T) = f(x+T) = f(x)$, so that f is periodic with period $2T$. Of all these possible periods, if there exists a *smallest* one, that period is called the **fundamental period**. Thus, $\sin x$ (in Example 2) is periodic with period $2\pi, 4\pi, 6\pi, \ldots$, and its fundamental period is 2π; f in Example 3 is periodic with period $4, 8, 12, \ldots$, and its fundamental period is 4.

In contrast, observe that if $f(x) =$ constant, then f is periodic and *every* $T > 0$ is a period. Thus, there exists no smallest period, so f does not have a fundamental period.

Closure. Even, odd, and periodic functions will be basic to our study of Fourier series, to follow. The defining properties are (1), (2), and (7), respectively.

EXERCISES 17.2

1. (a) Prove (3a). (b) Prove (3b).
(c) Prove (3c). (d) Prove (3d).

2. Provide a proof of (4a) that is analogous to the proof of (4b) given in (5).

3. Prove that

(a) f is both even and odd if and only if it is identically zero
(b) if f is even (and integrable), then $F(x) = \int_0^x f(t)\,dt$ is odd
(c) if f is odd (and integrable), then $F(x) = \int_0^x f(t)\,dt$ is even
(d) if f is even (and differentiable), then $F(x) = df/dx$ is odd
(e) if f is odd (and differentiable), then $F(x) = df/dx$ is even

4. Prove the decomposition formula (6). HINT: Start with $f(x) = f_e(x) + f_o(x)$ and change x to $-x$.

5. Determine $f_e(x)$ and $f_o(x)$. Is f even? Odd? Neither?

(a) $2 - 5x$ (b) $\sin(x+2)$
(c) $x/(x^2+x+3)$ (d) xe^{-x}
(e) $x/(x+2)$ (f) $x^2\cos(x^3) - 8$
(g) $x^4 + x^3 + x^2 + x + 1$ (h) $\ln(1+x^2)$
(i) $e^{-2\sin x}$ (j) $\sin(\sin x)$
(k) $\cos(\sin x)$ (l) $e^{-x}/(x^2+1)$

6. If F is even and G is odd, show that

(a) $1/F(x)$ is even (b) $1/G(x)$ is odd
(c) $F(G(x))$ is even (d) $G(F(x))$ is even

7. Show that if f is odd, then it is necessarily true that $f(0) = 0$.

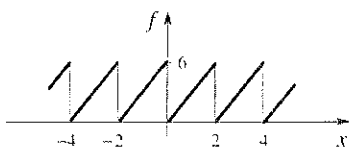
8. Let f be even (and not identically zero), and let g be odd (and not identically zero). If f and g are defined on a common interval, show that f and g are necessarily linearly independent on that interval.

9. Show that if $f(x) = g(x)$ (over a common x interval), we can equate even and odd parts: $f_e(x) = g_e(x)$, and $f_o(x) = g_o(x)$.

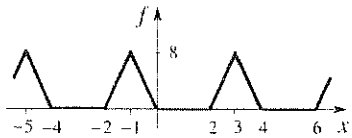
10. Show that if f is even and g is odd, and $f + g = 0$ (over a common x interval), then $f(x) = 0$ and $g(x) = 0$.

11. Determine the fundamental period in each case. Also, draw the graphs of f_e and f_o.

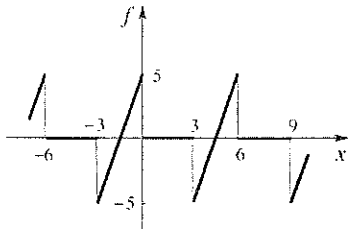
(a)



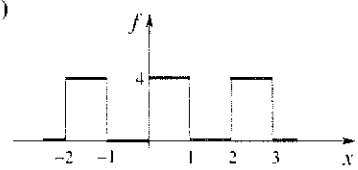
(b)



(c)



(d)



12. Determine whether or not the given function (defined on $-\infty < x < \infty$) is periodic. If it is, find its fundamental period (if it has one).

(a) x^4 (b) e^x (c) e^{-x}
(d) $\sin(\omega x + \phi)$ (e) $\cos 6x$ (f) $\sin 2x$
(g) $\tan x$ (h) $\sinh x$ (i) $\cosh x$
(j) $\cos^2 x$ (k) $\sin^2 x$ (l) $\sin x \cos 2x$
(m) $e^{\sin 3x}$ (n) $\sinh(\cos 2x)$ (o) $\sin(\sin x)$
(p) $\cos(\sin x)$ (q) $\sin(8\pi\cos x)$ (r) $\sin(\sin 4x)$
(s) $\sin|x|$ (t) $\cos|x|$ (u) $x\sin x$

13. The following functions are periodic. Determine the fundamental period in each case. (The a_n's and b_n's are nonzero constants.)

(a) $a_0 + a_1\cos x$
(b) $a_0 + a_1\cos x + a_2\cos 2x$
(c) $6\cos x - 4\sin 3x$
(d) $\cos 5x + \sin 5x$
(e) $a_0 + a_1\cos x + b_1\sin x + a_2\cos 2x + b_2\sin 2x$
(f) $a_0 + \sum_{n=1}^{\infty}(a_n\cos nx + b_n\sin nx)$
(g) $a_0 + \sum_{n=1}^{\infty}\left(a_n\cos\dfrac{n\pi x}{\ell} + b_n\sin\dfrac{n\pi x}{\ell}\right)$

14. Show that if f is periodic with period T, then

$$\int_0^T f(x)\,dx = \int_A^{A+T} f(x)\,dx$$

for any finite value of A.

15. Show that if $f(x)$ is periodic with period T, then

(a) so is its derivative $f'(x)$

(b) so is its integral $\int_0^x f(t)\,dt$, if and only if $\int_0^T f(t)\,dt = 0$

16. Let $f(x) = 1$ whenever x is rational, and let $f(x) = 0$ whenever x is irrational. Show that f is periodic with period T where T is any (positive) rational number. (Thus there exists no smallest period, so f does not have a fundamental period.) Is it also true that f is periodic with period T, where T is any (positive) *irrational* number? Explain.

17. Show that if f is periodic with period T, then $g(f(x))$ is, too. Give two examples.

17.3 Fourier Series of a Periodic Function

17.3.1. Fourier series. Recall that if $f(x)$ is infinitely differentiable at a point $x = a$, then it has a Taylor series about that point,

$$\text{TS}\, f\Big|_{x=a} = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n. \tag{1}$$

For that series to be useful, we need two things. First, we need it to converge on some interval I. Second, if it does converge on I, then we need its sum function to be the same as the original function $f(x)$.* If that is the case, then we say that the Taylor series $\text{TS}\, f\Big|_{x=a}$ **represents** f on I, and we can write

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n \tag{2}$$

on I. We know from experience that Taylor series representations are useful in many ways.

Similarly, there are other types of representations that are useful as well. In this chapter we are concerned with the representation of periodic functions, not by Taylor series but by **trigonometric series**, that is, by series of the form

$$a_0 + \sum_{n=1}^{\infty}\left(a_n \cos\frac{n\pi x}{\ell} + b_n \sin\frac{n\pi x}{\ell}\right). \tag{3}$$

First, observe that *(3) is a periodic function with fundamental period 2ℓ.* To see this, note that (3) is a linear combination of the functions 1, $\cos(\pi x/\ell)$, $\sin(\pi x/\ell)$,

*For instance, we saw in Example 2 of Section 13.5 that it is possible for $\text{TS}\, f\Big|_{x=a}$ to converge, but not to $f(x)$.

$\cos(2\pi x/\ell)$, $\sin(2\pi x/\ell)$, and so on, and that these functions are periodic with the following periods:

$$1: \quad \text{arbitrary}, \tag{4a}$$

$$\cos\frac{\pi x}{\ell} \quad \text{and} \quad \sin\frac{\pi x}{\ell}: \quad \underline{2\ell}, 4\ell, 6\ell, 8\ell, \ldots, \tag{4b}$$

$$\cos\frac{2\pi x}{\ell} \quad \text{and} \quad \sin\frac{2\pi x}{\ell}: \quad \ell, \underline{2\ell}, 3\ell, 4\ell, \ldots, \tag{4c}$$

$$\cos\frac{3\pi x}{\ell} \quad \text{and} \quad \sin\frac{3\pi x}{\ell}: \quad \frac{2\ell}{3}, \frac{4\ell}{3}, \underline{2\ell}, \frac{8\ell}{3}, \ldots, \tag{4d}$$

and so on. The smallest period shared by all the terms is 2ℓ [underlined in (4)], so the fundamental period of (3) is 2ℓ.* Thus, perhaps the trigonometric series (3) can be used to represent periodic functions of period 2ℓ. (When we say of period 2ℓ, we shall mean of fundamental period 2ℓ.)

Specifically, if $f(x)$ is periodic, of period 2ℓ, then we define the **Fourier series of f**, say FS f, as

$$\boxed{\text{FS}\, f = a_0 + \sum_{n=1}^{\infty}\left(a_n \cos\frac{n\pi x}{\ell} + b_n \sin\frac{n\pi x}{\ell}\right),} \tag{5a}$$

where the coefficients are given by the **Euler formulas**

$$\boxed{\begin{aligned} a_0 &= \frac{1}{2\ell}\int_{-\ell}^{\ell} f(x)\,dx, \\ a_n &= \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\cos\frac{n\pi x}{\ell}\,dx, \qquad n = 1, 2, \ldots \\ b_n &= \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\sin\frac{n\pi x}{\ell}\,dx, \qquad n = 1, 2, \ldots \end{aligned}} \tag{5b,c,d}$$

and are known as the **Fourier coefficients** of f.

For FS f to represent f we need the series to converge, and we need its sum function to be the same as the original function $f(x)$. Various theorems are available, that give sufficient conditions on f for FS f to represent f. One such theorem, that is easily applied and which covers the vast majority of periodic functions that arise in applications, is as follows.†

THEOREM 17.3.1 *Fourier Convergence Theorem*
Let f be 2ℓ-periodic, and let f and f' be piecewise continuous on $[-\ell, \ell]$. Then

*Actually, the fundamental period of the function (3) *may* be less than 2ℓ. For example, if all the coefficients except for a_3 are zero, then the fundamental period is $2\ell/3$. Thus, we should say that (3) is always periodic with period 2ℓ: its fundamental period is at most 2ℓ, but may be less.

†For proof of a slightly stronger version of Theorem 17.3.1, see R. V. Churchill and J. W. Brown, *Fourier Series and Boundary Value Problems*, 3rd ed. (New York: McGraw-Hill, 1978), Sec.41.

the Fourier series given by (5) converges to $f(x)$ at every point x at which f is continuous, and to the mean value $[f(x+) + f(x-)]/2$ at every point x at which f is discontinuous.

Piecewise continuity is defined in Section 5.2. By $f(x+)$ and $f(x-)$ we mean the **right-** and **left-hand limits** of f,

$$f(x+) \equiv \lim_{h\to 0} f(x+h) \qquad \text{and} \qquad f(x-) \equiv \lim_{h\to 0} f(x-h), \tag{6}$$

where $h \to 0$ through positive values. [If $f(x+) = f(x-) = f(x)$, then f is continuous at x, otherwise it is discontinuous there.]

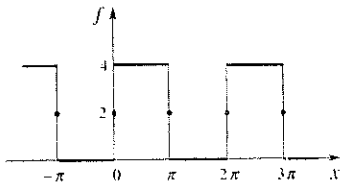


Figure 1. Square wave.

EXAMPLE 1. *Square Wave.* Consider the "square wave" f shown in Fig. 1, where $f(x)$ is defined as 2 at $x = 0, \pm\pi, \pm 2\pi, \ldots$, as indicated by the heavy dots. Since the period referred to in Theorem 17.3.1 and in (5) is 2ℓ, and the period is seen from Fig. 1 to be 2π, it follows that $\ell = \pi$. Both f and f' are piecewise continuous on $[-\pi, \pi]$, so the theorem applies. Let us use (5) to work out the Fourier series of f and examine its convergence to the square wave f using computer plots of the partial sums of the series.

First, (5b) gives

$$\begin{aligned} a_0 &= \frac{1}{2\ell}\int_{-\ell}^{\ell} f(x)\,dx = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\,dx \\ &= \frac{1}{2\pi}\left[\int_{-\pi}^{0} 0\,dx + \int_{0}^{\pi} 4\,dx\right] = 2. \end{aligned} \tag{7}$$

Actually, $a_0 = 2$ could have been seen by inspection because the right-hand side of (5b) is, by definition, the **average value** of f over one period and, from Fig. 1, we can see that for our square wave that average value is 2.

Next,

$$\begin{aligned} a_n &= \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\cos\frac{n\pi x}{\ell}\,dx = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos nx\,dx \\ &= \frac{1}{\pi}\left[\int_{-\pi}^{0} 0\,dx + \int_{0}^{\pi} 4\cos nx\,dx\right] = \left.\frac{4\sin nx}{n\pi}\right|_0^{\pi} = 0 \end{aligned} \tag{8}$$

since $\sin n\pi = 0$ and $\sin 0 = 0$. Finally,

$$\begin{aligned} b_n &= \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\sin\frac{n\pi x}{\ell}\,dx = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin nx\,dx \\ &= \frac{1}{\pi}\left[\int_{-\pi}^{0} 0\,dx + \int_{0}^{\pi} 4\sin nx\,dx\right] = \frac{4}{n\pi}(1-\cos n\pi). \end{aligned} \tag{9}$$

We can leave (9) as it stands, but it is best to note that $\cos n\pi = (-1)^n$, so that we can reduce (9) to

$$b_n = \frac{4}{n\pi}[1-(-1)^n] = \begin{cases} \dfrac{8}{n\pi}, & n = 1, 3, \ldots \\ 0, & n = 2, 4, \ldots. \end{cases} \tag{10}$$

Thus the Fourier series representation of f is

$$f(x) = 2 + \frac{8}{\pi} \sum_{1,3,\ldots}^{\infty} \frac{1}{n} \sin nx, \tag{11}$$

where "$1, 3, \ldots$" tells us to omit the terms corresponding to $n = 2, 4, \ldots$. If preferred, (11) can be expressed, equivalently, as

$$f(x) = 2 + \frac{8}{\pi} \sum_{1}^{\infty} \frac{\sin(2n-1)x}{2n-1}, \tag{12}$$

with $\sum_1^\infty$ denoting the usual summation over $n = 1, 2, 3, \ldots$.

Observe from Fig. 1 that the defined values of $f(x)$ at the jump discontinuities ($x = n\pi$) coincide with the mean values $[f(x+) + f(x-)]/2 = 2$, so Theorem 17.3.1 assures us that (12) is indeed an equality for every value of x.

Is it not remarkable that a linear combination of sines and cosines, each of which is beautifully smooth (infinitely differentiable for all x, i.e., of class C^∞), can sum to a function with jump discontinuities? To obtain insight as to how this convergence to f is accomplished, it will be illuminating to plot some of the partial sums of the series. Writing out (12) as

$$f(x) = 2 + \frac{8}{\pi} \sin x + \frac{8}{3\pi} \sin 3x + \frac{8}{5\pi} \sin 5x + \cdots, \tag{13}$$

we define the partial sums of the series as

$$\begin{aligned} s_1(x) &= 2, \\ s_2(x) &= 2 + \frac{8}{\pi} \sin x, \\ s_3(x) &= 2 + \frac{8}{\pi} \sin x + \frac{8}{3\pi} \sin 3x, \end{aligned}$$

and so on: that is, $s_N(x)$ is the sum of the first N terms.

We see from Fig. 2 that although the first partial sum $s_1(x)$ is merely a constant it "does the best it can" by equalling the average value of f. If we are to provide the needed correction to $s_1(x)$ we need to pull it up on the right and push it down on the left, as

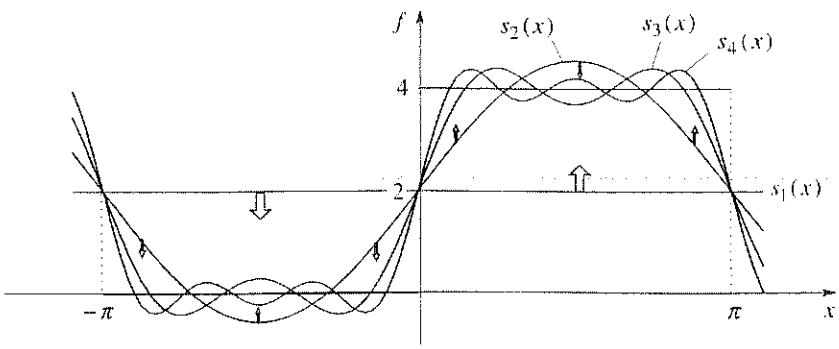


Figure 2. The first few partial sums.

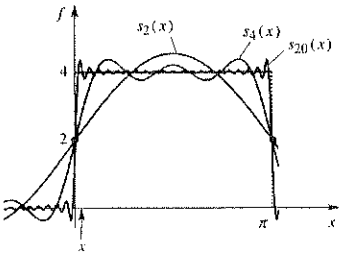


Figure 3. Gibbs phenomenon.

indicated by the broad arrows in the figure. A $\sin x$ term can accomplish that (a $\cos x$ term is not appropriate since we need an antisymmetric correction, not a symmetric one) and an amplitude a bit greater than 2 seems optimal. That correction is precisely what is provided by the second term in the series, so $s_2(x)$ begins to take on the desired shape. To correct $s_2(x)$, in turn, we need to push it up and down according to the six thin arrows. Such correction dictates the need for a third sine harmonic, and such a term is indeed forthcoming as the third term in the series. By adding more and more terms of the series, the graph comes closer and closer to the square wave f. Observe also that the graph of every partial sum passes through the mean value 2 at each jump discontinuity, which result is also clear from (12) since $\sin n\pi = 0$ for each n.

Next, consider the graph of $s_N(x)$ for larger N, for instance for $N = 20$ (Fig. 3). As x increases from zero, $s_{20}(x)$ rises sharply from the mean value 2, overshoots the value 4, and settles down close to 4 – until x approaches π, where the same sort of overshoot is followed by a steep descent to the mean value 2 at $x = \pi$. By periodicity, these results repeat over each period.

Strangely, the overshoot does not diminish as N increases. For instance, the peak values of $s_2(x)$, $s_4(x)$, and $s_{20}(x)$ are 4.548, 4.376, and 4.360, respectively, and it can be shown that the overshoot approaches a limiting value of around 9% of the jump (the jump is 4, so 9% of 4 is 0.36) as $N \to \infty$. This persistent 9% overshoot occurs not only in this example but in the Fourier series representation of any function with a jump discontinuity, and is known as the **Gibbs phenomenon.***

In view of this overshoot, one may well wonder how convergence to the square wave is attained, for are there not always x locations at which the error is around 9% no matter how many terms are summed? The key point is that when we say that $\lim_{N\to\infty} s_N(x) = f(x)$ it is N that is varying; x is *fixed*. Picking any x point (Fig. 3), as close to the origin as we like, as N is increased the overshoot "spike" eventually moves to the left of x, and subsequent values of $s_N(x)$ do settle down and converge to $f(x)$. Nonetheless, the Gibbs phenomenon is of great practical importance because it implies that the convergence of Fourier series may be painfully slow (and expensive) in the vicinity of a jump discontinuity.†

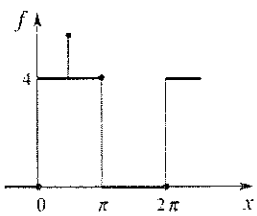


Figure 4. Modified f.

COMMENT 1. We have already noted that the Fourier series of the square wave $f(x)$ shown in Fig. 1 converges to $f(x)$ at every point x. Suppose, instead, we define f as shown in Fig. 4. That is, the modified f is 0 at $x = 0, \pm 2\pi, \pm 4\pi, \ldots$, it is 4 at $x = \pm\pi, \pm 3\pi, \ldots$, and it is 10^6 (not shown to scale) at $x = \frac{\pi}{2}, -\frac{3\pi}{2}, \frac{5\pi}{2}, -\frac{7\pi}{2}, \ldots$. The Fourier series of the modified f, say f_{mod}, will be identical to the Fourier series of f, because the Fourier coefficients are computed, according to (5b,c,d), by integrals, and changing the value of the integrands at a number of isolated points does not change the value of those integrals.

*Josiah Willard Gibbs (1839–1903) was one of the first important American mathematical physicists, and is especially well known for his work on vector analysis and thermodynamics. The Gibbs phenomenon was noted by Gibbs in 1899 (*Nature*, Vol. 59, p. 606) and explained in 1906 by Maxime Bôcher (*Annals of Mathematics*, Vol. 2, No. 7, p. 81).

†When studying infinite series in the calculus one learns to classify series as convergent or divergent, with the implication that convergent series are "good" and divergent series are "bad." Practically speaking, however, a series may converge – but so slowly as to be almost worthless. Thus, interest exists in developing *acceleration* techniques that permit us to transform a given slowly convergent series into a more rapidly convergent one. For a numerical technique to suppress the Gibbs phenomenon, see Forman Acton, *Numerical Methods That Work* (Washington, D.C.: Mathematical Assn. of America, 1990).

Thus, the Fourier series of f_{mod} will converge to $f_{\text{mod}}(x)$ for every point x except points such as $x = 0$, $\frac{\pi}{2}$, and π, where it will converge to 2, 4, and 2, respectively. Such pointwise discrepancies will cause no problem in applications. Thus, from this point forward (except in the exercises to this section) we will not bother to show the heavy dots, as we did in Figs. 1 and 4, and will simply show these graphs as in Fig. 5. ∎

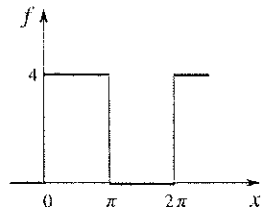


Figure 5. No circles or dots.

Thus far, the concepts of even and odd functions, introduced in Section 17.2, have not been used here. Suppose now that f is an *even* 2ℓ-periodic function. Then the integrands $f(x)$ and $f(x)\cos(n\pi x/\ell)$ in (5b) and (5c) are even functions, and the integrand $f(x)\sin(n\pi x/\ell)$ in (5c) is odd, so these Euler formulas simplify to

$$a_0 = \frac{1}{2\ell}\int_{-\ell}^{\ell} f(x)\,dx = \frac{1}{\ell}\int_0^{\ell} f(x)\,dx,$$

$$a_n = \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\cos\frac{n\pi x}{\ell}\,dx = \frac{2}{\ell}\int_0^{\ell} f(x)\cos\frac{n\pi x}{\ell}\,dx, \tag{14}$$

$$b_n = \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\sin\frac{n\pi x}{\ell}\,dx = 0.$$

Observe that only the constant and cosine terms survive, there being no need for sine terms in representing an even function. Similarly, if f is *odd*, then

$$a_0 = \frac{1}{2\ell}\int_{-\ell}^{\ell} f(x)\,dx = 0,$$

$$a_n = \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\cos\frac{n\pi x}{\ell}\,dx = 0, \tag{15}$$

$$b_n = \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\sin\frac{n\pi x}{\ell}\,dx = \frac{2}{\ell}\int_0^{\ell} f(x)\sin\frac{n\pi x}{\ell}\,dx,$$

and the constant and cosine terms drop out.

More generally, f is neither even nor odd. The constant and cosine terms in the Fourier series of $f(x)$ represent the even part $f_e(x)$, and the sine terms represent the odd part $f_o(x)$. That is,

$$\begin{aligned} f(x) &= \left[a_0 + \sum_1^{\infty} a_n\cos\frac{n\pi x}{\ell}\right] + \left[\sum_1^{\infty} b_n\sin\frac{n\pi x}{\ell}\right] \\ &= \qquad f_e(x) \qquad\qquad + \qquad f_o(x), \end{aligned} \tag{16}$$

and

$$f_e(x) = a_0 + \sum_1^{\infty} a_n\cos\frac{n\pi x}{\ell}, \tag{17a}$$

$$f_o(x) = \sum_1^{\infty} b_n\sin\frac{n\pi x}{\ell}. \tag{17b}$$

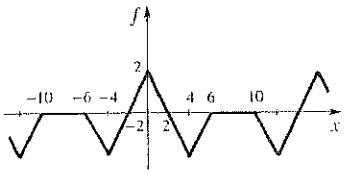


Figure 6. f in Example 2.

EXAMPLE 2. Let us work out the Fourier series of the periodic function f, the graph of which is given in Fig. 6. Its period is 16, so $\ell = 8$. We can compute a_0 by (5b), but we can see from the figure (dividing the net area over one period by 16) that the average height is $-\frac{1}{4}$ so $a_0 = -\frac{1}{4}$. Further, we see that f is even, in which case (14) gives $b_n = 0$ and

$$\begin{aligned}
a_n &= \frac{2}{\ell}\int_0^{\ell} f(x)\cos\frac{n\pi x}{\ell}\,dx = \frac{2}{8}\int_0^{8} f(x)\cos\frac{n\pi x}{8}\,dx \\
&= \frac{1}{4}\left[\int_0^4 (2-x)\cos\frac{n\pi x}{8}\,dx + \int_4^6 (x-6)\cos\frac{n\pi x}{8}\,dx + \int_6^8 0\cos\frac{n\pi x}{8}\,dx\right] \\
&= \frac{1}{4}\left[2\frac{\sin(n\pi x/8)}{n\pi/8} - \frac{1}{(n\pi/8)^2}\left(\cos\frac{n\pi x}{8} + \frac{n\pi x}{8}\sin\frac{n\pi x}{8}\right)\right]\Bigg|_0^4 \\
&\quad + \frac{1}{4}\left[\frac{1}{(n\pi/8)^2}\left(\cos\frac{n\pi x}{8} + \frac{n\pi x}{8}\sin\frac{n\pi x}{8}\right) - 6\frac{\sin(n\pi x/8)}{n\pi/8}\right]\Bigg|_4^6 \\
&= \frac{16}{n^2\pi^2}\left(1 - 2\cos\frac{n\pi}{2} + \cos\frac{3n\pi}{4}\right). \qquad (18)
\end{aligned}$$

Thus,

$$f(x) = -\frac{1}{4} + \frac{16}{\pi^2}\sum_{n=1}^{\infty}\frac{1 - 2\cos\frac{n\pi}{2} + \cos\frac{3n\pi}{4}}{n^2}\cos\frac{n\pi x}{8}, \qquad (19)$$

with the equality holding for all x, without exception.

COMMENT. In Example 1 we simplified $\sin n\pi$ and $\cos n\pi$ as 0 and $(-1)^n$, respectively, but the $\cos(n\pi/2)$ and $\cos(3n\pi/4)$ terms in (19) are trickier (depending on n, $\cos(n\pi/2)$ takes on the values $0, \pm 1$, and $\cos(3n\pi/4)$ takes on the values $0, \pm 1, \pm\sqrt{2}/2$), so we will leave them intact. ∎

Recall from the calculus that the ***p*-series**,

$$\sum_1^{\infty}\frac{1}{n^p}, \qquad (20)$$

converges if $p > 1$ and diverges if $p \le 1$, the borderline divergent case of $p = 1$ corresponding to the **harmonic series**

$$\sum_1^{\infty}\frac{1}{n}. \qquad (21)$$

Further, you may recall that the *alternating* harmonic series

$$\sum_1^{\infty}\frac{(-1)^n}{n} \qquad (22)$$

is convergent – barely convergent in that it converges extremely slowly,* and is only conditionally convergent rather than absolutely convergent. The Fourier series (11) is similar to the alternating series (22) and, like (22), converges "by the skin of its teeth,"† which result should not be surprising since (11) amounts to the representation of a function having jump discontinuities in terms of a collection of infinitely smooth sines. The function f in Example 2 is better behaved in that it is continuous and, sure enough, the coefficients in (19) die out faster, proportional to $1/n^2$. That pattern persists: as we consider periodic functions that are better and better behaved ($C^0, C^1, C^2, \ldots$) we find that their Fourier coefficients die out (as $n \to \infty$) more and more rapidly. In general, if a periodic function is C^N, we can expect its Fourier coefficients to tend to zero at least as fast as $1/n^{N+2}$.‡ As an extreme case, observe that the periodic function $f(x) = 5 \sin 3x$, say, is infinitely smooth (i.e., it is of class C^∞), and its Fourier coefficients do tend to zero infinitely fast. That is, its Fourier series is simply one term, $\mathrm{FS}\{5 \sin 3x\} = 5 \sin 3x$.

17.3.2. Euler's formulas. We chose to state the Fourier convergence theorem and to move quickly into examples to solidify the Fourier series concept. Now let us back up and show where Euler's formulas (5b,c,d) come from. Accepting that a (sufficiently well-behaved) 2ℓ-periodic function f can be represented in the Fourier series form

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi x}{\ell} + b_n \sin \frac{n\pi x}{\ell} \right), \tag{23}$$

we focus attention on how to compute the coefficients a_0, a_n, and b_n. We will need the following elementary integral formulas:

$$\int_{-\ell}^{\ell} \cos \frac{m\pi x}{\ell} \cos \frac{n\pi x}{\ell}\, dx = \begin{cases} 0 & m \neq n \\ \ell & m = n \neq 0 \\ 2\ell & m = n = 0 \end{cases} \tag{24a}$$

$$\int_{-\ell}^{\ell} \sin \frac{m\pi x}{\ell} \sin \frac{n\pi x}{\ell}\, dx = \begin{cases} 0 & m \neq n \\ \ell & m = n \neq 0 \end{cases} \tag{24b}$$

$$\int_{-\ell}^{\ell} \cos \frac{m\pi x}{\ell} \sin \frac{n\pi x}{\ell}\, dx = 0 \qquad \text{for all } m, n, \tag{24c}$$

where m and n are integers. The three zero results in (24) are due to cancellation of positive and negative areas; the two nonzero results, in (24a) and (24b), occur when $m = n$, in which case the integrands are squared quantities and no such cancellation can occur.

*For three-significant-figure accuracy we need to sum around 10^3 terms; for six-significant-figure accuracy we need around 10^6 terms, and so on.

†The $1/n$ decay, in (22) and (11), is not enough to induce convergence. Rather, these series converge (according to the *Dirichlet test*) because $(-1)^n$ and $\sin nx$ have *bounded partial sums*; that is, there exist finite numbers A and B such that $|\sum_{n=1}^{N}(-1)^n| < A$ and $|\sum_{n=1}^{N} \sin nx| < B$ for all N's, no matter how large.

‡See, for instance, H. S. Carslaw, *Fourier Series* (New York: Dover, 1930), pp. 269–271.

For definiteness, let us solve (23) for a_2. To do so, multiply both sides of (23) by $\cos(2\pi x/\ell)$ and integrate over one period. That step gives

$$\begin{aligned}
\int_{-\ell}^{\ell} f(x)\cos\frac{2\pi x}{\ell}\,dx &= \int_{-\ell}^{\ell}\left[a_0+\sum_{n=1}^{\infty}\left(a_n\cos\frac{n\pi x}{\ell}+b_n\sin\frac{n\pi x}{\ell}\right)\right]\cos\frac{2\pi x}{\ell}\,dx\\
&= a_0\int_{-\ell}^{\ell}\cos\frac{2\pi x}{\ell}\,dx + a_1\int_{-\ell}^{\ell}\cos\frac{\pi x}{\ell}\cos\frac{2\pi x}{\ell}\,dx\\
&\quad +b_1\int_{-\ell}^{\ell}\sin\frac{\pi x}{\ell}\cos\frac{2\pi x}{\ell}\,dx + a_2\int_{-\ell}^{\ell}\cos\frac{2\pi x}{\ell}\cos\frac{2\pi x}{\ell}\,dx\\
&\quad +b_2\int_{-\ell}^{\ell}\sin\frac{2\pi x}{\ell}\cos\frac{2\pi x}{\ell}\,dx + a_3\int_{-\ell}^{\ell}\cos\frac{3\pi x}{\ell}\cos\frac{2\pi x}{\ell}\,dx\\
&\quad +\cdots\\
&= 0+0+0+a_2\ell+0+0+\cdots\\
&= a_2\ell,
\end{aligned} \tag{25}$$

so

$$a_2 = \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\cos\frac{2\pi x}{\ell}\,dx, \tag{26}$$

where the zeros following the third equal sign follow from (24a,b,c), and the $a_2\ell$ term follows from (24a) with $m = n = 2$. [The first zero follows from (24a) with $m = 0$ and $n = 2$.] More generally, to solve for a_n multiply (23) by $\cos(n\pi x/\ell)$ and integrate over one period; to solve for b_n multiply by $\sin(n\pi x/\ell)$ instead, and to solve for a_0 multiply by 1 instead. These steps, with the help of (24), give the Euler formulas (5b,c,d).

The interchange in the order of summation and integration, in the second step in (25) needs justification, but we will postpone discussion of such technical points until Section 17.5.

There is a strong analogy between the calculation of the coefficients c_n in the expansion

$$\mathbf{v} = \sum_{n=1}^{N} c_n\mathbf{e}_n \tag{27}$$

of a vector $\mathbf{v}$ in an N-dimensional vector space, in terms of an orthogonal basis $\{\mathbf{e}_1,\ldots,\mathbf{e}_N\}$. To solve for c_2, for instance, we take advantage of the orthogonality of the basis vectors, and dot both sides with $\mathbf{e}_2$. That step gives

$$\begin{aligned}
\mathbf{v}\cdot\mathbf{e}_2 &= c_1\mathbf{e}_1\cdot\mathbf{e}_2 + c_2\mathbf{e}_2\cdot\mathbf{e}_2 + c_3\mathbf{e}_3\cdot\mathbf{e}_2 + \cdots + c_N\mathbf{e}_N\cdot\mathbf{e}_2\\
&= 0 + c_2\mathbf{e}_2\cdot\mathbf{e}_2 + 0 + \cdots + 0\\
&= c_2\mathbf{e}_2\cdot\mathbf{e}_2,
\end{aligned} \tag{28}$$

so

$$c_2 = \frac{\mathbf{v}\cdot\mathbf{e}_2}{\mathbf{e}_2\cdot\mathbf{e}_2} \tag{29}$$

and similarly for $c_1, c_3, c_4, \ldots, c_N$. In fact, in Sections 17.6 and 17.7 we will take another look at Fourier series from exactly that point of view. There, we will understand (23) as the expansion of a vector f in an infinite-dimensional function space in terms of the base vectors $1, \cos(\pi x/\ell), \sin(\pi x/\ell), \cos(2\pi x/\ell), \sin(2\pi x/\ell), \ldots$, which are orthogonal by virtue of equations (24); in fact, (24) are often called **orthogonality relations**.

17.3.3. Applications. Fourier series are indispensable in our study of PDE's (partial differential equations) in the next three chapters. Here, we give two applications to physical systems that are governed by ODE's and subjected to periodic forcing functions. Our approach will be *formal*, by which we mean that although we believe the results to be correct, we will not rigorously justify all of the steps involved in their derivation.

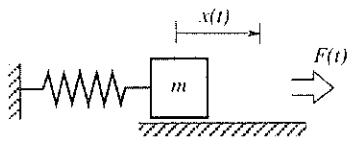


Figure 7. Forced mechanical oscillator.

EXAMPLE 3. *Periodically Driven Oscillator.* Consider the driven mechanical oscillator shown in Fig. 7 and governed by the differential equation of motion

$$mx'' + cx' + kx = F(t), \tag{30}$$

where m, c, k are the mass, damping coefficient (associated with some combination of viscous damping due to a film of lubricating oil between the mass and the table, and air resistance), and spring stiffness, respectively. Let $m = 1$ kg, $c = 0.04$ kg/sec, and $k = 15$ kg/sec^2, and let $F(t)$ (in newtons) be as shown in Fig. 8. F consists of an endless sequence of pulses, each having unit area (except the first, which is only a "half pulse"). In mechanics, $\int_{t_1}^{t_2} F(t)\,dt$ is the **impulse** delivered by the force F between times t_1 and t_2, so F consists of a periodic sequence of unit impulses, of period 2π. Thus, $\ell = \pi$ in (5). Even though the starting time is $t = 0$, so $t \geq 0$, we can think of F as the extended function shown in Fig. 9, which is even. Thus, if we expand F in a Fourier series we have, for its coefficients, $a_0 =$ average value $= 1/(2\pi)$, $b_n = 0$ because F is even, and, from (14),

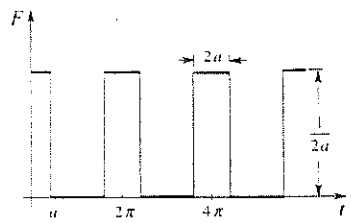


Figure 8. Forcing function F.

$$a_n = \frac{2}{\pi}\int_0^{\pi} F(t)\cos nt\,dt = \frac{2}{\pi}\int_0^{a} \frac{1}{2a}\cos nt\,dt = \frac{\sin na}{n\pi a}. \tag{31}$$

Thus, (30) becomes

$$x'' + 0.04x' + 15x = \frac{1}{2\pi} + \sum_1^{\infty} \frac{\sin na}{n\pi a}\cos nt. \tag{32}$$

Recall, from the general theory in Chapter 3, that if

$$L[x] = f_1 + \cdots + f_N, \tag{33}$$

where L is shorthand for a linear second-order differential operator, then the general solution of (33) is of the form

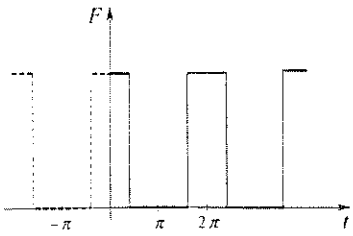


Figure 9. F is even.

$$x(t) = C_1 x_{h1}(t) + C_2 x_{h2}(t) + x_{p1}(t) + \cdots + x_{pN}(t), \tag{34}$$

where C_1, C_2 are arbitrary constants, x_{h1} and x_{h2} are linearly independent homogeneous solutions, and $x_{p1}, \ldots, x_{pN}$ are particular solutions corresponding to the forcing functions

$f_1, \ldots, f_N$, respectively. If two initial conditions are prescribed, they enable us to determine C_1 and C_2.

In the present example x_{h1} and x_{h2} are oscillatory, but with a slow exponential decay due to the $0.04x'$ damping term in (32).* As $t \to \infty$ those terms decay to zero, leaving us with the particular solution as the steady-state response. Let us limit our attention to finding, and discussing, the steady-state response.

Our plan is to find a particular solution corresponding to each forcing term on the right side of (32), using the method of undetermined coefficients, and then to superimpose those solutions, as in (34). Consider

$$x'' + 0.04x' + 15x = \cos nt, \tag{35}$$

and seek x_p in the form $A \cos nt + B \sin nt$. Putting that form into the left-hand side of (35) enables us to solve for A and B, and we obtain

$$x_p(t) = \frac{15 - n^2}{(15 - n^2)^2 + 0.0016n^2} \cos nt + \frac{0.04n}{(15 - n^2)^2 + 0.0016n^2} \sin nt. \tag{36}$$

Further, we find that a particular solution corresponding to the $1/2\pi$ forcing term in (32) is $x_p(t) = 1/30\pi$, so the desired steady-state response is, by linearity and superposition,

$$x(t) = \frac{1}{30\pi} + \sum_{1}^{\infty} \frac{\sin na}{n\pi a} \left[\frac{15 - n^2}{(15 - n^2)^2 + 0.0016n^2} \cos nt \right.$$
$$\left. + \frac{0.04n}{(15 - n^2)^2 + 0.0016n^2} \sin nt \right]. \tag{37}$$

Choosing a specific value for a, we can use (37) to compute $x(t)$. However, for purposes of understanding it will be useful to express the square-bracketed term in (37), equivalently, in the form $A_n \cos(nt + \phi_n)$, where the amplitude A_n and the phase ϕ_n are

$$A_n = \frac{1}{\sqrt{(15 - n^2)^2 + 0.0016n^2}} \qquad \text{and} \qquad \phi_n = \tan^{-1}\left(\frac{0.04n}{n^2 - 15}\right). \tag{38}$$

Let $a = \pi/8$, say. Then, using (38) we can write out (37) as

$$\begin{aligned} x(t) = {} & 0.0106 + 0.0222 \cos(t - 0.0029) + 0.0261 \cos(2t - 0.0073) \\ & + 0.0416 \cos(3t - 0.0200) + 0.2001 \cos(4t + 3.3002) \\ & + 0.0150 \cos(5t + 3.1616) + \cdots. \end{aligned} \tag{39}$$

Consider the terms on the right-hand side of (39). The first, 0.0106, is the response to the $1/2\pi$ forcing term in (32), the second is the response to the $n = 1$ term in (32), the third is the response to the $n = 2$ term in (32), and so on. Observe that if we ignore the $0.04x'$ damping term in (35) then the natural frequency of the oscillator is $\sqrt{15}$ and that the $\cos nt$ forcing function comes close to that natural frequency when $n = 4$. Sure enough, the $n = 4$ term in (39) has by far the largest amplitude, 0.2001. Observe further that the phase ϕ_n is rather small for $n < 4$. That is, the response term is almost in phase with the forcing term

*If this sounds unfamiliar we urge you to review Section 3.5.

– because the damping is light. However, for $n \geq 4$ the ϕ_n's are approximately π. That is, when the frequency of the forcing terms exceeds the approximate natural frequency $\sqrt{15}$ the phase increases to around π, so that the response is around 180° out of phase with the forcing function.*

Plotting the steady-state response $x(t)$, given by (39), in Fig. 10, we can see how the response is dominated by the $0.2001 \cos(4t + 3.3002)$ term, as discussed above. Observe that it suffices to plot $x(t)$ over any 2π interval since it is 2π-periodic.

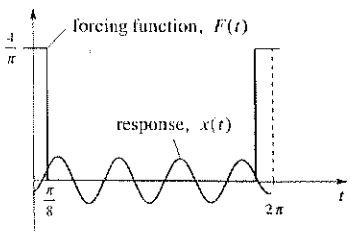


Figure 10. Steady-state response.

COMMENT 1. Why did we expand $F(t)$ in a Fourier series? That step gave us $F(t)$ as a linear combination of elementary functions, the response to each of which was readily found, with the total response then built up by superposition. Alternatively, the Laplace transform method would also have been convenient. (See Exercise 17 for other ideas.)

COMMENT 2. We stated that our solution would be formal rather than rigorous. The point that we did not justify is as follows. The expression (34) satisfies (33) because

$$\begin{aligned} L[x] &= L\left[C_1 x_{h1} + C_1 x_{h2} + \sum_1^N x_{pn}\right] \\ &= C_1 L[x_{h1}] + C_2 L[x_{h2}] + L\left[\sum_1^N x_{pn}\right] \\ &= 0 + 0 + \sum_1^N L[x_{pn}] = \sum_1^N f_n. \end{aligned} \tag{40}$$

However, in the present case $N = \infty$, so the step

$$L\left[\sum_1^\infty x_{pn}\right] = \sum_1^\infty L[x_{pn}] \tag{41}$$

amounts to an interchange in the order of two limit operations, the derivatives in L and the infinite series. The validity of such interchange will not be covered until Section 17.5.

COMMENT 3. Two limiting cases are of interest (and are available to us because we left a as a parameter). As $a \to \pi$, $F(t)$ tends to the constant $1/2\pi$. In that case the series in (32) vanishes and the steady-state response is simply $x(t) = 1/30\pi$. More subtle is the case where $a \to 0$ since then $F(t)$ becomes a sequence of hammer blows, each imparting unit impulse. Discussion of this case is left for the exercises.

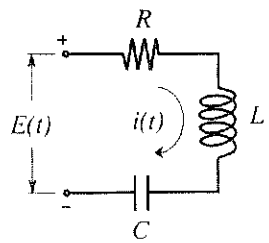


Figure 11. Electrical analog.

COMMENT 4. The electrical analog of the forced mechanical oscillator, governed by (30) and shown in Fig. 7, is the equation

$$LQ'' + RQ' + \frac{1}{C}Q = E(t) \tag{42}$$

governing the circuit shown in Fig. 11, where $Q(t)$ is the charge on the capacitor. ∎

*See Fig. 2 in Section 3.8. To understand the present example you may need to review both Sections 3.5 and 3.8.

EXAMPLE 4. *Infinite Beam on Elastic Foundation.* Consider an infinitely long beam on an elastic foundation, sketched in Fig. 12. The constant k is called the *foundation modulus* (i.e., the spring stiffness per unit x-length) and the square wave $w(x)$ is a prescribed periodic loading (force per unit length).* Physically, the beam might be a train track, with the elastic foundation used to model the track bed. We wish to determine the vertical deflection $u(x)$.

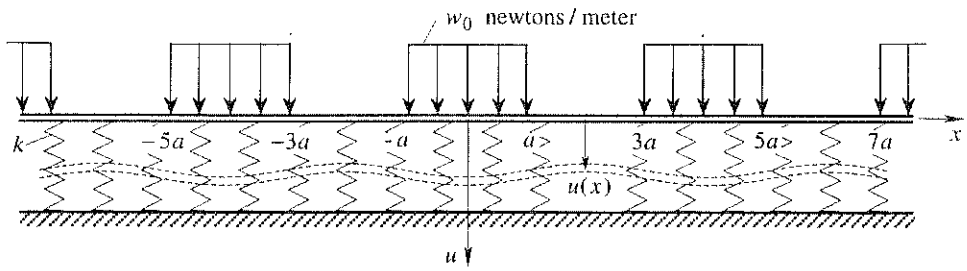


Figure 12. Infinite beam on elastic foundation.

According to the classical *Euler beam theory*† the deflection $u(x)$ resulting from a load distribution $p(x)$ newtons per meter satisfies the fourth-order differential equation

$$EIu'''' = p(x), \tag{43}$$

where E and I are physical constants of the beam; EI is called the *flexural rigidity* of the beam (and is considered here as a known constant) since u'''', and hence the deflection u, is inversely proportional to EI. Now $p(x)$ is the *net* loading and consists of the applied periodic loading $w(x)$ downward and the spring force $ku(x)$ upward. (We neglect the weight of the beam, for simplicity.) Thus, $p(x) = w(x) - ku(x)$, and (43) becomes

$$EIu'''' + ku = w(x). \tag{44}$$

This problem is similar to the preceding one in that they both involve differential equations with periodic forcing functions, but in this case the independent variable is x and the interval is $-\infty < x < \infty$. As in Example 3, we begin by expanding the forcing function in a Fourier series,

$$w(x) = \frac{w_0}{2} + \frac{2w_0}{\pi} \sum_{n=1}^{\infty} \frac{\sin(n\pi/2)}{n} \cos \frac{n\pi x}{2a}. \tag{45}$$

Rather than find the response to each forcing term and adding them up, let us use a slightly different procedure (which could have been used in Example 3 as well). Namely, anticipating that $u(x)$ will be an even periodic function, of the same period as $w(x)$ (i.e., 4a), let us seek $u(x)$ in the form

$$u(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos \frac{n\pi x}{2a}. \tag{46}$$

*That is, the downward load between any points x_1 and x_2 ($> x_1$) is $\int_{x_1}^{x_2} w(x)\,dx$.

†See, for example, S. Timoshenko, *Strength of Materials*, Part I (Princeton, NJ: D. Van Nostrand, 1955).

Formally differentiating (46) termwise four times, putting this result and (45) into (44), and formally equating coefficients of each cosine harmonic gives

$$ka_0 = \frac{w_0}{2} \tag{47}$$

and

$$\left(EI\,\frac{n^4\pi^4}{16a^4} + k\right) a_n = \frac{2w_0}{\pi}\,\frac{\sin(n\pi/2)}{n}, \tag{48}$$

so

$$a_0 = \frac{w_0}{2k}, \qquad a_n = \frac{2w_0}{\pi}\,\frac{\sin(n\pi/2)}{n[EI\,(n^4\pi^4/16a^4) + k]}. \quad (n \geq 1) \tag{49}$$

Thus we have the formal solution

$$u(x) = \frac{w_0}{2k} + \frac{32w_0a^4}{\pi}\sum_{n=1}^{\infty}\frac{\sin(n\pi/2)}{n(EIn^4\pi^4 + 16a^4k)}\cos\frac{n\pi x}{2a}. \tag{50}$$

COMMENT 1. It is striking that the terms in the series die out so rapidly, proportional to $1/n^5$ as $n \to \infty$, so that we can expect merely the first couple of terms of (50) to give a good approximation to $u(x)$,

$$u(x) \approx \frac{w_0}{2k} + \frac{32w_0a^4}{\pi(\pi^4EI + 16a^4k)}\cos\frac{\pi x}{2a}. \tag{51}$$

How are we to understand the input $w(x)$ being discontinuous and the output $u(x)$ being quite smooth? Physically, we don't expect $u(x)$ to be discontinuous or "kinky" just because the loading $w(x)$ is, because a train track is too "stiff" for that. Mathematically, observe that in essence (though not procedurally, unless $k = 0$) we solve (44) for u by four integrations of w. Now, what happens when we repeatedly integrate a discontinuous function? Consider a Heaviside step function, for simplicity, in place of w. Integrating from $-\infty$, say, to a variable point x, gives

$$\begin{aligned}\int_{-\infty}^{x} H(\xi)\,d\xi &= \begin{cases} 0, & x < 0 \\ x, & x > 0 \end{cases} \\ &= xH(x),\end{aligned} \tag{52}$$

which is a "ramp" function. Integrating $xH(x)$, in turn, gives $(x^2/2)H(x)$; integrating $(x^2/2)H(x)$ gives $(x^3/6)H(x)$, and so on, as displayed in Fig. 13. That is, *integration is a smoothing operation*: $H(x)$ is discontinuous, its integral $xH(x)$ is C^0 (continuous), the integral of the latter is C^1, the integral of the latter is C^2, and so on.* From this rough argument, we expect the response $u(x)$ to the discontinuous load $w(x)$ to be C^3, so its Fourier coefficients should tend to zero like $1/n^5$, and that is precisely what is revealed by (50).

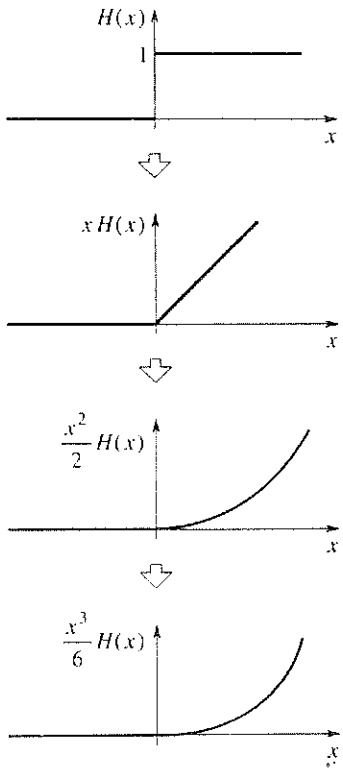


Figure 13. The smoothing effect of integration.

COMMENT 2. Actually, (50) is a *particular* solution of (44). To obtain the general solution we need to add the homogeneous solution

$$e^{\beta x}(A\sin\beta x + B\sin\beta x) + e^{-\beta x}(C\sin\beta x + D\sin\beta x), \tag{53}$$

*Conversely, differentiation has the opposite effect. For instance, $(x^3/6)H(x)$ is smooth but has a singular behavior at the origin, that is brought to light by repeated differentiations.

where $\beta \equiv \sqrt[4]{k/EI}/\sqrt{2}$. To determine A, B, C, D we need some sort of boundary conditions at the ends of the interval, that is, as $x \to \pm\infty$. By way of such conditions, let us require u to be *bounded* as $x \to \pm\infty$, as would be a reasonable requirement for a loaded train track. Then the behavior $e^{\beta x} \to \infty$ as $x \to +\infty$ implies that we need $A = B = 0$, and the behavior $e^{-\beta x} \to \infty$ as $x \to -\infty$ implies that we need $C = D = 0$. Thus, (53) is eliminated entirely, and we are left with (50).

COMMENT 3. Thus, this example is a *boundary-value* problem in which the homogeneous solution drops out by virtue of the boundary conditions. Example 3 is an *initial-value* problem in which the homogeneous solution is not zero, but tended to zero as $t \to \infty$, leaving the particular solution as the steady-state solution. ∎

17.3.4. Complex exponential form for Fourier series. Using the definitions

$$\cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2} \qquad \text{and} \qquad \sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2i}, \tag{54}$$

it is possible to re-express the Fourier series formula (5) in terms of complex exponentials, as follows:

$$\boxed{\text{FS}\, f = \sum_{n=-\infty}^{\infty} c_n\, e^{in\pi x/\ell},} \tag{55a}$$

where

$$\boxed{c_n = \frac{1}{2\ell}\int_{-\ell}^{\ell} f(x)\, e^{-in\pi x/\ell}\, dx.} \tag{55b}$$

Although the c_n's and exponentials in (55a) are complex, the series does have a real-valued sum.

Notice that the usual definition

$$\sum_{n=1}^{\infty} A_n \equiv \lim_{N\to\infty} \sum_{n=1}^{N} A_n$$

does not apply to (55a) because the lower limit is infinite as well. From our derivation of (55), below, we will see that the appropriate meaning of the series in (55a) is

$$\sum_{n=-\infty}^{\infty} c_n\, e^{in\pi x/\ell} \equiv \lim_{N\to\infty} \sum_{n=-N}^{N} c_n\, e^{in\pi x/\ell}. \tag{56}$$

Let us proceed:

$$\begin{aligned}\text{FS}\, f &= a_0 + \sum_{n=1}^{\infty}\left(a_n \cos\frac{n\pi x}{\ell} + b_n \sin\frac{n\pi x}{\ell}\right)\\ &= \lim_{N\to\infty}\sum_{n=0}^{N}\left(a_n \cos\frac{n\pi x}{\ell} + b_n \sin\frac{n\pi x}{\ell}\right)\end{aligned}$$

$$= \lim_{N\to\infty} \sum_{n=0}^{N} \left(a_n \frac{e^{in\pi x/\ell} + e^{-in\pi x/\ell}}{2} + b_n \frac{e^{in\pi x/\ell} - e^{-in\pi x/\ell}}{2i} \right)$$

$$= \lim_{N\to\infty} \left[\sum_{n=0}^{N} \left(\frac{a_n - ib_n}{2} \right) e^{in\pi x/\ell} + \sum_{n=0}^{N} \left(\frac{a_n + ib_n}{2} \right) e^{-in\pi x/\ell} \right]. \quad (57)$$

Changing n to $-n$ in the second sum, which is permissible because n is merely a dummy summation index,

$$\begin{aligned} \text{FS}\, f &= \lim_{N\to\infty} \left[\sum_{n=0}^{N} \left(\frac{a_n - ib_n}{2} \right) e^{in\pi x/\ell} + \sum_{n=0}^{-N} \left(\frac{a_{-n} + ib_{-n}}{2} \right) e^{in\pi x/\ell} \right] \\ &= \lim_{N\to\infty} \sum_{n=-N}^{N} c_n\, e^{in\pi x/\ell} \\ &= \sum_{n=-\infty}^{\infty} c_n\, e^{in\pi x/\ell}. \end{aligned} \quad (58)$$

For $n = 0$,

$$c_0 = a_0 = \frac{1}{2\ell} \int_{-\ell}^{\ell} f(x)\, dx, \quad (59)$$

for $n > 0$,

$$\begin{aligned} c_n = \frac{a_n - ib_n}{2} &= \frac{1}{2\ell} \int_{-\ell}^{\ell} f(x) \left(\cos \frac{n\pi x}{\ell} - i \sin \frac{n\pi x}{\ell} \right) dx \\ &= \frac{1}{2\ell} \int_{-\ell}^{\ell} f(x)\, e^{-in\pi x/\ell}\, dx, \end{aligned} \quad (60)$$

and for $n < 0$,

$$\begin{aligned} c_n = \frac{a_{-n} + ib_{-n}}{2} &= \frac{1}{2\ell} \int_{-\ell}^{\ell} f(x) \left[\cos \left(-\frac{n\pi x}{\ell} \right) + i \sin \left(-\frac{n\pi x}{\ell} \right) \right] dx \\ &= \frac{1}{2\ell} \int_{-\ell}^{\ell} f(x)\, e^{-in\pi x/\ell}\, dx. \end{aligned} \quad (61)$$

All three cases, (59)–(61), agree with (55b), as was to be shown.

The complex form (55) is sometimes favored, especially for electrical engineering applications. An advantage of (55) over (5) is that (55) contains only the exponentials and one set of coefficients (the c_n's), whereas (5) contains both cosines and sines and two sets of coefficients (the a_n's and b_n's).

EXAMPLE 5. Let us return to the square wave of Example 1 and work out its complex Fourier series. With $\ell = \pi$, and f displayed in Fig. 1,

$$c_n = \frac{1}{2\ell} \int_{-\ell}^{\ell} f(x)\, e^{-in\pi x/\ell}\, dx = \frac{1}{2\pi} \left[\int_{-\pi}^{0} 0\, dx + \int_{0}^{\pi} 4\, e^{-inx}\, dx \right]$$

$$= \frac{2}{\pi}\left(\frac{e^{-inx}}{-in}\right)\Bigg|_{x=0}^{x=\pi} = \begin{cases} 0, & n = \pm 2, \pm 4, \ldots \\ -\dfrac{4i}{n\pi}, & n = \pm 1, \pm 3, \ldots . \end{cases} \tag{62}$$

For $n = 0$, the quantity following the third equal sign is indeterminate; it is $(1-1)/0 = 0/0$. In that case, in place of $\int 4e^{-inx}\,dx = 4e^{-inx}/(-in)$ we should use $\int 4e^0\,dx = 4x$, which gives $c_0 = 2$. Thus,

$$f(x) = 2 + \sum_{n=-\infty\,(\text{nodd})}^{\infty} \left(-\frac{4i}{n\pi}\right) e^{inx}, \tag{63}$$

and it is not difficult to verify that (63) is equivalent to (11). ∎

Closure. The Fourier series of a 2ℓ-periodic function f, given by (5), converges to $f(x)$ at every point x at which f is continuous, if f and f' are piecewise continuous on $[-\ell, \ell]$. These conditions are sufficient, not necessary, and are met by virtually all functions of applied interest. If f is discontinuous at x, then the series converges to the mean value $[f(x+) + f(x-)]/2$. If the latter does not equal the defined value of $f(x)$, then a discrepancy exists at that point. However, such pointwise discrepancies will be inconsequential in applications.

Beyond the question of convergence, the *speed* of convergence depends on how well-behaved the function is and is of great practical importance. Surely it matters to us whether we need to add 100,000 terms to achieve a desired accuracy, or whether three or four terms will suffice.

Specifically, if a periodic function is C^N, then we can expect its Fourier coefficients to tend to zero at least as fast as $1/n^{N+2}$. In the loaded beam example (Example 4), for instance, the solution $u(x)$ is C^3, so its Fourier coefficients decay like $1/n^5$. Because of that rapid decay, it suffices to retain only the first two or three terms of the series.

The Fourier coefficients of a given periodic function often contain such quantities as $\cos n\pi$, $\sin n\pi$, $\cos(n\pi/2)$, $\sin(n\pi/2)$, $\cos(3n\pi/4)$, and so on. Of these, $cosn\pi = (-1)^n$ and $\sin n\pi = 0$, but the others are not so readily expressed in algebraic form, so we will leave them intact, and we suggest that you do the same – in working out the exercises.

EXERCISES 17.3

1. Are these functions piecewise continuous on $[0, \pi]$? Explain.

(a) $\sin^2 x$ (b) $\tan x$
(c) $\sin \frac{1}{x}$ (d) $\cos \frac{1}{x}$
(e) e^{-x^2} (f) $1/(x-1)$
(g) $1/x$ (h) $\sqrt{x}$
(i) $\begin{cases} 100, & x \neq 2 \\ 50, & x = 2 \end{cases}$ (j) $\begin{cases} x, & 0 \le x \le 2 \\ 3 - x, & 2 < x \le \pi \end{cases}$

2. (a) Derive (24a). HINT: $\cos A \cos B = \frac{1}{2}[\cos(A+B) + \cos(A-B)]$
(b) Derive (24b). HINT: $\sin A \sin B = \frac{1}{2}[\cos(A-B) - \cos(A+B)]$
(c) Derive (24c).

3. If both left- and right-hand derivatives of f exist at x_0 and are equal, does that imply that f is differentiable at x_0? Explain.

4. Work out the Fourier series of f, given over one period as follows. At which values of x, if any, does the series fail to converge to $f(x)$? To what values does it converge at those points?

(a) x on $(-\pi, \pi]$ (b) $|x|$ on $(-\pi, \pi]$
(c) $|x|$ on $(-2\pi, 2\pi]$ (d) 50 on $(0, 2]$
(e) 50 on $(0, 2)$, 100 at $x = 2$
(f) 50 on $-8 < x < -2$, 0 on $-2 \le x \le 4$
(g) $|\sin x|$ (for all x)
(h) $|\cos x|$ (for all x)
(i) $\sin x$ on $0 \le x < \pi$, 0 on $\pi \le x < 2\pi$
(j) $20 + 3\sin 4x$ (for all x)
(k) x on $0 < x < 1$, 1 on $1 \le x \le 2$
(l) $\cos^2 x$ on $0 < x \le \pi$ HINT: $\cos^2 x = (1 + \cos 2x)/2$
(m) $\sin^2 x$ on $0 < x \le \pi$
(n) e^{-x} on $0 \le x < 2$
(o) 100 on $0 \le x < 1$, 50 on $1 \le x < 2$, 0 on $2 \le x < 3$

5. Work out the Fourier series of the 2π-periodic function f defined on $-\pi < x \le \pi$ as follows, using computer software (such as the *Maple* int command) to evaluate the integrals. HINT: For parts (c) and (d) you will need to use l'Hôpital's rule to simplify the result.

(a) x^2 (b) x^3 (c) $\cos^2 x$ (d) $\sin^2 x$

6. Obtain a computer plot of the partial sums of the Fourier series (19) of the periodic function shown in Fig. 6, for

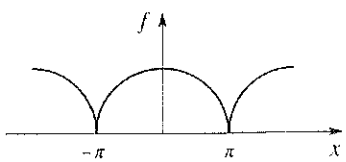
(a) $n = 2$ (b) $n = 5$ (c) $n = 10$
(d) $n = 20$ (e) $n = 30$ (f) $n = 50$

7. (a) Use (11) to show that

$$\frac{\pi}{4} = \sum_{n=1,3,\ldots}^{\infty} \frac{1}{n} \sin\frac{n\pi}{2} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots. \tag{7.1}$$

(b) If you were to use (7.1) to compute π, how many terms would be needed to obtain accuracy to six significant figures? Explain.

8. Let f be the periodic function shown in the figure, each segment of which is a semicircle of radius π. Show that its Fourier series is



$$f(x) = \frac{\pi^2}{4} + \frac{\pi^2}{2}\sum_{n=1}^{\infty}[J_0(n\pi) + J_2(n\pi)]\cos nx, \tag{8.1}$$

where the J_n's denote *Bessel functions of the first kind.* HINT: You may use any of these formulas:

$$\begin{aligned}
\cos(a\sin\theta) &= J_0 + 2J_2\cos 2\theta + 2J_4\cos 4\theta + \cdots,\\
\sin(a\sin\theta) &= 2J_1\sin\theta + 2J_3\sin 3\theta + 2J_5\sin 5\theta + \cdots,\\
\cos(a\cos\theta) &= J_0 - 2J_2\cos 2\theta + 2J_4\cos 4\theta - \cdots,\\
\sin(a\cos\theta) &= 2J_1\cos\theta - 2J_3\cos 3\theta + 2J_5\cos 5\theta - \cdots,
\end{aligned}$$

where the J_n's are shorthand for $J_n(a)$. NOTE: Observe that Theorem 17.3.1 does not apply in this case since f' is not piecewise continuous on $[-\pi, \pi]$, but it follows from a stronger version of that theorem that the series in (8.1) does converge to $f(x)$ for all x.

9. If the Fourier coefficients of a periodic function $f(x)$ are a_n $(n = 0, 1, 2, \ldots)$ and b_n $(n = 1, 2, \ldots)$, what are the Fourier coefficients A_n, B_n, say, of the shifted periodic function $f(x - c)$?

10. We have been interpreting the period 2ℓ, in equations (5), to be the *fundamental* period. However, surely (5) should yield the same result if we use twice the fundamental period, and so on. The purpose of this exercise is not to prove that claim but merely to illustrate its truth through a concrete example. Specifically, use (5) to determine the Fourier series of the square wave shown in Fig. 1, using $2\ell = 4\pi$ (rather than 2π), and show that you obtain exactly the same final result as was given in (11).

11. (*Polynomials*) It is a useful fact that if $p(x)$ is an even polynomial (i.e., containing only even-integer powers of x) on $(-\ell, \ell)$, then the b_n's are zero and

$$a_n = \frac{2\ell}{n^2\pi^2}(-1)^n\left[p'(\ell) - \frac{\ell^2}{n^2\pi^2}p'''(\ell) + \frac{\ell^4}{n^4\pi^4}p^{(v)}(\ell) - \cdots\right] \tag{11.1}$$

for $n \ge 1$; and if $p(x)$ is an odd polynomial on $(-\ell, \ell)$, then the a_n's are zero and

$$b_n = -\frac{2}{n\pi}(-1)^n\left[p(\ell) - \frac{\ell^2}{n^2\pi^2}p''(\ell) + \frac{\ell^4}{n^4\pi^4}p^{(iv)}(\ell) - \cdots\right] \tag{11.2}$$

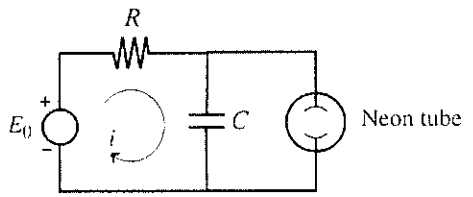
for $n \ge 1$.

(a) Derive (11.1) through the p''' term.
(b) Derive (11.2) through the p'' term.

12. Use (11.1) and/or (11.2) in Exercise 11 to obtain the Fourier series of the given periodic function. NOTE: (11.1) does not hold for $n = 0$, so you need to compute a_0 in the usual way.

(a) $f(x) = x$ on $(-3, 3)$
(b) $f(x) = x^2$ on $(-3, 3)$
(c) $f(x) = 6 + 2x - x^3$ on $(-2, 2)$
(d) $f(x) = x^2(x^2 - 8)$ on $(-2, 2)$
(e) $f(x) = x(x^2 - 3)$ on $(-1, 1)$
(f) $f(x) = 2 + x - x^2$ on $(-3, 3)$
(g) $f(x) = x + x^2$ on $(-1, 1)$

13. (*Relaxation oscillator*) There exist a great many periodic occurrences, called **relaxation oscillations**, which are characterized by a slow buildup and a rapid discharge. Examples include epidemics, economic crises, the sound generated by the bowing of a violin string, shivering from the cold, menstruation, the beating of the heart, and the discharge of a capacitor through a neon tube. Here we consider the latter (see the following figure). The resistance R, capacitance

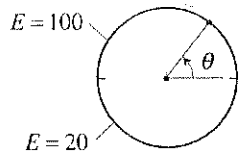


C, and voltage E_0 are constants. If $Q(t)$ is the charge on the capacitor, then $i(t) = dQ/dt$ is the current in the E_0, R, C loop, and Kirchhoff's law gives

$$R\frac{dQ}{dt} + \frac{1}{C}Q = E_0. \tag{13.1}$$

(a) Solve (13.1) subject to the initial condition $Q(0) = 0$.
(b) When the voltage Q/C on the capacitor, and hence on the neon tube, reaches a certain level, say $E_0/2$, the neon tube fires, and the charge Q on the capacitor drops to zero. The firing is so rapid compared to the "buildup time" that we may assume it to be instantaneous. From your solution to part (a), show that the neon tube fires at $t = RC \ln 2$.
(c) The cycle then repeats indefinitely. Sketch and label the graph of the periodic function $Q(t)$, and work out its Fourier series.

14. The voltage $E(\theta)$ is maintained at 100 volts on the top edge of the disk and at 20 volts on the bottom edge. Expand $E(\theta)$ in a Fourier series. HINT: Sketch the graph of $E(\theta)$ on $-\infty < \theta < \infty$.



15. (*Rms current*) If a steady electric current i flows through a resistor of resistance R, the power delivered (i.e., the rate of doing work) is equal to i^2R. In many applications i is not a constant, but a periodic function of the time t. In such cases one defines the *average power* as

$$\text{average power} = \frac{1}{T}\int_{-T/2}^{T/2} i^2(t)R\,dt, \tag{15.1}$$

where T is the fundamental period of $i(t)$. Expressing $i(t)$ as a Fourier series,

$$i(t) = a_0 + \sum_{n=1}^{\infty}\left(a_n \cos\frac{2n\pi t}{T} + b_n \sin\frac{2n\pi t}{T}\right), \tag{15.2}$$

show that the

$$\text{average power} = \left[a_0^2 + \frac{1}{2}\sum_{n=1}^{\infty}(a_n^2 + b_n^2)\right]R, \tag{15.3}$$

and that the steady current that is equivalent to $i(t)$, in that it will deliver the same power, is

$$I_{\text{rms}} = \sqrt{a_0^2 + \frac{1}{2}\sum_{n=1}^{\infty}(a_n^2 + b_n^2)}. \tag{15.4}$$

I_{rms} is known as the **root-mean square (rms) current.**

16. (*Complex form*) Work out the complex exponential form of the Fourier series (55), for the given periodic function, defined over one period as follows.

(a) $f(x) = 50$ on $|x| \le 1$, and 0 on $1 < |x| \le 2$
(b) $f(x) = e^x$ on $0 \le x < 2$
(c) $f(x) = e^{-x}$ on $-3 \le x < 3$
(d) $f(x) = 6\sin x$ (for all x)
(e) $f(x) = 4 - 5\cos 2x$ (for all x)
(f) $f(x) = -100$ on $-5 \le x < 0$, and 100 on $0 \le x < 5$
(g) $f(x) = 20 + 5x$ on $-2 \le x < 2$

17. (*Example 3*) (a) In Example 3 we expand $F(t)$ in a Fourier series. Could we have used a Taylor series, profitably, instead?
(b) It is proposed that we solve (30) in a piecewise manner. That is, since $F(t)$ is piecewise constant we might do well to consider the t intervals $0 < t < a$, $a < t < 2\pi - a$, $2\pi - a < t < 2\pi + a$, and so on. Solving $mx'' + cx' + kx$

$= \frac{1}{2a}$ on $0 < t < a$, subject to initial conditions $x(0) = x_0$ and $x'(0) = x'_0$, say, we can then use the final values $x(a)$ and $x'(a)$ from the first interval as the initial conditions for the next interval, $a < t < 2\pi - a$, on which $mx'' + cx' + kx = 0$, and so on. You don't need to carry out this method; rather, we merely ask you to assess whether it is a good idea for determining the steady-state response.
(c) Would the method used in Example 3 work if (30) were modified to

$$mx'' + cx' + \alpha x + \beta x^3 = F(t)?$$

Explain.
(d) If we let $a \to 0$ in Example 3, then (see Fig. 8) $F(t)$ becomes a series of delta functions (unit impulses), at $t = 0, 2\pi, 4\pi, \ldots$. (Actually, the one at $t = 0$ will be *half* a delta function.) Find the Fourier series representation of the steady-state response.
(e) Same as part (d), but find a closed form expression of the steady-state response $x(t)$ over one period, and sketch its graph.

18. Consider an undamped, driven mechanical oscillator, governed by the equation

$$mx'' + kx = F(t).$$

Solve for the steady-state response $x(t)$, if $F(t)$ is periodic and defined (over one period) as follows. Let $m = k = 1$.

(a) $F(t) = 100$ on $0 \le t < 2$, and 0 on $2 \le t < 4$
(b) $F(t) = 30$ on $0 \le t < 2$, and -30 on $2 \le t < 4$
(c) $F(t) = 5t$ on $0 \le t < 1$, and $10 - 5t$ on $1 \le t < 2$
(d) $F(t) = 5t$ on $0 \le t < 1$, and $5t - 10$ on $1 \le t < 2$
(e) $F(t) = 20$ on $0 \le t < 2$, and 10 on $2 \le t < 4$, and 0 on $4 \le t < 6$
(f) $F(t) = 10t$ on $0 \le t < 3$, and 0 on $3 \le t < 4$

17.4 Half- and Quarter-Range Expansions

It often happens in applications, especially when we solve partial differential equations by the method of separation of variables (Chapters 18–20), that we need to expand a given function f in a Fourier series, where f is defined only on a finite interval such as the function f whose graph is shown in Fig. 1, and which is defined only on $0 < x < L$.* But in that case f is not periodic, so how can we expand it in a Fourier series?

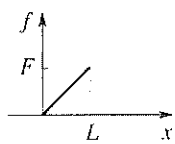


Figure 1. f on finite interval.

The idea is to extend the domain of definition of f to $-\infty < x < \infty$ and to define an "extended function," say f_{ext}, so that f_{ext} is periodic, with $f_{\text{ext}}(x) = f(x)$ on the original interval $0 < x < L$. There are an infinite number of such extensions, two of which are shown in Fig. 2. Each of these is periodic, the first with period $2L$ and the second with period L. Their Fourier series are different, but each of them converges to the original function f on the original interval $0 < x < L$.

How, then, do we know which extension to use? We shall see that the choice will be dictated by the context, so let us not worry about that right now. We will always need to choose from among four extensions which are known as half- and quarter-range cosine and sine extensions and which are based on symmetry or antisymmetry about the endpoints $x = 0$ and $x = L$. These are shown, for the representative function f of Fig. 1, in Fig. 3. For instance, in Fig. 3a f_{ext} is symmetric about $x = 0$ and also about $x = L$, hence the two S's below the x axis.

*We use the open interval notation throughout this section since the values of $f(x)$ at the endpoints will not affect the Fourier coefficients, as we learned in the foregoing section.

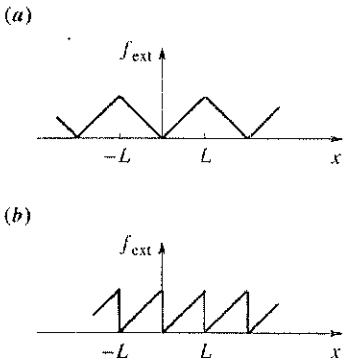


Figure 2. Possible extensions.

Because of its symmetry about $x = 0$, f_{ext} is an even function, and its Fourier series will contain only cosines, no sines. Further, its period is $2L$, so L is half the period. Thus, it is customary to designate this case as the "half-range cosine extension," which we denote in this text by the letters HRC. In Fig. 3b, f_{ext} is antisymmetric about $x = 0$ and $x = L$, the period is $2L$, and we have the half-range sine extension, denoted by HRS. In Fig. 3c f_{ext} is symmetric about $x = 0$ and antisymmetric about $x = L$, the period is $4L$ (so L is only a quarter of the period), and we have the quarter-range cosine extension, denoted by QRS. Similarly for the quarter-range sine extension shown in Fig. 3d and denoted by QRS.

Let us derive the Fourier series for these cases. For the half-range cosine case the period is $2L$, so $\ell = L$ (where we carry over the ℓ notation from Section 17.3) and

$$a_0 = \frac{1}{2L}\int_{-L}^{L} f_{\text{ext}}(x)\,dx = \frac{1}{L}\int_0^L f(x)\,dx, \tag{1a}$$

$$a_n = \frac{1}{L}\int_{-L}^{L} f_{\text{ext}}(x)\cos\frac{n\pi x}{L}\,dx = \frac{2}{L}\int_0^L f(x)\cos\frac{n\pi x}{L}\,dx, \tag{1b}$$

$$b_n = \frac{1}{L}\int_{-L}^{L} f_{\text{ext}}(x)\sin\frac{n\pi x}{L}\,dx = 0. \tag{1c}$$

The last step in (1a) follows because $f_{\text{ext}}(x)$ is symmetric about $x = 0$. Similarly, the symmetry of $f_{\text{ext}}(x)\cos(n\pi x/L)$ about $x = 0$ gives (1b), and the antisymmetry of $f_{\text{ext}}(x)\sin(n\pi x/L)$ about $x = 0$ gives (1c). Thus, we can write

$$f_{\text{ext}}(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos\frac{n\pi x}{L}, \qquad (-\infty < x < \infty) \tag{2}$$

with a_0 and a_n given by (1a) and (1b).

Understand that the extension was only an artifice, to make possible a Fourier representation of the original function f on the original interval $0 < x < L$. With the expansion in hand, we can now limit our attention to the $0 < x < L$ interval and write

$$\boxed{\begin{gathered} f(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos\frac{n\pi x}{L}, \qquad (0 < x < L) \\ a_0 = \frac{1}{L}\int_0^L f(x)\,dx, \qquad a_n = \frac{2}{L}\int_0^L f(x)\cos\frac{n\pi x}{L}\,dx. \end{gathered}} \tag{3}$$

We call (3) the **half-range cosine expansion** of f. Observe that the final formulas in (3) contain no artifacts of the extension, only f defined on $0 < x < L$.

By a similar process we obtain the **half-range sine expansion**

$$\boxed{\begin{gathered} f(x) = \sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{L}, \qquad (0 < x < L) \\ b_n = \frac{2}{L}\int_0^L f(x)\sin\frac{n\pi x}{L}\,dx, \end{gathered}} \tag{4}$$

the **quarter-range cosine expansion**

$$\boxed{\begin{aligned} f(x) &= \sum_{n=1,3,\ldots}^{\infty} a_n \cos\frac{n\pi x}{2L}, \qquad (0 < x < L) \\ a_n &= \frac{2}{L}\int_0^L f(x)\cos\frac{n\pi x}{2L}\,dx, \end{aligned}} \tag{5}$$

and the **quarter-range sine expansion**

$$\boxed{\begin{aligned} f(x) &= \sum_{n=1,3,\ldots}^{\infty} b_n \sin\frac{n\pi x}{2L}, \qquad (0 < x < L) \\ b_n &= \frac{2}{L}\int_0^L f(x)\sin\frac{n\pi x}{2L}\,dx. \end{aligned}} \tag{6}$$

We have already derived (3). Let us now derive (6) and leave (4) and (5) for the exercises. We see from Fig. 3d that the period is $4L$, so $\ell = 2L$. Thus

$$f_{\text{ext}}(x) = a_0 + \sum_{n=1}^{\infty}\left(a_n \cos\frac{n\pi x}{2L} + b_n \sin\frac{n\pi x}{2L}\right). \tag{7}$$

The antisymmetry of f_{ext} about $x = 0$ implies that $a_0 = 0$ and $a_n = 0$ for each $n = 1, 2, \ldots$. Next,

$$b_n = \frac{1}{2L}\int_{-2L}^{2L} f_{\text{ext}}(x)\,\sin\frac{n\pi x}{2L}\,dx = \frac{2}{2L}\int_0^{2L} f_{\text{ext}}(x)\,\sin\frac{n\pi x}{2L}\,dx \tag{8}$$

because both $f_{\text{ext}}(x)$ and $\sin(n\pi x/2L)$ are symmetric about $x = 0$, so their product is symmetric about $x = 0$.* Now consider the symmetry or antisymmetry about $x = L$. We see from Fig. 3d that $f_{\text{ext}}(x)$ is symmetric about that point. Further, it is easily verified that $\sin(n\pi x/2L)$ is symmetric about that point if n is odd, and antisymmetric about that point if n is even. Thus, the integrand in the second integral in (8) is symmetric about $x = L$ (which is the midpoint of the integration interval) if n is odd, and antisymmetric about that point if n is even. Hence (8) gives

$$\begin{aligned} b_n &= \frac{1}{L}\int_0^{2L} f_{\text{ext}}(x)\,\sin\frac{n\pi x}{2L}\,dx \\ &= \begin{cases} 0 & , \quad n \text{ even} \\ \dfrac{2}{L}\displaystyle\int_0^L f_{\text{ext}}(x)\,\sin\frac{n\pi x}{2L}\,dx\ , & n \text{ odd.} \end{cases} \end{aligned} \tag{9}$$

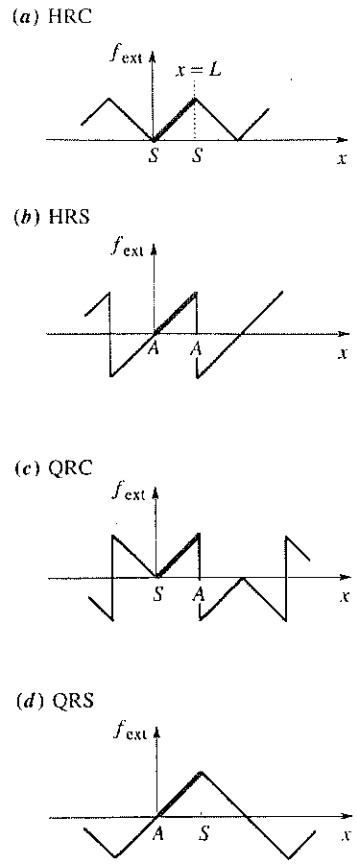


Figure 3. Half- and quarter-range extensions.

*Understand the difference between symmetric/antisymmetric and even/odd. A graph can be symmetric about *any* x point. If, in particular, the graph of f is symmetric/antisymmetric about the *origin*, then f is even/odd, respectively.

Finally, we can drop the subscripted "ext" in the final integral in (9) because the interval of integration is $0 < x < L$, over which interval $f_{\text{ext}}(x) = f(x)$. Putting these expressions for a_0, a_n, and b_n into (7) gives (6).

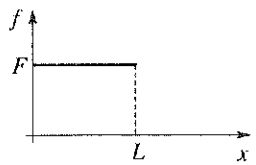


Figure 4. f in Example 1.

EXAMPLE 1. To illustrate, let us expand the function f, displayed in Fig. 4, in half- and quarter-range cosine and sine series.

HRC: From (3),

$$\begin{aligned} a_0 &= \frac{1}{L}\int_0^L F\,dx = \frac{FL}{L} = L, \\ a_n &= \frac{2}{L}\int_0^L F\cos\frac{n\pi x}{L}\,dx = \frac{2F}{n\pi}\sin\frac{n\pi x}{L}\Big|_{x=0}^{x=L} = 0, \end{aligned} \tag{10}$$

so the HRC expansion of f is simply

$$f(x) = F, \tag{11}$$

which results from the extension shown in Fig. 5a.

HRS: From (4),

$$b_n = \frac{2}{L}\int_0^L F\sin\frac{n\pi x}{L}\,dx = -\frac{2F}{n\pi}(\cos n\pi - 1), \tag{12}$$

so the HRS expansion of f is

$$f(x) = \frac{2F}{\pi}\sum_{n=1}^{\infty}\frac{1 - cosn\pi}{n}\sin\frac{n\pi x}{L}, \tag{13}$$

which results from the extension shown in Fig. 5b.

QRC: From (5),

$$a_n = \frac{2}{L}\int_0^L F\cos\frac{n\pi x}{2L}\,dx = \frac{4F}{n\pi}\sin\frac{n\pi}{2}, \tag{14}$$

so the QRC expansion of f is

$$f(x) = \frac{4F}{\pi}\sum_{n=1,3,\ldots}^{\infty}\frac{sin(n\pi/2)}{n}\cos\frac{n\pi x}{2L}, \tag{15}$$

which results from the extension shown in Fig. 5c.

QRS: From (6),

$$b_n = \frac{2}{L}\int_0^L F\sin\frac{n\pi x}{2L}\,dx = \frac{4F}{n\pi}(1 - \cos\frac{n\pi}{2}) = \frac{4F}{n\pi} \tag{16}$$

because $\cos\frac{n\pi}{2} = 0$ if n is odd, and n is indeed odd in (6). Thus, the QRS expansion of f is

$$f(x) = \frac{4F}{\pi}\sum_{n=1,3,\ldots}^{\infty}\frac{1}{n}\sin\frac{n\pi x}{2L}, \tag{17}$$

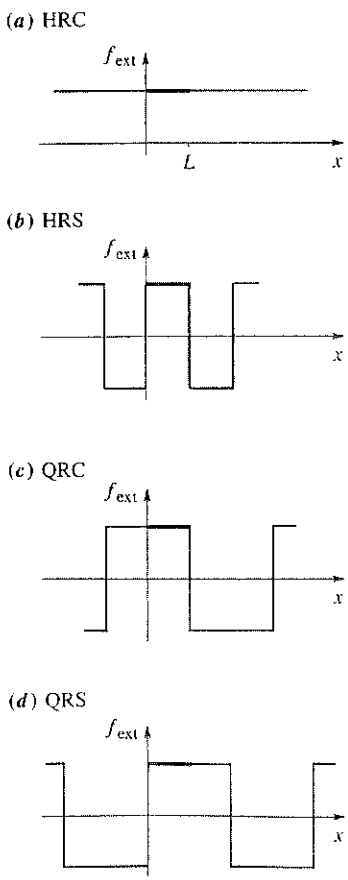


Figure 5. Half- and quarter-range extensions of f.

which results from the extension shown in Fig. 5d.

The series (11), (13), (15), and (17) converge to the functions shown in Fig. 5a,b,c,d, respectively, but on the interval $0 < x < L$ they all converge to the function f defined in Fig. 4.

COMMENT. Observe that n runs from 1 to infinity continuously in the half-range formulas (3) and (4), but only through the odd values $1, 3, \ldots$ in the quarter-range formulas (5) and (6). In the present example, the sum in the HRS expansion (13) runs through $n = 1, 3, \ldots$ because $1 - \cos n\pi = 1 - (-1)^n = 0$ for even n's, but in general the sums in the half-range expansions run through $n = 1, 2, 3, \ldots$. ∎

Closure. Functions defined only on a finite interval, say $0 < x < L$, can be periodically extended in an infinite number of ways. Of these, only four possible extensions will prove to be of interest, the half- and quarter-range extensions, which lead to the half- and quarter-range cosine and sine series (3)–(6). Application of these results occur repeatedly in Chapters 18–20 when we solve partial differential equations by the method of separation of variables.

EXERCISES 17.4

1. We derived the HRC and QRS formulas, (3) and (6), respectively. In similar fashion, derive the

(a) HRS formulas (4) (b) QRC formulas (5)

2. For the given function, prepare labeled sketches of the HRC, HRS, QRC, and QRS extensions, and derive the corresponding expansions.

(a) $f(x) = 25$ on $0 < x < 1$, and 0 on $1 < x < 2$

(b) $f(x) = 5x$ on $0 < x < 4$

(c) $f(x) = \sin x$ on $0 < x < \pi$

(d) $f(x) = \sin x$ on $0 < x < \pi/2$

(e) $f(x) = 1 - x$ on $0 < x < 1$

(f) $f(x) = 0$ on $0 < x < 4$, and 50 on $4 < x < 5$

(g) $f(x) = \sin x$ on $0 < x < \pi$, and 0 on $\pi < x < 2\pi$

(h) $f(x) = 2 + x$ on $0 < x < 3$

17.5 Manipulation of Fourier Series (Optional)

In Examples 3 and 4 of Section 17.4 we differentiated Fourier series termwise and equated coefficients of trigonometric series on the left- and right-hand sides of an equation. Our approach is formal (i.e., we do not rigorously justify the steps) since we did not yet have theorems in place with which to justify those steps. This section is intended to provide the necessary theorems.

Fundamental is the concept of uniform convergence.* Consider a series

$$\sum_{n=1}^{\infty} a_n(x), \tag{1}$$

and let $s_n(x)$ be its nth partial sum

$$s_n(x) = a_1(x) + \cdots + a_n(x). \tag{2}$$

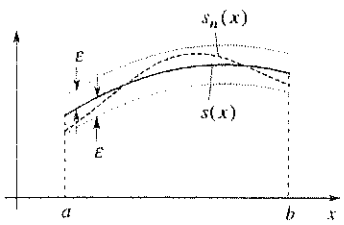


Figure 1. Uniform convergence.

We say that (1) **converges uniformly** to $s(x)$ on an interval $a \le x \le b$ if to each $\epsilon > 0$ (i.e., no matter how small) there corresponds an $N(\epsilon)$, independent of x, such that $|s_n(x) - s(x)| < \epsilon$ for all $n > N$ and for all x in the interval $a \le x \le b$.† The situation is illustrated in Fig. 1. Supposing that $s(x)$ is as shown, choose an arbitrarily small $\epsilon > 0$ and draw an "ϵ band" about $s(x)$, the band between $s(x) - \epsilon$ and $s(x) + \epsilon$. If (1) converges uniformly, then (no matter how small we choose ϵ to be) there must be some integer N such that the graph of $s_n(x)$ lies entirely within the ϵ band for all n's greater than N.

EXAMPLE 1. Let (1) be the Fourier series

$$f(x) = 2 + \frac{8}{\pi} \sum_{n=1,,3,\ldots}^{\infty} \frac{1}{n} \sin nx \tag{3}$$

of the square wave, one period of which is shown in Fig. 2. We know that that series converges to the square wave shown in the figure: 0 for $-\pi < x < 0$, 2 at $x = 0$ (denoted by the heavy dot), and 4 for $0 < x < \pi$). Is that convergence uniform? The answer depends upon the interval under consideration. Over $1 \le x \le 2$, for instance, the answer is yes, which we state without proof; for any $\epsilon > 0$, no matter how small, we can force the graph of $s_n(x)$ to fall within the ϵ band by taking n sufficiently large.

However, over any interval containing a jump discontinuity of f, such as $-1 \le x \le 1$, the answer is no because the partial sums are continuous, so the graph of $s_n(x)$ must inevitably break out of the ϵ bands (at P and Q) in order to pass through the heavy dot. In addition, there are breakouts due to the Gibbs phenomenon, the spike-like overshoots and undershoots near the jump discontinuities. ∎

Of course we cannot rely on pictures, we need an analytical test for uniformity

*The concept of uniform convergence is due to *Karl Weierstrass* (1841) and *G. G. Stokes* (1847). Generally speaking, it is probably true that major theorems attract the most acclaim in mathematics but, in truth, the fundamental definitions – such as the definitions of the *limit* of a function, the *derivative* of a function, the *uniform convergence* of a series, and so on, are of comparable importance. It is easy to invent definitions, but not all definitions are fruitful. Consider, for instance, the way all of the differential and integral calculus, as well as other branches of analysis, rest upon the shoulders of the limit concept.

†Contrast this definition with the definition of convergence of (1) at x: (1) converges to $s(x)$ at x if to each $\epsilon > 0$ (i.e., no matter how small) there corresponds an $N(\epsilon)$ such that $|s_n(x) - s(x)| < \epsilon$ for all $n > N$.

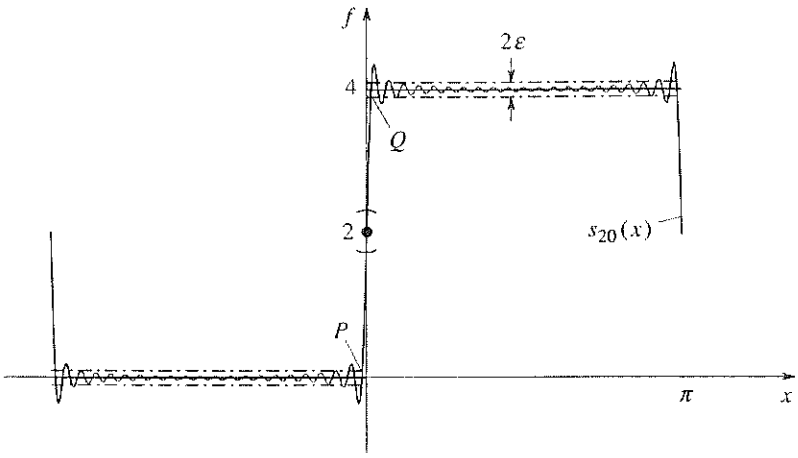


Figure 2. Square wave Fourier series.

of convergence. A useful sufficient condition for uniform convergence is as follows.

THEOREM 17.5.1 *Weierstrass M-Test*
If $\sum_{n=1}^{\infty} M_n$ is a convergent series of positive constants and $|a_n(x)| \leq M_n$ on an x interval I, then $\sum_{n=1}^{\infty} a_n(x)$ is uniformly (and absolutely) convergent on I.

EXAMPLE 2. The series

$$\sum_{n=1}^{\infty} \frac{\cos nx}{n^2} \tag{4}$$

converges uniformly on $-\infty < x < \infty$, according to the Weierstrass M-test, because $|(\cos nx)/n^2| \leq 1/n^2$ on $-\infty < x < \infty$, and $\sum_{n=1}^{\infty} 1/n^2$ is convergent; specifically, it is a convergent p-series.* That is, M_n is $1/n^2$ in this case. ∎

We can now state a useful theorem on the termwise differentiation of an infinite series.

THEOREM 17.5.2 *Termwise Differentiation of Series*
Let $\sum_{n=1}^{\infty} a_n(x)$ converge on an x interval I. Then

$$\boxed{\frac{d}{dx} \sum_{n=1}^{\infty} a_n(x) = \sum_{n=1}^{\infty} \frac{d}{dx} a_n(x)} \tag{5}$$

*Recall from the calculus that the **p-series** $\sum_{n=1}^{\infty} 1/n^p$ converges if $p > 1$ and diverges if $p \leq 1$. If $p = 1$ it becomes the well known (and divergent) **harmonic series**.

if the series on the right converges uniformly on I.*

That is, we can interchange the order of the two limit operations (the infinite series is the limit of the sequence of partial sums, and the derivative is the limit of a difference quotient) if the $a_n(x)$'s are well enough behaved, specifically, if $\sum a_n'(x)$ converges uniformly on I. That condition is stated as sufficient, not necessary. To put (5) into perspective, it might be helpful to recall the analogous result for interchanging the order of differentiation and integration: if a and b are constants, then

$$\frac{d}{dx}\int_a^b f(t,x)\,dt = \int_a^b \frac{\partial f}{\partial x}(t,x)\,dt \tag{6}$$

if f is sufficiently well behaved, namely, if f and $\partial f/\partial x$ are continuous on the relevant rectangle in the x, t plane (see the Leibniz rule, Theorem 13.8.1).

EXAMPLE 3. The series

$$\sum_{n=1}^{\infty} \frac{\sin nx}{n^3} \tag{7}$$

can be differentiated termwise on $-\infty < x < \infty$. That is,

$$\frac{d}{dx}\sum_{n=1}^{\infty}\frac{\sin nx}{n^3} = \sum_{n=1}^{\infty}\frac{d}{dx}\left(\frac{\sin nx}{n^3}\right) = \sum_{n=1}^{\infty}\frac{\cos nx}{n^2}, \tag{8}$$

because the series on the right converges uniformly for all x, as shown in Example 2. ∎

Besides the termwise differentiations employed in the representative Examples 3 and 4 of Section 17.3, we also equated the corresponding coefficients of trigonometric series on the left- and right-hand sides of an equation. Justification of that step is provided by the following theorem.

THEOREM 17.5.3 *Uniqueness of Trigonometric Series*
If

$$a_0+\sum_{n=1}^{\infty}\left(a_n\cos\frac{n\pi x}{\ell}+b_n\sin\frac{n\pi x}{\ell}\right) = A_0+\sum_{n=1}^{\infty}\left(A_n\cos\frac{n\pi x}{\ell}+B_n\sin\frac{n\pi x}{\ell}\right), \tag{9}$$

where the trigonometric series on the left- and right-hand sides converge to the

*For proof of this theorem and additional discussion of this issue, see T. M. Apostol, *Mathematical Analysis* (Reading, MA: Addison–Wesley, 1957), Chap.13.

same sum for all x, then $a_0 = A_0$, $a_n = A_n$, and $b_n = B_n$ for each n.*

Notice the wording "trigonometric series" for the theorem. Recall that a trigonometric series is a series of the form given by the left- and right-hand sides of (9). Why don't we say Fourier series instead? Because not every convergent trigonometric series is a Fourier series. For example,

$$\sum_{n=1}^{\infty} \frac{1}{\ln(n+1)} \sin nx \tag{10}$$

is a trigonometric series, and it can be shown (by a Dirichlet test) to converge for all x, yet it is not a Fourier series.* That is, there does not exist a 2π-periodic function f such that

$$\int_{-\pi}^{\pi} f(x) \cos nx \, dx = 0, \qquad (n = 0, 1, 2, \ldots) \tag{11a}$$

$$\int_{-\pi}^{\pi} f(x) \sin nx \, dx = \frac{1}{\ln(n+1)}. \qquad (n = 1, 2, \ldots) \tag{11b}$$

Thus, every Fourier series is a trigonometric series, but not every convergent trigonometric series is a Fourier series. This result is in interesting contrast with the result, encountered later in this text (Theorem 24.2.8), that every convergent power series is the Taylor series of its sum function. However, if a trigonometric series, with period 2ℓ, converges *uniformly* on $[-\ell, \ell]$ (or, equivalently, on $-\infty < x < \infty$), then it is the Fourier series of its sum function.†

Let us illustrate these results with a practical application that is similar to Examples 3 and 4 in Section 17.3.

EXAMPLE 4. Find a particular solution to the differential equation

$$x'' + 0.5x = f(t), \qquad (0 < t < \infty) \tag{12}$$

where the forcing function $f(t)$ is the 2π-periodic function shown in Fig. 3. The Fourier of f is

$$f(t) = \frac{8F}{\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{(2n-1)^2} \sin(2n-1)t, \tag{13}$$

and if we seek $x(t)$ in the form

$$x(t) = \sum_{n=1}^{\infty} b_n \sin(2n-1)t \tag{14}$$

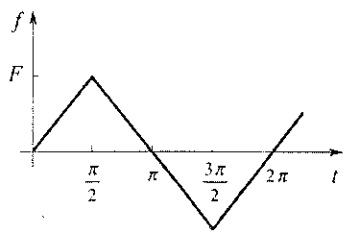


Figure 3. $f(t)$ in Example 4.

*Proof of this sophisticated theorem can be found in E. C. Titchmarsh, *The Theory of Functions*, 2nd ed. (London: Oxford University Press, 1939), pp. 427–432.

*See I. S. Sokolnikoff and R. M. Redheffer, *Mathematics of Physics and Modern Engineering*, 2nd ed. (New York: McGraw Hill, 1966), p. 86, Exer.3.

†See G. P. Tolstov, *Fourier Series* (Englewood Cliffs, NJ: Prentice–Hall, 1962), pp. 14–15.

we find (Exercise 1), formally, that

$$x(t) = \frac{8F}{\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^n}{(2n-1)^2[(2n-1)^2 - 0.5]} \sin(2n-1)t. \tag{15}$$

We can make the process rigorous either by justifying each of the steps in the derivation of (15a), using Theorems 17.5.1–17.5.3, or we can proceed formally to (15) and then rigorously verify that the latter does satisfy the given differential equation (12). Let us do the latter.

According to Theorem 17.5.2 we can differentiate (15) termwise and obtain

$$x'(t) = \frac{8F}{\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^n}{(2n-1)[(2n-1)^2 - 0.5]} \cos(2n-1)t \tag{16}$$

because the latter series is uniformly convergent on $0 < t < \infty$. That is,

$$\left| \frac{(-1)^n}{(2n-1)[(2n-1)^2 - 0.5]} \cos(2n-1)t \right| \le \frac{1}{(2n-1)[(2n-1)^2 - 0.5]} = M_n, \tag{17}$$

and $\sum_{n=1}^{\infty} M_n$ is convergent because $M_n \sim 1/8n^3$ as $n \to \infty$, where $\sum_{n=1}^{\infty}(1/8n^3) = \frac{1}{8}\sum_{n=1}^{\infty} 1/n^3$ is $\frac{1}{8}$ times a convergent p-series, with $p = 3$.* Likewise, we can differentiate (16) termwise and obtain

$$x''(t) = -\frac{8F}{\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^n}{(2n-1)^2 - 0.5} \sin(2n-1)t \tag{18}$$

because the latter series is uniformly convergent on $0 < t < \infty$. That is,

$$\left| \frac{(-1)^n}{(2n-1)^2 - 0.5} \sin(2n-1)t \right| \le \frac{(-1)^n}{(2n-1)^2 - 0.5} = M_n, \tag{19}$$

where $\sum_{n=1}^{\infty} M_n$ is convergent because $M_n \sim 1/4n^2$ as $n \to \infty$.

Finally, putting (18), (15), and (13) into (12) and adding the two series on the left-hand side termwise† is found to produce an identity, so the verification of (15) is complete.

COMMENT. Observe that each time we differentiate (15) we weaken the convergence because we pull out a $2n-1$ from the sine or cosine. Since f was C^0 (i.e., continuous), the terms in (13) were of order $O(1/n^2)$ and, consequently, the terms in (15) were $O(1/n^4)$, as $n \to \infty$. Thus, the terms in $x'(t)$ were $O(1/n^3)$ and those in $x''(t)$ were $O(1/n^2)$, the n^2 being sufficient to render the series uniformly convergent so as to justify the termwise differentiation of (16). If $f(t)$ had been even better behaved such as C^1 or C^2, then so much the better, but if it had been *dis*continuous such as a square wave, then our efforts at justification would have failed because then the terms in the $x''(t)$ series would die out like

*Recall from the calculus that if $\sum a_n$ is a series of positive constants and $a_n \sim Kb_n$ as $n \to \infty$, for some positive constant K, then the series $\sum a_n$ and $\sum b_n$ both converge or both diverge.

†Recall from the calculus that if $\sum a_n$ and $\sum b_n$ are convergent, then $\sum a_n + \sum b_n = \sum(a_n + b_n)$, and the latter is convergent as well.

$O(1/n)$, which would not be fast enough to ensure uniform convergence by the Weierstrass M-test (since the series $\sum 1/n$ is divergent). ∎

Actually, the idea of finding solutions to differential equations that are in the form of infinite series is not new to us since we have already studied *power series* solutions in Chapter 4. Indeed, the technical issues that arose there are the same: the termwise differentiation of a power series, the termwise addition of two series, and the fact that if two power series are equal, then their corresponding coefficients must be equal. These matters are handled by parts (a), (b), (c), respectively, of Theorem 4.2.3. There, we saw that power series are very easy to work with, especially in that a power series may be differentiated (or integrated) any number of times (within its interval of convergence). Of course, that fact is not surprising since a convergent power series is the Taylor series of its sum function, and for a function to have a Taylor series it must be extremely well-behaved, for instance, infinitely differentiable (C^∞). However, in applications like Example 4, above, power series are of no help because $f(t)$ is merely continuous; it is not even once differentiable because of the kinks in its graph at $t = \pi/2, 3\pi/2, 5\pi/2, \ldots$. But since f is periodic we are able to use Fourier series.

The fact that the Fourier series (15) of $x(t)$ can be differentiated termwise only twice is not surprising since $x(t)$ is not infinitely differentiable, it is only C^2 (since the input f is C^0 and x is obtained from f, in effect, by two integrations).

Though we have not used termwise integration of Fourier series we would be remiss if we did not include the following theorem.

THEOREM 17.5.4 *Termwise Integration of Fourier Series*
If a Fourier series is integrated termwise between any finite limits, the resulting series converges to the integral of the periodic function corresponding to the original series.

EXAMPLE 5. Let us integrate the square wave f shown in Fig. 4, say, from 0 to any point x. From Example 1 of Section 17.3 we have, for the Fourier series of f,

$$f(x) = 2 + \frac{8}{\pi} \sum_{n=1,3,\ldots}^{\infty} \frac{1}{n} \sin nx. \tag{20}$$

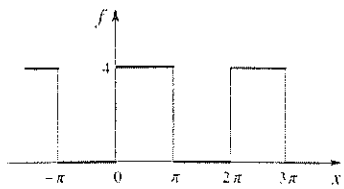


Figure 4. Square wave f.

Thus,

$$\int_0^x f(\xi)\,d\xi = \int_0^x \left[2 + \frac{8}{\pi} \sum_{n=1,3,\ldots}^{\infty} \frac{1}{n} \sin n\xi\right] d\xi$$

$$= \int_0^x 2\,d\xi + \frac{8}{\pi} \sum_{n=1,3,\ldots}^{\infty} \frac{1}{n} \int_0^x \sin n\xi\,d\xi$$

$$= 2x + \frac{8}{\pi} \sum_{n=1,3,\ldots}^{\infty} \frac{1 - \cos nx}{n^2}, \tag{21}$$

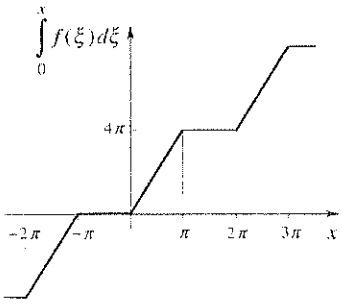


Figure 5. Sum of the integrated series.

where the second equality follows from Theorem 17.5.4. Further, the final right-hand side, in (21) must, according to the theorem, converge to $\int_0^x f(\xi)\,d\xi$, the graph of which can be inferred by inspection of Fig. 4 and is drawn in Fig. 5.

COMMENT. The series (21) is not quite a Fourier series because of the $2x$ term, which arises because the average value of f [corresponding to the 2 in (20)] is nonzero. ∎

Closure. This section is about technical matters related to the manipulation of Fourier series, especially regarding their termwise differentiation. The latter is important because functions represented by Fourier series are often not very well behaved; for instance, they may be C^1, C^0, or even discontinuous. And since differentiation makes bad behavior even worse,* one is well advised to approach the differentiation of Fourier series with caution. Theorem 17.5.2 tells us that termwise differentiation is permissible if the resulting series is uniformly convergent, although that condition is stated as sufficient, not necessary. In turn, Theorem 17.5.1 gives us a simple and well known test for uniformity of convergence, the Weierstrass M-test.

Whereas each termwise differentiation of a Fourier series pulls a factor of n out of the sine and cosine terms, thus retarding the speed of convergence, termwise integration introduces factors of $1/n$ and therefore enhances the convergence. In fact, Theorem 17.5.4 tells us that *every* Fourier series can be safely integrated termwise.

EXERCISES 17.5

1. Assuming the form (14), derive the solution (15).

2. Show that the given series is uniformly convergent on the given interval.

(a) $\sum_1^\infty e^{-nx} \sin nx$ on $2 < x < 5$

(b) $\sum_1^\infty \dfrac{\sin 2nx}{n^2 - 2n + 2}$ on $-\infty < x < \infty$

(c) $\sum_0^\infty \dfrac{e^{-nx}}{n^2 + 5}$ on $0 < x < \infty$

(d) $\sum_3^\infty \dfrac{1}{n^2 + x^2}$ on $-\infty < x < \infty$

(e) $\sum_1^\infty \ln\left(1 + \dfrac{x^3}{n^3}\right)$ on every finite interval

2. Show that $\sum_1^\infty n^{-x}$ converges uniformly on $x_0 \le x < \infty$, for every $x_0 > 1$.

3. Determine an x interval on which the series is uniformly convergent. NOTE: You may need to use one of the standard convergence tests, from the calculus.

(a) $\sum_1^\infty \dfrac{1}{1 + x^n}$

(b) $\sum_1^\infty n(\cos x)^n$

(c) $\sum_1^\infty n^5(\sin x)^n$

(d) $\sum_6^\infty (5x)^n$

(e) $\sum_1^\infty \dfrac{x^n}{n^3}$

(f) $\sum_1^\infty (x^2 + 2x - 1)^n$

(g) $\sum_3^\infty (x^2 - 2x)^{3n}$

4. Differentiate the given series termwise, and verify the validity of that step if x is in the given interval (which is not necessarily the broadest one possible).

(a) $\sum_1^\infty \dfrac{\sin 2nx}{n^4}$, $-\infty < x < \infty$

*See Comment 1 in Example 4, Section 17.3.

(b) $4 + \sum_1^\infty \frac{\cos n\pi x}{n^3+1}, \quad -\infty < x < \infty$

(c) $\sum_1^\infty (4x)^n, \quad -0.2 \le x \le 0.1$

(d) $\sum_1^\infty \frac{x^n}{n^3}, \quad -1 \le x \le 1$

(e) $\sum_1^\infty e^{-n} \sin nx, \quad -\infty < x < \infty$

5. Derive a particular solution to the forced oscillator equation

$$x'' + x' + x = f(t), \qquad (0 < t < \infty) \tag{5.1}$$

where $f(t)$ is periodic and defined over one period as follows, and then rigorously verify that your solution does indeed satisfy (5.1).

(a) $f(t) = \sin t$ on $0 \le t \le \pi$
(b) $f(t) = |\cos t|$ on $0 \le t \le \pi$
(c) $f(t) = \sin t$ on $0 \le t \le \pi$ and 0 on $\pi \le t \le 2\pi$
(d) $f(t) = |5 - 5t|$ on $0 \le t \le 2$
(e) $f(t) = 0$ on $0 \le t \le \pi/2$ and $\cos t$ on $\pi/2 \le t \le 3\pi/2$

6. Evaluate $\int_0^1 f(x)\,dx$ to three significant figures, and verify the validity of the termwise integration.

(a) $f(x) = \frac{4}{\pi} \sum_{n=1}^\infty \frac{\cos(2n-1)x}{(2n-1)^2}$

(b) $f(x) = \frac{2}{\pi} - \frac{4}{\pi} \sum_{n=1}^\infty \frac{\cos 2nx}{4n^2 - 1}$

17.6 Vector Space Approach

An elegant and more modern approach to Fourier series is available within the vector space context. Vector space is the subject of Section 9.6, which section you may wish to review before continuing.

Specifically, consider the **function space** $C_p[a, b]$ of all real-valued piecewise-continuous functions defined on $[a, b]$, that is, on $a \le x \le b$. First, let us verify that $C_p[a, b]$ is indeed a vector space. Let $\mathbf{f} = f(x)$ and $\mathbf{g} = g(x)$ be any two functions* in $C_p[a, b]$, and let α be any (real) scalar. We define the sum $\mathbf{f} + \mathbf{g}$ and the scalar multiple $\alpha\mathbf{f}$ by

$$\mathbf{f} + \mathbf{g} \equiv f(x) + g(x), \qquad \alpha\mathbf{f} \equiv \alpha f(x), \tag{1}$$

respectively. Observe that if f and g are piecewise continuous on $[a, b]$ then so is $f + g$, so that $C_p[a, b]$ is closed under vector addition. Similarly, if f is piecewise continuous on $[a, b]$ then so is αf, so that $C_p[a, b]$ is closed under scalar multiplication. Furthermore, we define the zero vector $\mathbf{0}$ as the function which is identically zero, so that $\mathbf{f} + \mathbf{0} = f(x) + 0 = f(x) = \mathbf{f}$. And we define the negative inverse of $\mathbf{f} = f(x)$ as $-\mathbf{f} \equiv -f(x)$, in which case we have $\mathbf{f} + (-\mathbf{f}) = f(x) + [-f(x)] = 0 = \mathbf{0}$.

With these definitions of vector addition, scalar multiplication, the zero vector, and the negative inverse, we can see that all of the requirements listed in Definition 9.6.1 are satisfied, so that $C_p[a, b]$ is indeed a vector space.

Next, we wish to introduce an inner product for $C_p[a, b]$ and we choose the

*By "$\mathbf{f} = f(x)$" we mean that the vector $\mathbf{f}$ is the function whose values are $f(x)$. Notation becomes tricky here since there are now three quantities to be distinguished: the function f considered as a mapping, the values $f(x)$ of that function, and the vector $\mathbf{f}$.

inner product of f and g as†

$$\langle \mathbf{f}, \mathbf{g} \rangle \equiv \int_a^b f(x)g(x)\,dx, \tag{2}$$

which is introduced in Example 4 of Section 9.6. Recall that f and g are **orthogonal** if $\langle \mathbf{f}, \mathbf{g} \rangle = 0$, that the **norm** $\|\mathbf{f}\|$ of f is defined as $\|\mathbf{f}\| = \sqrt{\langle \mathbf{f}, \mathbf{f} \rangle}$, and that f is said to be **normalized** (i.e., scaled so as to have unit norm) if $\|\mathbf{f}\| = 1$.

Finally, recall from our study of best approximation, in Section 9.10, the following important result: If f is any vector in a normed inner product vector space $\mathcal{S}$ with natural norm $\|\mathbf{f}\| = \sqrt{\langle \mathbf{f}, \mathbf{f} \rangle}$, and $\{\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_N\}$ is an ON (orthonormal) set in $\mathcal{S}$, then the **best approximation** of f within span $\{\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_N\}$ is given by the **orthogonal projection** of f onto span $\{\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_N\}$, namely, by

$$\mathbf{f} \approx \langle \mathbf{f}, \hat{\mathbf{e}}_1 \rangle \hat{\mathbf{e}}_1 + \cdots + \langle \mathbf{f}, \hat{\mathbf{e}}_N \rangle \hat{\mathbf{e}}_N = \sum_{n=1}^{N} \langle \mathbf{f}, \hat{\mathbf{e}}_n \rangle \hat{\mathbf{e}}_n. \tag{3}$$

To apply these results to Fourier series, let $\mathcal{S}$ be $C_p[a, b]$, with the inner product and norm defined above, let $a = -\ell$ and $b = \ell$, and consider the vectors

$$\mathbf{e}_1 = 1,\ \mathbf{e}_2 = \cos\frac{\pi x}{\ell},\ \mathbf{e}_3 = \sin\frac{\pi x}{\ell},\ \cdots,\ \mathbf{e}_{2k} = \cos\frac{k\pi x}{\ell},\ \mathbf{e}_{2k+1} = \sin\frac{k\pi x}{\ell} \tag{4}$$

in $C_p[-\ell, \ell]$. The set $\{\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_N\}$ is orthogonal by virtue of the inner product definition (2) and the integrals (24a,b,c) in Section 17.3. For instance,

$$\langle \mathbf{e}_2, \mathbf{e}_3 \rangle = \int_{-\ell}^{\ell} \cos\frac{\pi x}{\ell} \sin\frac{\pi x}{\ell}\,dx = 0. \tag{5}$$

Furthermore,

$$\begin{aligned}
\langle \mathbf{e}_1, \mathbf{e}_1 \rangle &= \int_{-\ell}^{\ell} (1)(1)\,dx = 2\ell, \\
\langle \mathbf{e}_2, \mathbf{e}_2 \rangle &= \int_{-\ell}^{\ell} \cos^2\frac{\pi x}{\ell}\,dx = \ell, \\
&\vdots \\
\langle \mathbf{e}_{2k+1}, \mathbf{e}_{2k+1} \rangle &= \int_{-\ell}^{\ell} \sin^2\frac{k\pi x}{\ell}\,dx = \ell,
\end{aligned} \tag{6}$$

† If we were considering a complex function space (i.e., where the functions are complex-valued and the scalars are complex), then we would use the inner product

$$\langle \mathbf{f}, \mathbf{g} \rangle \equiv \int_a^b f(x)\overline{g}(x)\,dx$$

instead of (2), in accordance with the discussion in Section 12.2, where $\overline{g}(x)$ is the complex conjugate of $g(x)$. For example, if $g(x) = 6 + 5i$, then $\overline{g}(x) = 6 - 5i$.

so that the normalized $\mathbf{e}_n$'s are

$$\begin{aligned}
\hat{\mathbf{e}}_1 &= \frac{1}{\|\mathbf{e}_1\|}\,\mathbf{e}_1 = \frac{1}{\sqrt{\langle \mathbf{e}_1, \mathbf{e}_1\rangle}}\,\mathbf{e}_1 = \frac{1}{\sqrt{2\ell}},\\
\hat{\mathbf{e}}_2 &= \frac{1}{\|\mathbf{e}_2\|}\,\mathbf{e}_2 = \frac{1}{\sqrt{\langle \mathbf{e}_2, \mathbf{e}_2\rangle}}\,\mathbf{e}_2 = \frac{1}{\sqrt{\ell}}\cos\frac{\pi x}{\ell},\\
&\vdots\\
\hat{\mathbf{e}}_{2k+1} &= \frac{1}{\sqrt{\ell}}\sin\frac{k\pi x}{\ell}.
\end{aligned} \tag{7}$$

Since the set $\{\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_{2k+1}\}$ is ON, it is LI (linearly independent), and since k (and hence $2k+1$) is arbitrarily large, it follows that we have an arbitrarily large number of LI vectors in $C_p[-\ell, \ell]$, so the latter is *infinite dimensional*.

Returning to (3), we are approximating a given $\mathbf{f} = f(x)$, in $C_p[-\ell, \ell]$, in the form

$$\begin{aligned}
f(x) \approx c_1\frac{1}{\sqrt{2\ell}} + c_2\frac{1}{\sqrt{\ell}}\cos\frac{\pi x}{\ell} + c_3\frac{1}{\sqrt{\ell}}\sin\frac{\pi x}{\ell} + \cdots\\
+ c_{2k}\frac{1}{\sqrt{\ell}}\cos\frac{k\pi x}{\ell} + c_{2k+1}\frac{1}{\sqrt{\ell}}\sin\frac{k\pi x}{\ell},
\end{aligned} \tag{8}$$

where

$$\begin{aligned}
c_1 &= \langle \mathbf{f}, \hat{\mathbf{e}}_1\rangle = \int_{-\ell}^{\ell} f(x)\,\frac{1}{\sqrt{2\ell}}\,dx = \frac{1}{\sqrt{2\ell}}\int_{-\ell}^{\ell} f(x)\,dx\\
c_2 &= \langle \mathbf{f}, \hat{\mathbf{e}}_2\rangle = \int_{-\ell}^{\ell} f(x)\,\frac{1}{\sqrt{\ell}}\cos\frac{\pi x}{\ell}\,dx = \frac{1}{\sqrt{\ell}}\int_{-\ell}^{\ell} f(x)\cos\frac{\pi x}{\ell}\,dx\\
&\vdots\\
c_{2k+1} &= \langle \mathbf{f}, \hat{\mathbf{e}}_{2k+1}\rangle = \frac{1}{\sqrt{\ell}}\int_{-\ell}^{\ell} f(x)\sin\frac{k\pi x}{\ell}\,dx.
\end{aligned} \tag{9}$$

Equivalently, we can write

$$\begin{aligned}
f(x) &\approx a_0 + a_1\cos\frac{\pi x}{\ell} + b_1\sin\frac{\pi x}{\ell} + \cdots + a_k\cos\frac{k\pi x}{\ell} + b_k\sin\frac{k\pi x}{\ell}\\
&= a_0 + \sum_{n=1}^{k}\left(a_n\cos\frac{n\pi x}{\ell} + b_n\sin\frac{n\pi x}{\ell}\right),
\end{aligned} \tag{10}$$

where

$$a_0 = \frac{1}{2\ell}\int_{-\ell}^{\ell} f(x)\,dx, \tag{11a}$$

$$a_n = \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\cos\frac{n\pi x}{\ell}\,dx, \tag{11b}$$

$$b_n = \frac{1}{\ell}\int_{-\ell}^{\ell} f(x)\sin\frac{n\pi x}{\ell}\,dx. \tag{11c}$$

Let us review the situation. The right-hand side of (10) is the orthogonal projection of $\mathbf{f}$ on the $(2k+1)$-dimensional subspace of $C_p[-\ell, \ell]$, which is the span of the ON vectors

$$\frac{1}{\sqrt{2\ell}}, \quad \frac{1}{\sqrt{\ell}} \cos \frac{\pi x}{\ell}, \quad \frac{1}{\sqrt{\ell}} \sin \frac{\pi x}{\ell}, \ldots, \frac{1}{\sqrt{\ell}} \cos \frac{k\pi x}{\ell}, \quad \frac{1}{\sqrt{\ell}} \sin \frac{k\pi x}{\ell}. \tag{12}$$

With the coefficients in (10) equal to the by-now-familiar Fourier coefficients, as given in (11), (10) is the best possible approximation to $f(x)$ of the form (8). That is, it is the best approximation in the sense of minimizing the norm of the error vector

$$\begin{aligned} &\left\| \mathbf{f} - \left(c_1 \frac{1}{\sqrt{2\ell}} + c_2 \frac{1}{\sqrt{\ell}} \cos \frac{\pi x}{\ell} + \cdots + c_{2k+1} \frac{1}{\sqrt{\ell}} \sin \frac{k\pi x}{\ell} \right) \right\|^2 \\ &= \int_{-\ell}^{\ell} \left[f(x) - \left(c_1 \frac{1}{\sqrt{2\ell}} + c_2 \frac{1}{\sqrt{\ell}} \cos \frac{\pi x}{\ell} + \cdots + c_{2k+1} \frac{1}{\sqrt{\ell}} \sin \frac{k\pi x}{\ell} \right) \right]^2 dx. \end{aligned} \tag{13}$$

The integral in (13) is known as the **squared error**. Thus (10) is the best approximation in the sense of minimizing the squared error and is called the **least-square approximation** of $f(x)$ with respect to the ON set (4).

We can well expect the squared error to diminish as we increase k, but the key question is: Does it tend to zero as $k \to \infty$? It does.

THEOREM 17.6.1 *Vector Convergence*
If $f(x)$ is piecewise continuous on $[-\ell, \ell]$ and $a_0, a_1, b_1, a_2, b_2, \ldots$ are the Fourier coefficients defined by (11), then

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi x}{\ell} + b_n \sin \frac{n\pi x}{\ell} \right) \tag{14}$$

holds in the sense of vector (least-square) convergence, namely,

$$\lim_{k \to \infty} \int_{-\ell}^{\ell} \left[f(x) - a_0 - \sum_{n=1}^{k} \left(a_n \cos \frac{n\pi x}{\ell} + b_n \sin \frac{n\pi x}{\ell} \right) \right]^2 dx = 0. \tag{15}$$

Proof of this sophisticated theorem is well beyond our present scope.

Note that the meaning of (14) here is different from its meaning in preceding sections. Specifically, in preceding sections equation (14) held in the pointwise sense, namely, that at a specific single fixed value of x we have

$$\lim_{k \to \infty} \left[a_0 + \sum_{n=1}^{k} \left(a_n \cos \frac{n\pi x}{\ell} + b_n \sin \frac{n\pi x}{\ell} \right) \right] = f(x) \tag{16}$$

or, equivalently,

$$\boxed{\lim_{k\to\infty}\left[f(x)-a_0-\sum_{n=1}^{k}\left(a_n\cos\frac{n\pi x}{\ell}+b_n\sin\frac{n\pi x}{\ell}\right)\right]=0.} \tag{17}$$

In contrast, in the present section (14) is understood as a vector equation, which holds in the sense that

$$\boxed{\lim_{k\to\infty}\int_{-\ell}^{\ell}\left[f(x)-a_0-\sum_{n=1}^{k}\left(a_n\cos\frac{n\pi x}{\ell}+b_n\sin\frac{n\pi x}{\ell}\right)\right]^2dx=0.} \tag{18}$$

The truth of (17) does not imply the truth of (18), nor does the truth of (18) imply the truth of (17); they are logically independent statements (Exercise 4).

To better appreciate the distinction, observe that the pointwise convergence addressed in Theorem 17.3.1 is convergence in a *local* sense, at a specific point x, whereas the vector convergence addressed in Theorem 17.6.1 is convergence in a *global* sense. That is, the squared error *integrated over the entire interval* tends to zero as $k\to\infty$.

We do not wish to imply that one form of convergence is inherently better or more correct than the other; they are simply different.

Closure. In effect, Theorem 17.6.1 tells us that the infinite set of orthogonal vectors $\{1,\cos\frac{\pi x}{\ell},\sin\frac{\pi x}{\ell},\cos\frac{2\pi x}{\ell},\sin\frac{2\pi x}{\ell},\ldots\}$ comprises an orthogonal *basis* for $C_p[-\ell,\ell]$. Couldn't we say that that set is a basis because $C_p[-\ell,\ell]$ is infinite-dimensional and there is an infinite number of vectors in the set? No. For suppose we remove one (or more) of the vectors from the set. Then we still have an infinite set of orthogonal vectors, but they are not a basis. For instance, if we remove the $\cos(\pi x/\ell)$ vector and take $f(x)$ to be $6\cos(\pi x/\ell)$, say, then (14) would be

$$6\cos\frac{\pi x}{\ell}=0+0+0+\cdots,$$

which is surely incorrect.

EXERCISES 17.6

1. Corresponding to the approximation (10), let the error vector be

$$\mathbf{E}=f(x)-a_0-\sum_{n=1}^{k}\left(a_n\cos\frac{n\pi x}{\ell}+b_n\sin\frac{n\pi x}{\ell}\right). \tag{1.1}$$

(a) Show that

$$\|\mathbf{E}\|^2=\int_{-\ell}^{\ell}[f(x)]^2\,dx-\ell\left[2a_0^2+\sum_{n=1}^{k}(a_n^2+b_n^2)\right]. \tag{1.2}$$

(b) Deduce, from (1.2), the **Bessel inequality**

$$\boxed{2a_0^2 + \sum_{n=1}^{k}(a_n^2 + b_n^2) \le \frac{1}{\ell}\int_{-\ell}^{\ell}[f(x)]^2\,dx,} \tag{1.3}$$

which holds for each $k = 1, 2, \ldots$. NOTE: Since $\|\mathbf{E}\| \to 0$ as $k \to \infty$, according to Theorem 17.6.1, we obtain

$$\boxed{2a_0^2 + \sum_{n=1}^{\infty}(a_n^2 + b_n^2) = \frac{1}{\ell}\int_{-\ell}^{\ell}[f(x)]^2\,dx,} \tag{1.4}$$

which is known as the **Parseval equality**, and which is an infinite-dimensional function space version of the familiar Pythagorean theorem. We can draw an interesting and useful conclusion from (1.4). The integral on the right converges because f has been assumed piecewise continuous on $[-\ell, \ell]$. Thus, the series on the left converges and has nonnegative terms. For such a series we know that its nth term must tend to zero as $n \to \infty$. Hence, it follows that $a_n \to 0$ and $b_n \to 0$ as $n \to \infty$.

2. Use equation (1.2), above, to compute $\|\mathbf{E}\|$, for $k = 1, 2, \ldots, 8$, where the 2π-periodic function f is defined on $-\pi < x < \pi$ as follows.

(a) $f(x) = |x|$ (b) $f(x) = x$
(c) $f(x) = \cos^2 x$ (d) $f(x) = x^2$
(e) $f(x) = |\sin x|$ (f) $f(x) = |\cos x|$

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

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

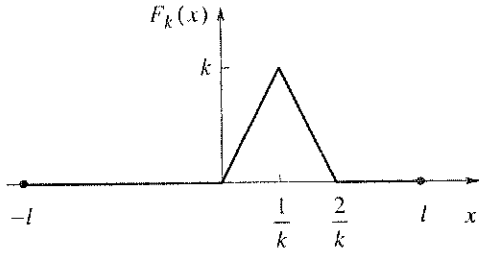
(i) $f(x) = \begin{cases} 100, & -\pi < x < \pi/2 \\ 50, & \pi/2 < x < \pi \end{cases}$

(j) $f(x) = \begin{cases} \sin x, & -\pi < x < 0 \\ 0, & 0 < x < \pi \end{cases}$

3. Let f and g both be members of $C_p[-\ell, \ell]$, and $f(x) = g(x)$ on $[-\ell, \ell]$ except at a finite number of points, at which $f(x) \neq g(x)$. Show that $\mathbf{f} = \mathbf{g}$ in spite of these pointwise differences; i.e., show that $\|\mathbf{f} - \mathbf{g}\| = 0$, so that within the vector space framework $\mathbf{f}$ and $\mathbf{g}$ are indistinguishable.

4. (*Vector convergence and pointwise convergence*) Below Theorem 17.6.1, we emphasized that vector and pointwise convergence are independent, neither one implies the other. The purpose of this exercise is to illustrate that claim through a simple example. Consider, in place of the messy $[\ ]^2$ integrand in (18), the sequence $F_k(x)$ defined in parts (a), (b), and (c).

(a) First, consider the sequence displayed below.



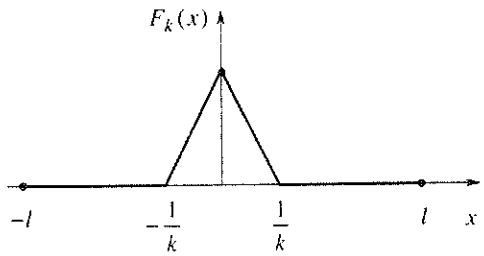
As k increases, the "mountain" becomes taller and narrower. Show that

$$\lim_{k\to\infty} F_k(x) = 0 \tag{4.1a}$$

on $[-\ell, \ell]$, whereas

$$\lim_{k\to\infty}\int_{-\ell}^{\ell} F_k(x)\,dx = 1 \neq 0. \tag{4.1b}$$

(b) With $F_k(x)$ as shown below, instead, show that



$$\lim_{k\to\infty}\int_{-\ell}^{\ell} F_k(x)\,dx = 0, \tag{4.1c}$$

whereas

$$\lim_{k\to\infty} F_k(x) = \begin{cases} 0, & x \neq 0 \\ 1, & x = 0 \end{cases} \quad \neq 0. \tag{4.1d}$$

(c) Let $F_k(x)$ be the same as in part (b), except that the height of the "mountain" is $1/k$ instead of 1. Show that in this case *both* limits are zero. CONCLUSION: The statements $\lim_{k\to\infty}\int_{-\ell}^{\ell} F_k(x)\,dx = 0$ and $\lim_{k\to\infty} F_k(x) = 0$ on $[-\ell, \ell]$ are *independent*; the truth of one does not imply the truth of the other.

5. Beginning with the expression

$$\|\mathbf{E}\|^2 = \int_{-\ell}^{\ell}\left[f(x) - a_0 - \sum_{n=1}^{k}\left(a_n \cos\frac{n\pi x}{\ell} + b_n \sin\frac{n\pi x}{\ell}\right)\right]^2 dx$$

for the square of the norm of the error vector $\mathbf{E}$ associated with the approximation (10), seek the optional choice of a_0, a_n, b_n by setting $\partial \|\mathbf{E}\|^2 / \partial a_0 = 0$, $\partial \|\mathbf{E}\|^2 / \partial a_n = 0$, and $\partial \|\mathbf{E}\|^2 / \partial b_n = 0$. Show that that step produces the expressions for a_0, a_n, b_n given in (11).

17.7 The Sturm–Liouville Theory

17.7.1. Sturm–Liouville problem. In Section 17.6 we found that the sines and cosines present in Fourier series constitute an orthogonal basis for the relevant infinite-dimensional function space. Where do such bases come from? Are there others as well? In this section we discover that such bases arise as the eigenfunctions of Sturm–Liouville eigenvalue problems, just as real symmetric $n \times n$ matrices provide us with sets of eigenvectors that are orthogonal bases for n-space. Thus, Fourier series will be found to be just one (extremely important) example within a broader Sturm–Liouville theory.*

By a **Sturm–Liouville problem** we mean a linear homogeneous second-order differential equation

$$\boxed{[p(x)y']' + q(x)y + \lambda w(x)y = 0, \qquad (a < x < b)} \tag{1a}$$

with homogeneous boundary conditions of the form

$$\boxed{\begin{aligned} \alpha y(a) + \beta y'(a) &= 0, \\ \gamma y(b) + \delta y'(b) &= 0, \end{aligned}} \tag{1b}$$

where a, b are finite, where p, p', q, w are continuous on $[a, b]$, and where $p(x) > 0$ and $w(x) > 0$ on $[a, b]$. These conditions, as well as the precise form of (1a) and (1b), are important, and should be carefully noted. Further, α and β are not both zero, γ and δ are not both zero, and $a, b, p(x), q(x), w(x), \alpha, \beta, \gamma, \delta$ are all real.

Also critical is that (1) is a *boundary-value* problem, not an initial-value problem, because the conditions (1b) are imposed at both ends of the interval. If, in place of (1b), we imposed homogeneous initial conditions $y(a) = 0$ and $y(b) = 0$, then (1) would admit the trivial solution $y(x) = 0$, and that solution would be unique (Theorem 3.3.1). However, we saw in Section 3.3.2 that boundary-value problems can admit no solution, a unique solution, or a nonunique solution. Thus, even though the Sturm–Liouville boundary-value problem (1) surely

*The theory presented in this section was discovered by *Charles Sturm* (1803–1855) and *Joseph Liouville* (1809–1882), and published in 1836–1837. Sturm, a professor of mechanics at the Sorbonne, became involved in this work as an outgrowth of his studies of the partial differential equations governing the flow of heat in a bar of nonuniform density. He was joined in this work by his friend Liouville, who was a professor of mathematics at the Collège de France and who is also well known for his work on complex variable theory and on transcendental numbers. Note the spelling Liouville (not Louisville), and the pronunciation lē-ōō′vil (not lōō′ē-vil).

admits the trivial solution, our interest is in finding *nontrivial* solutions. Though $a, b, p(x), q(x), w(x), \alpha, \beta, \gamma, \delta$ are all specified, λ is a free parameter. Any value of λ that permits the existence of nontrivial solutions of (1) is called an **eigenvalue** of (1), and the corresponding nontrivial solution is called an **eigenfunction** of (1). Thus, (1) is an **eigenvalue problem**, analogous to the matrix eigenvalue problem studied in Chapter 11.

EXAMPLE 1. Consider the case

$$y'' + \lambda y = 0, \qquad (0 < x < L) \tag{2a}$$

$$y(0) = 0, \quad y(L) = 0. \tag{2b}$$

Comparing (2) with (1) we see that $p(x) = w(x) = 1$, $q(x) = 0$, $a = 0$, $b = L$, $\alpha = \gamma = 1$, and $\beta = \delta = 0$, all of which satisfy the conditions listed below (1).

To see if (2) admits nontrivial solutions, let us not be intimidated by the fact it is an "eigenvalue problem." After all, (2a) is a simple differential equation. Its general solution is evidently*

$$y(x) = A\cos\sqrt{\lambda}\,x + B\sin\sqrt{\lambda}\,x. \tag{3}$$

However, if $\lambda = 0$, then (3) reduces to $y(x) = A$, which is not a general solution of (2a). Thus, for the special case $\lambda = 0$ we return to (2a), which becomes $y'' = 0$, and determine the general solution to be $C + Dx$. Thus, in place of (3) we write

$$y(x) = \begin{cases} A\cos\sqrt{\lambda}\,x + B\sin\sqrt{\lambda}\,x, & \lambda \neq 0 \\ C + Dx. & \lambda = 0 \end{cases} \tag{4a.b}$$

Treating the cases $\lambda = 0$ and $\lambda \neq 0$ separately, consider $\lambda = 0$ first. Applying the boundary conditions (2b) to $y(x) = C + Dx$ gives

$$y(0) = 0 = C, \tag{5a}$$

$$y(L) = 0 = C + DL, \tag{5b}$$

so $C = D = 0$. Thus, the only solution corresponding to $\lambda = 0$ is the trivial solution $y(x) = 0$, so $\lambda = 0$ is not an eigenvalue (in this example).

---

*Note that we don't yet know λ. If $\lambda > 0$, then (3) is the general solution of (2), but does (3) hold if $\lambda < 0$? Shouldn't we have a cosh and sinh in that case? If $\lambda < 0$, (3) becomes,

$$\begin{aligned} y(x) &= A\cos(i\sqrt{|\lambda|}\,x) + B\sin(i\sqrt{|\lambda|}\,x) \\ &= A\cosh(\sqrt{|\lambda|}\,x) + iB\sinh\sqrt{|\lambda|}\,x, \end{aligned}$$

so we do have an arbitrary linear combination of cosh and sinh, as we should. (We could rename iB as C if we don't like the i.) Thus, there is no need to treat the cases $\lambda > 0$ and $\lambda < 0$ separately, although some authors prefer to do that. In fact, (3) holds even if λ is complex, but we will find that for Sturm–Liouville problems λ is always real.

Turning to the case $\lambda \neq 0$, the boundary conditions give

$$y(0) = 0 = A, \tag{6a}$$

$$y(L) = 0 = A\cos\sqrt{\lambda}\,L + B\sin\sqrt{\lambda}\,L. \tag{6b}$$

Since (6a) gives $A = 0$, (6b) gives $B\sin\sqrt{\lambda}\,L = 0$, so either $B = 0$ or $\sin\sqrt{\lambda}\,L = 0$ and B is arbitrary. We rule out the choice $B = 0$ since then we would have $A = B = 0$ and hence the trivial solution $y(x) = 0$. Rather, B is arbitrary and

$$\sin\sqrt{\lambda}\,L = 0. \tag{7}$$

Solving (7), we have $\sqrt{\lambda}L = n\pi$ for $n = 0, \pm1, \pm2, \ldots$. Of these values, discard $n = 0$ because it gives $\lambda = 0$, which case has already been considered. Thus we have the eigenfunctions

$$y(x) = B\sin\frac{n\pi x}{L}, \tag{8}$$

where B is arbitrary and $n = \pm1, \pm2, \ldots$. The negative values of n can be discarded as well since the positive and negative choices do not lead to distinct (i.e., linearly independent) eigenfunctions. For example, $n = +2$ gives $\sin(2\pi x/L)$, and $n = -2$ gives $\sin(-2\pi x/L) = -\sin(2\pi x/L)$, and since the scale factor B is arbitrary it can absorb the minus sign. Let us set $B = 1$, say, for definiteness.

The upshot is that we have the infinite set of eigenvalues and eigenfunctions

$$\lambda_n = \frac{n^2\pi^2}{L^2} \qquad \text{and} \qquad \phi_n(x) = \sin\frac{n\pi x}{L} \tag{9}$$

for $n = 1, 2, \ldots$, where we use the symbol ϕ_n to denote the nth eigenfunction (analogous to the special symbol $\mathbf{e}_n$ that we used in the matrix case).

COMMENT 1. It may be useful to recast the solution of (6) in matrix form because the matrix approach is more convenient in algebraically-more-difficult cases. Re-expressing (6) as

$$\begin{bmatrix} 1 & 0 \\ \cos\sqrt{\lambda}\,L & \sin\sqrt{\lambda}\,L \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{10}$$

we see that we have a nontrivial solution of (10) [and hence of (2)] if and only if the determinant of the coefficient matrix is zero:

$$\begin{vmatrix} 1 & 0 \\ \cos\sqrt{\lambda}\,L & \sin\sqrt{\lambda}\,L \end{vmatrix} = \sin\sqrt{\lambda}\,L = 0, \tag{11}$$

which is the same as (7). Next, solve (11) for the λ's, put them back in (10), and solve (10), say by Gauss elimination, for the nontrivial solutions for A and B. That step gives $A = 0$ and $B =$ arbitrary, as above.

COMMENT 2. We call (11) the **characteristic equation** corresponding to the eigenvalue problem (2). In the $n \times n$ matrix case the characteristic equation is always an nth degree algebraic equation; in the Sturm–Liouville case it is always transcendental (an infinite-degree algebraic equation), with an infinite number of distinct roots.* ∎

Sturm–Liouville problems such as (2) arise throughout Chapters 18–20, when

*To see that (11) is an infinite-degree algebraic equation, replace $\sin\sqrt{\lambda}\,L$ by its Taylor series.

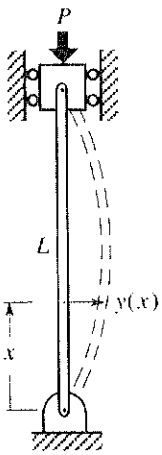


Figure 1. Column buckling.

we solve partial differential equations by the method of separation of variables – for instance in connection with unsteady heat conduction in a rod and the lateral motion of a vibrating string. Sometimes, however, they arise directly. For instance, a classical problem in structural mechanics is the determination of the buckling load of a structure. Consider, as a simple structure, the column shown in Fig. 1, of length L, pinned at both ends and subjected to a downward load (i.e., force) P. As we increase P nothing happens – the column remains straight, until a certain value of P is reached, which we call the critical load or *buckling load* and which we denote as P_{cr}. Under that load the column bends (and probably collapses). Clearly, it is important to be able to predict P_{cr}. To do so, we use *Euler beam theory*,* which tells us that the lateral deflection $y(x)$ is governed by the boundary-value problem

$$EIy'' + Py = 0, \qquad (0 < x < L) \tag{12a}$$

$$y(0) = 0, \quad y(L) = 0, \tag{12b}$$

where E and I are physical constants of the column: E is Young's modulus of the material and I is the cross-sectional inertia. We see that (12) is the same as (2), with $\lambda = P/EI$. Surely, $y(x) = 0$ satisfies (12), but that solution is of no interest because it does not correspond to buckling. From Example 1 we recall that nontrivial solutions occur for $\lambda = P/EI = \pi^2/L^2, 4\pi^2/L^2, 9\pi^2/L^2, \ldots$. The smallest of these, $P/EI = \pi^2/L^2$, gives the buckling load

$$P_{cr} = \frac{\pi^2 EI}{L^2}, \tag{13}$$

and the corresponding eigenfunction $\sin(\pi x/L)$ gives (to within an arbitrary scale factor) the shape of the corresponding buckling mode, which is somewhat as sketched in Fig. 1. The analysis gives the inception of buckling and does not give insight into the dynamical process of collapse. The formula (13) was published first by Euler in 1757.

Since we will be concerned with orthogonal bases in function space, we will need an inner product $\langle f, g\rangle$ between two functions (i.e., vectors) f and g. It will be convenient to use the inner product†

$$\boxed{\langle f, g\rangle \equiv \int_a^b f(x)\, g(x)\, w(x)\, dx,} \tag{14}$$

where the weight function $w(x)$ is the $w(x)$ in the Sturm–Liouville equation (1a). If $\langle f, g\rangle = 0$, then f and g are *orthogonal*.

*See S. Timoshenko, *Strength of Materials*, Part I (Princeton, NJ: D. Van Nostrand, 1955).

†If f and g are complex-valued functions such as e^{ix}, then (14) fails to meet the conditions required of any dot or inner product [(16) in Section 9.6] and should be modified as $\int_a^b f(x)\,\overline{g}(x)\,w(x)\,dx$. We will face up to that detail in the optional Section 17.7.2 but, otherwise, will continue to use (14) because it will turn out that complex-valued functions need not arise. Also, note that whether we write $\langle f, g\rangle$ or $\langle \mathbf{f}, \mathbf{g}\rangle$ doesn't matter, as long as we understand that f and g are here being considered as vectors.

We have the following major theorem regarding the eigenvalues and eigenfunctions of the Sturm–Liouville eigenvalue problem (1), with the restrictions on $p(x), q(x), w(x), a, b, \alpha, \beta, \gamma, \delta$ stated earlier.

THEOREM 17.7.1 *Sturm–Liouville Theorem*
Let λ_n and $\phi_n(x)$ denote any eigenvalue and corresponding eigenfunction of the Sturm–Liouville eigenvalue problem (1), respectively.

(a) The eigenvalues are real.

(b) The eigenvalues are simple. That is, to each eigenvalue there corresponds only one linearly independent eigenfunction. Further, there are an infinite number of eigenvalues, and they can be ordered so that $\lambda_1 < \lambda_2 < \lambda_3 < \cdots$, where $\lambda_n \to \infty$ as $n \to \infty$.

(c) Eigenfunctions corresponding to distinct eigenvalues are orthogonal. That is, if $\lambda_j \neq \lambda_k$, then $\langle \phi_j, \phi_k \rangle = 0$.

(d) Let f and f' be piecewise continuous on $a \leq x \leq b$. If $a_n = \langle f, \phi_n \rangle / \langle \phi_n, \phi_n \rangle$, then the series $\sum_{n=1}^{\infty} a_n \phi_n(x)$ converges to $f(x)$ if f is continuous at x, and to the mean value $[f(x+) + f(x-)]/2$ if f is discontinuous at x, for each point x in the open interval $a < x < b$.

This theorem is analogous to the several individual theorems given in Section 11.3 for $n \times n$ real symmetric matrix eigenvalue problems. Parts of it are proved in the optional Section 17.7.2.

Part (d) says that

$$\boxed{f(x) = \sum_{n=1}^{\infty} \frac{\langle f, \phi_n \rangle}{\langle \phi_n, \phi_n \rangle} \phi_n(x)} \tag{15}$$

holds at each point x in $a < x < b$ at which f is continuous. If, at a discontinuity, $f(x)$ does not happen to equal the mean value$[f(x+)+f(x-)]/2$, then the equality in (15) does not hold at that point. To remind us of the possibility of such pointwise discrepancies, some authors write an "ae" above the equal sign, to mean equal "almost everywhere," but we will simply write the eigenfunction expansion of f as we have in (15), without such reminders. Observe that (d) is a pointwise convergence statement. Various other statements are available regarding both pointwise and vector convergence, but (d) will suffice for most practical purposes and for our purposes in this text.

EXAMPLE 2. Consider the results of Example 1 in the light of Theorem 17.7. Since $\lambda_n = n^2\pi^2/L^2$, the eigenvalues are real, there are an infinite number of them, and $\lambda_n \to \infty$ as $n \to \infty$. Further, each eigenvalue is simple because the single eigenfunction

$\phi_n(x) = \sin(n\pi x/L)$ corresponds to each eigenvalue λ_n. Finally, with the weight function $w(x) = 1$ we have

$$\langle \phi_m, \phi_n \rangle = \int_0^L \sin\frac{m\pi x}{L} \sin\frac{n\pi x}{L}\, dx = 0 \tag{16}$$

for $m \neq n$, by virtue of the Euler formula (24b) in Section 17.3. Each of these results is in accord with Theorem 17.7.1.

To illustrate the eigenfunction expansion (15), let $f(x) = x$. Then

$$\langle f, \phi_n \rangle = \int_0^L x \sin\frac{n\pi x}{L}\, dx = (-1)^{n+1}\frac{L^2}{n\pi}, \tag{17a}$$

$$\langle \phi_n, \phi_n \rangle = \int_0^L \sin^2\frac{n\pi x}{L}\, dx = \frac{L}{2}, \tag{17b}$$

so we have

$$f(x) = x = \frac{2L}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}\sin\frac{n\pi x}{L}. \tag{18}$$

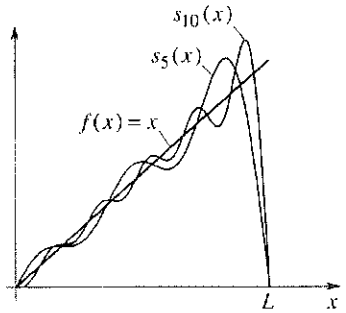


Figure 2. Convergence of (18).

COMMENT 1. Carefully observe that part (d) of the theorem does NOT require f to satisfy the homogeneous Sturm–Liouville boundary conditions, which are $y(0) = 0$ and $y(L) = 0$ in the present example. In fact, $f(L)$ is L, not 0. Nonetheless, we do obtain the convergence that is guaranteed by the theorem, over $0 < x < L$ (actually, over $0 \leq x < L$ in this example), as hinted at in Fig. 2, where we compare $f(x) = x$ with the fifth and tenth partial sums, $s_5(x)$ and $s_{10}(x)$.

COMMENT 2. Observe also that, corresponding to the present Sturm–Liouville problem, the eigenfunction expansion (15), namely,

$$f(x) = \sum_{n=1}^{\infty} a_n \sin\frac{n\pi x}{L}, \tag{19a}$$

where

$$a_n = \frac{2}{L}\int_0^L f(x)\sin\frac{n\pi x}{L}\, dx, \tag{19b}$$

is actually the half-range sine expansion of f, studied in Section 17.4. Other choices of the boundary conditions, in place of $y(0) = 0$ and $y(L) = 0$, will result in eigenfunctions that produce the half-range cosine and quarter-range sine and cosine expansions (Exercise 2). ∎

There is a small flaw in our procedure. To solve the characteristic equation (7), we recalled that $\sin x = 0$ has roots at $x = n\pi$ on the real axis. Might there be complex roots as well? It's true that we know in advance that λ is real, but if it is real and negative then the argument of the sine is purely imaginary. Thus, we need to search for roots of $\sin z = 0$ (where $z = x + iy$) not only along the real axis but along the imaginary axis as well. Doing so, observe that $\sin iy = i\sinh y = 0$ is equivalent to $\sinh y = 0$, which admits only the root $y = 0$. Thus, we did not miss

any roots, and all is well. The following theorem could have saved us this extra trouble.

THEOREM 17.7.2 *Nonnegative Eigenvalues*
If $q(x) \leq 0$ on $[a, b]$ and $[p(x)\phi_n(x)\phi_n'(x)]|_a^b \leq 0$ for the eigenfunction $\phi_n(x)$, then not only is λ_n real, $\lambda_n \geq 0$.

Applying Theorem 17.7.2 to the problem in Example 1, observe that $q(x) \leq 0$ on $[0, L]$ because $q(x) = 0$. Further, $p(x) = 1$ and the $\phi_n(x)$'s satisfy the boundary conditions (2b), so

$$\begin{aligned}[p(x)\phi_n(x)\phi_n'(x)]\Big|_a^b &= \phi_n(L)\phi_n'(L) - \phi_n(0)\phi_n'(0) \\ &= (0)\phi_n'(L) - (0)\phi_n'(0) = 0. \end{aligned} \tag{20}$$

Hence, not only are the λ_n's real, they are also nonnegative.

EXAMPLE 3. Find the eigenvalues and eigenfunctions for the Sturm–Liouville problem

$$y'' + \lambda y = 0, \qquad (0 < x < 1) \tag{21a}$$

$$y(0) - 2y'(0) = 0, \quad y(1) = 0. \tag{21b}$$

We speak of the boundary condition at $x = 0$ as being of **mixed** type because both y and y' are present: that is, both α and β are nonzero in (1b). Such boundary conditions do arise in applications such as unsteady heat conduction, as we shall see in Chapter 18.

As in Example 1, solution of (21a) gives

$$y(x) = \begin{cases} A\cos\sqrt{\lambda}\,x + B\sin\sqrt{\lambda}\,x, & \lambda \neq 0 \\ C + Dx, & \lambda = 0. \end{cases} \tag{22a,b}$$

Imposing the boundary conditions (21b) on the $C + Dx$ solution gives $C - 2D = 0$ and $C + D = 0$, so $C = D = 0$. Hence, $\lambda = 0$ is not an eigenvalue. For the $\lambda \neq 0$ case, the boundary conditions give

$$\begin{bmatrix} 1 & -2\sqrt{\lambda} \\ \cos\sqrt{\lambda} & \sin\sqrt{\lambda} \end{bmatrix} \begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{23}$$

To obtain nontrivial solutions (i.e., where A and B are not both zero), set the determinant equal to zero, which step gives the characteristic equation $\sin\sqrt{\lambda} + 2\sqrt{\lambda}\cos\sqrt{\lambda} = 0$ or, more conveniently,

$$\tan\sqrt{\lambda} = -2\sqrt{\lambda}. \tag{24}$$

Applying Theorem 17.7.2, with $q(x) = 0$ and $p(x) = 1$,

$$\begin{aligned}(p\phi_n\phi_n')\Big|_0^1 &= \phi_n(1)\phi_n'(1) - \phi_n(0)\phi_n'(0) \\ &= (0) - 2[\phi_n'(0)]^2 \qquad \text{[from (21b)]} \\ &\leq 0, \end{aligned} \tag{25}$$

so the λ_n's are nonnegative. Hence, it suffices to look for solutions of (24) along a real $\sqrt{\lambda}$ axis. Plotting the left- and right-hand sides of (24) in Fig. 3, we can see the roots $\sqrt{\lambda_n}$ fall as

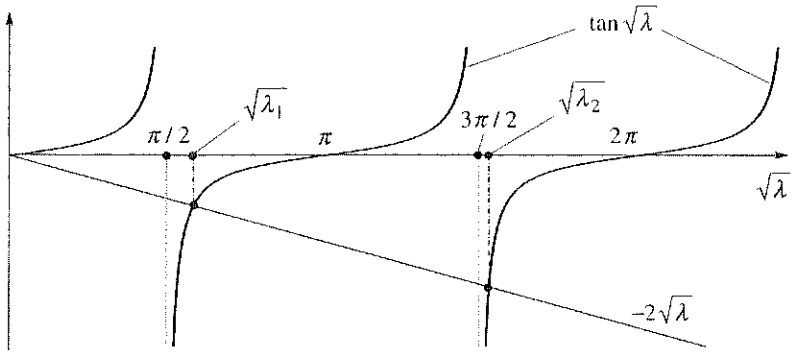


Figure 3. Roots of (24).

indicated. (The intersection at the origin can be discarded because we already determined that $\lambda = 0$ is not an eigenvalue.) We can find any number of roots using computer software such as *Maple* and obtain, for the first few,

$$\sqrt{\lambda_1} = 1.83660, \quad \sqrt{\lambda_2} = 4.81584, \quad \sqrt{\lambda_3} = 7.91705, \ldots, \tag{26}$$

or

$$\lambda_1 = 3.3731, \quad \lambda_2 = 23.1923, \quad \lambda_3 = 62.6797, \ldots, \tag{27}$$

and so on. The graphs in Fig. 3, or even a freehand sketch of them, reveal that $\sqrt{\lambda_n} \sim (2n-1)\pi/2$ as $n \to \infty$.

With the eigenvalues determined, we return to (23) to find the nontrivial solutions for A and B and hence for $y(x)$. With λ satisfying (24), the second row of the coefficient matrix is a multiple of the first, so of the two scalar equations implied by (23) the second can be discarded, leaving the one equation

$$A - 2\sqrt{\lambda_n}\, B = 0 \tag{28}$$

on the two unknowns A and B. Thus, $A = 2\sqrt{\lambda_n}\, B$, where B remains arbitrary, so

$$\begin{aligned} y(x) &= A\cos\sqrt{\lambda_n}\, x + B\sin\sqrt{\lambda_n}\, x \\ &= B(2\sqrt{\lambda_n}\cos\sqrt{\lambda_n}\, x + \sin\sqrt{\lambda_n}\, x), \end{aligned} \tag{29}$$

and hence the eigenfunctions are

$$\phi_n(x) = 2\sqrt{\lambda_n}\cos\sqrt{\lambda_n}\, x + \sin\sqrt{\lambda_n}\, x, \tag{30}$$

where the λ_n's are given by (27).

COMMENT. Actually, if we extended the graphs in Fig. 3 over $-\infty < \sqrt{\lambda} < 0$ we would find the additional roots $-\sqrt{\lambda_1}$ and $-\sqrt{\lambda_2}$ and so on. These can be discarded just as we discarded the $n = -1, -2, \ldots$ cases in Example 1 because they contribute nothing new.

For instance, the right-hand side of (30) is an odd function of $\sqrt{\lambda_n}$, so changing $\sqrt{\lambda_n}$ to $-\sqrt{\lambda_n}$ merely scales $\phi_n(x)$ by a factor of -1. ∎

EXAMPLE 4. As a final example, consider the problem

$$y'' - 2y' + \lambda y = 0, \qquad (0 < x < \pi) \tag{31a}$$

$$y(0) = 0, \quad y(\pi) = 0. \tag{31b}$$

It appears that (31) may not be a Sturm–Liouville problem at all since the written-out version of (1a) is

$$py'' + p'y' + qy + \lambda wy = 0. \tag{32}$$

That is, *the coefficient of y' needs to be the derivative of the coefficient of y''*. Yet, -2 in (31a) is not the derivative of 1. However, let us multiply (31a) by a yet-to-be-determined function $\sigma(x)$, giving

$$\sigma y'' - 2\sigma y' + \lambda\sigma y = 0 \tag{33}$$

such that the coefficient -2σ of y' is the derivative of the coefficient σ of y'':

$$\sigma' = -2\sigma. \tag{34}$$

Solving, $\sigma(x) = Ce^{-2x}$ and we can take $C = 1$ without loss. Putting that σ into (33) does give the standard Sturm–Liouville form,

$$(e^{-2x}y')' + \lambda e^{-2x}y = 0. \tag{35}$$

Since the factor $\sigma = e^{-2x}$ in (33) is everywhere nonzero, the solution of (31a) and (35) are identical, so the two equations are equivalent.

That step was important for two reasons. First, it establishes the problem as being of Sturm–Liouville type so that we can make use of Theorems 17.7.1 and 17.7.2. Second, it enables us to identify $p(x)$ and $w(x)$: $p(x) = w(x) = e^{-2x}$, so we see that $p(x) > 0$ and $w(x) > 0$ on the closed interval $0 \le x \le \pi$, as required by the theory. And of course we will need to know the weight function $w(x)$ in the inner product if we are to carry out any eigenfunction expansions.

To find the eigenvalues and eigenfunctions, seek an exponential solution form $y(x) = e^{rx}$. Putting that form into (31a), we find that $r = 1 \pm \sqrt{1-\lambda}$, so the general solution of (31a) is

$$y(x) = e^x\left(Ae^{\sqrt{1-\lambda}\,x} + Be^{-\sqrt{1-\lambda}\,x}\right), \tag{36}$$

unless $\lambda = 1$, in which case the two solutions in (36) coalesce into one. Thus, we need to distinguish the two cases $\lambda \neq 1$ and $\lambda = 1$, and write the general solution as

$$y(x) = \begin{cases} e^x[C\sinh(\sqrt{1-\lambda}\,x) + D\cosh(\sqrt{1-\lambda}\,x)], & \lambda \neq 1 \\ e^x(E + Fx), & \lambda = 1 \end{cases} \tag{37a,b}$$

where the sinh, cosh combination will be a bit more convenient than the positive and negative exponentials in (36) because the $y(0) = 0$ boundary condition will give $D = 0$ and will thereby knock out one of the two terms.

Applying the boundary conditions to (37b) gives $E = F = 0$, so $\lambda = 1$ is not an eigenvalue of (31). Applying them to (37a) gives $D = 0$ and the characteristic equation

$$\sinh(\sqrt{1-\lambda}\,\pi) = 0, \tag{38}$$

with C remaining arbitrary. For $\lambda < 1$, (38) has no roots. For $\lambda > 1$, write*

$$\sinh(\sqrt{1-\lambda}\,\pi) = \sinh(i\sqrt{\lambda-1}\,\pi) = i\sin(\sqrt{\lambda-1}\,\pi) = 0, \tag{39}$$

so $\sqrt{\lambda-1}\,\pi = n\pi$ for $n = 1, 2, \ldots$. Thus, the eigenvalues are

$$\lambda_n = 1 + n^2. \qquad (n = 1, 2, \ldots) \tag{40}$$

Further,

$$\begin{aligned} y(x) &= Ce^x \sinh(\sqrt{1-\lambda}\,x) = Ce^x \sinh(i\sqrt{\lambda-1}x) \\ &= iCe^x \sin(\sqrt{\lambda-1}\,x) = iCe^x \sin nx, \end{aligned} \tag{41}$$

so the eigenfunctions are

$$\phi_n(x) = e^x \sin nx. \tag{42}$$

Finally, the eigenfunction expansion of a given function $f(x)$ on $0 < x < \pi$ is

$$f(x) = \sum_{n=1}^{\infty} a_n \phi_n(x), \qquad (0 < x < \pi) \tag{43a}$$

$$\begin{aligned} a_n &= \frac{\langle f, \phi_n \rangle}{\langle \phi_n, \phi_n \rangle} = \frac{\displaystyle\int_0^\pi f(x)(e^x \sin nx)e^{-2x}\,dx}{\displaystyle\int_0^\pi (e^x \sin nx)^2 e^{-2x}\,dx} \\ &= \frac{\displaystyle\int_0^\pi f(x)e^{-x}\sin nx\,dx}{\displaystyle\int_0^\pi \sin^2 nx\,dx} = \frac{2}{\pi}\int_0^\pi f(x)e^{-x}\sin nx\,dx. \end{aligned} \tag{43b}$$

COMMENT. We did not use Theorem 17.7.2 in this example because it addresses the distinction $\lambda > 0$, $\lambda < 0$, whereas here we were concerned with the cases $\lambda > 1$, $\lambda < 1$. ∎

In Examples 1 – 4 it turned out that the separately-considered λ's [$\lambda = 0$ in (4), $\lambda = 0$ in (22), and $\lambda = 1$ in (37)] turned out not to be eigenvalues. Do not discard such values out of hand because they might, in other examples, turn out to be eigenvalues. For instance, you will find that if the boundary conditions in Example

*Here we use the definitions of sinh ( ) and sin ( ):

$$\sinh it = \frac{e^{it} - e^{-it}}{2} = i\,\frac{e^{it} - e^{-it}}{2i} = i\sin it.$$

1 are changed to $y'(0) = 0$ and $y'(L) = 0$, then $\lambda = 0$ is indeed an eigenvalue, with the eigenfunction $\phi(x) = 1$.

17.7.2. Lagrange identity and proofs. (Optional) We will derive a Lagrange identity, and use it to prove parts of Theorem 17.7.1. Proof of Theorem 17.7.2 is left for the exercises. We assume elementary knowledge of the Cartesian representation of complex numbers $z = x + iy$ and the complex conjugate $\overline{z} = x - iy$, as covered in Section 21.2.

When we introduced the inner product (14), we noted that if we are to admit complex-valued functions then we should modify the inner product as

$$\langle f, g\rangle = \int_a^b f(x)\,\overline{g}(x)\,w(x)\,dx. \tag{44}$$

That is, we continue to ask $p(x), q(x), w(x), \alpha, \beta, \gamma, \delta$ to be real, but it is not at all obvious that assumption implies that the eigenvalues and eigenfunctions must be real. For instance, the real matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

has the complex eigenvalues $\lambda = 1 \pm i$ and complex eigenvectors as well.

To begin, observe that the properties

$$\langle f, g\rangle = \overline{\langle g, f\rangle}, \tag{45a}$$
$$\langle \mu f, g\rangle = \mu\langle f, g\rangle, \tag{45b}$$
$$\langle f, \mu g\rangle = \overline{\mu}\langle f, g\rangle, \tag{45c}$$

follow immediately from (44), where μ is any scalar. Since $\langle f, g\rangle \neq \langle g, f\rangle$ in general, according to (45a), it is sometimes useful to specify that f comes first and g comes second in $\langle f, g\rangle$. In what follows we say that $\langle f, g\rangle$ is g **pre-dotted** with f or f **post-dotted** with g, which terminology is not standard.

Let us express (1a) in operator form as

$$\boxed{L[y] = \lambda y,} \tag{46}$$

where L is the differential operator

$$\boxed{L = -\frac{1}{w}\left[\frac{d}{dx}\left(p\frac{d}{dx}\right) + q\right],} \tag{47}$$

and let u and v be any functions having continuous second derivatives on $[a, b]$ and satisfying the homogeneous boundary conditions (1b). Then

$$\begin{aligned}\langle L[u], v\rangle &= \int_a^b -\frac{1}{w}\,[(pu')' + qu]\overline{v}w\,dx \\ &= -\int_a^b [(pu')' + qu]\overline{v}\,dx,\end{aligned} \tag{48}$$

and integrating the $(pu')'\overline{v}$ term by parts twice, so as to undo the derivatives on u, gives

$$\langle L[u], v\rangle = [p(u\overline{v}' - u'\overline{v})]\Big|_a^b - \int_a^b [(p\overline{v}')' + q\overline{v}]u\,dx. \tag{49}$$

Remembering that $p(x)$, $q(x)$, and $w(x)$ are real, we can re-express the last term in (49) as

$$\begin{aligned}\int_a^b [(p\overline{v}')' + q\overline{v}]u\,dx &= \int_a^b \frac{1}{w}[(p\overline{v}')' + q\overline{v}]u\,w\,dx \\ &= \int_a^b \overline{\frac{1}{w}[(pv')' + qv]}\;u\,w\,dx \\ &= -\int_a^b u\,\overline{L[v]}\,w\,dx \\ &= -\langle u, L[v]\rangle. \end{aligned}\tag{50}$$

Further, the boundary term in (49) is zero because u and v satisfy the homogeneous boundary conditions (1b). For instance, suppose $\alpha = \gamma = 1$ and $\beta = \delta = 0$ in (1b), so $u(a) = 0$, $u(b) = 0$, $v(a) = 0$, and $v(b) = 0$. It follows from the latter two that $\overline{v}(a) = 0$ and $\overline{v}(b) = 0$, so

$$[p(u\overline{v}' - u'\overline{v})]\Big|_a^b = (0 - 0) - (0 - 0) = 0; \tag{51}$$

similarly for any α and β (not both zero) and any γ and δ (not both zero), verification of which claim is left for the exercises. Thus, (49) becomes

$$\boxed{\langle L[u], v\rangle = \langle u, L[v]\rangle,} \tag{52}$$

which formula is known as the **Lagrange identity**.

Proof of Theorem 17.7.1, part (a): Let λ be an eigenvalue of (1) and ϕ a corresponding eigenfunction, so

$$L[\phi] = \lambda\phi. \tag{53}$$

Post-dotting (53) with ϕ and pre-dotting (53) with ϕ gives

$$\langle L[\phi], \phi\rangle = \langle \lambda\phi, \phi\rangle = \lambda\langle \phi, \phi\rangle \tag{54a}$$

from (45b), and

$$\langle \phi, L[\phi]\rangle = \langle \phi, \lambda\phi\rangle = \overline{\lambda}\langle \phi, \phi\rangle \tag{54b}$$

from (45c), respectively. Subtracting (54) from (54a) gives, by virtue of the Lagrange identity,

$$0 = (\lambda - \overline{\lambda})\langle \phi, \phi\rangle = (\lambda - \overline{\lambda})\,\|\phi\|^2. \tag{55}$$

The factor $\|\phi\|^2$ is nonzero because ϕ is an eigenfunction, so it follows from (55) that $\lambda - \overline{\lambda} = 0$. Thus, $\overline{\lambda} = \lambda$ so λ is real, as was to be proved. ■

Proof of Theorem 17.7.1, part (b): Let ϕ_j and ϕ_k be eigenfunctions corresponding to distinct eigenvalues λ_j and λ_k, respectively. Thus,

$$L[\phi_j] = \lambda_j \phi_j \qquad \text{and} \qquad L[\phi_k] = \lambda_k \phi_k. \tag{56a,b}$$

If we dot ϕ_k into each side of (56a) and dot each side of (56b) into ϕ_j [i.e., we pre-dot (56a) with ϕ_k and post-dot (56b) with ϕ_j], we obtain

$$\begin{aligned} \langle \phi_k, L[\phi_j] \rangle &= \langle \phi_k, \lambda_j \phi_j \rangle \\ &= \overline{\lambda}_j \langle \phi_k, \phi_j \rangle, \end{aligned} \tag{57}$$

and

$$\begin{aligned} \langle L[\phi_k], \phi_j \rangle &= \langle \lambda_k \phi_k, \phi_j \rangle \\ &= \lambda_k \langle \phi_k, \phi_j \rangle, \end{aligned} \tag{58}$$

respectively. The left-hand sides are equal by virtue of the Lagrange identity, and $\overline{\lambda}_j = \lambda_j$ because the λ's are real, so subtraction of (57) from (58) gives

$$(\lambda_j - \lambda_k)\langle \phi_k, \phi_j \rangle = 0. \tag{59}$$

Finally, $\lambda_j - \lambda_k \neq 0$ by assumption, so it follows that $\langle \phi_k, \phi_j \rangle = 0$, as was to be proved. ■

Before closing this section let us explain the significance of the Lagrange identity (52). The operator L is the differential operator

$$L = \frac{1}{w}\left[\frac{d}{dx}\left(p\frac{d}{dx}\right) + q\right] \tag{60a}$$

on the domain $\mathcal{D}$ of functions, u say, that are defined and have continuous second derivatives on $[a, b]$ and that satisfy the homogeneous boundary conditions*

$$\begin{aligned} \alpha u(a) + \beta u'(a) &= 0, \\ \gamma u(b) + \delta u'(b) &= 0. \end{aligned} \tag{60b}$$

*Note that the definition of the domain $\mathcal{D}$ is part of the definition of L, just as the domain of definition is part of the definition of a function. For instance, the function whose values are $\sin x$ on $0 \le x \le \pi$ is different from the function whose values are $\sin x$ on $-\pi \le x \le 23$. Thus, the operator is the differential operator (60a) plus the domain of definition. In this discussion, although not elsewhere in this text, we distinguish between the *differential operator* [namely, the "action" (60a)] and *operator* [namely, the action (60a) plus the domain of definition $\mathcal{D}$]. By the way, why do we ask functions $u(x)$ in $\mathcal{D}$ to have continuous second derivatives? They need to be twice differentiable so that $L[u]$ exists because L is a second-order differential operator. Further, we wish to ensure that the inner product integral $\langle L[u], v \rangle$ exists, and we can do that by asking the integrand to be continuous on $[a, b]$. In fact, $L[u]$ will be continuous if u has a continuous second derivative.

More generally, the equation

$$\boxed{\langle L[u], v\rangle = \langle u, L^*[v]\rangle} \tag{61}$$

is used to *define* the **Hermitian conjugate** (or **adjoint**) L^* of the operator L, relative to whatever inner product is chosen. Let us illustrate.

EXAMPLE 5. Find the Hermitian conjugate of the operator consisting of the differential operator

$$L = \frac{d^2}{dx^2} + \frac{d}{dx} + 1 \tag{62a}$$

on the domain $\mathcal{D}$ of real-valued functions defined and having continuous second derivatives on $[0, \pi]$ and satisfying the homogeneous initial conditions

$$u(0) = 0, \qquad u'(0) = 0, \tag{62b}$$

subject to the inner product definition

$$\langle u, v\rangle = \int_0^\pi u(x)v(x)\,dx, \tag{62c}$$

say. Begin with the left-hand side of (61),

$$\langle L[u], v\rangle = \int_0^\pi (u'' + u' + u)v\,dx. \tag{63}$$

Integrating the $u''v$ term by parts twice, the $u'v$ term once, leaving the uv term intact, and using (62b), gives

$$\begin{aligned}\langle L[u], v\rangle &= (u'v - uv' + uv)\Big|_0^\pi + \int_0^\pi u(v'' - v' + v)\,dx \\ &= [u'(\pi) + u(\pi)]v(\pi) - [u(\pi)]v'(\pi) + \langle u, L^*[v]\rangle,\end{aligned} \tag{64}$$

where, from the $v'' - v' + v$ in the integral, we can infer that

$$L^* = \frac{d^2}{dx^2} - \frac{d}{dx} + 1. \tag{65a}$$

To obtain the boundary conditions associated with L^* we see, by comparing (64) with (61), that we need the boundary terms in (64) to drop out. Whereas $u(0) = 0$ and $u'(0) = 0$, the bracketed quantities $u'(\pi) + u(\pi)$ and $u(\pi)$ are not prescribed, so we must have both

$$v(\pi) = 0, \qquad v'(\pi) = 0. \tag{65b}$$

Thus, the Hermitian conjugate operator is the differential operator (65a) on the domain $\mathcal{D}^*$ of real-valued functions defined and having continuous second derivatives on $[0, \pi]$ and satisfying the conditions (65b). ∎

If the operator and its Hermitian conjugate (or adjoint) are identical, then we say that it is **Hermitian** (or **self-adjoint**). Thus, the operator in Example 5 is not

Hermitian because the L^* in (65a) differs from the L in (62a) (by the minus sign in front of the d/dx) and also because the boundary conditions (65b) on functions in $\mathcal{D}^*$ are different from the boundary conditions (62b) on functions in $\mathcal{D}$. Either of these differences would be sufficient to conclude that the operator is not Hermitian.

EXAMPLE 6. *Matrix Case.* Find the Hermitian conjugate of a real $n \times n$ matrix operator $\mathbf{A}$, defined on the vector space $\mathbb{R}^n$, with the dot product $\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^{\mathrm{T}}\mathbf{y}$, where $\mathbf{x}$ and $\mathbf{y}$ are n-dimensional column vectors. $\mathbf{A}^*$ is defined by requiring that

$$(\mathbf{A}\mathbf{x}) \cdot \mathbf{y} = \mathbf{x} \cdot (\mathbf{A}^*\mathbf{y}) \tag{66}$$

for all $\mathbf{x}$'s in the domain $\mathcal{D}$ of $\mathbf{A}$ and for all $\mathbf{y}$'s in the domain $\mathcal{D}^*$ of $\mathbf{A}^*$. Since $\mathbf{A}$ is $n \times n$ and $\mathbf{x}$ is $n \times 1$, the $\mathbf{A}\mathbf{x}$ in (66) is $n \times 1$. For the dot product on the left to be defined we need $\mathbf{y}$ to be $n \times 1$ as well. On the right, $\mathbf{x}$ is $n \times 1$, so we need $\mathbf{A}^*\mathbf{y}$ to be $n \times 1$. Since $\mathbf{y}$ is $n \times 1$, $\mathbf{A}^*$ needs to be $n \times n$. Thus, like $\mathbf{A}$, $\mathbf{A}^*$ is an $n \times n$ matrix operator defined on $\mathbb{R}^n$. To determine $\mathbf{A}^*$, begin with the left-hand side of (66),

$$(\mathbf{A}\mathbf{x}) \cdot \mathbf{y} = (\mathbf{A}\mathbf{x})^{\mathrm{T}}\mathbf{y} = \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{y} = \mathbf{x} \cdot (\mathbf{A}^{\mathrm{T}}\mathbf{y}), \tag{67}$$

so the Hermitian conjugate $\mathbf{A}^*$ of $\mathbf{A}$ is

$$\mathbf{A}^* = \mathbf{A}^{\mathrm{T}}. \tag{68}$$

Thus, $\mathbf{A}$ is Hermitian if and only if $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$, that is, if $\mathbf{A}$ is symmetric. Just as the eigenvalue problem for the Hermitian Sturm-Liouville problem is of great importance, so is the eigenvalue problem for real symmetric (hence Hermitian) matrices and, indeed, that case is singled out for study in Section 11.3. ∎

Closure. The Sturm-Liouville eigenvalue problem is the differential equation analog of the matrix eigenvalue problem $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, where $\mathbf{A}$ is real and symmetric (hence Hermitian). In both cases the eigenvalues are real and the eigenvectors provide an orthogonal basis for the relevant vector space. The expansion formula (15) is the analog of formula (23) in Section 9.9.

Although the Sturm-Liouville and matrix eigenvalue problems are closely related (if $\mathbf{A}$ is Hermitian), the Sturm-Liouville case is much more subtle because the vector space is infinite-dimensional and expansions are, in general, infinite series. For instance, in a five-dimensional space five orthogonal vectors necessarily constitute a basis, but in an infinite-dimensional space an infinite number of orthogonal vectors need not constitute a basis. Thus, part (d) of Theorem 17.7.1 is a deep result, and we put it forward without proof.

If we were to generalize the real symmetric matrix and Sturm-Liouville eigenvalue problems further, we would study Hermitian operators, with matrix, ordinary differential, partial differential, and integral operators merely occurring as special cases. Such study is beyond our present scope and falls within the domain of mathematics known as functional analysis. Rather than being exceptional, the operators encountered in applications are often Hermitian.

Computer software. To find the roots of (24) using *Maple* we can use the fsolve command, but it is best to include the option of specifying the interval in which to search. From Fig. 3, it is evident that the first $\sqrt{\lambda}$ root falls in the interval $[\pi/2, 3]$. Thus, there is a λ root in $[2.46, 9]$, so enter

$$\text{fsolve}(\tan(\text{sqrt}(x)) + 2 * \text{sqrt}(x) = 0,\ x,\ 2.46..9);$$

and return. The result is the first root, $\lambda = 3.3731$. For the second root enter

$$\text{fsolve}(\tan(\text{sqrt}(x)) + 2 * \text{sqrt}(x) = 0,\ x,\ 22..36);$$

and so on.

EXERCISES 17.7

1. Identify $p(x), q(x), w(x), \alpha, \beta, \gamma, \delta$, solve for the eigenvalues and eigenfunctions, and work out the eigenfunction expansion of the given function f. If the characteristic equation is too difficult to solve analytically, state that and proceed with the rest of the problem as though the λ_n's were known.

(a) $y'' + \lambda y = 0,$
$y(0) = 0, \quad y'(L) = 0, \quad f(x) = 100$

(b) $y'' + \lambda y = 0,$
$y'(0) = 0, \quad y(L) = 0, \quad f(x) = 1$

(c) $y'' + \lambda y = 0,$
$$y'(0) = 0, \quad y'(L) = 0, \quad f(x) = \begin{cases} 1, & 0 \le x < L/2 \\ 0, & L/2 \le x \le L \end{cases}$$

(d) $y'' + \lambda y = 0,$
$y'(0) = 0, \quad y(L) + y'(L) = 0, \quad f(x) = 50$

(e) $y'' + \lambda y = 0,$
$y(0) + y'(0) = 0, \quad y(\pi) = 0, \quad f(x) = 10$

(f) $y'' + \lambda y = 0,$
$$y'(-1) = 0,\ y'(1) = 0,\ f(x) = \begin{cases} 0, & -1 \le x \le 0 \\ 50, & 0 < x \le 1 \end{cases}$$

(g) $y'' + \lambda y = 0,$
$y(0) - 2y'(0) = 0, \quad y'(2) = 0, \quad f(x) = 100$

2. We pointed out that the Sturm-Liouville problem in Example 1 generated the half-range sine expansion studied in Section 17.4. Modify the boundary conditions in that example so as to generate, instead, the

(a) half-range cosine expansion
(b) quarter-range sine expansion
(c) quarter-range cosine expansion

3. (*Obtaining Sturm–Liouville form*) We observed, in Example 4, that the equation

$$A(x)y'' + B(x)y' + C(x)y + \lambda D(x)y = 0 \qquad (3.1)$$

is in the standard Sturm-Liouville form (1a) only if $B(x) = A'(x)$. Show that if $A(x) \neq 0$ on $[a, b]$ and $(B - A')/A$ is continuous on $[a, b]$, then we can recast (3.1) in the form (1a) by multiplying (3.1) by

$$\sigma(x) = e^{\int [(B-A')/A]\,dx}. \qquad (3.2)$$

4. Use the results of Exercise 3 to recast each of the following differential equations in the Sturm–Liouville form (1a). Identify $p(x)$, $q(x)$, and $w(x)$.

(a) $xy'' + 5y' + \lambda xy = 0$
(b) $y'' + 2y' + xy + \lambda x^2 y = 0$
(c) $y'' + y' + \lambda y = 0$
(d) $y'' - y' + \lambda xy = 0$
(e) $x^2 y'' + xy' + \lambda x^2 y = 0$
(f) $y'' + (\cot x)y' + \lambda y = 0$

5. Use computer software to find $\lambda_1, \ldots, \lambda_8$ from (24), to seven significant figures.

6. Consider the eigenvalue problem

$$y'' + \lambda y = 0, \qquad (0 < x < 1)$$

$$2y(0) - y(1) + 4y'(1) = 0, \qquad y(0) + 2y'(1) = 0.$$

Explain why the latter problem is not of Sturm-Liouville type. Using computer software, determine any two eigenvalues.

HINT: You should obtain the characteristic equation

$$\sin\sqrt{\lambda} = 2\sqrt{\lambda}.$$

Although the latter equation has no roots on a real $\sqrt{\lambda}$ axis, we need to search in the complex plane. With $z = x + iy$ write $\sin z = 2z$, use the identity $\sin z = \sin(x+iy) = \sin x \cosh y + i \cos x \sinh y$ and obtain the equations $\sin x \cosh y = 2x$, $\cos x \sinh y = 2y$ on x and y. Then, use computer software to find any two solution pairs for x and y, and hence for λ.

7. Show that for

$$\begin{aligned} y'' + \lambda y &= 0, \qquad (0 < x < 1) \\ y(0) - y(1) &= 0, \qquad y'(0) + y'(1) = 0 \end{aligned}$$

every λ (real or complex) is an eigenvalue! Is the latter a Sturm–Liouville system? Explain.

8. Consider the eigenvalue problem

$$\begin{aligned} x^2y'' + xy' + \lambda y &= 0, \qquad (1 < x < a) \\ y(1) &= 0, \qquad y(a) = 0. \end{aligned}$$

(a) Show that the eigenvalues and eigenfunctions are

$$\lambda_n = \frac{n^2\pi^2}{(\ln a)^2}, \qquad \phi_n(x) = \sin\left(n\pi \frac{\ln x}{\ln a}\right)$$

for $n = 1, 2, \ldots$.
(b) Show that the eigenfunction expansion of a given function f is of the form

$$f(x) = \sum_{n=1}^{\infty} c_n \sin\left(n\pi \frac{\ln x}{\ln a}\right),$$

where

$$c_n = \frac{\displaystyle\int_1^a f(x) \sin\left(n\pi \frac{\ln x}{\ln a}\right) \frac{dx}{x}}{\displaystyle\int_1^a \sin^2\left(n\pi \frac{\ln x}{\ln a}\right) \frac{dx}{x}}.$$

HINT: You will need to get the differential equation into Sturm–Liouville form, as discussed in Exercise 3, before you can identify the weight function $w(x)$ for the inner product.

9. Expand the function

$$f(x) = \begin{cases} x^4, & 0 \le x < 2 \\ 0, & 2 \le x \le \pi \end{cases}$$

in terms of the eigenfunctions of the given eigenvalue problem. Use computer software, such as the *Maple* int command, to evaluate the expansion coefficient a_n as a function of n.

(a) $y'' + \lambda y = 0, \quad y'(0) = 0, \; y'(\pi) = 0$
(b) $y'' + \lambda y = 0, \quad y'(0) = 0, \; y(\pi) = 0$
(c) $y'' + \lambda y = 0, \quad y(0) = 0, \; y'(\pi) = 0$

10. Prove that the Sturm–Liouville eigenvalues are simple, as stated in part (b) of Theorem 17.7.1. HINT: Suppose that ϕ_1 and ϕ_2 are two eigenfunctions corresponding to an eigenvalue λ of (1), and suppose $\alpha \neq 0$ in (1b). Then the Wronskian $W(x)$ of ϕ_1 and ϕ_2, evaluated at $x = a$, is

$$\begin{aligned} W(a) &= \begin{vmatrix} \phi_1(a) & \phi_2(a) \\ \phi_1'(a) & \phi_2'(a) \end{vmatrix} \\ &= \begin{vmatrix} -\dfrac{\beta}{\alpha}\phi_1'(a) & -\dfrac{\beta}{\alpha}\phi_2'(a) \\ \phi_1'(a) & \phi_2'(a) \end{vmatrix} = 0. \end{aligned} \tag{10.1}$$

On the other hand, if α does equal zero in (1b) then (1b) becomes $y'(0) = 0$, so

$$W(a) = \begin{vmatrix} \phi_1(a) & \phi_2(a) \\ \phi_1'(a) & \phi_2'(a) \end{vmatrix} = \begin{vmatrix} \phi_1(a) & \phi_2(a) \\ 0 & 0 \end{vmatrix} = 0 \tag{10.2}$$

once again. Show that it follows from these results and Liouville's formula that $W(x) = 0$ on $[a, b]$, and cite an appropriate theorem which then implies that ϕ_1 and ϕ_2 must be linearly dependent on $[a, b]$.

11. (*Real eigenfunctions*) Show that if $\phi(x)$ is an eigenfunction of a Sturm–Liouville problem, then $\phi(x)$ is either a real-valued function or else it is a complex constant times a real-valued function. HINT: Show that if $\phi(x)$ is an eigenfunction corresponding to an eigenvalue λ, then so is $\overline{\phi}(x)$. Then use part (b) of Theorem 17.7.1 (namely, that the eigenvalues are simple) to show that $\phi(x) = c\overline{\phi}(x)$, where c is a constant. Expressing the latter equation in the (polar) form $A(x)e^{iB(x)} = Ce^{iD}A(x)e^{-iB(x)}$, show that $B(x)$ is, at most, a constant.

12. Prove Theorem 17.7.2. HINT: You may assume that the eigenfunctions are real (proof of which is outlined in Exercise 10). Multiply each term in $(py')' + qy + \lambda wy = 0$ by $\overline{y}$, and integrate over the $[a, b]$ interval. Thus, show that

$$\lambda \|y\|^2 = -(py'\overline{y})\Big|_a^b + \int_a^b p\,|y'|^2\,dx - \int_a^b q\,|y|^2\,dx, \tag{12.1}$$

and examine the signs of the individual terms.

13. (*Column buckling with lateral restraint*) Consider the buckling of the column of length L and stiffness EI shown in the figure. It is fixed into the floor such that $y(0) = y'(0) = 0$, and it is restrained laterally, at the free end, by a spring of stiffness k. Then it turns out that the lateral deflection $y(x)$ is governed by the eigenvalue problem

$$\begin{aligned} y'''' + \lambda y'' &= 0, \qquad (0 < x < L) \\ y(0) = y'(0) &= y''(L) = 0, \\ y'''(L) &= -\lambda y'(L) + \kappa y(L), \end{aligned} \tag{13.1}$$

where $\lambda \equiv P/EI$ and $\kappa = k/EI$.

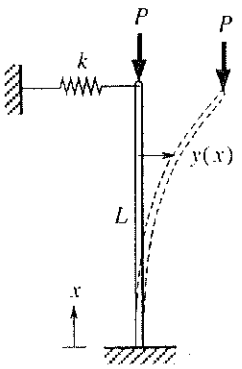
(a) Show that the characteristic equation is

$$(\Lambda^2 - \kappa L)\Lambda \cos \Lambda L + \kappa \sin \Lambda L = 0, \tag{13.2}$$

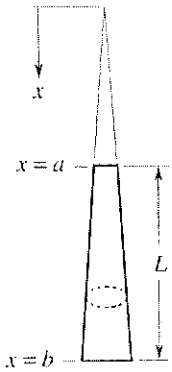
where $\Lambda = \sqrt{\lambda}$, and that the corresponding eigenfunctions (buckling modes) are (to within an arbitrary scale factor)

$$y = \sin \Lambda x - \tan \Lambda L \cos \Lambda x - \Lambda x + \tan \Lambda L. \tag{13.3}$$

(b) Solve (13.2) for the critical buckling load, P_{cr}, for the case where $\kappa = 0$.
(c) Is (13.1) a Sturm–Liouville problem? Explain.



14. (*Buckling of linearly tapered column*) Consider a column of circular cross section, the radius of which varies linearly with x. It extends over $a < x < b$, as shown in the figure,



is pinned at both ends, and is loaded axially by a force P. Then the cross-sectional moment of inertia I is not a constant; it is given by $I(x) = I_0(x/b)^4$, where the constant I_0 is the value of $I(x)$ at $x = b$, so the eigenvalue problem governing buckling is found to be

$$\begin{aligned} x^4 y'' + \lambda y &= 0, \qquad (a < x < b) \\ y(a) = 0, &\quad y(b) = 0, \end{aligned} \tag{14.1}$$

where $\lambda \equiv b^4 P/EI_0$.

(a) Verify that (for the case $\lambda \neq 0$)the general solution of the differential equation can be expressed as

$$y(x) = x\left[A \sin\left(\frac{\sqrt{\lambda}}{x}\right) + B \cos\left(\frac{\sqrt{\lambda}}{x}\right)\right]. \tag{14.2}$$

(b) Applying the boundary conditions, show that the eigenvalues and eigenfunctions are

$$\lambda_n = \left(\frac{n\pi ab}{L}\right)^2, \tag{14.3a}$$

$$\phi_n(x) = x \sin\left[n\pi \frac{b}{L}\left(1 - \frac{a}{x}\right)\right] \tag{14.3b}$$

for $n = 1, 2, \ldots$, and that the buckling load is $P_{cr} = \pi^2 EI_0 a^2/b^2 L^2$, where $L = b - a$.

15. (*Buckling of nonlinearly tapered column*) Although not wishing to give undue prominence to the subject of the buckling of columns, we include this final exercise on buckling, which we believe is interesting and challenging. If, in the problem of Exercise 14, the column radius is proportional to $\sqrt{x}$ rather than to x, then $I(x) = I_0(x/b)^2$, and we have

$$\begin{aligned} x^2 y'' + \lambda y &= 0, \qquad (a < x < b) \\ y(a) = 0, &\quad y(b) = 0, \end{aligned} \tag{15.1}$$

where $\lambda \equiv b^2 P/EI_0$. Show that the buckling load is

$$P_{cr} = \frac{EI_0}{b^2}\left\{\frac{1}{4} + \frac{\pi^2}{[\ln(b/a)]^2}\right\}. \tag{15.2}$$

16. We show in (51) that the boundary term in (49) is zero for the special simple case where $\alpha = \gamma = 1$ and $\beta = \delta = 0$. Prove that the boundary term is zero for *any* α and β (not both zero) and for *any* γ and δ (not both zero).

17. Find the Hermitian conjugate (i.e., the adjoint) of the given operator and state whether the given operator is Hermitian (self-adjoint). If it is not, state why it is not. In each case, the interval is $0 \le x \le 1$, and the inner product is

$$\langle f, g\rangle = \int_0^1 f(x)g(x)\,dx.$$

(a) $L = \dfrac{d}{dx}, \quad u(0) = 0$

(b) $L = \dfrac{d}{dx}, \quad u(1) = 0$

(c) $L = \dfrac{d^2}{dx^2}, \quad u(0) = u'(0) = 0$

(d) $L = \dfrac{d^2}{dx^2}, \quad u'(0) = u'(1) = 0$

(e) $L = \dfrac{d^2}{dx^2} + 3, \quad u'(0) = u'(1) = 0$

(f) $L = \dfrac{d^2}{dx^2} + \dfrac{d}{dx}, \quad u(0) = u(1) = 0$

(g) $L = \dfrac{d^3}{dx^3} - \dfrac{d^2}{dx^2} + 2\dfrac{d}{dx}, \quad u(0) = u'(0) = u'(1) = 0$

(h) $L = \dfrac{d^2}{dx^2} - 1, \quad u(0) + u'(0) = u'(1) = 0$

(i) $L = \dfrac{d^2}{dx^2}, \quad u'(0) = u(1) + 5u'(1) = 0$

(j) $L = \dfrac{d^2}{dx^2}, \quad u(0) - u(1) = u'(0) - u'(1) = 0$

17.8 Periodic and Singular Sturm–Liouville Problems

The Sturm–Liouville problem studied in Section 17.7 consists of the linear homogeneous second-order differential equation

$$[p(x)y']' + q(x)y + \lambda w(x)y = 0, \qquad (a < x < b) \tag{1a}$$

with homogeneous boundary conditions of the form

$$\begin{aligned} \alpha y(a) + \beta y'(a) &= 0, \\ \gamma y(b) + \delta y'(b) &= 0, \end{aligned} \tag{1b}$$

where a, b are finite, where p, p', q, w are continuous on $[a, b]$, and where $p(x) > 0$ and $w(x) > 0$ on $[a, b]$. The latter is generally known as a **regular** Sturm–Liouville problem, and many powerful results followed, as given in Theorems 17.7.1 and 17.7.2. We say that the boundary conditions (1b) are **separated** since one condition applies at $x = a$ and the other at $x = b$.

If any of the conditions cited above are not met, then the results obtained in Theorems 17.7.1 and 17.7.2 may not hold. Among the nonregular versions of the Sturm–Liouville problem, two are especially prominent and are the subject of this section: the Sturm–Liouville problem with **periodic boundary conditions**, and the **singular** Sturm–Liouville problem. In these cases the conditions cited above are met, except as noted below.

Periodic boundary conditions. In this case we have, in place of the separated boundary conditions (1b), the nonseparated conditions

$$\boxed{\begin{aligned} y(a) &= y(b), \\ y'(a) &= y'(b), \end{aligned}} \tag{2}$$

which are known as *periodic boundary conditions*, for reasons that will become clear when we work an example.

Singular case. In this case $p(x)$ [and possibly $w(x)$] vanishes at one or both endpoints, so that $p(x) > 0$ and $w(x) > 0$ holds on the open interval (a, b) rather than on the closed interval $[a, b]$. Further, the boundary conditions are modified as follows.

$\boldsymbol{p(a) = 0}$ [and $p(b) \neq 0$]: Then the boundary conditions are

$$\boxed{\begin{aligned} &y \text{ bounded at } a, \\ &\gamma y(b) + \delta y'(b) = 0. \end{aligned}} \tag{3}$$

$\boldsymbol{p(b) = 0}$ [and $p(a) \neq 0$]: Then the boundary conditions are

$$\boxed{\begin{aligned} &\alpha y(a) + \beta y'(a) = 0, \\ &y \text{ bounded at } b. \end{aligned}} \tag{4}$$

$\boldsymbol{p(a) = p(b) = 0}$: Then the boundary conditions are

$$\boxed{\begin{aligned} &y \text{ bounded at } a, \\ &y \text{ bounded at } b. \end{aligned}} \tag{5}$$

By y being bounded at a, for example, we mean that $\lim_{x \to a} y(x)$ exists (and is therefore finite).

For these cases we have the following results.

THEOREM 17.8.1 *Periodic and Singular Cases*
Let λ_n and $\phi_n(x)$ denote any eigenvalue and corresponding eigenfunction of a Sturm–Liouville problem with periodic boundary conditions, given by (2), or a singular Sturm–Liouville problem (as defined above).

(a) The eigenvalues are real.

(b) If $q(x) \leq 0$ on $[a, b]$ and $[p(x)\phi_n(x)\phi_n'(x)]|_a^b \leq 0$ for the eigenfunction $\phi_n(x)$, then not only is λ_n real, it is also nonnegative: $\lambda_n \geq 0$.

(c) Eigenfunctions corresponding to distinct eigenvalues are orthogonal. That is, if $\lambda_j \neq \lambda_k$, then $\langle \phi_j, \phi_k \rangle = 0$.

As for the regular case, positive statements can be made about the completeness of the sets of orthogonal eigenfunctions generated by these problems, in the sense of their being bases for the eigenfunction expansion representation of sufficiently well behaved functions on the interval $a < x < b$.

EXAMPLE 1. *Periodic Boundary Conditions.* Consider the Sturm–Liouville problem

$$y'' + \lambda y = 0, \qquad (-L < x < L) \tag{6a}$$

$$y(-L) = y(L), \quad y'(-L) = y'(L). \tag{6b}$$

We begin with the general solution

$$y(x) = \begin{cases} A\cos\sqrt{\lambda}\,x + B\sin\sqrt{\lambda}\,x, & \lambda \neq 0 \\ C + Dx, & \lambda = 0. \end{cases} \tag{7a,b}$$

For $\lambda = 0$ the boundary conditions (6b) give $C - DL = C + DL$ and $D = D$, so $D = 0$ and C is arbitrary. Thus, $y(x) = C$, so $\lambda = 0$ is an eigenvalue and has the eigenfunction $\phi(x) = 1$. For $\lambda \neq 0$ the boundary conditions (6b) give

$$(\sin\sqrt{\lambda}\,L)B = 0 \tag{8a}$$

$$(\sin\sqrt{\lambda}\,L)A = 0. \tag{8b}$$

Thus, either $A = B = 0$, which result we reject because it gives only the trivial solution $y(x) = 0$, or else

$$\sin\sqrt{\lambda}\,L = 0 \tag{9}$$

and A, B are arbitrary. Since $q(x) = 0$, and

$$\begin{aligned} [p(x)\phi_n(x)\phi_n'(x)]\Big|_{-L}^{L} &= \phi_n(L)\phi_n'(L) - \phi_n(-L)\phi_n'(-L) \\ &= \phi_n(L)\phi_n'(L) - \phi_n(L)\phi_n'(L) = 0, \end{aligned}$$

it follows from part (b) of Theorem 17.8.1 that $\lambda_n \geq 0$. Thus, $\sqrt{\lambda}$ in (9) is real and (9) has the roots $\sqrt{\lambda}\,L = n\pi$, so the eigenvalues and eigenfunctions are

$$\lambda_0 = 0, \qquad \phi_0(x) = 1 \tag{10a}$$

$$\lambda_n = \left(\frac{n\pi}{L}\right)^2, \qquad \phi_n(x) = \cos\frac{n\pi x}{L} \text{ and } \sin\frac{n\pi x}{L} \tag{10b}$$

for $n = 1, 2, \ldots$.

Remember that there is nothing inappropriate about an eigenvalue being zero. It is the eigen*function* that is to be nontrivial, and $\lambda_0 = 0$ does give the nontrivial solution $\phi_0(x) = 1$.

Observe that the eigenvalues $\lambda_1, \lambda_2, \ldots$ are nonsimple since each one has two linearly independent eigenfunctions. This result could not have occurred in a regular Sturm–Liouville problem, which must have simple eigenvalues [part (b) of Theorem 17.7.1], and is due to the periodic boundary conditions.

Noting that the weight function is $w(x) = 1$, the eigenfunction expansion of a given function f on $(-L, L)$ takes the form

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi x}{L} + b_n \sin \frac{n\pi x}{L} \right), \tag{11a}$$

$$a_0 = \frac{\langle f, 1 \rangle}{\langle 1, 1 \rangle} = \frac{\displaystyle\int_{-L}^{L} f(x)\, 1\, dx}{\displaystyle\int_{-L}^{L} 1^2\, dx} = \frac{1}{2L} \int_{-L}^{L} f(x)\, dx, \tag{11b}$$

$$a_n = \frac{\langle f, \cos \frac{n\pi x}{L} \rangle}{\langle \cos \frac{n\pi x}{L}, \cos \frac{n\pi x}{L} \rangle} = \frac{\displaystyle\int_{-L}^{L} f(x) \cos \frac{n\pi x}{L}\, dx}{\displaystyle\int_{-L}^{L} \cos^2 \frac{n\pi x}{L}\, dx}$$

$$= \frac{1}{L} \int_{-L}^{L} f(x) \cos \frac{n\pi x}{L}\, dx, \tag{11c}$$

$$b_n = \frac{\langle f, \sin \frac{n\pi x}{L} \rangle}{\langle \sin \frac{n\pi x}{L}, \sin \frac{n\pi x}{L} \rangle} = \frac{\displaystyle\int_{-L}^{L} f(x) \sin \frac{n\pi x}{L}\, dx}{\displaystyle\int_{-L}^{L} \sin^2 \frac{n\pi x}{L}\, dx}$$

$$= \frac{1}{L} \int_{-L}^{L} f(x) \sin \frac{n\pi x}{L}\, dx. \tag{11d}$$

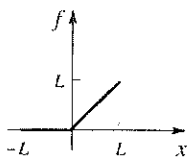


Figure 1. Graph of f.

The result (11) is seen to be the same as the classical Fourier series of a $2L$-periodic function, which we studied in Section 17.3. Thus the expansion (11a) of the function f shown in Fig. 1, for example, will be the same as the classical Fourier series of the $2L$-periodic function shown in Fig. 2, which result explains why we call (2) "periodic boundary conditions."

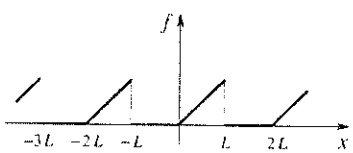


Figure 2. $2L$-periodic f.

COMMENT. As usual, the function f being expanded (e.g., f in Fig. 1) does not itself need to satisfy the boundary conditions imposed on the eigenfunctions [(6b) in this case]. ∎

EXAMPLE 2. *A Bessel Equation.* Consider the singular Sturm–Liouville problem

$$(xy')' + \lambda xy = 0, \qquad (0 < x < L) \tag{12a}$$

$$y(0) \text{ bounded}, \quad y(L) = 0. \tag{12b}$$

Since (12a) is already in the standard Sturm–Liouville form $(py')' + qy + \lambda wy = 0$, we can see that $p(x) = x$, $q(x) = 0$, and $w(x) = x$. And since $p(x)$ and $w(x)$ vanish at the left endpoint $x = 0$, the problem (12) is singular; hence the boundary condition adopted at $x = 0$ is simply a boundedness condition.

To solve (12a), we notice that (12a) would be a Bessel equation of order zero if the λ were not present. Let us try to convert (12a) to a Bessel equation by a simple scaling, $x = \alpha t$, where α is to be determined. Under that change of variables (12a) becomes

$$(tT')' + \alpha^2 \lambda t T = 0, \tag{13}$$

where $T(t) \equiv y(x(t)) = y(\alpha t)$, and where the primes denote d/dt. Thus, we can remove the λ by choosing $\alpha = 1/\sqrt{\lambda}$ (tentatively assuming that $\lambda \neq 0$), so $t = \sqrt{\lambda}\, x$. Then (13) becomes $(tT')' + tT = 0$, with the general solution $T(t) = AJ_0(t) + BY_0(t)$. Or, reverting to x and y,

$$y(x) = AJ_0(\sqrt{\lambda}\, x) + BY_0(\sqrt{\lambda}\, x). \tag{14}$$

However, for $\lambda = 0$ the latter fails to provide the general solution of (12a) because $Y_0(0) = -\infty$ is undefined. But if $\lambda = 0$ then (12a) becomes $(xy')' = 0$ which can be integrated to give $y(x) = C + D \ln x$. Thus, let us write the general solution of (12a) as

$$y(x) = \begin{cases} AJ_0(\sqrt{\lambda}\, x) + BY_0(\sqrt{\lambda}\, x), & \lambda \neq 0 \\ C + D \ln x, & \lambda = 0. \end{cases} \tag{15a,b}$$

For $\lambda = 0$, the boundedness condition requires that $D = 0$, and then $y(L) = 0$ gives $C = 0$, so $y(x) = 0$. Therefore, $\lambda = 0$ is not an eigenvalue. For $\lambda \neq 0$, the boundedness condition requires that $B = 0$ (because $Y_0 \to -\infty$ as its argument tends to zero), so

$$y(x) = AJ_0(\sqrt{\lambda}\, x). \tag{16}$$

Then, the other boundary condition gives

$$y(L) = 0 = AJ_0(\sqrt{\lambda}\, L). \tag{17}$$

If $A = 0$, then (16) becomes the trivial solution, so let us satisfy (17) by asking that

$$J_0(\sqrt{\lambda}\, L) = 0 \tag{18}$$

instead. Now, $q(x) = 0$, and

$$\left[p(x)\phi_n(x)\phi_n'(x)\right]\Big|_0^L = \left[x\phi_n(x)\phi_n'(x)\right]\Big|_0^L \tag{19}$$

is zero because $\phi_n(L) = 0$ and the x factor is zero at $x = 0$,* so Theorem 17.8.1 tells us that not only is λ real, it is also nonnegative. Thus, the argument of J_0 in (18) is real, and it suffices to look for roots of (18) on the real axis. If we denote the zeros of $J_0(x)$ as $x = z_1, z_2, \ldots$ (Fig. 3), then (18) gives $\sqrt{\lambda}\, L = z_n$, so the eigenvalues and eigenfunctions of (12) are

$$\lambda_n = \left(\frac{z_n}{L}\right)^2 \quad \text{and} \quad \phi_n(x) = J_0\left(z_n \frac{x}{L}\right) \tag{20}$$

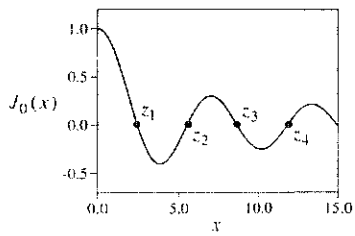


Figure 3. The zeros z_n of J_0.

*We can conclude that $x\phi_n\phi_n'$ vanishes at $x = 0$, provided that ϕ_n and ϕ_n' are finite there. We know that ϕ_n is finite there because that is our boundary condition, at $x = 0$, in (12b). Thus, we need to also ask that ϕ_n' be finite there. But by the time we reach (19) we already have the form (16) for the eigenfunctions, even if we don't yet know the λ's, and the derivative of the right-hand side of (16) is bounded at $x = 0$; indeed, it is even zero.

for $n = 1, 2, \ldots$. The zeros z_n are tabulated, and the first several are as follows:

$$z_1 = 2.405, \quad z_2 = 5.520, \quad z_3 = 8.654, \quad z_4 = 11.792, \quad z_5 = 14.931, \quad \ldots. \tag{21}$$

Further, the eigenfunction expansion of a given function f, on $0 < x < L$, is given by

$$f(x) = \sum_{n=1}^{\infty} a_n J_0\left(z_n \frac{x}{L}\right), \tag{22}$$

where, recalling that the weight function in the inner product is $w(x) = x$,

$$a_n = \frac{\langle f(x), J_0\left(z_n \frac{x}{L}\right)\rangle}{\langle J_0\left(z_n \frac{x}{L}\right), J_0\left(z_n \frac{x}{L}\right)\rangle} = \frac{\displaystyle\int_0^L f(x) J_0\left(z_n \frac{x}{L}\right) x\, dx}{\displaystyle\int_0^L \left[J_0\left(z_n \frac{x}{L}\right)\right]^2 x\, dx}. \tag{23}$$

The integral in the denominator of (23) is evaluated in Exercise 7 of Section 4.6, so the final expression for a_n is

$$a_n = \frac{2}{L^2[J_1(z_n)]^2} \int_0^L f(x) J_0\left(z_n \frac{x}{L}\right) x\, dx, \tag{24}$$

where J_0 and J_1 are Bessel functions of the first kind, of orders 0 and 1, respectively. ∎

EXAMPLE 3. *A Legendre Equation.* The Sturm–Liouville problem

$$(1 - x^2)y'' - 2xy' + \lambda y = 0 \qquad (-1 < x < 1) \tag{25a}$$

$$y(-1) \text{ bounded}. \quad y(1) \text{ bounded} \tag{25b}$$

is also singular because $p(x) = 1 - x^2$ vanishes at both endpoints. Hence, we apply the boundedness boundary conditions. In fact, (25a) is the Legendre equation, which is the subject of Section 4.4. There, we found that solutions of (25a) that are bounded on $-1 \le x \le 1$ are possible only if $\lambda = n(n+1)$, for $n = 0, 1, 2, \ldots$, and those nontrivial solutions are the Legendre polynomials $P_n(x)$. Thus, the eigenvalues and eigenfunctions of (25) are

$$\lambda_n = n(n+1), \qquad \phi_n(x) = P_n(x), \tag{26}$$

for $n = 0, 1, 2, \ldots$.

The eigenfunction expansion of a given function f, on $-1 \le x \le 1$, is given by

$$f(x) = \sum_{n=0}^{\infty} a_n P_n(x), \tag{27}$$

where, since the weight function is $w(x) = 1$,

$$a_n = \frac{\langle f(x), P_n(x)\rangle}{\langle P_n(x), P_n(x)\rangle} = \frac{\displaystyle\int_{-1}^{1} f(x) P_n(x)\, dx}{\displaystyle\int_{-1}^{1} P_n^2(x)\, dx}. \tag{28}$$

The integral in the denominator of (28) was found [(18) in Section 4.4] to be $2/(2n+1)$, so

$$a_n = \frac{2n+1}{2}\int_{-1}^{1} f(x)P_n(x)\,dx. \tag{29}$$

For instance, let f be the "ramp" $f(x) = xH(x)$, where H is the Heaviside function. Then (29) becomes

$$a_n = \frac{2n+1}{2}\int_{0}^{1} xP_n(x)\,dx, \tag{30}$$

so

$$a_0 = \frac{1}{2}\int_0^1 x\,dx = \frac{1}{4},$$
$$a_1 = \frac{3}{2}\int_0^1 x^2\,dx = \frac{1}{2},$$

and so on. Thus,

$$f(x) = \frac{1}{4}P_0(x) + \frac{1}{2}P_1(x) + \frac{5}{16}P_2(x) - \frac{3}{32}P_4(x) + \frac{13}{256}P_6(x) - \cdots. \tag{31}$$

For comparison, we have plotted both $f(x)$ and the partial sum $s_n(x)$ of the first n nonvanishing terms on the right-hand side of (31), for $n = 2$ and $n = 5$, in Fig. 4. ∎

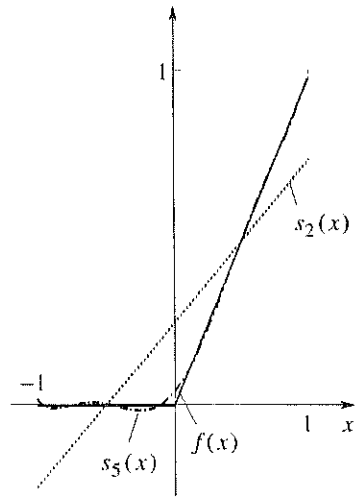


Figure 4. Convergence of (31) to $f(x)$.

Because of their close relationship with Fourier series, we call (22) and (31) **Fourier–Bessel** and **Fourier–Legendre** series, respectively. Such expansions will be needed in Chapters 18–20.

Closure. We have studied the Sturm–Liouville problem with periodic boundary conditions [specifically, the nonseparated conditions (2)], and the singular Sturm–Liouville problem [where $p(x) > 0$ and $w(x) > 0$ hold on $a < x < b$ rather than on $a \le x \le b$, and where a boundedness boundary condition is imposed at an endpoint at which $p(x)$ vanishes] because those cases are not covered by the theorems given in Section 17.7, and because they are important cases. Essentially, the upshot is that "all is well": we still obtain real eigenvalues and sets of orthogonal eigenfunctions that can be used to expand functions over the (a, b) interval.

Observe that for the interval $0 < x < 1$, say, each of the Sturm–Liouville problems

$$y'' + \lambda y = 0, \qquad y(0) = 0,\ y(1) = 0$$

and

$$(1 - x^2)y'' - 2xy' + \lambda y = 0, \qquad y'(0) = 0,\ y(1) = \text{bounded}$$

generates an orthogonal basis of eigenfunctions, and that one could write down an unlimited number of other Sturm–Liouville problems that generate orthogonal bases over the same interval. If we wish to expand a given function on that interval,

then how do we know which basis to use? As we shall see in Chapters 18–20, that decision will be based on the mathematical context.

EXERCISES 17.8

1. We derived the solution (14) of (12a) by introducing a change of variables $x = t/\sqrt{\lambda}$. Derive it using the method explained in Section 4.6.6, instead.

2. Find the eigenvalues and eigenfunctions, and work out the eigenfunction expansion of the given f. NOTE: As usual, $H(x)$ denotes the Heaviside step function. Use computer software, if you wish, to evaluate any needed integrals.

(a) $y'' + \lambda y = 0$, $y(0) = y(4)$, $y'(0) = y'(4)$, $f(x) = H(x-2)$
(b) $y'' + \lambda y = 0$, $y(-1) = y(5)$, $y'(-1) = y'(5)$, $f(x) = x + 2$
(c) $x^2y'' + xy' + \lambda y = 0$, $y(1) = y(2)$, $y'(1) = y'(2)$, $f(x) = 6$
(d) $(1 - x^2)y'' - 2xy' + \lambda y = 0$, $y(0) = 0$, $y(1)$ bounded, $f(x) = 4$; evaluate only the first three nonvanishing terms in the expansion of f
(e) $(1 - x^2)y'' - 2xy' + \lambda y = 0$, $y'(0) = 0$, $y(1)$ bounded, $f(x) = x$; evaluate only the first three nonvanishing terms in the expansion of f
(f) $(1 - x^2)y'' - 2xy' + \lambda y = 0$, $y(-1)$ bounded, $y'(0) = 0$, $f(x) = 5x^2$
(g) $(4 - x^2)y'' - 2xy' + \lambda y = 0$, $y(-2)$ bounded, $y(2)$ bounded, $f(x) = 5 - 2x$

3. Expand $f(x) = H(x)$, on $-1 < x < 1$, in terms of the eigenfunctions of the Sturm–Liouville problem

$$(1 - x^2)y'' - 2xy' + \lambda y = 0,$$

where $y(-1)$ and $y(1)$ are bounded. Plot both $f(x)$ and the sum of the first four nonvanishing terms of that expansion.

4. Expand $f(x) = 1 - x$ on $0 < x < 1$, in terms of the eigenfunctions of the Sturm–Liouville problem

$$(1 - x^2)y'' - 2xy' + \lambda y = 0,$$

where $y'(0) = 0$ and $y(1)$ is bounded. Plot both $f(x)$ and the sum of the first three nonvanishing terms of that expansion.

5. Determine the eigenvalues (or at least the characteristic equation for them), eigenfunctions, and weight function of the Sturm–Liouville problem

$$x^2y'' + xy' + (\lambda x^2 - 9)y = 0,$$

where $y(0)$ is bounded and $y(L) = 0$.

6. (*Chebyshev polynomials*) Consider the eigenvalue problem

$$\boxed{(1 - x^2)y'' - xy' + \lambda y = 0, \qquad (-1 < x < 1)} \tag{6.1}$$

where $y(-1)$, $y'(-1)$, $y(1)$, and $y'(1)$ are to be bounded; (6.1) is the **Chebyshev equation**, after the Russian mathematician *Pafnuti Chebyshev* (1821–1894), often transliterated as Tchebichef.

(a) Show that under the change of variables $x = \cos\theta$ the equation (6.1) becomes

$$\Theta'' + \lambda\Theta = 0, \qquad (0 < \theta < \pi) \tag{6.2}$$

where $\Theta(\theta) \equiv y(x(\theta)) = y(\cos\theta)$. Thus, the general solution, in terms of θ, is

$$\Theta(\theta) = \begin{cases} A\cos\sqrt{\lambda}\,\theta + B\sin\sqrt{\lambda}\,\theta, & \lambda \neq 0 \\ C + D\theta, & \lambda = 0. \end{cases} \tag{6.3}$$

(b) Surely, the solutions $\cos\sqrt{\lambda}\,\theta$, $\sin\sqrt{\lambda}\,\theta$, 1, and θ in (6.3) are bounded at $\theta = 0$ ($x = 1$) and $\theta = \pi$ ($x = -1$). However, show (by chain differentiation) that $y'(x)$ is bounded at $x = \pm 1$ only if $B = D = 0$ and $\sqrt{\lambda} = n = 1, 2, \ldots$. Thus, the eigenfunctions of (6.1) are, in terms of θ, $\cos n\theta$ ($n = 1, 2, \ldots$) and 1 or, equivalently, $\cos n\theta$ for $n = 0, 1, 2, \ldots$. In terms of the original x variable, the eigenvalues and eigenfunctions of (6.1) are

$$\boxed{\lambda_n = n^2, \quad T_n(x) = \cos(n\cos^{-1}x), \quad (n = 0, 1, 2, \ldots)} \tag{6.4}$$

the T in honor of Chebyshev.

(c) Though not obvious, it turns out that $T_n(x)$ is an nth-degree polynomial in x. Show that the first several are as follows:

$$\begin{aligned} T_0(x) &= 1, \\ T_1(x) &= x, \\ T_2(x) &= 2x^2 - 1, \\ T_3(x) &= 4x^3 - 3x, \\ T_4(x) &= 8x^4 - 8x^2 + 1, \\ T_5(x) &= 16x^5 - 20x^3 + 5x. \end{aligned} \tag{6.5}$$

HINT: Use the trigonometric identities

$$\begin{aligned} \cos 2\theta &= 2\cos^2\theta - 1, \\ \cos 3\theta &= 4\cos^3\theta - 3\cos\theta, \\ \cos 4\theta &= 8\cos^4\theta - 8\cos^2\theta + 1, \\ \cos 5\theta &= 16\cos^5\theta - 20\cos^3\theta + 5\cos\theta, \end{aligned}$$

and so on.

(d) Get (6.1) into the standard Sturm–Liouville form by multiplying through by a suitably chosen function $\sigma(x)$ (that is, nonzero on $-1 < x < 1$), and thus show that the weight function is

$$w(x) = \frac{1}{\sqrt{1-x^2}}. \tag{6.6}$$

(e) Theorem 17.8.1 guarantees that, with respect to the weight function (6.6), $\langle T_m(x), T_n(x)\rangle = 0$ for $m \neq n$. Nonetheless, prove that result directly, by evaluating the integral

$$\langle T_m, T_n\rangle = \int_{-1}^{1} T_m(x)T_n(x)\,\frac{1}{\sqrt{1-x^2}}\,dx.$$

Further, show by direct integration that for $m = n$ we have

$$\langle T_n, T_n\rangle = \int_{-1}^{1} T_n^2(x)\,\frac{1}{\sqrt{1-x^2}}\,dx = \begin{cases} \pi, & m = n = 0 \\ \dfrac{\pi}{2}, & m = n \neq 0. \end{cases} \tag{6.7}$$

(f) Thus, the eigenfunction expansion of a given function f, defined on $-1 < x < 1$, is

$$\boxed{f(x) = \sum_{n=0}^{\infty} a_n T_n(x), \qquad (-1 < x < 1)} \tag{6.8}$$

where

$$a_n = \frac{\langle f, T_n\rangle}{\langle T_n, T_n\rangle} = \begin{cases} \dfrac{1}{\pi}\displaystyle\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\,dx, & n = 0 \\ \dfrac{2}{\pi}\displaystyle\int_{-1}^{1} \frac{f(x)T_n(x)}{\sqrt{1-x^2}}\,dx, & n = 1, 2, \ldots. \end{cases} \tag{6.9}$$

Use (6.9) to evaluate the a_n's for the case where $f(x) = H(x)$.

(g) Plot $H(x)$ and (by computer) the partial sum of the series obtained in part (f), through $n = 5$.

(h) Derive the values

$$T_n(1) = 1, \quad T_n(-1) = (-1)^n, \tag{6.10}$$

and the recursion formula

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \tag{6.11}$$

(i) Use (6.11), and the T_n's given in (6.5), to derive $T_6(x)$ and $T_7(x)$.

(j) It can be shown that

$$\frac{1 - xt}{1 - 2xt + t^2} = \sum_{n=0}^{\infty} T_n(x)t^n, \qquad (-1 < t < 1) \tag{6.12}$$

so the left-hand member of (6.12) is called a **generating function** of the T_n's. By working out the Taylor series of the left-hand side, verify (6.12) through $n = 2$.

17.9 Fourier Integral

If a function f defined on $-\infty < x < \infty$ is periodic (and sufficiently well-behaved), then it can be represented by a Fourier series. We have begun to see, and will continue to see in Chapters 18–20, that Fourier series representation is of great importance. Sometimes we work with functions, defined on $-\infty < x < \infty$, that are not periodic, such as $f(x) = e^{-x^2}$, the graph of which is given in Fig. 1.

Evidently, we cannot expand such functions in Fourier series if they are not periodic. Yet, we can think of f as periodic but with an infinite period.

Thus, to extend the Fourier series concept to nonperiodic functions we will now consider the limiting case of the classical Fourier series *

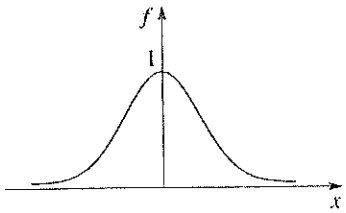


Figure 1. Graph of $f(x) = e^{-x^2}$.

$$f(x) = \sum_{n=0}^{\infty} \left(a_n \cos \frac{n\pi x}{\ell} + b_n \sin \frac{n\pi x}{\ell} \right) \tag{1a}$$

where

$$a_0 = \frac{1}{2\ell} \int_{-\ell}^{\ell} f(x)\, dx, \qquad a_n = \frac{1}{\ell} \int_{-\ell}^{\ell} f(x) \cos \frac{n\pi x}{\ell}\, dx,$$
$$b_n = \frac{1}{\ell} \int_{-\ell}^{\ell} f(x) \sin \frac{n\pi x}{\ell}\, dx, \tag{1b}$$

as $\ell \to \infty$. We cannot simply *set* $\ell = \infty$ in (1), as that would yield nonsense. Rather one needs to carry out a careful limit process, as ℓ *tends* to ∞.

First, note that the $n\pi/\ell$'s in (1a) are the *frequencies* – spatial or temporal, depending on whether x is a space variable or time. The set of all of the frequencies,

$$0, \frac{\pi}{\ell}, \frac{2\pi}{\ell}, \frac{3\pi}{\ell}, \ldots$$

is called the **frequency spectrum**. To see what happens to the frequency spectrum as ℓ increases, consider the cases where $\ell = \pi$, 2π, and 10π. The corresponding frequency spectra are as follows:

$$\begin{aligned} \ell = \pi: &\quad n\pi/\ell = 0, 1, 2, 3, 4, \ldots \\ \ell = 2\pi: &\quad n\pi/\ell = 0, 0.5, 1.0, 1.5, 2.0, \ldots \\ \ell = 10\pi: &\quad n\pi/\ell = 0, 0.1, 0.2, 0.3, 0.4, \ldots \end{aligned}$$

Observe that as ℓ increases the discrete spectrum becomes more and more dense, and approaches a *continuous spectrum* (from 0 to ∞) as $\ell \to \infty$. Therefore, we can expect that as $\ell \to \infty$ the summation in (1a), on the discrete variable n, will give way to an integration on a continuous variable, say ω. In fact, if f is sufficiently well-behaved (e.g., see Theorem 17.9.1 below) one can show that

$$\boxed{f(x) = \int_0^{\infty} [a(\omega) \cos \omega x + b(\omega) \sin \omega x]\, d\omega,} \tag{2a}$$

*Although we usually split out the a_0 term, here we include it in the sum, merely to increase the resemblance between (1a) and (2a), below.

where*

$$\boxed{\begin{aligned} a(\omega) &= \frac{1}{\pi}\int_{-\infty}^{\infty} f(x)\cos\omega x\,dx, \\ b(\omega) &= \frac{1}{\pi}\int_{-\infty}^{\infty} f(x)\sin\omega x\,dx. \end{aligned}} \tag{2b}$$

The right-hand side of (2a) is called the **Fourier integral** of f, which we denote as FI f, and (2a) is called the Fourier integral representation of f.

THEOREM 17.9.1 *Fourier Integral Theorem*
Let f be defined on $-\infty < x < \infty$, let f and f' be piecewise continuous on every finite interval $[-\ell, \ell]$ (i.e., for ℓ arbitrarily large), and let $\int_{-\infty}^{\infty} |f(x)|\,dx$ be convergent. Then the Fourier integral of f converges to $f(x)$ at every point x at which f is continuous, and to the mean value $[f(x+) + (f(x-)]/2$ at every point x at which f is discontinuous.

Proof: Rigorous proof of this theorem is well beyond our present scope, and the following is put forward only as a heuristic derivation. If we change the dummy integration variable to ξ in (1b), to distinguish it from the fixed point x in (1a), insert (1b) into (1a), and use the identity $\cos A \cos B + \sin A \sin B = \cos(A - B)$ for greater compactness, then the right-hand side of (1a), namely, FS f, becomes

$$\text{FS } f = \frac{1}{2\ell}\int_{-\ell}^{\ell} f(\xi)\,d\xi + \sum_{n=1}^{\infty}\frac{1}{\ell}\int_{-\ell}^{\ell} f(\xi)\cos\frac{n\pi}{\ell}(\xi - x)\,d\xi. \tag{3}$$

Denote the two terms on the right-hand side as I and J, respectively. For I we have

$$|I| = \left|\frac{1}{2\ell}\int_{-\ell}^{\ell} f(\xi)\,d\xi\right| \le \frac{1}{2\ell}\int_{-\infty}^{\infty} |f(\xi)|\,d\xi \to 0 \tag{4}$$

*Normally, a singular integral of the form $\int_{-\infty}^{\infty} F(x)\,dx$ is understood to mean

$$\int_{-\infty}^{\infty} F(x)\,dx = \lim_{A\to\infty, B\to\infty}\int_{-A}^{B} F(x)\,dx,$$

where A and B tend to infinity independently. However, the integrals in (2b) are to be understood in the more forgiving sense,

$$\int_{-\infty}^{\infty} F(x)\,dx = \lim_{A\to\infty}\int_{-A}^{A} F(x)\,dx.$$

To see that the latter is "more forgiving" than the former, observe that if $F(x) = x$, for instance, then with the former interpretation the integral is divergent but with the latter interpretation it converges to zero.

as $\ell \to \infty$ because $\int_{-\infty}^{\infty} |f(\xi)|\, d\xi$ is finite, by assumption, and $1/(2\ell) \to 0$ as $\ell \to \infty$. Thus, $I \to 0$ as $\ell \to \infty$. In J, let $\pi/\ell \equiv \Delta\omega$. Then

$$\begin{aligned} J &= \frac{1}{\pi} \sum_{n=1}^{\infty} \left[\int_{-\ell}^{\ell} f(\xi) \cos n\Delta\omega(\xi - x)\, d\xi \right] \Delta\omega \\ &\to \frac{1}{\pi} \int_0^{\infty} \left[\int_{-\infty}^{\infty} f(\xi) \cos \omega(\xi - x)\, d\xi \right] d\omega \end{aligned} \tag{5}$$

as $\ell \to \infty$. To understand the last step, observe that the sum in (5) is a Riemann sum, and as $\Delta\omega \to 0$ (i.e., as $\ell \to \infty$, because $\Delta\omega = \pi/\ell$) it yields a Riemann integral. That is, we partition the interval $0 < \omega < \infty$ into equal parts, of dimension $\Delta\omega = \pi/\ell$, and call $\omega_1 = \pi/\ell$, $\omega_2 = 2\pi/\ell$, $\omega_3 = 3\pi/\ell$, and so on, and use the general Riemann integral formula

$$\lim_{\Delta\omega \to 0} \sum_{n=1}^{\infty} F(n\Delta\omega)\Delta\omega = \lim_{\Delta\omega \to 0} \sum_{n=1}^{\infty} F(\omega_n)\Delta\omega = \int_0^{\infty} F(\omega)\, d\omega. \tag{6}$$

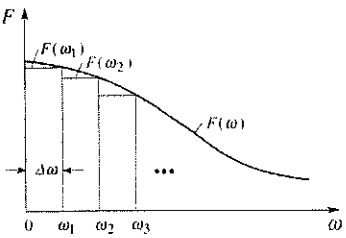


Figure 2. Riemann sum.

(See Fig. 2.) Finally, expressing $\cos\omega(\xi - x) = \cos\omega\xi \cos\omega x + \sin\omega\xi \sin\omega x$, we have

$$\begin{aligned} \text{FS}\, f &= I + J \\ &\to 0 + \int_0^{\infty} \left[\frac{1}{\pi} \int_{-\infty}^{\infty} f(\xi) \cos\omega\xi\, d\xi \right] \cos\omega x d\omega \\ &\quad + \int_0^{\infty} \left[\frac{1}{\pi} \int_{-\infty}^{\infty} f(\xi) \sin\omega\xi\, d\xi \right] \sin\omega x d\omega \\ &= \int_0^{\infty} [a(\omega) \cos\omega x + b(\omega) \sin\omega x]\, d\omega \end{aligned} \tag{7}$$

as $\ell \to \infty$. That is, as $\ell \to \infty$ the Fourier series tends to the Fourier integral. We reiterate that our approach has been heuristic, not rigorous. ■

EXAMPLE 1. *Rectangular Pulse.* Let f be the rectangular pulse shown in Fig. 3. This f does satisfy the conditions of the theorem. According to (2b),

$$a(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} f(x) \cos\omega x\, dx = \frac{1}{\pi} \int_{-1}^{1} \cos\omega x\, dx = \frac{2}{\pi} \frac{\sin\omega}{\omega} \tag{8a}$$

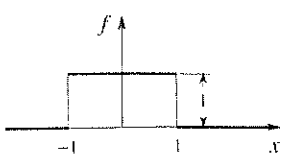


Figure 3. Rectangular pulse.

and

$$b(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} f(x) \sin\omega x\, dx = 0 \tag{8b}$$

because the integrand $f(x) \sin\omega x$ is an odd function (recall that even $\times$ odd $=$ odd). Thus, the Fourier integral representation of f is, from (2) and (8),

$$f(x) = \frac{2}{\pi} \int_0^{\infty} \frac{\sin\omega}{\omega} \cos\omega x\, d\omega. \tag{9}$$

Just as it was illuminating to plot the *partial sums* $s_n(x)$ of Fourier series, it should be illuminating to see how the *partial integral*

$$f_\Omega(x) = \frac{2}{\pi}\int_0^\Omega \frac{\sin\omega}{\omega}\cos\omega x\,d\omega \tag{10}$$

converges to $f(x)$ as $\Omega \to \infty$. Though the latter integral is not elementary, it can be evaluated in terms of the **sine integral** function

$$\boxed{Si(x) = \int_0^x \frac{\sin t}{t}\,dt,} \tag{11}$$

either analytically (Exercise 1) or using computer software. Using *Maple*, for instance, the command

$$\text{int}((2/\text{Pi}) * (1/y) * \sin(y) * \cos(y * x),\ y = 0..z);$$

(where z is used in place of Ω) gives

$$f_\Omega(x) = \frac{1}{\pi}\left\{Si[\Omega(x+1)] - Si[\Omega(x-1)]\right\}, \tag{12}$$

which we've plotted in Fig. 4 for $\Omega = 4, 16$, and 128. Evidently $f_\Omega(x)$ does converge to the pulse $f(x)$, but the convergence is slow near the jump discontinuities; the limiting case as $\Omega \to \infty$ is left for Exercise 3. As in the case of Fourier series, the Fourier integral exhibits the Gibbs phenomenon at such discontinuities.

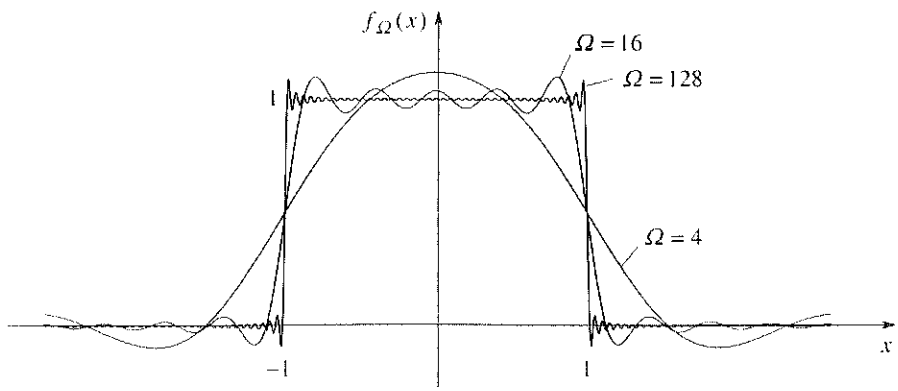


Figure 4. Convergence of $f_\Omega(x)$ to $f(x)$ as $\Omega \to \infty$.

It is also interesting to plot the Fourier coefficient $a(\omega)$, given by (8a), since $a(\omega)$ tells us the harmonic content of f, that is, the amplitude or "amount" of each $\cos\omega x$ harmonic present (Fig. 5). ∎

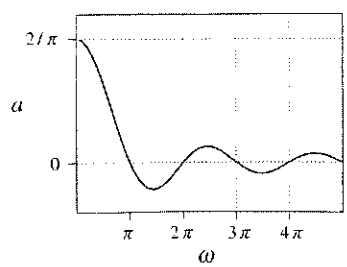


Figure 5. Harmonic content, $a(\omega)$.

EXAMPLE 2. *Infinite Beam on Elastic Foundation.* As a physical application of the Fourier integral let us consider the same problem as contained in Example 4 of Section 17.3 (which we urge you to review), but instead of the periodic loading suppose we have the nonperiodic rectangular pulse loading shown here in Fig. 6. Recall that the beam's

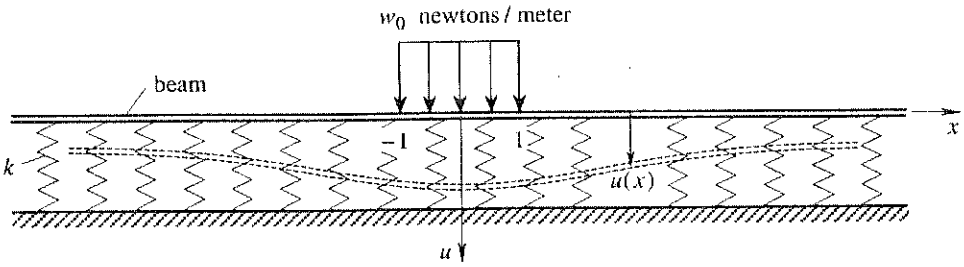


Figure 6. Infinite beam on elastic foundation.

deflection $u(x)$ is governed by the differential equation

$$EIu'''' + ku = w(x), \tag{13}$$

where E, I, k are physical constants, and

$$w(x) = \begin{cases} w_0, & |x| < 1 \\ 0, & |x| > 1 \end{cases} \tag{14}$$

is the rectangular pulse applied load distribution. Proceeding essentially as in Example 4 of Section 17.3, we first express $w(x)$ in Fourier integral form. Since $w(x)$ is merely $f(x)$ in Example 1, scaled by w_0, we conclude from (9) that

$$w(x) = \frac{2w_0}{\pi} \int_0^\infty \frac{\sin\omega}{\omega} \cos\omega x \, d\omega. \tag{15}$$

Next, we seek $u(x)$ in the Fourier integral form

$$u(x) = \int_0^\infty a(\omega) \cos\omega x \, d\omega, \tag{16}$$

where $a(\omega)$ remains to be determined. We have omitted the $b(\omega)\sin\omega x$ since $u(x)$ will evidently be a symmetric (even) function of x; that is, if we did include that term we would find that $b(\omega) = 0$.

Formally differentiating (16) under the integral sign four times, and putting that result and (15) and (16) into (13) gives

$$\int_0^\infty (EI\omega^4 + k)a(\omega) \cos\omega x \, d\omega = \frac{2w_0}{\pi} \int_0^\infty \frac{\sin\omega}{\omega} \cos\omega x \, d\omega. \tag{17}$$

Then, formally equating the coefficients of each cosine harmonic gives

$$(EI\omega^4 + k)a(\omega) = \frac{2w_0}{\pi} \frac{\sin\omega}{\omega} \tag{18}$$

or

$$a(\omega) = \frac{2w_0}{\pi} \frac{\sin\omega}{\omega} \frac{1}{(EI\omega^4 + k)}, \tag{19}$$

so that (16) becomes

$$u(x) = \frac{2w_0}{\pi} \int_0^\infty \frac{\sin\omega}{\omega} \frac{\cos\omega x}{(EI\omega^4 + k)} \, d\omega. \tag{20}$$

This integral can be evaluated analytically using complex variable techniques (the residue theorem), but for our present purposes it will suffice to let (20) stand as it is. ∎

Closure. We obtain the Fourier integral representation of nonperiodic functions defined on $-\infty < x < \infty$ by taking the limit of the Fourier series formula as the period tends to infinity. We limit discussion to two examples because we plan to use the Fourier integral only as a stepping stone to the *Fourier transform*. The latter is more highly developed as a methodology, like the Laplace transform, and is the subject of the next section.

EXERCISES 17.9

1. Use (11) to derive (12) from (10).

2. Derive the Fourier integral representations of the following functions. At which points, if any, does the Fourier integral fail to converge to $f(x)$? To what value does the integral converge at those points?

(a) $f(x) = \begin{cases} 100, & 0 \le x \le 2 \\ 0, & x < 0,\ x > 2 \end{cases}$

(b) $f(x) = \begin{cases} x, & 0 \le x < L \\ 0, & x < 0,\ x \ge L \end{cases}$

(c) $f(x) = \begin{cases} x, & |x| \le L \\ 0, & |x| > L \end{cases}$

(d) $f(x) = \begin{cases} -x, & -5 < x \le 0 \\ 0, & x \le -5,\ x > 0 \end{cases}$

(e) $f(x) = \begin{cases} |x|, & -1 < x < 2 \\ 0, & x \le -1,\ x \ge 2 \end{cases}$

(f) $f(x) = \begin{cases} 10, & 0 \le x \le 3 \\ 5, & 6 \le x \le 9 \\ 0, & x < 0,\ 3 < x < 6,\ x > 9 \end{cases}$

(g) $f(x) = \begin{cases} e^{-x}, & x \ge 0 \\ 0, & x < 0 \end{cases}$

(h) $f(x) = \begin{cases} e^{x}, & x \ge 0 \\ 0, & x < 0 \end{cases}$

(i) $f(x) = e^{-|x|}$

3. (a) Show that $Si(x)$ is an odd function of x.
(b) Using the known integral

$$Si(\infty) = \int_0^\infty \frac{\sin t}{t}\,dt = \frac{\pi}{2}$$

and recalling (11), show that

$$\lim_{\Omega\to\infty} f_\Omega(x) = \begin{cases} 1, & |x| < 1 \\ 1/2, & |x| = 1 \\ 0, & |x| > 1 \end{cases}$$

so that the Fourier integral (9) does converge to the rectangular pulse and (in accordance with Theorem 17.9.1) to the average values at the two jump discontinuities.

4. Comparing (2) with the classical Fourier series (1), it might appear that $a(0)$ is analogous to the a_0 term in the Fourier series and represents the average value of f. Is it true that $a(0)$ is the average value of f where, by the average value of f we mean $\lim_{A\to\infty} \frac{1}{2A}\int_{-A}^{A} f(x)\,dx$? Explain.

17.10 Fourier Transform

Our purpose in this section is to recast the Fourier integral representation of a function f as a pair of formulas, the first giving the so-called Fourier transform of f, and the second giving the inverse of that transform. Once the transform and its inverse

are derived, in Section 17.10.1, discussion will closely parallel our discussion, in Chapter 5, of the Laplace transform.

17.10.1. Transition from Fourier integral to Fourier transform. Our starting point is the Fourier integral formula

$$f(x) = \int_0^\infty [a(\omega)\cos\omega x + b(\omega)\sin\omega x]\,d\omega, \tag{1a}$$

where

$$\begin{aligned} a(\omega) &= \frac{1}{\pi}\int_{-\infty}^{\infty} f(x)\cos\omega x\,dx, \\ b(\omega) &= \frac{1}{\pi}\int_{-\infty}^{\infty} f(x)\sin\omega x\,dx. \end{aligned} \tag{1b}$$

Just as one can express a Fourier series in complex exponential form (Section 17.3.4), one can express the Fourier integral (1) in complex exponential form. To obtain that form put (1b) into (1a). [First we change the dummy integration variable x in (1b) to ξ, say, to avoid confusing that variable with the x's occurring in (1a), which denote the fixed point at which $f(x)$ is being computed.] Thus,

$$\begin{aligned} f(x) &= \frac{1}{\pi}\int_0^\infty \left\{\int_{-\infty}^{\infty} f(\xi)[\cos\omega\xi\cos\omega x + \sin\omega\xi\sin\omega x]\,d\xi\right\} d\omega \\ &= \frac{1}{\pi}\int_0^\infty \int_{-\infty}^{\infty} f(\xi)\cos\omega(\xi - x)\,d\xi\,d\omega, \end{aligned} \tag{2}$$

since $\cos(A - B) = \cos A\cos B + \sin A\sin B$. To introduce complex exponentials, re-express (2) as

$$\begin{aligned} f(x) &= \frac{1}{\pi}\int_0^\infty \int_{-\infty}^{\infty} f(\xi)\frac{e^{i\omega(\xi-x)} + e^{-i\omega(\xi-x)}}{2}\,d\xi\,d\omega \\ &= \frac{1}{2\pi}\int_0^\infty \int_{-\infty}^{\infty} f(\xi)e^{i\omega(\xi-x)}\,d\xi\,d\omega + \frac{1}{2\pi}\int_0^\infty \int_{-\infty}^{\infty} f(\xi)e^{-i\omega(\xi-x)}\,d\xi\,d\omega. \end{aligned}$$

To combine the two terms on the right-hand side, let us change the dummy integration variable from ω to $-\omega$ in the first. Thus,

$$\begin{aligned} f(x) &= \frac{1}{2\pi}\int_0^{-\infty} \int_{-\infty}^{\infty} f(\xi)e^{-i\omega(\xi-x)}\,d\xi\,(-d\omega) \\ &\qquad + \frac{1}{2\pi}\int_0^\infty \int_{-\infty}^{\infty} f(\xi)e^{-i\omega(\xi-x)}\,d\xi\,d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{0} \int_{-\infty}^{\infty} f(\xi)e^{-i\omega(\xi-x)}\,d\xi\,d\omega + \frac{1}{2\pi}\int_0^\infty \int_{-\infty}^{\infty} f(\xi)e^{-i\omega(\xi-x)}\,d\xi\,d\omega, \end{aligned}$$

so

$$\begin{aligned} f(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\xi)e^{-i\omega(\xi-x)}\,d\xi\,d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} f(\xi)e^{-i\omega\xi}\,d\xi\right] e^{i\omega x}\,d\omega. \end{aligned} \tag{3}$$

The latter can be split apart as

$$f(x) = a \int_{-\infty}^{\infty} c(\omega) e^{i\omega x}\, d\omega \tag{4a}$$

and

$$c(\omega) = b \int_{-\infty}^{\infty} f(\xi) e^{-i\omega\xi}\, d\xi, \tag{4b}$$

if the constants a, b are such that $ab = 1/2\pi$. We can make (4) resemble (1) more closely by choosing $a = 1$ and $b = 1/2\pi$, but we will choose $a = 1/2\pi$ and $b = 1$.* There is no longer a need to distinguish x and ξ, because the x's are confined to (4a) and the ξ's to (4b). Thus, to minimize nomenclature and to mimic the form of (1), we write

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} c(\omega) e^{i\omega x}\, d\omega, \tag{5a}$$

$$c(\omega) = \int_{-\infty}^{\infty} f(x) e^{-i\omega x}\, dx. \tag{5b}$$

Rather than thinking of (5a) as the Fourier integral of f and (5b) as giving (give or take the factor of $1/2\pi$) the Fourier coefficients $c(\omega)$, we can think of (5a,b) as a transform pair: (5b) defines the **Fourier transform** $c(\omega)$ of the given function $f(x)$, and (5a) is called the **inversion formula** because putting $c(\omega)$ in and integrating gives us back $f(x)$. It is standard to use the notation $\hat{f}(\omega)$, in place of $c(\omega)$, for the transform, so we rewrite (5) in final form as

$$\boxed{F\{f(x)\} = \hat{f}(\omega) = \int_{-\infty}^{\infty} f(x) e^{-i\omega x}\, dx,} \tag{6a}$$

and

$$\boxed{F^{-1}\{\hat{f}(\omega)\} = f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\omega) e^{i\omega x}\, d\omega.} \tag{6b}$$

Thus, the Fourier transform and inversion formulas are not mysterious; together, they simply amount to the Fourier integral representation, expressed in complex exponential form, and conditions imposed on f are the same as in Theorem 17.9.1.

Let us illustrate the calculation of the transform $\hat{f}(\omega)$ of $f(x)$.

EXAMPLE 1. *Rectangular Pulse.* Consider the rectangular pulse $f(x) = H(x+1) - H(x-1)$, where H denotes the Heaviside function. The graph of f is given in Fig. 1. Using (6a), the Fourier transform of f is

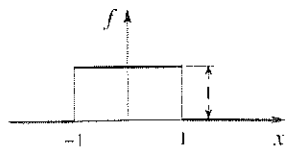


Figure 1. Rectangular pulse.

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} [H(x+1) - H(x-1)] e^{-i\omega x}\, dx = \int_{-1}^{1} e^{-i\omega x}\, dx$$
$$= \left. \frac{e^{-i\omega x}}{-i\omega} \right|_{-1}^{1} = 2\frac{\sin\omega}{\omega}. \tag{7}$$

*Some authors, perhaps out of a greater sense of fair play, choose $a = b = 1/\sqrt{2\pi}$.

COMMENT. We could illustrate the inversion formula as well, by putting $\hat{f}(\omega) = (2\sin\omega)/\omega$ into the integrand of (6b), integrating, and showing that the result is the rectangular pulse f that we started with. In fact, that integral can be evaluated by using the residue theorem of the complex integral calculus, but we won't study that theorem until Chapter 24. ∎

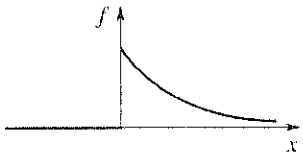


Figure 2. $f(x) = H(x)e^{-ax}$.

EXAMPLE 2. Evaluate the Fourier transform of

$$f(x) = H(x)e^{-ax}, \qquad (a > 0) \tag{8}$$

the graph of which is given in Fig. 2. From (6a),

$$\begin{aligned}\hat{f}(\omega) &= \int_{-\infty}^{\infty} H(x)e^{-ax}e^{-i\omega x}\,dx = \int_0^{\infty} e^{-(a+i\omega)x}\,dx \\ &= -\frac{e^{-(a+i\omega)x}}{a+i\omega}\bigg|_0^{\infty} = \frac{1}{a+i\omega} - \frac{1}{a+i\omega}\lim_{x\to\infty} e^{-(a+i\omega)x} \\ &= \frac{1}{a+i\omega}.\end{aligned} \tag{9}$$

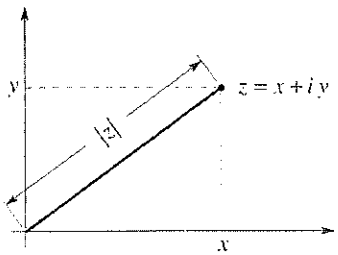


Figure 3. Modulus of z.

To explain the last step in (9), recall that the magnitude (or modulus) $|z|$ of a complex number $z = x + iy$ (Fig. 3) is $|z| = \sqrt{x^2 + y^2}$, and that the modulus of a product is the product of the moduli ($|z_1 z_2| = |z_1|\,|z_2|$). Then

$$\begin{aligned}\left|e^{-(a+i\omega)x}\right| &= \left|e^{-ax}e^{-i\omega x}\right| = \left|e^{-ax}\right|\left|e^{-i\omega x}\right| \\ &= e^{-ax}\left|\cos\omega x - i\sin\omega x\right| \\ &= e^{-ax}\sqrt{\cos^2\omega x + \sin^2\omega x} \\ &= e^{-ax} \to 0\end{aligned} \tag{10}$$

as $x \to \infty$. ∎

As with the Laplace transform, it is convenient to use tables for both the transform and its inverse, insofar as possible. The table provided here, in Appendix D, is brief, but will suffice for present purposes. Much more extensive tables are available,* as well as powerful computer software; the relevant commands, using *Maple*, are given at the end of this section.

17.10.2. Properties and applications. The Fourier transform admits a number of useful properties, our discussion of which will closely parallel our analogous discussion for the Laplace transform, given in Chapter 5. We will assume, without reiteration, that the functions being transformed satisfy the conditions given in Theorem 17.9.1.

*See, for example, A. Erdélyi (ed.), *Tables of Integral transforms*, Vols. 1 and 2 (New York: McGraw-Hill, 1954). Volume 1 contains Fourier, Laplace, and Mellin transforms, and Volume 2 contains Hankel and various other transforms.

Linearity of the transform and its inverse. For any scalars α and β, and any functions f and g,

$$\boxed{F\{\alpha f + \beta g\} = \alpha F\{f\} + \beta F\{g\}} \tag{11}$$

and

$$\boxed{F^{-1}\{\alpha \hat{f} + \beta \hat{g}\} = \alpha F^{-1}\{\hat{f}\} + \beta F^{-1}\{\hat{g}\}.} \tag{12}$$

Of course, $F^{-1}\{\hat{f}\}$ is f and $F^{-1}\{\hat{g}\}$ is g. Proofs of (11) and (12) follow the same lines as the proofs of Theorems 5.3.1 and 5.3.2, respectively.

Transform of nth derivative. If $f(x), f'(x), \ldots, f^{(n-1)}(x)$ all tend to zero as $x \to \pm\infty$, and $\int_{-\infty}^{\infty} |f^{(j)}(x)|\, dx$ converges for each $j = 0, 1, \ldots, n$, then

$$\boxed{F\{f^{(n)}(x)\} = (i\omega)^n \hat{f}(\omega). \qquad (n = 0, 1, 2, \ldots)} \tag{13}$$

Proof: Let us prove (13) by induction. First, observe that (13) holds for $n = 0$, by definition. To complete the proof by induction we need to establish that if it holds for $n = k$ then it also holds for $n = k + 1$. To do so, integrate by parts:

$$\begin{aligned} F\{f^{(k+1)}(x)\} &= \int_{-\infty}^{\infty} f^{(k+1)}(x) e^{-i\omega x}\, dx \\ &= [f^{(k)}(x) e^{-i\omega x}]\Big|_{-\infty}^{\infty} + i\omega \int_{-\infty}^{\infty} f^{(k)}(x) e^{-i\omega x}\, dx \\ &= [f^{(k)}(x) e^{-i\omega x}]\Big|_{-\infty}^{\infty} + i\omega[(i\omega)^k \hat{f}(\omega)] \\ &= [f^{(k)}(x) e^{-i\omega x}]\Big|_{-\infty}^{\infty} + (i\omega)^{k+1} \hat{f}(\omega), \end{aligned} \tag{14}$$

the third equality following from the assumption that (13) holds for $n = k$. Now, $|f^{(k)}(x) e^{-i\omega x}| = |f^{(k)}(x)|\, |e^{-i\omega x}| = |f^{(k)}(x)|$, and since $f^{(k)}(x) \to 0$ as $x \to \pm\infty$, by assumption, the boundary term in (14) drops out. Thus $F\{f^{(k+1)}(x)\} = (i\omega)^{k+1} \hat{f}(\omega)$ so (13) does hold for $n = k + 1$, which result completes our proof by induction. ■

Fourier convolution. We denote the **Fourier convolution** of functions f and g as $f * g$. It too is a function of x, defined as

$$\boxed{(f * g)(x) \equiv \int_{-\infty}^{\infty} f(x - \xi) g(\xi)\, d\xi.} \tag{15}$$

Then the **Fourier convolution theorem** states that

$$\boxed{F\{f * g\} = \hat{f}(\omega) \hat{g}(\omega),} \tag{16}$$

or, equivalently, that

$$\boxed{F^{-1}\{\hat{f}\hat{g}\} = f * g.} \tag{17}$$

Proof: It suffices to prove (16) or (17), because of their equivalence. Let us prove (16). We have

$$\begin{aligned} F\{f * g\} &= \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} f(x-\xi) g(\xi)\, d\xi \right\} e^{-i\omega x}\, dx \\ &= \int_{-\infty}^{\infty} g(\xi)\, d\xi \int_{-\infty}^{\infty} f(x-\xi) e^{-i\omega x}\, dx \\ &= \int_{-\infty}^{\infty} g(\xi)\, d\xi \int_{-\infty}^{\infty} f(\eta) e^{-i\omega(\xi+\eta)}\, d\eta \\ &= \left(\int_{-\infty}^{\infty} g(\xi) e^{-i\omega\xi}\, d\xi \right) \left(\int_{-\infty}^{\infty} f(\eta) e^{-i\omega\eta}\, d\eta \right) \\ &= \hat{g}(\omega)\hat{f}(\omega), \end{aligned} \tag{18}$$

as was to be shown. In the second equality we reversed the order of integration,* and the third equality follows from the change of variables $x - \xi = \eta$. ■

It is easy to show that the Fourier convolution is commutative,

$$f * g = g * f, \tag{19}$$

proof of which is left for the exercises. That is, it doesn't matter whether we take the argument of f to be $x - \xi$ and the argument of g to be ξ, in (15), or visa versa.

Translation formulas, x-shift and ω-shift. It is also easy to derive the shift (or translation) formulas

$$\boxed{F\{f(x-a)\} = e^{-ia\omega}\hat{f}(\omega)} \tag{20}$$

and

$$\boxed{F^{-1}\{\hat{f}(\omega-a)\} = e^{iax} f(x),} \tag{21}$$

proofs of which are left for the exercises.

Each of the properties (11), (12), (13), (16), (17), (20), and (21) corresponds to an analogous property of the Laplace transform. Properties (11) and (12) are identical to the corresponding properties of the Laplace transform. The derivative property (13) is similar to the property

$$L\{f^{(n)}(t)\} = s^n \overline{f}(s) - s^{n-1} f(0) - s^{n-2} f'(0) - \cdots - f^{(n-1)}(0), \tag{22}$$

*Sufficient conditions for the validity of this reversal are that f and g both be **absolutely integrable**, i.e., that $\int_{-\infty}^{\infty} |f(x)|\, dx$ and $\int_{-\infty}^{\infty} |g(x)|\, dx$ both converge. But these conditions have already been assumed. For detailed discussion, see T. M. Apostol, *Mathematical Analysis* (Reading, MA: Addison-Wesley, 1957), p.491.

but does not include any boundary values analogous to the initial conditions $f(0), \ldots, f^{(n-1)}(0)$ in (22), because the boundaries are at $x = \pm\infty$, and it is assumed that $f(x), f'(x), \ldots, f^{(n-1)}(x)$ all tend to zero as $x \to \pm\infty$. The convolution properties (16) and (17) are identical to the corresponding ones for the Laplace transform, but keep in mind that the Fourier and Laplace convolutions are not quite the same; in contrast with the Fourier convolution (15), the Laplace convolution was defined as

$$(f * g)(t) \equiv \int_0^t f(t-\tau)g(\tau)\,d\tau. \tag{23}$$

That is, the integration limits are different.

In the examples to follow, we illustrate the use of these various properties and the table in Appendix D.

EXAMPLE 3. Given

$$f(x) = 4e^{-|x|} - 5e^{-3|x+2|}, \tag{24}$$

evaluate its transform $\hat{f}(\omega)$. First, the linearity property (11) gives

$$F\{4e^{-|x|} - 5e^{-3|x+2|}\} = 4F\{e^{-|x|}\} - 5F\{e^{-3|x+2|}\}. \tag{25}$$

Next, entry 4 in Appendix D gives

$$F\{e^{-|x|}\} = \frac{2}{\omega^2+1}, \tag{26}$$

and entry 11 (with $a = 3$ and $b = 6$) gives

$$\begin{aligned} F\{e^{-3|x+2|}\} &= F\{e^{-|3x+6|}\} \\ &= \frac{1}{3}\,e^{i6\omega/3}\left(F\{e^{-|x|}\}\right)\Big|_{\omega\to\omega/3} \\ &= \frac{1}{3}\,e^{i2\omega}\left(\frac{2}{\omega^2+1}\right)\Bigg|_{\omega\to\omega/3} \\ &= \frac{1}{3}\,e^{i2\omega}\frac{2}{\left(\frac{\omega}{3}\right)^2+1}. \end{aligned} \tag{27}$$

From (25)–(27), it follows that

$$\begin{aligned} \hat{f}(\omega) &= 4\frac{2}{\omega^2+1} - 5\left(\frac{2}{3}\right)\frac{e^{i2\omega}}{\left(\frac{\omega}{3}\right)^2+1} \\ &= \frac{8}{\omega^2+1} - \frac{30e^{i2\omega}}{\omega^2+9}. \quad \blacksquare \end{aligned}$$

EXAMPLE 4. Given

$$f(x) = xe^{-4x^2}, \tag{28}$$

evaluate $\hat{f}(\omega)$. First,

$$\begin{aligned} F\{e^{-4x^2}\} &= F\{e^{-(2x)^2}\} \\ &= \frac{1}{2}\left(F\{e^{-x^2}\}\right)\Big|_{\omega\to\omega/2} \qquad \text{(entry 11, } a=2, b=0) \\ &= \frac{1}{2}\sqrt{\pi}\, e^{-\omega^2/4}\Big|_{\omega\to\omega/2} \qquad \text{(entry 5)} \\ &= \frac{\sqrt{\pi}}{2}\, e^{-\omega^2/16}. \end{aligned} \tag{29}$$

Next, entry ~~16~~ 17 (with $n = 1$) gives

$$\hat{f}(\omega) = i\,\frac{d}{d\omega}\left(\frac{\sqrt{\pi}}{2}\, e^{-\omega^2/16}\right) = -i\,\frac{\sqrt{\pi}}{16}\,\omega e^{-\omega^2/16}. \quad \blacksquare$$

EXAMPLE 5. Given $\hat{f}(\omega) = e^{-2|\omega|}$, evaluate $f(x)$. With $a = 2$, entry 1 gives

$$F^{-1}\left\{\frac{\pi}{2}\, e^{-2|\omega|}\right\} = \frac{1}{x^2+4}. \tag{30}$$

Then it follows from the linearity property (12), with $\alpha = \pi/2$ and $\beta = 0$, that (30) can be re-expressed as

$$\frac{\pi}{2}F^{-1}\left\{e^{-2|\omega|}\right\} = \frac{1}{x^2+4}, \tag{31}$$

so

$$F^{-1}\left\{e^{-2|\omega|}\right\} = \frac{2}{\pi}\,\frac{1}{x^2+4}, \tag{32}$$

is the desired inverse function $f(x)$. $\blacksquare$

EXAMPLE 6. Evaluate

$$F^{-1}\left\{\frac{1}{\omega^2+4\omega+13}\right\}. \tag{33}$$

First, completing the square, write $\omega^2 + 4\omega = (\omega^2 + 4\omega + 4) - 4 = (\omega+2)^2 - 4$, so (33) becomes

$$F^{-1}\left\{\frac{1}{(\omega+2)^2+9}\right\}. \tag{34}$$

Now, entry 4 (with $a = 3$) and linearity give

$$F^{-1}\left\{\frac{1}{\omega^2+9}\right\} = \frac{1}{6}\, e^{-3|x|}. \tag{35}$$

Finally, from entry 12, with $a = 1$ and $b = 2$, and (35), we have

$$F^{-1}\left\{\frac{1}{(\omega+2)^2+9}\right\} = e^{-i2x}\left(\frac{1}{6}\, e^{-3|x|}\right) \tag{36}$$

as the desired result.

COMMENT. Alternatively, we could have used partial fractions, as follows. Solving $\omega^2 + 4\omega + 13 = 0$ gives $\omega = -2 \pm 3i$. Thus,

$$\begin{aligned}\frac{1}{\omega^2+4\omega+13} &= \frac{1}{(\omega+2-3i)(\omega+2+3i)}\\ &= \frac{1}{6i}\frac{1}{\omega+2-3i} - \frac{1}{6i}\frac{1}{\omega+2+3i}\\ &= \frac{1}{6}\frac{1}{i\omega+(3+2i)} - \frac{1}{6}\frac{1}{i\omega+(-3+2i)}\\ &= \frac{1}{6}\frac{1}{i\omega+(3+2i)} + \frac{1}{6}\frac{1}{-i\omega+(3-2i)}.\end{aligned} \tag{37}$$

Observe that the first term after the third equal sign can be inverted by entry 2, because $\mathrm{Re}\,(3+2i) = 3 > 0$,* but the second cannot, because $\mathrm{Re}\,(-3+2i) = -3 < 0$. Thus, in the final step we rearranged that term so that it can be inverted by entry 3, because $\mathrm{Re}\,(3-2i) = 3 > 0$. Finally, using the linearity property (12), together with entries 2 and 3 gives

$$\begin{aligned}F^{-1}\left\{\frac{1}{\omega^2+4\omega+13}\right\} &= \frac{1}{6}H(x)e^{-(3+2i)x} + \frac{1}{6}H(-x)e^{(3-2i)x}\\ &= \frac{1}{6}e^{-i2x}[H(x)e^{-3x} + H(-x)e^{3x}]\\ &= \frac{1}{6}e^{-i2x}e^{-3|x|},\end{aligned} \tag{38}$$

as obtained in (36). ∎

EXAMPLE 7. Evaluate

$$F^{-1}\left\{\frac{1}{\omega^2+1}\right\}. \tag{39}$$

The latter can be evaluated directly from entry 4 (together with the linearity property), as $e^{-|x|}/2$. However, for pedagogical purposes let us use this example to illustrate the convolution property. First, factor $1/(\omega^2+1)$ as

$$\frac{1}{\omega^2+1} = \frac{1}{(\omega+i)(\omega-i)} = \frac{1}{(1-i\omega)(1+i\omega)}, \tag{40}$$

the final form in (40) being more convenient than the intermediate form because each factor can be inverted, using entries 3 and 2, respectively:

$$F^{-1}\left\{\frac{1}{1-i\omega}\right\} = H(-x)e^{x}, \tag{41a}$$

$$F^{-1}\left\{\frac{1}{1+i\omega}\right\} = H(x)e^{-x}. \tag{41b}$$

*If $z = x + iy$ is a complex number, then x and y are called the **real** and **imaginary parts** of z, and are denoted as $\mathrm{Re}\,z$ and $\mathrm{Im}\,z$, respectively.

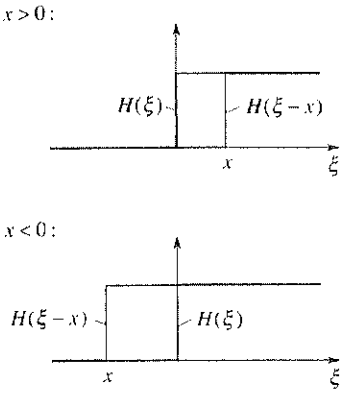


Figure 4. $H(\xi - x)$ and $H(\xi)$.

Then the Fourier convolution formula (17) gives

$$
\begin{aligned}
F^{-1}\left\{\frac{1}{\omega^2+1}\right\} &= [H(-x)e^{x}] * [H(x)e^{-x}] \\
&= \int_{-\infty}^{\infty} H(-x+\xi)e^{x-\xi}H(\xi)e^{-\xi}\,d\xi \\
&= \begin{cases} e^{x}\displaystyle\int_{x}^{\infty} e^{-2\xi}\,d\xi, & x>0 \\ e^{x}\displaystyle\int_{0}^{\infty} e^{-2\xi}\,d\xi, & x<0 \end{cases} \\
&= \begin{cases} \dfrac{1}{2}e^{-x} & x>0 \\ \dfrac{1}{2}e^{x} & x<0 \end{cases} \\
&= \frac{1}{2}e^{-|x|}.
\end{aligned}
\tag{42}
$$

To understand the third equality in (42), it is useful to plot $H(\xi - x)$ and $H(\xi)$, both for $x > 0$ and for $x < 0$, as we have in Fig. 4. Namely, we see from Fig. 4 that their product $H(\xi - x)H(\xi)$ is $H(\xi - x)$ if $x > 0$, and $H(\xi)$ if $x < 0$.

COMMENT. Alternatively, we could have applied partial fractions to the right-hand side of (40) and proceeded as in the comment in Example 6. ∎

EXAMPLE 8. *Infinite Beam on Elastic Foundation, Revisited.* In Example 2 in Section 17.9 we investigated the deflection $u(x)$ of an infinitely long beam resting on an elastic foundation and subjected to a load $w(x)$ newtons per meter (Fig. 5), governed by the differential equation

$$EIu'''' + ku = w(x), \tag{43}$$

where E, I, and k are physical constants. There we used the Fourier integral method, whereby $w(x)$ was expanded in a Fourier integral, and $u(x)$ was sought in the form of a Fourier integral. Here, we use the Fourier transform instead.

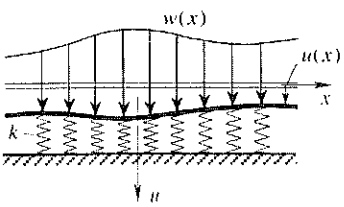


Figure 5. Beam on elastic foundation.

Taking the Fourier transform of (43) (i.e., multiplying each term by $e^{-i\omega x}\,dx$ and integrating from $x = -\infty$ to $x = +\infty$), we have

$$F\{EIu'''' + ku\} = F\{w\}. \tag{44}$$

It follows that

$$EI\,F\{u''''\} + k\,F\{u\} = F\{w\} \tag{45}$$

by the linearity property (11), and

$$EI(i\omega)^4\hat{u} + k\hat{u} = \hat{w} \tag{46}$$

by (13), assuming that u, u', u'', u''' all tend to zero as $x \to \pm\infty$ and that $\int_{-\infty}^{\infty} |u^{(j)}(x)|\,dx$ converges for $j = 0, 1, 2, 3, 4$. Now solving (46) for $\hat{u}$,

$$\hat{u} = \frac{\hat{w}}{EI\omega^4 + k}. \tag{47}$$

To invert (47), let us write the right-hand side in the more suggestive product form

$$\hat{u} = \left(\frac{1}{EI\omega^4 + k}\right)(\hat{w}). \tag{48}$$

The inverse of the product is not the product of the two inverses, but is the Fourier convolution of the two inverses,

$$u(x) = F^{-1}\left\{\frac{1}{EI\omega^4 + k}\right\} * F^{-1}\{\hat{w}\}. \tag{49}$$

From entry 8 in Appendix D, and the linearity of F^{-1}, we obtain

$$F^{-1}\left\{\frac{1}{EI\omega^4 + k}\right\} = \frac{\alpha}{\sqrt{2}\,k} e^{-\alpha|x|} \sin\left(\alpha|x| + \frac{\pi}{4}\right), \tag{50}$$

where $\alpha = [k/(4EI)]^{1/4}$, and of course $F^{-1}\{\hat{w}\} = w(x)$, so (49) gives

$$u(x) = \frac{\alpha}{\sqrt{2}\,k} \int_{-\infty}^{\infty} e^{-\alpha|x-\xi|} \sin\left(\alpha|x-\xi| + \frac{\pi}{4}\right) w(\xi)\, d\xi \tag{51}$$

as the desired solution.

COMMENT 1. It is always important to check our results. An excellent partial check of a result is provided by any special case for which the exact solution is known. In the present example, such a special case is provided by the case where $w(x) = \text{constant} \equiv W$, for then surely the solution $u(x)$ will be a constant too. Specifically, with $w(x) = W$, and u a constant, (43) gives $u(x) = W/k$. If we set $w(\xi) = W$ in (51), and integrate, do we obtain the same result? Yes, but we leave that calculation for the exercises.

In fact, observe that the correctness of (51) for the case where $w(x)$ is a constant is a surprise since some of the assumptions that were built into our solution are violated in that case. For instance, the transform $\hat{w}$ of $w(x) = W$ does not even exist, because the transform integral does not converge. That is,

$$\begin{aligned} \hat{w} &= \int_{-\infty}^{\infty} W e^{-i\omega x}\, dx = W \lim_{A\to\infty} \int_{-A}^{A} e^{-i\omega x}\, dx \\ &= \frac{2W}{\omega} \lim_{A\to\infty} \sin \omega A, \end{aligned} \tag{52}$$

which limit does not exist. Without elaborating, we state that (51) ends up being correct, for this case, even though (44) is not (since $F\{w\} = \hat{w}$ does not exist), thanks to the interchange in the order of integration that underlies the convolution property.* In any case, the moral is that it is often best to proceed formally to a solution. If we can verify that the solution thus obtained does satisfy all of the specified requirements (such as a differential equation and boundary conditions), then there is no need to worry about any lack of rigor in the intermediate steps.

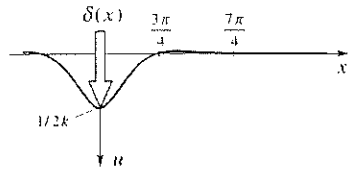


Figure 6. Response to a point unit load at $x = 0$.

COMMENT 2. If w is a delta function, $w(x) = \delta(x)$, then $u(x)$ is the response to a unit load at $x = 0$ (Fig. 6).† Remember that

*See our proof of the convolution formula, following (17).

†A *unit* load because $\int_{-\infty}^{\infty} w(x)\, dx = \int_{-\infty}^{\infty} \delta(x)\, dx = 1$.

$$\int_{-\infty}^{\infty} f(x)\delta(x)\,dx = f(0) \tag{53}$$

if f is continuous, by the definition of the delta function. Thus, if $w(x) = \delta(x)$ in (51) then we obtain the response

$$u(x) = \frac{\alpha}{\sqrt{2}\,k} e^{-\alpha|x|} \sin\left(\alpha|x| + \frac{\pi}{4}\right), \tag{54}$$

which is plotted in Fig. 6 for the representative case where $\alpha = 1$.

COMMENT 3. Let us re-express (51) in the compact form

$$u(x) = \int_{-\infty}^{\infty} K(x-\xi)w(\xi)\,d\xi. \tag{55}$$

That is, the output function u is given by the action of an integral operator on the input function w, the operator being multiplication by K followed by integration from $-\infty$ to ∞. We have written $K(x-\xi)$ rather than $K(\xi, x)$ to show that K depends only on the difference between x and ξ. We call K the **kernel** of the integral operator.* Since K is a function of the difference $x - \xi$, we call it a **difference kernel**. The physical significance of K is revealed by the fact that if the loading $w(\xi)$ in (51) is a point unit load $\delta(\xi)$ at the origin, then the resulting deflection $u(x)$ is, according to (53),

$$u(x) = \int_{-\infty}^{\infty} K(x-\xi)\delta(\xi)\,d\xi = K(x). \tag{56}$$

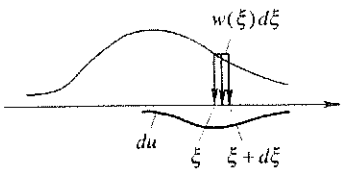


Figure 7. Incremental deflection du, due to load $w(\xi)\,d\xi$.

That is, $K(x)$ is the response function (54), the graph of which is shown in Fig. 6 (for $\alpha = 1$), so $K(x-\xi)$ in (55) is the deflection due to a point unit load at ξ. If the load between ξ and $\xi + d\xi$ is $w(\xi)\,d\xi$ rather than unity (Fig. 7), then the contribution du to the deflection u, due to that bit of loading, is $K(x-\xi)$ scaled by $w(\xi)\,d\xi$,

$$du = K(x-\xi)w(\xi)\,d\xi. \tag{57}$$

Adding all of these du's gives the integral in (55). Thus, we can now understand (55) as a *superposition* principle, whereby $u(x)$ is the sum, or superposition, of the individual contributions du due to each point load $w(\xi)\,d\xi$. That superposition result is a consequence of the linearity of (43). ∎

Additional physical ODE (ordinary differential equation) applications, to the deflection of a loaded string and to the steady-state concentration distribution in a stream subjected to an input of pollutant are given in the exercises. We urge you to at least read through those exercises even if you do not work them.

Closure. We have seen that the Fourier transform and the corresponding inversion formula are actually just a restatement of the Fourier integral representation of a nonperiodic function defined on $-\infty < x < \infty$. We have also seen that the

*To further illustrate the idea of the kernel of an integral operator, we note that the kernel of the Laplace transform operator is e^{-st}.

Fourier transform methodology is analogous, and in some aspects identical, to the Laplace transform methodology studied in Chapter 5. Given that similarity, the important question arises, as to which of the two transform methods to use in a given ODE application. The general guideline is as follows:

- The **Laplace transform** is tailored to **initial-value problems** on a **semi-infinite interval** $0 < t < \infty$.
- The **Fourier transform** is tailored to **boundary-value problems** on an **infinite interval** $-\infty < x < \infty$.

The independent variables are usually t (time) and x (linear dimension), but not always. Further, there does exist a "bilateral" Laplace transform defined on an infinite interval, the Laplace transform can sometimes be used for boundary-value problems on finite intervals, and the Fourier transform can (and will, in Section 17.11) be adapted to boundary-value problems on a semi-infinite interval $0 < x < \infty$. But the guideline given above provides the general rule of thumb for selecting one transform or the other.

In closing, we note that to compute the Fourier transform $\hat{f}(\omega)$ of $f(x)$ requires an integration of f over $(-\infty, \infty)$. One can evaluate the integral numerically by sampling the integrand at discrete x points and obtain what is known as the **discrete Fourier transform**. Further, there exist algorithms known collectively as the **fast Fourier transform** (i.e., the **FFT**) that provide a more efficient method of calculating the sum in the discrete Fourier transform. For an introduction to these methods we refer the interested reader to Peter V. O'Neil's *Advanced Engineering Mathematics*, 3rd ed. (Belmont, CA: Wadsworth, 1991).

Computer software. To obtain the Fourier transform of $e^{-5|x|}$, say, using *Maple*, enter

```
readlib(fourier):
```

to access the Fourier transform and inverse transform commands. Then enter

```
fourier(exp(-5 * abs(x)), x, w);
```

and return. The result is

$$10\,\frac{1}{25 + w^2}$$

which agrees with entry 4 in Appendix D. To invert the latter, enter

```
invfourier(10/(25 + w^2), w, x);
```

and return. The result is

$$e^{-5x}\,\text{Heaviside}(x) + e^{5x}\,\text{Heaviside}(-x)$$

which does reduce to $e^{-5|x|}$.

EXERCISES 17.10

1. Using (6a), derive the result

$$F\{H(x)e^{-ax}\} = \frac{1}{a+i\omega}$$

if $\operatorname{Re} a > 0$. (This case was worked in Example 2, but there a was considered to be real. Here, allow for a to be complex, with $\operatorname{Re} a > 0$.)

2. Using (6a), derive the result

$$F\{H(-x)e^{ax}\} = \frac{1}{a-i\omega}$$

if $\operatorname{Re} a > 0$.

3. Using (6a), derive the result

$$F\{e^{-a|x|}\} = \frac{2a}{\omega^2+a^2}$$

if $a > 0$.

4. Given $\hat{f}(\omega)$, use (6b) to evaluate the inverse, $f(x)$.

(a) $e^{-a|\omega|}$ $(a > 0)$
(b) $H(\omega+a) - H(\omega-a)$ $(a > 0)$
(c) $[H(\omega) - H(\omega-1)]\omega$
(d) $H(\omega)e^{-a\omega}$ $(\operatorname{Re} a > 0)$
(e) $[H(\omega+1) - H(\omega-1)]|\omega|$
(f) $\delta(\omega - a)$

5. Derive the following entry in Appendix D, by showing that the transform of the function given in the f column is the result given in the $\hat{f}$ column.

(a) Entry 10 (b) Entry 11
(c) Entry 12 (d) Entry 13
(e) Entry 14 (f) Entry 15
(g) Entry 17, by formally differentiating both sides of

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(x)e^{-i\omega x}\,dx$$

a sufficient number of times, with respect to ω
(h) Entry 19, using integration by parts

6. Evaluate the following using Appendix D. You may need to use more than one entry. Cite, by number, any entries that you use.

(a) $F\left\{4x^2e^{-3|x|}\right\}$ (b) $F\left\{xe^{-x^2}\right\}$
(c) $F\left\{\dfrac{\cos 3x}{x^2+2}\right\}$ (d) $F\left\{\dfrac{\sin 2x}{4x^2+3}\right\}$
(e) $F\left\{\dfrac{3}{2x^2+1} - 5e^{-|x|}\right\}$ (f) $F\left\{e^{-|x|} + e^{-3|x+2|}\right\}$
(g) $F^{-1}\left\{\dfrac{4\sin\omega}{\omega} - \dfrac{1}{\sqrt{|\omega|}}\right\}$ (h) $F^{-1}\left\{e^{-2|\omega-3|}\right\}$
(i) $F^{-1}\left\{\dfrac{9}{2\omega+i}\right\}$ (j) $F^{-1}\left\{e^{-\omega^2+4\omega}\right\}$
(k) $F^{-1}\left\{e^{-|\omega|}\cos\omega\right\}$ (l) $F^{-1}\left\{\omega e^{-3\omega^2}\right\}$
(m) $F^{-1}\left\{\dfrac{1}{\omega^2+i\omega+2}\right\}$ (n) $F^{-1}\left\{\dfrac{\omega}{\omega^2+1}\right\}$

7. (a)–(n) Use computer software to obtain (if possible) the transform or inverse transform called for in the corresponding part of Exercise 6.

8. We claimed, in Comment 1 in Example 8, that if $w(x) = W$ then the integral on the right-hand side of (51) is W/k. Verify that claim by evaluating the integral. HINT: Break the integral into two parts, one from $-\infty$ to x and the other from x to ∞, then make a suitable change of variables in each.

9. (*Preservation of evenness and oddness*)

(a) Show that $f(x)$ is an even function of x if and only if $\hat{f}(\omega)$ is an even function of ω.
(b) Show that $f(x)$ is an odd function of x if and only if $\hat{f}(\omega)$ is an odd function of ω.

10. (*Extension of transform tables*) It follows from the results in Exercise 9 that the transform of the even and odd parts of $f(x)$ are the even and odd parts of $\hat{f}(\omega)$, respectively. This result can be used to obtain more information from a given transform table. For example, consider entry 2 in Appendix D. Breaking f and $\hat{f}$ into even and odd parts, show that

$$F\left\{e^{-a|x|}\right\} = \frac{2a}{\omega^2+a^2} \tag{10.1}$$

and

$$F\left\{(\operatorname{sgn} x)e^{-a|x|}\right\} = -\frac{2i\omega}{\omega^2+a^2}, \tag{10.2}$$

where

$$\operatorname{sgn} x \equiv \begin{cases} +1, & x > 0 \\ -1, & x < 0, \end{cases} \tag{10.3}$$

which is read as "sign of x." Of these two results, observe that (10.1) is identical to entry 4, and that (10.2) is not contained in Appendix D.

11. (*Extension of transform tables*) Another idea that enables us to extend a given Fourier transform table is that of reciprocity, namely, the reciprocity relations

$$F\{\hat{f}(x)\} = 2\pi f(-\omega) \tag{11.1}$$

and

$$F^{-1}\{f(-\omega)\} = \frac{\hat{f}(x)}{2\pi}. \tag{11.2}$$

(a) Derive the relations (11.1) and (11.2).
(b) To illustrate, use (11.1) and entry 4, in Appendix D, to show that

$$F\left\{\frac{2a}{x^2+a^2}\right\} = 2\pi e^{-a|\omega|}, \qquad (a > 0)$$

or, equivalently,

$$F\left\{\frac{1}{x^2+a^2}\right\} = \frac{\pi}{a} e^{-a|\omega|}. \qquad (a > 0)$$

(In this case the result does not extend our table since it already appears as Entry 1.)
(c) Use (11.1) and entry 9 to show that

$$F\left\{\frac{\sin ax}{x}\right\} = \pi[H(\omega+a) - H(\omega-a)]. \qquad (a > 0)$$

(d) Use (11.1) and entry 3 to show that

$$F\left\{\frac{1}{a-ix}\right\} = 2\pi H(\omega) e^{-a\omega}. \qquad (\operatorname{Re} a > 0)$$

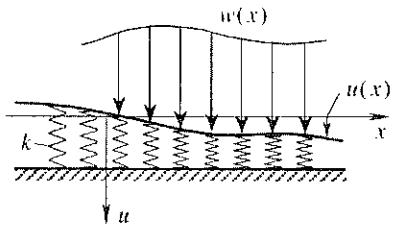
12. (*Deflection of loaded string*) Related to the problem of a beam on an elastic foundation is the analogous problem for a flexible string. Imagine a string (of negligible mass per unit length) stretched along the x axis, over $-\infty < x < \infty$, by a tension T newtons, and let $w(x)$ be an applied load distribution (newtons/meter), as sketched in the figure. If the displacement $u(x)$ of the string is resisted by a distributed spring of stiffness k (newtons per meter per meter), and the slope $u'(x)$ is sufficiently small over $-\infty < x < \infty$, then $u(x)$ is accurately governed by the differential equation $Tu'' - ku = -w$, or,

$$u'' - \alpha^2 u = -f(x), \qquad (-\infty < x < \infty) \tag{12.1}$$

where $\alpha^2 = k/T$ and $f(x) = w(x)/T$. (Here we consider the static deflection. In Chapter 20 we will derive the governing differential equation for the not-necessarily-static case.) Assume that $w(x)$ is sufficiently localized for $u(x)$ to satisfy the boundary conditions

$$u \to 0 \quad \text{and} \quad u' \to 0 \tag{12.2}$$

as $x \to \pm\infty$, as well.



(a) Solving (12.1) and (12.2) using a Fourier transform, derive the solution

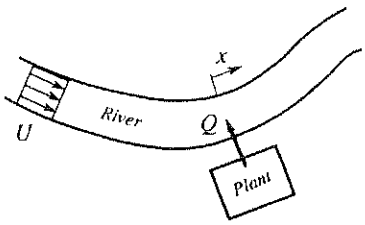
$$u(x) = -\frac{1}{2\alpha}\int_{-\infty}^{\infty} e^{-\alpha|x-\xi|} f(\xi)\, d\xi. \tag{12.3}$$

(b) Evaluate $u(x)$ from (12.3) for the case where $f(x)$ is a point unit load at the origin, $f(x) = \delta(x)$, and sketch the graph of u.
(c) Verify, by formally using the Leibniz rule, that (12.3) satisfies the differential equation (12.1).

13. (*Pollution in river*) Suppose that a manufacturing plant discharges a certain pollutant into an initially clear river at the rate Q grams/second. We wish to determine the resulting steady-state distribution of pollutant in the river, i.e., its concentration c (grams/meter^3). Measure x along the river, positive downstream, with origin at the plant site, as shown in the figure. The river flows with velocity U (meters/second), and has a cross-sectional area A, both of which, for simplicity, we assume to be constant. Also for simplicity, suppose that c is a function of x only; i.e., it is a constant over each cross section of the stream. This is evidently a poor approximation near the plant, where we expect appreciable across-stream and vertical variations in c, but it should suffice if we are concerned mostly with the concentration variation far upstream and downstream (for $|x|$ greater than several river widths, say). Then it can be shown that $c(x)$ is governed by the differential equation

$$kc'' - Uc' - \beta c = -\frac{Q}{A}\,\delta(x), \tag{13.1}$$

where k (meters^2/second) is a diffusion constant, β (grams per second per gram) is a chemical decay constant, and $\delta(x)$ is a delta function. [Physically, (13.1) expresses a mass balance between the *input* $Q\delta(x)/A$, the transport of pollutant by *diffusion*, kc'', the transport of pollutant by *convection* with the moving stream, Uc', and by disappearance through *chemical decay*, βc.]



(a) Applying the Fourier transform to (13.1), show that

$$c(x) = F^{-1}\left\{\frac{Q}{A}\,\frac{1}{k\omega^2 + iU\omega + \beta}\right\}. \tag{13.2}$$

(b) Expanding $1/(k\omega^2 + iU\omega + \beta)$ in partial fractions, and then using Appendix D, show that (13.2) gives

$$c(x) = \begin{cases} c_0 e^{-\Omega_- x}, & x < 0 \\ c_0 e^{-\Omega_+ x}, & x > 0 \end{cases} \tag{13.3}$$

where

$$c_0 = \frac{Q}{A\sqrt{U^2 + 4k\beta}},$$

$$\Omega_\pm = \frac{1}{2k}\left(\pm\sqrt{U^2 + 4k\beta} - U\right).$$

(c) Sketch the graph of $c(x)$ and state the qualitative effect of increasing β.

17.11 Fourier Cosine and Sine Transforms, and Passage from Fourier Integral to Laplace Transform (Optional)

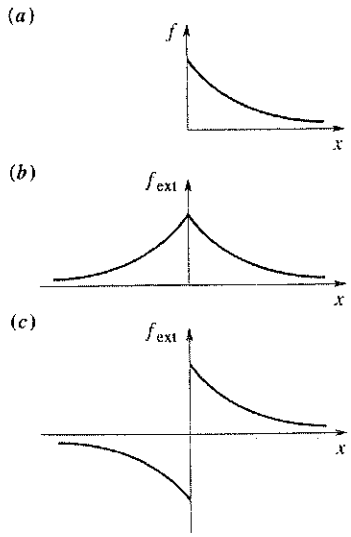


Figure 1. Even and odd extensions of f.

17.11.1. Cosine and sine transforms. Recall from Section 17.4 that if a problem is defined on a finite interval, say $0 < x < L$, then the concepts of periodicity and Fourier series are not directly applicable. However, by fictitiously extending both the domain and the functions involved to the infinite interval $-\infty < x < \infty$, so that the extended functions are periodic, we are able to use Fourier series representations of those functions. Depending upon the symmetries and/or antisymmetries about $x = 0$ and $x = L$, we obtain half- or quarter-range cosine or sine expansions.

Similarly, we sometimes encounter problems defined on a semi-infinite interval, say $0 < x < \infty$. Fictitiously extending both the domain and the functions involved, to the infinite interval $-\infty < x < \infty$, we will be able to use Fourier integral representations of those functions. Specifically, given a function f defined on $0 < x < \infty$ (such as the one shown in Fig. 1a), we shall be interested in two particular extensions, one that is even and one that is odd, as indicated in Figs.1b and 1c, respectively.

Denoting the extended functions as f_{ext}, consider first the even extension (Fig. 1b). Then the Fourier transform of f_{ext} is

$$\widehat{f_{\text{ext}}}(\omega) = \int_{-\infty}^{\infty} f_{\text{ext}}(x)e^{-i\omega x}\,dx$$

$$= \int_{-\infty}^{\infty} f_{\text{ext}}(x)(\cos\omega x - i\sin\omega x)\,dx = 2\int_0^{\infty} f_{\text{ext}}(x)\cos\omega x\,dx, \quad (1a)$$

where the last equality holds because $f_{\text{ext}}(x)\cos\omega x$ is an even function of x and $f_{\text{ext}}(x)\sin\omega x$ is odd, and the inversion formula gives

$$\begin{aligned} f_{\text{ext}}(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{f_{\text{ext}}}(\omega)e^{i\omega x}\,d\omega \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \widehat{f_{\text{ext}}}(\omega)(\cos\omega x + i\sin\omega x)\,d\omega \\ &= \frac{2}{2\pi}\int_0^{\infty} \widehat{f_{\text{ext}}}(\omega)\cos\omega x\,d\omega, \end{aligned} \quad (1b)$$

where the last equality holds because $\widehat{f_{\text{ext}}}(\omega)$ is an even function of ω [as can be seen from (1a)] and $\sin\omega x$ is an odd function of ω, so that the term $\widehat{f_{\text{ext}}}(\omega)\sin\omega x$ in the integrand is odd. Putting (1a) into (1b) gives the single statement

$$f_{\text{ext}}(x) = \int_0^{\infty}\left\{\frac{2}{\pi}\int_0^{\infty} f_{\text{ext}}(\xi)\cos\omega\xi\,d\xi\right\}\cos\omega x\,d\omega. \quad (-\infty < x < \infty) \quad (2)$$

Since (2) holds on $-\infty < x < \infty$, it also holds on the original interval $0 < x < \infty$. For $0 < x < \infty$ we can drop the subscripted "ext" on the left-hand side, because on that interval $f_{\text{ext}}(x) = f(x)$. Further, we can drop the "ext" on the right-hand side because the ξ integral is on $0 < x < \infty$, not $-\infty < x < \infty$. Then (2) becomes

$$f(x) = \frac{2}{\pi}\int_0^{\infty}\left\{\int_0^{\infty} f(\xi)\cos\omega\xi\,d\xi\right\}\cos\omega x\,d\omega, \qquad (0 < x < \infty) \quad (3)$$

which can be re-expressed, equivalently, as the **Fourier cosine transform**

$$\boxed{F_C\{f(x)\} = \hat{f}_C(\omega) = \int_0^{\infty} f(x)\cos\omega x\,dx} \quad (4a)$$

and its inverse

$$\boxed{F_C^{-1}\{\hat{f}_C(\omega)\} = f(x) = \frac{2}{\pi}\int_0^{\infty} \hat{f}_C(\omega)\cos\omega x\,d\omega,} \quad (4b)$$

on $0 < x < \infty$.

Similarly, an odd extension of f gives the **Fourier sine transform**

$$\boxed{F_S\{f(x)\} = \hat{f}_S(\omega) = \int_0^{\infty} f(x)\sin\omega x\,dx} \quad (5a)$$

and its inverse

$$\boxed{F_S^{-1}\{\hat{f}_S(\omega)\} = f(x) = \frac{2}{\pi}\int_0^{\infty} \hat{f}_S(\omega)\sin\omega x\,d\omega,} \quad (5b)$$

on $0 < x < \infty$ (Exercise 1). Sufficient conditions on f, for (4) and (5) to hold, are that f and f' be piecewise continuous on $0 \le x < \infty$, and that $\int_0^\infty |f(x)|\,dx$ converge (i.e., that f be *absolutely integrable* on $0 < x < \infty$).

Next, we could derive the various properties of the cosine and sine transforms, that are analogous to those derived for the Fourier transform. However, for brevity let it suffice to state that both the cosine and sine transforms, and their inverses, are linear, and that the transforms of the derivative f' are

$$\boxed{\begin{aligned} F_C\{f'(x)\} &= \omega \hat{f}_S(\omega) - f(0), \\ F_S\{f'(x)\} &= -\omega \hat{f}_C(\omega), \end{aligned}} \tag{6a,b}$$

if we assume additionally that $f(x) \to 0$ as $x \to \infty$. Let us derive (6a) and leave (6b) for the exercises. Integrating by parts,

$$\begin{aligned} F_C\{f'(x)\} &= \int_0^\infty f'(x)\cos\omega x\,dx \\ &= [f(x)\cos\omega x]\Big|_0^\infty + \omega\int_0^\infty f(x)\sin\omega x\,dx \\ &= 0 - f(0) + \omega\hat{f}_S(\omega). \end{aligned} \tag{7}$$

The transforms of higher-order derivatives can be obtained by repeated use of (6). For instance, replacing the function f by the function f' in (6) gives

$$F_C\{f''(x)\} = \omega F_S\{f'(x)\} - f'(0) = \omega[-\omega\hat{f}_C(\omega)] - f'(0) \tag{8a}$$

and

$$F_S\{f''(x)\} = -\omega F_C\{f'(x)\} = -\omega[\omega\hat{f}_S(\omega) - f(0)], \tag{8b}$$

so

$$\boxed{\begin{aligned} F_C\{f''(x)\} &= -\omega^2\hat{f}_C(\omega) - f'(0), \\ F_S\{f''(x)\} &= -\omega^2\hat{f}_S(\omega) + \omega f(0), \end{aligned}} \tag{9a,b}$$

if f and f' tend to zero as $x \to \infty$. Similarly for higher-order derivatives. For convolution properties, see Exercise 10 and Appendix E.

Short Fourier cosine and sine transform tables are given in Appendix E.

EXAMPLE 1. Consider the boundary-value problem

$$u'' - 9u = 50e^{-2x}, \qquad (0 < x < \infty) \tag{10a}$$

$$u(0) = u_0, \quad u(\infty) \text{ bounded.} \tag{10b}$$

To solve (10) using an integral transform we need to choose among the Laplace, Fourier cosine, and Fourier sine transforms, all of these being candidates because they are semi-infinite transforms; that is, they apply when the domain is semi-infinite ($0 < x < \infty$ in

this case). The Laplace transform will be inconvenient at best, because it is tailored to initial value problems whereas (10) is of boundary-value type. In choosing between the Fourier cosine and sine transforms, the key is in (9). Taking a cosine transform of (10a) we will, according to (9a), need to know $u'(0)$, yet the latter is not prescribed; taking a sine transform of (10a) we will, according to (9a), need to know $u(0)$, and the latter *is* prescribed.

Thus, take the Fourier sine transform of (10a), using the linearity of the transform, property (9b), and entry $1S$ in Appendix E:

$$F_S\left\{u'' - 9u\right\} = F_S\left\{50e^{-2x}\right\}, \tag{11a}$$

$$F_S\left\{u''\right\} - 9F_S\left\{u\right\} = 50F_S\left\{e^{-2x}\right\}, \tag{11b}$$

$$-\omega^2\hat{u}_S + \omega u_0 - 9\hat{u}_S = 50\,\frac{\omega}{\omega^2+4}. \tag{11c}$$

Solving this linear algebraic equation for $\hat{u}_S$ gives

$$\hat{u}_S(\omega) = u_0\,\frac{\omega}{\omega^2+9} - 50\,\frac{\omega}{(\omega^2+4)(\omega^2+9)}. \tag{12}$$

By partial fractions,

$$\frac{1}{(\omega^2+4)(\omega^2+9)} = \frac{1}{5}\,\frac{1}{\omega^2+4} - \frac{1}{5}\,\frac{1}{\omega^2+9}, \tag{13}$$

so

$$\hat{u}_S(\omega) = (u_0+10)\frac{\omega}{\omega^2+9} - 10\frac{\omega}{\omega^2+4}. \tag{14}$$

Then, using the linearity of the inverse transform and entry $1S$, gives

$$\begin{aligned} u(x) &= F_S^{-1}\left\{(u_0+10)\frac{\omega}{\omega^2+9} - 10\frac{\omega}{\omega^2+4}\right\} \\ &= (u_0+10)F_S^{-1}\left\{\frac{\omega}{\omega^2+9}\right\} - 10F_S^{-1}\left\{\frac{\omega}{\omega^2+4}\right\} \\ &= (u_0+10)e^{-3x} - 10e^{-2x}. \end{aligned} \tag{15}$$

COMMENT 1. Note that we used property (9b) tentatively since (9b) presupposes that both $u(x) \to 0$ and $u'(x) \to 0$ as $x \to \infty$. Now that $u(x)$ is in hand, in (15), we can check, and we see that these conditions are met. Or, more directly, we can simply verify that (15) does satisfy (10a) and (10b).

COMMENT 2. If $u'(0)$ were prescribed, in (10b), in place of $u(0)$, then we would use the Fourier cosine transform instead. ∎

17.11.2. Passage from Fourier integral to Laplace transform. We have seen that the Fourier transform is merely a restatement of the Fourier integral, the Fourier integral is a limiting case of the Fourier series of a periodic function (as the period tends to infinity), and the Fourier series is as an eigenfunction expansion corresponding to a periodic Sturm–Liouville problem. On the other hand, the Laplace transform and its inversion formula were given in Chapter 5 without derivation and

may seem unrelated to these Fourier methods. In this final subsection we will show that the Laplace transform and its inverse can be derived from the Fourier integral!

We begin with the Fourier integral in the complex exponential form given by equation (3) in Section 17.10,

$$F(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} F(\tau)e^{-i\omega\tau}\,d\tau\right\}e^{i\omega t}\,d\omega, \qquad (-\infty < t < \infty) \tag{16}$$

where we use t in place of x because we will end up with the Laplace transform, in which the independent variable is traditionally taken to be t (because in applications it usually corresponds to the time). Further, it will be convenient to use F in place of f. In (16), let

$$F(t) = \begin{cases} e^{-\gamma t}f(t), & t > 0 \\ 0, & t < 0, \end{cases} \tag{17}$$

where γ is a real constant which is sufficiently positive so that $e^{-\gamma t}$"clobbers" $f(t)$ as $t \to \infty$. Specifically, suppose that f is of exponential order (defined in section 5.2) as $t \to \infty$. Then a sufficiently positive γ can indeed be found such that $e^{-\gamma t}f(t)$ dies out exponentially fast as $t \to \infty$. Of course, whereas $e^{-\gamma t}$ helps as $t \to +\infty$, it hurts as $t \to -\infty$. Thus, we simply "shut off F" for $t < 0$ by defining it to be zero for $t < 0$, in (17). The resulting F easily satisfies the condition $\int_{-\infty}^{\infty}|F(t)|\,dt < \infty$ contained in the Fourier integral theorem.

Putting (17) into (16) gives

$$H(t)e^{-\gamma t}f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left\{\int_{0}^{\infty} e^{-\gamma\tau}f(\tau)e^{-i\omega\tau}\,d\tau\right\}e^{i\omega t}\,d\omega \tag{18}$$

where H is the Heaviside function. Thus,

$$H(t)f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left\{\int_{0}^{\infty} e^{-(\gamma+i\omega)\tau}f(\tau)\,d\tau\right\}e^{(\gamma+i\omega)t}\,d\omega \tag{19}$$

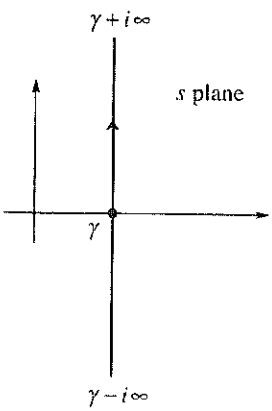


Figure 2. Line of integration in s plane.

which form suggests changing variables from ω to s according to

$$s = \gamma + i\omega. \tag{20}$$

Thus,

$$H(t)f(t) = \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty}\left\{\int_{0}^{\infty} e^{-s\tau}f(\tau)\,d\tau\right\}e^{st}\,ds, \tag{21}$$

where $\int_{\gamma-i\infty}^{\gamma+i\infty}$ denotes an integration along a vertical line in a complex s plane (Fig. 2). If we define the **Laplace transform** of f as

$$\boxed{L\{f(t)\} = \overline{f}(s) = \int_{0}^{\infty} f(t)e^{-st}\,dt,} \tag{22}$$

where we have changed the dummy integration variable from τ to t, then (21) gives the inversion formula as

$$\boxed{L^{-1}\{\overline{f}(s)\} = H(t)f(t) = \frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} \overline{f}(s)e^{st}\,ds.} \tag{23}$$

Normally we are interested only in $0 < t < \infty$, so we can replace $H(t)$ by 1 in (23). However, for completeness we observe that while the integral in (23) gives $f(t)$ for $t > 0$, it gives 0 for $t < 0$, which result is, after all, in accordance with our definition $F(t) = 0$ for $t < 0$, in (17).

Closure. In this section we have combined two topics, each being too brief to justify a full section. First we introduced the Fourier cosine and sine transforms, for problems defined on the semi-infinite interval $0 < x < \infty$, which results are analogous to the half-range cosine and sine Fourier series. The Laplace transform is also a semi-infinite transform, but it is tailored to *initial-value* problems whereas the Fourier cosine and sine transforms are tailored to *boundary-value* problems, because $L\{f^{(n)}(t)\}$ involves values of $f, f', \ldots, f^{(n-1)}$ at $t = 0$, whereas $F_C\{f^{(n)}(x)\}$ and $F_S\{f^{(n)}(x)\}$ involve values both at $x = 0$ and at $x = \infty$. Whether to use a cosine transform or a sine transform, in solving a differential equation boundary-value problem, depends (as we saw in Example 1) on the type of boundary conditions prescribed at $x = 0$.

Finally, we used the complex exponential form of the Fourier integral to derive the Laplace transform and its inversion formula. Normally, in using the Laplace transform, we have the t domain of interest is $t > 0$. However, we noted that if one *were* to obtain an inverse Laplace transform by evaluating the inversion integral in (23), for $t < 0$, then one would inevitably find that the inverse function is identically zero for $t < 0$.

EXERCISES 17.11

1. Derive the formulas (5a) and (5b) for the Fourier sine transform and its inverse, respectively.

2. Derive (6b), that $F_S\{f'(x)\} = -\omega\hat{f}_C(\omega)$.

3. Derive these results:

(a) $F_C\{f''''(x)\} = \omega^4\hat{f}_C(\omega) + \omega^2 f'(0) - f'''(0)$
(b) $F_S\{f''''(x)\} = \omega^4\hat{f}_S(\omega) - \omega^3 f(0) + \omega f''(0)$ if $f(x), f'(x), f''(x)$, and $f'''(x)$ all tend to zero as $x \to \infty$.

4. Given the rectangular pulse $f(x) = 50[1 - H(x-4)]$, evaluate $\hat{f}_C(\omega)$ and $\hat{f}_S(\omega)$.

5. Use computer software, such as the int command on *Maple*, to evaluate each, where $a > 0$. Show that your result agrees with the corresponding result obtained from Appendix E.

(a) $F_C\{e^{-ax}\}$ (b) $F_S\{e^{-ax}\}$
(c) $F_C\{xe^{-ax}\}$ (d) $F_S\{xe^{-ax}\}$
(e) $F_C\{x^2e^{-ax}\}$ (f) $F_S\{x^2e^{-ax}\}$

NOTE: Be careful, in using tables or software, to be clear on the author's definition of the transform. For instance, in contrast with our definitions (4) and (5) of the Fourier cosine and sine transforms, some authors use the more symmetric versions

$$F_C\{f(x)\} = \sqrt{\frac{2}{\pi}}\int_0^\infty f(x)\cos\omega x\,dx \tag{5.1}$$

$$F_C^{-1}\{\hat{f}_C(\omega)\} = \sqrt{\frac{2}{\pi}} \int_0^\infty \hat{f}_C(\omega) \cos \omega x \, d\omega \qquad (5.2)$$

and

$$F_S\{f(x)\} = \sqrt{\frac{2}{\pi}} \int_0^\infty f(x) \sin \omega x \, dx \qquad (5.3)$$

$$F_S^{-1}\{\hat{f}_S(\omega)\} = \sqrt{\frac{2}{\pi}} \int_0^\infty \hat{f}_S(\omega) \sin \omega x \, d\omega. \qquad (5.4)$$

6. In Example 1 we avoided the Laplace transform because (10) is of boundary-value type, not of initial-value type. Show that the Laplace transform can be used nevertheless, and does give the solution (15), though not as conveniently. HINT: When you take the transform of u'' you will be faced with a $u'(0)$ term, which is not prescribed in (10b). Thus, call that quantity C, say, and evaluate it by imposing on your solution the condition that $u(\infty)$ be bounded, at the end.

7. We used a sine transform to solve Example 1. Try to solve (10) using a cosine transform instead, and explain why that method does not work.

8. Modify (10) by changing $u(0) = u_0$ to $u'(0) = u_0'$, and solve by a cosine or sine transform.

9. Solve, using a cosine or sine transform.

(a) $u'' - 9u = 50e^{-3x}, \quad (0 < x < \infty)$
$u(0) = 0, \; u(\infty)$ bounded

(b) $u'' - 9u = 50e^{-3x}, \quad (0 < x < \infty)$
$u'(0) = 0, \; u(\infty)$ bounded

10. (*Convolution Theorem*) As for the Fourier and Laplace transforms there are convolution theorems for the Fourier cosine and sine transforms, and these are given in Appendix E.

(a) Prove the Fourier cosine transform convolution theorem (entry 7C), either by showing that the transform of the given integral is $\hat{f}_C(\omega)\hat{g}_S(\omega)$ or by showing that the inverse of $\hat{f}_C(\omega)\hat{g}_S(\omega)$ is the given integral.
(b) Prove the Fourier sine transform convolution theorem (entry 7S).
(c) Verify entry 7C for the case where $f(x) = e^{-x}$ and $g(x) = e^{-3x}$.
(d) Verify entry 7S for the case where $f(x) = e^{-x}$ and $g(x) = e^{-3x}$.

Chapter 17 Review

We began with the Fourier series representation

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi x}{\ell} + b_n \sin \frac{n\pi x}{\ell} \right) \qquad (-\infty < x < \infty) \qquad (1)$$

of any 2ℓ-periodic function f defined on $-\infty < x < \infty$, which representation is valid subject to very mild conditions on f. For instance, f can even have jump discontinuities, whereas to represent f by a Taylor series f needs to be infinitely differentiable over the interval under consideration (and even that condition does not quite suffice). In applications, periodic functions arise in a number of ways. For instance an offshore structure is probably subjected to wave forces that are periodic in time, and the temperature distribution around the edge of a circular disk is a 2π-periodic function of the polar angle θ.

If, instead, the x domain is finite, say $0 < x < L$, then Fourier series can still be employed – by fictitiously extending the domain to $-\infty < x < \infty$, and extending the definition of f onto that domain so that f_{ext} is periodic. Such extensions can be accomplished in an infinite number of ways, but the four that will be needed, in applications, correspond to extensions that are symmetric or antisymmetric about

the ends $x = 0$ and $x = L$, and these are the half- and quarter-range cosine and sine expansions, which we denote by HRC, HRS, QRC, and QRS, respectively:

$$\text{HRC:} \qquad f(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos \frac{n\pi x}{L} \qquad (0 < x < L) \tag{2}$$

$$\text{HRS:} \qquad f(x) = \sum_{n=1}^{\infty} b_n \sin \frac{n\pi x}{L} \qquad (0 < x < L) \tag{3}$$

$$\text{QRC:} \qquad f(x) = \sum_{n=1,3,\ldots}^{\infty} a_n \cos \frac{n\pi x}{2L} \qquad (0 < x < L) \tag{4}$$

$$\text{QRS:} \qquad f(x) = \sum_{n=1,3,\ldots}^{\infty} b_n \sin \frac{n\pi x}{2L} \qquad (0 < x < L) \tag{5}$$

(For brevity, we do not repeat, here, the formulas for the a_n's and b_n's given in the text.) We emphasized that the choice, as to which of these four expansions to use, will be dictated by the context.

From a vector space viewpoint, (1)–(5) amount to expansions of f in terms of an infinite set of orthogonal base vectors. In (1), for instance, the base vectors are $1, \cos(\pi x/\ell), \sin(\pi x/\ell), \cos(2\pi x/\ell), \sin(2\pi x/\ell), \ldots$. Such sets of orthogonal base vectors arise as the eigenfunctions of Sturm–Liouville problems, namely, eigenvalue problems of the type

$$(py')' + qy + \lambda wy = 0, \qquad (a < x < b) \tag{6}$$

with homogeneous boundary conditions at $x = a$ and $x = b$, with the inner product

$$\langle u, v \rangle = \int_a^b u(x)v(x)w(x)\,dx \tag{7}$$

with weight function w. For instance, the base vectors in (1) are the eigenfunctions of the Sturm–Liouville problem

$$\begin{gathered} y'' + \lambda y = 0, \quad (-\ell < x < \ell) \\ y(-\ell) - y(\ell) = 0, \quad y'(-\ell) - y'(\ell) = 0, \end{gathered} \tag{8}$$

and the base vectors in (2)–(5) are the eigenfunctions of the Sturm–Liouville problems

$$\begin{gathered} y'' + \lambda y = 0, \qquad (0 < x < L) \\ \begin{aligned} \text{HRC:} \quad & y'(0) = 0, \quad y'(L) = 0, \\ \text{HRS:} \quad & y(0) = 0, \quad y(L) = 0, \\ \text{QRC:} \quad & y'(0) = 0, \quad y(L) = 0, \\ \text{QRS:} \quad & y(0) = 0, \quad y'(L) = 0, \end{aligned} \end{gathered} \tag{9}$$

respectively. The four cases in (9) are examples of regular Sturm–Liouville systems, while the Sturm–Liouville problem in (8) is of periodic type. Singular Sturm–Liouville problems were discussed in Section 17.8, prominent examples involving the Bessel and Legendre equations.

We showed that if we let $\ell \to \infty$ in (1), then the frequency spectrum $\{n\pi/\ell\}$ becomes a continuous spectrum from 0 to ∞ and we obtain, in place of the Fourier series (1), the Fourier integral

$$f(x) = \int_0^\infty [a(\omega)\cos\omega x + b(\omega)\sin\omega x]\,d\omega. \qquad (-\infty < x < \infty) \tag{10}$$

Expressing (10) in complex exponential form, we obtained the equation pair

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(x)e^{-i\omega x}\,dx, \tag{11a}$$

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(\omega)e^{i\omega x}\,d\omega, \tag{11b}$$

which are equivalent to (10) and which are the Fourier transform and its inverse, respectively. Finally, we obtained the Fourier cosine and sine transforms from (11), for problems defined on the semi-infinite interval $0 < x < \infty$, these being analogous to the half-range cosine and sine representations of functions defined on $0 < x < L$.

Chapter 18

Diffusion Equation

18.1 Introduction

Chapters 18–20 are about partial differential equations (PDE's). Two of the most important PDE's of mathematical physics have already been encountered in Chapter 16. In Example 3 in Section 16.8, we derived the equation

$$\alpha^2 \nabla^2 T = \frac{\partial T}{\partial t} \tag{1}$$

governing the unsteady diffusion of heat by conduction, where α^2 is a physical constant known as the diffusivity of the material, and $T(x, y, z, t)$ is the temperature field. The latter is known as the **heat equation** but is also called the **diffusion equation** because it governs diffusion process in general. For instance, whereas (1) governs the diffusion of *heat*, the equation

$$D\nabla^2 c = \frac{\partial c}{\partial t} \tag{2}$$

governs the unsteady diffusion of *material* (such as a particular chemical pollutant within a body of water or of an anti-cancer drug within an organ such as the liver), where D, like α^2, is a concentration of the material (i.e., the mass of material per unit volume of medium); (2) is of the same form as (1).

If a steady state is achieved, then $\partial T/\partial t = 0$ and (1) reduces to the **Laplace equation**

$$\nabla^2 T = 0. \tag{3}$$

In Section 16.10 we found that the Laplace equation also governs the velocity potential $\Phi(x, y, z)$ for irrotational incompressible flows; other applications of the Laplace equation are discussed in Chapter 20.

Arguably, equations (1) and (3) are two of the three most prominent PDE's of mathematical physics, the third being the **wave equation** of the form

$$c^2 \nabla^2 u = \frac{\partial^2 u}{\partial t^2} \tag{4}$$

where c^2 is a constant.

The diffusion, wave, and Laplace equations are the subjects of Chapters 18, 19, and 20, respectively. They could be presented in any order, but we have chosen this sequence for pedagogical reasons that will be explained as we proceed.

Our approach in these three chapters is substantially different from our approach in the early chapters on ODE's (ordinary differential equations). For ODE's we proceeded systematically, beginning with first-order equations and then moving on to equations of second order and higher and developing the general theory – which covered existence, uniqueness, and methods of solution. Though our emphasis was on linear equations, we considered nonlinear equations as well. For PDE's, however, our scope is more limited as we focus almost entirely on the diffusion, wave, and Laplace equations and emphasize solution technique. These three are by no means the only PDE's encountered in applications, but they are extremely important, and the solution methods that we develop can be applied to various other (linear) PDE's as well.

18.2 Preliminary Concepts

18.2.1. Definitions. Recall (from Section 1.2) that a differential equation is a **partial differential equation** if it contains partial derivatives of the dependent variable with respect to two or more independent variables. For most applications the independent variables are one or more space variables (Cartesian or non-Cartesian), and possibly the time t, but the dependent variables encountered are much more varied and include temperature, concentration, deflection of a string or membrane or beam, velocity potential, and electric potential, to name just a few. As for ODE's, there may be more than one dependent variable, and we may have a system of PDE's in two or more unknowns. Here, however, we consider only the case of one equation in one unknown.

Also as for ODE's, we define the **order** of a PDE as the order of the highest derivative therein, and we say that a function is a **solution** of a PDE, over a particular domain of the independent variables, if its substitution into the equation reduces that equation to an identity everywhere within that domain. For instance, the PDE

$$u_{xx} - 5u_y + u = xy^2(y - 15) \tag{1}$$

(where subscripts denote partial derivatives) is of second order and admits the solutions

$$u_1(x, y) = 6e^{2x+y} + xy^3 \qquad \text{and} \qquad u_2(x, y) = -e^{3x+2y} + xy^3$$

over the entire x, y plane (and other solutions as well). For instance, putting u_1 into (1) gives

$$24e^{2x+y} - 5(6e^{2x+y} + 3xy^2) + 6e^{2x+y} + xy^3 = xy^2(y - 15),$$

which is indeed an identity for all values of x and y.

It is standard and convenient to use the differential operator notation $L[\ ]$ for PDE's, as we do for ODE's. Accordingly, (1) may be written more compactly as

$$L[u] = f, \tag{2}$$

where

$$L = \frac{\partial^2}{\partial x^2} - 5\frac{\partial}{\partial y} + 1 \tag{3}$$

is the second-order *partial differential operator*, and $f(x, y) = xy^2(y - 15)$. Just as the definition of a function is not complete until we specify the domain on which it acts [e.g., the function $\sin x$ defined on $0 \le x \le \pi$ is not the same as the function $\sin x$ defined on $-\infty < x < 5$], likewise the definition of a differential operator is not complete until we specify the domain of functions on which it acts. We do not "figure out" what the domain is; we specify it. For L given by (3), for instance, we might specify its domain as the set of functions $u(x, y)$ that are defined on the first quadrant ($0 < x < \infty$, $0 < y < \infty$) and that are twice differentiable in x and once in y.

We say that a differential operator, be it an ordinary differential operator or a partial differential operator, is **linear** if

$$\boxed{L[\alpha u + \beta v] = \alpha L[u] + \beta L[v]} \tag{4}$$

for any functions u and v in the domain of L and for any constants α and β; otherwise it is **nonlinear**.

EXAMPLE 1. The operator L given by (3) is linear because

$$\begin{aligned} L[\alpha u + \beta v] &= \left(\frac{\partial^2}{\partial x^2} - 5\frac{\partial}{\partial y} + 1\right)(\alpha u + \beta v) \\ &= \alpha(u_{xx} - 5u_y + u) + \beta(v_{xx} - 5v_y + v) \\ &= \alpha L[u] + \beta L[v]. \quad \blacksquare \end{aligned}$$

EXAMPLE 2. The operator defined by

$$L[u] = u_{xx} + uu_y$$

is nonlinear because the difference

$$\begin{aligned} &L[\alpha u + \beta v] - \alpha L[u] - \beta L[v] \\ &\quad = (\alpha u_{xx} + \beta v_{xx}) + (\alpha u + \beta v)(\alpha u_y + \beta v_y) - \alpha(u_{xx} + uu_y) - \beta(v_{xx} + vv_y) \\ &\quad = (\alpha^2 - \alpha)uu_y + (\beta^2 - \beta)vv_y + \alpha\beta(uv_y + u_y v) \end{aligned}$$

is not identically zero. For instance, if $u(x, y) = y$, $v(x, y) = 0$, $\alpha = 2$, and $\beta = 6$, then the right-hand side is $3y$, not zero. $\blacksquare$

If an operator L is linear, then it follows immediately from (4) that

$$L[\alpha_1 u_1 + \cdots + \alpha_k u_k] = \alpha_1 L[u_1] + \cdots + \alpha_k L[u_k] \tag{5}$$

for any functions $u_1, \ldots, u_k$ in the domain of L and for any constants $\alpha_1, \ldots, \alpha_k$, for any finite k.

Consider a linear differential equation

$$L[u] = f, \tag{6}$$

where f is a prescribed function of the independent variables. If $f = 0$ then (6) is **homogeneous**, and if $f \neq 0$ then (6) is **nonhomogeneous** with f as a forcing function. Any solution of the full equation $L[u] = f$ is called a **particular solution** of (6). The power of the linearity property (5) is that it enables us to build a more robust solution from a collection of individual solutions by superposition. For suppose that $u_1, \ldots, u_k$ are solutions of the homogeneous version $L[u] = 0$ of (6), and that u_p is a particular solution of (6). Then

$$u = C_1 u_1 + \cdots + C_k u_k + u_p \tag{7}$$

is a solution of (6) for any constants $C_1, \ldots, C_k$ since

$$\begin{aligned} L[C_1 u_1 + \cdots + C_k u_k + u_p] &= C_1 L[u_1] + \cdots + C_k L[u_k] + L[u_p] \\ &= C_1(0) + \cdots + C_k(0) + f \\ &= f. \end{aligned} \tag{8}$$

We say that (7) is robust in the sense that it contains k arbitrary constants, which are available to help u to satisfy the various boundary conditions that may be prescribed.

However, there is a major difference in the application of the foregoing idea to ODE's and PDE's. If (6) is a linear kth-order ODE and $u_1, \ldots, u_k$ are linearly independent solutions of the homogeneous equation $L[u] = 0$, then (7) is a *general* solution of (6). For the PDE's that we study in these chapters we will be able to find even an infinite number of solutions $u_1, u_2, \ldots$, yet

$$u = u_p + \sum_{j=1}^{\infty} C_j u_j \tag{9}$$

may fall short of being a general solution of (6). That is, (6) may admit solutions that cannot be expressed in the form (9) for any choice of the C_j's. Though it would be nice to obtain general solutions of our PDE's (and indeed we will for the wave equation in Chapter 19) our objective is more limited than that. Specifically, we will be content to be able to obtain solutions satisfying specific boundary conditions, and it will turn out that the infinite series solutions that we develop will indeed be robust enough for that, even if they fall short of being general solutions.

18.2.2. Second-order linear equations and their classification. We are especially, though not exclusively, interested in linear second-order PDE's of the form*

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = f, \tag{10}$$

*We assume that A, B, C are not all zero for then (10) would not be of second order. Further, we assume that if $A = B = 0$ then $D \neq 0$, for if $A = B = D = 0$ then only y derivatives appear and (10) would more reasonably be regarded as an ODE rather than as a PDE. Similarly, we assume that if $B = C = 0$ then $E \neq 0$.

where $A, \ldots, F, f$ are prescribed functions of x and y and where the 2 is included for subsequent convenience. Often $A, \ldots, F$ are constants, but not always. The independent variables x and y will be Cartesian space variables, or else x will be a Cartesian space variable and y will be the time – in which case we will use t in place of the generic y.

We classify (10) as one of three types, depending on the sign of the discriminant $B^2 - AC$: (10) is

$$\boxed{\begin{array}{ll} \textbf{parabolic} & \text{if } B^2 - AC = 0, \\ \textbf{hyperbolic} & \text{if } B^2 - AC > 0, \\ \textbf{elliptic} & \text{if } B^2 - AC < 0 \end{array}} \tag{11}$$

in the region under consideration. This terminology is by analogy with the general equation of a conic section, $ax^2 + 2bxy + cy^2 + dx + ey + f = 0$, which gives a parabola, a hyperbola, or an ellipse, according to the sign of the discriminant $b^2 - ac$. Just as parabolas, hyperbolas, and ellipses are governed by distinct geometrical theories, so are parabolic, hyperbolic, and elliptic PDE's governed by distinct theories. Prototype examples of these three types are as follows:

1. The **diffusion equation**

$$\alpha^2 u_{xx} = u_t \qquad (\alpha^2 = \text{constant}) \tag{12}$$

is **parabolic** (with $y \to t$) since $B^2 - AC = 0^2 - (\alpha^2)(0) = 0$.

2. The **wave equation**

$$c^2 u_{xx} = u_{tt} \qquad (c^2 = \text{constant}) \tag{13}$$

is **hyperbolic** (with $y \to t$) since $B^2 - AC = 0^2 - (c^2)(-1) = c^2 > 0$.

3. The **Laplace equation**

$$u_{xx} + u_{yy} = 0 \tag{14}$$

is **elliptic** since $B^2 - AC = 0^2 - (1)(1) = -1 < 0$.

Thus, keep in mind in Chapters 18–20 that the diffusion, wave, and Laplace equations are not only of importance in their own right, but that they are also representative of the three equation types – parabolic, hyperbolic, and elliptic, respectively. For instance, the PDE

$$\alpha^2 u_{xx} = u_t + V u_x + H u \tag{15}$$

governs the diffusion of heat in a one-dimensional rod, but differs from the basic diffusion equation (12) by virtue of the $V u_x$ term (which is due to the rod being in motion with constant speed) and the Hu term (which is due to heat loss from the lateral surface of the rod to the environment). However, (15) is still parabolic, like (12), since $B^2 - AC = 0 - (\alpha^2)(0) = 0$, so its solution should in fundamental ways

be similar to the solution of (12) that is discussed in this chapter. Such variations from the basic diffusion, wave, or Laplace equation, are considered in a number of the end-of-section exercises.

EXAMPLE 3. Since $A, \ldots, F, f$ in (10) may be functions of x and y, the discriminant $B^2 - AC$ may be a function of x and y. Thus, besides the possibility $B^2 - AC$ is zero, positive, or negative everywhere in the x, y plane, it is also possible that it is zero, positive, or negative in different parts of the x, y plane. To illustrate, consider the **Tricomi equation**

$$u_{xx} + xu_{yy} = 0, \tag{16}$$

which arises in the study of the two-dimensional steady transonic flow past a body such as a wing. (Transonic means that flight speed is close to the speed of sound.) Then

$$B^2 - AC = 0 - (1)(x) = -x$$

so the Tricomi equation is elliptic in the right half plane $x > 0$ and hyperbolic in the left half plane $x < 0$. Thus the Tricomi equation is a **change-of-type** equation, with solutions that are qualitatively different in the two half planes. For ODE's, an analogous behavior is exhibited by the **Airy equation**

$$y'' + xy = 0, \tag{17}$$

which has oscillatory solutions for $x > 0$ and nonoscillatory solutions for $x < 0$.

The moral of this example is that a given example of equation (10) might be of one type in one region and of another type in another region. Such cases are more difficult and are not studied here. ∎

18.2.3. Diffusion equation and modeling. In Example 3 of Section 16.8 we derive the heat equation $\alpha^2\nabla^2 u = u_t$ by considering a heat balance for an arbitrary control volume and using the divergence theorem. (There we used T for the temperature field, but here we use u, since we would like to have the letter T available for another purpose.) In slight contrast with the arbitrary control volume approach, engineering textbooks usually consider infinitesimal elements in such derivations. Let us present such a derivation for a one-dimensional rod and, at the same time, include the additional effects of translation of the rod and heat loss to the environment from its lateral surface.

Specifically, consider a uniform rod of cross-sectional area A, circumference s, mass density σ (mass per unit volume), thermal conductivity k, and specific heat c, and suppose it is in uniform motion with speed v in the positive x direction. Consider an element of the rod between x and $x + \Delta x$, where the x axis is fixed in space and the rod is moving relative to it. Assume, merely for definiteness and with no loss of generality, that the derivative u_x is positive at x and at $x + \Delta x$. Then, according to the Fourier law of heat conduction (Example 3, Section 16.8), heat enters the element through its right-hand face at the rate $kAu_x|_{x+\Delta x}$ and leaves through its left-hand face at the rate $kAu_x|_x$ (Fig. 1).

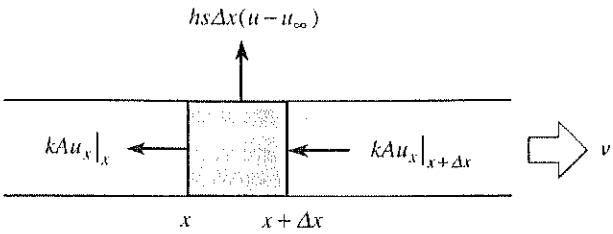


Figure 1. Heat balance for infinitesimal element.

Additionally, suppose there is a heat loss to the environment from the exposed lateral surface of the element, that is proportional to the surface area $s\Delta x$ and the temperature difference $u - u_\infty$, where u_∞ is the temperature of the environment (which we assume to be constant). If the constant of proportionality is some known heat transfer coefficient h, then by *Newton's law of cooling* the rate of heat loss is $hs\Delta x(u - u_\infty)$ (Fig. 1). Finally, there is a transport of heat, in at x and out at $x + \Delta x$, because the rod is translating. Recall that the heat contained in a mass m at (absolute) temperature u is mcu. In time Δt the mass m entering at the left and leaving at the right is the dimension $v\Delta t$ times the area A times the density σ or, on a per unit time basis, $vA\sigma$. Thus, the heat in at the left, per unit time, is $vA\sigma cu|_x$, and the heat out at the right, per unit time, is $vA\sigma cu|_{x+\Delta x}$ (not shown in Fig. 1). Then the net heat influx into the element, per unit time, is

$$kAu_x\Big|_{x+\Delta x} - kAu_x\Big|_x - hs\Delta x(u - u_\infty) + vA\sigma cu\Big|_x - vA\sigma cu\Big|_{x+\Delta x}. \tag{18}$$

The latter must equal the rate of change of the heat mcu contained in the element, where $m = A\Delta x\sigma$, so

$$kA\left(u_x\Big|_{x+\Delta x} - u_x\Big|_x\right) - hs\Delta x(u-u_\infty) - vA\sigma c\left(u\Big|_{x+\Delta x} - u\Big|_x\right) = \frac{\partial}{\partial t}(A\Delta x\sigma cu). \tag{19}$$

Dividing through by $A\Delta x\sigma c$ and letting $\Delta x \to 0$ gives the desired field equation

$$\frac{k}{c\sigma}u_{xx} - \frac{hs}{Ac\sigma}(u - u_\infty) - v\,u_x = u_t \tag{20}$$

or, more concisely,

$$\alpha^2 u'_{xx} = u'_t + vu'_x + Hu', \tag{21}$$

where $\alpha^2 = k/(c\sigma)$ is the thermal diffusivity of the material, v is the translational speed, $H = hs/(Ac\sigma)$ is proportional to the heat transfer coefficient h, and $u' = u - u_\infty$ is the temperature at any point in the rod *relative* to the ambient temperature u_∞.* If we remember that u' is the relative temperature we can, for notational

*Some common values of α^2 (cm^2/sec) are: silver, 1.70; copper, 1.14; aluminum, 0.86; cast iron, 0.16; brick, 0.0052; glass, 0.0034; and water, 0.0014.

convenience, drop the primes in (21). Of course, (21) is the same as (15), mentioned above.

The derivation given above is only heuristic and is typical of the elemental approach to deriving various field equations, as it is found in engineering science textbooks. But it is not hard to render the derivation rigorous. First, the $A\Delta x\sigma cu$ on the right-hand side of (19) should really have been an integral over the element, but

$$\frac{\partial}{\partial t}\int_x^{x+\Delta x} A\sigma cu(\xi,t)\,d\xi = \frac{\partial}{\partial t}A\sigma cu(x+\mu\Delta x,t)\Delta x \sim A\Delta x\sigma cu_t(x,t) \quad (22)$$

as $\Delta x \to 0$, so we obtain the same result as before. The equality in (22) follows from the mean value theorem of the integral calculus, for which we merely need $u(\xi,t)$ to be a continuous function of x, where μ is some value between 0 and 1. Similarly, the $hs\Delta x(u-u_\infty)$ term in (19) should actually be an integral over the lateral surface, but once again we can use the mean value theorem and obtain the same result as before.

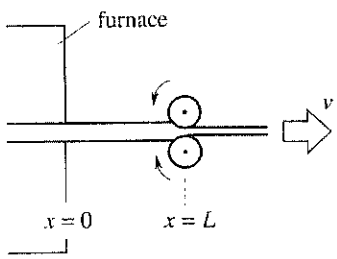


Figure 2. Extrusion.

The Vu_x term in (15) is called a **convection** term, and the Hu term is called a **Newton cooling** term; of course, it will be a heating term rather than a cooling term if the environment is hotter than the rod. As a physical application where both terms would be important, consider a hot continuous metal rod being drawn from a furnace and entering an extrusion die at some distance from the furnace (Fig. 2). In designing the facility it is important to predict the temperature of the rod when it reaches the die, as a function of the various design parameters, so that the metal is still sufficiently hot to be extruded.

Let us set $V = 0$ and $H = 0$ and limit our subsequent attention to the basic one-dimensional diffusion equation (12). As a typical application of (12), consider the heat conduction through the outer wall of a house (Fig. 3), and let x, y, z be normal, horizontal, and vertical axes, respectively, with $x > L$ corresponding to the interior of the house and $x < 0$ corresponding to the exterior. Actually, we should be working with the *three*-dimensional diffusion equation

$$\alpha^2(u_{xx}+u_{yy}+u_{zz}) = u_t. \quad (23)$$

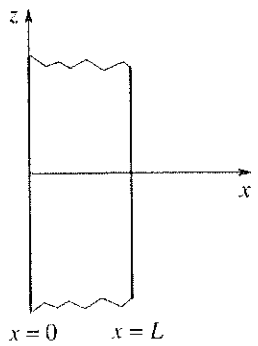


Figure 3. Conduction through a wall.

However, on a cold (or hot) day we expect u to vary very little with y and z compared to its variation with x (over $0 < x < L$). Thus, as a reasonable *approximation* we can neglect the u_{yy} and u_{zz} terms, in which case (23) simplifies to the one-dimensional equation

$$\alpha^2 u_{xx} = u_t \quad (24)$$

or, in operator form,

$$L[u] = \left(\alpha^2\frac{\partial^2}{\partial x^2} - \frac{\partial}{\partial t}\right)[u] = \alpha^2 u_{xx} - u_t = 0. \quad (25)$$

To complete the problem formulation it is useful to observe that the domain of interest is the semi-infinite strip $0 < x < L$ and $0 < t < \infty$ in the x, t plane,

as shown in Fig. 4. Clearly, (25) admits an infinite number of solutions such as $4x$, $5x - 30$, $\sin x \exp(-\alpha^2 t)$, $x + 20 - \cos 3x \exp(-9\alpha^2 t)$, and so on. Our expectation, if only intuitive at this stage, is that if we append to (25) a suitable set of boundary conditions, then the resulting problem will have a unique solution. To motivate such a set of conditions, observe that (25) is a second-order equation with respect to x, so we expect two x boundary conditions to be appropriate, one at $x = 0$ and one at $x = L$. If we call the outside temperature u_1, and call the inside temperature u_2, which is the temperature setting on our thermostat, then the boundary conditions on x are

$$u(0,t) = u_1, \qquad (0 < t < \infty) \tag{26a}$$

$$u(L,t) = u_2. \qquad (0 < t < \infty) \tag{26b}$$

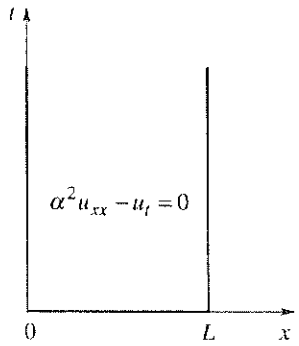


Figure 4. The x, t plane.

Further, (25) is a first-order equation with respect to t, so we expect one t boundary condition to be appropriate, at $t = 0$ (i.e., an initial condition). If we prescribe some initial temperature distribution

$$u(x,0) = f(x), \qquad (0 < x < L) \tag{27}$$

then the full problem statement consists of (25)–(27) and is summarized in Fig. 5. We solve that problem for $u(x,t)$ in Section 18.3, and also establish its uniqueness.

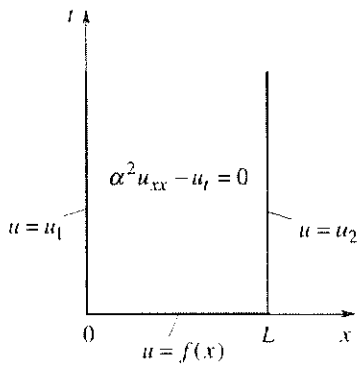


Figure 5. The problem.

Thus, in formulating the problem the idea is to impose boundary and initial conditions that are sufficient to reduce the solution set to a unique solution, but not excessive so that there is *no* solution. For instance, if we impose not only the conditions shown in Fig. 5 but also the condition $u_x(0,t) = 0$, then that problem will have no solution. Whereas existence and uniqueness questions were prominent in our study of ODE's, here we are not so concerned about the existence question, because we will generally be able to *find* a solution. The nagging question that remains, then, is uniqueness. Thus, representative uniqueness theorems are presented in these chapters.

Besides thinking of our problem (25)–(27) as governing heat conduction in a wall, we can think of it as governing heat conduction in a rod that is thermally insulated everywhere on its lateral surface – but not at its two ends, which are subjected to temperatures u_1 and u_2 for all $0 < t < \infty$ (Fig. 6).*

Before closing, let us consider the possible boundary conditions at $x = 0$ and $x = L$ more fully. We distinguish three types. Conditions (26a,b) are examples of **boundary conditions of the first kind**, or **Dirichlet boundary conditions**, because they are of the form

$$u \text{ prescribed} \tag{28}$$

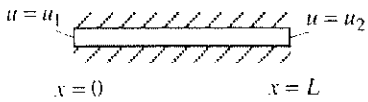


Figure 6. Insulated rod.

on the boundary. In (26a,b) u is prescribed to be a constant, but Dirichlet boundary conditions need not be constant. For instance, if the time of interest is short compared to a day, then it may well suffice to take $u(0,t) = u_1$ to be a constant, but if

*In fact, even for the wall shown in Fig. 3, any "pencil" of material, parallel to the x axis and extending from $x = 0$ to $x = L$, is essentially an insulated rod, insulated by virtue of the lack of temperature variation with y and z.

the time of interest is on the order of a day or longer, then we really need to take $u(0,t) = u_1(t)$ to be a function of time.

A **boundary condition of the second kind**, or **Neumann boundary condition**, is a derivative boundary condition of the form

$$u_n \text{ prescribed,} \tag{29}$$

where u_n denotes the derivative $\partial u/\partial n$ of u normal to the boundary under consideration. In the present example u_n is $-u_x$ on the $x = 0$ boundary and $+u_x$ on the $x = L$ boundary. Physically, (29) amounts to prescribing the *heat flux* rather than the temperature. For instance, the heat flux $Q(t)$ crossing the left end of the rod shown in Fig. 6, counted positive if it flows from left to right is, according to Fourier's law of heat conduction,

$$Q(t) = -kAu_x(0,t). \tag{30}$$

Thus, if we specify the normal derivative $u_x(0,t)$ we are, in effect, specifying the heat flux $Q(t)$. We see from (30) that a homogeneous Neumann boundary condition

$$u_x(0,t) = 0 \tag{31}$$

[or $u_x(L,t) = 0$] amounts to a stipulation that that end is thermally *insulated*, for then $Q(t) = 0$.

Finally, a **boundary condition of the third kind**, or **Robin boundary condition**, occurs when a linear combination of u and u_n is prescribed. To illustrate, within the context of the present example, consider the heat flux crossing the end $x = L$, say. According to the Fourier law of conduction the flux crossing $x = L$ from the left is $-kAu_x(L,t)$, and according to Newton's law of cooling the flux crossing $x = L$ into the environment is $hA[u(L,t) - u_2]$. Since these must be equal we have the boundary condition

$$-kAu_x(L,t) = hA[u(L,t) - u_2] \tag{32}$$

or

$$u_x(L,t) + \frac{h}{k}\,u(L,t) = -\frac{h}{k}\,u_2, \tag{33}$$

which is a boundary condition of Robin type, also called a "mixed" boundary condition. It is useful to nondimensionalize terms in this equation. Nondimensionalizing u with respect to the reference temperature u_2 and x with respect to the reference length L, the nondimensional quantities are

$$\overline{x} = x/L \qquad \text{and} \qquad \overline{u} = u/u_2. \tag{34}$$

Then (32) becomes $-\dfrac{kAu_2}{L}\,\overline{u}_{\overline{x}}\Big|_{\overline{x}=1} = hAu_2\left[\overline{u}\Big|_{\overline{x}=1} - 1\right]$ or

$$-\overline{u}_{\overline{x}}\Big|_{\overline{x}=1} = \text{Bi}\left[\overline{u}\Big|_{\overline{x}=1} - 1\right], \tag{35}$$

where the dimensionless parameter

$$\text{Bi} = \frac{hL}{k} \tag{36}$$

is known in heat transfer as the *Biot number*. If Bi $\gg 1$, then (35) implies that $\overline{u}\Big|_{\overline{x}=1} - 1 \approx 0$ or, returning to dimensional terms,

$$u(L,t) = u_2, \tag{37}$$

which was the boundary condition naively adopted in (26b). However, if the values of h, L, and k give Bi $\ll 1$, then (35) implies that $\overline{u}_{\overline{x}}\Big|_{\overline{x}=1} \approx 0$ or, in dimensional terms,

$$u_x(L,t) = 0. \tag{38}$$

If Bi is neither very large nor very small, then we should leave the mixed boundary condition (33) intact.

With the foregoing remarks completed we will simply specify boundary conditions such as (37), (38), or (33) without further discussion as we now turn our attention to the solution of such problems.

Closure. Keep in mind, as we embark on our study of PDE's in Chapters 18–20, that we are focusing on the extremely important class of PDE's of the form

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = f, \tag{39}$$

where $A, \ldots, F, f$ are prescribed functions of x and y. Within that class we distinguish three types. Specifically, we say that (39) is parabolic if $B^2 - AC = 0$, hyperbolic if $B^2 - AC > 0$, and elliptic if $B^2 - AC < 0$. Representative of these types are the diffusion, wave, and Laplace equations, respectively, and we will devote one chapter to each of these types.

Since (39) is of second order, appropriate boundary conditions will involve u and possibly first-order derivatives of u. Specifically, if u, u_n, or a linear combination of u and u_n are prescribed, then we say that the boundary condition is of the first kind (Dirichlet type), second kind (Neumann type), or third kind (Robin type), respectively. Of these, Robin boundary conditions are the most difficult.

EXERCISES 18.2

1. Show that (5) follows from (4). HINT: Use mathematical induction.

2. Show whether the equation is linear or nonlinear: k, α, and β are constants. HINT: See Examples 1 and 2.

(a) Helmholtz equation. $\nabla^2 u + k^2 u = 0$

(b) Korteweg-de Vries (KdV) equation, $u_t + \alpha u u_x + \beta u_{xxx} = 0$

(c) biharmonic equation. $\nabla^4 u = u_{xxxx} + 2u_{xxyy} + u_{yyyy} = 0$

(d) Tricomi equation. $u_{xx} + xu_{yy} = 0$

(e) $u_{xx} + u_{yy} = e^u$

(f) $x^4 u_{xx} = u_y$

(g) $u_{xx} + 5u_{xy} - xu = e^x$
(h) $xu_x + uu_y = 6$

3. Classify the following PDE's, defined over $-\infty < x < \infty$ and $-\infty < y < \infty$, as elliptic, parabolic, or hyperbolic. If the equation is of mixed type, identify the relevant regions and give the classification within each region.

(a) $u_{xx} + u_{xy} - x^2 u_y = e^{xy}$
(b) $xu_{xx} - u_{xy} + yu_{yy} + 3u_y = 1$
(c) $u_{xy} + u_x - 4u_y = 6u$
(d) $xu_{xx} - (\sin^2 y + 1)u_{yy} = x^2 u$
(e) $u_{xx} + u_{xy} + u_{yy} + u_x + u_y + u = 1$
(f) $u_{xx} + (\cos x)u_{yy} = 2xy$
(g) $u_{xx} + u_x + u_y = x^3 u$
(h) $u_{xy} - u_{yy} + e^x u = f(x, y)$

4. In deriving the diffusion equation (21), we assumed that the cross-sectional shape of the rod does not vary with x. Reconsider our derivation for the basic case where there is no Newton cooling (i.e., the lateral surface is insulated, $h = 0$) and no translation of the rod ($v = 0$), but allow for the cross-sectional area A to vary with x. Show that the revised diffusion equation is

$$\frac{\alpha^2}{A(x)}[A(x)u_x]_x = u_t. \tag{4.1}$$

18.3 Separation of Variables

18.3.1. The method of separation of variables. We explain the method of separation of variables by a sequence of examples.

EXAMPLE 1. Consider the diffusion problem

$$L[u] = \alpha^2 u_{xx} - u_t = 0, \qquad (0 < x < L,\ 0 < t < \infty) \tag{1a}$$
$$u(0,t) = u_1, \quad u(L,t) = u_2, \qquad (0 < t < \infty) \tag{1b}$$
$$u(x,0) = f(x), \qquad (0 < x < L) \tag{1c}$$

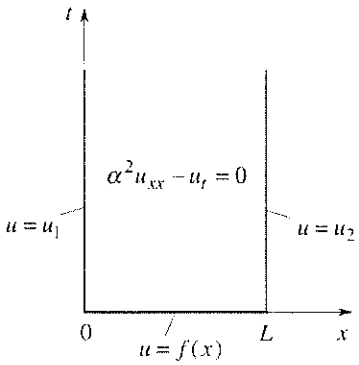


Figure 1. The problem (1).

that is derived in Section 18.2 and that governs the temperature field $u(x,t)$ in a rod with insulated lateral surface (or in a wall or slab of thickness L); see Fig. 1.

According to the method of **separation of variables** we begin by seeking solutions of (1a) in the product form

$$\boxed{u(x,t) = X(x)T(t).} \tag{2}$$

Putting (2) into (1a) gives

$$\alpha^2 X''T = XT', \tag{3}$$

where primes denote ordinary differentiation. To separate the variables, divide both sides of (3) by XT and obtain

$$\frac{X''}{X} = \frac{1}{\alpha^2}\frac{T'}{T}. \tag{4}$$

Actually, we divided by α^2 as well, but whether we have $1/\alpha^2$ on the right-hand side of (4) or α^2 on the left-hand side will not affect the final result. Observe that (4) is of the form

$$F(x) = G(t). \tag{5}$$

Since x and t are independent variables, $G(t)$ does not vary with x. But $F(x) = G(t)$, so $F(x)$ does not vary with x either. Hence $F(x)$ is a constant. From (5), $G(t)$ is a constant also, the same constant.* Thus,

$$\frac{X''}{X} = \frac{1}{\alpha^2}\frac{T'}{T} = \text{constant} = -\kappa^2, \tag{6}$$

say, where we have written κ^2 for convenience because we will soon need to take the square root of that quantity. Motivation for including the minus sign in (6) is given in the end-of-example comments.

The beauty of the separation procedure is that in place of the *partial* differential equation $\alpha^2 u_{xx} = u_t$ we now have two *ordinary* differential equations,

$$\frac{X''}{X} = -\kappa^2 \qquad \text{and} \qquad \frac{1}{\alpha^2}\frac{T'}{T} = -\kappa^2,$$

or,

$$X'' + \kappa^2 X = 0, \tag{7a}$$

$$T' + \kappa^2\alpha^2 T = 0. \tag{7b}$$

The *separation constant* κ remains to be determined. Solving (7a,b) gives

$$X = A\cos\kappa x + B\sin\kappa x, \tag{8a}$$

$$T = Ce^{-\kappa^2\alpha^2 t}. \tag{8b}$$

However, observe that (8a) is the general solution of (7a) only in the event that $\kappa \neq 0$, for if $\kappa = 0$ then the $\sin\kappa x$ term drops out. Since we do not yet know the value(s) of κ, we must allow for the possibility that the value $\kappa = 0$ will be needed. Setting $\kappa = 0$ in (7a) gives $X'' = 0$, with the general solution $X = D + Ex$, so we replace (8a) by the two-tier statement

$$X = \begin{cases} A\cos\kappa x + B\sin\kappa x, & \kappa \neq 0 \\ D + Ex, & \kappa = 0. \end{cases} \tag{9}$$

Apparently, we don't need to revise (8b) the way we revised (8a) because (8b) is a general solution of (7b) whether κ is zero or not. However, having already comitted ourselves in (9) to the separate treatment of these two cases, we replace (8b) by the two-tier statement †

$$T = \begin{cases} Fe^{-\kappa^2\alpha^2 t}, & \kappa \neq 0 \\ G, & \kappa = 0. \end{cases} \tag{10}$$

Thus far we have discussed the product solutions

$$u = XT = (D + Ex)G \tag{11}$$

*An alternative approach that you might prefer is to take $\partial/\partial x$ of (5), which step gives $F'(x) = 0$, so $F(x) =$ constant. From (5), $G(t)$ must also be constant, the same constant. [Or, $\partial/\partial t$ of (5) gives $G'(t) = 0$, so $G(t) =$ constant, and so on.] However, this approach is a bit weaker because it requires an assumption that X''/X is differentiable, an assumption that is not needed.

†$F\exp(-\kappa^2\alpha^2 t)$ can be expressed as $F\exp(-t/\tau)$, where $\tau = 1/(\kappa^2\alpha^2)$ is a **time constant**, namely, the time it takes the exponential to decay by 63% (from 1 to e^{-1}).

corresponding to $\kappa = 0$, and

$$u = XT = (A\cos\kappa x + B\sin\kappa x)Fe^{-\kappa^2\alpha^2 t} \tag{12}$$

for any $\kappa \neq 0$. Since D, E, G, A, B, F are arbitrary constants, we can combine DG as H and EG as I and simplify (11) as

$$u = H + Ix$$

for $\kappa = 0$, and we can combine AF as J and BF as K and simplify (12) as

$$u = (J\cos\kappa x + K\sin\kappa x)e^{-\kappa^2\alpha^2 t}$$

for $\kappa \neq 0$. Since (1a) is linear, the sum of these solutions must also be a solution, so we can write

$$u = H + Ix + (J\cos\kappa x + K\sin\kappa x)e^{-\kappa^2\alpha^2 t}, \tag{13}$$

where the constants H, I, J, K, and κ are arbitrary. But it is understood that $\kappa \neq 0$ in (13) because the $H + Ix$ part of (13) already accounts for the $\kappa = 0$ case. By the linearity of (1a) we can superimpose any number of such terms for different values of κ. With $\kappa = 1, 2$, and $\sqrt{5}$, for instance, we can write

$$\begin{aligned} u(x,t) = (H + Ix) &+ (J_1\cos x + K_1\sin x)e^{-\alpha^2 t} \\ &+ (J_2\cos 2x + K_2\sin 2x)e^{-4\alpha^2 t} + (J_3\cos\sqrt{5}x + K_3\sin\sqrt{5}x)e^{-5\alpha^2 t}. \end{aligned} \tag{14}$$

The latter expression satisfies (1a) because each term does, and because (1a) is linear. [We urge you to verify that (14) satisfies (1a), by direct substitution.] But since we do not yet know what κ values to choose, let us continue with the more compact form (13).

We are ready to apply the boundary conditions (1b) and the initial condition (1c). Beginning with the left end condition, (13) gives

$$u(0,t) = u_1 = H + Je^{-\kappa^2\alpha^2 t} \qquad (0 < t < \infty)$$

or, in a more suggestive form,

$$(H - u_1)(1) + J(e^{-\kappa^2\alpha^2 t}) = 0. \qquad (0 < t < \infty) \tag{15}$$

Since the functions 1 and $\exp(-\kappa^2\alpha^2 t)$ are linearly independent on the t interval,* it follows from (15) that we need $H - u_1 = 0$ (so $H = u_1$) and $J = 0$. Updating (13) to incorporate these results we have, thus far,

$$u(x,t) = u_1 + Ix + K\sin\kappa x\, e^{-\kappa^2\alpha^2 t}. \tag{16}$$

Applying the right end condition next, (16) gives

$$u(L,t) = u_2 = u_1 + IL + K\sin\kappa L\, e^{-\kappa^2\alpha^2 t}$$

or

$$(IL + u_1 - u_2)(1) + K\sin\kappa L\, e^{-\kappa^2\alpha^2 t} = 0. \tag{17}$$

*Recall that *two* functions are linearly dependent if and only if one is a scalar multiple of the other, and neither 1 nor $\exp(-\kappa^2\alpha^2 t)$ is a scalar multiple of the other.

Again invoking the linear independence of 1 and $\exp(-\kappa^2\alpha^2 t)$, it follows from (17) that

$$IL + u_1 - u_2 = 0 \tag{18a}$$

and

$$K \sin \kappa L = 0. \tag{18b}$$

Equation (18a) gives $I = (u_2 - u_1)/L$, but (18b) presents a choice: either $K = 0$ or $\sin \kappa L = 0$ (or both). Here, and at analogous points in examples to follow, the rule is to *make that choice so as to maintain as robust a solution as possible* [because we still have the initial condition $u(x, 0) = f(x)$ to satisfy, and we will need all the help we can get]. If we choose $K = 0$, then we lose the $\sin \kappa x \exp(-\kappa^2\alpha^2 t)$ term in (16) and are left with

$$u(x, t) = u_1 + (u_2 - u_1)\frac{x}{L}, \tag{19}$$

which is capable of satisfying the initial condition only in the unlikely event that $f(x)$ happens to be $u_1 + (u_2 - u_1)(x/L)$. However, if we choose

$$\sin \kappa L = 0, \tag{20}$$

then K need not be zero, we retain the $\sin \kappa x \exp(-\kappa^2\alpha^2 t)$ term in (16), and (20) serves to identify the allowable values of κ, namely,

$$\kappa = \frac{n\pi}{L} \tag{21}$$

for $n = 0, \pm 1, \pm 2, \ldots$.* Of these values we discard $n = 0$ because it gives $\kappa = 0$, whereas it was understood that the κ's in (16) were to be nonzero. Further, $n = -1, -2, \ldots$ can be discarded since the $K \sin(n\pi x/L) \exp[-(n\pi\alpha/L)^2 t]$ combination in (16) is insensitive to the sign of n – to within a factor of ± 1, which factor can be absorbed by K anyhow.†

Using superposition as in (14) but for the values $\kappa = n\pi/L$ ($n = 1, 2, \ldots$), (16) gives

$$u(x, t) = u_1 + (u_2 - u_1)\frac{x}{L} + \sum_{n=1}^{\infty} K_n \sin \frac{n\pi x}{L}\, e^{-(n\pi\alpha/L)^2 t}. \tag{22}$$

Before completing the solution let us review where we stand. The right-hand side of (22) satisfies the boundary conditions (1b) because $\sin(n\pi x/L) = 0$ at $x = 0$ and at $x = L$, for each $n = 1, 2, \ldots$. Further, it appears to satisfy the PDE $L[u] = 0$ because it is a linear combination of product solutions

$$u_1 + (u_2 - u_1)\frac{x}{L}, \quad \sin \frac{\pi x}{L}\, e^{-(\pi\alpha/L)^2 t}, \quad \sin \frac{2\pi x}{L}\, e^{-(2\pi\alpha/L)^2 t}, \quad \ldots,$$

each of which satisfies $L[u] = 0$. That is, if $L[\phi_1(x, t)] = 0, \ldots, L[\phi_k(x, t)] = 0$, then $L[\alpha_1\phi_1(x, t) + \cdots + \alpha_k\phi_k(x, t)] = 0$ too because

$$L[\alpha_1\phi_1 + \cdots + \alpha_k\phi_k] = \alpha_1 L[\phi_1] + \cdots + \alpha_k L[\phi_k] \tag{23}$$

*Surely, $\sin z = 0$ has the roots $z = n\pi$ on the real axis. In fact, even if we broaden the search and look in the complex z plane, we find only the roots $n\pi$ on the real axis.

†Put differently, observe that only κ^2 appears in (7a) and (7b). Since positive and negative values of κ are therefore indistinguishable in the ODE's, their general solutions for a given positive κ, say $\kappa = \kappa_0$, must be the same as their general solutions for $\kappa = -\kappa_0$.

by virtue of the linearity of L. However, k in (23) is finite, whereas k in (22) is infinite, so the step

$$L\left[\sum_{n=1}^{\infty} \alpha_n \phi_n\right] = \sum_{n=1}^{\infty} \alpha_n L[\phi_n] \tag{24}$$

amounts to an interchange in the order of two limit operations, the derivative in L and the infinite series, and that step needs to be rigorously justified. This point of rigor comes up in Chapter 17 when we use Fourier series solution forms to satisfy ODE's (e.g., Comment 2 of Example 3 in Section 17.3), and is addressed in the optional Section 17.5. We address it here too, in the optional subsection 18.3.2. Generally, however, in Chapters 18–20, we omit rigorous justification and are satisfied with **formal solutions**.

Finally, we impose the initial condition (1c) on (22):

$$u(x,0) = f(x) = u_1 + (u_2 - u_1)\frac{x}{L} + \sum_{n=1}^{\infty} K_n \sin\frac{n\pi x}{L}. \qquad (0 < x < L) \tag{25}$$

Two questions arise: given $f(x)$, is it *possible* to find coefficients K_n so as to satisfy (25) on $0 < x < L$ and, if it is possible, then *how* do we determine the K_n's? First, put all the known terms on the left-hand side:

$$f(x) - u_1 - (u_2 - u_1)\frac{x}{L} = \sum_{n=1}^{\infty} K_n \sin\frac{n\pi x}{L}, \qquad (0 < x < L) \tag{26}$$

and let us denote $f(x) - u_1 - (u_2 - u_1)(x/L)$ as $F(x)$, for brevity. Observing that

$$F(x) = \sum_{n=1}^{\infty} K_n \sin\frac{n\pi x}{L} \qquad (0 < x < L) \tag{27}$$

is of the form of a half-range sine expansion of F, we can conclude that if F is sufficiently well-behaved,* then (27) is indeed possible, and [according to (4) in Section 17.4] the K_n's are computed as

$$K_n = \frac{2}{L}\int_0^L F(x)\sin\frac{n\pi x}{L}\,dx. \tag{28}$$

Thus, our formal solution consists of (22), with the K_n's computed from (28).

To illustrate, consider a 10 cm-long copper rod held in boiling water until its temperature is 100° C throughout. At time $t = 0$ it is removed and its ends are quenched with ice for all $t > 0$. (We can either neglect heat loss from its lateral surface or specify that that surface be insulated.) Then $\alpha^2 = 1.14\,\text{cm}^2/\text{sec}$ (for copper), $L = 10\,\text{cm}$, $u_1 = u_2 = 0$, and $F(x) = 100°$ C, so (28) gives

$$K_n = \frac{2}{10}\int_0^{10} 100\sin\frac{n\pi x}{10}\,dx = \begin{cases} \dfrac{400}{n\pi}, & n \text{ odd} \\ 0, & n \text{ even} \end{cases} \tag{29}$$

*Let $F_{\text{ext}}(x)$ be the extended version of $F(x)$ that is $2L$-periodic and symmetric about $x = 0$ and $x = L$. If $F_{\text{ext}}(x)$ and $F'_{\text{ext}}(x)$ are piecewise continuous on $[-L, L]$, then (27) converges in the sense of Theorem 17.3.1.

and (22) becomes

$$u(x,t) = \frac{400}{\pi} \sum_{n=1,3,\ldots}^{\infty} \frac{1}{n} \sin \frac{n\pi x}{10} e^{-0.1125n^2 t}. \tag{30}$$

We have summed the series in (30) at a number of x's and at several representative times and have plotted the results in Fig. 2.

Since this is our first example there are numerous points to clarify, and we address them in the following comments.

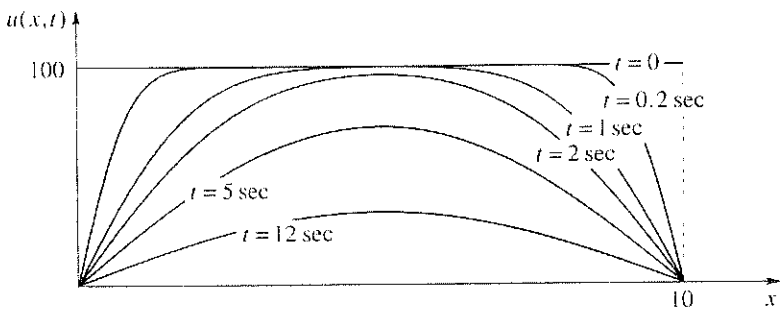


Figure 2. $u(x,t)$ for the case where $\alpha^2 = 1.14$, $L = 10$, $u_1 = u_2 = 0$, and $f(x) = 100$.

COMMENT 1. Notice the physical significance of the terms in (22). As $t \to \infty$ the exponential terms tend to zero, leaving the **steady-state** solution

$$\lim_{t\to\infty} u(x,t) \equiv u_s(x) = u_1 + (u_2 - u_1)\frac{x}{L}, \tag{31}$$

which is a linear variation from u_1 at $x = 0$ to u_2 at $x = L$ (Fig. 3). Thus, the summation term in (22) represents the **transient** part of the solution, that links the initial distribution $u(x,0) = f(x)$ to the steady-state distribution $u_s(x)$. For the specific case represented in Fig. 2, the steady-state solution is simply $u_s(x) = 0$ because $u_1 = u_2 = 0$.

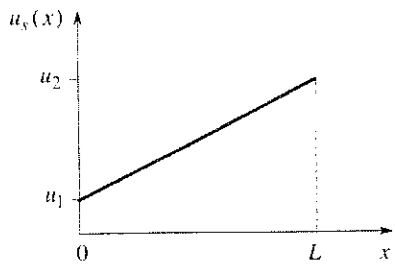


Figure 3. Steady-state.

COMMENT 2. Notice further that within the transient part of (22) each term dies out exponentially faster (as $t \to \infty$) than the preceding term because of the n^2 in the exponent. To illustrate the significance of that point let us write out (30) at the representative times $t = 0.2$, 1, and 12 seconds:

$$u(x,0.2) = 124.5 \sin \frac{\pi x}{10} + 34.7 \sin \frac{3\pi x}{10} + 14.5 \sin \frac{5\pi x}{10} + \cdots,$$
$$u(x,1) = 113.8 \sin \frac{\pi x}{10} + 15.4 \sin \frac{3\pi x}{10} + 1.5 \sin \frac{5\pi x}{10} + \cdots,$$
$$u(x,12) = 33.0 \sin \frac{\pi x}{10} + 0.0002 \sin \frac{3\pi x}{10} + 6\times 10^{-14} \sin \frac{5\pi x}{10} + \cdots,$$

which results are among those shown in Fig. 2. Computationally speaking, then, (30) [more generally, (22)] is an excellent result for "large" times because if t is large enough then only one or two of the terms in the series are needed for engineering accuracy. Conversely, for "small" times a great many terms in the series may be needed. For instance, at $t = 0$ the

terms in (30) die out only like $1/n$ because the exponential factor is unity at $t = 0$.

COMMENT 3. Why did we write $-\kappa^2$ in (16), rather than $+\kappa^2$? That is, how did we anticipate that the separation constant would be negative? The rule of thumb is to choose either sign and then examine the resulting ODE's for clues. Specifically, the minus sign chosen in (6) causes plus signs in (7a) and (7b), both of which look good: the plus sign in (7a) looks good because it results in sine and cosine solutions, which will be needed for the eventual Fourier series expansion of $f(x) - u_s(x)$, and the plus sign in (7b) looks good because it results in exponential decay rather than physically unreasonable exponential growth. (See Exercise 2.)

COMMENT 4. It is essential to *apply the boundary conditions before the initial condition*. To see why, let us try imposing the initial condition first, instead. Then, (13) gives

$$u(x,0) = f(x) = H + Ix + J\cos\kappa x + K\sin\kappa x,$$

which cannot be satisfied unless $f(x)$ happens to be a linear combination of 1, x, $\cos\kappa x$, and $\sin\kappa x$, for some κ. Applying the boundary conditions first enabled us to obtain the solution form (22), which form was powerful enough to handle any given initial condition $u(x,0) = f(x)$. This sequencing – boundary conditions first and initial condition second – will be appropriate all through Chapters 18 and 19.

COMMENT 5. As emphasized in Section 18.2.1, we did *not* find a *general solution* of the diffusion equation and then apply the boundary and initial conditions. Rather, we used the method of separation of variables to develop a solution that was sufficiently robust to handle the boundary conditions and initial condition.

COMMENT 6. It is interesting that the assumed product form $u(x,t) = X(x)T(t)$ seems a bad choice because it maintains the same shape, $X(x)$, for all time and is merely scaled in magnitude by the time-varying factor $T(t)$. Rather, we would expect the shape of $u(x,t)$ to change with time – as in Fig. 2, for instance, where $u(x,t)$ is initially a constant but approaches a sinusoidal shape as heat diffuses out of the rod at both ends due to the end conditions $u(0,t) = u(L,t) = 0$. However, understand that the *superposition* of product solutions is not itself a product solution; that is, the final solution (22) is *not* a function of x times a function of t.

COMMENT 7. Still retaining the boundary conditions $u(0,t) = u(L,t) = 0$, what would happen if we changed the initial condition from $u(x,0) = 100$ to

$$u(x,0) = \begin{cases} 100, & 0 < x < x_0 \text{ and } x_0 < x < L \\ 5\times 10^6, & x = x_0 \end{cases} \tag{32}$$

that is, if we changed the initial temperature at a single point x_0 to $5{,}000{,}000°$? Nothing; the solution would still be given by (30), because a change in the integrand of (28) at a single point, from one finite value to another, cannot change the value of the integral.

COMMENT 8. The solution of (1) is not necessarily an infinite series. For instance, suppose that $u_1 = u_2 = 0$ and

$$u(x,0) = f(x) = 40\sin\frac{\pi x}{L}. \tag{33}$$

Then, application of the initial condition (33) to (22) gives

$$40 \sin \frac{\pi x}{L} = \sum_{n=1}^{\infty} K_n \sin \frac{n\pi x}{L}$$
$$= K_1 \sin \frac{\pi x}{L} + K_2 \sin \frac{2\pi x}{L} + \cdots,$$

which is satisfied, as can be seen by inspection, by setting $K_1 = 40$ and all the other K_n's equal to zero. Thus, we obtain the one-term solution

$$u(x,t) = 40 \sin \frac{\pi x}{L} e^{-(\pi\alpha/L)^2 t}. \tag{34}$$

We can display (34) in two dimensions by plotting u versus x at representative times, as we have done in Fig. 4a [and as we did earlier in Fig. 2 for the case where $f(x)$ was 100], or in three dimensions by plotting the u surface above the x, t plane, as we have done in Fig. 4b. Similarly, if

$$u(x,0) = f(x) = 30 \sin \frac{2\pi x}{L} - 25 \sin \frac{5\pi x}{L}, \tag{35}$$

then

$$u(x,t) = 30 \sin \frac{2\pi x}{L} e^{-(2\pi\alpha/L)^2 t} - 25 \sin \frac{5\pi x}{L} e^{-(5\pi\alpha/L)^2 t}, \tag{36}$$

and so on. ∎

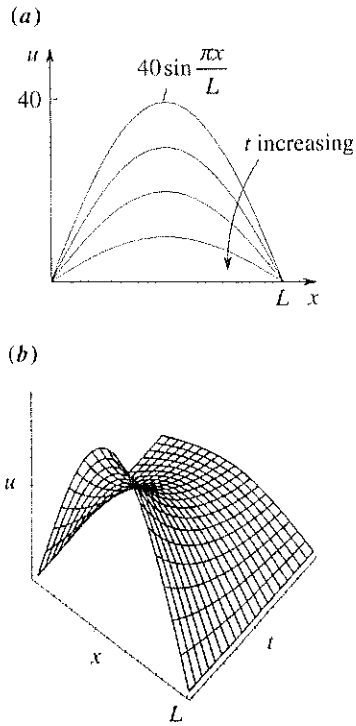


Figure 4. Graph of (34).

In dwelling on so many details, in Example 1, we have shown no mercy. However, Example 1 provides the basic model for the application of the method of separation of variables, as it is used here in Chapters 18–20, so careful study of that example is well worth the effort.

EXAMPLE 2. *Different Boundary Conditions.* This time suppose that the left end of the rod is insulated, and the right end is held at a constant temperature 100 for all $t > 0$, so the problem is as follows:

$$L[u] = \alpha^2 u_{xx} - u_t = 0, \qquad (0 < x < L,\ 0 < t < \infty) \tag{37a}$$
$$u_x(0,t) = 0, \quad u(L,t) = 100, \qquad (0 < t < \infty) \tag{37b}$$
$$u(x,0) = f(x), \qquad (0 < x < L) \tag{37c}$$

and suppose f is the piecewise constant function

$$f(x) = \begin{cases} 60, & 0 < x < L/2 \\ 0, & L/2 < x < L \end{cases} \tag{38}$$

shown in Fig. 5.

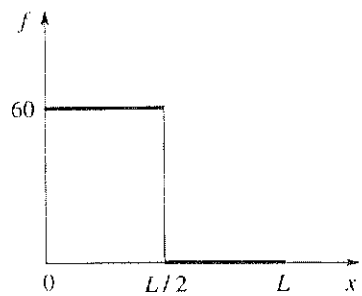


Figure 5. $f(x)$.

Up until (13) the story is unchanged, so let us begin with

$$u(x,t) = H + Ix + (J \cos \kappa x + K \sin \kappa x) e^{-\kappa^2 \alpha^2 t}. \tag{39}$$

Applying the left-hand boundary condition,

$$u_x(x,t) = I + (-\kappa J \sin \kappa x + \kappa K \cos \kappa x)e^{-\kappa^2\alpha^2 t},$$

so

$$u_x(0,t) = 0 = I + \kappa K e^{-\kappa^2\alpha^2 t}, \tag{40}$$

which gives $I = K = 0$. Updating (39) accordingly,

$$u(x,t) = H + J \cos \kappa x \, e^{-\kappa^2\alpha^2 t}. \tag{41}$$

Next, the right-hand boundary condition gives

$$u(L,t) = 100 = H + J \cos \kappa L \, e^{-\kappa^2\alpha^2 t}, \tag{42}$$

so $H = 100$ and $J \cos \kappa L = 0$. We cannot afford to satisfy the latter by setting $J = 0$ because if $J = 0$ then we lose the $\cos \kappa L e^{-\kappa^2\alpha^2 t}$ term in (41), and the latter reduces to $u(x,t) = H = 100$, which cannot satisfy the initial condition. Rather, we satisfy $J \cos \kappa L = 0$ by setting

$$\cos \kappa L = 0 \tag{43}$$

and letting J remain arbitrary. Thus, $\kappa L = \pi/2, 3\pi/2, 5\pi/2, \ldots$, so

$$\kappa = \frac{n\pi}{2L}, \qquad (n = 1, 3, \ldots) \tag{44}$$

and we have

$$u(x,t) = 100 + \sum_{n=1,3,\ldots}^{\infty} J_n \cos \frac{n\pi x}{2L} \, e^{-(n\pi\alpha/2L)^2 t}. \tag{45}$$

Finally, the initial condition is

$$u(x,0) = f(x) = 100 + \sum_{n=1,3,\ldots}^{\infty} J_n \cos \frac{n\pi x}{2L}$$

or

$$f(x) - 100 = \sum_{n=1,3,\ldots}^{\infty} J_n \cos \frac{n\pi x}{2L}. \qquad (0 < x < L) \tag{46}$$

Comparing the form of the right-hand side of (46) with the half- and quarter-range cosine and sine expansion formulas, we see that (46) amounts to a quarter-range cosine expansion of $f(x) - 100$, so

page 844

$$\begin{aligned} J_n &= \frac{2}{L} \int_0^L [f(x) - 100] \cos \frac{n\pi x}{2L} \, dx \\ &= \frac{2}{L} \left[\int_0^{L/2} (-40) \cos \frac{n\pi x}{2L} \, dx + \int_{L/2}^{L} (-100) \cos \frac{n\pi x}{2L} \, dx \right] \\ &= \frac{80}{n\pi} \left(3 \sin \frac{n\pi}{4} - 5 \sin \frac{n\pi}{2} \right). \end{aligned} \tag{47}$$

The solution is given by (45) and (47).

To plot the results, let the material be glass (so $\alpha^2 = 0.0034\,\text{cm}^2/\text{sec}$), and let $L = 10\,\text{cm}$. In Fig. 6 we plot u versus x at representative times.

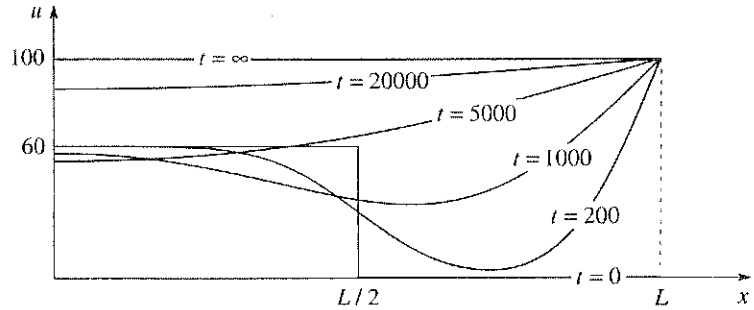


Figure 6. Graph of (45) for $\alpha^2 = 0.0034$ and $L = 10$.

COMMENT 1. Notice that each graph of u is flat at $x = 0$, in accord with the boundary condition $u_x(0,t) = 0$, and that the steady-state solution $u_s(x) = 100$ is approached as $t \to \infty$. That the initial temperature $u(x,0) = f(x)$ also satisfies the boundary condition $u_x(0,t) = 0$ is only by coincidence and is not required since we require the boundary conditions to hold only for $0 < t < \infty$, not for $0 \le t < \infty$.

COMMENT 2. You might be tempted to break (46) into two parts,

$$-40 = \sum_{n=1,3,\ldots}^{\infty} J_n \cos\frac{n\pi x}{2L} \qquad \left(0 < x < \frac{L}{2}\right) \tag{48a}$$

and

$$-100 = \sum_{n=1,3,\ldots}^{\infty} J_n \cos\frac{n\pi x}{2L}, \qquad \left(\frac{L}{2} < x < L\right) \tag{48b}$$

and to use the quarter-range cosine expansion formulas to solve for the J_n's in each case. That procedure would be INCORRECT for two reasons. First, it will give J_n's that have different values in the intervals $0 < x < L/2$ and $L/2 < x < L$, in which case they will be functions of x, not constants – as they were supposed to be. Second, we cannot use the quarter-range cosine formulas to solve (48a) and (48b) for the J_n's because $\sum_{n=1,3,\ldots}^{\infty} J_n \cos(n\pi x/2L)$ is a quarter-range cosine expansion on $0 < x < L$, not on $0 < x < L/2$ or on $L/2 < x < L$. ∎

In Example 1, $u(0,t)$ and $u(L,t)$ are prescribed constants and we end up expanding $f(x) - u_s(x)$ in a half-range sine series. In Example 2, $u_x(0,t)$ and $u(L,t)$ are prescribed constants and we use a quarter-range cosine series. You will find that if instead $u(0,t)$ and $u_x(L,t)$ are prescribed constants, then a quarter-range sine series will be appropriate. If $u_x(0,t)$ and $u_x(L,t)$ are prescribed constants that are equal, then a half-range cosine series will be appropriate, but if they are unequal then the solution is more difficult (Exercise 19).

18.3.2. Verification of solution. (Optional) As stated, we can claim only to have found *formal* solutions of the diffusion problems in Examples 1 and 2. To illustrate the verification process let us verify that the formal solution (22) of the problem (1), with the K_n's given by (28), does satisfy the requirements in (1).

First let us show that (22) satisfies the PDE $\alpha^2 u_{xx} = u_t$. If we differentiate (22) termwise we obtain

$$u_x = \frac{u_2 - u_1}{L} + \sum_{n=1}^{\infty} \frac{n\pi}{L} K_n \cos \frac{n\pi x}{L} e^{-(n\pi\alpha/L)^2 t}, \tag{49a}$$

$$u_{xx} = -\sum_{n=1}^{\infty} \left(\frac{n\pi}{L}\right)^2 K_n \sin \frac{n\pi x}{L} e^{-(n\pi\alpha/L)^2 t}, \tag{49b}$$

$$u_t = -\sum_{n=1}^{\infty} \left(\frac{n\pi\alpha}{L}\right)^2 K_n \sin \frac{n\pi x}{L} e^{-(n\pi\alpha/L)^2 t}, \tag{49c}$$

so we see that $\alpha^2 u_{xx}$ does equal u_t, provided that we can rigorously justify the termwise differentiations that produced (49a,b,c). Theorem 17.5.2 tells us that those steps are permissible if the resulting series [on the right-hand sides of (49a,b,c)] converge *uniformly* on the problem domain $0 < x < L$, $0 < t < \infty$, and Theorem 17.5.1 gives us the Weierstrass M-test as a test for uniform convergence. To apply the latter to the series in (49a), note that

$$\left| \frac{n\pi}{L} K_n \cos \frac{n\pi x}{L} e^{-(n\pi\alpha/L)^2 t} \right| \le Q n e^{-(n\pi\alpha/L)^2 t_0} \equiv M_n \tag{50}$$

for all $t > t_0$ and $0 < x < L$ because $|\cos(n\pi x/L)| \le 1$ and $K_n \to 0$ as $n \to \infty$,* so that the K_n's must be bounded: thus there must exist a finite positive constant Q such that $|\pi K_n/L| \le Q$ for all n's. It follows easily from the ratio test that $\sum_{n=1}^{\infty} M_n$ converges, so the series in (49a) does indeed converge uniformly in $0 < x < L$, $t_0 \le t < \infty$, for arbitrarily small t_0. Similarly for the series in (49b) and (49c), the only difference being that those series contain a factor of n^2 rather than n; but the ratio test shows that $\sum_{n=1}^{\infty} n^2 \exp[-(n\pi\alpha/L)^2 t_0]$ converges, just as $\sum_{n=1}^{\infty} n \exp[-(n\pi\alpha/L)^2 t_0]$ does.

Turning to the boundary conditions (1b) and the initial condition (1c), it might appear that the satisfaction of these conditions does not even need verification. For instance, is not

$$u(0,t) = u_1 + 0 + \sum_{n=1}^{\infty} 0 = u_1$$

obviously true? The point to emphasize is that in posing PDE problems we require satisfaction of the PDE in the *open* region, in this case $0 < x < L$, $0 < t < \infty$. We do that so that we can allow for not-so-well-behaved boundary and initial data. For if $\alpha^2 u_{xx} = u_t$ were to be satisfied in the region $0 \le x \le L$, $0 \le t < \infty$, then we would need $u(x,0) = f(x)$ to be twice differentiable, and $u(0,t) = g(t)$

*See the last sentence of Exercise 1, Section 17.6.

and $u(L,t) = h(t)$, say, to be once differentiable, whereas we want to allow for boundary and initial data that are not even continuous, let alone differentiable. In Example 2, for instance, $f(x)$ is discontinuous. The way we link the solution $u(x,t)$ in the open domain to the boundary and initial conditions is through limits. That is, by (1b) we mean

$$\lim_{x\to 0+} u(x,t) = u_1 \quad \text{and} \quad \lim_{x\to L-} u(x,t) = u_2, \qquad (0 < t < \infty) \tag{51}$$

and by (1c) we mean

$$\lim_{t\to 0+} u(x,t) = f(x). \qquad (0 < x < L) \tag{52}$$

To verify that (22) satisfies (51) we can use the following theorem.

THEOREM 18.3.1 *Continuity of Sum Function*
If $\sum_{n=1}^{\infty} a_n(x)$ converges uniformly to $s(x)$ on some x interval I and the $a_n(x)$'s are continuous on I, then $s(x)$ is continuous on I.

That is, if the convergence is uniform, then the continuity of the partial sums is passed on to the sum function $s(x)$.

Applying that result, observe first that the series in (22) converges uniformly on $0 \le x \le L$ for each t such that $0 < t_0 \le t < \infty$ because there is a finite constant P such that

$$\left| K_n \sin \frac{n\pi x}{L} e^{-(n\pi\alpha/L)^2 t} \right| \le P e^{-(n\pi\alpha/L)^2 t_0} \equiv M_n$$

there, and $\sum_{n=1}^{\infty} M_n$ is convergent. Thus the right-hand side of (22) is a continuous function of x on $0 \le x \le L$, so its values at $x = 0$ and at $x = L$, namely u_1 and u_2, respectively, are the same as their limiting values as $x \to 0+$ and as $x \to L-$, respectively.

Verification of the initial condition (52) can be accomplished with the help of a theorem of Abel. For that, and a generally more detailed discussion, we refer you to Churchill and Brown.*

Finally, there is the question of uniqueness: is (22) the *only* solution of (1)? A formal proof of uniqueness is outlined in Exercise 25; for detailed discussion we refer you, again, to Churchill and Brown.

18.3.3. Use of Sturm–Liouville theory. (Optional) In subsection 18.3.1 we found that the final step of the separation-of-variables solution involves the expansion of a given function. For the diffusion equation $\alpha^2 u_{xx} = u_t$ with Dirichlet or Neumann boundary conditions that expansion was a half- or quarter-range cosine or

*R. V. Churchill and J. W. Brown, *Fourier Series and Boundary Value Problems*, 3rd ed. (New York: McGraw Hill, 1978, p. 129).

sine series, so we were guided by our knowledge of such series. More generally, the necessary expansion is an *eigenfunction expansion* in terms of the orthogonal eigenfunctions of a Sturm–Liouville problem that is "built in." Thus, a more powerful approach is to appeal to the Sturm–Liouville theory, which includes the half- and quarter-range expansions as special cases. Let us illustrate with three examples.

EXAMPLE 3. *Example 1 Reconsidered.* The Sturm–Liouville theory becomes relevant only when we reach the point of needing to carry out the expansion of a given function. Thus, in reconsidering Example 1 in the light of the Sturm–Liouville theory we can begin with equation (26). Our claim is that the $\sin(n\pi x/L)$ functions in (26) are the orthogonal eigenfunctions of a Sturm–Liouville problem governing $X(x)$, namely, the equation (7a) together with the homogeneous boundary conditions $X(0) = 0$ and $X(L) = 0$:

$$X'' + \kappa^2 X = 0, \qquad (0 < x < L) \tag{53a}$$

$$X(0) = 0, \quad X(L) = 0, \tag{53b}$$

which problem is indeed of the Sturm–Liouville form

$$(py')' + qy + \lambda wy = 0, \qquad (a < x < b) \tag{54a}$$

$$\alpha y(a) + \beta y'(a) = 0, \quad \gamma y(b) + \delta y'(b) = 0, \tag{54b}$$

where $y(x)$ is $X(x)$, $p(x) = w(x) = 1$, $q(x) = 0$, $\lambda = \kappa^2$, $a = 0$, $b = L$, $\alpha = \gamma = 1$, and $\beta = \delta = 0$. Considering the boundary conditions $u(0,t) = u_1$ and $u(L,t) = u_2$, where did we get $X(0) = 0$ and $X(L) = 0$, in (53b)? Recall that the $u_1 + (u_2 - u_1)x/L$ terms in (22) make up the steady-state solution $u_s(x)$. The burden of satisfying the nonhomogeneous boundary conditions, $u(0,t) = u_1$ and $u(L,t) = u_2$, is carried by the steady-state solution $u_s(x)$, so the transient part, $\sum_{n=1}^{\infty} K_n \sin\frac{n\pi x}{L} \exp[-(n\pi\alpha/L)^2 t]$, must be *zero* at $x = 0$ and $x = L$. Thus, the $X(x) = \sin(n\pi x/L)$ eigenfunctions, contained therein, actually satisfy the *homogeneous conditions* (53b), as is easily verified by evaluating $\sin(n\pi x/L)$ at $x = 0$ and $x = L$.

According to the Sturm–Liouville theory, then, (27) can indeed be satisfied, and the expansion coefficients are given by

$$K_n = \frac{\langle F(x), \sin\frac{n\pi x}{L}\rangle}{\langle \sin\frac{n\pi x}{L}, \sin\frac{n\pi x}{L}\rangle} = \frac{\int_0^L F(x)\sin\frac{n\pi x}{L}\,dx}{\int_0^L \sin^2\frac{n\pi x}{L}\,dx}$$

$$= \frac{2}{L}\int_0^L F(x)\sin\frac{n\pi x}{L}\,dx, \tag{55}$$

which result is the same as (28). [Recall that the weight function, in the two integrals, in this case is $w(x) = 1$.] ∎

Thus, the idea is that satisfaction of the initial condition $u(x,0) = f(x)$ requires the expansion of $F(x) = f(x) - u_s(x)$. The latter will inevitably be an eigenfunction expansion in terms of the eigenfunctions $\phi_n(x)$ [namely, $\sin(n\pi x/L)$

in Example 3] of a Sturm–Liouville problem on $X(x)$. The Sturm–Liouville theory assures us that the desired expansion is possible, and it tells us how to compute the expansion coefficients – namely, as $\langle F(x), \phi_n\rangle / \langle \phi_n, \phi_n\rangle$, where the inner product has weight function $w(x)$.

In Example 3 we were able to use either the half-range sine expansion concept *or* the Sturm–Liouville theory. In the next example the expansion will not be of half- or quarter-range type, so we will *have to* use the Sturm–Liouville theory.

EXAMPLE 4. This time consider the problem

$$L[u] = u_{xx} - u_t = 0, \qquad (0 < x < 1,\ 0 < t < \infty) \tag{56a}$$

$$u(0,t) - 2u_x(0,t) = 5, \quad u(1,t) = 35, \qquad (0 < t < \infty) \tag{56b}$$

$$u(x,0) = f(x), \qquad (0 < x < 1) \tag{56c}$$

where we have set $\alpha^2 = 1$ and $L = 1$ for simplicity. Observe that $u(0,t) - 2u_x(0,t) = 5$ is a Robin boundary condition, a boundary condition of the third kind.

Separating variables as usual, let us begin with equation (13):

$$u(x,t) = H + Ix + (J\cos\kappa x + K\sin\kappa x)e^{-\kappa^2 t}. \tag{57}$$

Applying the left end condition gives

$$\begin{aligned} u(0,t) - 2u_x(0,t) = 5 &= H + Je^{-\kappa^2 t} - 2(I + \kappa K e^{-\kappa^2 t}) \\ &= (H - 2I) + (J - 2\kappa K)e^{-\kappa^2 t}, \qquad (0 < t < \infty) \end{aligned}$$

so

$$H - 2I = 5, \tag{58a}$$

$$J - 2\kappa K = 0. \tag{58b}$$

And the right end condition gives

$$u(1,t) = 35 = H + I + (J\cos\kappa + K\sin\kappa)e^{-\kappa^2 t}, \qquad (0 < t < \infty)$$

so

$$H + I = 35, \tag{59a}$$

$$J\cos\kappa + K\sin\kappa = 0. \tag{59b}$$

Equations (58a) and (59a) give $H = 25$ and $I = 10$. Equations (58b) and (59b) give the unique trivial solution $J = K = 0$, unless we choose κ so that the determinant of the coefficient matrix vanishes:

$$\begin{vmatrix} 1 & -2\kappa \\ \cos\kappa & \sin\kappa \end{vmatrix} = 0. \tag{60}$$

We cannot accept the trivial solution $J = K = 0$ because it reduces (57) to $u(x,t) = 25 + 10x$, which does satisfy the PDE and boundary conditions but which cannot satisfy

the initial condition $u(x,0) = f(x)$ unless $f(x)$ happens to be $25 + 10x$. Thus, we impose the determinant condition (60), or

$$\tan\kappa = -2\kappa, \tag{61}$$

and designate the positive roots of (61) as $\kappa_1, \kappa_2, \ldots$. Next, we need to *find* the corresponding nontrivial solutions of (58b) and (59b) for J and K. With (61) satisfied, (58b) and (59b) are redundant, so we can drop (59b) and use (58b) to obtain $J = 2\kappa K$.

With these results, and superposition, (57) gives

$$u(x,t) = 25 + 10x + \sum_{n=1}^{\infty} K_n \phi_n(x) e^{-\kappa_n^2 t}, \tag{62}$$

where

$$\phi_n(x) = 2\kappa_n \cos\kappa_n x + \sin\kappa_n x. \tag{63}$$

Finally, the initial condition requires that

$$u(x,0) = f(x) = 25 + 10x + \sum_{n=1}^{\infty} K_n \phi_n(x),$$

or,

$$F(x) = \sum_{n=1}^{\infty} K_n \phi_n(x), \qquad (0 < x < 1) \tag{64}$$

where $F(x) = f(x) - (25 + 10x)$, $25 + 10x$ being the steady-state solution $u_s(x)$. The ϕ_n's in (64) are the eigenfunctions generated by the Sturm–Liouville problem

$$X'' + \kappa^2 X = 0, \qquad (0 < x < 1) \tag{65a}$$

$$X(0) - 2X'(0) = 0, \quad X(1) = 0, \tag{65b}$$

with weight function $w(x) = 1$, so

$$K_n = \frac{\langle F, \phi_n \rangle}{\langle \phi_n, \phi_n \rangle} = \frac{\displaystyle\int_0^1 F(x)(2\kappa_n \cos\kappa_n x + \sin\kappa_n x)\,dx}{\displaystyle\int_0^1 (2\kappa_n \cos\kappa_n x + \sin\kappa_n x)^2\,dx}. \tag{66}$$

In fact, this Sturm–Liouville problem, including the determination of the κ_n's, is the subject of Example 3 in Section 17.7. ∎

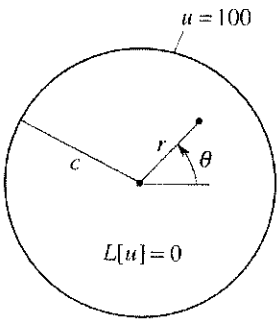


Figure 7. Conduction in a disk.

EXAMPLE 5. *Unsteady Conduction in a Disk.* Consider the unsteady conduction of heat in a circular disk such as a coin of radius c, the flat faces of which are insulated (Fig. 7). Then the temperature field $u(r,t)$ in the disk is governed by the problem

$$\begin{gathered} \alpha^2 \nabla^2 u = \alpha^2 \left(u_{rr} + \frac{1}{r} u_r + \frac{1}{r^2} u_{\theta\theta} \right) = u_t, \qquad (0 \le r < c, \;\; 0 < t < \infty) \\ u(c,t) = 100, \quad u(r,0) = f(r). \end{gathered} \tag{67}$$

That is, the disk is initially at a prescribed temperature $f(r)$, and then we hold the outer edge (with boiling water, for example) at $u = 100°$ for all $t > 0$. Because the domain is circular, and the initial and boundary temperature are independent of θ, it appears that the resulting temperature field u will be independent of θ as well, and will be a function only of r and t. Hence, we can strike out the $u_{\theta\theta}$ term, and the PDE reduces to

$$\alpha^2(u_{rr} + \frac{1}{r}u_r) = u_t. \tag{68}$$

To solve by separation of variables, seek u in the product form

$$u(r,t) = R(r)T(t). \tag{69}$$

Putting (69) into the PDE and dividing by RT (and α^2) gives

$$\frac{R'' + \frac{1}{r}R'}{R} = \frac{1}{\alpha^2}\frac{T'}{T} = \text{constant} = -\kappa^2, \tag{70}$$

and hence the ODE's

$$R'' + \frac{1}{r}R' + \kappa^2 R = 0, \tag{71a}$$

$$T' + \kappa^2\alpha^2 T = 0. \tag{71b}$$

Equation (71a) is the subject of Example 1 in Section 4.6, and its general solution is

$$R(r) = AJ_0(\kappa r) + BY_0(\kappa r), \tag{72}$$

where J_0, Y_0 are the Bessel functions of first and second kind, respectively, of order zero (Fig. 8). Are there any values of κ for which (71) fails to provide a general solution? Yes, for $\kappa = 0$ because $Y_0(0) = -\infty$ does not exist. But for $\kappa = 0$ (71a) can be expressed as $(rR')' = 0$, which can be integrated to give $R(r) = C + D\ln r$. Thus, we distinguish the cases $\kappa \neq 0$ and $\kappa = 0$, and write

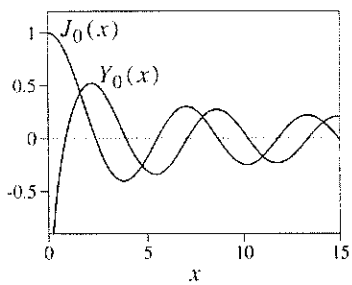


Figure 8. $J_0(x)$ and $Y_0(x)$.

$$R(r) = \begin{cases} AJ_0(\kappa r) + BY_0(\kappa r), & \kappa \neq 0 \\ C + D\ln r, & \kappa = 0 \end{cases} \tag{73}$$

$$T(t) = \begin{cases} Ee^{-\kappa^2\alpha^2 t}, & \kappa \neq 0 \\ F, & \kappa = 0. \end{cases} \tag{74}$$

Thus far, we have

$$\begin{aligned} u(r,t) &= (C + D\ln r)F + [AJ_0(\kappa r) + BY_0(\kappa r)]Ee^{-\kappa^2\alpha^2 t} \\ &= G + H\ln r + [PJ_0(\kappa r) + QY_0(\kappa r)]e^{-\kappa^2\alpha^2 t}, \end{aligned} \tag{75}$$

where we have combined CF as G, DF as H, and so on.

In the preceding examples, $X(x)$ is governed by a second-order ODE and there are two x boundary conditions (at $x = 0$ and at $x = L$). In the present example, likewise, $R(r)$ is governed by a second-order ODE, but we find only one r boundary condition in (67),

$u = 100$ at $r = c$. As a second r boundary condition it seems appropriate to require that u be bounded at $r = 0$.

As in preceding examples, we save the initial condition for last and begin by applying the boundary conditions. Of the two boundary conditions (u bounded at $r = 0$ and $u = 100$ at $r = c$), we recommend applying any boundedness condition first because it gives an immediate simplification. Specifically, for $u(0, t)$ to be bounded we need $H = Q = 0$ in (74) because both $\ln r$ and $Y_0(\kappa r)$ are unbounded at $r = 0$. Then (74) simplifies to

$$u(r, t) = G + PJ_0(\kappa r)e^{-\kappa^2\alpha^2 t}. \tag{76}$$

Next,

$$u(c, t) = 100 = G + PJ_0(\kappa c)e^{-\kappa^2\alpha^2 t},$$

or,

$$(G - 100)(1) + PJ_0(\kappa c)e^{-\kappa^2\alpha^2 t} = 0. \qquad (0 < t < \infty) \tag{77}$$

Since 1 and $\exp(-\kappa^2\alpha^2 t)$ are linearly independent on the t interval, it follows from (77) that $G - 100 = 0$ and $PJ_0(\kappa c) = 0$. The former gives $G = 100$ and the latter gives $P = 0$ or $J_0(\kappa c) = 0$. We reject the choice $P = 0$ because we cannot afford to lose the $PJ_0(\kappa r)\exp(-\kappa^2\alpha^2 t)$ term in (76), and adopt the choice

$$J_0(\kappa c) = 0 \tag{78}$$

with positive roots $\kappa_n c = z_n$ for $n = 1, 2, \ldots$, where the z_n's are the (known) zeros of J_0 (Fig. 9). With these choices, and the help of superposition, we have

$$u(r, t) = 100 + \sum_{n=1}^{\infty} P_n J_0\left(z_n \frac{r}{c}\right) e^{-(z_n\alpha/c)^2 t}. \tag{79}$$

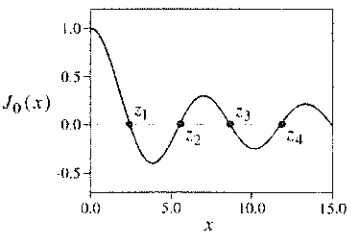


Figure 9. The zeros z_n of $J_0(x)$.

(The P_n's are arbitrary constants, not Legendre polynomials.)

Finally, the initial condition requires that

$$u(r, 0) = f(r) = 100 + \sum_{n=1}^{\infty} P_n J_0\left(z_n \frac{r}{c}\right),$$

or

$$F(r) \equiv f(r) - 100 = \sum_{n=1}^{\infty} P_n J_0\left(z_n \frac{r}{c}\right). \qquad (0 \le r < c) \tag{80}$$

The expansion functions $J_0(z_n r/c)$ are the eigenfunctions of the singular Sturm–Liouville problem

$$\begin{aligned} (rR')' + \kappa^2 rR &= 0, \qquad (0 < r < c) \\ R(0) \text{ bounded}, \quad R(c) &= 0, \end{aligned} \tag{81}$$

which problem is the subject of Example 2 in Section 17.8. Observe that we multiplied (71a) through by r in order to obtain the standard form given in (81). That step is important because it is only when the equation is in the standard form that we can identify from it

the weight function – that will be needed in our inner product. From (81) we see that the weight function is $w(r) = r$, so we can write

$$P_n = \frac{\langle F(r), J_0(z_n r/c)\rangle}{\langle J_0(z_n r/c), J_0(z_n r/c)\rangle} = \frac{\displaystyle\int_0^c F(r) J_0\left(z_n \frac{r}{c}\right) r\,dr}{\displaystyle\int_0^c \left[J_0\left(z_n \frac{r}{c}\right)\right]^2 r\,dr}$$

$$= \frac{2}{c^2[J_1(z_n)]^2} \int_0^c F(r) J_0\left(z_n \frac{r}{c}\right) r\,dr. \tag{82}$$

For explanation of the last step, see Example 2 of Section 17.8.

COMMENT 1. As a concrete example, let $f(r) = 0$. Then (Exercise 31)

$$P_n = -\frac{200}{c^2[J_1(z_n)]^2} \int_0^c J_0\left(z_n \frac{r}{c}\right) r\,dr = -\frac{200}{z_n J_1(z_n)}, \tag{83}$$

so

$$u(r,t) = 100 - 200 \sum_{n=1}^{\infty} \frac{1}{z_n J_1(z_n)} J_0\left(z_n \frac{r}{c}\right) e^{-(z_n \alpha/c)^2 t}. \tag{84}$$

For instance, the temperature history at the center of the disk is

$$u(0,t) = 100 - 200 \sum_{n=1}^{\infty} \frac{1}{z_n J_1(z_n)} e^{-(z_n \alpha/c)^2 t}, \tag{85}$$

which is plotted in Fig. 10 for the case where the material is glass ($\alpha^2 = 0.0034\,\text{cm}^2/\text{sec}$ and $c = 10\,\text{cm}$).

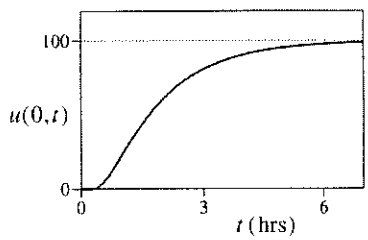


Figure 10. Temperature history at center; $\alpha^2 = 0.0034$, $c = 10$.

COMMENT 2. Observe that the point $r = 0$ is the left endpoint of the r interval $0 \le r < c$, but, physically, it corresponds to an interior point of the disk, the center of the disk. Thus, the PDE must be satisfied there. In particular, if $u_r(0,t)$ is to exist, then it must be zero. That is, the u surface (i.e., the graph of u above the r, θ plane) must be flat at $r = 0$ because otherwise the u surface will be conical there and $u_r(0,t)$ will not exist. Thus, in place of the condition that $R(0)$ be bounded we could have used the condition

$$R'(0) = 0, \tag{86}$$

and the latter would have produced the same results.

COMMENT 3. Recall that we argued that u does not vary with θ in this example. Hence, we dropped the $u_{\theta\theta}$ term in the PDE and solved the reduced equation $\alpha^2(u_{rr} + \frac{1}{r}u_r) = u_t$. If you still have doubts about that step, observe that (79) does satisfy the boundary conditions, the initial condition [if the P_n's are computed using (81)], and the *full* equation $\alpha^2 \nabla^2 u = u_t$ including the $u_{\theta\theta}$ term because $\partial^2/\partial\theta^2$ of the right-hand side of (79) is zero. ∎

Closure. This section covers almost all aspects of the method of separation of variables, which method is used in each of Chapters 18–20. The starting point is to assume a procedure form for the solution, for instance $u(x,t) = X(x)T(t)$ if the

independent variables are x and t. For the diffusion equation, that step enable us to separate the variables and obtain ODE's on $X(x)$ and $T(t)$. That simplification, from a PDE to two ODE's, is the point of the method.

The essential ingredients, for the success of the method in a given application, are that the equation needs to be separable, the boundary and initial conditions need to be given on constant coordinate curves, and the PDE needs to be linear – so that a sufficiently robust solution can be built up by the superposition of various product solutions. Only two PDE's are studied in the foregoing examples, $\alpha^2 u_{xx} = u_t$ and $\alpha^2(u_{rr} + \frac{1}{r}u_r) = u_t$, and these three conditions are met: the equations are separable, conditions are on constant coordinate curves ($x = 0, x = L, t = 0, r = c$, and $r = 0$), and the equations are linear.

In contrast, an example of a PDE that is *not* separable is the two-dimensional **biharmonic equation** in Cartesian coordinates,

$$\nabla^4 u = u_{xxxx} + 2u_{xxyy} + u_{yyyy} = 0 \tag{87}$$

because $u(x, y) = X(x)Y(y)$ gives

$$\frac{X''''}{X} + 2\frac{X''}{X}\frac{Y''}{Y} + \frac{Y''''}{Y} = 0, \tag{88}$$

which cannot be rearranged in the separated form $F(x) = G(y)$ because of the $(X''/X)(Y''/Y)$ term.

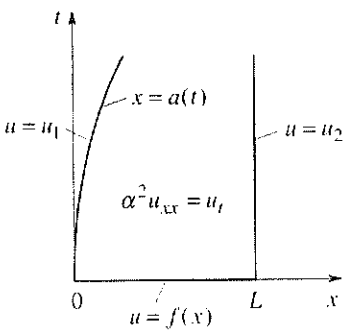


Figure 11. Moving boundary problem.

And as an example where the boundary conditions are not given on constant coordinate curves consider the problem shown in Fig. 11. There, the temperature u_1 is so high that the left end of the rod melts and drips away, so that the boundary condition $u = u_1$ is applied not on the line $x = 0$ but on some curve $x = a(t)$. The latter is an example of a **moving boundary problem**. Generally, such problems defy analytical solution, and we resort to numerical solution techniques – such as the finite difference method that is presented in Section 18.6. An interesting application of the problem shown in Fig. 11 is in the design of a space vehicle that reenters the earth's atmosphere at hypersonic velocity. To protect the craft from the heat thereby generated (remember that meteorites often burn up before reaching the earth's surface) one can design the nose cone to be long enough so that part of it melts away during reentry. A simple one-dimensional model of the heat conduction in the nose cone would be somewhat like the problem shown in Fig. 11.

In this chapter and the next, the sign of the separation constant is always negative, but in Chapter 20, on the Laplace equation, it is negative *or* positive, depending on the specific application.

In subsection 18.3.2 we discuss the rigorous verification of a solution. Normally, our solutions will be only formal, in the sense that such verification will not be carried out.

Finally, in subsection 18.3.3 we show how to use the Sturm–Liouville theory to handle the expansion process that is needed to satisfy the initial condition. Using that theory we are not limited to half- and quarter-range cosine and sine expansions, as we were in subsection 18.3.1. In fact, the half- and quarter-range cases are but special cases of the Sturm–Liouville theory. Remember: Liouville, not Louisville.

Computer software. We can use the *Maple* **sum** command to obtain the infinite series solutions that we generate by separation of variables. As a first illustration, consider the series

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots, \tag{89}$$

which has the sum $\ln 2$ because it is the Taylor series of $\ln(1+x)$ about $x = 1$. To sum the series, enter

$$\text{sum}\,((-1)\hat{}(n+1)/n, n = 1..\text{infinity}\,); \tag{90}$$

and return. The result is $\ln 2$. Remarkably, the software not only gives the numerical value, it gives that value in the closed form $\ln 2$ rather than the open form $0.693147\ldots$.

Next, consider the solution (30) given in Example 1, namely,

$$u(x,t) = \frac{400}{\pi} \sum_{n=1,3,\ldots}^{\infty} \frac{1}{n} \sin \frac{n\pi x}{10}\, e^{-0.1125n^2 t}. \tag{91}$$

To illustrate, let us use the sum command to sum this series at $x = 1$ and $t = 0.2$. First, change the dummy summation index from n to $2n - 1$ so that the new index n runs continuously from 1 to infinity:

$$u(x,t) = \frac{400}{\pi} \sum_{n=1}^{\infty} \frac{1}{2n-1} \sin \frac{(2n-1)\pi x}{10}\, e^{-0.1125(2n-1)^2 t}. \tag{92}$$

If, to compute $u(1, 0.2)$, we enter

$$\begin{aligned} &\text{sum}\,(400/(Pi * (2 * n - 1)) * \sin((2 * n - 1) * \text{Pi}/10) \\ &* \exp(-0.1125 * (2 * n - 1)\hat{}2 * 0.2), n = 1..\text{infinity}); \end{aligned} \tag{93}$$

the computer is unable to obtain the result in closed form and merely prints the series itself, in a form similar to the right-hand side of (92). Thus, we rerun the command (93) with "infinity" changed to a finite number such as 20. The result, 86.13830281, does not change if we increase the upper summation limit from 20 to 30, say. Thus, it is reasonable to assume that the answer $u(1, 0.2) = 86.13830281$ is correct to that many significant figures. Of course, in a practical sense it would be foolish to insist on 10 significant figure accuracy in a problem like this if $\alpha^2 = 0.0034$ and the other data are known only to two or three significant figures.

If we run the sum command from $n = 1$ to $n = 20$, say, and the result is merely a printout of the 20 terms rather than their numerical sum, we can use the evalf(") command to evaluate their sum.

EXERCISES 18.3

1. Verify by direct substitution that (14) does indeed satisfy the diffusion equation (1a).

2. (*On the sign of separation constant*) Writing the separation constant as $-\kappa^2$ in (6), rather than as κ^2, worked out well, and we also give arguments in Comment 3 as to why the separation constant should be negative. In this exercise we explore that point further.

(a) Show that if we use κ^2 in (6), instead of $-\kappa^2$, then we eventually arrive at the same solution as before, given by (22), although it will be a bit more challenging because complex quantities will arise. HINT: Recall the identity $\sin ix = i \sinh x$ (or, $\sinh ix = i \sin x$). You may use the fact that the only roots of the equation $\sin z = 0$, in the complex plane, are the points $z = n\pi$ on the real axis.
(b) Show that if we use $-\kappa^2$ in (6), as we did, then the relevant Sturm–Liouville problem is given by (53). Show that it follows from Theorem 17.7.2 that κ^2 *must* be nonnegative.

3. In Comment 6 at the end of Example 1 we note that although we began by seeking $u(x,t)$ in product form, the final solution (22) is not itself of that form. Give conditions on u_1, u_2, and $f(x)$ in (1) such that the final solution *will* be of product form (so that its graph does not change its shape, with time, though its magnitude may change with time).

4. (*Separation*) Seeking a solution $u(x,t) = X(x)T(t)$ for the given PDE, carry out steps analogous to equations (3)–(6), and derive ODE's analogous to (7a,b). Take the separation constant to be $-\kappa^2$, as we do in (6). Obtain general solutions of those ODE's (distinguishing any special κ values, as necessary) and use superposition to obtain a solution analogous to the solution (13) of (1a). If the PDE cannot be separated, state that.

(a) $u_{xx} = u_t + 3u$
(b) $u_{xx} + 2u_x = u_t$ HINT: In this case you should find that the value of κ that needs to be distinguished [as we distinguished the case $\kappa = 0$ in (9) and (10)] is $\kappa = 1$, not $\kappa = 0$.
(c) $u_{xx} + 2u_{xt} = u_t$
(d) $u_{xx} + 2u_{xt} = u_{tt}$

5. Can we use superposition to conclude from (9) and (10) that

$$X = A\cos\kappa x + B\sin\kappa x + D + Ex,$$
$$T = Fe^{-\kappa^2\alpha^2 t} + G,$$

and

$$u(x,t) = (A\cos\kappa x + B\sin\kappa x + D + Ex)(Fe^{-\kappa^2\alpha^2 t} + G)?$$

Explain.

6. (*Continuation of Examples 1 and 2*) In each case solve (1), with the boundary conditions (1b) changed as indicated, and for the specified $f(x)$. Use a half- or quarter-range cosine or sine expansion, as appropriate. Evaluate the expansion coefficients explicitly, rather than leaving them in integral form. Also, identify the steady-state solution $u_s(x)$.

(a) $u(0,t) = 20,\ u_x(\pi,t) = 3$, (i.e., $L = \pi$), $f(x) = 0$
(b) $u(0,t) = 10,\ u_x(2,t) = -5,\ f(x) = 10$
(c) $u(0,t) = 0,\ u_x(2,t) = 0,\ f(x) = 50\sin(\pi x/2)$
(d) $u(0,t) = 0,\ u_x(2,t) = 0,$
$f(x) = 5\sin(\pi x/4) - 12\sin(5\pi x/4)$
(e) $u(0,t) = 25,\ u_x(4,t) = 0,\ f(x) = 25$
(f) $u(0,t) = 25,\ u_x(2,t) = 0,\ f(x) = 0$ for $0 < x < 1$, $f(x) = 25$ for $1 < x < 2$
(g) $u_x(0,t) = u_x(\pi,t) = 0,\ f(x) = 300$
(h) $u_x(0,t) = u_x(3\pi,t) = 0,\ f(x) = 0$ for $0 < x < 2\pi$, $f(x) = 60$ for $2\pi < x < 3\pi$
(i) $u_x(0,t) = u_x(10,t) = 5,\ f(x) = 45 + 5x$
(j) $u_x(0,t) = u_x(5,t) = 3,\ f(x) = 2x$
(k) $u(0,t) = 0,\ u(5,t) = 0,$
$f(x) = \sin\pi x - 37\sin(\pi x/5) + 6\sin(9\pi x/5)$
(l) $u(0,t) = 0,\ u(10,t) = 100,\ f(x) = 0$
(m) $u_x(0,t) = 2,\ u(6,t) = 12,\ f(x) = 0$
(n) $u_x(0,t) = 0,\ u(6,t) = 0,\ f(x) = \sin x$

7. Use (45) and (47) to compute $u(0,t)$, and plot it versus t. At the least, take t = 1,000, 2,000, 5,000, 10,000, 15,000, 20,000 and 30,000. Recall that $\alpha^2 = 0.0034$ and $L = 10$.

8. We stated in Comment 8 that it can be seen by inspection that $K_1 = 40$ and that all the other K_n's are zero. Alternatively, obtain that same resulting by working out the integral

$$K_n = \frac{2}{L}\int_0^L \left(40\sin\frac{\pi x}{L}\right)\sin\frac{n\pi x}{L}\,dx.$$

9. The temperature distribution $u(x,t)$ in a 2-m long brass rod is governed by the problem

$$\alpha^2 u_{xx} = u_t, \qquad (0 < x < 2,\ 0 < t < \infty)$$
$$u(0,t) = u(2,t) = 0, \qquad (t > 0)$$
$$u(x,0) = \begin{cases} 50x, & (0 < x < 1) \\ 100 - 5x, & (1 < x < 2) \end{cases}$$

where $\alpha^2 = 2.9 \times 10^{-5}$ m^2/sec.

(a) Determine the solution for $u(x,t)$.
(b) Compute the temperature at the midpoint of the rod at the end of 1 hour.
(c) Compute the time it will take for the temperature at that point to diminish to 5° C.
(d) Compute the time it will take for the temperature at that point to diminish to 1°C.

10. (*Steady-state solution*) If the solution $u(x,t)$ tends to a steady-state solution $u_s(x)$ as $t \to \infty$, then we can determine $u_s(x)$ from $u(x,t)$ as

$$u_s(x) = \lim_{t\to\infty} u(x,t). \tag{10.1}$$

However, if we are interested *only* in $u_s(x)$ then it is wasteful to first solve for $u(x,t)$. To solve for $u_s(x)$ directly, merely set $u_t = 0$ in the PDE, which step reduces the PDE to an ODE on $u_s(x)$. Solve that ODE subject to the boundary conditions (which, we assume here, do not vary with t). In Example 1, for instance, $u_s(x)$ is governed by the problem

$$\begin{aligned} &\alpha^2 u_s'' = 0, \qquad (0 < x < L) \\ &u_s(0) = u_1, \quad u_s(L) = u_2, \end{aligned}$$

which boundary-value problem is readily solved, its solution being $u_s(x) = u_1 + (u_2 - u_1)x/L$, as obtained in Example 1 by letting $t \to \infty$ in $u(x,t)$. Use this method to find $u_s(x)$ in each case: u_1, u_2, Q_1, Q_2, V, H are constants, and the initial condition is $u(x,0) = f(x)$.

(a) $\alpha^2 u_{xx} = u_t,\ u(0,t) = u_1,\ u_x(L,t) = Q_2$
(b) $\alpha^2 u_{xx} = u_t,\ u_x(0,t) = Q_1,\ u(L,t) = u_2$
(c) $\alpha^2 u_{xx} = u_t,\ u_x(0,t) = Q_1,\ u_x(L,t) = Q_2$ HINT: Show that $u_s(x)$ does not exist if $Q_2 \neq Q_1$, and explain why that result makes sense in physical terms. Show that if $Q_2 = Q_1 \equiv Q$, however, then $u_s(x)$ does exist but contains an undetermined constant, say C. To determine C, integrate the PDE on x, from 0 to L:

$$\alpha^2 \int_0^L u_{xx}\,dx = \int_0^L u_t\,dx, \tag{10.2}$$

and show that

$$\frac{d}{dt}\int_0^L u(x,t)\,dx = 0, \tag{10.3}$$

so that we have the conservation principle

$$\int_0^L u(x,t)\,dx = \text{constant}. \tag{10.4}$$

Use (10.4) to solve for C, thus completing the solution for $u_s(x)$.

(d) $\alpha^2 u_{xx} = u_t + Hu,\ u(0,t) = u_1,\ u(L,t) = u_2$
(e) $\alpha^2 u_{xx} = u_t + Hu,\ u_x(0,t) = Q_1,\ u(L,t) = u_2$
(f) $\alpha^2 u_{xx} = u_t + Hu,\ u(0,t) = u_1,\ u_x(L,t) = Q_2$
(g) $\alpha^2 u_{xx} = u_t + Hu,\ u_x(0,t) = Q_1,\ u_x(L,t) = Q_2$
(h) $\alpha^2 u_{xx} - Vu_x = u_t,\ u(0,t) = u_1,\ u(L,t) = u_2$
(i) $\alpha^2 u_{xx} - Vu_x = u_t,\ u_x(0,t) = Q_1,\ u(L,t) = u_2$
(j) $\alpha^2 u_{xx} - Vu_x = u_t,\ u(0,t) + 5u_x(0,t) = 3,\ u(L,t) = 10$
(k) $\alpha^2 u_{xx} - Vu_x = u_t,\ u_x(0,t) = 0,$
$u(L,t) + 2u_x(L,t) = -5$

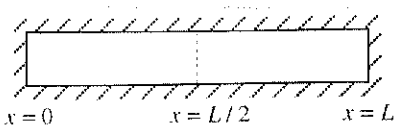
11. (*Existence of steady state*) For the problem

$$\begin{aligned} &\alpha^2 u_{xx} = u_t + F(x), \qquad (0 < x < L,\ 0 < t < \infty) \\ &u_x(0,t) = Q_1, \quad u_x(L,t) = Q_2, \quad u(x,0) = f(x), \end{aligned}$$

show that a steady state does not exist unless a certain condition is satisfied by Q_1, Q_2, and $F(x)$. Assuming that that condition is satisfied, solve for $u_s(x)$.

12. (*Steady-state extrusion*) In Section 18.2 we derive equation (20) governing the temperature distribution $u(x,t)$ in a heated rod being drawn through an extrusion die, as sketched there in Fig. 2. Actually, (20) holds both inside the furnace ($x < 0$) and outside the furnace ($x > 0$), but with different h's and u_∞'s. Let $h = h_f$ and h_a, and let $u_\infty = u_f$ and u_a inside and outside the furnace, respectively. Assuming steady-state operation so that $u = u_s(x)$, propose a suitable set of boundary conditions and solve for $u_s(L)$, the rod temperature at the die. HINT: It will be a helpful approximation to consider the rod to extend from $-\infty$ to $+\infty$. Over $-\infty < x < 0$ use h_f and u_f in the ODE, and over $0 < x < \infty$ use h_a and u_a. Solve over $x < 0$ and $x > 0$ separately, and apply suitable boundary conditions at $x = -\infty$ and $x = +\infty$, as well as suitable "matching conditions" at $x = 0$.

13. (*Diffusion of one gas into another*) Consider a cylindrical compressed-gas container of length L, divided in half by a baffle (see sketch). To the left of the baffle is a gas of species A,



and to the right of it is a different gas of species B. Suppose they are at the same pressure, so that when the baffle is removed at time $t = 0$ the two gases proceed to mix by diffusion

alone. Considering species A, say, its concentration $c_A(x,t)$ moles/cm^3 is governed by the problem

$$D\frac{\partial^2 c_A}{\partial x^2} = \frac{\partial c_A}{\partial t}, \qquad (0 < x < L,\ 0 < t < \infty) \tag{13.1}$$

$$\frac{\partial c_A}{\partial x}(0,t) = \frac{\partial c_A}{\partial x}(L,t) = 0, \qquad (0 < t < \infty) \tag{13.2}$$

$$c_A(x,0) = \begin{cases} c_0 & 0 < x < L/2 \\ 0, & L/2 < x < L \end{cases} \tag{13.3}$$

where D is the diffusion coefficient and D and c_0 are constants.

(a) Solve for $c_A(x,t)$. From $c_A(x,t)$ determine the steady-state solution

$$c_{As}(x) = \lim_{t\to\infty} c_A(x,t). \tag{13.4}$$

Draw a neat sketch of what you expect the graph of $c_A(x,t)$ versus x to look like, at $t = 0$, $t = \infty$ and at a couple of intermediate times. Label any key values and features of those graphs.

(b) Integrating equation (13.1) with respect to x, from 0 to L, show that

$$\int_0^L c_A(x,t)\,dx = \text{constant}, \tag{13.5}$$

which says that the total amount of A is conserved. This result makes sense physically since the container is sealed [note (13.2)] and the gas is being neither created nor destroyed. (The point here is that it may be possible to learn something about the solution, from the PDE and boundary and initial conditions, without actually obtaining the full solution.)

(c) Solve for $c_{As}(x)$ directly, i.e., by solving

$$Dc''_{As}(x) = 0; \quad c'_{As}(0) = c'_{As}(L) = 0$$

and using equation (13.5). Your result should, of course, be the same as in part (a).

14. (*Conduction in metal ring*) Consider the conduction of heat in a circular metal ring, the surface of which is insulated. Actually, whether the shape is a circle or an ellipse or whatever is irrelevant insofar as the heat conduction is concerned. If we measure x along the ring, from some starting point, and denote the length of the ring as L, then the temperature $u(x,t)$ is defined on $-\infty < x < \infty$ and is an L-periodic function of x.

(a) Solve the heat equation $\alpha^2 u_{xx} = u_t$ $(-\infty < x < \infty,\ 0 < t < \infty)$ subject to the initial condition $u(x,0) = f(x)$, where $f(x)$ is L-periodic on $-\infty < x < \infty$. Letting $t \to \infty$ in your solution, show that the steady-state solution is a constant temperature that is the average value of f. HINT: Since u is L-periodic in x, it must be expressible in the Fourier series form

$$u(x,t) = a_0(t) + \sum_{n=1}^{\infty}\left[a_n(t)\cos\frac{2n\pi x}{L} + b_n(t)\sin\frac{2n\pi x}{L}\right], \tag{14.1}$$

where a_0, a_n, b_n vary with time. Putting (14.1) into the PDE and matching the coefficients of 1, $\cos(2n\pi x/L)$, and $\sin(2n\pi x/L)$ on the left- and right-hand sides, derive simple ODE's governing $a_0(t), a_n(t), b_n(t)$. Solve these ODE's and then apply the initial condition $u(x,0) = f(x)$. Give integral formulas for the evaluation of any constants. (Exercise 29 re-examines this problem using the Sturm–Liouville theory.)

(b) Integrating $\alpha^2 u_{xx} = u_t$ from 0 to L, derive the conservation principle

$$\int_0^L u(x,t)\,dx = \text{constant}. \tag{14.2}$$

(c) Derive the steady-state solution $u_s(x)$ again, this time by solving

$$\begin{aligned} \alpha^2 u''_s(x) &= 0, \qquad (0 < x < L) \\ u_s(0) = u_s(L), &\quad u'_s(0) = u'_s(L), \end{aligned}$$

and using (14.3). Your result should be the same as found in part (a).

15. (*Presence of a constant source term*) Consider the problem

$$\begin{aligned} \alpha^2 u_{xx} &= u_t - F, && (0 < x < L,\ 0 < t < \infty) \\ u(0,t) &= 0, \quad u(L,t) = 50, && (0 < t < \infty) \\ u(x,0) &= f(x), && (0 < x < L) \end{aligned} \tag{15.1,2,3}$$

where the source term F is assumed to be a constant. Solve for $u(x,t)$. Expansion coefficients may be left in integral form. HINT: This is the first problem in which the PDE is *nonhomogeneous*, namely, $L[u] = \alpha^2 u_{xx} - u_t = -F$. Observe that if we seek $u(x,t) = X(x)T(t)$ as in the text examples, and attempt to carry out the separation process, we obtain

$$\frac{X''}{X} = \frac{1}{\alpha^2}\frac{T'}{T} - \frac{F}{\alpha^2}\frac{1}{XT}. \tag{15.4}$$

Because of the last term in (15.4), which contains both x