The left-hand side of this equation is the vector

$$\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}.$$

Hence $c_1=0, c_2=0,\ldots,$ and $c_n=0$. Consequently, $\mathbf{e}^1,\mathbf{e}^2,\ldots,\mathbf{e}^n$ are linearly independent vectors in R^n.

Definition. The *dimension* of a vector space $\mathbf{V}$, denoted by $\dim\mathbf{V}$, is the fewest number of linearly independent vectors which span $\mathbf{V}$. $\mathbf{V}$ is said to be a finite dimensional space if its dimension is finite. On the other hand, $\mathbf{V}$ is said to be an infinite dimensional space if no set of finitely many elements span $\mathbf{V}$.

The dimension of a space $\mathbf{V}$ can be characterized as the fewest number of elements that we have to find in order to know all the elements of $\mathbf{V}$. In this sense, the definition of dimension captures our intuitive feeling. However, it is extremely difficult to compute the dimension of a space $\mathbf{V}$ from this definition alone. For example, let $\mathbf{V}=\mathsf{R}^n$. We have shown in Examples 2 and 4 that the vectors $\mathbf{e}^1,\mathbf{e}^2,\ldots,\mathbf{e}^n$ are linearly independent and span $\mathbf{V}$. Moreover, it seems intuitively obvious to us that we cannot generate R^n from fewer than n vectors. Thus, the dimension of R^n should be n. But how can we prove this rigorously? To wit, how can we prove that it is impossible to find a set of $(n-1)$ linearly independent vectors that span R^n? Thus, our definition of dimension is not, as yet, a very useful one. However, it will become extremely useful after we prove the following theorem.

Theorem 2. *If n linearly independent vectors span $\mathbf{V}$, then $\dim\mathbf{V}=n$.*

We will need two lemmas to prove Theorem 2. The first lemma concerns itself with the solutions of simultaneous linear equations and can be motivated as follows. Suppose that we are interested in determining n unknown numbers $x_1,x_2,\ldots,x_n$ uniquely. It seems pretty reasonable that we should be given n equations satisfied by these unknowns. If we are given too few equations then there may be many different solutions, that is, many different sets of values for $x_1,x_2,\ldots,x_n$ which satisfy the given equations. Lemma 1 proves this in the special case that we have m homogeneous linear equations for $n>m$ unknowns.

Lemma 1. *A set of m homogeneous linear equations for n unknowns $x_1,x_2,\ldots,x_n$ always admits a nontrivial solution if $m<n$. That is to say,*

the set of m equations in n unknowns

$$\begin{aligned} a_{11}x_1+a_{12}x_2+\dots+a_{1n}x_n&=0\\ a_{21}x_1+a_{22}x_2+\dots+a_{2n}x_n&=0\\ &\vdots\\ a_{m1}x_1+a_{m2}x_2+\dots+a_{mn}x_n&=0 \end{aligned} \tag{1}$$

always has a solution $x_1,x_2,\dots,x_n$, *other than* $x_1=\dots=x_n=0$, *if* $m<n$.

Remark. Notice that $x_1=0,x_2=0,\dots,x_n=0$ is certainly one solution of the system of equations (1). Thus, Lemma 1 is telling us that these equations have more than one solution.

PROOF OF LEMMA 1. We will prove Lemma 1 by induction on m. To this end, observe that the lemma is certainly true if $m=1$, for in this case we have a single equation of the form $a_{11}x_1+a_{12}x_2+\dots+a_{1n}x_n=0$, with $n\geqslant 2$. We can find a nontrivial solution of this equation, if $a_{11}=0$, by taking $x_1=1,x_2=0,\dots,x_n=0$. We can find a nontrivial solution of this equation, if $a_{11}\neq 0$, by taking $x_2=1,\dots,x_n=1$ and $x_1=-(a_{12}+\dots+a_{1n})/a_{11}$.

For the next step in our induction proof, we assume that Lemma 1 is true for some integer $m=k$ and show that this implies that Lemma 1 is true for $m=k+1$, and $k+1<n$. To this end, consider the $k+1$ equations for the n unknowns $x_1,x_2,\dots,x_n$

$$\begin{aligned} a_{11}x_1\ +\ a_{12}x_2\ +\dots+\ a_{1n}x_n\ &=0\\ a_{21}x_1\ +\ a_{22}x_2\ +\dots+\ a_{2n}x_n\ &=0\\ &\vdots\\ a_{k+1,1}x_1+a_{k+1,2}x_2+\dots+a_{k+1,n}x_n&=0 \end{aligned} \tag{2}$$

with $k+1<n$. If $a_{11},a_{21},\dots,a_{k+1,1}$ are all zero, then $x_1=1,x_2=0,\dots,x_n=0$ is clearly a non-trivial solution. Hence, we may assume that at least one of these coefficients is not zero. Without any loss of generality, we may assume that $a_{11}\neq 0$, for otherwise we can take the equation with the nonzero coefficient of x_1 and relabel it as the first equation. Then

$$x_1=-\frac{a_{12}}{a_{11}}x_2-\frac{a_{13}}{a_{11}}x_3-\dots-\frac{a_{1n}}{a_{11}}x_n.$$

Substituting this value of x_1 in the second through the $(k+1)$st equations, we obtain the equivalent equations

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n &= 0 \\
b_{22}x_2 + b_{23}x_3 + \ldots + b_{2n}x_n &= 0 \\
&\vdots \\
b_{k2}x_2 + b_{k3}x_3 + \ldots + b_{kn}x_n &= 0 \\
b_{k+1,2}x_2 + b_{k+1,3}x_3 + \ldots + b_{k+1,n}x_n &= 0
\end{aligned}
\tag{3}
$$

where $b_{ij} = a_{ij} - a_{i1}a_{1j}/a_{11}$. Now, the last k equations of (3) are k homogeneous linear equations for the $(n-1)$ unknowns $x_2,\ldots,x_n$. Moreover, k is less than $n-1$ since $k+1$ is less than n. Hence, by the induction hypothesis, these equations have a nontrivial solution $x_2,\ldots,x_n$. Once $x_2,\ldots,x_n$ are known, we have as before $x_1 = -(a_{12}x_2 + \ldots + a_{1n}x_n)/a_{11}$ from the first equation of (3). This establishes Lemma 1 for $m = k+1$, and therefore for all m, by induction. □

If a vector space **V** has dimension m, then it has m linearly independent vectors $\mathbf{x}^1,\ldots,\mathbf{x}^m$ and every vector in the space can be written as a linear combination of the m vectors $\mathbf{x}^1,\mathbf{x}^2,\ldots,\mathbf{x}^m$. It seems intuitively obvious to us in this case that there cannot be more than m linear independent vectors in **V**. This is the content of Lemma 2.

Lemma 2. *In an m dimensional space, any set of $n > m$ vectors must be linearly dependent. In other words, the maximum number of linearly independent vectors in a finite dimensional space is the dimension of the space.*

Proof. Since **V** has dimension m, there exist m linearly independent vectors $\mathbf{x}^1,\mathbf{x}^2,\ldots,\mathbf{x}^m$ which span **V**. Let $\mathbf{y}^1,\mathbf{y}^2,\ldots,\mathbf{y}^n$ be a set of n vectors in **V**, with $n > m$. Since $\mathbf{x}^1,\mathbf{x}^2,\ldots,\mathbf{x}^m$ span **V**, all the $\mathbf{y}^j$ can be written as linear combinations of these vectors. That is to say, there exist constants a_{ij}, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant m$ such that

$$
\begin{aligned}
\mathbf{y}^1 &= a_{11}\mathbf{x}^1 + a_{12}\mathbf{x}^2 + \ldots + a_{1m}\mathbf{x}^m \\
\mathbf{y}^2 &= a_{21}\mathbf{x}^1 + a_{22}\mathbf{x}^2 + \ldots + a_{2m}\mathbf{x}^m \\
&\vdots \\
\mathbf{y}^n &= a_{n1}\mathbf{x}^1 + a_{n2}\mathbf{x}^2 + \ldots + a_{nm}\mathbf{x}^m.
\end{aligned}
\tag{4}
$$

To determine whether $\mathbf{y}^1, \mathbf{y}^2, \dots, \mathbf{y}^n$ are linearly dependent or linearly independent, we consider the equation

$$c_1\mathbf{y}^1 + c_2\mathbf{y}^2 + \dots + c_n\mathbf{y}^n = \mathbf{0}. \tag{5}$$

Using (4) we can rewrite (5) in the form

$$\begin{aligned}\mathbf{0} &= c_1\mathbf{y}^1 + c_2\mathbf{y}^2 + \dots + c_n\mathbf{y}^n \\ &= (c_1a_{11} + \dots + c_na_{n1})\mathbf{x}^1 + (c_1a_{12} + \dots + c_na_{n2})\mathbf{x}^2 \\ &\quad + \dots + (c_1a_{1m} + \dots + c_na_{nm})\mathbf{x}^m.\end{aligned}$$

This equation states that a linear combination of $\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^m$ is zero. Since $\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^m$ are linearly independent, all these coefficients must be zero. Hence,

$$\begin{aligned} c_1a_{11} + c_2a_{21} + \dots + c_na_{n1} &= 0 \\ c_1a_{12} + c_2a_{22} + \dots + c_na_{n2} &= 0 \\ &\vdots \\ c_1a_{1m} + c_2a_{2m} + \dots + c_na_{nm} &= 0. \end{aligned} \tag{6}$$

Now, observe that the system of Equations (6) is a set of m homogeneous linear equations for n unknowns $c_1, c_2, \dots, c_n$, with $n > m$. By Lemma 1, these equations have a nontrivial solution. Thus, there exist constants $c_1, c_2, \dots, c_n$, not all zero, such that $c_1\mathbf{y}^1 + c_2\mathbf{y}^2 + \dots + c_n\mathbf{y}^n = \mathbf{0}$. Consequently, $\mathbf{y}^1, \mathbf{y}^2, \dots, \mathbf{y}^n$ are linearly dependent. □

We are now in a position to prove Theorem 2.

PROOF OF THEOREM 2. If n linearly independent vectors span $\mathbf{V}$, then, by the definition of dimension, $\dim \mathbf{V} \leqslant n$. By Lemma 2, $n \leqslant \dim \mathbf{V}$. Hence, $\dim \mathbf{V} = n$. □

Example 5. The dimension of R^n is n since $\mathbf{e}^1, \mathbf{e}^2, \dots, \mathbf{e}^n$ are n linearly independent vectors which span R^n.

Example 6. Let $\mathbf{V}$ be the set of all 3×3 matrices

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix},$$

and let $\mathbf{E}_{ij}$ denote the matrix with a one in the ith row, jth column and zeros everywhere else. For example,

$$\mathbf{E}_{23} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

To determine whether these matrices are linearly dependent or linearly independent, we consider the equation

$$\sum_{i,j=1}^{3} c_{ij}\mathbf{E}_{ij} = \mathbf{0} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{7}$$

Now, observe that the left-hand side of (7) is the matrix

$$c_{11}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + c_{12}\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \ldots + c_{33}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix}.$$

Equating this matrix to the zero matrix gives $c_{11}=0, c_{12}=0,\ldots,c_{33}=0$. Hence the 9 matrices $\mathbf{E}_{ij}$ are linearly independent. Moreover, these 9 matrices also span $\mathbf{V}$ since any matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

can obviously be written in the form $\mathbf{A} = \sum_{i,j=1}^{3} a_{ij}\mathbf{E}_{ij}$. Hence $\dim \mathbf{V} = 9$.

Definition. If a set of linearly independent vectors span a vector space $\mathbf{V}$, then this set of vectors is said to be a *basis* for $\mathbf{V}$. A basis may also be called a *coordinate system*. For example, the vectors

$$\mathbf{e}^1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{e}^2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{e}^3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad \text{and} \quad \mathbf{e}^4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

are a basis for R^4. If

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix},$$

then $\mathbf{x} = x_1\mathbf{e}^1 + x_2\mathbf{e}^2 + x_3\mathbf{e}^3 + x_4\mathbf{e}^4$, and relative to this basis the x_i are called "components" or "coordinates."

Corollary. *In a finite dimensional vector space, each basis has the same number of vectors, and this number is the dimension of the space.*

The following theorem is extremely useful in determining whether a set of vectors is a basis for $\mathbf{V}$.

Theorem 3. *Any n linearly independent vectors in an n dimensional space* $\mathbf{V}$ *must also span* $\mathbf{V}$. *That is to say, any n linearly independent vectors in an n dimensional space* $\mathbf{V}$ *are a basis for* $\mathbf{V}$.

PROOF. Let $\mathbf{x}^1,\mathbf{x}^2,\ldots,\mathbf{x}^n$ be n linearly independent vectors in an n dimensional space $\mathbf{V}$. To show that they span $\mathbf{V}$, we must show that every $\mathbf{x}$ in $\mathbf{V}$ can be written as a linear combination of $\mathbf{x}^1,\mathbf{x}^2,\ldots,\mathbf{x}^n$. To this end, pick any $\mathbf{x}$ in $\mathbf{V}$ and consider the set of vectors $\mathbf{x},\mathbf{x}^1,\mathbf{x}^2,\ldots,\mathbf{x}^n$. This is a set of $(n+1)$ vectors in the n dimensional space $\mathbf{V}$; by Lemma 2, they must be linearly dependent. Consequently, there exist constants $c,c_1,c_2,\ldots,c_n$, not all zero, such that

$$c\mathbf{x}+c_1\mathbf{x}^1+c_2\mathbf{x}^2+\ldots+c_n\mathbf{x}^n=\mathbf{0}. \tag{8}$$

Now $c\neq 0$, for otherwise the set of vectors $\mathbf{x}^1,\mathbf{x}^2,\ldots,\mathbf{x}^n$ would be linearly dependent. Therefore, we can divide both sides of (8) by c to obtain that

$$\mathbf{x}=-\frac{c_1}{c}\mathbf{x}^1-\frac{c_2}{c}\mathbf{x}^2-\ldots-\frac{c_n}{c}\mathbf{x}^n.$$

Hence, any n linearly independent vectors in an n dimensional space $\mathbf{V}$ must also span $\mathbf{V}$. □

Example 7. Prove that the vectors

$$\mathbf{x}^1=\begin{pmatrix}1\\1\end{pmatrix} \quad\text{and}\quad \mathbf{x}^2=\begin{pmatrix}1\\-1\end{pmatrix}$$

form a basis for R^2.

Solution. To determine whether $\mathbf{x}^1$ and $\mathbf{x}^2$ are linearly dependent or linearly independent, we consider the equation

$$c_1\mathbf{x}^1+c_2\mathbf{x}^2=c_1\begin{pmatrix}1\\1\end{pmatrix}+c_2\begin{pmatrix}1\\-1\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix}. \tag{9}$$

Equation (9) implies that $c_1+c_2=0$ and $c_1-c_2=0$. Adding these two equations gives $c_1=0$ while subtracting these two equations gives $c_2=0$. Consequently, $\mathbf{x}^1$ and $\mathbf{x}^2$ are two linearly independent vectors in the two dimensional space R^2. Hence, by Theorem 3, they must also span $\mathbf{V}$.

EXERCISES

In each of Exercises 1–4, determine whether the given set of vectors is linearly dependent or linearly independent.

1. $\begin{pmatrix}1\\1\\1\end{pmatrix}$, $\begin{pmatrix}1\\-1\\1\end{pmatrix}$ and $\begin{pmatrix}-4\\0\\-4\end{pmatrix}$ **2.** $\begin{pmatrix}1\\1\\0\end{pmatrix}$, $\begin{pmatrix}0\\1\\1\end{pmatrix}$ and $\begin{pmatrix}1\\0\\1\end{pmatrix}$

3. $\begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$, $\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}$ **4.** $\begin{pmatrix} 1 \\ 2 \\ 6 \end{pmatrix}$, $\begin{pmatrix} 3 \\ -13 \\ 7 \end{pmatrix}$, $\begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} -1 \\ -1 \\ -1 \end{pmatrix}$

5. Let **V** be the set of all 2×2 matrices. Determine whether the following sets of matrices are linearly dependent or linearly independent in **V**.

(a) $\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$, $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$

(b) $\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$, $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ and $\begin{pmatrix} 2 & -2 \\ -1 & 1 \end{pmatrix}$.

6. Let **V** be the space of all polynomials in t of degree $\leqslant 2$.
(a) Show that $\dim \mathbf{V}=3$.
(b) Let p_1, p_2 and p_3 be the three polynomials whose values at any time t are $(t-1)^2$, $(t-2)^2$, and $(t-1)(t-2)$ respectively. Show that p_1, p_2, and p_3 are linearly independent. Hence, conclude from Theorem 3 that p_1, p_2, and p_3 form a basis for **V**.

7. Let **V** be the set of all solutions of the differential equation $d^2y/dt^2-y=0$.
(a) Show that **V** is a vector space.
(b) Find a basis for **V**.

8. Let **V** be the set of all solutions of the differential equation $(d^3y/dt^3)+y=0$ which satisfy $y(0)=0$. Show that **V** is a vector space and find a basis for it.

9. Let **V** be the set of all polynomials $p(t)=a_0+a_1t+a_2t^2$ which satisfy

$$p(0)+2p'(0)+3p''(0)=0.$$

Show that **V** is a vector space and find a basis for it.

10. Let **V** be the set of all solutions

$$\mathbf{x}=\begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix}$$

of the differential equation

$$\dot{\mathbf{x}}=\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 6 & -11 & 6 \end{pmatrix}\mathbf{x}.$$

Show that

$$\mathbf{x}^1(t)=\begin{bmatrix} e^t \\ e^t \\ e^t \end{bmatrix}, \quad \mathbf{x}^2(t)=\begin{bmatrix} e^{2t} \\ 2e^{2t} \\ 4e^{2t} \end{bmatrix}, \quad \text{and} \quad \mathbf{x}^3(t)=\begin{bmatrix} e^{3t} \\ 3e^{3t} \\ 9e^{3t} \end{bmatrix}$$

form a basis for **V**.

11. Let **V** be a vector space. We say that **W** is a subspace of **V** if **W** is a subset of **V** which is itself a vector space. Let **W** be the subset of $\mathbf{R}^3$ which consists of all vectors

$$\mathbf{x}=\begin{pmatrix} x_1\\ x_2\\ x_3\end{pmatrix}$$

which satisfy the equations

$$\begin{aligned} x_1+x_2+2x_3&=0\\ 2x_1-x_2+\ x_3&=0\\ 6x_1\qquad +6x_3&=0.\end{aligned}$$

Show that **W** is a subspace of $\mathbf{R}^3$ and find a basis for it.

12. Prove that any n vectors which span an n dimensional vector space **V** must be linearly independent. *Hint*: Show that any set of linearly dependent vectors contains a linearly independent subset which also spans **V**.

13. Let $\mathbf{v}^1,\mathbf{v}^2,\dots,\mathbf{v}^n$ be n vectors in a vector space **V**. Let **W** be the subset of **V** which consists of all linear combinations $c_1\mathbf{v}^1+c_2\mathbf{v}^2+\dots+c_n\mathbf{v}^n$ of $\mathbf{v}^1,\mathbf{v}^2,\dots,\mathbf{v}^n$. Show that **W** is a subspace of **V**, and that $\dim\mathbf{W}\leqslant n$.

14. Let **V** be the set of all functions $f(t)$ which are analytic for $|t|<1$, that is, $f(t)$ has a power series expansion $f(t)=a_0+a_1t+a_2t^2+\dots$ which converges for $|t|<1$. Show that **V** is a vector space, and that its dimension is infinite. *Hint*: **V** contains all polynomials.

15. Let $\mathbf{v}^1,\mathbf{v}^2,\dots,\mathbf{v}^m$ be m linearly independent vectors in an n dimensional vector space **V**, with $n>m$. Show that we can find vectors $\mathbf{v}^{m+1},\dots,\mathbf{v}^n$ so that $\mathbf{v}^1,\mathbf{v}^2,\dots,\mathbf{v}^m,\mathbf{v}^{m+1},\dots,\mathbf{v}^n$ form a basis for **V**. That is to say, any set of m linearly independent vectors in an $n>m$ dimensional space **V** can be completed to form a basis for **V**.

16. Find a basis for $\mathbf{R}^3$ which includes the vectors

$$\begin{pmatrix}1\\1\\0\end{pmatrix}\quad\text{and}\quad\begin{pmatrix}1\\3\\4\end{pmatrix}.$$

17. (a) Show that

$$\mathbf{v}^1=\begin{pmatrix}1\\0\\0\end{pmatrix},\qquad \mathbf{v}^2=\frac{1}{\sqrt2}\begin{pmatrix}0\\1\\1\end{pmatrix},\quad\text{and}\quad \mathbf{v}^3=\frac{1}{\sqrt2}\begin{pmatrix}0\\-1\\1\end{pmatrix}$$

are linearly independent in $\mathbf{R}^3$.

(b) Let

$$\mathbf{x}=\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=x_1\mathbf{e}^1+x_2\mathbf{e}^2+x_3\mathbf{e}^3.$$

Since $\mathbf{v}^1, \mathbf{v}^2$, and $\mathbf{v}^3$ are linearly independent they are a basis and $\mathbf{x} = y_1\mathbf{v}^1 + y_2\mathbf{v}^2 + y_3\mathbf{v}^3$. What is the relationship between the original coordinates x_i and the new coordinates y_j?

(c) Express the relations between coordinates in the form $\mathbf{x} = \mathbf{B}\mathbf{y}$. Show that the columns of $\mathbf{B}$ are $\mathbf{v}^1$, $\mathbf{v}^2$, and $\mathbf{v}^3$.

3.4 Applications of linear algebra to differential equations

Recall that an important tool in solving the second-order linear homogeneous equation $(d^2y/dt^2) + p(t)(dy/dt) + q(t)y = 0$ was the existence–uniqueness theorem stated in Section 2.1. In a similar manner, we will make extensive use of Theorem 4 below in solving the homogeneous linear system of differential equations

$$\frac{d\mathbf{x}}{dt} = \mathbf{A}\mathbf{x}, \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \qquad \mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}. \tag{1}$$

The proof of this theorem will be indicated in Section 4.6.

Theorem 4 (Existence–uniqueness theorem). *There exists one, and only one, solution of the initial-value problem*

$$\frac{d\mathbf{x}}{dt} = \mathbf{A}\mathbf{x}, \qquad \mathbf{x}(t_0) = \mathbf{x}^0 = \begin{bmatrix} x_1^0 \\ x_2^0 \\ \vdots \\ x_n^0 \end{bmatrix}. \tag{2}$$

Moreover, this solution exists for $-\infty < t < \infty$.

Theorem 4 is an extremely powerful theorem, and has many implications. In particular, if $\mathbf{x}(t)$ is a nontrivial solution, then $\mathbf{x}(t) \neq \mathbf{0}$ for any t. (If $\mathbf{x}(t^*) = \mathbf{0}$ for some t^*, then $\mathbf{x}(t)$ must be identically zero, since it, and the trivial solution, satisfy the same differential equation and have the same value at $t = t^*$.)

We have already shown (see Example 7, Section 3.2) that the space $\mathbf{V}$ of all solutions of (1) is a vector space. Our next step is to determine the dimension of $\mathbf{V}$.

Theorem 5. *The dimension of the space* $\mathbf{V}$ *of all solutions of the homogeneous linear system of differential equations* (1) *is* n.

PROOF. We will exhibit a basis for $\mathbf{V}$ which contains n elements. To this end, let $\boldsymbol{\phi}^j(t), j=1,\dots,n$ be the solution of the initial-value problem

$$\frac{d\mathbf{x}}{dt}=\mathbf{A}\mathbf{x}, \qquad \mathbf{x}(0)=\mathbf{e}^j=\begin{array}{c} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \end{array} \begin{array}{l} \\ \\ \\ -j\text{th row.} \\ \\ \\ \\ \end{array} \tag{3}$$

For example, $\boldsymbol{\phi}^1(t)$ is the solution of the differential equation (1) which satisfies the initial condition

$$\boldsymbol{\phi}^1(0)=\mathbf{e}^1=\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Note from Theorem 4 that $\boldsymbol{\phi}^j(t)$ exists for all t and is unique. To determine whether $\boldsymbol{\phi}^1,\boldsymbol{\phi}^2,\dots,\boldsymbol{\phi}^n$ are linearly dependent or linearly independent vectors in $\mathbf{V}$, we consider the equation

$$c_1\boldsymbol{\phi}^1+c_2\boldsymbol{\phi}^2+\dots+c_n\boldsymbol{\phi}^n=\mathbf{0} \tag{4}$$

where the zero on the right-hand side of (4) stands for the zero vector in $\mathbf{V}$ (that is, the vector whose every component is the zero function). We want to show that (4) implies $c_1=c_2=\dots=c_n=0$. Evaluating both sides of (4) at $t=0$ gives

$$c_1\boldsymbol{\phi}^1(0)+c_2\boldsymbol{\phi}^2(0)+\dots+c_n\boldsymbol{\phi}^n(0)=\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}=\mathbf{0}$$

or

$$c_1\mathbf{e}^1+c_2\mathbf{e}^2+\dots+c_n\mathbf{e}^n=\mathbf{0}.$$

Since we know that $\mathbf{e}^1,\mathbf{e}^2,\dots,\mathbf{e}^n$ are linearly independent in R^n, $c_1=c_2=\dots=c_n=0$. We conclude, therefore, that $\boldsymbol{\phi}^1,\boldsymbol{\phi}^2,\dots,\boldsymbol{\phi}^n$ are linearly independent vectors in $\mathbf{V}$.

Next, we claim that $\boldsymbol{\phi}^1,\boldsymbol{\phi}^2,\dots,\boldsymbol{\phi}^n$ also span $\mathbf{V}$. To prove this, we must show that any vector $\mathbf{x}$ in $\mathbf{V}$ (that is, any solution $\mathbf{x}(t)$ of (1)) can be written as a linear combination of $\boldsymbol{\phi}^1,\boldsymbol{\phi}^2,\dots,\boldsymbol{\phi}^n$. To this end, pick any $\mathbf{x}$ in $\mathbf{V}$, and let

$$\mathbf{c}=\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}$$

be the value of $\mathbf{x}$ at $t=0$ ($\mathbf{x}(0)=\mathbf{c}$). With these constants $c_1,c_2,\dots,c_n$, construct the vector-valued function

$$\boldsymbol{\phi}(t)=c_1\boldsymbol{\phi}^1(t)+c_2\boldsymbol{\phi}^2(t)+\dots+c_n\boldsymbol{\phi}^n(t).$$

We know that $\boldsymbol{\phi}(t)$ satisfies (1) since it is a linear combination of solutions. Moreover,

$$\begin{aligned}\boldsymbol{\phi}(0)&=c_1\boldsymbol{\phi}^1(0)+c_2\boldsymbol{\phi}^2(0)+\dots+c_n\boldsymbol{\phi}^n(0)\\ &=c_1\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}+c_2\begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}+\dots+c_n\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}=\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}=\mathbf{x}(0).\end{aligned}$$

Now, observe that $\mathbf{x}(t)$ and $\boldsymbol{\phi}(t)$ satisfy the same homogeneous linear system of differential equations, and that $\mathbf{x}(t)$ and $\boldsymbol{\phi}(t)$ have the same value at $t=0$. Consequently, by Theorem 4, $\mathbf{x}(t)$ and $\boldsymbol{\phi}(t)$ must be identical, that is

$$\mathbf{x}(t)\equiv\boldsymbol{\phi}(t)=c_1\boldsymbol{\phi}^1(t)+c_2\boldsymbol{\phi}^2(t)+\dots+c_n\boldsymbol{\phi}^n(t).$$

Thus, $\boldsymbol{\phi}^1,\boldsymbol{\phi}^2,\dots,\boldsymbol{\phi}^n$ also span $\mathbf{V}$. Therefore, by Theorem 2 of Section 3.3, $\dim\mathbf{V}=n$. □

Theorem 5 states that the space $\mathbf{V}$ of all solutions of (1) has dimension n. Hence, we need only guess, or by some means find, n linearly independent solutions of (1). Theorem 6 below establishes a test for linear independence of solutions. It reduces the problem of determining whether n solutions $\mathbf{x}^1,\mathbf{x}^2,\dots,\mathbf{x}^n$ are linearly independent to the much simpler problem of determining whether their values $\mathbf{x}^1(t_0),\mathbf{x}^2(t_0),\dots,\mathbf{x}^n(t_0)$ at an appropriate time t_0 are linearly independent vectors in R^n.

Theorem 6 (Test for linear independence). *Let* $\mathbf{x}^1,\mathbf{x}^2,\dots,\mathbf{x}^k$ *be* k *solutions of* $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$. *Select a convenient* t_0. *Then,* $\mathbf{x}^1,\dots,\mathbf{x}^k$ *are linear independent solutions if, and only if,* $\mathbf{x}^1(t_0),\mathbf{x}^2(t_0),\dots,\mathbf{x}^k(t_0)$ *are linearly independent vectors in* R^n.

PROOF. Suppose that $\mathbf{x}^1,\mathbf{x}^2,\dots,\mathbf{x}^k$ are linearly dependent solutions. Then, there exist constants $c_1,c_2,\dots,c_k$, not all zero, such that

$$c_1\mathbf{x}^1+c_2\mathbf{x}^2+\dots+c_k\mathbf{x}^k=\mathbf{0}.$$

Evaluating this equation at $t=t_0$ gives

$$c_1\mathbf{x}^1(t_0)+c_2\mathbf{x}^2(t_0)+\ldots+c_k\mathbf{x}^k(t_0)=\begin{bmatrix}0\\0\\\vdots\\0\end{bmatrix}.$$

Hence $\mathbf{x}^1(t_0),\mathbf{x}^2(t_0),\ldots,\mathbf{x}^k(t_0)$ are linearly dependent vectors in R^n.

Conversely, suppose that the values of $\mathbf{x}^1,\mathbf{x}^2,\ldots,\mathbf{x}^k$ at some time t_0 are linearly dependent vectors in R^n. Then, there exist constants $c_1,c_2,\ldots,c_k$, not all zero, such that

$$c_1\mathbf{x}^1(t_0)+c_2\mathbf{x}^2(t_0)+\ldots+c_k\mathbf{x}^k(t_0)=\begin{bmatrix}0\\0\\\vdots\\0\end{bmatrix}=\mathbf{0}.$$

With this choice of constants $c_1,c_2,\ldots,c_k$, construct the vector-valued function

$$\boldsymbol{\phi}(t)=c_1\mathbf{x}^1(t)+c_2\mathbf{x}^2(t)+\ldots+c_k\mathbf{x}^k(t).$$

This function satisfies (1) since it is a linear combination of solutions. Moreover, $\boldsymbol{\phi}(t_0)=\mathbf{0}$. Hence, by Theorem 4, $\boldsymbol{\phi}(t)=\mathbf{0}$ for all t. This implies that $\mathbf{x}^1,\mathbf{x}^2,\ldots,\mathbf{x}^k$ are linearly dependent solutions. $\square$

Example 1. Consider the system of differential equations

$$\begin{aligned}\frac{dx_1}{dt}&=x_2\\ \frac{dx_2}{dt}&=-x_1-2x_2\end{aligned}\qquad\text{or}\qquad \frac{d\mathbf{x}}{dt}=\begin{pmatrix}0&1\\-1&-2\end{pmatrix}\mathbf{x},\quad \mathbf{x}=\begin{pmatrix}x_1\\x_2\end{pmatrix}. \tag{5}$$

This system of equations arose from the single second-order equation

$$\frac{d^2y}{dt^2}+2\frac{dy}{dt}+y=0 \tag{6}$$

by setting $x_1=y$ and $x_2=dy/dt$. Since $y_1(t)=e^{-t}$ and $y_2(t)=te^{-t}$ are two solutions of (6), we see that

$$\mathbf{x}^1(t)=\begin{pmatrix}e^{-t}\\-e^{-t}\end{pmatrix}$$

and

$$\mathbf{x}^2(t)=\begin{pmatrix}te^{-t}\\(1-t)e^{-t}\end{pmatrix}$$

are two solutions of (5). To determine whether $\mathbf{x}^1$ and $\mathbf{x}^2$ are linearly dependent or linearly independent, we check whether their initial values

$$\mathbf{x}^1(0)=\begin{pmatrix}1\\-1\end{pmatrix}$$

and

$$\mathbf{x}^2(0)=\begin{pmatrix}0\\1\end{pmatrix}$$

are linearly dependent or linearly independent vectors in R^2. Thus, we consider the equation

$$c_1\mathbf{x}^1(0)+c_2\mathbf{x}^2(0)=\begin{pmatrix}c_1\\-c_1+c_2\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix}.$$

This equation implies that both c_1 and c_2 are zero. Hence, $\mathbf{x}^1(0)$ and $\mathbf{x}^2(0)$ are linearly independent vectors in R^2. Consequently, by Theorem 6, $\mathbf{x}^1(t)$ and $\mathbf{x}^2(t)$ are linearly independent solutions of (5), and every solution $\mathbf{x}(t)$ of (5) can be written in the form

$$\mathbf{x}(t)=\begin{pmatrix}x_1(t)\\x_2(t)\end{pmatrix}=c_1\begin{pmatrix}e^{-t}\\-e^{-t}\end{pmatrix}+c_2\begin{pmatrix}te^{-t}\\(1-t)e^{-t}\end{pmatrix}$$
$$=\begin{pmatrix}(c_1+c_2t)e^{-t}\\(c_2-c_1-c_2t)e^{-t}\end{pmatrix}. \tag{7}$$

Example 2. Solve the initial-value problem

$$\frac{d\mathbf{x}}{dt}=\begin{pmatrix}0&1\\-1&-2\end{pmatrix}\mathbf{x},\qquad \mathbf{x}(0)=\begin{pmatrix}1\\1\end{pmatrix}.$$

Solution. From Example 1, every solution $\mathbf{x}(t)$ must be of the form (7). The constants c_1 and c_2 are determined from the initial conditions

$$\begin{pmatrix}1\\1\end{pmatrix}=\mathbf{x}(0)=\begin{pmatrix}c_1\\c_2-c_1\end{pmatrix}.$$

Therefore, $c_1=1$ and $c_2=1+c_1=2$. Hence

$$\mathbf{x}(t)=\begin{pmatrix}x_1(t)\\x_2(t)\end{pmatrix}=\begin{pmatrix}(1+2t)e^{-t}\\(1-2t)e^{-t}\end{pmatrix}.$$

Up to this point in studying (1) we have found the concepts of linear algebra such as vector space, dependence, dimension, basis, etc., and vector-matrix notation useful, but we might well ask is all this other than simply an appropriate and convenient language. If it were nothing else it would be worth introducing. Good notations are important in expressing mathematical ideas. However, it is more. It is a body of theory with many applications.

In Sections 3.8–3.10 we will reduce the problem of finding all solutions of (1) to the much simpler algebraic problem of solving simultaneous linear equations of the form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n &= b_n \end{aligned}$$

Therefore, we will now digress to study the theory of simultaneous linear equations. Here too we will see the role played by linear algebra.

Exercises

In each of Exercises 1–4 find a basis for the set of solutions of the given differential equation.

1. $\dot{\mathbf{x}} = \begin{pmatrix} 0 & 1 \\ -1 & -1 \end{pmatrix}\mathbf{x}$ (*Hint*: Find a second-order differential equation satisfied by $x_1(t)$.)

2. $\dot{\mathbf{x}} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 2 & -1 & 2 \end{pmatrix}\mathbf{x}$ (*Hint*: Find a third-order differential equation satisfied by $x_1(t)$.)

3. $\dot{\mathbf{x}} = \begin{pmatrix} 1 & 0 \\ 2 & 1 \end{pmatrix}\mathbf{x}$

4. $\dot{\mathbf{x}} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}\mathbf{x}$

For each of the differential equations 5–9 determine whether the given solutions are a basis for the set of all solutions.

5. $\dot{\mathbf{x}} = \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}\mathbf{x}$; $\mathbf{x}^1(t) = \begin{pmatrix} e^t \\ -e^t \end{pmatrix}$, $\mathbf{x}^2(t) = \begin{pmatrix} e^{-t} \\ e^{-t} \end{pmatrix}$

6. $\dot{\mathbf{x}} = \begin{pmatrix} 4 & -2 & 2 \\ -1 & 3 & 1 \\ 1 & -1 & 5 \end{pmatrix}\mathbf{x}$; $\mathbf{x}^1(t) = \begin{bmatrix} e^{2t} \\ e^{2t} \\ 0 \end{bmatrix}$, $\mathbf{x}^2(t) = \begin{bmatrix} 0 \\ e^{4t} \\ e^{4t} \end{bmatrix}$, $\mathbf{x}^3(t) = \begin{bmatrix} e^{6t} \\ 0 \\ e^{6t} \end{bmatrix}$

7. $\dot{\mathbf{x}} = \begin{pmatrix} -3 & -2 & 3 \\ 1 & 0 & -3 \\ 1 & -2 & -1 \end{pmatrix}\mathbf{x}$; $\mathbf{x}^1(t) = \begin{bmatrix} e^{-2t} \\ e^{-2t} \\ e^{-2t} \end{bmatrix}$, $\mathbf{x}^2(t) = \begin{bmatrix} e^{2t} \\ -e^{2t} \\ e^{2t} \end{bmatrix}$, $\mathbf{x}^3(t) = \begin{bmatrix} -e^{-4t} \\ e^{-4t} \\ e^{-4t} \end{bmatrix}$

8. $\dot{\mathbf{x}} = \begin{pmatrix} -5 & 2 & -2 \\ 1 & -4 & -1 \\ -1 & 1 & -6 \end{pmatrix}\mathbf{x}$; $\mathbf{x}^1(t) = \begin{bmatrix} e^{-3t} \\ e^{-3t} \\ 0 \end{bmatrix}$, $\mathbf{x}^2(t) = \begin{bmatrix} 0 \\ e^{-5t} \\ e^{-5t} \end{bmatrix}$

9. $\dot{\mathbf{x}}=\begin{pmatrix}-5 & 2 & -2\\ 1 & -4 & -1\\ -1 & 1 & -6\end{pmatrix}\mathbf{x};\quad \mathbf{x}^1(t)=\begin{bmatrix}e^{-3t}\\ e^{-3t}\\ 0\end{bmatrix},$

$\mathbf{x}^2(t)=\begin{bmatrix}0\\ e^{-5t}\\ e^{-5t}\end{bmatrix},\quad \mathbf{x}^3(t)=\begin{bmatrix}e^{-3t}+e^{-7t}\\ e^{-3t}\\ e^{-7t}\end{bmatrix},\quad \mathbf{x}^4(t)=\begin{bmatrix}2e^{-7t}\\ e^{-5t}\\ e^{-5t}+2e^{-7t}\end{bmatrix}$

10. Determine the solutions $\boldsymbol{\phi}^1,\boldsymbol{\phi}^2,\dots,\boldsymbol{\phi}^n$ (see proof of Theorem 5) for the system of differential equations in (a) Problem 5; (b) Problem 6; (c) Problem 7.

11. Let $\mathbf{V}$ be the vector space of all continuous functions on $(-\infty,\infty)$ to R^n (the values of $\mathbf{x}(t)$ lie in R^n). Let $\mathbf{x}^1,\mathbf{x}^2,\dots,\mathbf{x}^n$ be functions in $\mathbf{V}$.
(a) Show that $\mathbf{x}^1(t_0),\dots,\mathbf{x}^n(t_0)$ linearly independent vectors in R^n for some t_0 implies $\mathbf{x}^1,\mathbf{x}^2,\dots,\mathbf{x}^n$ are linearly independent functions in $\mathbf{V}$.
(b) Is it true that $\mathbf{x}^1(t_0),\dots,\mathbf{x}^n(t_0)$ linearly dependent in R^n for some t_0 implies $\mathbf{x}^1,\mathbf{x}^2,\dots,\mathbf{x}^n$ are linearly dependent functions in $\mathbf{V}$? Justify your answer.

12. Let $\mathbf{u}$ be a vector in $\mathsf{R}^n(\mathbf{u}\neq 0)$.
(a) Is $\mathbf{x}(t)=t\mathbf{u}$ a solution of a linear homogeneous differential equation $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$?
(b) Is $\mathbf{x}(t)=e^{\lambda t}\mathbf{u}$?
(c) Is $\mathbf{x}(t)=(e^t-e^{-t})\mathbf{u}$?
(d) Is $\mathbf{x}(t)=(e^t+e^{-t})\mathbf{u}$?
(e) Is $\mathbf{x}(t)=(e^{\lambda_1 t}+e^{\lambda_2 t})\mathbf{u}$?
(f) For what functions $\phi(t)$ can $\mathbf{x}(t)=\phi(t)\mathbf{u}$ be a solution of some $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$?

3.5 The theory of determinants

In this section we will study simultaneous equations of the form

$$\begin{aligned}
a_{11}x_1+a_{12}x_2+\dots+a_{1n}x_n&=b_1\\
a_{21}x_1+a_{22}x_2+\dots+a_{2n}x_n&=b_2\\
\vdots\qquad\qquad&\quad\vdots\\
a_{n1}x_1+a_{n2}x_2+\dots+a_{nn}x_n&=b_n.
\end{aligned}\tag{1}$$

Our goal is to determine a necessary and sufficient condition for the system of equations (1) to have a unique solution $x_1,x_2,\dots,x_n$.

To gain some insight into this problem, we begin with the simplest case $n=2$. If we multiply the first equation $a_{11}x_1+a_{12}x_2=b_1$ by a_{21}, the second equation $a_{21}x_1+a_{22}x_2=b_2$ by a_{11}, and then subtract the former from the latter, we obtain that

$$(a_{11}a_{22}-a_{12}a_{21})x_2=a_{11}b_2-a_{21}b_1.$$

Similarly, if we multiply the first equation by a_{22}, the second equation by a_{12}, and then subtract the latter from the former, we obtain that

$$(a_{11}a_{22}-a_{12}a_{21})x_1=a_{22}b_1-a_{12}b_2.$$

Consequently, the system of equations (1) has a unique solution

$$x_1=\frac{a_{22}b_1-a_{12}b_2}{a_{11}a_{22}-a_{12}a_{21}},\qquad x_2=\frac{a_{11}b_2-a_{21}b_1}{a_{11}a_{22}-a_{12}a_{21}}$$

if the number $a_{11}a_{22}-a_{12}a_{21}$ is unequal to zero. If this number equals zero, then we may, or may not, have a solution. For example, the system of equations

$$x_1-x_2=1, \qquad 2x_1-2x_2=4$$

obviously has no solutions, while the system of equations

$$x_1-x_2=0, \qquad 2x_1-2x_2=0$$

has an infinity of solutions $x_1=c, x_2=c$ for any number c. For both these systems of equations,

$$a_{11}a_{22}-a_{12}a_{21}=1(-2)-(-1)2=0.$$

The case of three equations

$$\begin{aligned} a_{11}x_1+a_{12}x_2+a_{13}x_3&=b_1\\ a_{21}x_1+a_{22}x_2+a_{23}x_3&=b_2\\ a_{31}x_1+a_{32}x_2+a_{33}x_3&=b_3 \end{aligned} \tag{2}$$

in three unknowns x_1,x_2,x_3 can also be handled quite easily. By eliminating one of the variables from two of the equations (2), and thus reducing ourselves to the case $n=2$, it is possible to show (see Exercise 1) that the system of equations (2) has a unique solution x_1,x_2,x_3 if, and only if, the number

$$a_{11}a_{22}a_{33}+a_{12}a_{23}a_{31}+a_{13}a_{21}a_{32}-a_{13}a_{22}a_{31}-a_{12}a_{21}a_{33}-a_{11}a_{23}a_{32} \tag{3}$$

is unequal to zero.

We now suspect that the system of equations (1), which we abbreviate in the form

$$\mathbf{Ax}=\mathbf{b}, \quad \mathbf{A}=\begin{bmatrix} a_{11} & \dots & a_{1n}\\ \vdots & & \vdots\\ a_{n1} & \dots & a_{nn}\end{bmatrix}, \quad \mathbf{x}=\begin{bmatrix} x_1\\ \vdots\\ x_n\end{bmatrix}, \quad \mathbf{b}=\begin{bmatrix} b_1\\ \vdots\\ b_n\end{bmatrix}, \tag{4}$$

has a unique solution $\mathbf{x}$ if, and only if, a certain number, which depends on the elements a_{ij} of the matrix $\mathbf{A}$, is unequal to zero. We can determine this number for $n=4$ by eliminating one of the variables from two of the equations (1). However, the algebra is so complex that the resulting number is unintelligible. Instead, we will generalize the number (3) so as to associate with each system of equations $\mathbf{Ax}=\mathbf{b}$ a single number called determinant $\mathbf{A}$ ($\det\mathbf{A}$ for short), which depends on the elements of the matrix $\mathbf{A}$. We will establish several useful properties of this association, and then use these properties to show that the system of equations (4) has a unique solution $\mathbf{x}$ if, and only if, $\det\mathbf{A}\neq 0$.

If we carefully analyze the number (3), we see that it can be described in the following interesting manner. First, we pick an element a_{1j_1} from the first row of the matrix

$$\mathbf{A}=\begin{bmatrix} a_{11} & a_{12} & a_{13}\\ a_{21} & a_{22} & a_{23}\\ a_{31} & a_{32} & a_{33}\end{bmatrix}.$$

This element can be either a_{11}, a_{12}, or a_{13}. Then, we multiply a_{1j_1} by an element a_{2j_2} from the second row of $\mathbf{A}$. However, j_2 must not equal j_1. For example, if we choose a_{12} from the first row of $\mathbf{A}$, then we must choose either a_{21} or a_{23} from the second row of $\mathbf{A}$. Next, we multiply these two numbers by the element in the third row of $\mathbf{A}$ in the remaining column. We do this for all possible choices of picking one element from each row of $\mathbf{A}$, never picking from the same column twice. In this manner, we obtain 6 different products of three elements of $\mathbf{A}$, since there are three ways of choosing an element from the first row of $\mathbf{A}$, then two ways of picking an element from the second row of $\mathbf{A}$, and then only one way of choosing an element from the third row of $\mathbf{A}$. Each of these products $a_{1j_1}a_{2j_2}a_{3j_3}$ is multiplied by ± 1, depending on the specific order $j_1 j_2 j_3$. The products $a_{1j_1}a_{2j_2}a_{3j_3}$ with $(j_1 j_2 j_3) = (123)$, (231), and (312) are multiplied by $+1$, while the products $a_{1j_1}a_{2j_2}a_{3j_3}$ with $(j_1 j_2 j_3) = (321)$, (213), and (132) are multiplied by -1. Finally, the resulting numbers are added together.

The six sets of numbers (123), (231), (312), (321), (213), and (132) are called *permutations*, or *scramblings*, of the integers 1, 2, and 3. Observe that each of the three permutations corresponding to the plus terms requires an even number of interchanges of adjacent integers to unscramble the permutation, that is, to bring the integers back to their natural order. Similarly, each of the three permutations corresponding to the minus terms requires an odd number of interchanges of adjacent integers to unscramble the permutation. To verify this, observe that

$$\begin{aligned}
&231\to 213\to 123 \quad &&(2 \text{ interchanges})\\
&312\to 132\to 123 \quad &&(2 \text{ interchanges})\\
&321\to 312\to 132\to 123 \quad &&(3 \text{ interchanges})\\
&213\to 123 \quad \text{and} \quad 132\to 123 \quad &&(1 \text{ interchange each}).
\end{aligned}$$

This motivates the following definition of the *determinant* of an $n\times n$ matrix $\mathbf{A}$.

Definition.

$$\det \mathbf{A} = \sum_{j_1,\dots,j_n} \varepsilon_{j_1 j_2 \dots j_n} a_{1j_1} a_{2j_2} \dots a_{nj_n} \tag{5}$$

where $\varepsilon_{j_1 j_2 \dots j_n} = 1$ if the permutation $(j_1 j_2 \dots j_n)$ is even, that is, if we can bring the integers $j_1 \dots j_n$ back to their natural order by an even number of interchanges of adjacent integers, and $\varepsilon_{j_1 j_2 \dots j_n} = -1$ if the permutation $(j_1 j_2 \dots j_n)$ is odd. In other words, pick an element a_{1j_1} from the first row of the matrix $\mathbf{A}$. Then, multiply it by an element a_{2j_2} from the second row of $\mathbf{A}$, with $j_2 \neq j_1$. Continue this process, going from row to row, and always picking from a new column. Finally, multiply the product $a_{1j_1}a_{2j_2}\ldots a_{nj_n}$ by $+1$ if the permutation $(j_1 j_2 \ldots j_n)$ is even and by -1 if the permutation is odd. Do this for all possible choices of picking one element from each row of $\mathbf{A}$, never picking from the same column twice. Then, add up all these contributions and denote the resulting number by $\det \mathbf{A}$.

Remark. There are many different ways of bringing a permutation of the integers $1,2,\ldots,n$ back to their natural order by successive interchanges of adjacent integers. For example,

$$4312 \to 4132 \to 1432 \to 1423 \to 1243 \to 1234$$

and

$$4312 \to 3412 \to 3142 \to 3124 \to 1324 \to 1234.$$

However, it can be shown that the number of interchanges of adjacent integers necessary to unscramble the permuation $j_1 \ldots j_n$ is always odd or always even. Hence $\varepsilon_{j_1 \ldots j_n}$ is perfectly well defined.

Example 1. Let

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

In this case, there are only two products $a_{11}a_{22}$ and $a_{12}a_{21}$ that enter into the definition of $\det \mathbf{A}$. Since the permutation (12) is even, and the permutation (21) is odd, the term $a_{11}a_{22}$ is multiplied by $+1$ and the term $a_{12}a_{21}$ is multiplied by -1. Hence, $\det \mathbf{A} = a_{11}a_{22} - a_{12}a_{21}$.

Example 2. Compute

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

Solution. A shorthand method for computing the determinant of a 3×3 matrix is to write the first two columns after the matrix and then take products along the diagonals as shown below.

$$\det \begin{bmatrix} 1 & 1 & 1 \\ 3 & 2 & 1 \\ -1 & -1 & 2 \end{bmatrix} = \begin{array}{ccccc} 1 & 1 & 1 & 1 & 1 \\ 3 & 2 & 1 & 3 & 2 \\ -1 & -1 & 2 & -1 & -1 \end{array}$$

$$= 1\cdot 2\cdot 2 + 1\cdot 1\cdot(-1) + 1\cdot 3(-1) - (-1)\cdot 2\cdot 1 - (-1)\cdot 1\cdot 1 - 2\cdot 3\cdot 1 = -3.$$

If $\mathbf{A}$ is an $n \times n$ matrix, then $\det \mathbf{A}$ will contain, in general, $n!$ products of n elements. The determinant of a 4×4 matrix contains, in general, 24 terms, while the determinant of a 10×10 matrix contains the unacceptably large figure of 3,628,800 terms. Thus, it is practically impossible to compute the determinant of a large matrix **A** using the definition (5) alone. The smart way, and the only way, to compute determinants is to (i) find special matrices whose determinants are easy to compute, and (ii) reduce the problem of finding any determinant to the much simpler problem of computing the determinant of one of these special matrices. To this end, observe that there are three special classes of matrices whose determinants are trivial to compute.

1. *Diagonal matrices*: A matrix

$$\mathbf{A}=\begin{bmatrix} a_{11} & 0 & 0 & \dots & 0 \\ 0 & a_{22} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & a_{nn} \end{bmatrix},$$

whose nondiagonal elements are all zero, is called a diagonal matrix. Its determinant is the product of the diagonal elements $a_{11},a_{22},\dots,a_{nn}$. This follows immediately from the observation that the only way we can choose a nonzero element from the first row of **A** is to pick a_{11}. Similarly, the only way we can choose a nonzero element from the jth row of **A** is to pick a_{jj}. Thus the only nonzero product entering into the definition of $\det\mathbf{A}$ is $a_{11}a_{22}\dots a_{nn}$, and this term is multiplied by $+1$ since the permutation $(12\dots n)$ is even.

2. *Lower diagonal matrices*: A matrix

$$\mathbf{A}=\begin{bmatrix} a_{11} & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

whose elements above the main diagonal are all zero, is called a lower diagonal matrix, and its determinant too is the product of the diagonal elements $a_{11},\dots,a_{nn}$. To prove this, observe that the only way we can choose a nonzero element from the first row of **A** is to pick a_{11}. The second row of **A** has two nonzero elements, but since we have already chosen from the first column, we are forced to pick a_{22} from the second row. Similarly, we are forced to pick a_{jj} from the jth row of **A**. Thus, $\det\mathbf{A}=a_{11}a_{22}\dots a_{nn}$.

3. *Upper diagonal matrices*: A matrix

$$\mathbf{A}=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & a_{nn} \end{bmatrix}$$

whose elements below the main diagonal are all zero is called an upper diagonal matrix and its determinant too is the product of the diagonal elements $a_{11},\dots,a_{nn}$. To prove this, we proceed backwards. The only way we can choose a nonzero element from the last row of **A** is to pick a_{nn}. This then forces us to pick $a_{n-1,n-1}$ from the $(n-1)$st row of **A**. Similarly, we are forced to pick a_{jj} from the jth row of **A**. Hence $\det\mathbf{A}=a_{nn}\dots a_{22}a_{11}$.

We now derive some simple but extremely useful properties of determinants.

Property 1. If we interchange two adjacent rows of **A**, then we change the sign of its determinant.

PROOF. Let **B** be the matrix obtained from **A** by interchanging the kth and $(k+1)$st rows. Observe that all of the products entering into the definition of $\det\mathbf{B}$ are exactly the same as the products entering into the definition of $\det\mathbf{A}$. The only difference is that the order in which we choose from the columns of **A** and **B** is changed. For example, let

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

and

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

The product $4\times2\times2$ appears in $\det\mathbf{A}$ by choosing first from the first row, third column; then from the second row, first column; and finally from the third row, second column. This same product appears in $\det\mathbf{B}$ by choosing first from the first row, third column; then from the second row, second column; and finally from the third row, first column. More generally, the term

$$a_{1j_1}\dots a_{kj_k}a_{k+1,j_{k+1}}\dots a_{nj_n}$$

in $\det\mathbf{A}$ corresponds to the term

$$b_{1j_1}\dots b_{k,j_{k+1}}b_{k+1,j_k}\dots b_{nj_n}$$

in $\det\mathbf{B}$. The sign of the first term is determined by the permutation $(j_1\dots j_k j_{k+1}\dots j_n)$ while the sign of the second term is determined by the permutation $(j_1\dots j_{k+1}j_k\dots j_n)$. Since the second permutation is obtained from the first by interchanging the kth and $(k+1)$st elements, we see that these two terms have opposite signs. Hence $\det\mathbf{B}=-\det\mathbf{A}$. □

Property 2. If we interchange any two rows of **A**, then we change the sign of its determinant.

PROOF. We will show that the number of interchanges of adjacent rows required to interchange the ith and jth rows of $\mathbf{A}$ is odd. Property 2 will then follow immediately from Property 1. To this end, assume that j is greater than i. We need $j-i$ successive interchanges of adjacent rows to get the jth row into the ith place, and then $j-i-1$ successive interchanges of adjacent rows to get the original ith row into the jth place. Thus the total number of interchanges required is $2(j-i)-1$, and this number is always odd. □

Property 3. If any two rows of $\mathbf{A}$ are equal, then $\det\mathbf{A}=0$.

PROOF. Let the ith and jth rows of $\mathbf{A}$ be equal, and let $\mathbf{B}$ be the matrix obtained from $\mathbf{A}$ by interchanging its ith and jth rows. Obviously, $\mathbf{B}=\mathbf{A}$. But Property 2 states that $\det\mathbf{B}=-\det\mathbf{A}$. Hence, $\det\mathbf{A}=-\det\mathbf{A}$ if two rows of $\mathbf{A}$ are equal, and this is possible only if $\det\mathbf{A}=0$. □

Property 4. $\det c\mathbf{A}=c^n\det\mathbf{A}$.

PROOF. Obvious. □

Property 5. Let $\mathbf{B}$ be the matrix obtained from $\mathbf{A}$ by multiplying its ith row by a constant c. Then, $\det\mathbf{B}=c\det\mathbf{A}$.

PROOF. Obvious. □

Property 6. Let $\mathbf{A}^T$ be the matrix obtained from $\mathbf{A}$ by switching rows and columns. The matrix $\mathbf{A}^T$ is called the transpose of $\mathbf{A}$. For example, if

$$\mathbf{A}=\begin{bmatrix} 1 & 3 & 2\\ 6 & 9 & 4\\ -1 & 2 & 7\end{bmatrix}, \quad \text{then} \quad \mathbf{A}^T=\begin{bmatrix} 1 & 6 & -1\\ 3 & 9 & 2\\ 2 & 4 & 7\end{bmatrix}.$$

A concise way of saying this is $(\mathbf{A}^T)_{ij}=a_{ji}$. Then,

$$\det\mathbf{A}^T=\det\mathbf{A}.$$

PROOF. It is clear that all the products entering into the definition of $\det\mathbf{A}$ and $\det\mathbf{A}^T$ are equal, since we always choose an element from each row and each column. The proof that these products have the same sign, though, is rather difficult, and will not be included here. (Frankly, whenever this author teaches determinants to his students he wishes that he were king, so that he could declare Property 6 true by edict.) □

Remark. It follows immediately from Properties 2, 3, and 6 that we change the sign of the determinant when we interchange two columns of $\mathbf{A}$, and that $\det\mathbf{A}=0$ if two columns of $\mathbf{A}$ are equal.

Property 7. If we add a multiple of one row of $\mathbf{A}$ to another row of $\mathbf{A}$, then we do not change the value of its determinant.

Proof. We first make the crucial observation that $\det \mathbf{A}$ is a linear function of each row of $\mathbf{A}$ separately. By this we mean the following. Write the matrix $\mathbf{A}$ in the form

$$\mathbf{A}=\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}=\begin{bmatrix} \mathbf{a}^1 \\ \mathbf{a}^2 \\ \vdots \\ \mathbf{a}^n \end{bmatrix}$$

where

$$\begin{aligned} \mathbf{a}^1 &= (a_{11}, a_{12}, \dots, a_{1n}) \\ \mathbf{a}^2 &= (a_{21}, a_{22}, \dots, a_{2n}) \\ &\vdots \\ \mathbf{a}^n &= (a_{n1}, a_{n2}, \dots, a_{nn}). \end{aligned}$$

Then

$$\det\begin{bmatrix} \mathbf{a}^1 \\ \vdots \\ c\mathbf{a}^k \\ \vdots \\ \mathbf{a}^n \end{bmatrix} = c\det\begin{bmatrix} \mathbf{a}^1 \\ \vdots \\ \mathbf{a}^k \\ \vdots \\ \mathbf{a}^n \end{bmatrix} \tag{i}$$

and

$$\det\begin{bmatrix} \mathbf{a}^1 \\ \vdots \\ \mathbf{a}^k+\mathbf{b} \\ \vdots \\ \mathbf{a}^n \end{bmatrix} = \det\begin{bmatrix} \mathbf{a}^1 \\ \vdots \\ \mathbf{a}^k \\ \vdots \\ \mathbf{a}^n \end{bmatrix} + \det\begin{bmatrix} \mathbf{a}^1 \\ \vdots \\ \mathbf{b} \\ \vdots \\ \mathbf{a}^n \end{bmatrix}. \tag{ii}$$

For example,

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

since $(4,1,9)+(4,2,-2)=(8,3,7)$. Now Equation (i) is Property 5. To derive Equation (ii), we compute

$$\det\begin{bmatrix}\mathbf{a}^1\\ \vdots\\ \mathbf{a}^k+\mathbf{b}\\ \vdots\\ \mathbf{a}^n\end{bmatrix} = \sum_{j_1,\ldots,j_n} \varepsilon_{j_1\ldots j_n} a_{1j_1}\ldots(a_{kj_k}+b_{j_k})\ldots a_{nj_n}$$

$$= \sum_{j_1,\ldots,j_n} \varepsilon_{j_1\ldots j_n} a_{1j_1}\ldots a_{kj_k}\ldots a_{nj_n} + \sum_{j_1,\ldots,j_n} \varepsilon_{j_1\ldots j_n} a_{1j_1}\ldots b_{j_k}\ldots a_{nj_n}$$

$$=\det\begin{bmatrix}\mathbf{a}^1\\ \vdots\\ \mathbf{a}^k\\ \vdots\\ \mathbf{a}^n\end{bmatrix}+\det\begin{bmatrix}\mathbf{a}^1\\ \vdots\\ \mathbf{b}\\ \vdots\\ \mathbf{a}^n\end{bmatrix}.$$

Property 7 now follows immediately from Equation (ii), for if $\mathbf{B}$ is the matrix obtained from $\mathbf{A}$ by adding a multiple c of the kth row of $\mathbf{A}$ to the jth row of $\mathbf{A}$, then

$$\det\mathbf{B}=\det\begin{bmatrix}\mathbf{a}^1\\ \vdots\\ \mathbf{a}^j+c\mathbf{a}^k\\ \vdots\\ \mathbf{a}^k\\ \vdots\\ \mathbf{a}^n\end{bmatrix}=\det\begin{bmatrix}\mathbf{a}^1\\ \vdots\\ \mathbf{a}^j\\ \vdots\\ \mathbf{a}^k\\ \vdots\\ \mathbf{a}^n\end{bmatrix}+\det\begin{bmatrix}\mathbf{a}^1\\ \vdots\\ c\mathbf{a}^k\\ \vdots\\ \mathbf{a}^k\\ \vdots\\ \mathbf{a}^n\end{bmatrix}=\det\mathbf{A}+c\det\begin{bmatrix}\mathbf{a}^1\\ \vdots\\ \mathbf{a}^k\\ \vdots\\ \mathbf{a}^k\\ \vdots\\ \mathbf{a}^n\end{bmatrix}.$$

But

$$\det\begin{bmatrix}\mathbf{a}^1\\ \vdots\\ \mathbf{a}^k\\ \vdots\\ \mathbf{a}^k\\ \vdots\\ \mathbf{a}^n\end{bmatrix}=0$$

since this matrix has two equal rows. Hence, $\det\mathbf{B}=\det\mathbf{A}$. □

Remark 1. Everything we say about rows applies to columns, since $\det \mathbf{A}^T = \det \mathbf{A}$. Thus, we do not change the value of the determinant when we add a multiple of one column of $\mathbf{A}$ to another column of $\mathbf{A}$.

Remark 2. The determinant is a linear function of each row of $\mathbf{A}$ separately. It is not a linear function of $\mathbf{A}$ itself, since, in general,

$$\det c\mathbf{A} \neq c \det \mathbf{A} \quad \text{and} \quad \det(\mathbf{A}+\mathbf{B}) \neq \det \mathbf{A} + \det \mathbf{B}.$$

For example, if

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

and

$$\mathbf{B} = \begin{pmatrix} -1 & 7 \\ 0 & 9 \end{pmatrix},$$

then

$$\det(\mathbf{A}+\mathbf{B}) = \det\begin{pmatrix} 0 & 5 \\ 0 & 12 \end{pmatrix} = 0,$$

while $\det \mathbf{A} + \det \mathbf{B} = 3 - 9 = -6$.

Property 7 is extremely important because it enables us to reduce the problem of computing any determinant to the much simpler problem of computing the determinant of an upper diagonal matrix. To wit, if

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

and $a_{11} \neq 0$, then we can add suitable multiples of the first row of $\mathbf{A}$ to the remaining rows of $\mathbf{A}$ so as to make the resulting values of $a_{21}, \dots, a_{n1}$ all zero. Similarly, we can add multiples of the resulting second row of $\mathbf{A}$ to the rows beneath it so as to make the resulting values of $a_{32}, \dots, a_{n2}$ all zero, and so on. We illustrate this method with the following example.

Example 3. Compute

$$\det\begin{bmatrix} 1 & -1 & 2 & 3 \\ 2 & 2 & 0 & 2 \\ 4 & 1 & -1 & -1 \\ 1 & 2 & 3 & 0 \end{bmatrix}.$$

Solution. Subtracting twice the first row from the second row; four times the first row from the third row; and the first row from the last row gives

$$\det\begin{bmatrix} 1 & -1 & 2 & 3\\ 2 & 2 & 0 & 2\\ 4 & 1 & -1 & -1\\ 1 & 2 & 3 & 0 \end{bmatrix} = \det\begin{bmatrix} 1 & -1 & 2 & 3\\ 0 & 4 & -4 & -4\\ 0 & 5 & -9 & -13\\ 0 & 3 & 1 & -3 \end{bmatrix}$$

$$= 4\det\begin{bmatrix} 1 & -1 & 2 & 3\\ 0 & 1 & -1 & -1\\ 0 & 5 & -9 & -13\\ 0 & 3 & 1 & -3 \end{bmatrix}.$$

Next, we subtract five times the second row of this latter matrix from the third row, and three times the second row from the fourth row. Then

$$\det\begin{bmatrix} 1 & -1 & 2 & 3\\ 2 & 2 & 0 & 2\\ 4 & 1 & -1 & -1\\ 1 & 2 & 3 & 0 \end{bmatrix} = 4\det\begin{bmatrix} 1 & -1 & 2 & 3\\ 0 & 1 & -1 & -1\\ 0 & 0 & -4 & -8\\ 0 & 0 & 4 & 0 \end{bmatrix}.$$

Finally, adding the third row of this matrix to the fourth row gives

$$\det\begin{bmatrix} 1 & -1 & 2 & 3\\ 2 & 2 & 0 & 2\\ 4 & 1 & -1 & -1\\ 1 & 2 & 3 & 0 \end{bmatrix} = 4\det\begin{bmatrix} 1 & -1 & 2 & 3\\ 0 & 1 & -1 & -1\\ 0 & 0 & -4 & -8\\ 0 & 0 & 0 & -8 \end{bmatrix}$$
$$= 4(-4)(-8) = 128.$$

(Alternately, we could have interchanged the third and fourth columns of the matrix

$$\begin{bmatrix} 1 & -1 & 2 & 3\\ 0 & 1 & -1 & -1\\ 0 & 0 & -4 & -8\\ 0 & 0 & 4 & 0 \end{bmatrix}$$

to yield the same result.)

Remark 1. If $a_{11}=0$ and $a_{j1}\neq 0$ for some j, then we can interchange the first and jth rows of $\mathbf{A}$ so as to make $a_{11}\neq 0$. (We must remember, of course, to multiply the determinant by -1.) If the entire first column of $\mathbf{A}$ is zero, that is, if $a_{11}=a_{21}=\ldots=a_{n1}=0$ then we need proceed no further for $\det\mathbf{A}=0$.

Remark 2. In exactly the same manner as we reduced the matrix $\mathbf{A}$ to an upper diagonal matrix, we can reduce the system of equations

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= b_1\\
a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n &= b_2\\
&\vdots\\
a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n &= b_n
\end{aligned}$$

to an equivalent system of the form

$$\begin{aligned}
c_{11}x_1 + c_{12}x_2 + \dots + c_{1n}x_n &= d_1\\
c_{22}x_2 + \dots + c_{2n}x_n &= d_2\\
&\vdots\\
c_{nn}x_n &= d_n.
\end{aligned}$$

We can then solve (if $c_{nn} \neq 0$) for x_n from the last equation, for x_{n-1} from the $(n-1)$st equation, and so on.

Example 4. Find all solutions of the system of equations

$$\begin{aligned}
x_1 + x_2 + x_3 &= 1\\
-x_1 + x_2 + x_3 &= 2\\
2x_1 - x_2 + x_3 &= 3.
\end{aligned}$$

Solution. Adding the first equation to the second equation and subtracting twice the first equation from the third equation gives

$$\begin{aligned}
x_1 + x_2 + x_3 &= 1\\
2x_2 + 2x_3 &= 3\\
-3x_2 - x_3 &= 1.
\end{aligned}$$

Next, adding $\frac{3}{2}$ the second equation to the third equation gives

$$\begin{aligned}
x_1 + x_2 + x_3 &= 1\\
2x_2 + 2x_3 &= 3\\
2x_3 &= \tfrac{11}{2}.
\end{aligned}$$

Consequently, $x_3 = \frac{11}{4}$, $x_2 = (3 - \frac{11}{2})/2 = -\frac{5}{4}$, and $x_1 = 1 + \frac{5}{4} - \frac{11}{4} = -\frac{1}{2}$.

Exercises

1. Show that the system of equations

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1\\
a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2\\
a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3
\end{aligned}$$

has a unique solution x_1,x_2,x_3 if, and only if,

$$\det\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \neq 0.$$

Hint: Solve for x_1 in terms of x_2 and x_3 from one of these equations.

2. (a) Show that the total number of permutations of the integers $1,2,\ldots,n$ is even.

(b) Prove that exactly half of these permutations are even, and half are odd.

In each of Problems 3–8 compute the determinant of the given matrix.

3. $\begin{pmatrix} 1 & 0 & 2 \\ 1 & 2 & 5 \\ 6 & 8 & 0 \end{pmatrix}$

4. $\begin{bmatrix} 1 & t & t^2 \\ t & t^2 & 1 \\ t^2 & t & 1 \end{bmatrix}$

5. $\begin{bmatrix} 0 & a & b & 0 \\ -a & 0 & c & 0 \\ -b & -c & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

6. $\begin{bmatrix} 2 & -1 & 6 & 3 \\ 1 & 0 & 1 & -1 \\ 1 & 3 & 0 & 2 \\ 1 & -1 & 1 & 0 \end{bmatrix}$

7. $\begin{bmatrix} 0 & 2 & 3 & -1 \\ 1 & 8 & -1 & 1 \\ 1 & -1 & 0 & -1 \\ 0 & 2 & 6 & 1 \end{bmatrix}$

8. $\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 2 \\ 0 & 1 & 0 & 0 & 3 \\ 0 & 0 & 1 & 0 & 4 \\ 0 & 0 & 0 & 1 & 5 \end{bmatrix}$

9. Without doing any computations, show that

$$\det\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & k \end{bmatrix} = \det\begin{bmatrix} e & b & h \\ d & a & g \\ f & c & k \end{bmatrix}.$$

In each of Problems 10–15, find all solutions of the given system of equations

10. $$\begin{aligned} x_1+x_2-x_3&=0 \\ 2x_1 \quad + \quad x_3&=14 \\ x_2+x_3&=13 \end{aligned}$$

11. $$\begin{aligned} x_1+x_2+x_3&=6 \\ x_1-x_2-x_3&=-4 \\ x_2+x_3&=-1 \end{aligned}$$

12. $$\begin{aligned} x_1+x_2+x_3&=0 \\ x_1-x_2-x_3&=0 \\ x_2+x_3&=0 \end{aligned}$$

13. $$\begin{aligned} x_1+\ x_2+\ x_3-x_4&=1 \\ x_1+2x_2-2x_3+x_4&=1 \\ x_1+3x_2-3x_3-x_4&=1 \\ x_1+4x_2-4x_3-x_4&=1 \end{aligned}$$

14. $$\begin{aligned} x_1+x_2+2x_3-x_4&=1 \\ x_1-x_2+2x_3+x_4&=2 \\ x_1+x_2+2x_3-x_4&=1 \\ -x_1-x_2-2x_3+x_4&=0 \end{aligned}$$

15. $$\begin{aligned} x_1-\ x_2+x_3+\ x_4&=0 \\ x_1+2x_2-x_3+3x_4&=0 \\ 3x_1+3x_2-x_3+7x_4&=0 \\ -x_1+2x_2+x_3-\ x_4&=0 \end{aligned}$$

3.6 Solutions of simultaneous linear equations

In this section we will prove that the system of equations

$$\mathbf{Ax}=\mathbf{b}, \qquad \mathbf{A}=\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}, \quad \mathbf{x}=\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{b}=\begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \tag{1}$$

has a unique solution $\mathbf{x}$ if $\det \mathbf{A} \neq 0$. To this end, we define the product of two $n\times n$ matrices and then derive some additional properties of determinants.

Definition. Let $\mathbf{A}$ and $\mathbf{B}$ be $n\times n$ matrices with elements a_{ij} and b_{ij} respectively. We define their product $\mathbf{AB}$ as the $n\times n$ matrix $\mathbf{C}$ whose ij element c_{ij} is the product of the ith row of $\mathbf{A}$ with the jth column of $\mathbf{B}$. That is to say,

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

Alternately, if we write $\mathbf{B}$ in the form $\mathbf{B}=(\mathbf{b}^1,\mathbf{b}^2,\dots,\mathbf{b}^n)$, where $\mathbf{b}^j$ is the jth column of $\mathbf{B}$, then we can express the product $\mathbf{C}=\mathbf{AB}$ in the form $\mathbf{C}=(\mathbf{Ab}^1,\mathbf{Ab}^2,\dots,\mathbf{Ab}^n)$, since the ith component of the vector $\mathbf{Ab}^j$ is $\sum_{k=1}^{n} a_{ik}b_{kj}$.

Example 1. Let

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

and

$$\mathbf{B}=\begin{bmatrix} 1 & -1 & 0 \\ 2 & -1 & 1 \\ -1 & 0 & 0 \end{bmatrix}.$$

Compute $\mathbf{AB}$.

Solution.

$$\begin{bmatrix} 3 & 1 & -1 \\ 0 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} 1 & -1 & 0 \\ 2 & -1 & 1 \\ -1 & 0 & 0 \end{bmatrix}=\begin{bmatrix} 3+2+1 & -3-1+0 & 0+1+0 \\ 0+4-1 & 0-2+0 & 0+2+0 \\ 1+2-1 & -1-1+0 & 0+1+0 \end{bmatrix}$$

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

Example 2. Let **A** and **B** be the matrices in Example 1. Compute **BA**.
Solution.

$$\begin{bmatrix} 1 & -1 & 0 \\ 2 & -1 & 1 \\ -1 & 0 & 0 \end{bmatrix}\begin{bmatrix} 3 & 1 & -1 \\ 0 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix}=\begin{bmatrix} 3+0+0 & 1-2+0 & -1-1+0 \\ 6+0+1 & 2-2+1 & -2-1+1 \\ -3+0+0 & -1+0+0 & 1+0+0 \end{bmatrix}$$

$$=\begin{bmatrix} 3 & -1 & -2 \\ 7 & 1 & -2 \\ -3 & -1 & 1 \end{bmatrix}$$

Remark 1. As Examples 1 and 2 indicate, it is generally not true that $\mathbf{AB}=\mathbf{BA}$. It can be shown, though, that

$$\mathbf{A}(\mathbf{BC})=(\mathbf{AB})\mathbf{C} \tag{2}$$

for any three $n\times n$ matrices **A**, **B**, and **C**. We will give an extremely simple proof of (2) in the next section.

Remark 2. Let **I** denote the diagonal matrix

$$\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}.$$

I is called the identity matrix since $\mathbf{IA}=\mathbf{AI}=\mathbf{A}$ (see Exercise 5) for any $n\times n$ matrix **A**.

The following two properties of determinants are extremely useful in many applications.

Property 8.

$$\det\mathbf{AB}=\det\mathbf{A}\times\det\mathbf{B}.$$

That is to say, the determinant of the product is the product of the determinants.

Property 9. Let $\mathbf{A}(i|j)$ denote the $(n-1)\times(n-1)$ matrix obtained from **A** by deleting the ith row and jth column of **A**. For example, if

$$\mathbf{A}=\begin{bmatrix} 1 & 2 & 6 \\ 0 & -1 & -3 \\ -4 & -5 & 0 \end{bmatrix},\quad\text{then}\quad \mathbf{A}(2|3)=\begin{pmatrix} 1 & 2 \\ -4 & -5 \end{pmatrix}.$$

Let $c_{ij}=(-1)^{i+j}\det\mathbf{A}(i|j)$. Then

$$\det\mathbf{A}=\sum_{i=1}^{n} a_{ij}c_{ij}$$

for any choice of j between 1 and n. This process of computing determinants is known as "expansion by the elements of columns," and Property 9 states that it does not matter which column we choose to expand about. For example, let

$$\mathbf{A}=\begin{bmatrix} 1 & 3 & 2 & 6\\ 9 & -1 & 0 & 7\\ 2 & 1 & 3 & 4\\ -1 & 6 & 3 & 5\end{bmatrix}.$$

Expanding about the first, second, third, and fourth columns of $\mathbf{A}$, respectively, gives

$$\begin{aligned}
\det\mathbf{A}&=\det\begin{bmatrix} -1 & 0 & 7\\ 1 & 3 & 4\\ 6 & 3 & 5\end{bmatrix}-9\det\begin{bmatrix} 3 & 2 & 6\\ 1 & 3 & 4\\ 6 & 3 & 5\end{bmatrix}\\
&\quad+2\det\begin{bmatrix} 3 & 2 & 6\\ -1 & 0 & 7\\ 6 & 3 & 5\end{bmatrix}+\det\begin{bmatrix} 3 & 2 & 6\\ -1 & 0 & 7\\ 1 & 3 & 4\end{bmatrix}\\
&=-3\det\begin{bmatrix} 9 & 0 & 7\\ 2 & 3 & 4\\ -1 & 3 & 5\end{bmatrix}-\det\begin{bmatrix} 1 & 2 & 6\\ 2 & 3 & 4\\ -1 & 3 & 5\end{bmatrix}\\
&\quad-\det\begin{bmatrix} 1 & 2 & 6\\ 9 & 0 & 7\\ -1 & 3 & 5\end{bmatrix}+6\det\begin{bmatrix} 1 & 2 & 6\\ 9 & 0 & 7\\ 2 & 3 & 4\end{bmatrix}\\
&=2\det\begin{bmatrix} 9 & -1 & 7\\ 2 & 1 & 4\\ -1 & 6 & 5\end{bmatrix}+3\det\begin{bmatrix} 1 & 3 & 6\\ 9 & -1 & 7\\ -1 & 6 & 5\end{bmatrix}\\
&\quad-3\det\begin{bmatrix} 1 & 3 & 6\\ 9 & -1 & 7\\ 2 & 1 & 4\end{bmatrix}\\
&=-6\det\begin{bmatrix} 9 & -1 & 0\\ 2 & 1 & 3\\ -1 & 6 & 3\end{bmatrix}+7\det\begin{bmatrix} 1 & 3 & 2\\ 2 & 1 & 3\\ -1 & 6 & 3\end{bmatrix}\\
&\quad-4\det\begin{bmatrix} 1 & 3 & 2\\ 9 & -1 & 0\\ -1 & 6 & 3\end{bmatrix}+5\det\begin{bmatrix} 1 & 3 & 2\\ 9 & -1 & 0\\ 2 & 1 & 3\end{bmatrix}.
\end{aligned}$$

We will derive Properties 8 and 9 with the aid of the following lemma.

Lemma 1. Let $D=D(\mathbf{A})$ be a function that assigns to each $n\times n$ matrix $\mathbf{A}$ a number $D(\mathbf{A})$. Suppose, moreover, that D is a linear function of each column (row) of $\mathbf{A}$ separately, i.e.

$$D(\mathbf{a}^1,\dots,\mathbf{a}^j+c\mathbf{b}^j,\dots,\mathbf{a}^n)=D(\mathbf{a}^1,\dots,\mathbf{a}^j,\dots,\mathbf{a}^n)+cD(\mathbf{a}^1,\dots,\mathbf{b}^j,\dots,\mathbf{a}^n),$$

and $D(\mathbf{B})=-D(\mathbf{A})$ if $\mathbf{B}$ is obtained from $\mathbf{A}$ by interchanging two columns (rows) of $\mathbf{A}$. Then

$$D(\mathbf{A})=\det\mathbf{A}\times D(\mathbf{I}).$$

A function D that assigns to each $n\times n$ matrix a number is called alternating if $D(\mathbf{B})=-D(\mathbf{A})$ whenever $\mathbf{B}$ is obtained from $\mathbf{A}$ by interchanging two columns (rows) of $\mathbf{A}$. Lemma 1 shows that the properties of being alternating and linear in the columns (rows) of $\mathbf{A}$ serve almost completely to characterize the determinant function $\det\mathbf{A}$. More precisely, any function $D(\mathbf{A})$ which is alternating and linear in the columns (rows) of $\mathbf{A}$ must be a constant multiple of $\det\mathbf{A}$. If, in addition, $D(\mathbf{I})=1$, then $D(\mathbf{A})=\det\mathbf{A}$ for all $n\times n$ matrices $\mathbf{A}$. It also follows immediately from Lemma 1 that if $D(\mathbf{A})$ is alternating and linear in the columns of $\mathbf{A}$, then it is also alternating and linear in the rows of $\mathbf{A}$.

Proof of lemma 1. We first write $\mathbf{A}$ in the form $\mathbf{A}=(\mathbf{a}^1,\mathbf{a}^2,\dots,\mathbf{a}^n)$ where

$$\mathbf{a}^1=\begin{bmatrix}a_{11}\\a_{21}\\ \vdots\\a_{n1}\end{bmatrix},\qquad \mathbf{a}^2=\begin{bmatrix}a_{12}\\a_{22}\\ \vdots\\a_{n2}\end{bmatrix},\dots,\mathbf{a}^n=\begin{bmatrix}a_{1n}\\a_{2n}\\ \vdots\\a_{nn}\end{bmatrix}.$$

Then, writing $\mathbf{a}^1$ in the form $\mathbf{a}^1=a_{11}\mathbf{e}^1+\dots+a_{n1}\mathbf{e}^n$ we see that

$$\begin{aligned}D(\mathbf{A})&=D\left(a_{11}\mathbf{e}^1+\dots+a_{n1}\mathbf{e}^n,\mathbf{a}^2,\dots,\mathbf{a}^n\right)\\&=a_{11}D(\mathbf{e}^1,\mathbf{a}^2,\dots,\mathbf{a}^n)+\dots+a_{n1}D(\mathbf{e}^n,\mathbf{a}^2,\dots,\mathbf{a}^n)\\&=\sum_{j_1}a_{1j_1}D(\mathbf{e}^{j_1},\mathbf{a}^2,\dots,\mathbf{a}^n).\end{aligned}$$

Similarly, writing $\mathbf{a}^2$ in the form $\mathbf{a}^2=a_{12}\mathbf{e}^1+\dots+a_{n2}\mathbf{e}^n$ we see that

$$D(\mathbf{A})=\sum_{j_1,j_2}a_{1j_1}a_{2j_2}D(\mathbf{e}^{j_1},\mathbf{e}^{j_2},\mathbf{a}^3,\dots,\mathbf{a}^n).$$

Proceeding inductively, we see that

$$D(\mathbf{A})=\sum_{j_1,\dots,j_n}a_{1j_1}a_{2j_2}\cdots a_{nj_n}D(\mathbf{e}^{j_1},\mathbf{e}^{j_2},\dots,\mathbf{e}^{j_n}).$$

Now, we need only sum over those integers $j_1,j_2,\dots,j_n$ with $j_i\neq j_k$ since $D(\mathbf{A})$ is zero if $\mathbf{A}$ has two equal columns. Moreover,

$$D(\mathbf{e}^{j_1},\mathbf{e}^{j_2},\dots,\mathbf{e}^{j_n})=\varepsilon_{j_1j_2\dots j_n}D(\mathbf{e}^1,\mathbf{e}^2,\dots,\mathbf{e}^n)=\varepsilon_{j_1\dots j_n}D(\mathbf{I}).$$

Consequently,

$$D(\mathbf{A}) = \sum_{j_1,\dots,j_n} \varepsilon_{j_1\dots j_n} a_{1j_1}\dots a_{nj_n} D(\mathbf{I}) = \det\mathbf{A} \times D(\mathbf{I}). \qquad \square$$

We are now in a position to derive Properties 8 and 9.

Proof of property 8. Let $\mathbf{A}$ be a fixed $n\times n$ matrix, and define the function $D(\mathbf{B})$ by the formula

$$D(\mathbf{B}) = \det\mathbf{AB}.$$

Observe that $D(\mathbf{B})$ is alternating and linear in the columns $\mathbf{b}^1,\dots,\mathbf{b}^n$ of $\mathbf{B}$. This follows immediately from the fact that the columns of $\mathbf{AB}$ are $\mathbf{Ab}^1,\dots,\mathbf{Ab}^n$. Hence, by Lemma 1.

$$D(\mathbf{B}) = \det\mathbf{B} \times D(\mathbf{I}) = \det\mathbf{B}\times\det\mathbf{AI} = \det\mathbf{A}\times\det\mathbf{B}. \qquad \square$$

Proof of property 9. Pick any integer j between 1 and n and let

$$D(\mathbf{A}) = \sum_{i=1}^{n} a_{ij}c_{ij},$$

where $c_{ij} = (-1)^{i+j}\det\mathbf{A}(i|j)$. It is trivial to verify that D is alternating and linear in the columns of $\mathbf{A}$. Hence, by Lemma 1,

$$D(\mathbf{A}) = \det\mathbf{A}\times D(\mathbf{I}) = \det\mathbf{A}. \qquad \square$$

The key to solving the system of equations (1) is the crucial observation that

$$\sum_{i=1}^{n} a_{ik}c_{ij} = 0 \quad \text{for} \quad k\neq j \tag{3}$$

where $c_{ij} = (-1)^{i+j}\det\mathbf{A}(i|j)$. The proof of (3) is very simple: Let $\mathbf{B}$ denote the matrix obtained from $\mathbf{A}$ by replacing the jth column of $\mathbf{A}$ by its kth column, leaving everything else unchanged. For example, if

$$\mathbf{A} = \begin{bmatrix} 3 & 5 & 6 \\ 4 & 2 & 1 \\ -1 & 0 & -1 \end{bmatrix}, \qquad j=2, \quad \text{and} \quad k=3,$$

then

$$\mathbf{B} = \begin{bmatrix} 3 & 6 & 6 \\ 4 & 1 & 1 \\ -1 & -1 & -1 \end{bmatrix}.$$

Now, the determinant of $\mathbf{B}$ is zero, since $\mathbf{B}$ has two equal columns. On the other hand, expanding about the jth column of $\mathbf{B}$ gives

$$\det\mathbf{B} = \sum_{i=1}^{n} b_{ij}\hat{c}_{ij} = \sum_{i=1}^{n} a_{ik}\hat{c}_{ij}$$

where

$$\hat{c}_{ij} = (-1)^{i+j} \det \mathbf{B}(i|j) = (-1)^{i+j} \det \mathbf{A}(i|j) = c_{ij}.$$

Hence, $\sum_{i=1}^{n} a_{ik} c_{ij} = 0$ if k is unequal to j.

Now, whenever we see a sum from 1 to n involving the product of terms with two fixed indices j and k (as in (3)), we try and write it as the jk element of the product of two matrices. If we let $\mathbf{C}$ denote the matrix whose ij element is c_{ij} and set

$$\operatorname{adj} \mathbf{A} \equiv \mathbf{C}^T,$$

then

$$\sum_{i=1}^{n} a_{ik} c_{ij} = \sum_{i=1}^{n} (\operatorname{adj} \mathbf{A})_{ji} a_{ik} = (\operatorname{adj} \mathbf{A} \times \mathbf{A})_{jk}.$$

Hence, from (3),

$$(\operatorname{adj} \mathbf{A} \times \mathbf{A})_{jk} = 0 \quad \text{for} \quad j \neq k.$$

Combining this result with the identity

$$\det \mathbf{A} = \sum_{i=1}^{n} a_{ij} c_{ij} = (\operatorname{adj} \mathbf{A} \times \mathbf{A})_{jj},$$

we see that

$$\operatorname{adj} \mathbf{A} \times \mathbf{A} = \begin{bmatrix} \det \mathbf{A} & 0 & \dots & 0 \\ 0 & \det \mathbf{A} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \det \mathbf{A} \end{bmatrix} = (\det \mathbf{A})\mathbf{I}. \tag{4}$$

Similarly, by working with $\mathbf{A}^T$ instead of $\mathbf{A}$ (see Exercise 8) we see that

$$\mathbf{A} \times \operatorname{adj} \mathbf{A} = \begin{bmatrix} \det \mathbf{A} & 0 & \dots & 0 \\ 0 & \det \mathbf{A} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \det \mathbf{A} \end{bmatrix} = (\det \mathbf{A})\mathbf{I}. \tag{5}$$

Example 3. Let

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

Compute $\operatorname{adj} \mathbf{A}$ and verify directly the identities (4) and (5).

Solution.

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

so that

$$\operatorname{adj}\mathbf{A}\times\mathbf{A}=\begin{bmatrix}3&-2&1\\-2&2&0\\1&-2&1\end{bmatrix}\begin{bmatrix}1&0&-1\\1&1&-1\\1&2&1\end{bmatrix}=\begin{bmatrix}2&0&0\\0&2&0\\0&0&2\end{bmatrix}=2\mathbf{I}$$

and

$$\mathbf{A}\times\operatorname{adj}\mathbf{A}=\begin{bmatrix}1&0&-1\\1&1&-1\\1&2&1\end{bmatrix}\begin{bmatrix}3&-2&1\\-2&2&0\\1&-2&1\end{bmatrix}=\begin{bmatrix}2&0&0\\0&2&0\\0&0&2\end{bmatrix}=2\mathbf{I}.$$

Now, $\det\mathbf{A}=1-2+1+2=2$. Hence,

$$\operatorname{adj}\mathbf{A}\times\mathbf{A}=\mathbf{A}\times\operatorname{adj}\mathbf{A}=(\det\mathbf{A})\mathbf{I}.$$

We return now to the system of equations

$$\mathbf{A}\mathbf{x}=\mathbf{b},\qquad \mathbf{A}=\begin{bmatrix}a_{11}&\dots&a_{1n}\\\vdots&&\vdots\\a_{n1}&\dots&a_{nn}\end{bmatrix},\quad \mathbf{x}=\begin{bmatrix}x_1\\\vdots\\x_n\end{bmatrix},\quad \mathbf{b}=\begin{bmatrix}b_1\\\vdots\\b_n\end{bmatrix}. \tag{6}$$

If $\mathbf{A}$ were a nonzero number instead of a matrix, we would divide both sides of (6) by $\mathbf{A}$ to obtain that $\mathbf{x}=\mathbf{b}/\mathbf{A}$. This expression, of course, does not make sense if $\mathbf{A}$ is a matrix. However, there is a way of deriving the solution $\mathbf{x}=\mathbf{b}/\mathbf{A}$ which *does* generalize to the case where $\mathbf{A}$ is an $n\times n$ matrix. To wit, if the number $\mathbf{A}$ were unequal to zero, then we can multiply both sides of (6) by the number $\mathbf{A}^{-1}$ to obtain that

$$\mathbf{A}^{-1}\mathbf{A}\mathbf{x}=\mathbf{x}=\mathbf{A}^{-1}\mathbf{b}.$$

Now, if we can define $\mathbf{A}^{-1}$ as an $n\times n$ matrix when $\mathbf{A}$ is an $n\times n$ matrix, then the expression $\mathbf{A}^{-1}\mathbf{b}$ would make perfectly good sense. This leads us to ask the following two questions.

Question 1: Given an $n\times n$ matrix $\mathbf{A}$, does there exist another $n\times n$ matrix, which we will call $\mathbf{A}^{-1}$, with the property that

$$\mathbf{A}^{-1}\mathbf{A}=\mathbf{A}\mathbf{A}^{-1}=\mathbf{I}?$$

Question 2: If $\mathbf{A}^{-1}$ exists, is it unique? That is to say, can there exist two *distinct* matrices $\mathbf{B}$ and $\mathbf{C}$ with the property that

$$\mathbf{BA}=\mathbf{AB}=\mathbf{I}\quad\text{and}\quad\mathbf{CA}=\mathbf{AC}=\mathbf{I}?$$

The answers to these two questions are supplied in Theorems 7 and 8.

Theorem 7. *An $n\times n$ matrix* $\mathbf{A}$ *has at most one inverse.*

PROOF. Suppose that $\mathbf{A}$ has two distinct inverses. Then, there exist two distinct matrices $\mathbf{B}$ and $\mathbf{C}$ with the property that

$$\mathbf{AB}=\mathbf{BA}=\mathbf{I}\quad\text{and}\quad\mathbf{AC}=\mathbf{CA}=\mathbf{I}.$$

If we multiply both sides of the equation $\mathbf{AC}=\mathbf{I}$ by $\mathbf{B}$ we obtain that

$$\mathbf{BI}=\mathbf{B(AC)}=\mathbf{(BA)C}=\mathbf{IC}.$$

Hence, $\mathbf{B}=\mathbf{C}$, which is a contradiction. □

Theorem 8. $\mathbf{A}^{-1}$ *exists if, and only if,* $\det\mathbf{A}\neq 0$, *and in this case*

$$\mathbf{A}^{-1}=\frac{1}{\det\mathbf{A}}\operatorname{adj}\mathbf{A}. \tag{7}$$

PROOF. Suppose that $\det\mathbf{A}\neq 0$. Then, we can divide both sides of the identities (4) and (5) by $\det\mathbf{A}$ to obtain that

$$\frac{\operatorname{adj}\mathbf{A}}{\det\mathbf{A}}\times\mathbf{A}=\mathbf{I}=\frac{1}{\det\mathbf{A}}\mathbf{A}\times\operatorname{adj}\mathbf{A}=\mathbf{A}\times\frac{\operatorname{adj}\mathbf{A}}{\det\mathbf{A}}.$$

Hence, $\mathbf{A}^{-1}=\operatorname{adj}\mathbf{A}/\det\mathbf{A}$.

Conversely, suppose that $\mathbf{A}^{-1}$ exists. Taking determinants of both sides of the equation $\mathbf{A}^{-1}\mathbf{A}=\mathbf{I}$, and using Property 8 gives

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

But this equation implies that $\det\mathbf{A}$ cannot equal zero. □

Finally, suppose that $\det\mathbf{A}\neq 0$. Then, $\mathbf{A}^{-1}$ exists, and multiplying both sides of (6) by this matrix gives

$$\mathbf{A}^{-1}\mathbf{Ax}=\mathbf{Ix}=\mathbf{x}=\mathbf{A}^{-1}\mathbf{b}.$$

Hence, if a solution exists, it must be $\mathbf{A}^{-1}\mathbf{b}$. Moreover, this vector is a solution of (6) since

$$\mathbf{A}(\mathbf{A}^{-1}\mathbf{b})=\mathbf{AA}^{-1}\mathbf{b}=\mathbf{Ib}=\mathbf{b}.$$

Thus, Equation (6) has a unique solution $\mathbf{x}=\mathbf{A}^{-1}\mathbf{b}$ if $\det\mathbf{A}\neq 0$.

Example 4. Find all solutions of the equation

$$\begin{bmatrix} 1 & 1 & 1\\ 2 & -2 & 2\\ 3 & 3 & -3\end{bmatrix}\mathbf{x}=\begin{bmatrix}0\\0\\0\end{bmatrix},\qquad \mathbf{x}=\begin{bmatrix}x_1\\x_2\\x_3\end{bmatrix}. \tag{8}$$

Solution.

$$\det\begin{bmatrix} 1 & 1 & 1\\ 2 & -2 & 2\\ 3 & 3 & -3\end{bmatrix}=\det\begin{bmatrix} 1 & 1 & 1\\ 0 & -4 & 0\\ 0 & 0 & -6\end{bmatrix}=24.$$

Hence, Equation (8) has a *unique* solution $\mathbf{x}$. But

$$\mathbf{x}=\begin{bmatrix}0\\0\\0\end{bmatrix}$$

is obviously one solution. Therefore,

$$\mathbf{x}=\begin{bmatrix}0\\0\\0\end{bmatrix}$$

is the unique solution of (8).

Remark. It is often quite cumbersome and time consuming to compute the inverse of an $n\times n$ matrix $\mathbf{A}$ from (7). This is especially true for $n\geqslant 4$. An alternate, and much more efficient way of computing $\mathbf{A}^{-1}$, is by means of "elementary row operations."

Definition. An elementary row operation on a matrix $\mathbf{A}$ is either
(i) an interchange of two rows,
(ii) the multiplication of one row by a nonzero number,
or
(iii) the addition of a multiple of one row to another row.

It can be shown that every matrix $\mathbf{A}$, with $\det\mathbf{A}\neq 0$, can be transformed into the identity $\mathbf{I}$ by a systematic sequence of these operations. Moreover, if the same sequence of operations is then performed upon $\mathbf{I}$, it is transformed into $\mathbf{A}^{-1}$. We illustrate this method with the following example.

Example 5. Find the inverse of the matrix

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

Solution. The matrix $\mathbf{A}$ can be transformed into $\mathbf{I}$ by the following sequence of elementary row operations. The result of each step appears below the operation performed.

(a) We obtain zeros in the off-diagonal positions in the first column by subtracting the first row from both the second and third rows.

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

(b) We obtain zeros in the off-diagonal positions in the second column by adding (-2) times the second row to the third row.

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

(c) We obtain a one in the diagonal position in the third column by multiplying the third row by $\frac{1}{2}$.

$$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

(d) Finally, we obtain zeros in the off-diagonal positions in the third column by adding the third row to the first row.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

If we perform the same sequence of elementary row operations upon **I**, we obtain the following sequence of matrices:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & -2 & 1 \end{bmatrix},$$

$$\begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ \frac{1}{2} & -1 & \frac{1}{2} \end{bmatrix}, \begin{bmatrix} \frac{3}{2} & -1 & \frac{1}{2} \\ -1 & 1 & 0 \\ \frac{1}{2} & -1 & \frac{1}{2} \end{bmatrix}.$$

The last of these matrices is $\mathbf{A}^{-1}$.

Exercises

In each of Problems 1–4, compute **AB** and **BA** for the given matrices **A** and **B**.

1. $\mathbf{A}=\begin{pmatrix} 3 & 2 & 0 \\ 1 & 1 & 1 \\ 2 & 6 & 0 \end{pmatrix}, \quad \mathbf{B}=\begin{pmatrix} -1 & 0 & 9 \\ 2 & 1 & -1 \\ 0 & 6 & 2 \end{pmatrix}$

2. $\mathbf{A}=\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \quad \mathbf{B}=\begin{pmatrix} 4 & 3 \\ 2 & 1 \end{pmatrix}$

3. $\mathbf{A}=\begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \quad \mathbf{B}=\begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$

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

5. Show that $\mathbf{IA}=\mathbf{AI}=\mathbf{A}$ for all matrices **A**.

6. Show that any two diagonal matrices **A** and **B** *commute*, that is $\mathbf{AB}=\mathbf{BA}$, if **A** and **B** are diagonal matrices.

7. Suppose that $\mathbf{AD}=\mathbf{DA}$ for all matrices $\mathbf{A}$. Prove that $\mathbf{D}$ is a multiple of the identity matrix.

8. Prove that $\mathbf{A}\times\operatorname{adj}\mathbf{A}=\det\mathbf{A}\times\mathbf{I}$.

In each of Problems 9–14, find the inverse, if it exists, of the given matrix.

9. $\begin{pmatrix} -2 & 3 & 2 \\ 6 & 0 & 3 \\ 4 & 1 & -1 \end{pmatrix}$

10. $\begin{pmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{pmatrix}$

11. $\begin{pmatrix} 1 & 1 & 1 \\ -1 & i & -i \\ 2 & 1 & 1 \end{pmatrix}$

12. $\begin{pmatrix} 1 & 2 & -1 \\ 3 & 1 & 2 \\ 2 & 3 & -1 \end{pmatrix}$

13. $\begin{pmatrix} 1 & 1+i & 1-i \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}$

14. $\begin{pmatrix} 0 & 1 & 1 \\ -1 & 0 & 1 \\ -1 & -1 & 0 \end{pmatrix}$

15. Let

$$\mathbf{A}=\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

Show that

$$\mathbf{A}^{-1}=\frac{1}{\det\mathbf{A}}\begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$$

if $\det\mathbf{A}\neq 0$.

16. Show that $(\mathbf{AB})^{-1}=\mathbf{B}^{-1}\mathbf{A}^{-1}$ if $\det\mathbf{A}\times\det\mathbf{B}\neq 0$.

In each of Problems 17–20 show that $\mathbf{x}=\mathbf{0}$ is the unique solution of the given system of equations.

17. $$\begin{aligned} x_1-x_2-x_3&=0\\ 3x_1-x_2+2x_3&=0\\ 2x_1+2x_2+3x_3&=0 \end{aligned}$$

18. $$\begin{aligned} x_1+2x_2+4x_3&=0\\ x_2+x_3&=0\\ x_1+x_2+x_3&=0 \end{aligned}$$

19. $$\begin{aligned} x_1+2x_2-x_3\qquad&=0\\ 2x_1+3x_2+x_3-x_4&=0\\ -x_1\qquad+2x_3+2x_4&=0\\ 3x_1-x_2+x_3+3x_4&=0 \end{aligned}$$

20. $$\begin{aligned} x_1+2x_2-x_3+3x_4&=0\\ 2x_1+3x_2\qquad-x_4&=0\\ -x_1+x_2+2x_3+x_4&=0\\ -x_2+2x_3+3x_4&=0 \end{aligned}$$

3.7 Linear transformations

In the previous section we approached the problem of solving the equation

$$\mathbf{Ax}=\mathbf{b},\quad \mathbf{A}=\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix},\quad \mathbf{x}=\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix},\quad \mathbf{b}=\begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \tag{1}$$

by asking whether the matrix $\mathbf{A}^{-1}$ exists. This approach led us to the conclusion that Equation (1) has a unique solution $\mathbf{x}=\mathbf{A}^{-1}\mathbf{b}$ if $\det\mathbf{A}\neq 0$. In order to determine what happens when $\det \mathbf{A}=0$, we will approach the problem of solving (1) in an entirely different manner. To wit, we will look at the set $\mathbf{V}$ of all vectors obtained by multiplying every vector in R^n by $\mathbf{A}$ and see if $\mathbf{b}$ is in this set. Obviously, Equation (1) has at least one solution $\mathbf{x}$ if, and only if, $\mathbf{b}$ is in $\mathbf{V}$. We begin with the following simple but extremely useful lemma.

Lemma 1. *Let* $\mathbf{A}$ *be an* $n\times n$ *matrix with elements* a_{ij}, *and let* $\mathbf{x}$ *be a vector with components* $x_1, x_2, \dots, x_n$. *Let*

$$\mathbf{a}^j = \begin{bmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{bmatrix}$$

denote the jth *column of* $\mathbf{A}$. *Then*

$$\mathbf{A}\mathbf{x} = x_1\mathbf{a}^1 + x_2\mathbf{a}^2 + \dots + x_n\mathbf{a}^n.$$

PROOF. We will show that the vectors $\mathbf{Ax}$ and $x_1\mathbf{a}^1 + \dots + x_n\mathbf{a}^n$ have the same components. To this end, observe that $(\mathbf{Ax})_j$, the jth component of $\mathbf{Ax}$, is $a_{j1}x_1 + \dots + a_{jn}x_n$, while the jth component of the vector $x_1\mathbf{a}^1 + \dots + x_n\mathbf{a}^n$ is

$$x_1\mathbf{a}_j^1 + \dots + x_n\mathbf{a}_j^n = x_1 a_{j1} + \dots + x_n a_{jn} = (\mathbf{Ax})_j.$$

Hence, $\mathbf{A}x = x_1\mathbf{a}^1 + x_2\mathbf{a}^2 + \dots + x_n\mathbf{a}^n$. □

Now, let $\mathbf{V}$ be the set of vectors obtained by multiplying every vector $\mathbf{x}$ in R^n by the matrix $\mathbf{A}$. It follows immediately from Lemma 1 that $\mathbf{V}$ is the set of all linear combinations of the vectors $\mathbf{a}^1, \mathbf{a}^2, \dots, \mathbf{a}^n$, that is, $\mathbf{V}$ is spanned by the columns of $\mathbf{A}$. Hence the equation $\mathbf{Ax}=\mathbf{b}$ has a solution if, and only if, $\mathbf{b}$ is a linear combination of the columns of $\mathbf{A}$. With the aid of this observation, we can now prove the following theorem.

Theorem 9. (a) *The equation* $\mathbf{Ax}=\mathbf{b}$ *has a unique solution if the columns of* $\mathbf{A}$ *are linearly independent.*

(b) *The equation* $\mathbf{Ax}=\mathbf{b}$ *has either no solution, or infinitely many solutions, if the columns of* $\mathbf{A}$ *are linearly dependent.*

PROOF. (a) Suppose that the columns of $\mathbf{A}$ are linearly independent. Then, $\mathbf{a}^1, \mathbf{a}^2, \dots, \mathbf{a}^n$ form a basis for R^n. In particular, every vector $\mathbf{b}$ can be written as a linear combination of $\mathbf{a}^1, \mathbf{a}^2, \dots, \mathbf{a}^n$. Consequently, Equation (1) has at least one solution. To prove that (1) has exactly one solution, we show that any two solutions

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

and

$$\mathbf{y}=\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

must be equal. To this end, observe that if $\mathbf{x}$ and $\mathbf{y}$ are two solutions of (1), then

$$\mathbf{A}(\mathbf{x}-\mathbf{y})=\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{y}=\mathbf{b}-\mathbf{b}=\mathbf{0}.$$

By Lemma 1, therefore,

$$(x_1-y_1)\mathbf{a}^1+(x_2-y_2)\mathbf{a}^2+\ldots+(x_n-y_n)\mathbf{a}^n=\mathbf{0}. \tag{2}$$

But this implies that $x_1=y_1$, $x_2=y_2,\ldots,$ and $x_n=y_n$, since $\mathbf{a}^1,\mathbf{a}^2,\ldots,\mathbf{a}^n$ are linearly independent. Consequently, $\mathbf{x}=\mathbf{y}$.

(b) If the columns of $\mathbf{A}$ are linearly dependent, then we can extract from $\mathbf{a}^1,\mathbf{a}^2,\ldots,\mathbf{a}^n$ a smaller set of independent vectors which also span $\mathbf{V}$ (see Exercise 12, Section 3.3). Consequently, the space $\mathbf{V}$ has dimension at most $n-1$. In other words, the space $\mathbf{V}$ is distinctly smaller than R^n. Hence, there are vectors in R^n which are not in $\mathbf{V}$. If $\mathbf{b}$ is one of these vectors, then the equation $\mathbf{A}\mathbf{x}=\mathbf{b}$ obviously has no solutions. On the other hand, if $\mathbf{b}$ is in $\mathbf{V}$, that is, there exists at least one vector $\mathbf{x}^*$ such that $\mathbf{A}\mathbf{x}^*=\mathbf{b}$, then Equation (1) has infinitely many solutions. To prove this, observe first that $\mathbf{x}=\mathbf{x}^*$ is certainly one solution of (1). Second, observe that if $\mathbf{A}\boldsymbol{\xi}=\mathbf{0}$ for some vector $\boldsymbol{\xi}$ in R^n, then $\mathbf{x}=\mathbf{x}^*+\boldsymbol{\xi}$ is also a solution of (1), since

$$\mathbf{A}(\mathbf{x}^*+\boldsymbol{\xi})=\mathbf{A}\mathbf{x}^*+\mathbf{A}\boldsymbol{\xi}=\mathbf{b}+\mathbf{0}=\mathbf{b}.$$

Finally, observe that there exist numbers $c_1,c_2,\ldots,c_n$ not all zero, such that $c_1\mathbf{a}^1+c_2\mathbf{a}^2+\ldots+c_n\mathbf{a}^n=\mathbf{0}$. By Lemma 1, therefore, $\mathbf{A}\boldsymbol{\xi}=\mathbf{0}$, where

$$\boldsymbol{\xi}=\begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}.$$

But if $\mathbf{A}\boldsymbol{\xi}$ equals zero, then $\mathbf{A}(\alpha\boldsymbol{\xi})$ also equals zero, for any constant α. Thus, there are infinitely many vectors $\boldsymbol{\xi}$ with the property that $\mathbf{A}\boldsymbol{\xi}=\mathbf{0}$. Consequently, the equation $\mathbf{A}\mathbf{x}=\mathbf{b}$ has infinitely many solutions. □

Example 1. (a) For which vectors

$$\mathbf{b}=\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

can we solve the equation

$$\mathbf{A}\mathbf{x}=\begin{bmatrix} 1 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 3 \end{bmatrix}\mathbf{x}=\mathbf{b}?$$

(b) Find all solutions of the equation

$$\begin{bmatrix} 1 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 3 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

(c) Find all solutions of the equation

$$\begin{bmatrix} 1 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 3 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

Solution. (a) The columns of the matrix $\mathbf{A}$ are

$$\mathbf{a}^1 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}, \quad \mathbf{a}^2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{a}^3 = \begin{bmatrix} 2 \\ 1 \\ 3 \end{bmatrix}.$$

Notice that $\mathbf{a}^1$ and $\mathbf{a}^2$ are linearly independent while $\mathbf{a}^3$ is the sum of $\mathbf{a}^1$ and $\mathbf{a}^2$. Hence, we can solve the equation $\mathbf{Ax} = \mathbf{b}$ if, and only if,

$$\mathbf{b} = c_1 \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} + c_2 \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} c_1 + c_2 \\ c_1 \\ 2c_1 + c_2 \end{bmatrix}$$

for some constants c_1 and c_2. Equivalently, (see Exercise 25) $b_3 = b_1 + b_2$.

(b) Consider the three equations

$$\begin{aligned} x_1 + x_2 + 2x_3 &= 0 \\ x_1 \qquad + x_3 &= 0 \\ 2x_1 + x_2 + 3x_3 &= 0. \end{aligned}$$

Notice that the third equation is the sum of the first two equations. Hence, we need only consider the first two equations. The second equation says that $x_1 = -x_3$. Substituting this value of x_1 into the first equation gives $x_2 = -x_3$. Hence, all solutions of the equation $\mathbf{Ax} = \mathbf{0}$ are of the form

$$\mathbf{x} = c \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}.$$

(c) Observe first that

$$\mathbf{x}^1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

is clearly one solution of this equation. Next, let $\mathbf{x}^2$ be any other solution of this equation. It follows immediately that $\mathbf{x}^2 = \mathbf{x}^1 + \boldsymbol{\xi}$, where $\boldsymbol{\xi}$ is a solution of the homogeneous equation $\mathbf{Ax} = \mathbf{0}$. Moreover, the sum of any solution of the nonhomogeneous equation

$$\mathbf{Ax}=\begin{bmatrix}1\\0\\1\end{bmatrix}$$

with a solution of the homogeneous equation is again a solution of the nonhomogeneous equation. Hence, any solution of the equation

$$\mathbf{Ax}=\begin{bmatrix}1\\0\\1\end{bmatrix}$$

is of the form

$$\mathbf{x}=\begin{bmatrix}0\\1\\0\end{bmatrix}+c\begin{bmatrix}-1\\-1\\1\end{bmatrix}.$$

Theorem 9 is an extremely useful theorem since it establishes necessary and sufficient conditions for the equation $\mathbf{Ax}=\mathbf{b}$ to have a unique solution. However, it is often very difficult to apply Theorem 9 since it is usually quite difficult to determine whether n vectors are linearly dependent or linearly independent. Fortunately, we can relate the question of whether the columns of $\mathbf{A}$ are linearly dependent or linearly independent to the much simpler problem of determining whether the determinant of $\mathbf{A}$ is zero or nonzero. There are several different ways of accomplishing this. In this section, we will present a very elegant method which utilizes the important concept of a linear transformation.

Definition. A linear transformation $\mathcal{Q}$ taking R^n into R^n is a function which assigns to each vector $\mathbf{x}$ in R^n a new vector which we call $\mathcal{Q}(\mathbf{x})=\mathcal{Q}(x_1,\dots,x_n)$. Moreover, this association obeys the following rules.

$$\mathcal{Q}(c\mathbf{x})=c\mathcal{Q}(\mathbf{x}) \tag{i}$$

and

$$\mathcal{Q}(\mathbf{x}+\mathbf{y})=\mathcal{Q}(\mathbf{x})+\mathcal{Q}(\mathbf{y}). \tag{ii}$$

Example 2. The transformation

$$\mathbf{x}=\begin{bmatrix}x_1\\ \vdots\\ x_n\end{bmatrix}\rightarrow\mathcal{Q}(\mathbf{x})=\begin{bmatrix}x_1\\ \vdots\\ x_n\end{bmatrix}$$

is obviously a linear transformation of R^n into R^n since

$$\mathcal{Q}(c\mathbf{x})=c\mathbf{x}=c\mathcal{Q}(\mathbf{x})$$

and

$$\mathcal{Q}(\mathbf{x}+\mathbf{y})=\mathbf{x}+\mathbf{y}=\mathcal{Q}(\mathbf{x})+\mathcal{Q}(\mathbf{y}).$$

Example 3. The transformation

$$\mathbf{x}=\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \to \mathcal{A}(\mathbf{x})=\begin{bmatrix} x_1+x_2+x_3 \\ x_1+x_2-x_3 \\ x_1 \end{bmatrix}$$

is a linear transformation taking R^3 into R^3 since

$$\mathcal{A}(c\mathbf{x})=\begin{bmatrix} cx_1+cx_2+cx_3 \\ cx_1+cx_2-cx_3 \\ cx_1 \end{bmatrix}=c\begin{bmatrix} x_1+x_2+x_3 \\ x_1+x_2-x_3 \\ x_1 \end{bmatrix}=c\mathcal{A}(\mathbf{x})$$

and

$$\mathcal{A}(\mathbf{x}+\mathbf{y})=\begin{bmatrix} (x_1+y_1)+(x_2+y_2)+(x_3+y_3) \\ (x_1+y_1)+(x_2+y_2)-(x_3+y_3) \\ x_1+y_1 \end{bmatrix}$$

$$=\begin{bmatrix} x_1+x_2+x_3 \\ x_1+x_2-x_3 \\ x_1 \end{bmatrix}+\begin{bmatrix} y_1+y_2+y_3 \\ y_1+y_2-y_3 \\ y_1 \end{bmatrix}=\mathcal{A}(\mathbf{x})+\mathcal{A}(\mathbf{y}).$$

Example 4. Let $\mathcal{A}(x_1,x_2)$ be the point obtained from $\mathbf{x}=(x_1,x_2)$ by rotating $\mathbf{x}$ 30° in a counterclockwise direction. It is intuitively obvious that any rotation is a linear transformation. If the reader is not convinced of this, though, he can compute

$$\mathcal{A}(x_1,x_2)=\begin{bmatrix} \frac{1}{2}\sqrt{3}\,x_1-\frac{1}{2}x_2 \\ \frac{1}{2}x_1+\frac{1}{2}\sqrt{3}\,x_2 \end{bmatrix}$$

and then verify directly that $\mathcal{A}$ is a linear transformation taking R^2 into R^2.

Example 5. The transformation

$$\mathbf{x}=\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\to\begin{pmatrix} 1 \\ x_1^2+x_2^2 \end{pmatrix}$$

takes R^2 into R^2 but is not linear, since

$$\mathcal{A}(2\mathbf{x})=\mathcal{A}(2x_1,2x_2)=\begin{pmatrix} 1 \\ 4x_1^2+4x_2^2 \end{pmatrix}\neq 2\mathcal{A}(\mathbf{x}).$$

Now, every $n\times n$ matrix $\mathbf{A}$ defines, in a very natural manner, a linear transformation $\mathcal{A}$ taking $\mathsf{R}^n\to\mathsf{R}^n$. To wit, consider the transformation of $\mathsf{R}^n\to\mathsf{R}^n$ defined by

$$\mathbf{x}\to\mathcal{A}(\mathbf{x})=\mathbf{A}\mathbf{x}.$$

In Section 3.1, we showed that $\mathbf{A}(c\mathbf{x})=c\mathbf{A}\mathbf{x}$ and $\mathbf{A}(\mathbf{x}+\mathbf{y})=\mathbf{A}\mathbf{x}+\mathbf{A}\mathbf{y}$. Hence, the association $\mathbf{x}\to\mathcal{A}(\mathbf{x})=\mathbf{A}\mathbf{x}$ is linear. Conversely, any linear transformation $\mathcal{A}$ taking $\mathsf{R}^n\to\mathsf{R}^n$ must be of the form $\mathcal{A}(\mathbf{x})=\mathbf{A}\mathbf{x}$ for some matrix $\mathbf{A}$. This is the content of the following theorem.

Theorem 10. *Any linear transformation $\mathbf{x}\to\mathcal{A}(\mathbf{x})$ taking R^n into R^n must be of the form $\mathcal{A}(\mathbf{x})=\mathbf{A}\mathbf{x}$. In other words, given any linear transformation $\mathcal{A}$ taking R^n into R^n, we can find an $n\times n$ matrix $\mathbf{A}$ such that*

$$\mathcal{A}(\mathbf{x})=\mathbf{A}\mathbf{x}$$

for all $\mathbf{x}$.

Proof. Let $\mathbf{e}^j$ denote the vector whose jth component is one, and whose remaining components are zero, and let

$$\mathbf{a}^j=\mathcal{A}(\mathbf{e}^j).$$

We claim that

$$\mathcal{A}(\mathbf{x})=\mathbf{A}\mathbf{x},\quad \mathbf{A}=(\mathbf{a}^1,\mathbf{a}^2,\ldots,\mathbf{a}^n) \tag{3}$$

for every

$$\mathbf{x}=\begin{bmatrix} x_1\\ \vdots\\ x_n\end{bmatrix}$$

in R^n. To prove this, observe that any vector $\mathbf{x}$ can be written in the form $\mathbf{x}=x_1\mathbf{e}^1+\ldots+x_n\mathbf{e}^n$. Hence, by the linearity of $\mathcal{A}$,

$$\begin{aligned}\mathcal{A}(\mathbf{x})&=\mathcal{A}\left(x_1\mathbf{e}^1+\ldots+x_n\mathbf{e}^n\right)=x_1\mathcal{A}(\mathbf{e}^1)+\ldots+x_n\mathcal{A}(\mathbf{e}^n)\\ &=x_1\mathbf{a}^1+\ldots+x_n\mathbf{a}^n=\mathbf{A}\mathbf{x}.\end{aligned}$$

□

Remark 1. The simplest way of evaluating a linear transformation $\mathcal{A}$ is to compute $\mathbf{a}^1=\mathcal{A}(\mathbf{e}^1),\ldots,\mathbf{a}^n=\mathcal{A}(\mathbf{e}^n)$ and then observe from Theorem 10 and Lemma 1 that $\mathcal{A}(\mathbf{x})=\mathbf{A}\mathbf{x}$, where $\mathbf{A}=(\mathbf{a}^1,\mathbf{a}^2,\ldots,\mathbf{a}^n)$. Thus, to evaluate the linear transformation in Example 4 above, we observe that under a rotation of 30° in a counterclockwise direction, the point (1,0) goes into the point $(\frac12\sqrt3\,,\frac12)$ and the point (0,1) goes into the point $(-\frac12,\frac12\sqrt3\,)$. Hence, any point $\mathbf{x}=(x_1,x_2)$ goes into the point

$$\begin{bmatrix}\frac12\sqrt3 & -\frac12\\ \frac12 & \frac12\sqrt3\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}\frac12\sqrt3\,x_1-\frac12x_2\\ \frac12x_1+\frac12\sqrt3\,x_2\end{bmatrix}.$$

Remark 2. If $\mathcal{A}$ and $\mathcal{B}$ are linear transformations taking R^n into R^n, then the composition transformation $\mathcal{A}\circ\mathcal{B}$ defined by the relation

$$\mathcal{A}\circ\mathcal{B}(\mathbf{x})=\mathcal{A}(\mathcal{B}(\mathbf{x}))$$

is again a linear transformation taking $\mathsf{R}^n \to \mathsf{R}^n$. To prove this, observe that

$$\begin{aligned}\mathcal{A} \circ \mathcal{B}(c\mathbf{x}) &= \mathcal{A}(\mathcal{B}(c\mathbf{x})) = \mathcal{A}(c\mathcal{B}(\mathbf{x})) = c\mathcal{A}(\mathcal{B}(\mathbf{x}))\\ &= c\mathcal{A} \circ \mathcal{B}(\mathbf{x})\end{aligned}$$

and

$$\begin{aligned}\mathcal{A} \circ \mathcal{B}(\mathbf{x}+\mathbf{y}) &= \mathcal{A}(\mathcal{B}(\mathbf{x}+\mathbf{y})) = \mathcal{A}(\mathcal{B}(\mathbf{x}) + \mathcal{B}(\mathbf{y}))\\ &= \mathcal{A}(\mathcal{B}(\mathbf{x})) + \mathcal{A}(\mathcal{B}(\mathbf{y})) = \mathcal{A} \circ \mathcal{B}(\mathbf{x}) + \mathcal{A} \circ \mathcal{B}(\mathbf{y}).\end{aligned}$$

Moreover, it is a simple matter to show (see Exercise 15) that if $\mathcal{A}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ and $\mathcal{B}(\mathbf{x}) = \mathbf{B}\mathbf{x}$, then

$$\mathcal{A} \circ \mathcal{B}(\mathbf{x}) = \mathbf{A}\mathbf{B}\mathbf{x}. \tag{4}$$

Similarly, if $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ are 3 linear transformations taking R^n into R^n, with $\mathcal{A}(\mathbf{x}) = \mathbf{A}\mathbf{x}$, $\mathcal{B}(\mathbf{x}) = \mathbf{B}\mathbf{x}$, and $\mathcal{C}(\mathbf{x}) = \mathbf{C}\mathbf{x}$ then

$$(\mathcal{A} \circ \mathcal{B}) \circ \mathcal{C}(\mathbf{x}) = (\mathbf{A}\mathbf{B})\mathbf{C}\mathbf{x} \tag{5}$$

and

$$\mathcal{A} \circ (\mathcal{B} \circ \mathcal{C})(\mathbf{x}) = \mathbf{A}(\mathbf{B}\mathbf{C})\mathbf{x}. \tag{6}$$

Now, clearly, $(\mathcal{A} \circ \mathcal{B}) \circ \mathcal{C}(\mathbf{x}) = \mathcal{A} \circ (\mathcal{B} \circ \mathcal{C})(\mathbf{x})$. Hence,

$$(\mathbf{A}\mathbf{B})\mathbf{C}\mathbf{x} = \mathbf{A}(\mathbf{B}\mathbf{C})\mathbf{x}$$

for all vectors $\mathbf{x}$ in R^n. This implies (see Exercise 14) that

$$(\mathbf{A}\mathbf{B})\mathbf{C} = \mathbf{A}(\mathbf{B}\mathbf{C})$$

for any three $n \times n$ matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$.

In most applications, it is usually desirable, and often absolutely essential, that the inverse of a linear transformation exists. Heuristically, the inverse of a transformation $\mathcal{A}$ undoes the effect of $\mathcal{A}$. That is to say, if $\mathcal{A}(\mathbf{x}) = \mathbf{y}$, then the inverse transformation applied to $\mathbf{y}$ must yield $\mathbf{x}$. More precisely, we define $\mathcal{A}^{-1}(\mathbf{y})$ as the unique element $\mathbf{x}$ in R^n for which $\mathcal{A}(\mathbf{x}) = \mathbf{y}$. Of course, the transformation $\mathcal{A}^{-1}$ may not exist. There may be some vectors $\mathbf{y}$ with the property that $\mathbf{y} \neq \mathcal{A}(\mathbf{x})$ for all $\mathbf{x}$ in R^n. Or, there may be some vectors $\mathbf{y}$ which come from more than one $\mathbf{x}$, that is, $\mathcal{A}(\mathbf{x}^1) = \mathbf{y}$ and $\mathcal{A}(\mathbf{x}^2) = \mathbf{y}$. In both these cases, the transformation $\mathcal{A}$ does not possess an inverse. In fact, it is clear that $\mathcal{A}$ possesses an inverse, which we will call $\mathcal{A}^{-1}$ if, and only if, the equation $\mathcal{A}(\mathbf{x}) = \mathbf{y}$ has a unique solution $\mathbf{x}$ for every $\mathbf{y}$ in R^n. In addition, it is clear that if $\mathcal{A}$ is a linear transformation and $\mathcal{A}^{-1}$ exists, then $\mathcal{A}^{-1}$ must also be linear. To prove this, observe first that $\mathcal{A}^{-1}(c\mathbf{y}) = c\mathcal{A}^{-1}(\mathbf{y})$ since $\mathcal{A}(c\mathbf{x}) = c\mathbf{y}$ if $\mathcal{A}(\mathbf{x}) = \mathbf{y}$. Second, observe that $\mathcal{A}^{-1}(\mathbf{y}^1 + \mathbf{y}^2) = \mathcal{A}^{-1}(\mathbf{y}^1) + \mathcal{A}^{-1}(\mathbf{y}^2)$ since $\mathcal{A}(\mathbf{x}^1 + \mathbf{x}^2) = \mathbf{y}^1 + \mathbf{y}^2$ if $\mathcal{A}(\mathbf{x}^1) = \mathbf{y}^1$ and $\mathcal{A}(\mathbf{x}^2) = \mathbf{y}^2$. Thus $\mathcal{A}^{-1}$, if it exists, must be linear.

At this point the reader should feel that there is an intimate connection between the linear transformation $\mathcal{A}^{-1}$ and the matrix $\mathbf{A}^{-1}$. This is the content of the following lemma.

Lemma 2. *Let* $\mathbf{A}$ *be an* $n \times n$ *matrix and let* $\mathcal{A}$ *be the linear transformation defined by the equation* $\mathcal{A}(\mathbf{x}) = \mathbf{A}\mathbf{x}$. *Then,* $\mathcal{A}$ *has an inverse if, and only if, the matrix* $\mathbf{A}$ *has an inverse. Moreover, if* $\mathbf{A}^{-1}$ *exists, then* $\mathcal{A}^{-1}(\mathbf{x}) = \mathbf{A}^{-1}\mathbf{x}$.

PROOF. Suppose that $\mathcal{A}^{-1}$ exists. Clearly,

$$\mathcal{A} \circ \mathcal{A}^{-1}(\mathbf{x}) = \mathcal{A}^{-1} \circ \mathcal{A}(\mathbf{x}) = \mathbf{x} \tag{7}$$

and $\mathcal{A}^{-1}$ is linear. Moreover, there exists a matrix $\mathbf{B}$ with the property that $\mathcal{A}^{-1}(\mathbf{x}) = \mathbf{B}\mathbf{x}$. Therefore, from (4) and (7)

$$\mathbf{A}\mathbf{B}\mathbf{x} = \mathbf{B}\mathbf{A}\mathbf{x} = \mathbf{x}$$

for all $\mathbf{x}$ in R^n. But this immediately implies (see Exercise 14) that

$$\mathbf{A}\mathbf{B} = \mathbf{B}\mathbf{A} = \mathbf{I}.$$

Hence $\mathbf{B} = \mathbf{A}^{-1}$.

Conversely, suppose that $\mathbf{A}^{-1}$ exists. Then, the equation

$$\mathcal{A}(\mathbf{x}) = \mathbf{A}\mathbf{x} = \mathbf{y}$$

has a unique solution $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$ for all $\mathbf{y}$ in R^n. Thus, $\mathbf{A}^{-1}$ also exists, and

$$\mathcal{A}^{-1}(\mathbf{x}) = \mathbf{A}^{-1}\mathbf{x}.$$ □

We are now ready to relate the problem of determining whether the columns of an $n \times n$ matrix $\mathbf{A}$ are linearly dependent or linearly independent to the much simpler problem of determining whether the determinant of $\mathbf{A}$ is zero or nonzero.

Lemma 3. *The columns of an* $n \times n$ *matrix* $\mathbf{A}$ *are linearly independent if, and only if,* $\det \mathbf{A} \neq 0$.

PROOF. We prove Lemma 3 by the following complex, but very clever argument.

(1) The columns of $\mathbf{A}$ are linearly independent if, and only if, the equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ has a unique solution $\mathbf{x}$ for every $\mathbf{b}$ in R^n. This statement is just a reformulation of Theorem 9.
(2) From the remarks preceding Lemma 2, we conclude that the equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ has a unique solution $\mathbf{x}$ for every $\mathbf{b}$ if, and only if, the linear transformation $\mathcal{A}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ has an inverse.
(3) From Lemma 2, the linear transformation $\mathcal{A}$ has an inverse if, and only if, the matrix $\mathbf{A}^{-1}$ exists.
(4) Finally, the matrix $\mathbf{A}^{-1}$ exists if, and only if, $\det \mathbf{A} \neq 0$. This is the content of Theorem 8, Section 3.6. Therefore, we conclude that the columns of $\mathbf{A}$ are linearly independent if, and only if, $\det \mathbf{A} \neq 0$. □

We summarize the results of this section by the following theorem.

Theorem 11. *The equation* $\mathbf{Ax}=\mathbf{b}$ *has a unique solution* $\mathbf{x}=\mathbf{A}^{-1}\mathbf{b}$ *if* $\det\mathbf{A}\neq 0$. *The equation* $\mathbf{Ax}=\mathbf{b}$ *has either no solutions, or infinitely many solutions if* $\det\mathbf{A}=0$.

PROOF. Theorem 11 follows immediately from Theorem 9 and Lemma 3. □

Corollary. *The equation* $\mathbf{Ax}=\mathbf{0}$ *has a nontrivial solution (that is, a solution*

$$\mathbf{x}=\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

with not all the x_i *equal to zero) if, and only if,* $\det\mathbf{A}=0$.

PROOF. Observe that

$$\mathbf{x}=\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

is always one solution of the equation $\mathbf{Ax}=\mathbf{0}$. Hence, it is the only solution if $\det\mathbf{A}\neq 0$. On the other hand, there exist infinitely many solutions if $\det\mathbf{A}=0$, and all but one of these are nontrivial. □

Example 6. For which values of λ does the equation

$$\begin{bmatrix} 1 & \lambda & \lambda \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}\mathbf{x}=\mathbf{0}$$

have a nontrivial solution?

Solution.

$$\det\begin{bmatrix} 1 & \lambda & \lambda \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}=\lambda+\lambda-1-\lambda=\lambda-1.$$

Hence, the equation

$$\begin{bmatrix} 1 & \lambda & \lambda \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}\mathbf{x}=\mathbf{0}$$

has a nontrivial solution if, and only if, $\lambda=1$.

Remark 1. Everything we've said about the equation $\mathbf{Ax}=\mathbf{b}$ applies equally well when the elements of $\mathbf{A}$ and the components of $\mathbf{x}$ and $\mathbf{b}$ are complex numbers. In this case, we interpret $\mathbf{x}$ and $\mathbf{b}$ as vectors in $\mathbf{C}^n$, and the matrix $\mathbf{A}$ as inducing a linear transformation of $\mathbf{C}^n$ into itself.

Remark 2. Suppose that we seek to determine n numbers $x_1,x_2,\ldots,x_n$. Our intuitive feeling is that we must be given n equations which are satisfied by these unknowns. This is certainly the case if we are given n linear equations of the form

$$a_{j1}x_1 + a_{j2}x_2 + \dots + a_{jn}x_n = b_j, \quad j = 1,2,\dots,n \tag{9}$$

and $\det \mathbf{A} \neq 0$, where

$$\mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}.$$

On the other hand, our intuition would seem to be wrong when $\det \mathbf{A} = 0$. This is not the case, though. To wit, if $\det \mathbf{A} = 0$, then the columns of $\mathbf{A}$ are linearly dependent. But then the columns of $\mathbf{A}^T$, which are the rows of $\mathbf{A}$, are also linearly dependent, since $\det \mathbf{A}^T = \det \mathbf{A}$. Consequently, one of the rows of $\mathbf{A}$ is a linear combination of the other rows. Now, this implies that the left-hand side of one of the equations (9), say the kth equation, is a linear combination of the other left-hand sides. Obviously, the equation $\mathbf{Ax} = \mathbf{b}$ has no solution if b_k is not the exact same linear combination of $b_1, \dots, b_{k-1}, b_{k+1}, \dots, b_n$. For example, the system of equations

$$\begin{aligned} x_1 - x_2 + x_3 &= 1 \\ x_1 + x_2 + x_3 &= 1 \\ 2x_1 \qquad + 2x_3 &= 3 \end{aligned}$$

obviously has no solution. On the other hand, if b_k is the same linear combination of $b_1, \dots, b_{k-1}, b_{k+1}, \dots, b_n$, then the kth equation is redundant. In this case, therefore, we really have only $n-1$ equations for the n unknowns $x_1, x_2, \dots, x_n$.

Remark 3. Once we introduce the concept of a linear transformation we no longer need to view an $n \times n$ matrix as just a square array of numbers. Rather, we can now view an $n \times n$ matrix $\mathbf{A}$ as inducing a linear transformation $\mathcal{A}(\mathbf{x}) = \mathbf{Ax}$ on R^n. The benefit of this approach is that we can derive properties of $\mathbf{A}$ by deriving the equivalent properties of the linear transformation $\mathcal{A}$. For example, we showed that $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$ for any 3 $n \times n$ matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ by showing that the induced linear transformations $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ satisfy the relation $(\mathcal{A} \circ \mathcal{B}) \circ \mathcal{C} = \mathcal{A} \circ (\mathcal{B} \circ \mathcal{C})$. Now, this result can be proven directly, but it requires a great deal more work (see Exercise 24).

Exercises

In each of Problems 1–3 find all vectors $\mathbf{b}$ for which the given system of equations has a solution.

1. $\begin{pmatrix} 3 & -1 & 1 \\ 1 & 1 & 1 \\ 5 & 1 & 3 \end{pmatrix} \mathbf{x} = \mathbf{b}$

2. $\begin{bmatrix} 1 & 2 & 2 & 3 \\ 2 & 4 & 4 & 6 \\ 3 & 6 & 6 & 9 \\ 4 & 8 & 8 & 12 \end{bmatrix} \mathbf{x} = \mathbf{b}$

3. $\begin{pmatrix} 11 & 4 & 8 \\ 1 & -1 & 1 \\ 3 & 2 & 2 \end{pmatrix}\mathbf{x}=\mathbf{b}$

In each of Problems 4–9, find all solutions of the given system of equations.

4. $\begin{pmatrix} 1 & 2 & -3 \\ 2 & 3 & -1 \\ 1 & 3 & 10 \end{pmatrix}\mathbf{x}=\mathbf{0}$

5. $\begin{pmatrix} 1 & 2 & 1 \\ -3 & -2 & 1 \\ 6 & 8 & 2 \end{pmatrix}\mathbf{x}=\begin{pmatrix} 3 \\ 5 \\ 4 \end{pmatrix}$

6. $\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{bmatrix}\mathbf{x}=\mathbf{0}$

7. $\begin{pmatrix} 2 & -1 & -1 \\ -5 & 3 & 1 \\ -1 & 1 & -1 \end{pmatrix}\mathbf{x}=\begin{pmatrix} -1 \\ 2 \\ 3 \end{pmatrix}$

8. $\begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & -1 & 6 \\ 2 & 0 & 0 & 1 \\ 2 & 4 & 6 & 8 \end{bmatrix}\mathbf{x}=\mathbf{0}$

9. $\begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix}\mathbf{x}=\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$

In each of Problems 10–12, determine all values of λ for which the given system of equations has a nontrivial solution.

10. $\begin{pmatrix} \lambda & 1 & 3 \\ 1 & \lambda & 3 \\ -1 & 3 & \lambda \end{pmatrix}\mathbf{x}=\mathbf{0}$

11. $\begin{bmatrix} 1 & 1 & \lambda & 1 \\ 1 & -1 & \lambda & \lambda \\ 0 & 1 & 0 & 1 \\ -1 & 0 & -1 & 1 \end{bmatrix}\mathbf{x}=\mathbf{0}$

12. $\begin{pmatrix} \lambda & -1 & -1 \\ 2 & -1 & 0 \\ 3 & \lambda & 5 \end{pmatrix}\mathbf{x}=\mathbf{0}.$

13. (a) For which value of λ does the system of equations

$$\begin{pmatrix} -2 & -4 & -5 \\ 1 & -1 & -1 \\ 4 & 2 & 3 \end{pmatrix}\mathbf{x}=\begin{pmatrix} \lambda \\ 1 \\ 3 \end{pmatrix}$$

have a solution?

(b) Find all solutions for this value of λ.

14. Suppose that $\mathbf{Ax}=\mathbf{Bx}$ for all vectors

$$\mathbf{x}=\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}.$$

Prove that $\mathbf{A}=\mathbf{B}$.

15. Let $\mathcal{A}$ and $\mathcal{B}$ be two linear transformations taking R^n into R^n. Then, there exist $n\times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ such that $\mathcal{A}(\mathbf{x})=\mathbf{Ax}$ and $\mathcal{B}(\mathbf{x})=\mathbf{Bx}$. Show that $\mathcal{A}\circ\mathcal{B}(\mathbf{x})=\mathbf{ABx}$. *Hint*: $\mathcal{A}\circ\mathcal{B}$ is a linear transformation taking R^n into R^n. Hence, there exists an $n\times n$ matrix $\mathbf{C}$ such that $\mathcal{A}\circ\mathcal{B}(\mathbf{x})=\mathbf{Cx}$. The jth column of $\mathbf{C}$ is $\mathcal{A}\circ\mathcal{B}(\mathbf{e}^j)$. Thus show that $\mathcal{A}\circ\mathcal{B}(\mathbf{e}^j)$ is the jth column of the matrix $\mathbf{AB}$.

16. Let $\mathcal{A}$ be a linear transformation taking $\mathsf{R}^n \to \mathsf{R}^n$. Show that $\mathcal{A}(\mathbf{0}) = \mathbf{0}$.

17. Let $\mathfrak{R}(\theta)$ be the linear transformation which rotates each point in the plane by an angle θ in the counterclockwise direction. Show that

$$\mathfrak{R}(\theta)(\mathbf{x}) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

18. Let $\mathfrak{R}_1$ and $\mathfrak{R}_2$ be the linear transformations which rotate each point in the plane by angles θ_1 and θ_2 respectively. Then the linear transformation $\mathfrak{R}_3 = \mathfrak{R}_1 \circ \mathfrak{R}_2$ rotates each point in the plane by an angle $\theta_1 + \theta_2$ (in the counterclockwise direction). Using Exercise (15), show that

$$\begin{pmatrix} \cos(\theta_1+\theta_2) & -\sin(\theta_1+\theta_2) \\ \sin(\theta_1+\theta_2) & \cos(\theta_1+\theta_2) \end{pmatrix} = \begin{pmatrix} \cos\theta_1 & -\sin\theta_1 \\ \sin\theta_1 & \cos\theta_1 \end{pmatrix} \begin{pmatrix} \cos\theta_2 & -\sin\theta_2 \\ \sin\theta_2 & \cos\theta_2 \end{pmatrix}.$$

Thus, derive the trigonometric identities

$$\sin(\theta_1+\theta_2) = \sin\theta_1\cos\theta_2 + \cos\theta_1\sin\theta_2$$
$$\cos(\theta_1+\theta_2) = \cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2.$$

19. Let

$$\mathcal{A}(x_1,x_2) = \begin{pmatrix} x_1 + x_2 \\ x_1 - x_2 \end{pmatrix}.$$

(a) Verify that $\mathcal{A}$ is linear.
(b) Show that every point (x_1,x_2) on the unit circle $x_1^2 + x_2^2 = 1$ goes into a point on the circle, $x_1^2 + x_2^2 = 2$.

20. Let $\mathbf{V}$ be the space of all polynomials $p(t)$ of degree less than or equal to 3 and let $(Dp)(t) = dp(t)/dt$.
(a) Show that D is a linear transformation taking $\mathbf{V}$ into $\mathbf{V}$.
(b) Show that D does not possess an inverse.

21. Let $\mathbf{V}$ be the space of all continuous functions $f(t)$, $-\infty < t < \infty$ and let $(Kf)(t) = \int_0^t f(s)\,ds$.
(a) Show that K is a linear transformation taking $\mathbf{V}$ into $\mathbf{V}$.
(b) Show that $(DK)f = f$ where $Df = f'$.
(c) Let $f(t)$ be differentiable. Show that

$$[(KD)f](t) = f(t) - f(0).$$

22. A linear transformation $\mathcal{A}$ is said to be $1-1$ if $\mathcal{A}(\mathbf{x}) \neq \mathcal{A}(\mathbf{y})$ whenever $\mathbf{x} \neq \mathbf{y}$. In other words, no two vectors go into the same vector under $\mathcal{A}$. Show that $\mathcal{A}$ is $1-1$ if, and only if, $\mathcal{A}(\mathbf{x}) = \mathbf{0}$ implies that $\mathbf{x} = \mathbf{0}$.

23. A linear transformation $\mathcal{A}$ is said to be *onto* if the equation $\mathcal{A}(\mathbf{x}) = \mathbf{y}$ has at least one solution for every $\mathbf{y}$ in R^n. Prove that $\mathcal{A}$ is onto if, and only if, $\mathcal{A}$ is $1-1$. *Hint*: Show first that $\mathcal{A}$ is onto if, and only if, the vectors $\mathcal{A}(\mathbf{e}^1), \ldots, \mathcal{A}(\mathbf{e}^n)$ are linearly independent. Then, use Lemma 1 to show that we can find a nonzero solution of the equation $\mathcal{A}(\mathbf{x}) = \mathbf{0}$ if $\mathcal{A}(\mathbf{e}^1), \ldots, \mathcal{A}(\mathbf{e}^n)$ are linearly dependent. Finally, show that $\mathcal{A}(\mathbf{e}^1), \ldots, \mathcal{A}(\mathbf{e}^n)$ are linearly dependent if the equation $\mathcal{A}(\mathbf{x}) = \mathbf{0}$ has a nonzero solution.

24. Prove directly that $(\mathbf{AB})\mathbf{C}=\mathbf{A}(\mathbf{BC})$. *Hint*: Show that these matrices have the same elements.

25. Show that

$$\mathbf{b}=\begin{bmatrix} b_1\\ b_2\\ b_3\end{bmatrix}=c_1\begin{pmatrix}1\\1\\2\end{pmatrix}+c_2\begin{pmatrix}1\\0\\1\end{pmatrix}$$

if, and only if, $b_3=b_1+b_2$.

3.8 The eigenvalue–eigenvector method of finding solutions

We return now to the first-order linear homogeneous differential equation

$$\dot{\mathbf{x}}=\mathbf{A}\mathbf{x},\quad \mathbf{x}=\begin{bmatrix} x_1\\ \vdots\\ x_n\end{bmatrix},\quad \mathbf{A}=\begin{bmatrix} a_{11} & \cdots & a_{1n}\\ \vdots & & \vdots\\ a_{n1} & \cdots & a_{nn}\end{bmatrix}. \tag{1}$$

Our goal is to find n linearly independent solutions $\mathbf{x}^1(t),\dots,\mathbf{x}^n(t)$. Now, recall that both the first-order and second-order linear homogeneous scalar equations have exponential functions as solutions. This suggests that we try $\mathbf{x}(t)=e^{\lambda t}\mathbf{v}$, where $\mathbf{v}$ is a constant vector, as a solution of (1). To this end, observe that

$$\frac{d}{dt}e^{\lambda t}\mathbf{v}=\lambda e^{\lambda t}\mathbf{v}$$

and

$$\mathbf{A}(e^{\lambda t}\mathbf{v})=e^{\lambda t}\mathbf{A}\mathbf{v}.$$

Hence, $\mathbf{x}(t)=e^{\lambda t}\mathbf{v}$ is a solution of (1) if, and only if, $\lambda e^{\lambda t}\mathbf{v}=e^{\lambda t}\mathbf{A}\mathbf{v}$. Dividing both sides of this equation by $e^{\lambda t}$ gives

$$\mathbf{A}\mathbf{v}=\lambda\mathbf{v}. \tag{2}$$

Thus, $\mathbf{x}(t)=e^{\lambda t}\mathbf{v}$ is a solution of (1) if, and only if, λ and $\mathbf{v}$ satisfy (2).

Definition. A nonzero vector $\mathbf{v}$ satisfying (2) is called an *eigenvector* of $\mathbf{A}$ with *eigenvalue* λ.

Remark. The vector $\mathbf{v}=\mathbf{0}$ is excluded because it is uninteresting. Obviously, $\mathbf{A0}=\lambda\cdot\mathbf{0}$ for any number λ.

An eigenvector of a matrix $\mathbf{A}$ is a rather special vector: under the linear transformation $\mathbf{x}\to\mathbf{A}\mathbf{x}$, it goes into a multiple λ of itself. Vectors which are transformed into multiples of themselves play an important role in many applications. To find such vectors, we rewrite Equation (2) in the form

$$\mathbf{0}=\mathbf{A}\mathbf{v}-\lambda\mathbf{v}=(\mathbf{A}-\lambda\mathbf{I})\mathbf{v}. \tag{3}$$

But, Equation (3) has a nonzero solution $\mathbf{v}$ only if $\det(\mathbf{A}-\lambda\mathbf{I})=0$. Hence the eigenvalues λ of $\mathbf{A}$ are the roots of the equation

$$0=\det(\mathbf{A}-\lambda\mathbf{I})=\det\begin{bmatrix} a_{11}-\lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22}-\lambda & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn}-\lambda \end{bmatrix},$$

and the eigenvectors of $\mathbf{A}$ are then the nonzero solutions of the equations $(\mathbf{A}-\lambda\mathbf{I})\mathbf{v}=\mathbf{0}$, for these values of λ.

The determinant of the matrix $\mathbf{A}-\lambda\mathbf{I}$ is clearly a polynomial in λ of degree n, with leading term $(-1)^n\lambda^n$. It is customary to call this polynomial the characteristic polynomial of $\mathbf{A}$ and to denote it by $p(\lambda)$. For each root λ_j of $p(\lambda)$, that is, for each number λ_j such that $p(\lambda_j)=0$, there exists at least one nonzero vector $\mathbf{v}^j$ such that $\mathbf{A}\mathbf{v}^j=\lambda_j\mathbf{v}^j$. Now, every polynomial of degree $n\geqslant 1$ has at least one (possibly complex) root. Therefore, every matrix has at least one eigenvalue, and consequently, at least one eigenvector. On the other hand, $p(\lambda)$ has at most n distinct roots. Therefore, every $n\times n$ matrix has at most n eigenvalues. Finally, observe that every $n\times n$ matrix has at most n linearly independent eigenvectors, since the space of all vectors

$$\mathbf{v}=\begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$$

has dimension n.

Remark. Let $\mathbf{v}$ be an eigenvector of $\mathbf{A}$ with eigenvalue λ. Observe that

$$\mathbf{A}(c\mathbf{v})=c\mathbf{A}\mathbf{v}=c\lambda\mathbf{v}=\lambda(c\mathbf{v})$$

for any constant c. Hence, any constant multiple $(c\neq 0)$ of an eigenvector of $\mathbf{A}$ is again an eigenvector of $\mathbf{A}$, with the same eigenvalue.

For each eigenvector $\mathbf{v}^j$ of $\mathbf{A}$ with eigenvalue λ_j, we have a solution $\mathbf{x}^j(t)=e^{\lambda_j t}\mathbf{v}^j$ of (1). If $\mathbf{A}$ has n linearly independent eigenvectors $\mathbf{v}^1,\dots,\mathbf{v}^n$ with eigenvalues $\lambda_1,\dots,\lambda_n$ respectively ($\lambda_1,\dots,\lambda_n$ need not be distinct), then $\mathbf{x}^j(t)=e^{\lambda_j t}\mathbf{v}^j$, $j=1,\dots,n$ are n linearly independent solutions of (1). This follows immediately from Theorem 6 of Section 3.4 and the fact that $\mathbf{x}^j(0)=\mathbf{v}^j$. In this case, then, every solution $\mathbf{x}(t)$ of (1) is of the form

$$\mathbf{x}(t)=c_1e^{\lambda_1 t}\mathbf{v}^1+c_2e^{\lambda_2 t}\mathbf{v}^2+\dots+c_ne^{\lambda_n t}\mathbf{v}^n. \tag{4}$$

This is sometimes called the "general solution" of (1).

The situation is simplest when $\mathbf{A}$ has n distinct real eigenvalues $\lambda_1,\lambda_2,\dots,\lambda_n$ with eigenvectors $\mathbf{v}^1,\mathbf{v}^2,\dots,\mathbf{v}^n$ respectively, for in this case we are guaranteed that $\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^n$ are linearly independent. This is the content of Theorem 12.

Theorem 12. *Any k eigenvectors $\mathbf{v}^1, \ldots, \mathbf{v}^k$ of $\mathbf{A}$ with distinct eigenvalues $\lambda_1, \ldots, \lambda_k$ respectively, are linearly independent.*

PROOF. We will prove Theorem 12 by induction on k, the number of eigenvectors. Observe that this theorem is certainly true for $k=1$. Next, we assume that Theorem 12 is true for $k=j$. That is to say, we assume that any set of j eigenvectors of $\mathbf{A}$ with distinct eigenvalues is linearly independent. We must show that any set of $j+1$ eigenvectors of $\mathbf{A}$ with distinct eigenvalues is also linearly independent. To this end, let $\mathbf{v}^1, \ldots, \mathbf{v}^{j+1}$ be $j+1$ eigenvectors of $\mathbf{A}$ with distinct eigenvalues $\lambda_1, \ldots, \lambda_{j+1}$ respectively. To determine whether these vectors are linearly dependent or linearly independent, we consider the equation

$$c_1\mathbf{v}^1 + c_2\mathbf{v}^2 + \ldots + c_{j+1}\mathbf{v}^{j+1} = \mathbf{0}. \tag{5}$$

Applying $\mathbf{A}$ to both sides of (5) gives

$$\lambda_1 c_1\mathbf{v}^1 + \lambda_2 c_2\mathbf{v}^2 + \ldots + \lambda_{j+1} c_{j+1}\mathbf{v}^{j+1} = \mathbf{0}. \tag{6}$$

Thus, if we multiply both sides of (5) by λ_1 and subtract the resulting equation from (6), we obtain that

$$(\lambda_2 - \lambda_1)c_2\mathbf{v}^2 + \ldots + (\lambda_{j+1} - \lambda_1)c_{j+1}\mathbf{v}^{j+1} = \mathbf{0}. \tag{7}$$

But $\mathbf{v}^2, \ldots, \mathbf{v}^{j+1}$ are j eigenvectors of $\mathbf{A}$ with distinct eigenvalues $\lambda_2, \ldots, \lambda_{j+1}$ respectively. By the induction hypothesis, they are linearly independent. Consequently,

$$(\lambda_2 - \lambda_1)c_2 = 0, \qquad (\lambda_3 - \lambda_1)c_3 = 0, \ldots, \quad \text{and} \quad (\lambda_{j+1} - \lambda_1)c_{j+1} = 0.$$

Since $\lambda_1, \lambda_2, \ldots, \lambda_{j+1}$ are distinct, we conclude that $c_2, c_3, \ldots, c_{j+1}$ are all zero. Equation (5) now forces c_1 to be zero. Hence, $\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^{j+1}$ are linearly independent. By induction, therefore, every set of k eigenvectors of $\mathbf{A}$ with distinct eigenvalues is linearly independent. □

Example 1. Find all solutions of the equation

$$\dot{\mathbf{x}} = \begin{pmatrix} 1 & -1 & 4 \\ 3 & 2 & -1 \\ 2 & 1 & -1 \end{pmatrix} \mathbf{x}.$$

Solution. The characteristic polynomial of the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 4 \\ 3 & 2 & -1 \\ 2 & 1 & -1 \end{pmatrix}$$

is

$$p(\lambda)=\det(\mathbf{A}-\lambda\mathbf{I})=\det\begin{bmatrix}1-\lambda & -1 & 4\\ 3 & 2-\lambda & -1\\ 2 & 1 & -1-\lambda\end{bmatrix}$$
$$=-(1+\lambda)(1-\lambda)(2-\lambda)+2+12-8(2-\lambda)+(1-\lambda)-3(1+\lambda)$$
$$=(1-\lambda)(\lambda-3)(\lambda+2).$$

Thus the eigenvalues of $\mathbf{A}$ are $\lambda_1=1$, $\lambda_2=3$, and $\lambda_3=-2$.

(i) $\lambda_1=1$: We seek a nonzero vector $\mathbf{v}$ such that

$$(\mathbf{A}-\mathbf{I})\mathbf{v}=\begin{bmatrix}0 & -1 & 4\\ 3 & 1 & -1\\ 2 & 1 & -2\end{bmatrix}\begin{bmatrix}v_1\\ v_2\\ v_3\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix}.$$

This implies that

$$-v_2+4v_3=0,\qquad 3v_1+v_2-v_3=0,\quad \text{and}\quad 2v_1+v_2-2v_3=0.$$

Solving for v_1 and v_2 in terms of v_3 from the first two equations gives $v_1=-v_3$ and $v_2=4v_3$. Hence, each vector

$$\mathbf{v}=c\begin{bmatrix}-1\\4\\1\end{bmatrix}$$

is an eigenvector of $\mathbf{A}$ with eigenvalue one. Consequently,

$$ce^{t}\begin{bmatrix}-1\\4\\1\end{bmatrix}$$

is a solution of the differential equation for any constant c. For simplicity, we take

$$\mathbf{x}^1(t)=e^{t}\begin{bmatrix}-1\\4\\1\end{bmatrix}.$$

(ii) $\lambda_2=3$: We seek a nonzero vector $\mathbf{v}$ such that

$$(\mathbf{A}-3\mathbf{I})\mathbf{v}=\begin{bmatrix}-2 & -1 & 4\\ 3 & -1 & -1\\ 2 & 1 & -4\end{bmatrix}\begin{bmatrix}v_1\\ v_2\\ v_3\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix}.$$

This implies that

$$-2v_1-v_2+4v_3=0,\qquad 3v_1-v_2-v_3=0,\quad \text{and}\quad 2v_1+v_2-4v_3=0.$$

Solving for v_1 and v_2 in terms of v_3 from the first two equations gives $v_1=v_3$ and $v_2=2v_3$. Consequently, each vector

$$\mathbf{v}=c\begin{bmatrix}1\\2\\1\end{bmatrix}$$

is an eigenvector of $\mathbf{A}$ with eigenvalue 3. Therefore,

$$\mathbf{x}^2(t)=e^{3t}\begin{bmatrix}1\\2\\1\end{bmatrix}$$

is a second solution of the differential equation.

(iii) $\lambda_3=-2$: We seek a nonzero vector $\mathbf{v}$ such that

$$(\mathbf{A}+2\mathbf{I})\mathbf{v}=\begin{bmatrix}3&-1&4\\3&4&-1\\2&1&1\end{bmatrix}\begin{bmatrix}v_1\\v_2\\v_3\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix}.$$

This implies that

$$3v_1-v_2+4v_3=0,\qquad 3v_1+4v_2-v_3=0\quad\text{and}\quad 2v_1+v_2+v_3=0.$$

Solving for v_1 and v_2 in terms of v_3 gives $v_1=-v_3$ and $v_2=v_3$. Hence, each vector

$$\mathbf{v}=c\begin{bmatrix}-1\\1\\1\end{bmatrix}$$

is an eigenvector of $\mathbf{A}$ with eigenvalue -2. Consequently,

$$\mathbf{x}^3(t)=e^{-2t}\begin{bmatrix}-1\\1\\1\end{bmatrix}$$

is a third solution of the differential equation. These solutions must be linearly independent, since $\mathbf{A}$ has distinct eigenvalues. Therefore, every solution $\mathbf{x}(t)$ must be of the form

$$\begin{aligned}\mathbf{x}(t)&=c_1e^{t}\begin{bmatrix}-1\\4\\1\end{bmatrix}+c_2e^{3t}\begin{bmatrix}1\\2\\1\end{bmatrix}+c_3e^{-2t}\begin{bmatrix}-1\\1\\1\end{bmatrix}\\&=\begin{bmatrix}-c_1e^{t}+c_2e^{3t}-c_3e^{-2t}\\4c_1e^{t}+2c_2e^{3t}+c_3e^{-2t}\\c_1e^{t}+c_2e^{3t}+c_3e^{-2t}\end{bmatrix}.\end{aligned}$$

Remark. If λ is an eigenvalue of $\mathbf{A}$, then the n equations

$$a_{j1}v_1+\ldots+(a_{jj}-\lambda)v_j+\ldots+a_{jn}v_n=0,\qquad j=1,\ldots,n$$

are not independent; at least one of them is a linear combination of the others. Consequently, we have at most $n-1$ independent equations for the n unknowns $v_1,\ldots,v_n$. This implies that at least one of the unknowns $v_1,\ldots,v_n$ can be chosen arbitrarily.

Example 2. Solve the initial-value problem

$$\dot{\mathbf{x}}=\begin{pmatrix}1 & 12\\ 3 & 1\end{pmatrix}\mathbf{x},\qquad \mathbf{x}(0)=\begin{pmatrix}0\\ 1\end{pmatrix}.$$

Solution. The characteristic polynomial of the matrix

$$\mathbf{A}=\begin{pmatrix}1 & 12\\ 3 & 1\end{pmatrix}$$

is

$$p(\lambda)=\det\begin{pmatrix}1-\lambda & 12\\ 3 & 1-\lambda\end{pmatrix}=(1-\lambda)^2-36=(\lambda-7)(\lambda+5).$$

Thus, the eigenvalues of A are $\lambda_1=7$ and $\lambda_2=-5$.

(i) $\lambda_1=7$: We seek a nonzero vector $\mathbf{v}$ such that

$$(\mathbf{A}-7\mathbf{I})\mathbf{v}=\begin{pmatrix}-6 & 12\\ 3 & -6\end{pmatrix}\begin{pmatrix}v_1\\ v_2\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix}.$$

This implies that $v_1=2v_2$. Consequently, every vector

$$\mathbf{v}=c\begin{pmatrix}2\\ 1\end{pmatrix}$$

is an eigenvector of $\mathbf{A}$ with eigenvalue 7. Therefore,

$$\mathbf{x}^1(t)=e^{7t}\begin{pmatrix}2\\ 1\end{pmatrix}$$

is a solution of the differential equation.

(ii) $\lambda_2=-5$: We seek a nonzero vector $\mathbf{v}$ such that

$$(\mathbf{A}+5\mathbf{I})\mathbf{v}=\begin{pmatrix}6 & 12\\ 3 & 6\end{pmatrix}\begin{pmatrix}v_1\\ v_2\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix}.$$

This implies that $v_1=-2v_2$. Consequently,

$$\begin{pmatrix}-2\\ 1\end{pmatrix}$$

is an eigenvector of $\mathbf{A}$ with eigenvalue -5, and

$$\mathbf{x}^2(t)=e^{-5t}\begin{pmatrix}-2\\ 1\end{pmatrix}$$

is a second solution of the differential equation. These solutions are linearly independent since $\mathbf{A}$ has distinct eigenvalues. Hence, $\mathbf{x}(t)=c_1\mathbf{x}^1(t)+c_2\mathbf{x}^2(t)$. The constants c_1 and c_2 are determined from the initial condition

$$\begin{pmatrix}0\\ 1\end{pmatrix}=c_1\mathbf{x}^1(0)+c_2\mathbf{x}^2(0)=\begin{pmatrix}2c_1\\ c_1\end{pmatrix}+\begin{pmatrix}-2c_2\\ c_2\end{pmatrix}.$$

Thus, $2c_1 - 2c_2 = 0$ and $c_1 + c_2 = 1$. The solution of these two equations is $c_1 = \frac{1}{2}$ and $c_2 = \frac{1}{2}$. Consequently,

$$\mathbf{x}(t) = \frac{1}{2}e^{7t}\begin{pmatrix} 2 \\ 1 \end{pmatrix} + \frac{1}{2}e^{-5t}\begin{pmatrix} -2 \\ 1 \end{pmatrix} = \begin{bmatrix} e^{7t} - e^{-5t} \\ \frac{1}{2}e^{7t} + \frac{1}{2}e^{-5t} \end{bmatrix}.$$

Example 3. Find all solutions of the equation

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 & 3 \\ 4 & 4 & 4 & 4 & 4 \\ 5 & 5 & 5 & 5 & 5 \end{bmatrix}\mathbf{x}.$$

Solution. It is not necessary to compute the characteristic polynomial of $\mathbf{A}$ in order to find the eigenvalues and eigenvectors of $\mathbf{A}$. To wit, observe that

$$\mathbf{A}\mathbf{x} = \mathbf{A}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = (x_1 + x_2 + x_3 + x_4 + x_5)\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}.$$

Hence, any vector $\mathbf{x}$ whose components add up to zero is an eigenvector of $\mathbf{A}$ with eigenvalue 0. In particular

$$\mathbf{v}^1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ -1 \end{bmatrix}, \quad \mathbf{v}^2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ -1 \end{bmatrix}, \quad \mathbf{v}^3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ -1 \end{bmatrix}, \quad \text{and} \quad \mathbf{v}^4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ -1 \end{bmatrix}$$

are four independent eigenvectors of $\mathbf{A}$ with eigenvalue zero. Moreover, observe that

$$\mathbf{v}^5 = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}$$

is an eigenvector of $\mathbf{A}$ with eigenvalue 15 since

$$\mathbf{A}\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} = (1+2+3+4+5)\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} = 15\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}.$$

The five vectors $\mathbf{v}^1$, $\mathbf{v}^2$, $\mathbf{v}^3$, $\mathbf{v}^4$, and $\mathbf{v}^5$ are easily seen to be linearly independent. Hence, every solution $\mathbf{x}(t)$ is of the form

$$\mathbf{x}(t)=c_1\begin{bmatrix}1\\0\\0\\0\\-1\end{bmatrix}+c_2\begin{bmatrix}0\\1\\0\\0\\-1\end{bmatrix}+c_3\begin{bmatrix}0\\0\\1\\0\\-1\end{bmatrix}+c_4\begin{bmatrix}0\\0\\0\\1\\-1\end{bmatrix}+c_5e^{15t}\begin{bmatrix}1\\2\\3\\4\\5\end{bmatrix}.$$

Exercises

In each of Problems 1–6 find all solutions of the given differential equation.

1. $\dot{\mathbf{x}}=\begin{pmatrix}6 & -3\\ 2 & 1\end{pmatrix}\mathbf{x}$

2. $\dot{\mathbf{x}}=\begin{pmatrix}-2 & 1\\ -4 & 3\end{pmatrix}\mathbf{x}$

3. $\dot{\mathbf{x}}=\begin{pmatrix}3 & 2 & 4\\ 2 & 0 & 2\\ 4 & 2 & 3\end{pmatrix}\mathbf{x}$

4. $\dot{\mathbf{x}}=\begin{pmatrix}7 & -1 & 6\\ -10 & 4 & -12\\ -2 & 1 & -1\end{pmatrix}\mathbf{x}$

5. $\dot{\mathbf{x}}=\begin{pmatrix}-7 & 0 & 6\\ 0 & 5 & 0\\ 6 & 0 & 2\end{pmatrix}\mathbf{x}$

6. $\dot{\mathbf{x}}=\begin{bmatrix}1 & 2 & 3 & 6\\ 3 & 6 & 9 & 18\\ 5 & 10 & 15 & 30\\ 7 & 14 & 21 & 42\end{bmatrix}\mathbf{x}$

In each of Problems 7–12, solve the given initial-value problem.

7. $\dot{\mathbf{x}}=\begin{pmatrix}1 & 1\\ 4 & 1\end{pmatrix}\mathbf{x},\quad \mathbf{x}(0)=\begin{pmatrix}2\\3\end{pmatrix}$

8. $\dot{\mathbf{x}}=\begin{pmatrix}1 & -3\\ -2 & 2\end{pmatrix}\mathbf{x},\quad \mathbf{x}(0)=\begin{pmatrix}0\\5\end{pmatrix}$

9. $\dot{\mathbf{x}}=\begin{pmatrix}3 & 1 & -1\\ 1 & 3 & -1\\ 3 & 3 & -1\end{pmatrix}\mathbf{x},\quad \mathbf{x}(0)=\begin{pmatrix}1\\-2\\-1\end{pmatrix}$

10. $\dot{\mathbf{x}}=\begin{pmatrix}1 & -1 & 0\\ 1 & 2 & 1\\ 1 & 10 & 2\end{pmatrix}\mathbf{x},\quad \mathbf{x}(0)=\begin{pmatrix}-1\\-4\\13\end{pmatrix}$

11. $\dot{\mathbf{x}}=\begin{pmatrix}1 & -3 & 2\\ 0 & -1 & 0\\ 0 & -1 & -2\end{pmatrix}\mathbf{x},\quad \mathbf{x}(0)=\begin{pmatrix}-2\\0\\3\end{pmatrix}$

12. $\dot{\mathbf{x}}=\begin{pmatrix}3 & 1 & -2\\ -1 & 2 & 1\\ 4 & 1 & -3\end{pmatrix}\mathbf{x},\quad \mathbf{x}(0)=\begin{pmatrix}1\\4\\-7\end{pmatrix}$

13. (a) Show that $e^{\lambda(t-t_0)}\mathbf{v}, t_0$ constant, is a solution of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$ if $\mathbf{A}\mathbf{v}=\lambda\mathbf{v}$.
(b) Solve the initial-value problem

$$\dot{\mathbf{x}}=\begin{pmatrix}3 & 1 & -2\\ -1 & 2 & 1\\ 4 & 1 & -3\end{pmatrix}\mathbf{x},\quad \mathbf{x}(1)=\begin{pmatrix}1\\4\\-7\end{pmatrix}$$

(see Exercise 12).

14. Three solutions of the equation $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$ are

$$\begin{bmatrix} e^{t}+e^{2t} \\ e^{2t} \\ 0 \end{bmatrix}, \quad \begin{bmatrix} e^{t}+e^{3t} \\ e^{3t} \\ e^{3t} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} e^{t}-e^{3t} \\ -e^{3t} \\ -e^{3t} \end{bmatrix}.$$

Find the eigenvalues and eigenvectors of $\mathbf{A}$.

15. Show that the eigenvalues of $\mathbf{A}^{-1}$ are the reciprocals of the eigenvalues of $\mathbf{A}$.

16. Show that the eigenvalues of $\mathbf{A}^{n}$ are the nth power of the eigenvalues of $\mathbf{A}$.

17. Show that $\lambda=0$ is an eigenvalue of $\mathbf{A}$ if $\det\mathbf{A}=0$.

18. Show, by example, that the eigenvalues of $\mathbf{A}+\mathbf{B}$ are not necessarily the sum of an eigenvalue of $\mathbf{A}$ and an eigenvalue of $\mathbf{B}$.

19. Show, by example, that the eigenvalues of $\mathbf{AB}$ are not necessarily the product of an eigenvalue of $\mathbf{A}$ with an eigenvalue of $\mathbf{B}$.

20. Show that the matrices $\mathbf{A}$ and $\mathbf{T}^{-1}\mathbf{AT}$ have the same characteristic polynomial.

21. Suppose that either $\mathbf{B}^{-1}$ or $\mathbf{A}^{-1}$ exists. Prove that $\mathbf{AB}$ and $\mathbf{BA}$ have the same eigenvalues. *Hint*: Use Exercise 20. (This result is true even if neither $\mathbf{B}^{-1}$ or $\mathbf{A}^{-1}$ exist; however, it is more difficult to prove then.)

3.9 Complex roots

If $\lambda=\alpha+i\beta$ is a complex eigenvalue of $\mathbf{A}$ with eigenvector $\mathbf{v}=\mathbf{v}^1+i\mathbf{v}^2$, then $\mathbf{x}(t)=e^{\lambda t}\mathbf{v}$ is a complex-valued solution of the differential equation

$$\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}. \tag{1}$$

This complex-valued solution gives rise to *two* real-valued solutions, as we now show.

Lemma 1. *Let* $\mathbf{x}(t)=\mathbf{y}(t)+i\mathbf{z}(t)$ *be a complex-valued solution of* (1). *Then, both* $\mathbf{y}(t)$ *and* $\mathbf{z}(t)$ *are real-valued solutions of* (1).

PROOF. If $\mathbf{x}(t)=\mathbf{y}(t)+i\mathbf{z}(t)$ is a complex-valued solution of (1), then

$$\dot{\mathbf{y}}(t)+i\dot{\mathbf{z}}(t)=\mathbf{A}(\mathbf{y}(t)+i\mathbf{z}(t))=\mathbf{A}\mathbf{y}(t)+i\mathbf{A}\mathbf{z}(t). \tag{2}$$

Equating real and imaginary parts of (2) gives $\dot{\mathbf{y}}(t)=\mathbf{A}\mathbf{y}(t)$ and $\dot{\mathbf{z}}(t)=\mathbf{A}\mathbf{z}(t)$. Consequently, both $\mathbf{y}(t)=\mathrm{Re}\{\mathbf{x}(t)\}$ and $\mathbf{z}(t)=\mathrm{Im}\{\mathbf{x}(t)\}$ are real-valued solutions of (1). □

The complex-valued function $\mathbf{x}(t)=e^{(\alpha+i\beta)t}(\mathbf{v}^1+i\mathbf{v}^2)$ can be written in the form

$$\begin{aligned}\mathbf{x}(t)&=e^{\alpha t}(\cos\beta t+i\sin\beta t)(\mathbf{v}^1+i\mathbf{v}^2)\\ &=e^{\alpha t}\left[(\mathbf{v}^1\cos\beta t-\mathbf{v}^2\sin\beta t)+i(\mathbf{v}^1\sin\beta t+\mathbf{v}^2\cos\beta t)\right].\end{aligned}$$

Hence, if $\lambda=\alpha+i\beta$ is an eigenvalue of $\mathbf{A}$ with eigenvector $\mathbf{v}=\mathbf{v}^1+i\mathbf{v}^2$, then

$$\mathbf{y}(t)=e^{\alpha t}(\mathbf{v}^1\cos\beta t-\mathbf{v}^2\sin\beta t)$$

and

$$\mathbf{z}(t)=e^{\alpha t}(\mathbf{v}^1\sin\beta t+\mathbf{v}^2\cos\beta t)$$

are two real-valued solutions of (1). Moreover, these two solutions must be linearly independent (see Exercise 10).

Example 1. Solve the initial-value problem

$$\dot{\mathbf{x}}=\begin{bmatrix}1&0&0\\0&1&-1\\0&1&1\end{bmatrix}\mathbf{x},\qquad \mathbf{x}(0)=\begin{bmatrix}1\\1\\1\end{bmatrix}.$$

Solution. The characteristic polynomial of the matrix

$$\mathbf{A}=\begin{bmatrix}1&0&0\\0&1&-1\\0&1&1\end{bmatrix}$$

is

$$p(\lambda)=\det(\mathbf{A}-\lambda\mathbf{I})=\det\begin{bmatrix}1-\lambda&0&0\\0&1-\lambda&-1\\0&1&1-\lambda\end{bmatrix}$$
$$=(1-\lambda)^3+(1-\lambda)=(1-\lambda)(\lambda^2-2\lambda+2).$$

Hence the eigenvalues of $\mathbf{A}$ are

$$\lambda=1\quad\text{and}\quad\lambda=\frac{2\pm\sqrt{4-8}}{2}=1\pm i.$$

(i) $\lambda=1$: Clearly,

$$\begin{bmatrix}1\\0\\0\end{bmatrix}$$

is an eigenvector of $\mathbf{A}$ with eigenvalue 1. Hence

$$\mathbf{x}^1(t)=e^t\begin{bmatrix}1\\0\\0\end{bmatrix}$$

is one solution of the differential equation $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$.

(ii) $\lambda=1+i$: We seek a nonzero vector $\mathbf{v}$ such that

$$\left[\mathbf{A}-(1+i)\mathbf{I}\right]\mathbf{v}=\begin{bmatrix}-i&0&0\\0&-i&-1\\0&1&-i\end{bmatrix}\begin{bmatrix}v_1\\v_2\\v_3\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix}.$$

This implies that $-iv_1=0$, $-iv_2-v_3=0$, and $v_2-iv_3=0$. The first equation says that $v_1=0$ and the second and third equations both say that $v_2=iv_3$. Consequently, each vector

$$\mathbf{v}=c\begin{bmatrix}0\\i\\1\end{bmatrix}$$

is an eigenvector of $\mathbf{A}$ with eigenvalue $1+i$. Thus,

$$\mathbf{x}(t)=e^{(1+i)t}\begin{bmatrix}0\\i\\1\end{bmatrix}$$

is a complex-valued solution of the differential equation $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$. Now,

$$\begin{aligned}
e^{(1+i)t}\begin{bmatrix}0\\i\\1\end{bmatrix}&=e^{t}(\cos t+i\sin t)\left[\begin{bmatrix}0\\0\\1\end{bmatrix}+i\begin{bmatrix}0\\1\\0\end{bmatrix}\right]\\
&=e^{t}\left[\cos t\begin{bmatrix}0\\0\\1\end{bmatrix}-\sin t\begin{bmatrix}0\\1\\0\end{bmatrix}+i\sin t\begin{bmatrix}0\\0\\1\end{bmatrix}+i\cos t\begin{bmatrix}0\\1\\0\end{bmatrix}\right]\\
&=e^{t}\begin{bmatrix}0\\-\sin t\\\cos t\end{bmatrix}+ie^{t}\begin{bmatrix}0\\\cos t\\\sin t\end{bmatrix}.
\end{aligned}$$

Consequently, by Lemma 1,

$$\mathbf{x}^2(t)=e^{t}\begin{bmatrix}0\\-\sin t\\\cos t\end{bmatrix}\quad\text{and}\quad\mathbf{x}^3(t)=e^{t}\begin{bmatrix}0\\\cos t\\\sin t\end{bmatrix}$$

are real-valued solutions. The three solutions $\mathbf{x}^1(t)$, $\mathbf{x}^2(t)$, and $\mathbf{x}^3(t)$ are linearly independent since their initial values

$$\mathbf{x}^1(0)=\begin{bmatrix}1\\0\\0\end{bmatrix},\qquad\mathbf{x}^2(0)=\begin{bmatrix}0\\1\\0\end{bmatrix},\quad\text{and}\quad\mathbf{x}^3(0)=\begin{bmatrix}0\\0\\1\end{bmatrix}$$

are linearly independent vectors in R^3. Therefore, the solution $\mathbf{x}(t)$ of our initial-value problem must have the form

$$\mathbf{x}(t)=c_1e^{t}\begin{bmatrix}1\\0\\0\end{bmatrix}+c_2e^{t}\begin{bmatrix}0\\-\sin t\\\cos t\end{bmatrix}+c_3e^{t}\begin{bmatrix}0\\\cos t\\\sin t\end{bmatrix}.$$

Setting $t=0$, we see that

$$\begin{bmatrix}1\\1\\1\end{bmatrix}=c_1\begin{bmatrix}1\\0\\0\end{bmatrix}+c_2\begin{bmatrix}0\\0\\1\end{bmatrix}+c_3\begin{bmatrix}0\\1\\0\end{bmatrix}=\begin{bmatrix}c_1\\c_3\\c_2\end{bmatrix}.$$

Consequently $c_1=c_2=c_3=1$ and

$$\mathbf{x}(t)=e^t\begin{bmatrix}1\\0\\0\end{bmatrix}+e^t\begin{bmatrix}0\\-\sin t\\\cos t\end{bmatrix}+e^t\begin{bmatrix}0\\\cos t\\\sin t\end{bmatrix}$$

$$=e^t\begin{bmatrix}1\\\cos t-\sin t\\\cos t+\sin t\end{bmatrix}.$$

Remark. If $\mathbf{v}$ is an eigenvector of $\mathbf{A}$ with eigenvalue λ, then $\bar{\mathbf{v}}$, the complex conjugate of $\mathbf{v}$, is an eigenvector of $\mathbf{A}$ with eigenvalue $\bar{\lambda}$. (Each component of $\bar{\mathbf{v}}$ is the complex conjugate of the corresponding component of $\mathbf{v}$.) To prove this, we take complex conjugates of both sides of the equation $\mathbf{A}\mathbf{v}=\lambda\mathbf{v}$ and observe that the complex conjugate of the vector $\mathbf{A}\mathbf{v}$ is $\mathbf{A}\bar{\mathbf{v}}$ if $\mathbf{A}$ is real. Hence, $\mathbf{A}\bar{\mathbf{v}}=\bar{\lambda}\bar{\mathbf{v}}$, which shows that $\bar{\mathbf{v}}$ is an eigenvector of $\mathbf{A}$ with eigenvalue $\bar{\lambda}$.

Exercises

In each of Problems 1–4 find the general solution of the given system of differential equations.

1. $\dot{\mathbf{x}}=\begin{pmatrix}-3&2\\-1&-1\end{pmatrix}\mathbf{x}$

2. $\dot{\mathbf{x}}=\begin{pmatrix}1&-5&0\\1&-3&0\\0&0&1\end{pmatrix}\mathbf{x}$

3. $\dot{\mathbf{x}}=\begin{pmatrix}1&0&0\\3&1&-2\\2&2&1\end{pmatrix}\mathbf{x}$

4. $\dot{\mathbf{x}}=\begin{pmatrix}1&0&1\\0&1&-1\\-2&0&-1\end{pmatrix}\mathbf{x}$

In each of Problems 5–8, solve the given initial-value problem.

5. $\dot{\mathbf{x}}=\begin{pmatrix}1&-1\\5&-3\end{pmatrix}\mathbf{x},\quad \mathbf{x}(0)=\begin{pmatrix}1\\2\end{pmatrix}$

6. $\dot{\mathbf{x}}=\begin{pmatrix}3&-2\\4&-1\end{pmatrix}\mathbf{x},\quad \mathbf{x}(0)=\begin{pmatrix}1\\5\end{pmatrix}$

7. $\dot{\mathbf{x}}=\begin{pmatrix}-3&0&2\\1&-1&0\\-2&-1&0\end{pmatrix}\mathbf{x},\quad \mathbf{x}(0)=\begin{pmatrix}0\\-1\\-2\end{pmatrix}$

8. $\dot{\mathbf{x}}=\begin{bmatrix}0&2&0&0\\-2&0&0&0\\0&0&0&-3\\0&0&3&0\end{bmatrix}\mathbf{x},\quad \mathbf{x}(0)=\begin{bmatrix}1\\1\\1\\0\end{bmatrix}$

9. Determine all vectors $\mathbf{x}^0$ such that the solution of the initial-value problem

$$\dot{\mathbf{x}}=\begin{pmatrix}1&0&-2\\0&1&0\\1&-1&-1\end{pmatrix}\mathbf{x},\qquad \mathbf{x}(0)=\mathbf{x}^0$$

is a periodic function of time.

10. Let $\mathbf{x}(t)=e^{\lambda t}\mathbf{v}$ be a solution of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$. Prove that $\mathbf{y}(t)=\operatorname{Re}\{\mathbf{x}(t)\}$ and $\mathbf{z}(t)=\operatorname{Im}\{\mathbf{x}(t)\}$ are linearly independent. *Hint*: Observe that $\mathbf{v}$ and $\bar{\mathbf{v}}$ are linearly independent in $\mathbf{C}^n$ since they are eigenvectors of $\mathbf{A}$ with distinct eigenvalues.

3.10 Equal roots

If the characteristic polynomial of $\mathbf{A}$ does not have n distinct roots, then $\mathbf{A}$ may not have n linearly independent eigenvectors. For example, the matrix

$$\mathbf{A}=\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

has only two distinct eigenvalues $\lambda_1=1$ and $\lambda_2=2$ and two linearly independent eigenvectors, which we take to be

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Consequently, the differential equation $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$ has only two linearly independent solutions

$$e^{t}\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad e^{2t}\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

of the form $e^{\lambda t}\mathbf{v}$. Our problem, in this case, is to find a third linearly independent solution. More generally, suppose that the $n\times n$ matrix $\mathbf{A}$ has only $k<n$ linearly independent eigenvectors. Then, the differential equation $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$ has only k linearly independent solutions of the form $e^{\lambda t}\mathbf{v}$. Our problem is to find an additional $n-k$ linearly independent solutions.

We approach this problem in the following ingenious manner. Recall that $x(t)=e^{at}c$ is a solution of the scalar differential equation $\dot{x}=ax$, for every constant c. Analogously, we would like to say that $\mathbf{x}(t)=e^{\mathbf{A}t}\mathbf{v}$ is a solution of the vector differential equation

$$\dot{\mathbf{x}}=\mathbf{A}\mathbf{x} \tag{1}$$

for every constant vector $\mathbf{v}$. However, $e^{\mathbf{A}t}$ is not defined if $\mathbf{A}$ is an $n\times n$ matrix. This is not a serious difficulty, though. There is a very natural way of defining $e^{\mathbf{A}t}$ so that it resembles the scalar exponential e^{at}; simply set

$$e^{\mathbf{A}t}\equiv\mathbf{I}+\mathbf{A}t+\frac{\mathbf{A}^2t^2}{2!}+\dots+\frac{\mathbf{A}^nt^n}{n!}+\dots. \tag{2}$$

It can be shown that the infinite series (2) converges for all t, and can be differentiated term by term. In particular

$$\begin{aligned}\frac{d}{dt}e^{\mathbf{A}t} &= \mathbf{A}+\mathbf{A}^2t+\ldots+\frac{\mathbf{A}^{n+1}}{n!}t^n+\ldots\\ &= \mathbf{A}\left[\mathbf{I}+\mathbf{A}t+\ldots+\frac{\mathbf{A}^nt^n}{n!}+\ldots\right]=\mathbf{A}e^{\mathbf{A}t}.\end{aligned}$$

This implies that $e^{\mathbf{A}t}\mathbf{v}$ is a solution of (1) for every constant vector $\mathbf{v}$, since

$$\frac{d}{dt}e^{\mathbf{A}t}\mathbf{v}=\mathbf{A}e^{\mathbf{A}t}\mathbf{v}=\mathbf{A}(e^{\mathbf{A}t}\mathbf{v}).$$

Remark. The matrix exponential $e^{\mathbf{A}t}$ and the scalar exponential e^{at} satisfy many similar properties. For example,

$$\left(e^{\mathbf{A}t}\right)^{-1}=e^{-\mathbf{A}t} \quad\text{and}\quad e^{\mathbf{A}(t+s)}=e^{\mathbf{A}t}e^{\mathbf{A}s}. \tag{3}$$

Indeed, the same proofs which show that $(e^{at})^{-1}=e^{-at}$ and $e^{a(t+s)}=e^{at}e^{as}$ can be used to establish the identities (3): we need only replace every a by $\mathbf{A}$ and every 1 by $\mathbf{I}$. However, $e^{\mathbf{A}t+\mathbf{B}t}$ equals $e^{\mathbf{A}t}e^{\mathbf{B}t}$ only if $\mathbf{AB}=\mathbf{BA}$ (see Exercise 15, Section 3.11).

There are several classes of matrices $\mathbf{A}$ (see Problems 9–11) for which the infinite series (2) can be summed exactly. In general, though, it does not seem possible to express $e^{\mathbf{A}t}$ in closed form. Yet, the remarkable fact is that we can always find n linearly independent vectors $\mathbf{v}$ for which the infinite series $e^{\mathbf{A}t}\mathbf{v}$ can be summed exactly. Moreover, once we know n linearly independent solutions of (1), we can even compute $e^{\mathbf{A}t}$ exactly. (This latter property will be proven in the next section.)

We now show how to find n linearly independent vectors $\mathbf{v}$ for which the infinite series $e^{\mathbf{A}t}\mathbf{v}$ can be summed exactly. Observe that

$$e^{\mathbf{A}t}\mathbf{v}=e^{(\mathbf{A}-\lambda\mathbf{I})t}e^{\lambda\mathbf{I}t}\mathbf{v}$$

for any constant λ, since $(\mathbf{A}-\lambda\mathbf{I})(\lambda\mathbf{I})=(\lambda\mathbf{I})(\mathbf{A}-\lambda\mathbf{I})$. Moreover,

$$\begin{aligned}e^{\lambda\mathbf{I}t}\mathbf{v} &= \left[\mathbf{I}+\lambda\mathbf{I}t+\frac{\lambda^2\mathbf{I}^2t^2}{2!}+\ldots\right]\mathbf{v}\\ &= \left[1+\lambda t+\frac{\lambda^2t^2}{2!}+\ldots\right]\mathbf{v}=e^{\lambda t}\mathbf{v}.\end{aligned}$$

Hence, $e^{\mathbf{A}t}\mathbf{v}=e^{\lambda t}e^{(\mathbf{A}-\lambda\mathbf{I})t}\mathbf{v}$.

Next, we make the crucial observation that if $\mathbf{v}$ satisfies $(\mathbf{A}-\lambda\mathbf{I})^m\mathbf{v}=\mathbf{0}$ for some integer m, then the infinite series $e^{(\mathbf{A}-\lambda\mathbf{I})t}\mathbf{v}$ terminates after m terms. If $(\mathbf{A}-\lambda\mathbf{I})^m\mathbf{v}=\mathbf{0}$, then $(\mathbf{A}-\lambda\mathbf{I})^{m+l}\mathbf{v}$ is also zero, for every positive integer l, since

$$(\mathbf{A}-\lambda\mathbf{I})^{m+l}\mathbf{v}=(\mathbf{A}-\lambda\mathbf{I})^l\left[(\mathbf{A}-\lambda\mathbf{I})^m\mathbf{v}\right]=\mathbf{0}.$$

Consequently,

$$e^{(\mathbf{A}-\lambda\mathbf{I})t}\mathbf{v}=\mathbf{v}+t(\mathbf{A}-\lambda\mathbf{I})\mathbf{v}+\ldots+\frac{t^{m-1}}{(m-1)!}(\mathbf{A}-\lambda\mathbf{I})^{m-1}\mathbf{v}$$

and

$$\begin{aligned}e^{\mathbf{A}t}\mathbf{v}&=e^{\lambda t}e^{(\mathbf{A}-\lambda\mathbf{I})t}\mathbf{v}\\&=e^{\lambda t}\left[\mathbf{v}+t(\mathbf{A}-\lambda\mathbf{I})\mathbf{v}+\ldots+\frac{t^{m-1}}{(m-1)!}(\mathbf{A}-\lambda\mathbf{I})^{m-1}\mathbf{v}\right].\end{aligned}$$

This suggests the following algorithm for finding n linearly independent solutions of (1).

(1) Find all the eigenvalues and eigenvectors of $\mathbf{A}$. If $\mathbf{A}$ has n linearly independent eigenvectors, then the differential equation $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$ has n linearly independent solutions of the form $e^{\lambda t}\mathbf{v}$. (Observe that the infinite series $e^{(\mathbf{A}-\lambda\mathbf{I})t}\mathbf{v}$ terminates after one term if $\mathbf{v}$ is an eigenvector of $\mathbf{A}$ with eigenvalue λ.)

(2) Suppose that $\mathbf{A}$ has only $k<n$ linearly independent eigenvectors. Then, we have only k linearly independent solutions of the form $e^{\lambda t}\mathbf{v}$. To find additional solutions we pick an eigenvalue λ of $\mathbf{A}$ and find all vectors $\mathbf{v}$ for which $(\mathbf{A}-\lambda\mathbf{I})^2\mathbf{v}=\mathbf{0}$, but $(\mathbf{A}-\lambda\mathbf{I})\mathbf{v}\neq\mathbf{0}$. For each such vector $\mathbf{v}$

$$e^{\mathbf{A}t}\mathbf{v}=e^{\lambda t}e^{(\mathbf{A}-\lambda\mathbf{I})t}\mathbf{v}=e^{\lambda t}\left[\mathbf{v}+t(\mathbf{A}-\lambda\mathbf{I})\mathbf{v}\right]$$

is an additional solution of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$. We do this for all the eigenvalues λ of $\mathbf{A}$.

(3) If we still do not have enough solutions, then we find all vectors $\mathbf{v}$ for which $(\mathbf{A}-\lambda\mathbf{I})^3\mathbf{v}=\mathbf{0}$, but $(\mathbf{A}-\lambda\mathbf{I})^2\mathbf{v}\neq\mathbf{0}$. For each such vector $\mathbf{v}$,

$$e^{\mathbf{A}t}\mathbf{v}=e^{\lambda t}\left[\mathbf{v}+t(\mathbf{A}-\lambda\mathbf{I})\mathbf{v}+\frac{t^2}{2!}(\mathbf{A}-\lambda\mathbf{I})^2\mathbf{v}\right]$$

is an additional solution of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$.

(4) We keep proceeding in this manner until, hopefully, we obtain n linearly independent solutions.

The following lemma from linear algebra, which we accept without proof, guarantees that this algorithm always works. Moreover, it puts an upper bound on the number of steps we have to perform in this algorithm.

Lemma 1. *Let the characteristic polynomial of* $\mathbf{A}$ *have* k *distinct roots* $\lambda_1,\ldots,\lambda_k$ *with multiplicities* $n_1,\ldots,n_k$ *respectively. (This means that* $p(\lambda)$ *can be factored into the form* $(\lambda_1-\lambda)^{n_1}\ldots(\lambda_k-\lambda)^{n_k}$.) *Suppose that* $\mathbf{A}$ *has only* $\nu_j<n_j$ *linearly independent eigenvectors with eigenvalue* λ_j. *Then the equation* $(\mathbf{A}-\lambda_j\mathbf{I})^2\mathbf{v}=\mathbf{0}$ *has at least* ν_j+1 *independent solutions. More generally, if the equation* $(\mathbf{A}-\lambda_j\mathbf{I})^m\mathbf{v}=\mathbf{0}$ *has only* $m_j<n_j$ *independent solutions, then the equation* $(\mathbf{A}-\lambda_j\mathbf{I})^{m+1}\mathbf{v}=\mathbf{0}$ *has at least* m_j+1 *independent solutions.*

Lemma 1 clearly implies that there exists an integer d_j with $d_j \leqslant n_j$, such that the equation $(\mathbf{A}-\lambda_j\mathbf{I})^{d_j}\mathbf{v}=\mathbf{0}$ has at least n_j linearly independent solutions. Thus, for each eigenvalue λ_j of $\mathbf{A}$, we can compute n_j linearly independent solutions of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$. All these solutions have the form

$$\mathbf{x}(t)=e^{\lambda_j t}\left[\mathbf{v}+t(\mathbf{A}-\lambda_j\mathbf{I})\mathbf{v}+\ldots+\frac{t^{d_j-1}}{(d_j-1)!}(\mathbf{A}-\lambda_j\mathbf{I})^{d_j-1}\mathbf{v}\right].$$

In addition, it can be shown that the set of $n_1+\ldots+n_k=n$ solutions thus obtained must be linearly independent.

Example 1. Find three linearly independent solutions of the differential equation

$$\dot{\mathbf{x}}=\begin{pmatrix}1&1&0\\0&1&0\\0&0&2\end{pmatrix}\mathbf{x}.$$

Solution. The characteristic polynomial of the matrix

$$\mathbf{A}=\begin{pmatrix}1&1&0\\0&1&0\\0&0&2\end{pmatrix}$$

is $(1-\lambda)^2(2-\lambda)$. Hence $\lambda=1$ is an eigenvalue of $\mathbf{A}$ with multiplicity two, and $\lambda=2$ is an eigenvalue of $\mathbf{A}$ with multiplicity one.

(i) $\lambda=1$: We seek all nonzero vectors $\mathbf{v}$ such that

$$(\mathbf{A}-\mathbf{I})\mathbf{v}=\begin{pmatrix}0&1&0\\0&0&0\\0&0&1\end{pmatrix}\begin{pmatrix}v_1\\v_2\\v_3\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix}.$$

This implies that $v_2=v_3=0$, and v_1 is arbitrary. Consequently,

$$\mathbf{x}^1(t)=e^t\begin{pmatrix}1\\0\\0\end{pmatrix}$$

is one solution of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$. Since $\mathbf{A}$ has only one linearly independent eigenvector with eigenvalue 1, we look for all solutions of the equation

$$(\mathbf{A}-\mathbf{I})^2\mathbf{v}=\begin{pmatrix}0&1&0\\0&0&0\\0&0&1\end{pmatrix}\begin{pmatrix}0&1&0\\0&0&0\\0&0&1\end{pmatrix}\mathbf{v}=\begin{pmatrix}0&0&0\\0&0&0\\0&0&1\end{pmatrix}\begin{pmatrix}v_1\\v_2\\v_3\end{pmatrix}=\begin{pmatrix}0\\0\\0\end{pmatrix}$$

This implies that $v_3=0$ and both v_1 and v_2 are arbitrary. Now, the vector

$$\mathbf{v}=\begin{pmatrix}0\\1\\0\end{pmatrix}$$

satisfies $(\mathbf{A}-\mathbf{I})^2\mathbf{v}=\mathbf{0}$, but $(\mathbf{A}-\mathbf{I})\mathbf{v}\neq\mathbf{0}$. (We could just as well choose any

$$\mathbf{v}=\begin{bmatrix} v_1\\ v_2\\ 0\end{bmatrix}$$

for which $v_2\neq 0$.) Hence,

$$\begin{aligned}
\mathbf{x}^2(t)&=e^{\mathbf{A}t}\begin{bmatrix}0\\1\\0\end{bmatrix}=e^{t}e^{(\mathbf{A}-\mathbf{I})t}\begin{bmatrix}0\\1\\0\end{bmatrix}\\
&=e^{t}\left[\mathbf{I}+t(\mathbf{A}-\mathbf{I})\right]\begin{bmatrix}0\\1\\0\end{bmatrix}=e^{t}\left[\begin{bmatrix}0\\1\\0\end{bmatrix}+t\begin{bmatrix}0&1&0\\0&0&0\\0&0&1\end{bmatrix}\begin{bmatrix}0\\1\\0\end{bmatrix}\right]\\
&=e^{t}\begin{bmatrix}0\\1\\0\end{bmatrix}+te^{t}\begin{bmatrix}1\\0\\0\end{bmatrix}=e^{t}\begin{bmatrix}t\\1\\0\end{bmatrix}
\end{aligned}$$

is a second linearly independent solution.

(ii) $\lambda=2$: We seek a nonzero vector $\mathbf{v}$ such that

$$(\mathbf{A}-2\mathbf{I})\mathbf{v}=\begin{bmatrix}-1&1&0\\0&-1&0\\0&0&0\end{bmatrix}\begin{bmatrix}v_1\\v_2\\v_3\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix}.$$

This implies that $v_1=v_2=0$ and v_3 is arbitrary. Hence

$$\mathbf{x}^3(t)=e^{2t}\begin{bmatrix}0\\0\\1\end{bmatrix}$$

is a third linearly independent solution.

Example 2. Solve the initial-value problem

$$\dot{\mathbf{x}}=\begin{bmatrix}2&1&3\\0&2&-1\\0&0&2\end{bmatrix}\mathbf{x},\qquad \mathbf{x}(0)=\begin{bmatrix}1\\2\\1\end{bmatrix}.$$

Solution. The characteristic polynomial of the matrix

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

is $(2-\lambda)^3$. Hence $\lambda=2$ is an eigenvalue of $\mathbf{A}$ with multiplicity three. The eigenvectors of $\mathbf{A}$ satisfy the equation

$$(\mathbf{A}-2\mathbf{I})\mathbf{v}=\begin{bmatrix}0&1&3\\0&0&-1\\0&0&0\end{bmatrix}\begin{bmatrix}v_1\\v_2\\v_3\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix}.$$

This implies that $v_2 = v_3 = 0$ and v_1 is arbitrary. Hence

$$\mathbf{x}^1(t) = e^{2t}\begin{bmatrix}1\\0\\0\end{bmatrix}$$

is one solution of $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$.

Since $\mathbf{A}$ has only one linearly independent eigenvector we look for all solutions of the equation

$$(\mathbf{A}-2\mathbf{I})^2\mathbf{v} = \begin{bmatrix}0&1&3\\0&0&-1\\0&0&0\end{bmatrix}\begin{bmatrix}0&1&3\\0&0&-1\\0&0&0\end{bmatrix}\mathbf{v} = \begin{bmatrix}0&0&-1\\0&0&0\\0&0&0\end{bmatrix}\begin{bmatrix}v_1\\v_2\\v_3\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix}.$$

This implies that $v_3 = 0$ and both v_1 and v_2 are arbitrary. Now, the vector

$$\mathbf{v} = \begin{bmatrix}0\\1\\0\end{bmatrix}$$

satisfies $(\mathbf{A}-2\mathbf{I})^2\mathbf{v} = \mathbf{0}$, but $(\mathbf{A}-2\mathbf{I})\mathbf{v} \neq \mathbf{0}$. Hence

$$\begin{aligned}\mathbf{x}^2(t) &= e^{\mathbf{A}t}\begin{bmatrix}0\\1\\0\end{bmatrix} = e^{2t}e^{(\mathbf{A}-2\mathbf{I})t}\begin{bmatrix}0\\1\\0\end{bmatrix}\\
&= e^{2t}\left[\mathbf{I}+t(\mathbf{A}-2\mathbf{I})\right]\begin{bmatrix}0\\1\\0\end{bmatrix} = e^{2t}\left[\mathbf{I}+t\begin{bmatrix}0&1&3\\0&0&-1\\0&0&0\end{bmatrix}\right]\begin{bmatrix}0\\1\\0\end{bmatrix}\\
&= e^{2t}\left[\begin{bmatrix}0\\1\\0\end{bmatrix}+t\begin{bmatrix}1\\0\\0\end{bmatrix}\right] = e^{2t}\begin{bmatrix}t\\1\\0\end{bmatrix}\end{aligned}$$

is a second solution of $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$.

Since the equation $(\mathbf{A}-2\mathbf{I})^2\mathbf{v} = \mathbf{0}$ has only two linearly independent solutions, we look for all solutions of the equation

$$(\mathbf{A}-2\mathbf{I})^3\mathbf{v} = \begin{bmatrix}0&1&3\\0&0&-1\\0&0&0\end{bmatrix}^3\mathbf{v} = \begin{bmatrix}0&0&0\\0&0&0\\0&0&0\end{bmatrix}\begin{bmatrix}v_1\\v_2\\v_3\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix}.$$

Obviously, every vector $\mathbf{v}$ is a solution of this equation. The vector

$$\mathbf{v} = \begin{bmatrix}0\\0\\1\end{bmatrix}$$

does not satisfy $(\mathbf{A}-2\mathbf{I})^2\mathbf{v}=\mathbf{0}$. Hence

$$\mathbf{x}^3(t)=e^{\mathbf{A}t}\begin{bmatrix}0\\0\\1\end{bmatrix}=e^{2t}e^{(\mathbf{A}-2\mathbf{I})t}\begin{bmatrix}0\\0\\1\end{bmatrix}$$

$$=e^{2t}\left[\mathbf{I}+t(\mathbf{A}-2\mathbf{I})+\frac{t^2}{2}(\mathbf{A}-2\mathbf{I})^2\right]\begin{bmatrix}0\\0\\1\end{bmatrix}$$

$$=e^{2t}\left[\begin{bmatrix}0\\0\\1\end{bmatrix}+t\begin{bmatrix}3\\-1\\0\end{bmatrix}+\frac{t^2}{2}\begin{bmatrix}-1\\0\\0\end{bmatrix}\right]=e^{2t}\begin{bmatrix}3t-\frac{1}{2}t^2\\-t\\1\end{bmatrix}$$

is a third linearly independent solution. Therefore,

$$\mathbf{x}(t)=e^{2t}\left[c_1\begin{bmatrix}1\\0\\0\end{bmatrix}+c_2\begin{bmatrix}t\\1\\0\end{bmatrix}+c_3\begin{bmatrix}3t-\frac{1}{2}t^2\\-t\\1\end{bmatrix}\right].$$

The constants c_1, c_2, and c_3 are determined from the initial conditions

$$\begin{bmatrix}1\\2\\1\end{bmatrix}=c_1\begin{bmatrix}1\\0\\0\end{bmatrix}+c_2\begin{bmatrix}0\\1\\0\end{bmatrix}+c_3\begin{bmatrix}0\\0\\1\end{bmatrix}.$$

This implies that $c_1=1$, $c_2=2$, and $c_3=1$. Hence

$$\mathbf{x}(t)=e^{2t}\begin{bmatrix}1+5t-\frac{1}{2}t^2\\2-t\\1\end{bmatrix}.$$

For the matrix $\mathbf{A}$ in Example 2, $p(\lambda)=(2-\lambda)^3$ and $(2\mathbf{I}-\mathbf{A})^3=\mathbf{0}$. This is not an accident. Every matrix $\mathbf{A}$ satisfies its own characteristic equation. This is the content of the following theorem.

Theorem 13 (Cayley–Hamilton). *Let $p(\lambda)=p_0+p_1\lambda+\ldots+(-1)^n\lambda^n$ be the characteristic polynomial of $\mathbf{A}$. Then,*

$$p(\mathbf{A})\equiv p_0\mathbf{I}+p_1\mathbf{A}+\ldots+(-1)^n\mathbf{A}^n=\mathbf{0}.$$

FAKE PROOF. Setting $\lambda=\mathbf{A}$ in the equation $p(\lambda)=\det(\mathbf{A}-\lambda\mathbf{I})$ gives $p(\mathbf{A})=\det(\mathbf{A}-\mathbf{A}\mathbf{I})=\det\mathbf{0}=0$. The fallacy in this proof is that we cannot set $\lambda=\mathbf{A}$ in the expression $\det(\mathbf{A}-\lambda\mathbf{I})$ since we cannot subtract a matrix from the diagonal elements of $\mathbf{A}$. However, there is a very clever way to make this proof kosher. Let $\mathbf{C}(\lambda)$ be the classical adjoint (see Section 3.6) of the matrix $(\mathbf{A}-\lambda\mathbf{I})$. Then,

$$(\mathbf{A}-\lambda\mathbf{I})\mathbf{C}(\lambda)=p(\lambda)\mathbf{I}. \tag{4}$$

Each element of the matrix $\mathbf{C}(\lambda)$ is a polynomial in λ of degree at most $(n-1)$. Therefore, we can write $\mathbf{C}(\lambda)$ in the form

$$\mathbf{C}(\lambda)=\mathbf{C}_0+\mathbf{C}_1\lambda+\dots+\mathbf{C}_{n-1}\lambda^{n-1}$$

where $\mathbf{C}_0,\dots,\mathbf{C}_{n-1}$ are $n\times n$ matrices. For example,

$$\begin{pmatrix}\lambda+\lambda^2 & 2\lambda\\ \lambda^2 & 1-\lambda\end{pmatrix}=\begin{pmatrix}0&0\\0&1\end{pmatrix}+\begin{pmatrix}\lambda & 2\lambda\\ 0 & -\lambda\end{pmatrix}+\begin{pmatrix}\lambda^2 & 0\\ \lambda^2 & 0\end{pmatrix}$$
$$=\begin{pmatrix}0&0\\0&1\end{pmatrix}+\lambda\begin{pmatrix}1&2\\0&-1\end{pmatrix}+\lambda^2\begin{pmatrix}1&0\\1&0\end{pmatrix}.$$

Thus, Equation (4) can be written in the form

$$(\mathbf{A}-\lambda\mathbf{I})\left[\mathbf{C}_0+\mathbf{C}_1\lambda+\dots+\mathbf{C}_{n-1}\lambda^{n-1}\right]=p_0\mathbf{I}+p_1\lambda\mathbf{I}+\dots+(-1)^n\lambda^n\mathbf{I}. \quad (5)$$

Observe that both sides of (5) are polynomials in λ, whose coefficients are $n\times n$ matrices. Since these two polynomials are equal for all values of λ, their coefficients must agree. But if the coefficients of like powers of λ agree, then we can put in anything we want for λ and still have equality. In particular, set $\lambda=\mathbf{A}$. Then,

$$p(\mathbf{A})=p_0\mathbf{I}+p_1\mathbf{A}+\dots+(-1)^n\mathbf{A}^n$$
$$=(\mathbf{A}-\mathbf{A}\mathbf{I})\left[\mathbf{C}_0+\mathbf{C}_1\mathbf{A}+\dots+\mathbf{C}_{n-1}\mathbf{A}^{n-1}\right]=\mathbf{0}. \qquad \square$$

Exercises

In each of Problems 1–4 find the general solution of the given system of differential equations.

1. $\dot{\mathbf{x}}=\begin{pmatrix}0&-1&1\\2&-3&1\\1&-1&-1\end{pmatrix}\mathbf{x}$ **2.** $\dot{\mathbf{x}}=\begin{pmatrix}1&1&1\\2&1&-1\\-3&2&4\end{pmatrix}\mathbf{x}$

3. $\dot{\mathbf{x}}=\begin{pmatrix}-1&-1&0\\0&-1&0\\0&0&-2\end{pmatrix}\mathbf{x}$ *Hint*: Look at Example 1 of text.

4. $\dot{\mathbf{x}}=\begin{bmatrix}2&0&-1&0\\0&2&1&0\\0&0&2&0\\0&0&-1&2\end{bmatrix}\mathbf{x}$

In each of Problems 5–8, solve the given initial-value problem

5. $\dot{\mathbf{x}}=\begin{pmatrix}-1&1&2\\-1&1&1\\-2&1&3\end{pmatrix}\mathbf{x},\quad \mathbf{x}(0)=\begin{pmatrix}1\\0\\1\end{pmatrix}$ **6.** $\dot{\mathbf{x}}=\begin{pmatrix}-4&-4&0\\10&9&1\\-4&-3&1\end{pmatrix}\mathbf{x},\quad \mathbf{x}(0)=\begin{pmatrix}2\\1\\-1\end{pmatrix}$

7. $\dot{\mathbf{x}}=\begin{pmatrix}1&2&-3\\1&1&2\\1&-1&4\end{pmatrix}\mathbf{x},\quad \mathbf{x}(0)=\begin{pmatrix}1\\0\\0\end{pmatrix}$ **8.** $\dot{\mathbf{x}}=\begin{bmatrix}3&0&0&0\\1&3&0&0\\0&0&3&0\\0&0&2&3\end{bmatrix}\mathbf{x},\quad \mathbf{x}(0)=\begin{bmatrix}1\\1\\1\\1\end{bmatrix}$

9. Let

$$\mathbf{A}=\begin{bmatrix}\lambda_1 & 0 & \dots & 0\\ 0 & \lambda_2 & \dots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \dots & \lambda_n\end{bmatrix}.$$

Show that

$$e^{\mathbf{A}t}=\begin{bmatrix}e^{\lambda_1 t} & 0 & 0\\ 0 & e^{\lambda_2 t} & 0\\ \vdots & \vdots & \vdots\\ 0 & 0 & e^{\lambda_n t}\end{bmatrix}.$$

10. Let

$$\mathbf{A}=\begin{pmatrix}\lambda & 1 & 0\\ 0 & \lambda & 1\\ 0 & 0 & \lambda\end{pmatrix}.$$

Prove that

$$e^{\mathbf{A}t}=e^{\lambda t}\begin{bmatrix}1 & t & \frac{1}{2}t^2\\ 0 & 1 & t\\ 0 & 0 & 1\end{bmatrix}.$$

Hint: Write **A** in the form

$$\mathbf{A}=\lambda\mathbf{I}+\begin{pmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ 0 & 0 & 0\end{pmatrix}$$

and observe that

$$e^{\mathbf{A}t}=e^{\lambda t}\exp\left[\begin{pmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ 0 & 0 & 0\end{pmatrix}t\right].$$

11. Let **A** be the $n\times n$ matrix

$$\begin{bmatrix}\lambda & 1 & 0 & \dots & 0\\ 0 & \lambda & 1 & \dots & 0\\ \vdots & \vdots & \vdots & & \vdots\\ 0 & 0 & 0 & \dots & 1\\ 0 & 0 & 0 & \dots & \lambda\end{bmatrix},$$

and let **P** be the $n\times n$ matrix

$$\begin{bmatrix}0 & 1 & 0 & \dots & 0\\ 0 & 0 & 1 & \dots & 0\\ \vdots & \vdots & \vdots & & \vdots\\ 0 & 0 & 0 & \dots & 1\\ 0 & 0 & 0 & \dots & 0\end{bmatrix}.$$

(a) Show that $\mathbf{P}^n = \mathbf{0}$. (b) Show that $(\lambda\mathbf{I})\mathbf{P} = \mathbf{P}(\lambda\mathbf{I})$.
(c) Show that

$$e^{\mathbf{A}t} = e^{\lambda t}\left[\mathbf{I} + t\mathbf{P} + \frac{t^2\mathbf{P}^2}{2!} + \dots + \frac{t^{n-1}}{(n-1)!}\mathbf{P}^{n-1}\right].$$

12. Compute $e^{\mathbf{A}t}$ if

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

(b) $\mathbf{A} = \begin{bmatrix} 2 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$, (c) $\mathbf{A} = \begin{bmatrix} 2 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$.

13. (a) Show that $e^{\mathbf{T}^{-1}\mathbf{A}\mathbf{T}} = \mathbf{T}^{-1}e^{\mathbf{A}}\mathbf{T}$.
(b) Given that

$$\mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

with

$$\mathbf{T} = \begin{pmatrix} 1 & -1 & -1 \\ 3 & -1 & 2 \\ 2 & 2 & 3 \end{pmatrix},$$

compute $e^{\mathbf{A}t}$.

14. Suppose that $p(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I})$ has n distinct roots $\lambda_1, \dots, \lambda_n$. Prove directly that $p(\mathbf{A}) \equiv (-1)^n(\mathbf{A} - \lambda_1\mathbf{I})\dots(\mathbf{A} - \lambda_n\mathbf{I}) = \mathbf{0}$. *Hint*: Write any vector $\mathbf{x}$ in the form $\mathbf{x} = x_1\mathbf{v}^1 + \dots + x_n\mathbf{v}^n$ where $\mathbf{v}^1, \dots, \mathbf{v}^n$ are n independent eigenvectors of $\mathbf{A}$ with eigenvalues $\lambda_1, \dots, \lambda_n$ respectively, and conclude that $p(\mathbf{A})\mathbf{x} = \mathbf{0}$ for all vectors $\mathbf{x}$.

15. Suppose that $\mathbf{A}^2 = \alpha\mathbf{A}$. Find $e^{\mathbf{A}t}$.

16. Let

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

(a) Show that $\mathbf{A}(\mathbf{A} - 5\mathbf{I}) = \mathbf{0}$.
(b) Find $e^{\mathbf{A}t}$.

17. Let

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

(a) Show that $\mathbf{A}^2 = -\mathbf{I}$.

(b) Show that

$$e^{\mathbf{A}t}=\begin{pmatrix} \cos t & \sin t\\ -\sin t & \cos t\end{pmatrix}.$$

In each of Problems 18–20 verify directly the Cayley–Hamilton Theorem for the given matrix **A**.

18. $\mathbf{A}=\begin{pmatrix} 1 & -1 & -1\\ 2 & 1 & -1\\ -1 & -1 & 1\end{pmatrix}$ **19.** $\mathbf{A}=\begin{pmatrix} 2 & 3 & 1\\ 2 & 3 & 1\\ 2 & 3 & 1\end{pmatrix}$ **20.** $\mathbf{A}=\begin{pmatrix} -1 & 0 & 1\\ -1 & 3 & 0\\ 2 & 4 & 6\end{pmatrix}$

3.11 Fundamental matrix solutions; $e^{\mathbf{A}t}$

If $\mathbf{x}^1(t),\ldots,\mathbf{x}^n(t)$ are n linearly independent solutions of the differential equation

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

then every solution $\mathbf{x}(t)$ can be written in the form

$$\mathbf{x}(t)=c_1\mathbf{x}^1(t)+c_2\mathbf{x}^2(t)+\ldots+c_n\mathbf{x}^n(t). \tag{2}$$

Let $\mathbf{X}(t)$ be the matrix whose columns are $\mathbf{x}^1(t),\ldots,\mathbf{x}^n(t)$. Then, Equation (2) can be written in the concise form $\mathbf{x}(t)=\mathbf{X}(t)\mathbf{c}$, where

$$\mathbf{c}=\begin{bmatrix} c_1\\ \vdots\\ c_n\end{bmatrix}.$$

Definition. A matrix $\mathbf{X}(t)$ is called a *fundamental matrix solution* of (1) if its columns form a set of n linearly independent solutions of (1).

Example 1. Find a fundamental matrix solution of the system of differential equations

$$\dot{\mathbf{x}}=\begin{bmatrix} 1 & -1 & 4\\ 3 & 2 & -1\\ 2 & 1 & -1\end{bmatrix}\mathbf{x}. \tag{3}$$

Solution. We showed in Section 3.8 (see Example 1) that

$$e^t\begin{bmatrix} -1\\ 4\\ 1\end{bmatrix},\qquad e^{3t}\begin{bmatrix} 1\\ 2\\ 1\end{bmatrix}\quad\text{and}\quad e^{-2t}\begin{bmatrix} -1\\ 1\\ 1\end{bmatrix}$$

are three linearly independent solutions of (3). Hence

$$\mathbf{X}(t)=\begin{bmatrix} -e^t & e^{3t} & -e^{-2t}\\ 4e^t & 2e^{3t} & e^{-2t}\\ e^t & e^{3t} & e^{-2t}\end{bmatrix}$$

is a fundamental matrix solution of (3).

In this section we will show that the matrix $e^{\mathbf{A}t}$ can be computed directly from any fundamental matrix solution of (1). This is rather remarkable since it does not appear possible to sum the infinite series $[\mathbf{I}+\mathbf{A}t+(\mathbf{A}t)^2/2!+\ldots]$ exactly, for an arbitrary matrix $\mathbf{A}$. Specifically, we have the following theorem.

Theorem 14. *Let $\mathbf{X}(t)$ be a fundamental matrix solution of the differential equation $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$. Then,*

$$e^{\mathbf{A}t}=\mathbf{X}(t)\mathbf{X}^{-1}(0). \tag{4}$$

In other words, the product of any fundamental matrix solution of (1) *with its inverse at $t=0$ must yield $e^{\mathbf{A}t}$.*

We prove Theorem 14 in three steps. First, we establish a simple test to determine whether a matrix-valued function is a fundamental matrix solution of (1). Then, we use this test to show that $e^{\mathbf{A}t}$ is a fundamental matrix solution of (1). Finally, we establish a connection between any two fundamental matrix solutions of (1).

Lemma 1. *A matrix $\mathbf{X}(t)$ is a fundamental matrix solution of* (1) *if, and only if, $\dot{\mathbf{X}}(t)=\mathbf{A}\mathbf{X}(t)$ and* $\det\mathbf{X}(0)\neq 0$. (*The derivative of a matrix-valued function $\mathbf{X}(t)$ is the matrix whose components are the derivatives of the corresponding components of* $\mathbf{X}(t)$.)

PROOF. Let $\mathbf{x}^1(t),\ldots,\mathbf{x}^n(t)$ denote the n columns of $\mathbf{X}(t)$. Observe that

$$\dot{\mathbf{X}}(t)=\left(\dot{\mathbf{x}}^1(t),\ldots,\dot{\mathbf{x}}^n(t)\right)$$

and

$$\mathbf{A}\mathbf{X}(t)=\left(\mathbf{A}\mathbf{x}^1(t),\ldots,\mathbf{A}\mathbf{x}^n(t)\right).$$

Hence, the n vector equations $\dot{\mathbf{x}}^1(t)=\mathbf{A}\mathbf{x}^1(t),\ldots,\dot{\mathbf{x}}^n(t)=\mathbf{A}\mathbf{x}^n(t)$ are equivalent to the single matrix equation $\dot{\mathbf{X}}(t)=\mathbf{A}\mathbf{X}(t)$. Moreover, n solutions $\mathbf{x}^1(t),\ldots,\mathbf{x}^n(t)$ of (1) are linearly independent if, and only if, $\mathbf{x}^1(0),\ldots,\mathbf{x}^n(0)$ are linearly independent vectors of R^n. These vectors, in turn, are linearly independent if, and only if, $\det\mathbf{X}(0)\neq 0$. Consequently, $\mathbf{X}(t)$ is a fundamental matrix solution of (1) if, and only if, $\dot{\mathbf{X}}(t)=\mathbf{A}\mathbf{X}(t)$ and $\det\mathbf{X}(0)\neq 0$. □

Lemma 2. *The matrix-valued function $e^{\mathbf{A}t}\equiv\mathbf{I}+\mathbf{A}t+\mathbf{A}^2t^2/2!+\ldots$ is a fundamental matrix solution of* (1).

PROOF. We showed in Section 3.10 that $(d/dt)e^{\mathbf{A}t}=\mathbf{A}e^{\mathbf{A}t}$. Hence $e^{\mathbf{A}t}$ is a solution of the matrix differential equation $\dot{\mathbf{X}}(t)=\mathbf{A}\mathbf{X}(t)$. Moreover, its determinant, evaluated at $t=0$, is one since $e^{\mathbf{A}0}=\mathbf{I}$. Therefore, by Lemma 1, $e^{\mathbf{A}t}$ is a fundamental matrix solution of (1). □

Lemma 3. *Let $\mathbf{X}(t)$ and $\mathbf{Y}(t)$ be two fundamental matrix solutions of* (1). *Then, there exists a constant matrix $\mathbf{C}$ such that $\mathbf{Y}(t)=\mathbf{X}(t)\mathbf{C}$.*

PROOF. By definition, the columns $\mathbf{x}^1(t),\dots,\mathbf{x}^n(t)$ of $\mathbf{X}(t)$ and $\mathbf{y}^1(t),\dots,$ $\mathbf{y}^n(t)$ of $\mathbf{Y}(t)$ are linearly independent sets of solutions of (1). In particular, therefore, each column of $\mathbf{Y}(t)$ can be written as a linear combination of the columns of $\mathbf{X}(t)$; i.e., there exist constants $c_1^j,\dots,c_n^j$ such that

$$\mathbf{y}^j(t)=c_1^j\mathbf{x}^1(t)+c_2^j\mathbf{x}^2(t)+\dots+c_n^j\mathbf{x}^n(t),\qquad j=1,\dots,n. \tag{5}$$

Let $\mathbf{C}$ be the matrix $(\mathbf{c}^1,\mathbf{c}^2,\dots,\mathbf{c}^n)$ where

$$\mathbf{c}^j=\begin{bmatrix} c_1^j\\ \vdots\\ c_n^j\end{bmatrix}.$$

Then, the n equations (5) are equivalent to the single matrix equation $\mathbf{Y}(t)$ $=\mathbf{X}(t)\mathbf{C}$. □

We are now in a position to prove Theorem 14.

PROOF OF THEOREM 14. Let $\mathbf{X}(t)$ be a fundamental matrix solution of (1). Then, by Lemmas 2 and 3 there exists a constant matrix $\mathbf{C}$ such that

$$e^{\mathbf{A}t}=\mathbf{X}(t)\mathbf{C}. \tag{6}$$

Setting $t=0$ in (6) gives $\mathbf{I}=\mathbf{X}(0)\mathbf{C}$, which implies that $\mathbf{C}=\mathbf{X}^{-1}(0)$. Hence, $e^{\mathbf{A}t}=\mathbf{X}(t)\mathbf{X}^{-1}(0)$. □

Example 1. Find $e^{\mathbf{A}t}$ if

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

Solution. Our first step is to find 3 linearly independent solutions of the differential equation

$$\dot{\mathbf{x}}=\begin{bmatrix}1&1&1\\0&3&2\\0&0&5\end{bmatrix}\mathbf{x}.$$

To this end we compute

$$p(\lambda)=\det(\mathbf{A}-\lambda\mathbf{I})=\det\begin{bmatrix}1-\lambda&1&1\\0&3-\lambda&2\\0&0&5-\lambda\end{bmatrix}=(1-\lambda)(3-\lambda)(5-\lambda).$$

Thus, $\mathbf{A}$ has 3 distinct eigenvalues $\lambda=1$, $\lambda=3$, and $\lambda=5$.

(i) $\lambda=1$: Clearly,

$$\mathbf{v}^1=\begin{bmatrix}1\\0\\0\end{bmatrix}$$

is an eigenvector of $\mathbf{A}$ with eigenvalue one. Hence

$$\mathbf{x}^1(t)=e^t\begin{bmatrix}1\\0\\0\end{bmatrix}$$

is one solution of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$.

(ii) $\lambda=3$: We seek a nonzero solution of the equation

$$(\mathbf{A}-3\mathbf{I})\mathbf{v}=\begin{bmatrix}-2&1&1\\0&0&2\\0&0&2\end{bmatrix}\begin{bmatrix}v_1\\v_2\\v_3\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix}.$$

This implies that $v_3=0$ and $v_2=2v_1$. Hence,

$$\mathbf{v}^2=\begin{bmatrix}1\\2\\0\end{bmatrix}$$

is an eigenvector of $\mathbf{A}$ with eigenvalue 3. Consequently,

$$\mathbf{x}^2(t)=e^{3t}\begin{bmatrix}1\\2\\0\end{bmatrix}$$

is a second solution of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$.

(iii) $\lambda=5$: We seek a nonzero solution of the equation

$$(\mathbf{A}-5\mathbf{I})\mathbf{v}=\begin{bmatrix}-4&1&1\\0&-2&2\\0&0&0\end{bmatrix}\begin{bmatrix}v_1\\v_2\\v_3\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix}.$$

This implies that $v_2=v_3$ and $v_1=v_3/2$. Hence,

$$\mathbf{v}^3=\begin{bmatrix}1\\2\\2\end{bmatrix}$$

is an eigenvector of $\mathbf{A}$ with eigenvalue 5. Consequently,

$$\mathbf{x}^3(t)=e^{5t}\begin{bmatrix}1\\2\\2\end{bmatrix}$$

is a third solution of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$. These solutions are clearly linearly independent. Therefore,

$$\mathbf{X}(t)=\begin{bmatrix}e^t&e^{3t}&e^{5t}\\0&2e^{3t}&2e^{5t}\\0&0&2e^{5t}\end{bmatrix}$$

is a fundamental matrix solution. Using the methods of Section 3.6, we compute

$$\mathbf{X}^{-1}(0)=\begin{bmatrix}1&1&1\\0&2&2\\0&0&2\end{bmatrix}^{-1}=\begin{bmatrix}1&-\frac{1}{2}&0\\0&\frac{1}{2}&-\frac{1}{2}\\0&0&\frac{1}{2}\end{bmatrix}.$$

Therefore,

$$\exp\left[\begin{pmatrix}1&1&1\\0&3&2\\0&0&5\end{pmatrix}t\right]=\begin{bmatrix}e^{t}&e^{3t}&e^{5t}\\0&2e^{3t}&2e^{5t}\\0&0&2e^{5t}\end{bmatrix}\begin{bmatrix}1&-\frac{1}{2}&0\\0&\frac{1}{2}&-\frac{1}{2}\\0&0&\frac{1}{2}\end{bmatrix}$$

$$=\begin{bmatrix}e^{t}&-\frac{1}{2}e^{t}+\frac{1}{2}e^{3t}&-\frac{1}{2}e^{3t}+\frac{1}{2}e^{5t}\\0&e^{3t}&-e^{3t}+e^{5t}\\0&0&e^{5t}\end{bmatrix}.$$

Exercises

Compute $e^{\mathbf{A}t}$ for $\mathbf{A}$ equal

1. $\begin{pmatrix}1&-1&-1\\1&3&1\\-3&1&-1\end{pmatrix}$ **2.** $\begin{pmatrix}1&1&-1\\2&3&-4\\4&1&-4\end{pmatrix}$ **3.** $\begin{pmatrix}2&0&1\\1&0&1\\1&-2&0\end{pmatrix}$

4. $\begin{pmatrix}1&0&0\\2&1&-2\\3&2&1\end{pmatrix}$ **5.** $\begin{pmatrix}0&2&1\\-1&-3&-1\\1&1&-1\end{pmatrix}$ **6.** $\begin{pmatrix}1&1&1\\2&1&-1\\0&-1&1\end{pmatrix}$

7. Find $\mathbf{A}$ if

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

In each of Problems 8–11, determine whether the given matrix is a fundamental matrix solution of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$, for some $\mathbf{A}$; if yes, find $\mathbf{A}$.

8. $\begin{bmatrix}e^{t}&e^{-t}&e^{t}+2e^{-t}\\e^{t}&-e^{-t}&e^{t}-2e^{-t}\\2e^{t}&e^{-t}&2(e^{t}+e^{-t})\end{bmatrix}$

9. $\begin{bmatrix}-5\cos 2t&-5\sin 2t&3e^{2t}\\-2(\cos 2t+\sin 2t)&2(\cos 2t-\sin 2t)&0\\\cos 2t&\sin 2t&e^{2t}\end{bmatrix}$

10. $e^t\begin{bmatrix} 1 & t+1 & t^2+1 \\ 1 & 2(t+1) & 4t^2 \\ 1 & t+2 & 3 \end{bmatrix}$ **11.** $\begin{bmatrix} e^{2t} & 2e^{-t} & e^{3t} \\ 2e^t & 2e^{-t} & e^{3t} \\ 3e^t & e^{-t} & 2e^{3t} \end{bmatrix}$

12. Let $\boldsymbol{\phi}^j(t)$ be the solution of the initial-value problem $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$, $\mathbf{x}(0)=\mathbf{e}^j$. Show that $e^{\mathbf{A}t}=(\boldsymbol{\phi}^1,\boldsymbol{\phi}^2,\dots,\boldsymbol{\phi}^n)$.

13. Suppose that $\mathbf{Y}(t)=\mathbf{X}(t)\mathbf{C}$, where $\mathbf{X}(t)$ and $\mathbf{Y}(t)$ are fundamental matrix solutions of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$, and $\mathbf{C}$ is a constant matrix. Prove that $\det\mathbf{C}\neq 0$.

14. Let $\mathbf{X}(t)$ be a fundamental matrix solution of (1), and $\mathbf{C}$ a constant matrix with $\det\mathbf{C}\neq 0$. Show that $\mathbf{Y}(t)=\mathbf{X}(t)\mathbf{C}$ is again a fundamental matrix solution of (1).

15. Let $\mathbf{X}(t)$ be a fundamental matrix solution of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$. Prove that the solution $\mathbf{x}(t)$ of the initial-value problem $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$, $\mathbf{x}(t_0)=\mathbf{x}^0$ is $\mathbf{x}(t)=\mathbf{X}(t)\mathbf{X}^{-1}(t_0)\mathbf{x}^0$.

16. Let $\mathbf{X}(t)$ be a fundamental matrix solution of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$. Prove that $\mathbf{X}(t)\mathbf{X}^{-1}(t_0)=e^{\mathbf{A}(t-t_0)}$.

17. Here is an elegant proof of the identity $e^{\mathbf{A}t+\mathbf{B}t}=e^{\mathbf{A}t}e^{\mathbf{B}t}$ if $\mathbf{AB}=\mathbf{BA}$.
(a) Show that $\mathbf{X}(t)=e^{\mathbf{A}t+\mathbf{B}t}$ satisfies the initial-value problem $\dot{\mathbf{X}}=(\mathbf{A}+\mathbf{B})\mathbf{X}$, $\mathbf{X}(0)=\mathbf{I}$.
(b) Show that $e^{\mathbf{A}t}\mathbf{B}=\mathbf{B}e^{\mathbf{A}t}$ if $\mathbf{AB}=\mathbf{BA}$. (*Hint*: $\mathbf{A}^j\mathbf{B}=\mathbf{B}\mathbf{A}^j$ if $\mathbf{AB}=\mathbf{BA}$). Then, conclude that $(d/dt)e^{\mathbf{A}t}e^{\mathbf{B}t}=(\mathbf{A}+\mathbf{B})e^{\mathbf{A}t}e^{\mathbf{B}t}$.
(c) It follows immediately from Theorem 4, Section 3.4 that the solution $\mathbf{X}(t)$ of the initial-value problem $\dot{\mathbf{X}}=(\mathbf{A}+\mathbf{B})\mathbf{X}$, $\mathbf{X}(0)=\mathbf{I}$, is unique. Conclude, therefore, that $e^{\mathbf{A}t+\mathbf{B}t}=e^{\mathbf{A}t}e^{\mathbf{B}t}$.

3.12 The nonhomogeneous equation; variation of parameters

Consider now the nonhomogeneous equation $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}+\mathbf{f}(t)$. In this case, we can use our knowledge of the solutions of the homogeneous equation

$$\dot{\mathbf{x}}=\mathbf{A}\mathbf{x} \tag{1}$$

to help us find the solution of the initial-value problem

$$\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}+\mathbf{f}(t), \qquad \mathbf{x}(t_0)=\mathbf{x}^0. \tag{2}$$

Let $\mathbf{x}^1(t),\dots,\mathbf{x}^n(t)$ be n linearly independent solutions of the homogeneous equation (1). Since the general solution of (1) is $c_1\mathbf{x}^1(t)+\dots+c_n\mathbf{x}^n(t)$, it is natural to seek a solution of (2) of the form

$$\mathbf{x}(t)=u_1(t)\mathbf{x}^1(t)+u_2(t)\mathbf{x}^2(t)+\dots+u_n(t)\mathbf{x}^n(t). \tag{3}$$

This equation can be written concisely in the form $\mathbf{x}(t)=\mathbf{X}(t)\mathbf{u}(t)$ where

$\mathbf{X}(t)=(\mathbf{x}^1(t),\dots,\mathbf{x}^n(t))$ and

$$\mathbf{u}(t)=\begin{bmatrix} u_1(t) \\ \vdots \\ u_n(t) \end{bmatrix}.$$

Plugging this expression into the differential equation $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}+\mathbf{f}(t)$ gives

$$\dot{\mathbf{X}}(t)\mathbf{u}(t)+\mathbf{X}(t)\dot{\mathbf{u}}(t)=\mathbf{A}\mathbf{X}(t)\mathbf{u}(t)+\mathbf{f}(t). \tag{4}$$

The matrix $\mathbf{X}(t)$ is a fundamental matrix solution of (1). Hence, $\dot{\mathbf{X}}(t)=\mathbf{A}\mathbf{X}(t)$, and Equation (4) reduces to

$$\mathbf{X}(t)\dot{\mathbf{u}}(t)=\mathbf{f}(t). \tag{5}$$

Recall that the columns of $\mathbf{X}(t)$ are linearly independent vectors of R^n at every time t. Hence $\mathbf{X}^{-1}(t)$ exists, and

$$\dot{\mathbf{u}}(t)=\mathbf{X}^{-1}(t)\mathbf{f}(t). \tag{6}$$

Integrating this expression between t_0 and t gives

$$\begin{aligned}\mathbf{u}(t)&=\mathbf{u}(t_0)+\int_{t_0}^{t}\mathbf{X}^{-1}(s)\mathbf{f}(s)\,ds\\ &=\mathbf{X}^{-1}(t_0)\mathbf{x}^0+\int_{t_0}^{t}\mathbf{X}^{-1}(s)\mathbf{f}(s)\,ds.\end{aligned}$$

Consequently,

$$\mathbf{x}(t)=\mathbf{X}(t)\mathbf{X}^{-1}(t_0)\mathbf{x}^0+\mathbf{X}(t)\int_{t_0}^{t}\mathbf{X}^{-1}(s)\mathbf{f}(s)\,ds. \tag{7}$$

If $\mathbf{X}(t)$ is the fundamental matrix solution $e^{\mathbf{A}t}$, then Equation (7) simplifies considerably. To wit, if $\mathbf{X}(t)=e^{\mathbf{A}t}$, then $\mathbf{X}^{-1}(s)=e^{-\mathbf{A}s}$. Hence

$$\begin{aligned}\mathbf{x}(t)&=e^{\mathbf{A}t}e^{-\mathbf{A}t_0}\mathbf{x}^0+e^{\mathbf{A}t}\int_{t_0}^{t}e^{-\mathbf{A}s}\mathbf{f}(s)\,ds\\ &=e^{\mathbf{A}(t-t_0)}\mathbf{x}^0+\int_{t_0}^{t}e^{\mathbf{A}(t-s)}\mathbf{f}(s)\,ds.\end{aligned} \tag{8}$$

Example 1. Solve the initial-value problem

$$\dot{\mathbf{x}}=\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & -2 \\ 3 & 2 & 1 \end{bmatrix}\mathbf{x}+\begin{bmatrix} 0 \\ 0 \\ e^t\cos 2t \end{bmatrix}, \qquad \mathbf{x}(0)=\begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}.$$

Solution. We first find $e^{\mathbf{A}t}$, where

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

To this end compute

$$\det(\mathbf{A}-\lambda\mathbf{I})=\det\begin{bmatrix}1-\lambda&0&0\\2&1-\lambda&-2\\3&2&1-\lambda\end{bmatrix}=(1-\lambda)(\lambda^2-2\lambda+5).$$

Thus the eigenvalues of $\mathbf{A}$ are

$$\lambda=1 \quad\text{and}\quad \lambda=\frac{2\pm\sqrt{4-20}}{2}=1\pm 2i.$$

(i) $\lambda=1$: We seek a nonzero vector $\mathbf{v}$ such that

$$(\mathbf{A}-\mathbf{I})\mathbf{v}=\begin{bmatrix}0&0&0\\2&0&-2\\3&2&0\end{bmatrix}\begin{bmatrix}v_1\\v_2\\v_3\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix}.$$

This implies that $v_1=v_3$ and $v_2=-3v_1/2$. Hence

$$\mathbf{v}=\begin{bmatrix}2\\-3\\2\end{bmatrix}$$

is an eigenvector of $\mathbf{A}$ with eigenvalue 1. Consequently,

$$\mathbf{x}^1(t)=e^t\begin{bmatrix}2\\-3\\2\end{bmatrix}$$

is a solution of the homogeneous equation $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$.

(ii) $\lambda=1+2i$: We seek a nonzero vector $\mathbf{v}$ such that

$$\left[\mathbf{A}-(1+2i)\mathbf{I}\right]\mathbf{v}=\begin{bmatrix}-2i&0&0\\2&-2i&-2\\3&2&-2i\end{bmatrix}\begin{bmatrix}v_1\\v_2\\v_3\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix}.$$

This implies that $v_1=0$ and $v_3=-iv_2$. Hence,

$$\mathbf{v}=\begin{bmatrix}0\\1\\-i\end{bmatrix}$$

is an eigenvector of $\mathbf{A}$ with eigenvalue $1+2i$. Therefore,

$$\mathbf{x}(t)=\begin{bmatrix}0\\1\\-i\end{bmatrix}e^{(1+2i)t}$$

is a complex-valued solution of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$. Now,

$$\begin{bmatrix}0\\1\\-i\end{bmatrix}e^{(1+2i)t}=e^{t}(\cos 2t+i\sin 2t)\left[\begin{bmatrix}0\\1\\0\end{bmatrix}-i\begin{bmatrix}0\\0\\1\end{bmatrix}\right]$$

$$=e^{t}\left[\cos 2t\begin{bmatrix}0\\1\\0\end{bmatrix}+\sin 2t\begin{bmatrix}0\\0\\1\end{bmatrix}\right]$$

$$+ie^{t}\left[\sin 2t\begin{bmatrix}0\\1\\0\end{bmatrix}-\cos 2t\begin{bmatrix}0\\0\\1\end{bmatrix}\right].$$

Consequently,

$$\mathbf{x}^2(t)=e^{t}\begin{bmatrix}0\\\cos 2t\\\sin 2t\end{bmatrix}\quad\text{and}\quad\mathbf{x}^3(t)=e^{t}\begin{bmatrix}0\\\sin 2t\\-\cos 2t\end{bmatrix}$$

are real-valued solutions of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$. The solutions $\mathbf{x}^1$, $\mathbf{x}^2$, and $\mathbf{x}^3$ are linearly independent since their values at $t=0$ are clearly linearly independent vectors of R^3. Therefore,

$$\mathbf{X}(t)=\begin{bmatrix}2e^{t}&0&0\\-3e^{t}&e^{t}\cos 2t&e^{t}\sin 2t\\2e^{t}&e^{t}\sin 2t&-e^{t}\cos 2t\end{bmatrix}$$

is a fundamental matrix solution of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$. Computing

$$\mathbf{X}^{-1}(0)=\begin{bmatrix}2&0&0\\-3&1&0\\2&0&-1\end{bmatrix}^{-1}=\begin{bmatrix}\frac{1}{2}&0&0\\\frac{3}{2}&1&0\\1&0&-1\end{bmatrix}$$

we see that

$$e^{\mathbf{A}t}=\begin{bmatrix}2e^{t}&0&0\\-3e^{t}&e^{t}\cos 2t&e^{t}\sin 2t\\2e^{t}&e^{t}\sin 2t&-e^{t}\cos 2t\end{bmatrix}\begin{bmatrix}\frac{1}{2}&0&0\\\frac{3}{2}&1&0\\1&0&-1\end{bmatrix}$$

$$=e^{t}\begin{bmatrix}1&0&0\\-\frac{3}{2}+\frac{3}{2}\cos 2t+\sin 2t&\cos 2t&-\sin 2t\\1+\frac{3}{2}\sin 2t-\cos 2t&\sin 2t&\cos 2t\end{bmatrix}.$$

Consequently,

$$\mathbf{x}(t)=e^{\mathbf{A}t}\begin{bmatrix}0\\1\\1\end{bmatrix}$$

$$+e^{\mathbf{A}t}\int_0^t e^{-s}\begin{bmatrix}1 & 0 & 0\\ -\frac{3}{2}+\frac{3}{2}\cos 2s-\sin 2s & \cos 2s & \sin 2s\\ 1-\frac{3}{2}\sin 2s-\cos 2s & -\sin 2s & \cos 2s\end{bmatrix}\begin{bmatrix}0\\0\\e^s\cos 2s\end{bmatrix}ds$$

$$=e^t\begin{bmatrix}0\\ \cos 2t-\sin 2t\\ \cos 2t+\sin 2t\end{bmatrix}+e^{\mathbf{A}t}\int_0^t\begin{bmatrix}0\\ \sin 2s\cos 2s\\ \cos^2 2s\end{bmatrix}ds$$

$$=e^t\begin{bmatrix}0\\ \cos 2t-\sin 2t\\ \cos 2t+\sin 2t\end{bmatrix}+e^{\mathbf{A}t}\begin{bmatrix}0\\ (1-\cos 4t)/8\\ t/2+(\sin 4t)/8\end{bmatrix}$$

$$=e^t\begin{bmatrix}0\\ \cos 2t-\sin 2t\\ \cos 2t+\sin 2t\end{bmatrix}$$

$$+e^t\begin{bmatrix}0\\ -\dfrac{t\sin 2t}{2}+\dfrac{\cos 2t-\cos 4t\cos 2t-\sin 4t\sin 2t}{8}\\ \dfrac{t\cos 2t}{2}+\dfrac{\sin 4t\cos 2t-\sin 2t\cos 4t+\sin 2t}{8}\end{bmatrix}$$

$$=e^t\begin{bmatrix}0\\ \cos 2t-\left(1+\frac{1}{2}t\right)\sin 2t\\ \left(1+\frac{1}{2}t\right)\cos 2t+\frac{5}{4}\sin 2t\end{bmatrix}.$$

As Example 1 indicates, the method of variation of parameters is often quite tedious and laborious. One way of avoiding many of these calculations is to "guess" a particular solution $\psi(t)$ of the nonhomogeneous equation and then to observe (see Exercise 9) that every solution $\mathbf{x}(t)$ of the nonhomogeneous equation must be of the form $\phi(t)+\psi(t)$ where $\phi(t)$ is a solution of the homogeneous equation.

Example 2. Find all solutions of the differential equation

$$\dot{\mathbf{x}}=\begin{bmatrix}1 & 0 & 0\\ 2 & 1 & -2\\ 3 & 2 & 1\end{bmatrix}\mathbf{x}+\begin{bmatrix}1\\0\\0\end{bmatrix}e^{ct},\qquad c\neq 1. \tag{9}$$

Solution. Let

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

We "guess" a particular solution $\psi(t)$ of the form $\psi(t)=\mathbf{b}e^{ct}$. Plugging this expression into (9) gives

$$c\mathbf{b}e^{ct}=\mathbf{A}\mathbf{b}e^{ct}+\begin{bmatrix}1\\0\\0\end{bmatrix}e^{ct},$$

or

$$(\mathbf{A}-c\mathbf{I})\mathbf{b}=\begin{bmatrix}-1\\0\\0\end{bmatrix}.$$

This implies that

$$\mathbf{b}=\frac{-1}{1-c}\begin{bmatrix}1\\ \dfrac{2(c-4)}{4+(1-c)^2}\\ \dfrac{1+3c}{4+(1-c)^2}\end{bmatrix}$$

Hence, every solution $\mathbf{x}(t)$ of (9) is of the form

$$\mathbf{x}(t)=e^{t}\left[c_1\begin{bmatrix}2\\-3\\2\end{bmatrix}+c_2\begin{bmatrix}0\\ \cos 2t\\ \sin 2t\end{bmatrix}+c_3\begin{bmatrix}0\\ -\sin 2t\\ \cos 2t\end{bmatrix}\right]$$

$$-\frac{e^{ct}}{1-c}\begin{bmatrix}1\\ \dfrac{2(c-4)}{4+(1-c)^2}\\ \dfrac{1+3c}{4+(1-c)^2}\end{bmatrix}$$

Remark. We run into trouble when $c=1$ because one is an eigenvalue of the matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & -2 \\ 3 & 2 & 1 \end{bmatrix}.$$

More generally, the differential equation $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}+\mathbf{v}e^{ct}$ may not have a solution of the form $\mathbf{b}e^{ct}$ if c is an eigenvalue of $\mathbf{A}$. In this case we have to guess a particular solution of the form

$$\boldsymbol{\psi}(t)=e^{ct}\left[\mathbf{b}_0+\mathbf{b}_1t+\ldots+\mathbf{b}_{k-1}t^{k-1}\right]$$

for some appropriate integer k. (See Exercises 10–18).

Exercises

In each of Problems 1–6 use the method of variation of parameters to solve the given initial-value problem.

1. $\dot{\mathbf{x}}=\begin{pmatrix} 4 & 5 \\ -2 & -2 \end{pmatrix}\mathbf{x}+\begin{pmatrix} 4e^t\cos t \\ 0 \end{pmatrix}, \quad \mathbf{x}(0)=\begin{pmatrix} 0 \\ 0 \end{pmatrix}$

2. $\dot{\mathbf{x}}=\begin{pmatrix} 3 & -4 \\ 1 & -1 \end{pmatrix}\mathbf{x}+\begin{pmatrix} 1 \\ 1 \end{pmatrix}e^t, \quad \mathbf{x}(0)=\begin{pmatrix} 1 \\ 1 \end{pmatrix}$

3. $\dot{\mathbf{x}}=\begin{pmatrix} 2 & -5 \\ 1 & -2 \end{pmatrix}\mathbf{x}+\begin{pmatrix} \sin t \\ \tan t \end{pmatrix}, \quad \mathbf{x}(0)=\begin{pmatrix} 0 \\ 0 \end{pmatrix}$

4. $\dot{\mathbf{x}}=\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\mathbf{x}+\begin{pmatrix} f_1(t) \\ f_2(t) \end{pmatrix}, \quad \mathbf{x}(0)=\begin{pmatrix} 0 \\ 0 \end{pmatrix}$

5. $\dot{\mathbf{x}}=\begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 1 & 3 \end{pmatrix}\mathbf{x}+\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}e^{2t}, \quad \mathbf{x}(0)=\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$

6. $\dot{\mathbf{x}}=\begin{pmatrix} -1 & -1 & -2 \\ 1 & 1 & 1 \\ 2 & 1 & 3 \end{pmatrix}\mathbf{x}+\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}e^t, \quad \mathbf{x}(0)=\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$

7. Consider the nth-order scalar differential equation

$$L[y]=\frac{d^ny}{dt^n}+a_1\frac{d^{n-1}y}{dt^{n-1}}+\ldots+a_ny=f(t). \tag{*}$$

Let $v(t)$ be the solution of $L[y]=0$ which satisfies the initial conditions $y(0)=\ldots=y^{(n-2)}(0)=0, y^{(n-1)}(0)=1$. Show that

$$y(t)=\int_0^t v(t-s)f(s)\,ds$$

is the solution of (*) which satisfies the initial conditions $y(0)=\ldots=y^{(n-1)}(0)=0$. *Hint*: Convert (*) to a system of n first-order equations of the form $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$, and show that

$$\begin{bmatrix} v(t) \\ v'(t) \\ \vdots \\ v^{(n-1)}(t) \end{bmatrix}$$

is the nth column of $e^{\mathbf{A}t}$.

8. Find the solution of the initial-value problem

$$\frac{d^3y}{dt^3}+\frac{dy}{dt}=\sec t\tan t, \qquad y(0)=y'(0)=y''(0)=0.$$

9. (a) Let $\psi(t)$ be a solution of the nonhomogeneous equation $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}+\mathbf{f}(t)$ and let $\phi(t)$ be a solution of the homogeneous equation $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$. Show that $\phi(t)+\psi(t)$ is a solution of the nonhomogeneous equation.

(b) Let $\psi_1(t)$ and $\psi_2(t)$ be two solutions of the nonhomogeneous equation. Show that $\psi_1(t)-\psi_2(t)$ is a solution of the homogeneous equation.

(c) Let $\psi(t)$ be a particular solution of the nonhomogeneous equation. Show that any other solution $\mathbf{y}(t)$ must be of the form $\mathbf{y}(t)=\phi(t)+\psi(t)$ where $\phi(t)$ is a solution of the homogeneous equation.

In each of Problems 10–14 use the method of judicious guessing to find a particular solution of the given differential equation.

10. $\dot{\mathbf{x}}=\begin{pmatrix} 2 & 1 \\ 3 & -2 \end{pmatrix}\mathbf{x}+\begin{pmatrix} 1 \\ 1 \end{pmatrix}e^{3t}$

11. $\dot{\mathbf{x}}=\begin{pmatrix} 1 & -1 \\ 1 & 3 \end{pmatrix}\mathbf{x}+\begin{pmatrix} -t^2 \\ 2t \end{pmatrix}$

12. $\dot{\mathbf{x}}=\begin{pmatrix} 1 & 3 & 2 \\ -1 & 2 & 1 \\ 4 & -1 & -1 \end{pmatrix}\mathbf{x}+\begin{pmatrix} \sin t \\ 0 \\ 0 \end{pmatrix}$

13. $\dot{\mathbf{x}}=\begin{pmatrix} 1 & 2 & -3 \\ 1 & 1 & 2 \\ 1 & -1 & 4 \end{pmatrix}\mathbf{x}+\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}e^{t}$

14. $\dot{\mathbf{x}}=\begin{pmatrix} -1 & -1 & 0 \\ 0 & -4 & -1 \\ 0 & 5 & 0 \end{pmatrix}\mathbf{x}+\begin{pmatrix} 1 \\ t \\ e^{t} \end{pmatrix}$

15. Consider the differential equation

$$\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}+\mathbf{v}e^{\lambda t} \tag{*}$$

where $\mathbf{v}$ is an eigenvector of $\mathbf{A}$ with eigenvalue λ. Suppose moreover, that $\mathbf{A}$ has n linearly independent eigenvectors $\mathbf{v}^1,\mathbf{v}^2,\ldots,\mathbf{v}^n$, with distinct eigenvalues $\lambda_1,\ldots,\lambda_n$ respectively.

(a) Show that (*) has no solution $\psi(t)$ of the form $\psi(t)=\mathbf{a}e^{\lambda t}$. *Hint*: Write $\mathbf{a}=a_1\mathbf{v}^1+\ldots+a_n\mathbf{v}^n$.

(b) Show that (*) has a solution $\psi(t)$ of the form

$$\psi(t)=\mathbf{a}e^{\lambda t}+\mathbf{b}te^{\lambda t}.$$

Hint: Show that $\mathbf{b}$ is an eigenvector of $\mathbf{A}$ with eigenvalue λ and choose it so that we can solve the equation

$$(\mathbf{A}-\lambda\mathbf{I})\mathbf{a}=\mathbf{b}-\mathbf{v}.$$

In each of Problems 16–18, find a particular solution $\psi(t)$ of the given differential equation of the form $\psi(t)=e^{\lambda t}(\mathbf{a}+\mathbf{b}t)$.

16. $\dot{\mathbf{x}}=\begin{pmatrix}1 & 1 & -1\\ 2 & 3 & -4\\ 4 & 1 & -4\end{pmatrix}\mathbf{x}+\begin{pmatrix}1\\ 2\\ 1\end{pmatrix}e^{2t}$

17. $\dot{\mathbf{x}}=\begin{pmatrix}1 & -1 & -1\\ 1 & 3 & 1\\ -3 & 1 & -1\end{pmatrix}\mathbf{x}+\begin{pmatrix}1\\ -1\\ -1\end{pmatrix}e^{3t}$

18. $\dot{\mathbf{x}}=\begin{pmatrix}3 & 2 & 4\\ 2 & 0 & 2\\ 4 & 2 & 3\end{pmatrix}\mathbf{x}+\begin{pmatrix}2\\ 1\\ 2\end{pmatrix}e^{8t}$

3.13 Solving systems by Laplace transforms

The method of Laplace transforms introduced in Chapter 2 can also be used to solve the initial-value problem

$$\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}+\mathbf{f}(t),\qquad \mathbf{x}(0)=\mathbf{x}^0. \tag{1}$$

Let

$$\mathbf{X}(s)=\begin{bmatrix}X_1(s)\\ \vdots\\ X_n(s)\end{bmatrix}=\mathcal{L}\{\mathbf{x}(t)\}=\begin{bmatrix}\int_0^\infty e^{-st}x_1(t)\,dt\\ \vdots\\ \int_0^\infty e^{-st}x_n(t)\,dt\end{bmatrix}$$

and

$$\mathbf{F}(s)=\begin{bmatrix}F_1(s)\\ \vdots\\ F_n(s)\end{bmatrix}=\mathcal{L}\{\mathbf{f}(t)\}=\begin{bmatrix}\int_0^\infty e^{-st}f_1(t)\,dt\\ \vdots\\ \int_0^\infty e^{-st}f_n(t)\,dt\end{bmatrix}.$$

Taking Laplace transforms of both sides of (1) gives

$$\begin{aligned}\mathcal{L}\{\dot{\mathbf{x}}(t)\}&=\mathcal{L}\{\mathbf{A}\mathbf{x}(t)+\mathbf{f}(t)\}=\mathbf{A}\mathcal{L}\{\mathbf{x}(t)\}+\mathcal{L}\{\mathbf{f}(t)\}\\ &=\mathbf{A}\mathbf{X}(s)+\mathbf{F}(s),\end{aligned}$$

and from Lemma 3 of Section 2.9,

$$\mathcal{L}\{\dot{\mathbf{x}}(t)\}=\begin{bmatrix}\mathcal{L}\{\dot{x}_1(t)\}\\ \vdots\\ \mathcal{L}\{\dot{x}_n(t)\}\end{bmatrix}=\begin{bmatrix}sX_1(s)-x_1(0)\\ \vdots\\ sX_n(s)-x_n(0)\end{bmatrix}$$

$$=s\mathbf{X}(s)-\mathbf{x}^0.$$

Hence,

$$s\mathbf{X}(s)-\mathbf{x}^0=\mathbf{A}\mathbf{X}(s)+\mathbf{F}(s)$$

or

$$(s\mathbf{I}-\mathbf{A})\mathbf{X}(s)=\mathbf{x}^0+\mathbf{F}(s),\qquad \mathbf{I}=\begin{bmatrix}1&0&\dots&0\\0&1&\dots&0\\ \vdots&\vdots&&\vdots\\0&0&\dots&1\end{bmatrix}. \tag{2}$$

Equation (2) is a system of n simultaneous equations for $X_1(s),\dots,X_n(s)$, and it can be solved in a variety of ways. (One way, in particular, is to multiply both sides of (2) by $(s\mathbf{I}-\mathbf{A})^{-1}$.) Once we know $X_1(s),\dots,X_n(s)$ we can find $x_1(t),\dots,x_n(t)$ by inverting these Laplace transforms.

Example 1. Solve the initial-value problem

$$\dot{\mathbf{x}}=\begin{pmatrix}1&4\\1&1\end{pmatrix}\mathbf{x}+\begin{pmatrix}1\\1\end{pmatrix}e^t,\qquad \mathbf{x}(0)=\begin{pmatrix}2\\1\end{pmatrix}. \tag{3}$$

Solution. Taking Laplace transforms of both sides of the differential equation gives

$$s\mathbf{X}(s)-\begin{pmatrix}2\\1\end{pmatrix}=\begin{pmatrix}1&4\\1&1\end{pmatrix}\mathbf{X}(s)+\frac{1}{s-1}\begin{pmatrix}1\\1\end{pmatrix}$$

or

$$\begin{aligned}(s-1)X_1(s)-4X_2(s)&=2+\frac{1}{s-1}-\\ X_1(s)+(s-1)X_2(s)&=1+\frac{1}{s-1}.\end{aligned}$$

The solution of these equations is

$$X_1(s)=\frac{2}{s-3}+\frac{1}{s^2-1},\qquad X_2(s)=\frac{1}{s-3}+\frac{s}{(s-1)(s+1)(s-3)}.$$

Now,

$$\frac{2}{s-3}=\mathcal{L}\{2e^{3t}\},\quad\text{and}\quad \frac{1}{s^2-1}=\mathcal{L}\{\sinh t\}=\mathcal{L}\left\{\frac{e^t-e^{-t}}{2}\right\}.$$

Hence,

$$x_1(t)=2e^{3t}+\frac{e^{t}-e^{-t}}{2}.$$

To invert $X_2(s)$, we use partial fractions. Let

$$\frac{s}{(s-1)(s+1)(s-3)}=\frac{A}{s-1}+\frac{B}{s+1}+\frac{C}{s-3}.$$

This implies that

$$A(s+1)(s-3)+B(s-1)(s-3)+C(s-1)(s+1)=s. \tag{4}$$

Setting $s=1$, -1, and 3 respectively in (4) gives $A=-\frac{1}{4}$, $B=-\frac{1}{8}$, and $C=\frac{3}{8}$. Consequently,

$$x_2(t)=\mathcal{L}^{-1}\left\{-\frac{1}{4}\frac{1}{s-1}-\frac{1}{8}\frac{1}{s+1}+\frac{11}{8}\frac{1}{s-3}\right\}$$
$$=-\frac{1}{8}e^{-t}-\frac{1}{4}e^{t}+\frac{11}{8}e^{3t}.$$

Exercises

Find the solution of each of the following initial-value problems.

1. $\dot{\mathbf{x}}=\begin{pmatrix}1 & -3\\ -2 & 2\end{pmatrix}\mathbf{x}, \quad \mathbf{x}(0)=\begin{pmatrix}0\\5\end{pmatrix}$

2. $\dot{\mathbf{x}}=\begin{pmatrix}1 & -1\\ 5 & -3\end{pmatrix}\mathbf{x}, \quad \mathbf{x}(0)=\begin{pmatrix}1\\2\end{pmatrix}$

3. $\dot{\mathbf{x}}=\begin{pmatrix}3 & -2\\ 2 & -2\end{pmatrix}\mathbf{x}+\begin{pmatrix}t\\ 3e^{t}\end{pmatrix}, \quad \mathbf{x}(0)=\begin{pmatrix}2\\1\end{pmatrix}$

4. $\dot{\mathbf{x}}=\begin{pmatrix}1 & 1\\ 4 & 1\end{pmatrix}\mathbf{x}+\begin{pmatrix}2\\ -1\end{pmatrix}e^{t}, \quad \mathbf{x}(0)=\begin{pmatrix}0\\0\end{pmatrix}$

5. $\dot{\mathbf{x}}=\begin{pmatrix}3 & -4\\ 1 & -1\end{pmatrix}\mathbf{x}+\begin{pmatrix}1\\ 1\end{pmatrix}e^{t}, \quad \mathbf{x}(0)=\begin{pmatrix}1\\1\end{pmatrix}$

6. $\dot{\mathbf{x}}=\begin{pmatrix}2 & -5\\ 1 & -2\end{pmatrix}\mathbf{x}+\begin{pmatrix}\sin t\\ \tan t\end{pmatrix}, \quad \mathbf{x}(0)=\begin{pmatrix}-1\\0\end{pmatrix}$

7. $\dot{\mathbf{x}}=\begin{pmatrix}4 & 5\\ -2 & -2\end{pmatrix}\mathbf{x}+\begin{pmatrix}4e^{t}\cos t\\ 0\end{pmatrix}, \quad \mathbf{x}(0)=\begin{pmatrix}1\\1\end{pmatrix}$

8. $\dot{\mathbf{x}}=\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}\mathbf{x}+\begin{pmatrix}f_1(t)\\ f_2(t)\end{pmatrix}, \quad \mathbf{x}(0)=\begin{pmatrix}0\\0\end{pmatrix}$

9. $\dot{\mathbf{x}}=\begin{pmatrix}2 & -2\\ 4 & -2\end{pmatrix}\mathbf{x}+\begin{pmatrix}0\\ \delta(t-\pi)\end{pmatrix}, \quad \mathbf{x}(0)=\begin{pmatrix}1\\0\end{pmatrix}$

10. $\dot{\mathbf{x}}=\begin{pmatrix}3 & -2\\ 2 & -2\end{pmatrix}\mathbf{x}+\begin{pmatrix}1-H_\pi(t)\\ 0\end{pmatrix}, \quad \mathbf{x}(0)=\begin{pmatrix}0\\0\end{pmatrix}$

11. $\dot{\mathbf{x}}=\begin{pmatrix} 1 & 2 & -3 \\ 1 & 1 & 2 \\ 1 & -1 & 4 \end{pmatrix}\mathbf{x}, \quad \mathbf{x}(0)=\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$

12. $\dot{\mathbf{x}}=\begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}\mathbf{x}+\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}e^{2t}, \quad \mathbf{x}(0)=\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$

13. $\dot{\mathbf{x}}=\begin{pmatrix} -1 & -1 & 2 \\ 1 & 1 & 1 \\ 2 & 1 & 3 \end{pmatrix}\mathbf{x}+\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}e^{t}, \quad \mathbf{x}(0)=\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$

14. $\dot{\mathbf{x}}=\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & -2 \\ 3 & 2 & 1 \end{pmatrix}\mathbf{x}+\begin{pmatrix} 0 \\ 0 \\ e^{t}\cos 2t \end{pmatrix}, \quad \mathbf{x}(0)=\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$

15. $\dot{\mathbf{x}}=\begin{bmatrix} 3 & 0 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 2 & 3 \end{bmatrix}\mathbf{x}, \quad \mathbf{x}(0)=\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

4 Qualitative theory of differential equations

4.1 Introduction

In this chapter we consider the differential equation

$$\dot{\mathbf{x}} = \mathbf{f}(t,\mathbf{x}) \tag{1}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix},$$

and

$$\mathbf{f}(t,\mathbf{x}) = \begin{bmatrix} f_1(t,x_1,\dots,x_n) \\ \vdots \\ f_n(t,x_1,\dots,x_n) \end{bmatrix}$$

is a nonlinear function of $x_1,\dots,x_n$. Unfortunately, there are no known methods of solving Equation (1). This, of course, is very disappointing. However, it is not necessary, in most applications, to find the solutions of (1) explicitly. For example, let $x_1(t)$ and $x_2(t)$ denote the populations, at time t, of two species competing amongst themselves for the limited food and living space in their microcosm. Suppose, moreover, that the rates of growth of $x_1(t)$ and $x_2(t)$ are governed by the differential equation (1). In this case, we are not really interested in the values of $x_1(t)$ and $x_2(t)$ at every time t. Rather, we are interested in the qualitative properties of $x_1(t)$ and $x_2(t)$. Specically, we wish to answer the following questions.

1. Do there exist values ξ_1 and ξ_2 at which the two species coexist together in a steady state? That is to say, are there numbers ξ_1, ξ_2 such that $x_1(t)\equiv\xi_1$, $x_2(t)\equiv\xi_2$ is a solution of (1)? Such values ξ_1,ξ_2, if they exist, are called *equilibrium points* of (1).

2. Suppose that the two species are coexisting in equilibrium. Suddenly, we add a few members of species 1 to the microcosm. Will $x_1(t)$ and $x_2(t)$ remain close to their equilibrium values for all future time? Or perhaps the extra few members give species 1 a large advantage and it will proceed to annihilate species 2.

3. Suppose that x_1 and x_2 have arbitrary values at $t=0$. What happens as t approaches infinity? Will one species ultimately emerge victorious, or will the struggle for existence end in a draw?

More generally, we are interested in determining the following properties of solutions of (1).

1. Do there exist equilibrium values

$$\mathbf{x}^0=\begin{bmatrix} x_1^0 \\ \vdots \\ x_n^0 \end{bmatrix}$$

for which $\mathbf{x}(t)\equiv\mathbf{x}^0$ is a solution of (1)?

2. Let $\boldsymbol{\phi}(t)$ be a solution of (1). Suppose that $\boldsymbol{\psi}(t)$ is a second solution with $\boldsymbol{\psi}(0)$ very close to $\boldsymbol{\phi}(0)$; that is, $\psi_j(0)$ is very close to $\phi_j(0), j=1,\ldots,n$. Will $\boldsymbol{\psi}(t)$ remain close to $\boldsymbol{\phi}(t)$ for all future time, or will $\boldsymbol{\psi}(t)$ diverge from $\boldsymbol{\phi}(t)$ as t approaches infinity? This question is often referred to as the problem of *stability*. It is the most fundamental problem in the qualitative theory of differential equations, and has occupied the attention of many mathematicians for the past hundred years.

3. What happens to solutions $\mathbf{x}(t)$ of (1) as t approaches infinity? Do all solutions approach equilibrium values? If they don't approach equilibrium values, do they at least approach a periodic solution?

This chapter is devoted to answering these three questions. Remarkably, we can often give satisfactory answers to these questions, even though we cannot solve Equation (1) explicitly. Indeed, the first question can be answered immediately. Observe that $\dot{\mathbf{x}}(t)$ is identically zero if $\mathbf{x}(t)\equiv\mathbf{x}^0$. Hence, $\mathbf{x}^0$ is an equilibrium value of (1), if, and only if,

$$\mathbf{f}(t,\mathbf{x}^0)\equiv\mathbf{0}. \tag{2}$$

Example 1. Find all equilibrium values of the system of differential equations

$$\frac{dx_1}{dt}=1-x_2, \qquad \frac{dx_2}{dt}=x_1^3+x_2.$$

Solution.

$$\mathbf{x}^0=\begin{pmatrix} x_1^0 \\ x_2^0 \end{pmatrix}$$

is an equilibrium value if, and only if, $1-x_2^0=0$ and $(x_1^0)^3+x_2^0=0$. This implies that $x_2^0=1$ and $x_1^0=-1$. Hence $\begin{pmatrix} -1 \\ 1 \end{pmatrix}$ is the only equilibrium value of this system.

Example 2. Find all equilibrium solutions of the system

$$\frac{dx}{dt}=(x-1)(y-1), \qquad \frac{dy}{dt}=(x+1)(y+1).$$

Solution.

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$$

is an equilibrium value of this system if, and only if, $(x_0-1)(y_0-1)=0$ and $(x_0+1)(y_0+1)=0$. The first equation is satisfied if either x_0 or y_0 is 1, while the second equation is satisfied if either x_0 or y_0 is -1. Hence, $x=1$, $y=-1$ and $x=-1, y=1$ are the equilibrium solutions of this system.

The question of stability is of paramount importance in all physical applications, since we can never measure initial conditions exactly. For example, consider the case of a particle of mass one kgm attached to an elastic spring of force constant 1 N/m which is moving in a frictionless medium. In addition, an external force $F(t)=\cos 2t$ N is acting on the particle. Let $y(t)$ denote the position of the particle relative to its equilibrium position. Then $(d^2y/dt^2)+y=\cos 2t$. We convert this second-order equation into a system of two first-order equations by setting $x_1=y$, $x_2=y'$. Then,

$$\frac{dx_1}{dt}=x_2, \qquad \frac{dx_2}{dt}=-x_1+\cos 2t. \tag{3}$$

The functions $y_1(t)=\sin t$ and $y_2(t)=\cos t$ are two independent solutions of the homogeneous equation $y''+y=0$. Moreover, $y=-\frac{1}{3}\cos 2t$ is a particular solution of the nonhomogeneous equation. Therefore, every solution

$$\mathbf{x}(t)=\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}$$

of (3) is of the form

$$\mathbf{x}(t)=c_1\begin{pmatrix} \sin t \\ \cos t \end{pmatrix}+c_2\begin{pmatrix} \cos t \\ -\sin t \end{pmatrix}+\begin{bmatrix} -\frac{1}{3}\cos 2t \\ \frac{2}{3}\sin 2t \end{bmatrix}. \tag{4}$$

At time $t=0$ we measure the position and velocity of the particle and obtain $y(0)=1, y'(0)=0$. This implies that $c_1=0$ and $c_2=\frac{4}{3}$. Consequently, the position and velocity of the particle for all future time are given by the equation

$$\begin{pmatrix} y(t) \\ y'(t) \end{pmatrix} = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = \begin{bmatrix} \frac{4}{3}\cos t - \frac{1}{3}\cos 2t \\ -\frac{4}{3}\sin t + \frac{2}{3}\sin 2t \end{bmatrix}. \tag{5}$$

However, suppose that our measurements permit an error of magnitude 10^{-4}. Will the position and velocity of the particle remain close to the values predicted by (5)? The answer to this question had better be yes, for otherwise, Newtonian mechanics would be of no practical value to us. Fortunately, it is quite easy to show, in this case, that the position and velocity of the particle remain very close to the values predicted by (5). Let $\hat{y}(t)$ and $\hat{y}'(t)$ denote the true values of $y(t)$ and $y'(t)$ respectively. Clearly,

$$y(t)-\hat{y}(t) = \left(\tfrac{4}{3}-c_2\right)\cos t - c_1 \sin t$$

$$y'(t)-\hat{y}'(t) = -c_1\cos t - \left(\tfrac{4}{3}-c_2\right)\sin t$$

where c_1 and c_2 are two constants satisfying

$$-10^{-4} \leqslant c_1 \leqslant 10^{-4}, \qquad \tfrac{4}{3}-10^{-4} \leqslant c_2 \leqslant \tfrac{4}{3}+10^{-4}.$$

We can rewrite these equations in the form

$$y(t)-\hat{y}(t) = \left[c_1^2 + \left(\tfrac{4}{3}-c_2\right)^2\right]^{1/2}\cos(t-\delta_1), \qquad \tan\delta_1 = \frac{c_1}{c_2-\frac{4}{3}}$$

$$y'(t)-\hat{y}'(t) = \left[c_1^2 + \left(\tfrac{4}{3}-c_2\right)^2\right]^{1/2}\cos(t-\delta_2), \qquad \tan\delta_2 = \frac{\frac{4}{3}-c_2}{c_1}.$$

Hence, both $y(t)-\hat{y}(t)$ and $y'(t)-\hat{y}'(t)$ are bounded in absolute value by $[c_1^2+(\frac{4}{3}-c_2)^2]^{1/2}$. This quantity is at most $\sqrt{2}\ 10^{-4}$. Therefore, the true values of $y(t)$ and $y'(t)$ are indeed close to the values predicted by (5).

As a second example of the concept of stability, consider the case of a particle of mass m which is supported by a wire, or inelastic string, of length l and of negligible mass. The wire is always straight, and the system is free to vibrate in a vertical plane. This configuration is usually referred to as a simple pendulum. The equation of motion of the pendulum is

$$\frac{d^2y}{dt^2} + \frac{g}{l}\sin y = 0$$

where y is the angle which the wire makes with the vertical line A0 (see

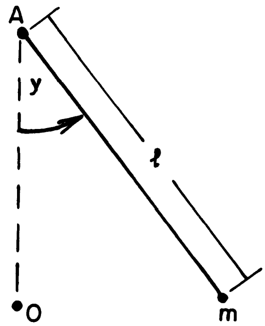


Figure 1

Figure 1). Setting $x_1 = y$ and $x_2 = dy/dt$ we see that

$$\frac{dx_1}{dt} = x_2, \qquad \frac{dx_2}{dt} = -\frac{g}{l}\sin x_1. \tag{6}$$

The system of equations (6) has equilibrium solutions $x_1 = 0$, $x_2 = 0$, and $x_1 = \pi$, $x_2 = 0$. (If the pendulum is suspended in the upright position $y = \pi$ with zero velocity, then it will remain in this upright position for all future time.) These two equilibrium solutions have very different properties. If we disturb the pendulum slightly from the equilibrium position $x_1 = 0$, $x_2 = 0$, by either displacing it slightly, or giving it a small velocity, then it will execute small oscillations about $x_1 = 0$. On the other hand, if we disturb the pendulum slightly from the equilibrium position $x_1 = \pi$, $x_2 = 0$, then it will either execute very large oscillations about $x_1 = 0$, or it will rotate around and around ad infinitum. Thus, the slightest disturbance causes the pendulum to deviate drastically from its equilibrium position $x_1 = \pi$, $x_2 = 0$. Intuitively, we would say that the equilibrium value $x_1 = 0$, $x_2 = 0$ of (6) is stable, while the equilibrium value $x_1 = \pi$, $x_2 = 0$ of (6) is unstable. This concept will be made precise in Section 4.2.

The question of stability is usually very difficult to resolve, because we cannot solve (1) explicitly. The only case which is manageable is when $\mathbf{f}(t,\mathbf{x})$ does not depend explicitly on t; that is, $\mathbf{f}$ is a function of $\mathbf{x}$ alone. Such differential equations are called *autonomous*. And even for autonomous differential equations, there are only two instances, generally, where we can completely resolve the stability question. The first case is when $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ and it will be treated in the next section. The second case is when we are only interested in the stability of an equilibrium solution of $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$. This case will be treated in Section 4.3.

Question 3 is extremely important in many applications since an answer to this question is a prediction concerning the long time evolution of the system under consideration. We answer this question, when possible, in Sections 4.6–4.8 and apply our results to some extremely important applications in Sections 4.9–4.12.

Exercises

In each of Problems 1–8, find all equilibrium values of the given system of differential equations.

1. $\dfrac{dx}{dt} = x - x^2 - 2xy$

$\dfrac{dy}{dt} = 2y - 2y^2 - 3xy$

2. $\dfrac{dx}{dt} = -\beta xy + \mu$

$\dfrac{dy}{dt} = \beta xy - \gamma y$

3. $\dfrac{dx}{dt} = ax - bxy$

$\dfrac{dy}{dt} = -cy + dxy$

$\dfrac{dz}{dt} = z + x^2 + y^2$

4. $\dfrac{dx}{dt} = -x - xy^2$

$\dfrac{dy}{dt} = -y - yx^2$

$\dfrac{dz}{dt} = 1 - z + x^2$

5. $\dfrac{dx}{dt} = xy^2 - x$

$\dfrac{dy}{dt} = x \sin \pi y$

6. $\dfrac{dx}{dt} = \cos y$

$\dfrac{dy}{dt} = \sin x - 1$

7. $\dfrac{dx}{dt} = -1 - y - e^x$

$\dfrac{dy}{dt} = x^2 + y(e^x - 1)$

$\dfrac{dz}{dt} = x + \sin z$

8. $\dfrac{dx}{dt} = x - y^2$

$\dfrac{dy}{dt} = x^2 - y$

$\dfrac{dz}{dt} = e^z - x$

9. Consider the system of differential equations

$$\frac{dx}{dt} = ax + by, \qquad \frac{dy}{dt} = cx + dy. \tag{*}$$

(i) Show that $x=0, y=0$ is the only equilibrium point of (*) if $ad - bc \neq 0$.
(ii) Show that (*) has a line of equilibrium points if $ad - bc = 0$.

10. Let $x = x(t), y = y(t)$ be the solution of the initial-value problem

$$\frac{dx}{dt} = -x - y, \qquad \frac{dy}{dt} = 2x - y, \qquad x(0) = y(0) = 1.$$

Suppose that we make an error of magnitude 10^{-4} in measuring $x(0)$ and $y(0)$. What is the largest error we make in evaluating $x(t), y(t)$ for $0 \leqslant t < \infty$?

11. (a) Verify that

$$\mathbf{x} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} e^{-t}$$

is the solution of the initial-value problem

$$\dot{\mathbf{x}} = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & -1 \\ -3 & 2 & 4 \end{pmatrix} \mathbf{x} - \begin{pmatrix} 2 \\ 2 \\ -3 \end{pmatrix} e^{-t}, \qquad \mathbf{x}(0) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

(b) Let $\mathbf{x}=\boldsymbol{\psi}(t)$ be the solution of the above differential equation which satisfies the initial condition

$$\mathbf{x}(0)=\mathbf{x}^0\neq\begin{pmatrix}1\\0\\0\end{pmatrix}.$$

Show that each component of $\boldsymbol{\psi}(t)$ approaches infinity, in absolute value, as $t\to\infty$.

4.2 Stability of linear systems

In this section we consider the stability question for solutions of autonomous differential equations. Let $\mathbf{x}=\boldsymbol{\phi}(t)$ be a solution of the differential equation

$$\dot{\mathbf{x}}=\mathbf{f}(\mathbf{x}). \tag{1}$$

We are interested in determining whether $\boldsymbol{\phi}(t)$ is stable or unstable. That is to say, we seek to determine whether every solution $\boldsymbol{\psi}(t)$ of (1) which starts sufficiently close to $\boldsymbol{\phi}(t)$ at $t=0$ must remain close to $\boldsymbol{\phi}(t)$ for all future time $t\geqslant 0$. We begin with the following formal definition of stability.

Definition. The solution $\mathbf{x}=\boldsymbol{\phi}(t)$ of (1) is stable if every solution $\boldsymbol{\psi}(t)$ of (1) which starts sufficiently close to $\boldsymbol{\phi}(t)$ at $t=0$ must remain close to $\boldsymbol{\phi}(t)$ for all future time t. The solution $\boldsymbol{\phi}(t)$ is unstable if there exists at least one solution $\boldsymbol{\psi}(t)$ of (1) which starts near $\boldsymbol{\phi}(t)$ at $t=0$ but which does not remain close to $\boldsymbol{\phi}(t)$ for all future time. More precisely, the solution $\boldsymbol{\phi}(t)$ is stable if for every $\varepsilon>0$ there exists $\delta=\delta(\varepsilon)$ such that

$$|\psi_j(t)-\phi_j(t)|<\varepsilon \quad \text{if} \quad |\psi_j(0)-\phi_j(0)|<\delta(\varepsilon), \qquad j=1,\dots,n$$

for every solution $\boldsymbol{\psi}(t)$ of (1).

The stability question can be completely resolved for each solution of the linear differential equation

$$\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}. \tag{2}$$

This is not surprising, of course, since we can solve Equation (2) exactly. We have the following important theorem.

Theorem 1. (a) *Every solution* $\mathbf{x}=\boldsymbol{\phi}(t)$ *of* (2) *is stable if all the eigenvalues of* $\mathbf{A}$ *have negative real part.*

(b) *Every solution* $\mathbf{x}=\boldsymbol{\phi}(t)$ *of* (2) *is unstable if at least one eigenvalue of* $\mathbf{A}$ *has positive real part.*

(c) *Suppose that all the eigenvalues of* $\mathbf{A}$ *have real part* $\leqslant 0$ *and* $\lambda_1=i\sigma_1,\dots,\lambda_l=i\sigma_l$ *have zero real part. Let* $\lambda_j=i\sigma_j$ *have multiplicity* k_j. *This means that the characteristic polynomial of* $\mathbf{A}$ *can be factored into the form*

$$p(\lambda)=(\lambda-i\sigma_1)^{k_1}\dots(\lambda-i\sigma_l)^{k_l}q(\lambda)$$

where all the roots of $q(\lambda)$ have negative real part. Then, every solution $\mathbf{x}=\boldsymbol{\phi}(t)$ of (1) *is stable if* $\mathbf{A}$ *has k_j linearly independent eigenvectors for each eigenvalue $\lambda_j = i\sigma_j$. Otherwise, every solution $\boldsymbol{\phi}(t)$ is unstable.*

Our first step in proving Theorem 1 is to show that every solution $\boldsymbol{\phi}(t)$ is stable if the equilibrium solution $\mathbf{x}(t)\equiv\mathbf{0}$ is stable, and every solution $\boldsymbol{\phi}(t)$ is unstable if $\mathbf{x}(t)\equiv\mathbf{0}$ is unstable. To this end, let $\boldsymbol{\psi}(t)$ be any solution of (2). Observe that $\mathbf{z}(t)=\boldsymbol{\phi}(t)-\boldsymbol{\psi}(t)$ is again a solution of (2). Therefore, if the equilibrium solution $\mathbf{x}(t)\equiv\mathbf{0}$ is stable, then $\mathbf{z}(t)=\boldsymbol{\phi}(t)-\boldsymbol{\psi}(t)$ will always remain small if $\mathbf{z}(0)=\boldsymbol{\phi}(0)-\boldsymbol{\psi}(0)$ is sufficiently small. Consequently, every solution $\boldsymbol{\phi}(t)$ of (2) is stable. On the other hand suppose that $\mathbf{x}(t)\equiv\mathbf{0}$ is unstable. Then, there exists a solution $\mathbf{x}=\mathbf{h}(t)$ which is very small initially, but which becomes large as t approaches infinity. The function $\boldsymbol{\psi}(t)=\boldsymbol{\phi}(t)+\mathbf{h}(t)$ is clearly a solution of (2). Moreover, $\boldsymbol{\psi}(t)$ is close to $\boldsymbol{\phi}(t)$ initially, but diverges from $\boldsymbol{\phi}(t)$ as t increases. Therefore, every solution $\mathbf{x}=\boldsymbol{\phi}(t)$ of (2) is unstable.

Our next step in proving Theorem 1 is to reduce the problem of showing that n quantities $\psi_j(t), j=1,\dots,n$ are small to the much simpler problem of showing that only one quantity is small. This is accomplished by introducing the concept of length, or magnitude, of a vector.

Definition. Let

$$\mathbf{x}=\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

be a vector with n components. The numbers $x_1,\dots,x_n$ may be real or complex. We define the length of $\mathbf{x}$, denoted by $\|\mathbf{x}\|$ as

$$\|\mathbf{x}\| \equiv \max\{|x_1|,|x_2|,\dots,|x_n|\}.$$

For example, if

$$\mathbf{x}=\begin{pmatrix} 1 \\ 2 \\ -3 \end{pmatrix},$$

then $\|\mathbf{x}\|=3$ and if

$$\mathbf{x}=\begin{pmatrix} 1+2i \\ 2 \\ -1 \end{pmatrix}$$

then $\|\mathbf{x}\|=\sqrt{5}$.

The concept of the length, or magnitude of a vector corresponds to the concept of the length, or magnitude of a number. Observe that $\|\mathbf{x}\|\geqslant 0$ for any vector $\mathbf{x}$ and $\|\mathbf{x}\|=0$ only if $\mathbf{x}=\mathbf{0}$. Second, observe that

$$\|\lambda\mathbf{x}\|=\max\{|\lambda x_1|,\dots,|\lambda x_n|\}=|\lambda|\max\{|x_1|,\dots,|x_n|\}=|\lambda|\|\mathbf{x}\|.$$

Finally, observe that

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|&=\max\{|x_1+y_1|,\dots,|x_n+y_n|\}\leqslant\max\{|x_1|+|y_1|,\dots,|x_n|+|y_n|\}\\&\leqslant\max\{|x_1|,\dots,|x_n|\}+\max\{|y_1|,\dots,|y_n|\}=\|\mathbf{x}\|+\|\mathbf{y}\|.\end{aligned}$$

Thus, our definition really captures the meaning of length.

In Section 4.7 we give a simple geometric proof of Theorem 1 for the case $n=2$. The following proof is valid for arbitrary n.

Proof of Theorem 1. (a) Every solution $\mathbf{x}=\boldsymbol{\psi}(t)$ of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$ is of the form $\boldsymbol{\psi}(t)=e^{\mathbf{A}t}\boldsymbol{\psi}(0)$. Let $\phi_{ij}(t)$ be the ij element of the matrix $e^{\mathbf{A}t}$, and let $\psi_1^0,\dots,\psi_n^0$ be the components of $\boldsymbol{\psi}(0)$. Then, the ith component of $\boldsymbol{\psi}(t)$ is

$$\psi_i(t)=\phi_{i1}(t)\psi_1^0+\dots+\phi_{in}(t)\psi_n^0\equiv\sum_{j=1}^{n}\phi_{ij}(t)\psi_j^0.$$

Suppose that all the eigenvalues of $\mathbf{A}$ have negative real part. Let $-\alpha_1$ be the largest of the real parts of the eigenvalues of $\mathbf{A}$. It is a simple matter to show (see Exercise 17) that for every number $-\alpha$, with $-\alpha_1<-\alpha<0$, we can find a number K such that $|\phi_{ij}(t)|\leqslant Ke^{-\alpha t}$, $t\geqslant 0$. Consequently,

$$|\psi_i(t)|\leqslant\sum_{j=1}^{n}Ke^{-\alpha t}|\psi_j^0|=Ke^{-\alpha t}\sum_{j=1}^{n}|\psi_j^0|$$

for some positive constants K and α. Now, $|\psi_j^0|\leqslant\|\boldsymbol{\psi}(0)\|$. Hence,

$$\|\boldsymbol{\psi}(t)\|=\max\{|\psi_1(t)|,\dots,|\psi_n(t)|\}\leqslant nKe^{-\alpha t}\|\boldsymbol{\psi}(0)\|.$$

Let $\varepsilon>0$ be given. Choose $\delta(\varepsilon)=\varepsilon/nK$. Then, $\|\boldsymbol{\psi}(t)\|<\varepsilon$ if $\|\boldsymbol{\psi}(0)\|<\delta(\varepsilon)$ and $t\geqslant 0$, since

$$\|\boldsymbol{\psi}(t)\|\leqslant nKe^{-\alpha t}\|\boldsymbol{\psi}(0)\|<nK\varepsilon/nK=\varepsilon.$$

Consequently, the equilibrium solution $\mathbf{x}(t)\equiv\mathbf{0}$ is stable.

(b) Let λ be an eigenvalue of $\mathbf{A}$ with positive real part and let $\mathbf{v}$ be an eigenvector of $\mathbf{A}$ with eigenvalue λ. Then, $\boldsymbol{\psi}(t)=ce^{\lambda t}\mathbf{v}$ is a solution of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$ for any constant c. If λ is real then $\mathbf{v}$ is also real and $\|\boldsymbol{\psi}(t)\|=|c|e^{\lambda t}\|\mathbf{v}\|$. Clearly, $\|\boldsymbol{\psi}(t)\|$ approaches infinity as t approaches infinity, for any choice of $c\neq 0$, no matter how small. Therefore, $\mathbf{x}(t)\equiv\mathbf{0}$ is unstable. If $\lambda=\alpha+i\beta$ is complex, then $\mathbf{v}=\mathbf{v}^1+i\mathbf{v}^2$ is also complex. In this case

$$\begin{aligned}e^{(\alpha+i\beta)t}(\mathbf{v}^1+i\mathbf{v}^2)&=e^{\alpha t}(\cos\beta t+i\sin\beta t)(\mathbf{v}^1+i\mathbf{v}^2)\\&=e^{\alpha t}\left[(\mathbf{v}^1\cos\beta t-\mathbf{v}^2\sin\beta t)+i(\mathbf{v}^1\sin\beta t+\mathbf{v}^2\cos\beta t)\right]\end{aligned}$$

is a complex-valued solution of (2). Therefore

$$\boldsymbol{\psi}^1(t)=ce^{\alpha t}(\mathbf{v}^1\cos\beta t-\mathbf{v}^2\sin\beta t)$$

is a real-valued solution of (2), for any choice of constant c. Clearly,

$\|\boldsymbol{\psi}^1(t)\|$ is unbounded as t approaches infinity if c and either $\mathbf{v}^1$ or $\mathbf{v}^2$ is nonzero. Thus, $\mathbf{x}(t)\equiv\mathbf{0}$ is unstable.

(c) If $\mathbf{A}$ has k_j linearly independent eigenvectors for each eigenvalue $\lambda_j = i\sigma_j$ of multiplicity k_j, then we can find a constant K such that $|(e^{\mathbf{A}t})_{ij}| \leqslant K$ (see Exercise 18). There, $\|\boldsymbol{\psi}(t)\| \leqslant nK\|\boldsymbol{\psi}(0)\|$ for every solution $\boldsymbol{\psi}(t)$ of (2). It now follows immediately from the proof of (a) that $\mathbf{x}(t)\equiv\mathbf{0}$ is stable.

On the other hand, if $\mathbf{A}$ has fewer than k_j linearly independent eigenvectors with eigenvalue $\lambda_j = i\sigma_j$, then $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$ has solutions $\boldsymbol{\psi}(t)$ of the form

$$\boldsymbol{\psi}(t) = ce^{i\sigma_j t}\left[\mathbf{v} + t(\mathbf{A} - i\sigma_j\mathbf{I})\mathbf{v}\right]$$

where $(\mathbf{A} - i\sigma_j\mathbf{I})\mathbf{v}\neq\mathbf{0}$. If $\sigma_j = 0$, then $\boldsymbol{\psi}(t) = c(\mathbf{v} + t\mathbf{A}\mathbf{v})$ is real-valued. Moreover, $\|\boldsymbol{\psi}(t)\|$ is unbounded as t approaches infinity for any choice of $c\neq 0$. Similarly, both the real and imaginary parts of $\boldsymbol{\psi}(t)$ are unbounded in magnitude for arbitrarily small $\boldsymbol{\psi}(0)\neq\mathbf{0}$, if $\sigma_j\neq 0$. Therefore, the equilibrium solution $\mathbf{x}(t)\equiv\mathbf{0}$ is unstable. □

If all the eigenvalues of $\mathbf{A}$ have negative real part, then every solution $\mathbf{x}(t)$ of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$ approaches zero as t approaches infinity. This follows immediately from the estimate $\|\mathbf{x}(t)\| \leqslant Ke^{-\alpha t}\|\mathbf{x}(0)\|$ which we derived in the proof of part (a) of Theorem 1. Thus, not only is the equilibrium solution $\mathbf{x}(t)\equiv\mathbf{0}$ stable, but every solution $\boldsymbol{\psi}(t)$ of (2) approaches it as t approaches infinity. This very strong type of stability is known as *asymptotic stability*.

Definition. A solution $\mathbf{x}=\boldsymbol{\phi}(t)$ of (1) is asymptotically stable if it is stable, and if every solution $\boldsymbol{\psi}(t)$ which starts sufficiently close to $\boldsymbol{\phi}(t)$ must approach $\boldsymbol{\phi}(t)$ as t approaches infinity. In particular, an equilibrium solution $\mathbf{x}(t)=\mathbf{x}^0$ of (1) is asymptotically stable if every solution $\mathbf{x}=\boldsymbol{\psi}(t)$ of (1) which starts sufficiently close to $\mathbf{x}^0$ at time $t=0$ not only remains close to $\mathbf{x}^0$ for all future time, but ultimately approaches $\mathbf{x}^0$ as t approaches infinity.

Remark. The asymptotic stability of any solution $\mathbf{x}=\boldsymbol{\phi}(t)$ of (2) is clearly equivalent to the asymptotic stability of the equilibrium solution $\mathbf{x}(t)\equiv\mathbf{0}$.

Example 1. Determine whether each solution $\mathbf{x}(t)$ of the differential equation

$$\dot{\mathbf{x}} = \begin{pmatrix} -1 & 0 & 0 \\ -2 & -1 & 2 \\ -3 & -2 & -1 \end{pmatrix}\mathbf{x}$$

is stable, asymptotically stable, or unstable.

Solution. The characteristic polynomial of the matrix

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

is

$$p(\lambda)=(\det\mathbf{A}-\lambda\mathbf{I})=\det\begin{bmatrix}-1-\lambda & 0 & 0\\ -2 & -1-\lambda & 2\\ -3 & -2 & -1-\lambda\end{bmatrix}$$
$$=-(1+\lambda)^3-4(1+\lambda)=-(1+\lambda)(\lambda^2+2\lambda+5).$$

Hence, $\lambda=-1$ and $\lambda=-1\pm 2i$ are the eigenvalues of $\mathbf{A}$. Since all three eigenvalues have negative real part, we conclude that every solution of the differential equation $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$ is asymptotically stable.

Example 2. Prove that every solution of the differential equation

$$\dot{\mathbf{x}}=\begin{pmatrix}1 & 5\\ 5 & 1\end{pmatrix}\mathbf{x}$$

is unstable.
Solution. The characteristic polynomial of the matrix

$$\mathbf{A}=\begin{pmatrix}1 & 5\\ 5 & 1\end{pmatrix}$$

is

$$p(\lambda)=\det(\mathbf{A}-\lambda\mathbf{I})=\det\begin{pmatrix}1-\lambda & 5\\ 5 & 1-\lambda\end{pmatrix}=(1-\lambda)^2-25.$$

Hence $\lambda=6$ and $\lambda=-4$ are the eigenvalues of $\mathbf{A}$. Since one eigenvalue of $\mathbf{A}$ is positive, we conclude that every solution $\mathbf{x}=\boldsymbol{\phi}(t)$ of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$ is unstable.

Example 3. Show that every solution of the differential equation

$$\dot{\mathbf{x}}=\begin{pmatrix}0 & -3\\ 2 & 0\end{pmatrix}\mathbf{x}$$

is stable, but not asymptotically stable.
Solution. The characteristic polynomial of the matrix

$$\mathbf{A}=\begin{pmatrix}0 & -3\\ 2 & 0\end{pmatrix}$$

is

$$p(\lambda)=\det(\mathbf{A}-\lambda\mathbf{I})=\det\begin{pmatrix}-\lambda & -3\\ 2 & -\lambda\end{pmatrix}=\lambda^2+6.$$

Thus, the eigenvalues of $\mathbf{A}$ are $\lambda=\pm\sqrt{6}\,i$. Therefore, by part (c) of Theorem 1, every solution $\mathbf{x}=\boldsymbol{\phi}(t)$ of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$ is stable. However, no solution is asymptotically stable. This follows immediately from the fact that the general solution of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$ is

$$\mathbf{x}(t)=c_1\begin{pmatrix}-\sqrt{6}\sin\sqrt{6}\,t\\ 2\cos\sqrt{6}\,t\end{pmatrix}+c_2\begin{pmatrix}\sqrt{6}\cos\sqrt{6}\,t\\ 2\sin\sqrt{6}\,t\end{pmatrix}.$$

Hence, every solution $\mathbf{x}(t)$ is periodic, with period $2\pi/\sqrt{6}$, and no solution $\mathbf{x}(t)$ (except $\mathbf{x}(t)\equiv\mathbf{0}$) approaches 0 as t approaches infinity.

Example 4. Show that every solution of the differential equation

$$\dot{\mathbf{x}}=\begin{bmatrix} 2 & -3 & 0\\ 0 & -6 & -2\\ -6 & 0 & -3\end{bmatrix}\mathbf{x}$$

is unstable.

Solution. The characteristic polynomial of the matrix

$$\mathbf{A}=\begin{bmatrix} 2 & -3 & 0\\ 0 & -6 & -2\\ -6 & 0 & -3\end{bmatrix}$$

is

$$p(\lambda)=\det(\mathbf{A}-\lambda\mathbf{I})=\det\begin{bmatrix} 2-\lambda & -3 & 0\\ 0 & -6-\lambda & -2\\ -6 & 0 & -3-\lambda\end{bmatrix}=-\lambda^2(\lambda+7).$$

Hence, the eigenvalues of $\mathbf{A}$ are $\lambda=-7$ and $\lambda=0$. Every eigenvector $\mathbf{v}$ of $\mathbf{A}$ with eigenvalue 0 must satisfy the equation

$$\mathbf{A}\mathbf{v}=\begin{bmatrix} 2 & -3 & 0\\ 0 & -6 & -2\\ -6 & 0 & -3\end{bmatrix}\begin{bmatrix} v_1\\ v_2\\ v_3\end{bmatrix}=\begin{bmatrix} 0\\ 0\\ 0\end{bmatrix}.$$

This implies that $v_1=3v_2/2$ and $v_3=-3v_2$, so that every eigenvector $\mathbf{v}$ of $\mathbf{A}$ with eigenvalue 0 must be of the form

$$\mathbf{v}=c\begin{bmatrix} 3\\ 2\\ -6\end{bmatrix}.$$

Consequently, every solution $\mathbf{x}=\boldsymbol{\phi}(t)$ of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$ is unstable, since $\lambda=0$ is an eigenvalue of multiplicity two and $\mathbf{A}$ has only one linearly independent eigenvector with eigenvalue 0.

Exercises

Determine the stability or instability of all solutions of the following systems of differential equations.

1. $\dot{\mathbf{x}}=\begin{pmatrix} 1 & 1\\ -2 & -2\end{pmatrix}\mathbf{x}$

2. $\dot{\mathbf{x}}=\begin{pmatrix} -3 & -4\\ 2 & 1\end{pmatrix}\mathbf{x}$

3. $\dot{\mathbf{x}}=\begin{pmatrix} -5 & 3\\ -1 & 1\end{pmatrix}\mathbf{x}$

4. $\dot{\mathbf{x}}=\begin{pmatrix} 1 & -4\\ 4 & -7\end{pmatrix}\mathbf{x}$

5. $\dot{\mathbf{x}}=\begin{pmatrix} -7 & 1 & -6\\ 10 & -4 & 12\\ 2 & -1 & 1\end{pmatrix}\mathbf{x}$

6. $\dot{\mathbf{x}}=\begin{pmatrix} 3 & 2 & 4\\ 2 & 0 & 2\\ 4 & 2 & 3\end{pmatrix}\mathbf{x}$

7. $\dot{\mathbf{x}}=\begin{pmatrix} 0 & 2 & 1 \\ -1 & -3 & -1 \\ 1 & 1 & -1 \end{pmatrix}\mathbf{x}$ **8.** $\dot{\mathbf{x}}=\begin{pmatrix} -2 & 1 & 1 \\ -3 & 2 & 3 \\ 1 & -1 & -2 \end{pmatrix}\mathbf{x}$

9. $\dot{\mathbf{x}}=\begin{bmatrix} 0 & 2 & 0 & 0 \\ -2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & -2 & 0 \end{bmatrix}\mathbf{x}$ **10.** $\dot{\mathbf{x}}=\begin{bmatrix} 0 & 2 & 1 & 0 \\ -2 & 0 & 0 & 1 \\ 0 & 0 & 0 & 2 \\ 0 & 0 & -2 & 0 \end{bmatrix}\mathbf{x}$

11. Determine whether the solutions $x(t)\equiv 0$ and $x(t)\equiv 1$ of the single scalar equation $\dot{x}=x(1-x)$ are stable or unstable.

12. Determine whether the solutions $x(t)\equiv 0$ and $x(t)\equiv 1$ of the single scalar equation $\dot{x}=-x(1-x)$ are stable or unstable.

13. Consider the differential equation $\dot{x}=x^2$. Show that all solutions $x(t)$ with $x(0)\geqslant 0$ are unstable while all solutions $x(t)$ with $x(0)<0$ are asymptotically stable.

14. Consider the system of differential equations

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\tfrac{1}{2}\left[x_1^2+\left(x_1^2+4x_2^2\right)^{1/2}\right]x_1. \end{aligned} \tag{*}$$

(a) Show that

$$\mathbf{x}(t)=\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}=\begin{pmatrix} c\sin(ct+d) \\ c^2\cos(ct+d) \end{pmatrix}$$

is a solution of (*) for any choice of constants c and d.

(b) Assume that a solution $\mathbf{x}(t)$ of (*) is uniquely determined once $x_1(0)$ and $x_2(0)$ are prescribed. Prove that (a) represents the general solution of (*).

(c) Show that the solution $\mathbf{x}=\mathbf{0}$ of (*) is stable, but not asymptotically stable.

(d) Show that every solution $\mathbf{x}(t)\not\equiv\mathbf{0}$ of (*) is unstable.

15. Show that the stability of any solution $\mathbf{x}(t)$ of the nonhomogeneous equation $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}+\mathbf{f}(t)$ is equivalent to the stability of the equilibrium solution $\mathbf{x}\equiv\mathbf{0}$ of the homogeneous equation $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$.

16. Determine the stability or instability of all solutions $\mathbf{x}(t)$ of the differential equation

$$\dot{\mathbf{x}}=\begin{pmatrix} -1 & -1 \\ 2 & -1 \end{pmatrix}\mathbf{x}+\begin{pmatrix} 1 \\ 5 \end{pmatrix}.$$

17. (a) Let $f(t)=t^a e^{-bt}$, for some positive constants a and b, and let c be a positive number smaller than b. Show that we can find a positive constant K such that $|f(t)|\leqslant Ke^{-ct}$, $0\leqslant t<\infty$. *Hint*: Show that $f(t)/e^{-ct}$ approaches zero as t approaches infinity.

(b) Suppose that all the eigenvalues of $\mathbf{A}$ have negative real part. Show that we can find positive constants K and α such that $|(e^{\mathbf{A}t})_{ij}|\leqslant Ke^{-\alpha t}$ for $1\leqslant i$, $j\leqslant n$. *Hint*: Each component of $e^{\mathbf{A}t}$ is a finite linear combination of functions of the form $q(t)e^{\lambda t}$, where $q(t)$ is a polynomial in t (of degree $\leqslant n-1$) and λ is an eigenvalue of $\mathbf{A}$.

18. (a) Let $\mathbf{x}(t)=e^{i\sigma t}\mathbf{v}$, σ real, be a complex-valued solution of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$. Show that both the real and imaginary parts of $\mathbf{x}(t)$ are bounded solutions of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$.
(b) Suppose that all the eigenvalues of $\mathbf{A}$ have real part $\leqslant 0$ and $\lambda_1=i\sigma_1,\dots,\lambda_l=i\sigma_l$ have zero real part. Let $\lambda_j=i\sigma_j$ have multiplicity k_j, and suppose that $\mathbf{A}$ has k_j linearly independent eigenvectors for each eigenvalue λ_j, $j=1,\dots,l$. Prove that we can find a constant K such that $|(e^{\mathbf{A}t})_{ij}|\leqslant K$.

19. Let

$$\mathbf{x}=\begin{bmatrix} x_1\\ \vdots\\ x_n\end{bmatrix}$$

and define $\|x\|_1=|x_1|+\dots+|x_n|$. Show that
(i) $\|\mathbf{x}\|_1\geqslant 0$ and $\|\mathbf{x}\|_1=0$ only if $\mathbf{x}=\mathbf{0}$
(ii) $\|\lambda\mathbf{x}\|_1=|\lambda|\,\|\mathbf{x}\|_1$
(iii) $\|\mathbf{x}+\mathbf{y}\|_1\leqslant\|\mathbf{x}\|_1+\|\mathbf{y}\|_1$

20. Let

$$\mathbf{x}=\begin{bmatrix} x_1\\ \vdots\\ x_n\end{bmatrix}$$

and define $\|\mathbf{x}\|_2=[|x_1|^2+\dots+|x_n|^2]^{1/2}$. Show that
(i) $\|\mathbf{x}\|_2\geqslant 0$ and $\|\mathbf{x}\|_2=0$ only if $\mathbf{x}=\mathbf{0}$
(ii) $\|\lambda\mathbf{x}\|_2=|\lambda|\,\|\mathbf{x}\|_2$
(iii) $\|\mathbf{x}+\mathbf{y}\|_2\leqslant\|\mathbf{x}\|_2+\|\mathbf{y}\|_2$

21. Show that there exist constants M and N such that

$$M\|\mathbf{x}\|_1\leqslant\|\mathbf{x}\|_2\leqslant N\|\mathbf{x}\|_1.$$

4.3 Stability of equilibrium solutions

In Section 4.2 we treated the simple equation $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$. The next simplest equation is

$$\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}+\mathbf{g}(\mathbf{x}) \tag{1}$$

where

$$\mathbf{g}(\mathbf{x})=\begin{bmatrix} g_1(\mathbf{x})\\ \vdots\\ g_n(\mathbf{x})\end{bmatrix}$$

is very small compared to $\mathbf{x}$. Specifically we assume that

$$\frac{g_1(\mathbf{x})}{\max\{|x_1|,\dots,|x_n|\}},\dots,\frac{g_n(\mathbf{x})}{\max\{|x_1|,\dots,|x_n|\}}$$

are continuous functions of $x_1,\dots,x_n$ which vanish for $x_1=\dots=x_n=0$. This is always the case if each component of $\mathbf{g}(\mathbf{x})$ is a polynomial in $x_1,\dots,x_n$ which begins with terms of order 2 or higher. For example, if

$$\mathbf{g}(\mathbf{x})=\begin{pmatrix} x_1x_2^2 \\ x_1x_2 \end{pmatrix},$$

then both $x_1x_2^2/\max\{|x_1|,|x_2|\}$ and $x_1x_2/\max\{|x_1|,|x_2|\}$ are continuous functions of x_1,x_2 which vanish for $x_1=x_2=0$.

If $\mathbf{g}(\mathbf{0})=\mathbf{0}$ then $\mathbf{x}(t)\equiv\mathbf{0}$ is an equilibrium solution of (1). We would like to determine whether it is stable or unstable. At first glance this would seem impossible to do, since we cannot solve Equation (1) explicitly. However, if $\mathbf{x}$ is very small, then $\mathbf{g}(\mathbf{x})$ is very small compared to $\mathbf{A}\mathbf{x}$. Therefore, it seems plausible that the stability of the equilibrium solution $\mathbf{x}(t)\equiv\mathbf{0}$ of (1) should be determined by the stability of the "approximate" equation $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$. This is almost the case as the following theorem indicates.

Theorem 2. *Suppose that the vector-valued function*

$$\mathbf{g}(\mathbf{x})/\|\mathbf{x}\|\equiv\mathbf{g}(\mathbf{x})/\max\{|x_1|,\dots,|x_n|\}$$

is a continuous function of $x_1,\dots,x_n$ *which vanishes for* $\mathbf{x}=\mathbf{0}$. *Then,*

(a) *The equilibrium solution* $\mathbf{x}(t)\equiv\mathbf{0}$ *of* (1) *is asymptotically stable if the equilibrium solution* $\mathbf{x}(t)\equiv\mathbf{0}$ *of the "linearized" equation* $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$ *is asymptotically stable. Equivalently, the solution* $\mathbf{x}(t)\equiv\mathbf{0}$ *of* (1) *is asymptotically stable if all the eigenvalues of* $\mathbf{A}$ *have negative real part.*

(b) *The equilibrium solution* $\mathbf{x}(t)\equiv\mathbf{0}$ *of* (1) *is unstable if at least one eigenvalue of* $\mathbf{A}$ *has positive real part.*

(c) *The stability of the equilibrium solution* $\mathbf{x}(t)\equiv\mathbf{0}$ *of* (1) *cannot be determined from the stability of the equilibrium solution* $\mathbf{x}(t)\equiv\mathbf{0}$ *of* $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$ *if all the eigenvalues of* $\mathbf{A}$ *have real part* $\leqslant 0$ *but at least one eigenvalue of* $\mathbf{A}$ *has zero real part.*

PROOF. (a) The key step in many stability proofs is to use the variation of parameters formula of Section 3.12. This formula implies that any solution $\mathbf{x}(t)$ of (1) can be written in the form

$$\mathbf{x}(t)=e^{\mathbf{A}t}\mathbf{x}(0)+\int_0^t e^{\mathbf{A}(t-s)}\mathbf{g}(\mathbf{x}(s))\,ds. \tag{2}$$

We wish to show that $\|\mathbf{x}(t)\|$ approaches zero as t approaches infinity. To this end recall that if all the eigenvalues of $\mathbf{A}$ have negative real part, then we can find positive constants K and α such that (see Exercise 17, Section 4.2).

$$\|e^{\mathbf{A}t}\mathbf{x}(0)\|\leqslant Ke^{-\alpha t}\|\mathbf{x}(0)\|$$

and

$$\|e^{\mathbf{A}(t-s)}\mathbf{g}(\mathbf{x}(s))\|\leqslant Ke^{-\alpha(t-s)}\|\mathbf{g}(\mathbf{x}(s))\|.$$

Moreover, we can find a positive constant σ such that

$$\|\mathbf{g}(\mathbf{x})\| \leqslant \frac{\alpha}{2K}\|\mathbf{x}\| \quad \text{if} \quad \|\mathbf{x}\| \leqslant \sigma.$$

This follows immediately from our assumption that $g(\mathbf{x})/\|\mathbf{x}\|$ is continuous and vanishes at $\mathbf{x}=\mathbf{0}$. Consequently, Equation (2) implies that

$$\begin{aligned}\|\mathbf{x}(t)\| &\leqslant \|e^{\mathbf{A}t}\mathbf{x}(0)\| + \int_0^t \|e^{\mathbf{A}(t-s)}\mathbf{g}(\mathbf{x}(s))\|\,ds \\ &\leqslant Ke^{-\alpha t}\|\mathbf{x}(0)\| + \frac{\alpha}{2}\int_0^t e^{-\alpha(t-s)}\|\mathbf{x}(s)\|\,ds\end{aligned}$$

as long as $\|\mathbf{x}(s)\| \leqslant \sigma$, $0 \leqslant s \leqslant t$. Multiplying both sides of this inequality by $e^{\alpha t}$ gives

$$e^{\alpha t}\|\mathbf{x}(t)\| \leqslant K\|\mathbf{x}(0)\| + \frac{\alpha}{2}\int_0^t e^{\alpha s}\|\mathbf{x}(s)\|\,ds. \tag{3}$$

The inequality (3) can be simplified by setting $z(t)=e^{\alpha t}\|\mathbf{x}(t)\|$, for then

$$z(t) \leqslant K\|\mathbf{x}(0)\| + \frac{\alpha}{2}\int_0^t z(s)\,ds. \tag{4}$$

We would like to differentiate both sides of (4) with respect to t. However, we cannot, in general, differentiate both sides of an inequality and still preserve the sense of the inequality. We circumvent this difficulty by the clever trick of setting

$$U(t) = \frac{\alpha}{2}\int_0^t z(s)\,ds.$$

Then

$$\frac{dU(t)}{dt} = \frac{\alpha}{2}z(t) \leqslant \frac{\alpha}{2}K\|\mathbf{x}(0)\| + \frac{\alpha}{2}U(t)$$

or

$$\frac{dU(t)}{dt} - \frac{\alpha}{2}U(t) \leqslant \frac{\alpha K}{2}\|\mathbf{x}(0)\|.$$

Multiplying both sides of this inequality by the integrating factor $e^{-\alpha t/2}$ gives

$$\frac{d}{dt}e^{-\alpha t/2}U \leqslant \frac{\alpha K}{2}\|\mathbf{x}(0)\|e^{-\alpha t/2},$$

or

$$\frac{d}{dt}e^{-\alpha t/2}\left[U(t) + K\|\mathbf{x}(0)\|\right] \leqslant 0.$$

Consequently,

$$e^{-\alpha t/2}\left[U(t) + K\|\mathbf{x}(0)\|\right] \leqslant U(0) + K\|\mathbf{x}(0)\| = K\|\mathbf{x}(0)\|,$$

so that $U(t) \leqslant -K\|\mathbf{x}(0)\| + K\|\mathbf{x}(0)\|e^{\alpha t/2}$. Returning to the inequality (4), we see that

$$\begin{aligned}\|\mathbf{x}(t)\| = e^{-\alpha t}z(t) &\leqslant e^{-\alpha t}\left[K\|\mathbf{x}(0)\| + U(t)\right]\\ &\leqslant K\|\mathbf{x}(0)\|e^{-\alpha t/2}\end{aligned} \tag{5}$$

as long as $\|\mathbf{x}(s)\| \leqslant \sigma$, $0 \leqslant s \leqslant t$. Now, if $\|\mathbf{x}(0)\| \leqslant \sigma/K$, then the inequality (5) guarantees that $\|\mathbf{x}(t)\| \leqslant \sigma$ for all future time t. Consequently, the inequality (5) is true for all $t \geqslant 0$ if $\|\mathbf{x}(0)\| \leqslant \sigma/K$. Finally, observe from (5) that $\|\mathbf{x}(t)\| \leqslant K\|\mathbf{x}(0)\|$ and $\|\mathbf{x}(t)\|$ approaches zero as t approaches infinity. Therefore, the equilibrium solution $\mathbf{x}(t) \equiv \mathbf{0}$ of (1) is asymptotically stable.
(b) The proof of (b) is too difficult to present here.
(c) We will present two differential equations of the form (1) where the nonlinear term $\mathbf{g}(\mathbf{x})$ determines the stability of the equilibrium solution $\mathbf{x}(t) \equiv \mathbf{0}$. Consider first the system of differential equations

$$\frac{dx_1}{dt} = x_2 - x_1\left(x_1^2 + x_2^2\right), \qquad \frac{dx_2}{dt} = -x_1 - x_2\left(x_1^2 + x_2^2\right). \tag{6}$$

The linearized equation is

$$\frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

and the eigenvalues of the matrix

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

are $\pm i$. To analyze the behavior of the nonlinear system (6) we multiply the first equation by x_1, the second equation by x_2 and add; this gives

$$\begin{aligned} x_1\frac{dx_1}{dt} + x_2\frac{dx_2}{dt} &= -x_1^2\left(x_1^2 + x_2^2\right) - x_2^2\left(x_1^2 + x_2^2\right)\\ &= -\left(x_1^2 + x_2^2\right)^2. \end{aligned}$$

But

$$x_1\frac{dx_1}{dt} + x_2\frac{dx_2}{dt} = \frac{1}{2}\frac{d}{dt}\left(x_1^2 + x_2^2\right).$$

Hence,

$$\frac{d}{dt}\left(x_1^2 + x_2^2\right) = -2\left(x_1^2 + x_2^2\right)^2.$$

This implies that

$$x_1^2(t) + x_2^2(t) = \frac{c}{1 + 2ct},$$

where

$$c = x_1^2(0) + x_2^2(0).$$

Thus, $x_1^2(t)+x_2^2(t)$ approaches zero as t approaches infinity for any solution $x_1(t)$, $x_2(t)$ of (6). Moreover, the value of $x_1^2+x_2^2$ at any time t is always less than its value at $t=0$. We conclude, therefore, that $x_1(t)\equiv 0$, $x_2(t)\equiv 0$ is asymptotically stable.

On the other hand, consider the system of equations

$$\frac{dx_1}{dt}=x_2+x_1\left(x_1^2+x_2^2\right), \qquad \frac{dx_2}{dt}=-x_1-x_2\left(x_1^2+x_2^2\right). \tag{7}$$

Here too, the linearized system is

$$\dot{\mathbf{x}}=\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\mathbf{x}.$$

In this case, though, $(d/dt)(x_1^2+x_2^2)=2(x_1^2+x_2^2)^2$. This implies that

$$x_1^2(t)+x_2^2(t)=\frac{c}{1-2ct}, \qquad c=x_1^2(0)+x_2^2(0).$$

Notice that every solution $x_1(t)$, $x_2(t)$ of (7) with $x_1^2(0)+x_2^2(0)\neq 0$ approaches infinity in finite time. We conclude, therefore, that the equilibrium solution $x_1(t)\equiv 0$, $x_2(t)\equiv 0$ is unstable. □

Example 1. Consider the system of differential equations

$$\begin{aligned}
\frac{dx_1}{dt}&=-2x_1+x_2+3x_3+9x_2^3\\
\frac{dx_2}{dt}&=-6x_2-5x_3+7x_3^5\\
\frac{dx_3}{dt}&=-x_3+x_1^2+x_2^2.
\end{aligned}$$

Determine, if possible, whether the equilibrium solution $x_1(t)\equiv 0$, $x_2(t)\equiv 0$, $x_3(t)\equiv 0$ is stable or unstable.

Solution. We rewrite this system in the form $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}+\mathbf{g}(\mathbf{x})$ where

$$\mathbf{x}=\begin{bmatrix} x_1\\ x_2\\ x_3\end{bmatrix}, \qquad \mathbf{A}=\begin{bmatrix} -2 & 1 & 3\\ 0 & -6 & -5\\ 0 & 0 & -1\end{bmatrix} \quad\text{and}\quad \mathbf{g}(\mathbf{x})=\begin{bmatrix} 9x_2^3\\ 7x_3^5\\ x_1^2+x_2^2\end{bmatrix}.$$

The function $\mathbf{g}(\mathbf{x})$ satisfies the hypotheses of Theorem 2, and the eigenvalues of $\mathbf{A}$ are -2, -6 and -1. Hence, the equilibrium solution $\mathbf{x}(t)\equiv\mathbf{0}$ is asymptotically stable.

Theorem 2 can also be used to determine the stability of equilibrium solutions of arbitrary autonomous differential equations. Let $\mathbf{x}^0$ be an equilibrium value of the differential equation

$$\dot{\mathbf{x}}=\mathbf{f}(\mathbf{x}) \tag{8}$$

and set $\mathbf{z}(t)=\mathbf{x}(t)-\mathbf{x}^0$. Then

$$\dot{\mathbf{z}}=\dot{\mathbf{x}}=\mathbf{f}(\mathbf{x}^0+\mathbf{z}). \tag{9}$$

Clearly, $\mathbf{z}(t)\equiv\mathbf{0}$ is an equilibrium solution of (9) and the stability of $\mathbf{x}(t)\equiv\mathbf{x}^0$ is equivalent to the stability of $\mathbf{z}(t)\equiv\mathbf{0}$.

Next, we show that $\mathbf{f}(\mathbf{x}^0+\mathbf{z})$ can be written in the form $\mathbf{f}(\mathbf{x}^0+\mathbf{z})=\mathbf{A}\mathbf{z}+\mathbf{g}(\mathbf{z})$ where $\mathbf{g}(\mathbf{z})$ is small compared to $\mathbf{z}$.

Lemma 1. *Let* $\mathbf{f}(\mathbf{x})$ *have two continuous partial derivatives with respect to each of its variables* $x_1,\dots,x_n$. *Then,* $\mathbf{f}(\mathbf{x}^0+\mathbf{z})$ *can be written in the form*

$$\mathbf{f}(\mathbf{x}^0+\mathbf{z})=\mathbf{f}(\mathbf{x}^0)+\mathbf{A}\mathbf{z}+\mathbf{g}(\mathbf{z}) \tag{10}$$

where $\mathbf{g}(\mathbf{z})/\max\{|z_1|,\dots,|z_n|\}$ *is a continuous function of* $\mathbf{z}$ *which vanishes for* $\mathbf{z}=\mathbf{0}$.

PROOF #1. Equation (10) is an immediate consequence of Taylor's Theorem which states that each component $f_j(\mathbf{x}^0+\mathbf{z})$ of $\mathbf{f}(\mathbf{x}^0+\mathbf{z})$ can be written in the form

$$f_j(\mathbf{x}^0+\mathbf{z})=f_j(\mathbf{x}^0)+\frac{\partial f_j(\mathbf{x}^0)}{\partial x_1}z_1+\dots+\frac{\partial f_j(\mathbf{x}^0)}{\partial x_n}z_n+g_j(\mathbf{z})$$

where $g_j(\mathbf{z})/\max\{|z_1|,\dots,|z_n|\}$ is a continuous function of $\mathbf{z}$ which vanishes for $\mathbf{z}=\mathbf{0}$. Hence,

$$\mathbf{f}(\mathbf{x}^0+\mathbf{z})=\mathbf{f}(\mathbf{x}^0)+\mathbf{A}\mathbf{z}+\mathbf{g}(\mathbf{z})$$

where

$$\mathbf{A}=\begin{bmatrix} \dfrac{\partial f_1(\mathbf{x}^0)}{\partial x_1} & \cdots & \dfrac{\partial f_1(\mathbf{x}^0)}{\partial x_n} \\ \vdots & & \vdots \\ \dfrac{\partial f_n(\mathbf{x}^0)}{\partial x_1} & \cdots & \dfrac{\partial f_n(\mathbf{x}^0)}{\partial x_n} \end{bmatrix}. \qquad \square$$

PROOF #2. If each component of $\mathbf{f}(\mathbf{x})$ is a polynomial (possibly infinite) in $x_1,\dots,x_n$, then each component of $\mathbf{f}(\mathbf{x}^0+\mathbf{z})$ is a polynomial in $z_1,\dots,z_n$. Thus,

$$f_j(\mathbf{x}^0+\mathbf{z})=a_{j0}+a_{j1}z_1+\dots+a_{jn}z_n+g_j(\mathbf{z}) \tag{11}$$

where $g_j(\mathbf{z})$ is a polynomial in $z_1,\dots,z_n$ beginning with terms of order two. Setting $\mathbf{z}=\mathbf{0}$ in (11) gives $f_j(\mathbf{x}^0)=a_{j0}$. Hence,

$$\mathbf{f}(\mathbf{x}^0+\mathbf{z})=\mathbf{f}(\mathbf{x}^0)+\mathbf{A}\mathbf{z}+\mathbf{g}(\mathbf{z}), \qquad \mathbf{A}=\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$$

and each component of $\mathbf{g}(\mathbf{z})$ is a polynomial in $z_1,\dots,z_n$ beginning with terms of order two. □

Theorem 2 and Lemma 1 provide us with the following algorithm for determining whether an equilibrium solution $\mathbf{x}(t)\equiv\mathbf{x}^0$ of $\dot{\mathbf{x}}=\mathbf{f}(\mathbf{x})$ is stable or unstable:

1. Set $\mathbf{z}=\mathbf{x}-\mathbf{x}^0$.
2. Write $\mathbf{f}(\mathbf{x}^0+\mathbf{z})$ in the form $\mathbf{A}\mathbf{z}+\mathbf{g}(\mathbf{z})$ where $\mathbf{g}(\mathbf{z})$ is a vector-valued polynomial in $z_1,\dots,z_n$ beginning with terms of order two or more.
3. Compute the eigenvalues of $\mathbf{A}$. If all the eigenvalues of $\mathbf{A}$ have negative real part, then $\mathbf{x}(t)\equiv\mathbf{x}^0$ is asymptotically stable. If one eigenvalue of $\mathbf{A}$ has positive real part, then $\mathbf{x}(t)\equiv\mathbf{x}^0$ is unstable.

Example 2. Find all equilibrium solutions of the system of differential equations

$$\frac{dx}{dt}=1-xy,\qquad \frac{dy}{dt}=x-y^3 \tag{12}$$

and determine (if possible) whether they are stable or unstable.

Solution. The equations $1-xy=0$ and $x-y^3=0$ imply that $x=1, y=1$ or $x=-1, y=-1$. Hence, $x(t)\equiv 1, y(t)\equiv 1$, and $x(t)\equiv -1, y(t)\equiv -1$ are the only equilibrium solutions of (12).

(i) $x(t)=1, y(t)=1$: Set $u=x-1$, $v=y-1$. Then,

$$\frac{du}{dt}=\frac{dx}{dt}=1-(1+u)(1+v)=-u-v-uv$$

$$\frac{dv}{dt}=\frac{dy}{dt}=(1+u)-(1+v)^3=u-3v-3v^2-v^3.$$

We rewrite this system in the form

$$\frac{d}{dt}\begin{pmatrix}u\\v\end{pmatrix}=\begin{pmatrix}-1&-1\\1&-3\end{pmatrix}\begin{pmatrix}u\\v\end{pmatrix}-\begin{pmatrix}uv\\3v^2+v^3\end{pmatrix}.$$

The matrix

$$\begin{pmatrix}-1&-1\\1&-3\end{pmatrix}$$

has a single eigenvalue $\lambda=-2$ since

$$\det\begin{pmatrix}-1-\lambda&-1\\1&-3-\lambda\end{pmatrix}=(1+\lambda)(3+\lambda)+1=(\lambda+2)^2.$$

Hence, the equilibrium solution $x(t)\equiv 1, y(t)\equiv 1$ of (12) is asymptotically stable.

(ii) $x(t)\equiv -1, y(t)\equiv -1$: Set $u=x+1$, $v=y+1$. Then,

$$\frac{du}{dt}=\frac{dx}{dt}=1-(u-1)(v-1)=u+v-uv$$

$$\frac{dv}{dt}=\frac{dy}{dt}=(u-1)-(v-1)^3=u-3v+3v^2-v^3.$$

We rewrite this system in the form

$$\frac{d}{dt}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -3 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} -uv \\ 3v^2 - v^3 \end{pmatrix}.$$

The eigenvalues of the matrix

$$\begin{pmatrix} 1 & 1 \\ 1 & -3 \end{pmatrix}$$

are $\lambda_1 = -1 - \sqrt{5}$, which is negative, and $\lambda_2 = -1 + \sqrt{5}$, which is positive. Therefore, the equilibrium solution $x(t) \equiv -1$, $y(t) \equiv -1$ of (12) is unstable.

Example 3. Find all equilibrium solutions of the system of differential equations

$$\frac{dx}{dt} = \sin(x+y), \qquad \frac{dy}{dt} = e^x - 1 \tag{13}$$

and determine whether they are stable or unstable.

Solution. The equilibrium points of (13) are determined by the two equations $\sin(x+y) = 0$ and $e^x - 1 = 0$. The second equation implies that $x = 0$, while the first equation implies that $x + y = n\pi$, n an integer. Consequently, $x(t) \equiv 0$, $y(t) \equiv n\pi$, $n = 0, \pm 1, \pm 2, \ldots$, are the equilibrium solutions of (13). Setting $u = x$, $v = y - n\pi$, gives

$$\frac{du}{dt} = \sin(u + v + n\pi), \qquad \frac{dv}{dt} = e^u - 1.$$

Now, $\sin(u + v + n\pi) = \cos n\pi \sin(u+v) = (-1)^n \sin(u+v)$. Therefore,

$$\frac{du}{dt} = (-1)^n \sin(u+v), \qquad \frac{dv}{dt} = e^u - 1.$$

Next, observe that

$$\sin(u+v) = u + v - \frac{(u+v)^3}{3!} + \ldots, \qquad e^u - 1 = u + \frac{u^2}{2!} + \ldots.$$

Hence,

$$\frac{du}{dt} = (-1)^n\left[(u+v) - \frac{(u+v)^3}{3!} + \ldots\right], \qquad \frac{dv}{dt} = u + \frac{u^2}{2!} + \ldots.$$

We rewrite this system in the form

$$\frac{d}{dt}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} (-1)^n & (-1)^n \\ 1 & 0 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} + \begin{matrix} \text{terms of order 2 or} \\ \text{higher in } u \text{ and } v. \end{matrix}$$

The eigenvalues of the matrix

$$\begin{pmatrix} (-1)^n & (-1)^n \\ 1 & 0 \end{pmatrix}$$

are

$$\lambda_1 = \frac{(-1)^n - \sqrt{1+4(-1)^n}}{2}, \qquad \lambda_2 = \frac{(-1)^n + \sqrt{1+4(-1)^n}}{2}.$$

When n is even, $\lambda_1 = (1-\sqrt{5})/2$ is negative and $\lambda_2 = (1+\sqrt{5})/2$ is positive. Hence, $x(t)\equiv 0, y(t)\equiv n\pi$ is unstable if n is even. When n is odd, both $\lambda_1 = (-1-\sqrt{3}\,i)/2$ and $\lambda_2 = (-1+\sqrt{3}\,i)/2$ have negative real part. Therefore, the equilibrium solution $x(t)\equiv 0, y(t)\equiv n\pi$ is asymptotically stable if n is odd.

Exercises

Find all equilibrium solutions of each of the following systems of equations and determine, if possible, whether they are stable or unstable.

1. $\dot{x} = x - x^3 - xy^2$
$\dot{y} = 2y - y^5 - yx^4$

2. $\dot{x} = x^2 + y^2 - 1$
$\dot{y} = x^2 - y^2$

3. $\dot{x} = x^2 + y^2 - 1$
$\dot{y} = 2xy$

4. $\dot{x} = 6x - 6x^2 - 2xy$
$\dot{y} = 4y - 4y^2 - 2xy$

5. $\dot{x} = \tan(x+y)$
$\dot{y} = x + x^3$

6. $\dot{x} = e^y - x$
$\dot{y} = e^x + y$

Verify that the origin is an equilibrium point of each of the following systems of equations and determine, if possible, whether it is stable or unstable.

7. $\dot{x} = y + 3x^2$
$\dot{y} = x - 3y^2$

8. $\dot{x} = y + \cos y - 1$
$\dot{y} = -\sin x + x^3$

9. $\dot{x} = e^{x+y} - 1$
$\dot{y} = \sin(x+y)$

10. $\dot{x} = \ln(1 + x + y^2)$
$\dot{y} = -y + x^3$

11. $\dot{x} = \cos y - \sin x - 1$
$\dot{y} = x - y - y^2$

12. $\dot{x} = 8x - 3y + e^y - 1$
$\dot{y} = \sin x^2 - \ln(1 - x - y)$

13. $\dot{x} = -x - y - (x^2 + y^2)^{3/2}$
$\dot{y} = x - y + (x^2 + y^2)^{3/2}$

14. $\dot{x} = x - y + z^2$
$\dot{y} = y + z - x^2$
$\dot{z} = z - x + y^2$

15. $\dot{x} = e^{x+y+z} - 1$
$\dot{y} = \sin(x+y+z)$
$\dot{z} = x - y - z^2$

16. $\dot{x} = \ln(1 - z)$
$\dot{y} = \ln(1 - x)$
$\dot{z} = \ln(1 - y)$

17. $\dot{x} = x - \cos y - z + 1$
$\dot{y} = y - \cos z - x + 1$
$\dot{z} = z - \cos x - y + 1$

18 (a) Find all equilibrium solutions of the system of differential equations

$$\frac{dx}{dt} = gz - hx, \qquad \frac{dy}{dt} = \frac{c}{a+bx} - ky, \qquad \frac{dz}{dt} = ey - fz.$$

(This system is a model for the control of protein synthesis.)

(b) Determine the stability or instability of these solutions if either g, e, or c is zero.

4.4 The phase-plane

In this section we begin our study of the "geometric" theory of differential equations. For simplicity, we will restrict ourselves, for the most part, to the case $n=2$. Our aim is to obtain as complete a description as possible of all solutions of the system of differential equations

$$\frac{dx}{dt}=f(x,y), \qquad \frac{dy}{dt}=g(x,y). \tag{1}$$

To this end, observe that every solution $x=x(t)$, $y=y(t)$ of (1) defines a curve in the three-dimensional space t,x,y. That is to say, the set of all points $(t,x(t),y(t))$ describe a curve in the three-dimensional space t,x,y. For example, the solution $x=\cos t$, $y=\sin t$ of the system of differential equations

$$\frac{dx}{dt}=-y, \qquad \frac{dy}{dt}=x$$

describes a helix (see Figure 1) in (t,x,y) space.

The geometric theory of differential equations begins with the important observation that every solution $x=x(t)$, $y=y(t)$, $t_0\leqslant t\leqslant t_1$, of (1) also defines a curve in the $x-y$ plane. To wit, as t runs from t_0 to t_1, the set of points $(x(t),y(t))$ trace out a curve C in the $x-y$ plane. This curve is called the *orbit*, or *trajectory*, of the solution $x=x(t)$, $y=y(t)$, and the $x-y$ plane is called the *phase-plane* of the solutions of (1). Equivalently, we can think of the orbit of $x(t)$, $y(t)$ as the path that the solution traverses in the $x-y$ plane.

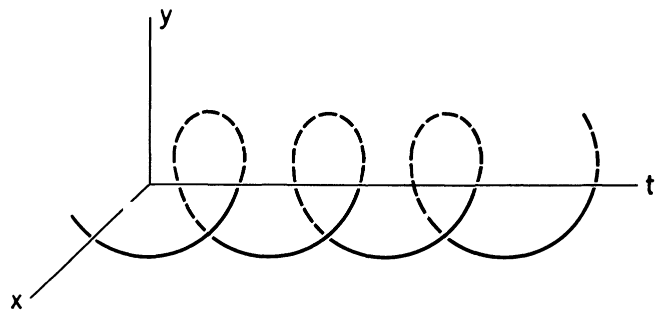


Figure 1. Graph of the solution $x=\cos t$, $y=\sin t$

Example 1. It is easily verified that $x=\cos t$, $y=\sin t$ is a solution of the system of differential equations $\dot{x}=-y$, $\dot{y}=x$. As t runs from 0 to 2π, the set of points $(\cos t,\sin t)$ trace out the unit circle $x^2+y^2=1$ in the $x-y$ plane. Hence, the unit circle $x^2+y^2=1$ is the orbit of the solution $x=\cos t$, $y=\sin t$, $0\leqslant t\leqslant 2\pi$. As t runs from 0 to ∞, the set of points $(\cos t,\sin t)$ trace out this circle infinitely often.

Example 2. It is easily verified that $x=e^{-t}\cos t, y=e^{-t}\sin t$, $-\infty<t<\infty$, is a solution of the system of differential equations $dx/dt=-x-y$, $dy/dt=x-y$. As t runs from $-\infty$ to ∞, the set of points $(e^{-t}\cos t, e^{-t}\sin t)$ trace out a spiral in the $x-y$ plane. Hence, the orbit of the solution $x=e^{-t}\cos t, y=e^{-t}\sin t$ is the spiral shown in Figure 2.

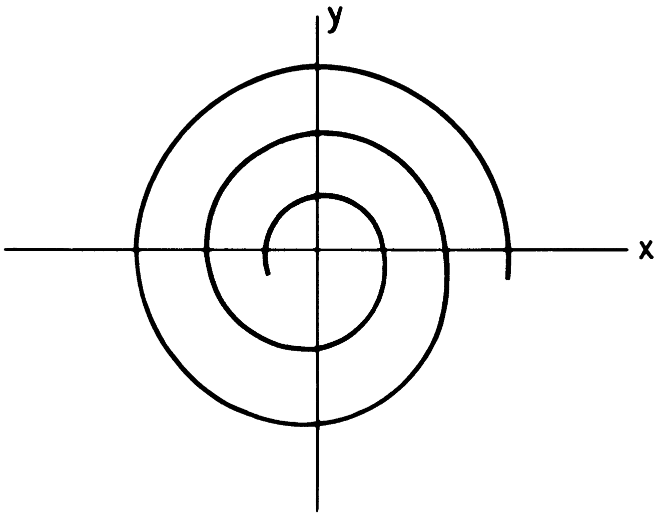


Figure 2. Orbit of $x=e^{-t}\cos t, y=e^{-t}\sin t$

Example 3. It is easily verified that $x=3t+2, y=5t+7$, $-\infty<t<\infty$ is a solution of the system of differential equations $dx/dt=3$, $dy/dt=5$. As t runs from $-\infty$ to ∞, the set of points $(3t+2,5t+7)$ trace out the straight line through the point $(2,7)$ with slope $\frac{5}{3}$. Hence, the orbit of the solution $x=3t+2, y=5t+7$ is the straight line $y=\frac{5}{3}(x-2)+7$, $-\infty<x<\infty$.

Example 4. It is easily verified that $x=3t^2+2, y=5t^2+7$, $0\leqslant t<\infty$ is a solution of the system of differential equations

$$\frac{dx}{dt}=6[(y-7)/5]^{1/2}, \qquad \frac{dy}{dt}=10[(x-2)/3]^{1/2}.$$

All of the points $(3t^2+2,5t^2+7)$ lie on the line through $(2,7)$ with slope $\frac{5}{3}$. However, x is always greater than or equal to 2, and y is always greater than or equal to 7. Hence, the orbit of the solution $x=3t^2+2, y=5t^2+7$, $0\leqslant t<\infty$, is the straight line $y=\frac{5}{3}(x-2)+7$, $2\leqslant x<\infty$.

Example 5. It is easily verified that $x=3t+2, y=5t^2+7$, $-\infty<t<\infty$, is a solution of the system of differential equations

$$\frac{dx}{dt}=y-\frac{5}{9}(x-2)^2-4, \qquad \frac{dy}{dt}=\frac{10}{3}(x-2).$$

The orbit of this solution is the set of all points $(x,y)=(3t+2,5t^2+7)$. Solving for $t=\frac{1}{3}(x-2)$, we see that $y=\frac{5}{9}(x-2)^2+7$. Hence, the orbit of the solution $x=3t+2, y=5t^2+7$ is the parabola $y=\frac{5}{9}(x-2)^2+7$, $|x|<\infty$.

One of the advantages of considering the orbit of the solution rather than the solution itself is that is is often possible to obtain the orbit of a solution without prior knowledge of the solution. Let $x=x(t), y=y(t)$ be a solution of (1). If $x'(t)$ is unequal to zero at $t=t_1$, then we can solve for $t=t(x)$ in a neighborhood of the point $x_1=x(t_1)$ (see Exercise 4). Thus, for t near t_1. the orbit of the solution $x(t), y(t)$ is the curve $y=y(t(x))$. Next, observe that

$$\frac{dy}{dx}=\frac{dy}{dt}\frac{dt}{dx}=\frac{dy/dt}{dx/dt}=\frac{g(x,y)}{f(x,y)}.$$

Thus, the orbits of the solutions $x=x(t)$, $y=y(t)$ of (1) are the solution curves of the first-order scalar equation

$$\frac{dy}{dx}=\frac{g(x,y)}{f(x,y)}. \tag{2}$$

Therefore, it is not necessary to find a solution $x(t), y(t)$ of (1) in order to compute its orbit; we need only solve the single first-order scalar differential equation (2).

Remark. From now on, we will use the phrase "the orbits of (1)" to denote the totality of orbits of solutions of (1).

Example 6. The orbits of the system of differential equations

$$\frac{dx}{dt}=y^2, \qquad \frac{dy}{dt}=x^2 \tag{3}$$

are the solution curves of the scalar equation $dy/dx=x^2/y^2$. This equation is separable, and it is easily seen that every solution is of the form $y(x)=(x^3-c)^{1/3}$, c constant. Thus, the orbits of (3) are the set of all curves $y=(x^3-c)^{1/3}$.

Example 7. The orbits of the system of differential equations

$$\frac{dx}{dt}=y(1+x^2+y^2), \qquad \frac{dy}{dt}=-2x(1+x^2+y^2) \tag{4}$$

are the solution curves of the scalar equation

$$\frac{dy}{dx}=-\frac{2x(1+x^2+y^2)}{y(1+x^2+y^2)}=-\frac{2x}{y}.$$

This equation is separable, and all solutions are of the form $\frac{1}{2}y^2+x^2=c^2$. Hence, the orbits of (4) are the families of ellipses $\frac{1}{2}y^2+x^2=c^2$.

Warning. A solution curve of (2) is an orbit of (1) only if dx/dt and dy/dt are not zero simultaneously along the solution. If a solution curve of (2) passes through an equilibrium point of (1), then the entire solution curve is not an orbit. Rather, it is the union of several distinct orbits. For example, consider the system of differential equations

$$\frac{dx}{dt}=y(1-x^2-y^2), \qquad \frac{dy}{dt}=-x(1-x^2-y^2). \tag{5}$$

The solution curves of the scalar equation

$$\frac{dy}{dx}=\frac{dy/dt}{dx/dt}=-\frac{x}{y}$$

are the family of concentric circles $x^2+y^2=c^2$. Observe, however, that every point on the unit circle $x^2+y^2=1$ is an equilibrium point of (5). Thus, the orbits of this system are the circles $x^2+y^2=c^2$, for $c\neq 1$, and all points on the unit circle $x^2+y^2=1$. Similarly, the orbits of (3) are the curves $y=(x^3-c)^{1/3}$, $c\neq 0$; the half-lines $y=x$, $x>0$, and $y=x$, $x<0$; and the point $(0,0)$.

It is not possible, in general, to explicitly solve Equation (2). Hence, we cannot, in general, find the orbits of (1). Nevertheless, it is still possible to obtain an accurate description of all orbits of (1). This is because the system of differential equations (1) sets up a *direction field* in the $x-y$ plane. That is to say, the system of differential equations (1) tells us how fast a solution moves along its orbit, and in what direction it is moving. More precisely, let $x=x(t), y=y(t)$ be a solution of (1). As t increases, the point $(x(t),y(t))$ moves along the orbit of this solution. Its velocity in the x-direction is dx/dt; its velocity in the y-direction is dy/dt; and the magnitude of its velocity is $[(dx(t)/dt)^2+(dy(t)/dt)^2]^{1/2}$. But $dx(t)/dt=f(x(t),y(t))$, and $dy(t)/dt=g(x(t),y(t))$. Hence, at each point (x,y) in the phase plane of (1) we know (i), the tangent to the orbit through (x,y) (the line through (x,y) with direction numbers $f(x,y)$, $g(x,y)$ respectively) and (ii), the speed $[f^2(x,y)+g^2(x,y)]^{1/2}$ with which the solution is traversing its orbit. As we shall see in Sections 4.8–13, this information can often be used to deduce important properties of the orbits of (1).

The notion of orbit can easily be extended to the case $n>2$. Let $\mathbf{x}=\mathbf{x}(t)$ be a solution of the vector differential equation

$$\dot{\mathbf{x}}=\mathbf{f}(\mathbf{x}), \qquad \mathbf{x}=\begin{bmatrix} x_1\\ \vdots\\ x_n\end{bmatrix}, \qquad \mathbf{f}(\mathbf{x})=\begin{bmatrix} f_1(x_1,\ldots,x_n)\\ \vdots\\ f_n(x_1,\ldots,x_n)\end{bmatrix} \tag{6}$$

on the interval $t_0\leqslant t\leqslant t_1$. As t runs from t_0 to t_1, the set of points $(x_1(t),\ldots,x_n(t))$ trace out a curve C in the n-dimensional space $x_1,x_2,\ldots,x_n$. This curve is called the orbit of the solution $\mathbf{x}=\mathbf{x}(t)$, for $t_0\leqslant t\leqslant t_1$, and the n-dimensional space $x_1,\ldots,x_n$ is called the phase-space of the solutions of (6).

EXERCISES

In each of Problems 1–3, verify that $x(t), y(t)$ is a solution of the given system of equations, and find its orbit.

1. $\dot{x}=1,\quad \dot{y}=2(1-x)\sin(1-x)^2$
$x(t)=1+t,\quad y(t)=\cos t^2$

2. $\dot{x}=e^{-x},\quad \dot{y}=e^{e^x-1}$
$x(t)=\ln(1+t),\quad y(t)=e^t$

3. $\dot{x}=1+x^2,\quad \dot{y}=(1+x^2)\sec^2 x$
$x(t)=\tan t,\quad y(t)=\tan(\tan t)$

4. Suppose that $x'(t_1)\neq 0$. Show that we can solve the equation $x=x(t)$ for $t=t(x)$ in a neighborhood of the point $x_1=x(t_1)$. *Hint*: If $x'(t_1)\neq 0$, then $x(t)$ is a strictly monotonic function of t in a neighborhood of $t=t_1$.

Find the orbits of each of the following systems.

5. $\dot{x}=y,$
$\dot{y}=-x$

6. $\dot{x}=y(1+x^2+y^2),$
$\dot{y}=-x(1+x^2+y^2)$

7. $\dot{x}=y(1+x+y),$
$\dot{y}=-x(1+x+y)$

8. $\dot{x}=y+x^2y,$
$\dot{y}=3x+xy^2$

9. $\dot{x}=xye^{-3x},$
$\dot{y}=-2xy^2$

10. $\dot{x}=4y,$
$\dot{y}=x+xy^2$

11. $\dot{x}=ax-bxy,$
$\dot{y}=cx-dxy$
(a,b,c,d positive)

12. $\dot{x}=x^2+\cos y,$
$\dot{y}=-2xy$

13. $\dot{x}=2xy,$
$\dot{y}=x^2-y^2$

14. $\dot{x}=y+\sin x,$
$\dot{y}=x-y\cos x$

4.5 Mathematical theories of war

4.5.1. L. F. Richardson's theory of conflict

In this section we construct a mathematical model which describes the relation between two nations, each determined to defend itself against a possible attack by the other. Each nation considers the possibility of attack quite real, and reasonably enough, bases its apprehensions on the readiness of the other to wage war. Our model is based on the work of Lewis Fry Richardson. It is not an attempt to make scientific statements about foreign politics or to predict the date at which the next war will break out. This, of course, is clearly impossible. Rather, it is a description of what people would do if they did not stop to think. As Richardson writes: "Why are so many nations reluctantly but steadily increasing their armaments as if they were mechanically compelled to do so? Because, I say, they follow their traditions which are fixtures and their instincts which are mechanical;

and because they have not yet made a sufficiently strenuous intellectual and moral effort to control the situation. The process described by the ensuing equations is not to be thought of as inevitable. It is what would occur if instinct and tradition were allowed to act uncontrolled."

Let $x=x(t)$ denote the war potential, or armaments, of the first nation, which we will call Jedesland, and let $y(t)$ denote the war potential of the second nation, which we will call Andersland. The rate of change of $x(t)$ depends, obviously, on the war readiness $y(t)$ of Andersland, and on the grievances that Jedesland feels towards Andersland. In the most simplistic model we represent these terms by ky and g respectively, where k and g are positive constants. These two terms cause x to increase. On the other hand, the cost of armaments has a restraining effect on dx/dt. We represent this term by $-\alpha x$, where α is a positive constant. A similar analysis holds for dy/dt. Consequently, $x=x(t), y=y(t)$ is a solution of the linear system of differential equations

$$\frac{dx}{dt}=ky-\alpha x+g, \qquad \frac{dy}{dt}=lx-\beta y+h. \tag{1}$$

Remark. The model (1) is not limited to two nations; it can also represent the relation between two alliances. For example, Andersland and Jedesland can represent the alliances of France with Russia, and Germany with Austria–Hungary during the years immediately prior to World War I.

Throughout history, there has been a constant debate on the cause of war. Over two thousand years ago, Thucydides claimed that armaments cause war. In his account of the Peloponnesian war he writes: "The real though unavowed cause I believe to have been the growth of Athenian power, which terrified the Lacedaemonians and forced them into war." Sir Edward Grey, the British Foreign Secretary during World War I agrees. He writes: "The increase of armaments that is intended in each nation to produce consciousness of strength, and a sense of security, does not produce these effects. On the contrary, it produces a consciousness of the strength of other nations and a sense of fear. The enormous growth of armaments in Europe, the sense of insecurity and fear caused by them—it was these that made war inevitable. This is the real and final account of the origin of the Great War."

On the other hand, L. S. Amery, a member of Britain's parliament during the 1930's vehemently disagrees. When the opinion of Sir Edward Grey was quoted in the House of Commons, Amery replied: "With all due respect to the memory of an eminent statesman, I believe that statement to be entirely mistaken. The armaments were only the symptoms of the conflict of ambitions and ideals, of those nationalist forces which created the War. The War was brought about because Serbia, Italy and Rumania passionately desired the incorporation in their states of territories which at that time belonged to the Austrian Empire and which the Austrian government was not prepared to abandon without a struggle. France was prepared, if the opportunity ever came, to make an effort to recover Alsace–Lorraine. It was in those facts, in those insoluble conflicts of ambitions, and not in the armaments themselves, that the cause of the War lay."

The system of equations (1) takes both conflicting theories into account. Thucydides and Sir Edward Grey would take g and h small compared to k and l, while Mr. Amery would take k and l small compared to g and h.

The system of equations (1) has several important implications. Suppose that g and h are both zero. Then, $x(t)\equiv 0, y(t)\equiv 0$ is an equilibrium solution of (1). That is, if x, y, g, and h are all made zero simultaneously, then $x(t)$ and $y(t)$ will always remain zero. This ideal condition is permanent peace by disarmament and satisfaction. It has existed since 1817 on the border between Canada and the United States, and since 1905 on the border between Norway and Sweden.

These equations further imply that mutual disarmament without satisfaction is not permanent. Assume that x and y vanish simultaneously at some time $t=t_0$. At this time, $dx/dt=g$ and $dy/dt=h$. Thus, x and y will not remain zero if g and h are positive. Instead, both nations will rearm.

Unilateral disarmament corresponds to setting $y=0$ at a certain instant of time. At this time, $dy/dt=lx+h$. This implies that y will not remain zero if either h or x is positive. Thus, unilateral disarmament is never permanent. This accords with the historical fact that Germany, whose army was reduced by the Treaty of Versailles to 100,000 men, a level far below that of several of her neighbors, insisting on rearming during the years 1933–36.

A race in armaments occurs when the "defense" terms predominate in (1). In this case,

$$\frac{dx}{dt}=ky, \qquad \frac{dy}{dt}=lx. \tag{2}$$

Every solution of (2) is of the form

$$x(t)=Ae^{\sqrt{kl}\,t}+Be^{-\sqrt{kl}\,t}, \qquad y(t)=\sqrt{\frac{l}{k}}\left[Ae^{\sqrt{kl}\,t}-Be^{-\sqrt{kl}\,t}\right].$$

Therefore, both $x(t)$ and $y(t)$ approach infinity if A is positive. This infinity can be interpreted as war.

Now, the system of equations (1) is not quite correct, since it does not take into effect the cooperation, or trade, between Andersland and Jedesland. As we see today, mutual cooperation between nations tends to decrease their fears and suspicions. We correct our model by changing the meaning of $x(t)$ and $y(t)$; we let the variables $x(t)$ and $y(t)$ stand for "threats" minus "cooperation." Specifically, we set $x=U-U_0$ and $y=V-V_0$, where U is the defense budget of Jedesland, V is the defense budget of Andersland, U_0 is the amount of goods exported by Jedesland to

Andersland and V_0 is the amount of goods exported by Andersland to Jedesland. Observe that cooperation evokes reciprocal cooperation, just as armaments provoke more armaments. In addition, nations have a tendency to reduce cooperation on account of the expense which it involves. Thus, the system of equations (1) still describes this more general state of affairs.

The system of equations (1) has a single equilibrium solution

$$x=x_0=\frac{kh+\beta g}{\alpha\beta-kl},\qquad y=y_0=\frac{lg+\alpha h}{\alpha\beta-kl} \tag{3}$$

if $\alpha\beta-kl\neq 0$. We are interested in determining whether this equilibrium solution is stable or unstable. To this end, we write (1) in the form $\dot{\mathbf{w}}=\mathbf{A}\mathbf{w}+\mathbf{f}$, where

$$\mathbf{w}(t)=\begin{pmatrix} x(t)\\ y(t)\end{pmatrix},\qquad \mathbf{f}=\begin{pmatrix} g\\ h\end{pmatrix},\quad\text{and}\quad \mathbf{A}=\begin{pmatrix} -\alpha & k\\ l & -\beta\end{pmatrix}.$$

The equilibrium solution is

$$\mathbf{w}=\mathbf{w}_0=\begin{pmatrix} x_0\\ y_0\end{pmatrix},$$

where $\mathbf{A}\mathbf{w}_0+\mathbf{f}=\mathbf{0}$. Setting $\mathbf{z}=\mathbf{w}-\mathbf{w}_0$, we obtain that

$$\dot{\mathbf{z}}=\dot{\mathbf{w}}=\mathbf{A}\mathbf{w}+\mathbf{f}=\mathbf{A}(\mathbf{z}+\mathbf{w}_0)+\mathbf{f}=\mathbf{A}\mathbf{z}+\mathbf{A}\mathbf{w}_0+\mathbf{f}=\mathbf{A}\mathbf{z}.$$

Clearly, the equilibrium solution $\mathbf{w}(t)=\mathbf{w}_0$ of $\dot{\mathbf{w}}=\mathbf{A}\mathbf{w}+\mathbf{f}$ is stable if, and only if, $\mathbf{z}=\mathbf{0}$ is a stable solution of $\dot{\mathbf{z}}=\mathbf{A}\mathbf{z}$. To determine the stability of $\mathbf{z}=\mathbf{0}$ we compute

$$p(\lambda)=\det\begin{pmatrix} -\alpha-\lambda & k\\ l & -\beta-\lambda\end{pmatrix}=\lambda^2+(\alpha+\beta)\lambda+\alpha\beta-kl.$$

The roots of $p(\lambda)$ are

$$\begin{aligned}\lambda&=\frac{-(\alpha+\beta)\pm\left[(\alpha+\beta)^2-4(\alpha\beta-kl)\right]^{1/2}}{2}\\ &=\frac{-(\alpha+\beta)\pm\left[(\alpha-\beta)^2+4kl\right]^{1/2}}{2}.\end{aligned}$$

Notice that both roots are real and unequal to zero. Moreover, both roots are negative if $\alpha\beta-kl>0$, and one root is positive if $\alpha\beta-kl<0$. Thus, $\mathbf{z}(t)\equiv\mathbf{0}$, and consequently the equilibrium solution $x(t)\equiv x_0$, $y(t)\equiv y_0$ is stable if $\alpha\beta-kl>0$ and unstable if $\alpha\beta-kl<0$.

Let us now tackle the difficult problem of estimating the coefficients α, β, k, l, g, and h. There is no way, obviously, of measuring g and h. However, it is possible to obtain reasonable estimates for α, β, k, and l. Observe that the units of these coefficients are reciprocal times. Physicists and engineers would call α^{-1} and β^{-1} relaxation times, for if y and g were identically zero, then $x(t)=e^{-\alpha(t-t_0)}x(t_0)$. This implies that $x(t_0+\alpha^{-1})=$

$x(t_0)/e$. Hence, α^{-1} is the time required for Jedesland's armaments to be reduced in the ratio 2.718 if that nation has no grievances and no other nation has any armaments. Richardson estimates α^{-1} to be the lifetime of Jedesland's parliament. Thus, $\alpha=0.2$ for Great Britain, since the lifetime of Britain's parliament is five years.

To estimate k and l we take a hypothetical case in which $g=0$ and $y=y_1$, so that $dx/dt=ky_1-\alpha x$. When $x=0$, $1/k=y_1/(dx/dt)$. Thus, $1/k$ is the time required for Jedesland to catch up to Andersland provided that (i) Andersland's armaments remain constant, (ii) there are no grievances, and (iii) the cost of armaments doesn't slow Jedesland down. Consider now the German rearmament during 1933–36. Germany started with nearly zero armaments and caught up with her neighbors in about three years. Assuming that the slowing effect of α nearly balanced the Germans' very strong grievances g, we take $k=0.3$ (year)$^{-1}$ for Germany. Further, we observe that k is obviously proportional to the amount of industry that a nation has. Thus, $k=0.15$ for a nation which has only half the industrial capacity of Germany, and $k=0.9$ for a nation which has three times the industrial capacity of Germany.

Let us now check our model against the European arms race of 1909–1914. France was allied with Russia, and Germany was allied with Austria–Hungary. Neither Italy or Britain was in a definite alliance with either party. Thus, let Jedesland represent the alliance of France with Russia, and let Andersland represent the alliance of Germany with Austria–Hungary. Since these two alliances were roughly equal in size we take $k=l$, and since each alliance was roughly three times the size of Germany, we take $k=l=0.9$. We also assume that $\alpha=\beta=0.2$. Then,

$$\frac{dx}{dt}=-\alpha x+ky+g, \qquad \frac{dy}{dt}=kx-\alpha y+h. \tag{4}$$

Equation (4) has a unique equilibrium point

$$x_0=\frac{kh+\alpha g}{\alpha^2-k^2}, \qquad y_0=\frac{kg+\alpha h}{\alpha^2-k^2}.$$

This equilibrium is unstable since

$$\alpha\beta-kl=\alpha^2-k^2=0.04-0.81=-0.77.$$

This, of course, is in agreement with the historical fact that these two alliances went to war with each other.

Now, the model we have constructed is very crude since it assumes that the grievances g and h are constant in time. This is obviously not true. The grievances g and h are not even continuous functions of time since they jump instantaneously by large amounts. (It's safe to assume, though, that g and h are relatively constant over long periods of time.) In spite of this, the system of equations (4) still provides a very accurate description of the arms race preceding World War I. To demonstrate this, we add the two equations of (4) together, to obtain that

$$\frac{d}{dt}(x+y)=(k-\alpha)(x+y)+g+h. \tag{5}$$

Recall that $x=U-U_0$ and $y=V-V_0$, where U and V are the defense budgets of the two alliances, and U_0 and V_0 are the amount of goods exported from each alliance to the other. Hence,

$$\frac{d}{dt}(U+V)=(k-\alpha)\left\{U+V-\left[U_0+V_0-\frac{g+h}{k-\alpha}-\frac{1}{k-\alpha}\frac{d}{dt}(U_0+V_0)\right]\right\}. \tag{6}$$

The defense budgets for the two alliances are set out in Table I.

Table 1. Defense budgets expressed in millions of £ sterling

	1909		1910		1911		1912		1913
France	48.6		50.9		57.1		63.2		74.7
Russia	66.7		68.5		70.7		81.8		92.0
Germany	63.1		62.0		62.5		68.2		95.4
Austria–Hungary	20.8		23.4		24.6		25.5		26.9
Total $U+V$	199.2		204.8		214.9		238.7		289.0
$\Delta(U+V)$		5.6		10.1		23.8		50.3	
$U+V$ at same date		202.0		209.8		226.8		263.8	

In Figure 1 we plot the annual increment of $U+V$ against the average of $U+V$ for the two years used in forming the increment. Notice how close these four points, denoted by $\circ$, are to the straight line

$$\Delta(U+V)=0.73(U+V-194). \tag{7}$$

Thus, foreign politics does indeed have a machine-like predictability. Equations (6) and (7) imply that

$$g+h=(k-\alpha)(U_0+V_0)-\Delta(U_0+V_0)-194$$

and $k-\alpha=0.73$. This is in excellent agreement with Richardson's estimates of 0.9 for k and 0.2 for α. Finally, observe from (7) that the total defense budgets of the two alliances will increase if $U+V$ is greater than 194 million, and will decrease otherwise. In actual fact, the defense expenditures of the two alliances was 199.2 million in 1909 while the trade between the two alliances amounted to only 171.8 million. Thus began an arms race which led eventually to World War I.

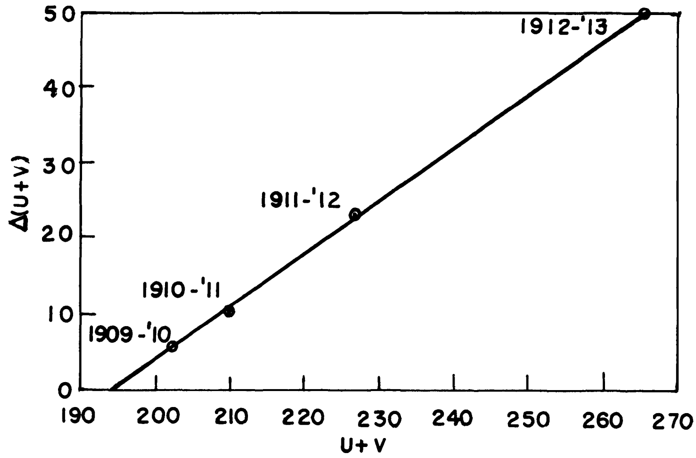


Figure 1. Graph of $\Delta(U+V)$ versus $U+V$

Reference

Richardson, L. F., "Generalized foreign politics," The British Journal of Psychology, monograph supplement #23, 1939.

Exercises

1. Suppose that what moves a government to arm is not the magnitude of other nations' armaments, but the difference between its own and theirs. Then,

$$\frac{dx}{dt} = k(y-x) - \alpha x + g, \qquad \frac{dy}{dt} = l(x-y) - \beta y + h.$$

Show that every solution of this system of equations is stable if $k_1 l_1 < (\alpha_1 + k_1)(\beta_1 + l_1)$ and unstable if $k_1 l_1 > (\alpha_1 + k_1)(\beta_1 + l_1)$.

2. Consider the case of three nations, each having the same defense coefficient k and the same restraint coefficient α. Then,

$$\frac{dx}{dt} = -\alpha x + ky + kz + g_1$$

$$\frac{dy}{dt} = kx - \alpha y + kz + g_2$$

$$\frac{dz}{dt} = kx + ky - \alpha z + g_3.$$

Setting

$$\mathbf{u} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \quad \mathbf{A} = \begin{pmatrix} -\alpha & k & k \\ k & -\alpha & k \\ k & k & -\alpha \end{pmatrix}, \quad \text{and} \quad \mathbf{g} = \begin{pmatrix} g_1 \\ g_2 \\ g_3 \end{pmatrix},$$

we see that $\dot{\mathbf{u}} = \mathbf{A}\mathbf{u} + \mathbf{g}$.

(a) Show that $p(\lambda) \equiv \det(\mathbf{A} - \lambda\mathbf{I}) = -(\alpha+\lambda)^3 + 3k^2(\alpha+\lambda) + 2k^3$.

(b) Show that $p(\lambda) = 0$ when $\lambda = -\alpha - k$. Use this information to find the remaining two roots of $p(\lambda)$.

(c) Show that every solution $\mathbf{u}(t)$ is stable if $2k<\alpha$, and unstable if $2k>\alpha$.

3. Suppose in Problem 2 that the z nation is a pacifist nation, while x and y are pugnacious nations. Then,

$$\begin{aligned} \frac{dx}{dt} &= -\alpha x + ky + kz + g_1 \\ \frac{dy}{dt} &= kx - \alpha y + kz + g_2 \\ \frac{dz}{dt} &= 0\cdot x + 0\cdot y - \alpha z + g_3. \end{aligned} \tag{*}$$

Show that every solution $x(t), y(t), z(t)$ of (*) is stable if $k<\alpha$, and unstable if $k>\alpha$.

4.5.2 *Lanchester's combat models and the battle of Iwo Jima*

During the first World War, F. W. Lanchester [4] pointed out the importance of the concentration of troops in modern combat. He constructed mathematical models from which the expected results of an engagement could be obtained. In this section we will derive two of these models, that of a conventional force versus a conventional force, and that of a conventional force versus a guerilla force. We will then solve these models, or equations, and derive "Lanchester's square law," which states that the *strength* of a combat force is proportional to the square of the number of combatants entering the engagement. Finally, we will fit one of these models, with astonishing accuracy, to the battle of Iwo Jima in World War II.

(a) *Construction of the models*

Suppose that an "x-force" and a "y-force" are engaged in combat. For simplicity, we define the strengths of these two forces as their number of combatants. (See Howes and Thrall [3] for another definition of combat strength.) Thus let $x(t)$ and $y(t)$ denote the number of combatants of the x and y forces, where t is measured in days from the start of the combat. Clearly, the rate of change of each of these quantities equals its *reinforcement rate* minus its *operational loss rate* minus its *combat loss rate*.

The operational loss rate. The operational loss rate of a combat force is its loss rate due to non-combat mishaps; i.e., desertions, diseases, etc. Lanchester proposed that the operational loss rate of a combat force is proportional to its strength. However, this does not appear to be very realistic. For example, the desertion rate in a combat force depends on a host of psychological and other intangible factors which are difficult even to describe, let alone quantify. We will take the easy way out here and consider only those engagements in which the operational loss rates are negligible.

The combat loss rate. Suppose that the x-force is a conventional force which operates in the open, comparatively speaking, and that every member of this force is within "kill range" of the enemy y. We also assume that as soon as the conventional force suffers a loss, fire is concentrated on the remaining combatants. Under these "ideal" conditions, the combat loss rate of a conventional force x equals $ay(t)$, for some positive constant a. This constant is called the *combat effectiveness coefficient* of the y-force.

The situation is very different if x is a guerilla force, invisible to its opponent y and occupying a region R. The y-force fires into R but cannot know when a kill has been made. It is certainly plausible that the combat loss rate for a guerilla force x should be proportional to $x(t)$, for the larger $x(t)$, the greater the probability that an opponent's shot will kill. On the other hand, the combat loss rate for x is also proportional to $y(t)$, for the larger y, the greater the number of x-casualties. Thus, the combat loss rate for a guerilla force x equals $cx(t)y(t)$, where the constant c is called the *combat effectiveness coefficient* of the opponent y.

The reinforcement rate. The reinforcement rate of a combat force is the rate at which new combatants enter (or are withdrawn from) the battle. We denote the reinforcement rates of the x- and y-forces by $f(t)$ and $g(t)$ respectively.

Under the assumptions listed above, we can now write down the following two Lanchestrian models for conventional–guerilla combat.

$$\text{Conventional combat:}\quad \begin{cases} \dfrac{dx}{dt} = -ay+f(t) \\[2mm] \dfrac{dy}{dt} = -bx+g(t) \end{cases} \tag{1a}$$

$$\begin{array}{c}\text{Conventional–guerilla combat:}\\ (x=\text{guerilla})\end{array}\quad \begin{cases} \dfrac{dx}{dt} = -cxy+f(t) \\[2mm] \dfrac{dy}{dt} = -dx+g(t) \end{cases} \tag{1b}$$

The system of equations (1a) is a linear system and can be solved explicitly once a, b, $f(t)$, and $g(t)$ are known. On the other hand, the system of equations (1b) is nonlinear, and its solution is much more difficult. (Indeed, it can only be obtained with the aid of a digital computer.)

It is very instructive to consider the special case where the reinforcement rates are zero. This situation occurs when the two forces are "isolated." In this case (1a) and (1b) reduce to the simpler systems

$$\frac{dx}{dt} = -ay, \qquad \frac{dy}{dt} = -bx \tag{2a}$$

and

$$\frac{dx}{dt} = -cxy, \qquad \frac{dy}{dt} = -dx. \tag{2b}$$

Conventional combat: *The square law*. The orbits of system (2a) are the solution curves of the equation

$$\frac{dy}{dx}=\frac{bx}{ay} \quad \text{or} \quad ay\frac{dy}{dx}=bx.$$

Integrating this equation gives

$$ay^2-bx^2=ay_0^2-bx_0^2=K. \tag{3}$$

The curves (3) define a family of hyperbolas in the x–y plane and we have indicated their graphs in Figure 1. The arrowheads on the curves indicate the direction of changing strengths as time passes.

Let us adopt the criterion that one force wins the battle if the other force vanishes first. Then, y wins if $K>0$ since the x-force has been annihilated by the time $y(t)$ has decreased to $\sqrt{K/a}$. Similarly, x wins if $K<0$.

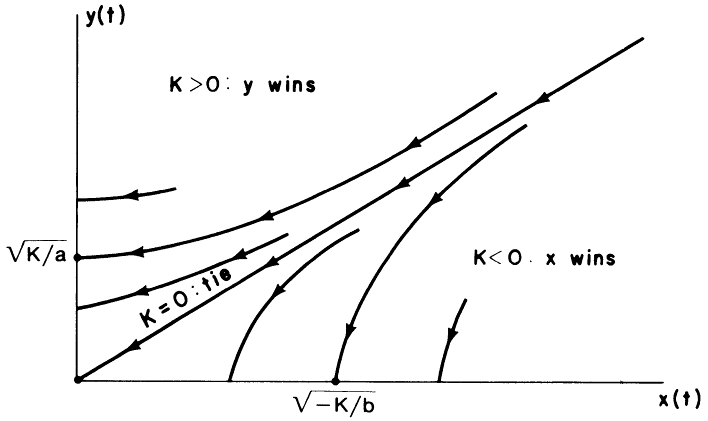


Figure 1. The hyperbolas defined by (3)

Remark 1. Equation (3) is often referred to as "Lanchester's square law," and the system (2a) is often called the square law model, since the strengths of the opposing forces appear *quadratically* in (3). This terminology is rather anomolous since the system (2a) is actually a linear system.

Remark 2. The y-force always seeks to establish a setting in which $K>0$. That is to say, the y-force wants the inequality

$$ay_0^2>bx_0^2$$

to hold. This can be accomplished by increasing a; i.e. by using stronger and more accurate weapons, or by increasing the initial force y_0. Notice though that a doubling of a results in a doubling of ay_0^2 while a doubling of y_0 results in a *four-fold* increase of ay_0^2. This is the essence of Lanchester's square law of conventional combat.

Conventional–guerilla combat. The orbits of system (2b) are the solution curves of the equation

$$\frac{dy}{dx} = \frac{dx}{cxy} = \frac{d}{cy}. \tag{4}$$

Multiplying both sides of (4) by cy and integrating gives

$$cy^2 - 2dx = cy_0^2 - 2dx_0 = M. \tag{5}$$

The curves (5) define a family of parabolas in the x–y plane, and we have indicated their graphs in Figure 2. The y-force wins if $M>0$, since the x-force has been annihilated by the time $y(t)$ has decreased to $\sqrt{M/c}$. Similarly, x wins if $M<0$.

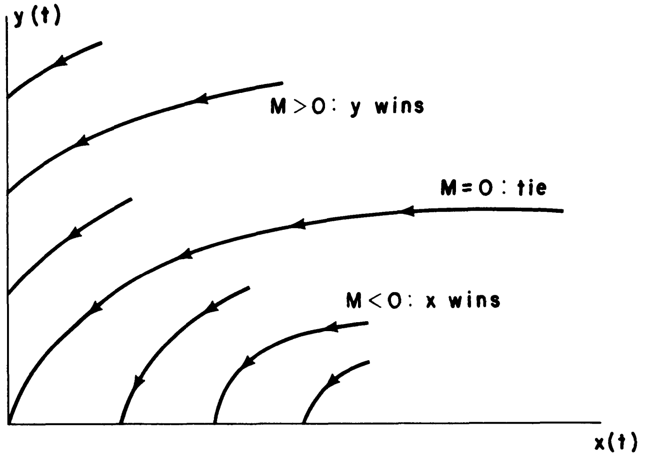


Figure 2. The parabolas defined by (5)

Remark. It is usually impossible to determine, a priori, the numerical value of the combat coefficients a, b, c, and d. Thus, it would appear that Lanchester's combat models have little or no applicability to real-life engagements. However, this is not so. As we shall soon see, it is often possible to determine suitable values of a and b (or c and d) using data from the battle itself. Once these values are established for one engagement, they are known for all other engagements which are fought under similar conditions.

(b) *The battle of Iwo Jima*

One of the fiercest battles of World War II was fought on the island of Iwo Jima, 660 miles south of Tokyo. Our forces coveted Iwo Jima as a bomber base close to the Japanese mainland, while the Japanese needed the island as a base for fighter planes attacking our aircraft on their way to bombing missions over Tokyo and other major Japanese cities. The American invasion of Iwo Jima began on February 19, 1945, and the fighting was intense throughout the month long combat. Both sides suffered heavy casualties (see Table 1). The Japanese had been ordered to fight to the last man, and this is exactly what they did. The island was declared "secure" by the American forces on the 28th day of the battle, and all active combat ceased on the 36th day. (The last two Japanese holdouts surrendered in 1951!)

Table 1. Casualties at Iwo Jima

	Total United States casualties at Iwo Jima			
	Killed, missing or died of wounds	Wounded	Combat Fatigue	Total
Marines	5,931	17,272	2,648	25,851
Navy units:				
Ships and air units	633	1,158		1,791
Medical corpsmen	195	529		724
Seabees	51	218		269
Doctors and dentists	2	12		14
Army units in battle	9	28		37
Grand totals	6,821	19,217	2,648	28,686
	Japanese casualties at Iwo Jima			
Defense forces (Estimated)		Prisoners		Killed
21,000		Marine 216		20,000
		Army 867		
		Total 1,083		

(Newcomb [6], page 296)

The following data is available to us from the battle of Iwo Jima.

1. *Reinforcement rates.* During the conflict Japanese troops were neither withdrawn nor reinforced. The Americans, on the other hand, landed 54,000 troops on the first day of the battle, none on the second, 6,000 on the third, none on the fourth and fifth, 13,000 on the sixth day, and none thereafter. There were no American troops on Iwo Jima prior to the start of the engagement.

2. *Combat losses.* Captain Clifford Morehouse of the United States Marine Corps (see Morehouse [5]) kept a daily count of all American combat losses. Unfortunately, no such records are available for the Japanese forces. Most probably, the casualty lists kept by General Kuribayashi (commander of the Japanese forces on Iwo Jima) were destroyed in the battle itself, while whatever records were kept in Tokyo were consumed in the fire bombings of the remaining five months of the war. However, we can infer from Table 1 that approximately 21,500 Japanese forces were on Iwo Jima at the start of the battle. (Actually, Newcomb arrived at the fig-

ure of 21,000 for the Japanese forces, but this is a little low since he apparently did not include some of the living and dead found in the caves in the final days.)

3. *Operational losses.* The operational losses on both sides were negligible.

Now, let $x(t)$ and $y(t)$ denote respectively, the active American and Japanese forces on Iwo Jima t days after the battle began. The data above suggests the following Lanchestrian model for the battle of Iwo Jima:

$$\begin{aligned}\frac{dx}{dt} &= -ay+f(t)\\ \frac{dy}{dt} &= -bx\end{aligned} \tag{6}$$

where a and b are the combat effectiveness coefficients of the Japanese and American forces, respectively, and

$$f(t)=\begin{cases} 54{,}000 & 0\leqslant t<1\\ 0 & 1\leqslant t<2\\ 6{,}000 & 2\leqslant t<3\\ 0 & 3\leqslant t<5\\ 13{,}000 & 5\leqslant t<6\\ 0 & t\geqslant 6\end{cases}$$

Using the method of variation of parameters developed in Section 3.12 or the method of elimination in Section 2.14, it is easily seen that the solution of (6) which satisfies $x(0)=0$, $y(0)=y_0=21{,}500$ is given by

$$x(t)=-\sqrt{\frac{a}{b}}\; y_0\cosh\sqrt{ab}\; t+\int_0^t \cosh\sqrt{ab}\;(t-s)f(s)\,ds \tag{7a}$$

and

$$y(t)=y_0\cosh\sqrt{ab}\; t-\sqrt{\frac{b}{a}}\;\int_0^t \sinh\sqrt{ab}\;(t-s)f(s)\,ds \tag{7b}$$

where

$$\cosh x\equiv(e^x+e^{-x})/2 \quad\text{and}\quad \sinh x\equiv(e^x-e^{-x})/2.$$

The problem before us now is this: Do there exist constants a and b so that (7a) yields a good fit to the data compiled by Morehouse? This is an extremely important question. An affirmative answer would indicate that Lanchestrian models do indeed describe real life battles, while a negative answer would shed a dubious light on much of Lanchester's work.

As we mentioned previously, it is extremely difficult to compute the combat effectiveness coefficients a and b of two opposing forces. However, it is often possible to determine suitable values of a and b once the data for the battle is known, and such is the case for the battle of Iwo Jima.

The calculation of a and b. Integrating the second equation of (6) between 0 and s gives

$$y(s)-y_0=-b\int_0^s x(t)\,dt$$

so that

$$b=\frac{y_0-y(s)}{\int_0^s x(t)\,dt}. \tag{8}$$

In particular, setting $s=36$ gives

$$b=\frac{y_0-y(36)}{\int_0^{36} x(t)\,dt}=\frac{21{,}500}{\int_0^{36} x(t)\,dt}. \tag{9}$$

Now the integral on the right-hand side of (9) can be approximated by the Riemann sum

$$\int_0^{36} x(t)\,dt \cong \sum_{i=1}^{36} x(i)$$

and for $x(i)$ we enter the number of effective American troops on the ith day of the battle. Using the data available from Morehouse, we compute for b the value

$$b=\frac{21{,}500}{2{,}037{,}000}=0.0106. \tag{10}$$

Remark. We would prefer to set $s=28$ in (8) since that was the day the island was declared secure, and the fighting was only sporadic after this day. However, we don't know $y(28)$. Thus, we are forced here to take $s=36$.

Next, we integrate the first equation of (6) between $t=0$ and $t=28$ and obtain that

$$\begin{aligned} x(28) &= -a\int_0^{28} y(t)\,dt+\int_0^{28} f(t)\,dt \\ &= -a\int_0^{28} y(t)\,dt+73{,}000. \end{aligned}$$

There were 52,735 effective American troops on the 28th day of the battle. Thus

$$a=\frac{73{,}000-52{,}735}{\int_0^{28} y(t)\,dt}=\frac{20{,}265}{\int_0^{28} y(t)\,dt}. \tag{11}$$

Finally, we approximate the integral on the right-hand side of (11) by the Riemann sum

$$\int_0^{28} y(t)\,dt \cong \sum_{j=1}^{28} y(j)$$

and we approximate $y(j)$ by

$$\begin{aligned} y(j) &= y_0 - b\int_0^j x(t)\,dt \\ &\cong 21{,}500 - b\sum_{i=i}^{j} x(i). \end{aligned}$$

Again, we replace $x(i)$ by the number of effective American troops on the ith day of the battle. The result of this calculation is (see Engel [2])

$$a = \frac{20{,}265}{372{,}500} = 0.0544. \tag{12}$$

Figure 3 below compares the actual American troop strength with the values predicted by Equation (7a) (with $a = 0.0544$ and $b = 0.0106$). The fit is remarkably good. Thus, it appears that a Lanchestrian model does indeed describe real life engagements.

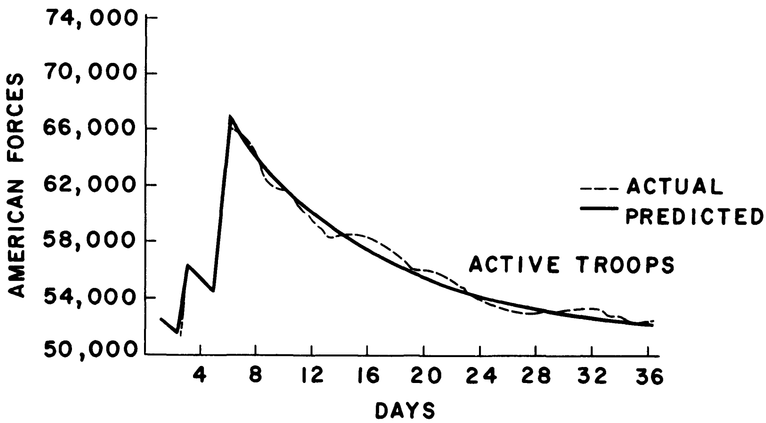


Figure 3. Comparison of actual troop strength with predicted troop strength

Remark. The figures we have used for American reinforcements include *all* the personnel put ashore, both combat troops and support troops. Thus the numbers a and b that we have computed should be interpreted as the *average* effectiveness per man ashore.

References

1. Coleman, C. S., Combat Models, MAA Workshop on Modules in Applied Math, Cornell University, Aug. 1976.
2. Engel, J. H., A verification of Lanchester's law, *Operations Research*, **2**, (1954), 163–171.
3. Howes, D. R., and Thrall, R. M., A theory of ideal linear weights for heterogeneous combat forces, *Naval Research Logistics Quarterly*, vol. 20, 1973, pp. 645–659.
4. Lanchester, F. W., *Aircraft in Warfare, the Dawn of the Fourth Arm*. Tiptree, Constable and Co., Ltd., 1916.
5. Morehouse, C. P., *The Iwo Jima Operation*, USMCR, Historical Division, Hdqr. USMC, 1946.
6. Newcomb, R. F., *Iwo Jima*. New York: Holt, Rinehart, and Winston, 1965.

Exercises

1. Derive Equations (7a) and (7b).

2. The system of equations

$$\begin{aligned} \dot{x} &= -ay \\ \dot{y} &= -by - cxy \end{aligned} \tag{13}$$

is a Lanchestrian model for conventional–guerilla combat, in which the operational loss rate of the guerilla force y is proportional to $y(t)$.
(a) Find the orbits of (13).
(b) Who wins the battle?

3. The system of equations

$$\begin{aligned} \dot{x} &= -ay \\ \dot{y} &= -bx - cxy \end{aligned} \tag{14}$$

is a Lanchestrian model for conventional–guerilla combat, in which the operational loss rate of the guerilla force y is proportional to the strength of the conventional force x. Find the orbits of (14).

4. The system of equations

$$\begin{aligned} \dot{x} &= -cxy \\ \dot{y} &= -dxy \end{aligned} \tag{15}$$

is a Lanchestrian model for guerilla–guerilla combat in which the operational loss rates are negligible.
(a) Find the orbits of (15).
(b) Who wins the battle?

5. The system of equations

$$\begin{aligned} \dot{x} &= -ay - cxy \\ \dot{y} &= -bx - dxy \end{aligned} \tag{16}$$

is a Lanchestrian model for guerilla–guerilla combat in which the operational loss rate of each force is proportional to the strength of its opponent. Find the orbits of (16).

6. The system of equations

$$\begin{aligned}\dot{x} &= -ax - cxy \\ \dot{y} &= -by - dxy\end{aligned} \tag{17}$$

is a Lanchestrian model for guerilla–guerilla combat in which the operational loss rate of each force is proportional to its strength.

(a) Find the orbits of (17).

(b) Show that the x and y axes are both orbits of (17).

(c) Using the fact (to be proved in Section 4.6) that two orbits of (17) cannot intersect, show that there is no clear-cut winner in this battle. *Hint*: Show that $x(t)$ and $y(t)$ can never become zero in finite time. (Using lemmas 1 and 2 of Section 4.8, it is easy to show that both $x(t)$ and $y(t)$ approach zero as $t \to \infty$.)

4.6 Qualitative properties of orbits

In this section we will derive two very important properties of the solutions and orbits of the system of differential equations

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}), \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \qquad \mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(x_1,\dots,x_n) \\ \vdots \\ f_n(x_1,\dots,x_n) \end{bmatrix}. \tag{1}$$

The first property deals with the existence and uniqueness of orbits, and the second property deals with the existence of periodic solutions of (1). We begin with the following existence–uniqueness theorem for the solutions of (1).

Theorem 3. *Let each of the functions $f_1(x_1,\dots,x_n),\dots,f_n(x_1,\dots,x_n)$ have continuous partial derivatives with respect to $x_1,\dots,x_n$. Then, the initial-value problem $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$, $\mathbf{x}(t_0) = \mathbf{x}^0$ has one, and only one solution $\mathbf{x} = \mathbf{x}(t)$, for every $\mathbf{x}^0$ in R^n.*

We prove Theorem 3 in exactly the same manner as we proved the existence–uniqueness theorem for the scalar differential equation $\dot{x} = f(t,x)$. Indeed, the proof given in Section 1.10 carries over here word for word. We need only interpret the quantity $|\mathbf{x}(t) - \mathbf{y}(t)|$, where $\mathbf{x}(t)$ and $\mathbf{y}(t)$ are vector-valued functions, as the length of the vector $\mathbf{x}(t) - \mathbf{y}(t)$. That is to say, if we interpret $|\mathbf{x}(t) - \mathbf{y}(t)|$ as

$$|\mathbf{x}(t) - \mathbf{y}(t)| = \max\{|x_1(t) - y_1(t)|,\dots,|x_n(t) - y_n(t)|\},$$

then the proof of Theorem 2, Section 1.10 is valid even for vector-valued functions $\mathbf{f}(t,\mathbf{x})$ (see Exercises 13–14).

Next, we require the following simple but extremely useful lemma.

Lemma 1. *If $\mathbf{x} = \boldsymbol{\phi}(t)$ is a solution of (1), then $\mathbf{x} = \boldsymbol{\phi}(t + c)$ is again a solution of (1).*

The meaning of Lemma 1 is the following. Let $\mathbf{x}=\boldsymbol{\phi}(t)$ be a solution of (1) and let us replace every t in the formula for $\boldsymbol{\phi}(t)$ by $t+c$. In this manner we obtain a new function $\hat{\mathbf{x}}(t)=\boldsymbol{\phi}(t+c)$. Lemma 1 states that $\hat{\mathbf{x}}(t)$ is again a solution of (1). For example, $x_1=\tan t$, $x_2=\sec^2 t$ is a solution of the system of differential equations $dx_1/dt=x_2$, $dx_2/dt=2x_1x_2$. Hence, $x_1=\tan(t+c)$, $x_2=\sec^2(t+c)$ is again a solution, for any constant c.

Proof of Lemma 1. If $\mathbf{x}=\boldsymbol{\phi}(t)$ is a solution of (1), then $d\boldsymbol{\phi}(t)/dt=\mathbf{f}(\boldsymbol{\phi}(t))$; that is, the two functions $d\boldsymbol{\phi}(t)/dt$ and $\mathbf{h}(t)\equiv\mathbf{f}(\boldsymbol{\phi}(t))$ agree at every single time. Fix a time t and a constant c. Since $d\boldsymbol{\phi}/dt$ and $\mathbf{h}$ agree at every time, they must agree at time $t+c$. Hence,

$$\frac{d\boldsymbol{\phi}}{dt}(t+c)=\mathbf{h}(t+c)=\mathbf{f}(\boldsymbol{\phi}(t+c)).$$

But, $d\boldsymbol{\phi}/dt$ evaluated at $t+c$ equals the derivative of $\hat{\mathbf{x}}(t)\equiv\boldsymbol{\phi}(t+c)$, evaluated at t. Therefore,

$$\frac{d}{dt}\boldsymbol{\phi}(t+c)=\mathbf{f}(\boldsymbol{\phi}(t+c)). \qquad \square$$

Remark 1. Lemma 1 can be verified explicitly for the linear equation $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$. Every solution $\mathbf{x}(t)$ of this equation is of the form $\mathbf{x}(t)=e^{\mathbf{A}t}\mathbf{v}$, for some constant vector $\mathbf{v}$. Hence,

$$\mathbf{x}(t+c)=e^{\mathbf{A}(t+c)}\mathbf{v}=e^{\mathbf{A}t}e^{\mathbf{A}c}\mathbf{v}$$

since $(\mathbf{A}t)\mathbf{A}c=\mathbf{A}c(\mathbf{A}t)$ for all values of t and c. Therefore, $\mathbf{x}(t+c)$ is again a solution of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$ since it is of the form $e^{\mathbf{A}t}$ times the constant vector $e^{\mathbf{A}c}\mathbf{v}$.

Remark 2. Lemma 1 is not true if the function $\mathbf{f}$ in (1) depends explicitly on t. To see this, suppose that $\mathbf{x}=\boldsymbol{\phi}(t)$ is a solution of the nonautonomous differential equation $\dot{\mathbf{x}}=\mathbf{f}(t,\mathbf{x})$. Then, $\dot{\boldsymbol{\phi}}(t+c)=\mathbf{f}(t+c,\boldsymbol{\phi}(t+c))$. Consequently, the function $\mathbf{x}=\boldsymbol{\phi}(t+c)$ satisfies the differential equation

$$\dot{\mathbf{x}}=\mathbf{f}(t+c,\mathbf{x}),$$

and this equation is different from (1) if $\mathbf{f}$ depends explicitly on t.

We are now in a position to derive the following extremely important properties of the solutions and orbits of (1).

Property 1. (Existence and uniqueness of orbits.) Let each of the functions $f_1(x_1,\ldots,x_n),\ldots,f_n(x_1,\ldots,x_n)$ have continuous partial derivatives with respect to $x_1,\ldots,x_n$. Then, there exists one, and only one, orbit through every point $\mathbf{x}^0$ in R^n. In particular, if the orbits of two solutions $\mathbf{x}=\boldsymbol{\phi}(t)$ and $\mathbf{x}=\boldsymbol{\psi}(t)$ of (1) have one point in common, then they must be identical.

Property 2. Let $\mathbf{x}=\boldsymbol{\phi}(t)$ be a solution of (1). If $\boldsymbol{\phi}(t_0+T)=\boldsymbol{\phi}(t_0)$ for some t_0 and $T>0$, then $\boldsymbol{\phi}(t+T)$ is identically equal to $\boldsymbol{\phi}(t)$. In other words, if a solution $\mathbf{x}(t)$ of (1) returns to its starting value after a time $T>0$, then it must be periodic, with period T (i.e. it must repeat itself over every time interval of length T.)

PROOF OF PROPERTY 1. Let $\mathbf{x}^0$ be any point in the n-dimensional phase space $x_1,\dots,x_n$, and let $\mathbf{x}=\boldsymbol{\phi}(t)$ be the solution of the initial-value problem $\dot{\mathbf{x}}=\mathbf{f}(\mathbf{x})$, $\mathbf{x}(0)=\mathbf{x}^0$. The orbit of this solution obviously passes through $\mathbf{x}^0$. Hence, there exists at least one orbit through every point $\mathbf{x}^0$. Now, suppose that the orbit of another solution $\mathbf{x}=\boldsymbol{\psi}(t)$ also passes through $\mathbf{x}^0$. This means that there exists $t_0(\neq 0)$ such that $\boldsymbol{\psi}(t_0)=\mathbf{x}^0$. By Lemma 1,

$$\mathbf{x}=\boldsymbol{\psi}(t+t_0)$$

is also a solution of (1). Observe that $\boldsymbol{\psi}(t+t_0)$ and $\boldsymbol{\phi}(t)$ have the same value at $t=0$. Hence, by Theorem 3, $\boldsymbol{\psi}(t+t_0)$ equals $\boldsymbol{\phi}(t)$ for all time t. This implies that the orbits of $\boldsymbol{\phi}(t)$ and $\boldsymbol{\psi}(t)$ are identical. To wit, if $\boldsymbol{\xi}$ is a point on the orbit of $\boldsymbol{\phi}(t)$; that is, $\boldsymbol{\xi}=\boldsymbol{\phi}(t_1)$ for some t_1, then $\boldsymbol{\xi}$ is also on the orbit of $\boldsymbol{\psi}(t)$, since $\boldsymbol{\xi}=\boldsymbol{\phi}(t_1)=\boldsymbol{\psi}(t_1+t_0)$. Conversely, if $\boldsymbol{\xi}$ is a point on the orbit of $\boldsymbol{\psi}(t)$; that is, there exists t_2 such that $\boldsymbol{\psi}(t_2)=\boldsymbol{\xi}$, then $\boldsymbol{\xi}$ is also on the orbit of $\boldsymbol{\phi}(t)$ since $\boldsymbol{\xi}=\boldsymbol{\psi}(t_2)=\boldsymbol{\phi}(t_2-t_0)$. □

PROOF OF PROPERTY 2. Let $\mathbf{x}=\boldsymbol{\phi}(t)$ be a solution of (1) and suppose that $\boldsymbol{\phi}(t_0+T)=\boldsymbol{\phi}(t_0)$ for some numbers t_0 and T. Then, the function $\boldsymbol{\psi}(t)=\boldsymbol{\phi}(t+T)$ is also a solution of (1) which agrees with $\boldsymbol{\phi}(t)$ at time $t=t_0$. By Theorem 3, therefore, $\boldsymbol{\psi}(t)=\boldsymbol{\phi}(t+T)$ is identically equal to $\boldsymbol{\phi}(t)$. □

Property 2 is extremely useful in applications, especially when $n=2$. Let $x=x(t)$, $y=y(t)$ be a periodic solution of the system of differential equations

$$\frac{dx}{dt}=f(x,y),\qquad \frac{dy}{dt}=g(x,y). \tag{2}$$

If $x(t+T)=x(t)$ and $y(t+T)=y(t)$, then the orbit of this solution is a closed curve C in the x–y plane. In every time interval $t_0\leqslant t\leqslant t_0+T$, the solution moves once around C. Conversely, suppose that the orbit of a solution $x=x(t)$, $y=y(t)$ of (2) is a closed curve containing no equilibrium points of (2). Then, the solution $x=x(t)$, $y=y(t)$ is periodic. To prove this, recall that a solution $x=x(t)$, $y=y(t)$ of (2) moves along its orbit with velocity $[f^2(x,y)+g^2(x,y)]^{1/2}$. If its orbit C is a closed curve containing no equilibrium points of (2), then the function $[f^2(x,y)+g^2(x,y)]^{1/2}$ has a positive minimum for (x,y) on C. Hence, the orbit of $x=x(t)$, $y=y(t)$ must return to its starting point $x_0=x(t_0)$, $y_0=y(t_0)$ in some finite time T. But this implies that $x(t+T)=x(t)$ and $y(t+T)=y(t)$ for all t.

Example 1. Prove that every solution $z(t)$ of the second-order differential equation $(d^2z/dt^2)+z+z^3=0$ is periodic.

PROOF. We convert this second-order equation into a system of two first-order equations by setting $x=z$, $y=dz/dt$. Then,

$$\frac{dx}{dt}=y, \qquad \frac{dy}{dt}=-x-x^3. \tag{3}$$

The orbits of (3) are the solution curves

$$\frac{y^2}{2}+\frac{x^2}{2}+\frac{x^4}{4}=c^2 \tag{4}$$

of the scalar equation $dy/dx=-(x+x^3)/y$. Equation (4) defines a closed curve in the x–y plane (see Exercise 7). Moreover, the only equilibrium point of (3) is $x=0, y=0$. Consequently, every solution $x=z(t), y=z'(t)$ of (3) is a periodic function of time. Notice, however, that we cannot compute the period of any particular solution. □

Example 2. Prove that every solution of the system of differential equations

$$\frac{dx}{dt}=ye^{1+x^2+y^2}, \qquad \frac{dy}{dt}=-xe^{1+x^2+y^2} \tag{5}$$

is periodic.
Solution. The orbits of (5) are the solution curves $x^2+y^2=c^2$ of the first-order scalar equation $dy/dx=-x/y$. Moreover, $x=0$, $y=0$ is the only equilibrium point of (5). Consequently, every solution $x=x(t), y=y(t)$ of (5) is a periodic function of time.

EXERCISES

1. Show that all solutions $x(t), y(t)$ of

$$\frac{dx}{dt}=x^2+y\sin x, \qquad \frac{dy}{dt}=-1+xy+\cos y$$

which start in the first quadrant ($x>0$, $y>0$) must remain there for all time (both backwards and forwards).

2. Show that all solutions $x(t), y(t)$ of

$$\frac{dx}{dt}=y(e^x-1), \qquad \frac{dy}{dt}=x+e^y$$

which start in the right half plane ($x>0$) must remain there for all time.

3. Show that all solutions $x(t), y(t)$ of

$$\frac{dx}{dt}=1+x^2+y^2, \qquad \frac{dy}{dt}=xy+\tan y$$

which start in the upper half plane ($y>0$) must remain there for all time.

4. Show that all solutions $x(t), y(t)$ of

$$\frac{dx}{dt}=-1-y+x^2, \qquad \frac{dy}{dt}=x+xy$$

which start inside the unit circle $x^2+y^2=1$ must remain there for all time. *Hint*: Compute $d(x^2+y^2)/dt$.

5. Let $x(t), y(t)$ be a solution of

$$\frac{dx}{dt}=y+x^2, \qquad \frac{dy}{dt}=x+y^2$$

with $x(t_0)\neq y(t_0)$. Show that $x(t)$ can never equal $y(t)$.

6. Can a figure 8 ever be an orbit of

$$\frac{dx}{dt}=f(x,y), \qquad \frac{dy}{dt}=g(x,y)$$

where f and g have continuous partial derivatives with respect to x and y?

7. Show that the curve $y^2+x^2+x^4/2=2c^2$ is closed. *Hint*: Show that there exist two points $y=0$, $x=\pm\alpha$ which lie on this curve.

Prove that all solutions of the following second-order equations are periodic.

8. $\dfrac{d^2z}{dt^2}+z^3=0$

9. $\dfrac{d^2z}{dt^2}+z+z^5=0$

10. $\dfrac{d^2z}{dt^2}+e^{z^2}=1$

11. $\dfrac{d^2z}{dt^2}+\dfrac{z}{1+z^2}=0$

12. Show that all solutions $z(t)$ of

$$\frac{d^2z}{dt^2}+z-2z^3=0$$

are periodic if $\dot{z}^2(0)+z^2(0)-z^4(0)<\frac{1}{4}$, and unbounded if

$$\dot{z}^2(0)+z^2(0)-z^4(0)>\tfrac{1}{4}.$$

13. (a) Let

$$L=n\times \max_{i,j=1,\dots,n} |\partial f_i/\partial x_j|, \quad \text{for} \quad |\mathbf{x}-\mathbf{x}^0|\leqslant b.$$

Show that $|\mathbf{f}(\mathbf{x})-\mathbf{f}(\mathbf{y})|\leqslant L|\mathbf{x}-\mathbf{y}|$ if $|\mathbf{x}-\mathbf{x}^0|\leqslant b$ and $|\mathbf{y}-\mathbf{x}^0|\leqslant b$.

(b) Let $M=\max|\mathbf{f}(\mathbf{x})|$ for $|\mathbf{x}-\mathbf{x}^0|\leqslant b$. Show that the Picard iterates

$$\mathbf{x}_{j+1}(t)=\mathbf{x}^0+\int_{t_0}^{t}\mathbf{f}(\mathbf{x}_j(s))\,ds, \qquad \mathbf{x}_0(t)=\mathbf{x}^0$$

converge to a solution $\mathbf{x}(t)$ of the initial-value problem $\dot{\mathbf{x}}=\mathbf{f}(\mathbf{x})$, $\mathbf{x}(t_0)=\mathbf{x}^0$ on the interval $|t-t_0|\leqslant b/M$. *Hint*: The proof of Theorem 2, Section 1.10 carries over here word for word.

14. Compute the Picard iterates $\mathbf{x}_j(t)$ of the initial-value problem $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$, $\mathbf{x}(0)=\mathbf{x}^0$, and verify that they approach $e^{\mathbf{A}t}\mathbf{x}^0$ as j approaches infinity.

4.7 Phase portraits of linear systems

In this section we present a complete picture of all orbits of the linear differential equation

$$\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}, \qquad \mathbf{x}=\begin{pmatrix} x_1\\ x_2\end{pmatrix}, \qquad \mathbf{A}=\begin{pmatrix} a & b\\ c & d\end{pmatrix}. \tag{1}$$

This picture is called a phase portrait, and it depends almost completely on the eigenvalues of the matrix **A**. It also changes drastically as the eigenvalues of **A** change sign or become imaginary.

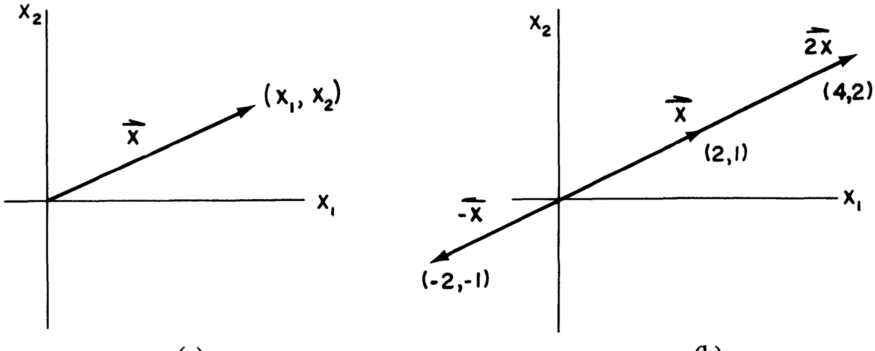
When analyzing Equation (1), it is often helpful to visualize a vector

$$\mathbf{x}=\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

in R^2 as a direction, or directed line segment, in the plane. Let

$$\mathbf{x}=\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

be a vector in R^2 and draw the directed line segment $\vec{x}$ from the point $(0,0)$ to the point (x_1,x_2), as in Figure 1a. This directed line segment is parallel to the line through $(0,0)$ with direction numbers x_1,x_2 respectively. If we visualize the vector $\mathbf{x}$ as being this directed line segment $\vec{x}$, then we see that the vectors $\mathbf{x}$ and $c\mathbf{x}$ are parallel if c is positive, and antiparallel if c is negative. We can also give a nice geometric interpretation of vector addition. Let $\mathbf{x}$ and $\mathbf{y}$ be two vectors in R^2. Draw the directed line segment $\vec{x}$, and place the vector $\vec{y}$ at the tip of $\vec{x}$. The vector $\vec{x}+\vec{y}$ is then the composi-



(a) (b)

Figure 1

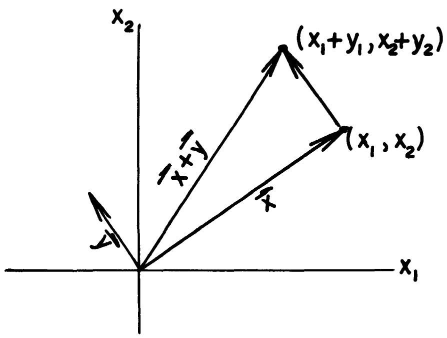


Figure 2

tion of these two directed line segments (see Figure 2). This construction is known as the parallelogram law of vector addition.

We are now in a position to derive the phase portraits of (1). Let λ_1 and λ_2 denote the two eigenvalues of $\mathbf{A}$. We distinguish the following cases.

1. $\lambda_2<\lambda_1<0$. Let $\mathbf{v}^1$ and $\mathbf{v}^2$ be eigenvectors of $\mathbf{A}$ with eigenvalues λ_1 and λ_2 respectively. In the x_1-x_2 plane we draw the four half-lines l_1, l_1', l_2, and l_2', as shown in Figure 3. The rays l_1 and l_2 are parallel to $\mathbf{v}^1$ and $\mathbf{v}^2$, while the rays l_1' and l_2' are parallel to $-\mathbf{v}^1$ and $-\mathbf{v}^2$. Observe first that $\mathbf{x}(t)=ce^{\lambda_1 t}\mathbf{v}^1$ is a solution of (1) for any constant c. This solution is always proportional to $\mathbf{v}^1$, and the constant of proportionality, $ce^{\lambda_1 t}$, runs from $\pm\infty$ to 0, depending as to whether c is positive or negative. Hence, the orbit of this solution is the half-line l_1 for $c>0$, and the half-line l_1' for $c<0$. Similarly, the orbit of the solution $\mathbf{x}(t)=ce^{\lambda_2 t}\mathbf{v}^2$ is the half-line l_2 for $c>0$, and the half-line l_2' for $c<0$. The arrows on these four lines in Figure 3 indicate in what direction $\mathbf{x}(t)$ moves along its orbit.

Next, recall that every solution $\mathbf{x}(t)$ of (1) can be written in the form

$$\mathbf{x}(t)=c_1e^{\lambda_1 t}\mathbf{v}^1+c_2e^{\lambda_2 t}\mathbf{v}^2 \tag{2}$$

for some choice of constants c_1 and c_2. Obviously, every solution $\mathbf{x}(t)$ of (1) approaches $\binom{0}{0}$ as t approaches infinity. Hence, every orbit of (1) approaches the origin $x_1=x_2=0$ as t approaches infinity. We can make an even stronger statement by observing that $e^{\lambda_2 t}\mathbf{v}^2$ is very small compared to $e^{\lambda_1 t}\mathbf{v}^1$ when t is very large. Therefore, $\mathbf{x}(t)$, for $c_1\neq 0$, comes closer and closer to $c_1e^{\lambda_1 t}\mathbf{v}^1$ as t approaches infinity. This implies that the tangent to the orbit of $\mathbf{x}(t)$ approaches l_1 if c_1 is positive, and l_1' if c_1 is negative. Thus,

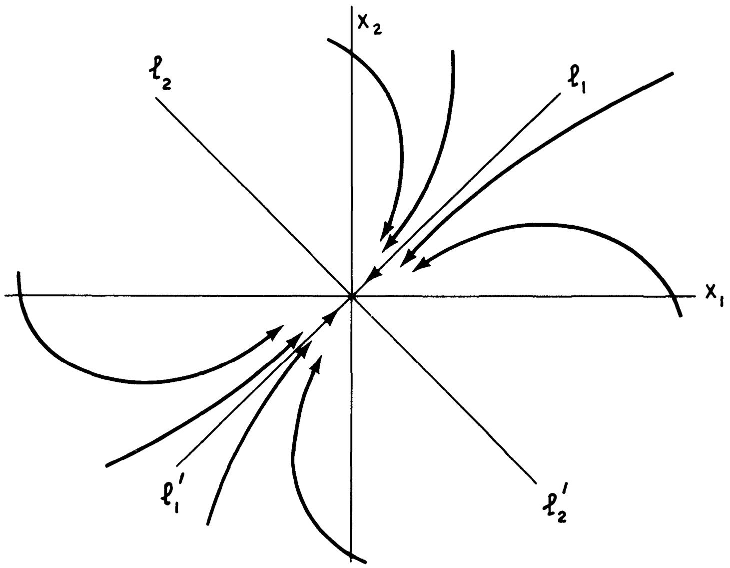


Figure 3. Phase portrait of a stable node

the phase portrait of (1) has the form described in Figure 3. The distinguishing feature of this phase portrait is that every orbit, with the exception of a single line, approaches the origin in a fixed direction (if we consider the directions $\mathbf{v}^1$ and $-\mathbf{v}^1$ equivalent). In this case we say that the equilibrium solution $\mathbf{x}=\mathbf{0}$ of (1) is a stable node.

Remark. The orbit of every solution $\mathbf{x}(t)$ of (1) approaches the origin $x_1 = x_2 = 0$ as t approaches infinity. However, this point does not belong to the orbit of any nontrivial solution $\mathbf{x}(t)$.

1′. $0<\lambda_1<\lambda_2$. The phase portrait of (1) in this case is exactly the same as Figure 3, except that the direction of the arrows is reversed. Hence, the equilibrium solution $\mathbf{x}(t)=\mathbf{0}$ of (1) is an unstable node if both eigenvalues of $\mathbf{A}$ are positive.

2. $\lambda_1=\lambda_2<0$. In this case, the phase portrait of (1) depends on whether $\mathbf{A}$ has one or two linearly independent eigenvectors. (a) Suppose that $\mathbf{A}$ has two linearly independent eigenvectors $\mathbf{v}^1$ and $\mathbf{v}^2$ with eigenvalue $\lambda<0$. In this case, every solution $\mathbf{x}(t)$ of (1) can be written in the form

$$\mathbf{x}(t)=c_1e^{\lambda t}\mathbf{v}^1+c_2e^{\lambda t}\mathbf{v}^2=e^{\lambda t}\left(c_1\mathbf{v}^1+c_2\mathbf{v}^2\right) \tag{2}$$

for some choice of constants c_1 and c_2. Now, the vector $e^{\lambda t}(c_1\mathbf{v}^1+c_2\mathbf{v}^2)$ is parallel to $c_1\mathbf{v}^1+c_2\mathbf{v}^2$ for all t. Hence, the orbit of every solution $\mathbf{x}(t)$ of (1) is a half-line. Moreover, the set of vectors $\{c_1\mathbf{v}^1+c_2\mathbf{v}^2\}$, for all choices of c_1 and c_2, cover every direction in the x_1-x_2 plane, since $\mathbf{v}^1$ and $\mathbf{v}^2$ are linearly independent. Hence, the phase portrait of (1) has the form described in Figure 4a. (b) Suppose that $\mathbf{A}$ has only one linearly independent eigenvector $\mathbf{v}$, with eigenvalue λ. In this case, $\mathbf{x}^1(t)=e^{\lambda t}\mathbf{v}$ is one solution of (1). To find a second solution of (1) which is independent of $\mathbf{x}^1$, we observe that $(\mathbf{A}-\lambda\mathbf{I})^2\mathbf{u}=\mathbf{0}$ for every vector $\mathbf{u}$. Hence,

$$\mathbf{x}(t)=e^{\mathbf{A}t}\mathbf{u}=e^{\lambda t}e^{(\mathbf{A}-\lambda\mathbf{I})t}\mathbf{u}=e^{\lambda t}\left[\mathbf{u}+t(\mathbf{A}-\lambda\mathbf{I})\mathbf{u}\right] \tag{3}$$

is a solution of (1) for any choice of $\mathbf{u}$. Equation (3) can be simplified by observing that $(\mathbf{A}-\lambda\mathbf{I})\mathbf{u}$ must be a multiple k of $\mathbf{v}$. This follows immediately from the equation $(\mathbf{A}-\lambda\mathbf{I})[(\mathbf{A}-\lambda\mathbf{I})\mathbf{u}]=\mathbf{0}$, and the fact that $\mathbf{A}$ has only one linearly independent eigenvector $\mathbf{v}$. Choosing $\mathbf{u}$ independent of $\mathbf{v}$, we see that every solution $\mathbf{x}(t)$ of (1) can be written in the form

$$\mathbf{x}(t)=c_1e^{\lambda t}\mathbf{v}+c_2e^{\lambda t}(\mathbf{u}+kt\mathbf{v})=e^{\lambda t}(c_1\mathbf{v}+c_2\mathbf{u}+c_2kt\mathbf{v}), \tag{4}$$

for some choice of constants c_1 and c_2. Obviously, every solution $\mathbf{x}(t)$ of (1) approaches $\binom{0}{0}$ as t approaches infinity. In addition, observe that $c_1\mathbf{v}+c_2\mathbf{u}$ is very small compared to $c_2kt\mathbf{v}$ if c_2 is unequal to zero and t is very large. Hence, the tangent to the orbit of $\mathbf{x}(t)$ approaches $\pm\mathbf{v}$ (depending on the sign of c_2) as t approaches infinity, and the phase portrait of (1) has the form described in Figure 4b.

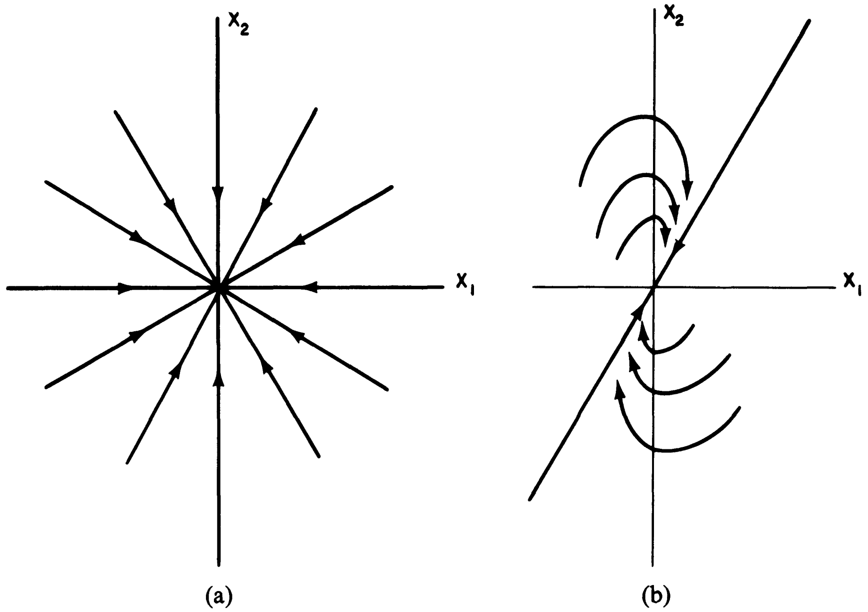


Figure 4

2′. $\lambda_1=\lambda_2>0$. The phase portraits of (1) in the cases (2a)′ and (2b)′ are exactly the same as Figures 4a and 4b, except that the direction of the arrows is reversed.

3. $\lambda_1<0<\lambda_2$. Let $\mathbf{v}^1$ and $\mathbf{v}^2$ be eigenvectors of $\mathbf{A}$ with eigenvalues λ_1 and λ_2 respectively. In the x_1-x_2 plane we draw the four half-lines l_1, l_1', l_2, and l_2'; the half-lines l_1 and l_2 are parallel to $\mathbf{v}^1$ and $\mathbf{v}^2$, while the half-lines l_1' and l_2' are parallel to $-\mathbf{v}^1$ and $-\mathbf{v}^2$. Observe first that every solution $\mathbf{x}(t)$ of (1) is of the form

$$\mathbf{x}(t)=c_1e^{\lambda_1t}\mathbf{v}^1+c_2e^{\lambda_2t}\mathbf{v}^2 \tag{5}$$

for some choice of constants c_1 and c_2. The orbit of the solution $\mathbf{x}(t)=c_1e^{\lambda_1t}\mathbf{v}^1$ is l_1 for $c_1>0$ and l_1' for $c_1<0$, while the orbit of the solution $\mathbf{x}(t)=c_2e^{\lambda_2t}\mathbf{v}^2$ is l_2 for $c_2>0$ and l_2' for $c_2<0$. Note, too, the direction of the arrows on l_1, l_1', l_2, and l_2'; the solution $\mathbf{x}(t)=c_1e^{\lambda_1t}\mathbf{v}^1$ approaches $\binom{0}{0}$ as t approaches infinity, whereas the solution $\mathbf{x}(t)=c_2e^{\lambda_2t}\mathbf{v}^2$ becomes unbounded (for $c_2\neq0$) as t approaches infinity. Next, observe that $e^{\lambda_1t}\mathbf{v}^1$ is very small compared to $e^{\lambda_2t}\mathbf{v}^2$ when t is very large. Hence, every solution $\mathbf{x}(t)$ of (1) with $c_2\neq0$ becomes unbounded as t approaches infinity, and its orbit approaches either l_2 or l_2'. Finally, observe that $e^{\lambda_2t}\mathbf{v}^2$ is very small compared to $e^{\lambda_1t}\mathbf{v}^1$ when t is very large negative. Hence, the orbit of any solution $\mathbf{x}(t)$ of (1), with $c_1\neq0$, approaches either l_1 or l_1' as t approaches minus infinity. Consequently, the phase portrait of (1) has the form described in Figure 5. This phase portrait resembles a "saddle" near $x_1=x_2=$

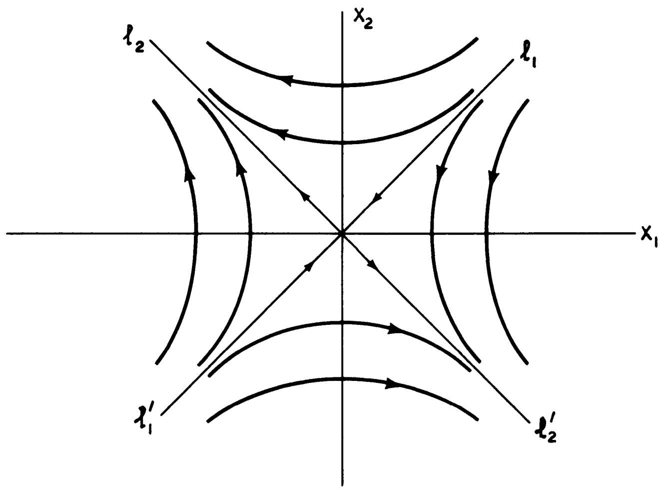


Figure 5. Phase portrait of a saddle point

0. For this reason, we say that the equilibrium solution $\mathbf{x}(t)=\mathbf{0}$ of (1) is a saddle point if the eigenvalues of $\mathbf{A}$ have opposite sign.

4. $\lambda_1=\alpha+i\beta, \lambda_2=\alpha-i\beta, \beta\neq 0$. Our first step in deriving the phase portrait of (1) is to find the general solution of (1). Let $\mathbf{z}=\mathbf{u}+i\mathbf{v}$ be an eigenvector of $\mathbf{A}$ with eigenvalue $\alpha+i\beta$. Then,

$$\begin{aligned}\mathbf{x}(t)&=e^{(\alpha+i\beta)t}(\mathbf{u}+i\mathbf{v})=e^{\alpha t}(\cos\beta t+i\sin\beta t)(\mathbf{u}+i\mathbf{v})\\&=e^{\alpha t}[\mathbf{u}\cos\beta t-\mathbf{v}\sin\beta t]+ie^{\alpha t}[\mathbf{u}\sin\beta t+\mathbf{v}\cos\beta t]\end{aligned}$$

is a complex-valued solution of (1). Therefore,

$$\mathbf{x}^1(t)=e^{\alpha t}[\mathbf{u}\cos\beta t-\mathbf{v}\sin\beta t]$$

and

$$\mathbf{x}^2(t)=e^{\alpha t}[\mathbf{u}\sin\beta t+\mathbf{v}\cos\beta t]$$

are two real-valued linearly independent solutions of (1), and every solution $\mathbf{x}(t)$ of (1) is of the form $\mathbf{x}(t)=c_1\mathbf{x}^1(t)+c_2\mathbf{x}^2(t)$. This expression can be written in the form (see Exercise 15)

$$\mathbf{x}(t)=e^{\alpha t}\begin{pmatrix}R_1\cos(\beta t-\delta_1)\\R_2\cos(\beta t-\delta_2)\end{pmatrix}\tag{6}$$

for some choice of constants $R_1\geq 0$, $R_2\geq 0$, δ_1, and δ_2. We distinguish the following cases.

(a) $\alpha=0$: Observe that both

$$x_1(t)=R_1\cos(\beta t-\delta_1)\quad\text{and}\quad x_2(t)=R_2\cos(\beta t-\delta_2)$$

are periodic functions of time with period $2\pi/\beta$. The function $x_1(t)$ varies between $-R_1$ and $+R_1$, while $x_2(t)$ varies between $-R_2$ and $+R_2$. Consequently, the orbit of any solution $\mathbf{x}(t)$ of (1) is a closed curve surrounding the origin $x_1 = x_2 = 0$, and the phase portrait of (1) has the form described in Figure 6a. For this reason, we say that the equilibrium solution $\mathbf{x}(t) = \mathbf{0}$ of (1) is a center when the eigenvalues of $\mathbf{A}$ are pure imaginary.

The direction of the arrows in Figure 6a must be determined from the differential equation (1). The simplest way of doing this is to check the sign of $\dot{x}_2$ when $x_2 = 0$. If $\dot{x}_2$ is greater than zero for $x_2 = 0$ and $x_1 > 0$, then all solutions $\mathbf{x}(t)$ of (1) move in the counterclockwise direction; if $\dot{x}_2$ is less than zero for $x_2 = 0$ and $x_1 > 0$, then all solutions $\mathbf{x}(t)$ of (1) move in the clockwise direction.

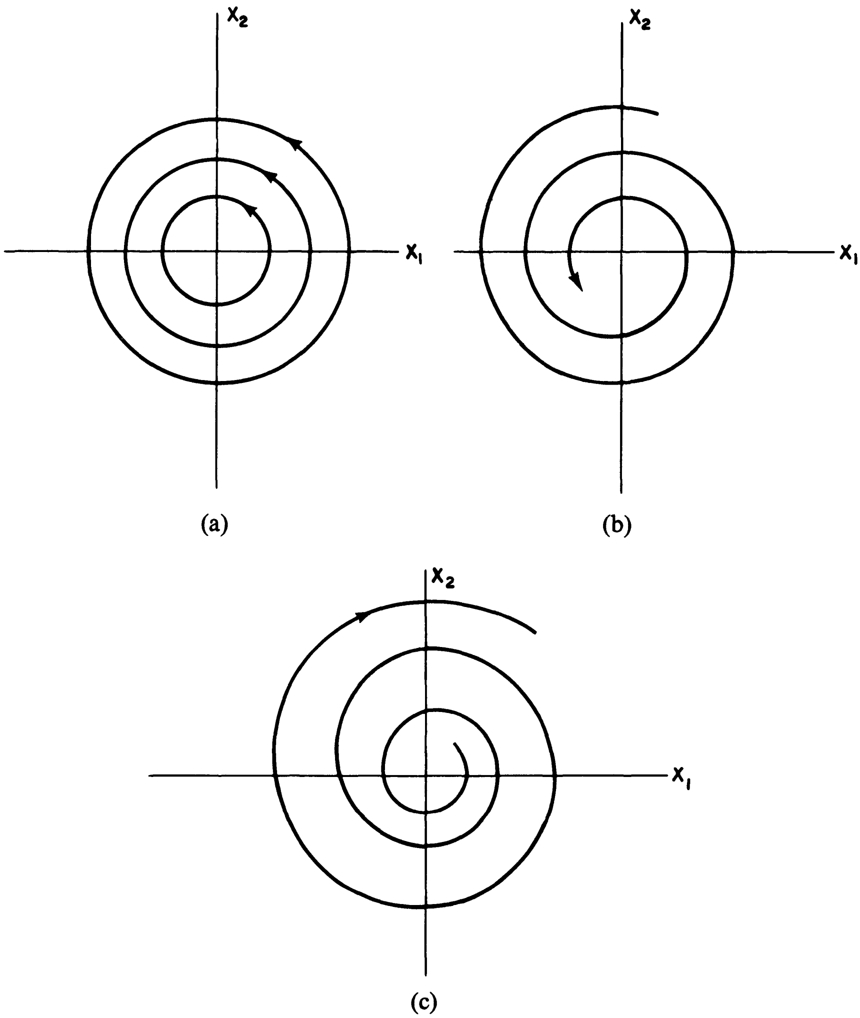


Figure 6. (a) $\alpha = 0$; (b) $\alpha < 0$; (c) $\alpha > 0$

(b) $\alpha<0$: In this case, the effect of the factor $e^{\alpha t}$ in Equation (6) is to change the simple closed curves of Figure 6a into the spirals of Figure 6b. This is because the point $\mathbf{x}(2\pi/\beta)=e^{2\pi\alpha/\beta}\mathbf{x}(0)$ is closer to the origin than $\mathbf{x}(0)$. Again, the direction of the arrows in Figure 6b must be determined directly from the differential equation (1). In this case, we say that the equilibrium solution $\mathbf{x}(t)=\mathbf{0}$ of (1) is a stable focus.

(c) $\alpha>0$: In this case, all orbits of (1) spiral away from the origin as t approaches infinity (see Figure 6c), and the equilibrium solution $\mathbf{x}(t)=\mathbf{0}$ of (1) is called an unstable focus.

Finally, we mention that the phase portraits of nonlinear systems, in the neighborhood of an equilibrium point, are often very similar to the phase portraits of linear systems. More precisely, let $\mathbf{x}=\mathbf{x}^0$ be an equilibrium solution of the nonlinear equation $\dot{\mathbf{x}}=\mathbf{f}(\mathbf{x})$, and set $\mathbf{u}=\mathbf{x}-\mathbf{x}^0$. Then, (see Section 4.3) we can write the differential equation $\dot{\mathbf{x}}=\mathbf{f}(\mathbf{x})$ in the form

$$\dot{\mathbf{u}}=\mathbf{A}\mathbf{u}+\mathbf{g}(\mathbf{u}) \tag{7}$$

where $\mathbf{A}$ is a constant matrix and $\mathbf{g}(\mathbf{u})$ is very small compared to $\mathbf{u}$. We state without proof the following theorem.

Theorem 4. *Suppose that* $\mathbf{u}=\mathbf{0}$ *is either a node, saddle, or focus point of the differential equation* $\dot{\mathbf{u}}=\mathbf{A}\mathbf{u}$. *Then, the phase portrait of the differential equation* $\dot{\mathbf{x}}=\mathbf{f}(\mathbf{x})$, *in a neighborhood of* $\mathbf{x}=\mathbf{x}^0$, *has one of the forms described in Figures* 3, 5, *and* 6 (b *and* c), *depending as to whether* $\mathbf{u}=\mathbf{0}$ *is a node, saddle, or focus.*

Example 1. Draw the phase portrait of the linear equation

$$\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}=\begin{pmatrix}-2 & -1\\ 4 & -7\end{pmatrix}\mathbf{x}. \tag{8}$$

Solution. It is easily verified that

$$\mathbf{v}^1=\begin{pmatrix}1\\1\end{pmatrix} \quad\text{and}\quad \mathbf{v}^2=\begin{pmatrix}1\\4\end{pmatrix}$$

are eigenvectors of $\mathbf{A}$ with eigenvalues -3 and -6, respectively. Therefore, $\mathbf{x}=\mathbf{0}$ is a stable node of (8), and the phase portrait of (8) has the form described in Figure 7. The half-line l_1 makes an angle of 45° with the x_1 axis, while the half-line l_2 makes an angle of θ degrees with the x_1-axis, where $\tan\theta=4$.

Example 2. Draw the phase portrait of the linear equation

$$\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}=\begin{pmatrix}1 & -3\\ -3 & 1\end{pmatrix}\mathbf{x}. \tag{9}$$

Solution. It is easily verified that

$$\mathbf{v}^1=\begin{pmatrix}1\\1\end{pmatrix} \quad\text{and}\quad \mathbf{v}^2=\begin{pmatrix}-1\\1\end{pmatrix}$$

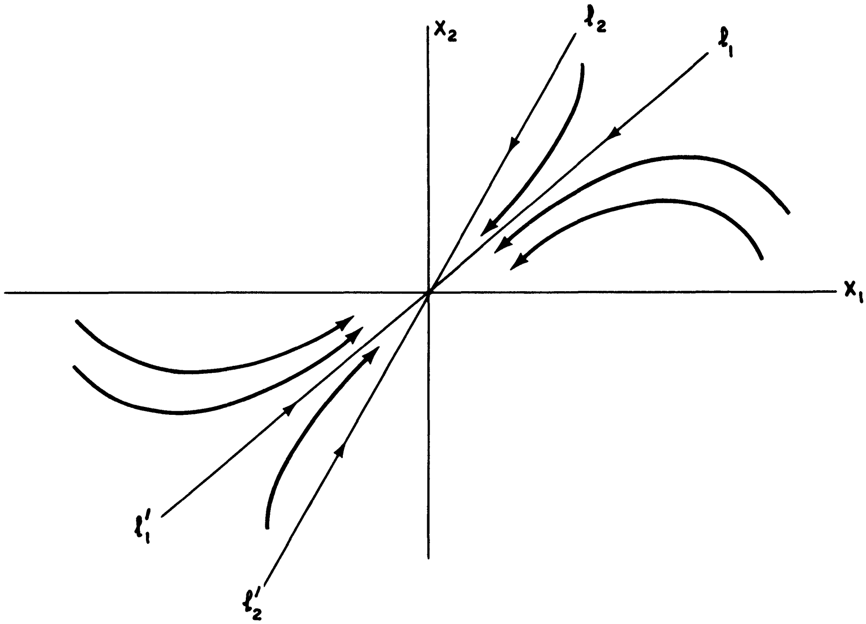


Figure 7. Phase portrait of (8)

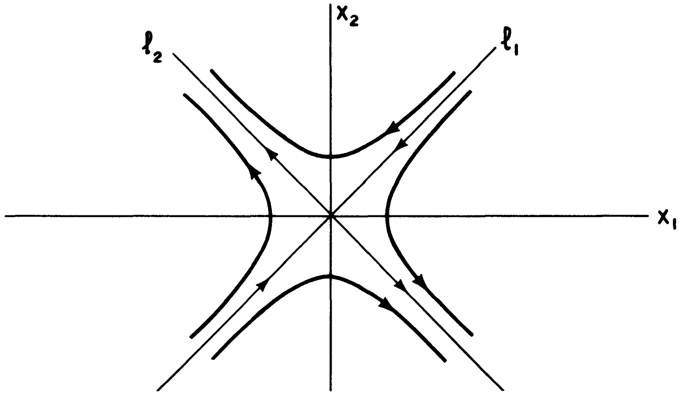


Figure 8. Phase portrait of (9)

are eigenvectors of **A** with eigenvalues -2 and 4, respectively. Therefore, $\mathbf{x}=\mathbf{0}$ is a saddle point of (9), and its phase portrait has the form described in Figure 8. The half-line l_1 makes an angle of 45° with the x_1-axis, and the half-line l_2 is at right angles to l_1.

Example 3. Draw the phase portrait of the linear equation

$$\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}=\begin{pmatrix} -1 & 1 \\ -1 & -1 \end{pmatrix}\mathbf{x}. \tag{10}$$

Solution. The eigenvalues of $\mathbf{A}$ are $-1 \pm i$. Hence, $\mathbf{x}=\mathbf{0}$ is a stable focus of (10) and every nontrivial orbit of (10) spirals into the origin as t approaches infinity. To determine the direction of rotation of the spiral, we observe that $\dot{x}_2 = -x_1$ when $x_2=0$. Thus, $\dot{x}_2$ is negative for $x_1>0$ and $x_2=0$. Consequently, all nontrivial orbits of (10) spiral into the origin in the clockwise direction, as shown in Figure 9.

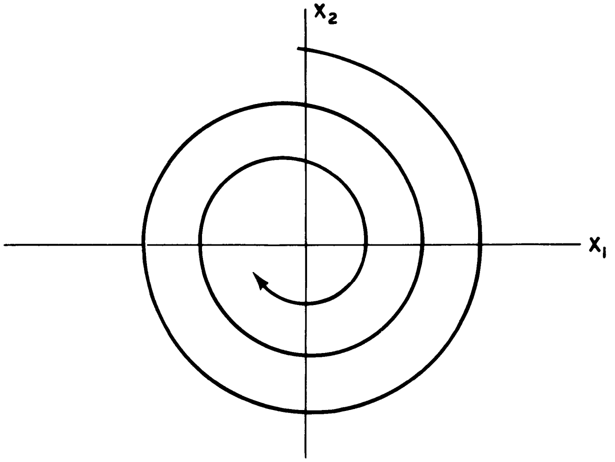


Figure 9. Phase portrait of (10)

Exercises

Draw the phase portraits of each of the following systems of differential equations.

1. $\dot{\mathbf{x}}=\begin{pmatrix} -5 & 1 \\ 1 & -5 \end{pmatrix}\mathbf{x}$ **2.** $\dot{\mathbf{x}}=\begin{pmatrix} 0 & -1 \\ 8 & -6 \end{pmatrix}\mathbf{x}$ **3.** $\dot{\mathbf{x}}=\begin{pmatrix} 4 & -1 \\ -2 & 5 \end{pmatrix}\mathbf{x}$

4. $\dot{\mathbf{x}}=\begin{pmatrix} -4 & -1 \\ 1 & -6 \end{pmatrix}\mathbf{x}$ **5.** $\dot{\mathbf{x}}=\begin{pmatrix} 1 & -4 \\ -8 & 4 \end{pmatrix}\mathbf{x}$ **6.** $\dot{\mathbf{x}}=\begin{pmatrix} 3 & -1 \\ 5 & -3 \end{pmatrix}\mathbf{x}$

7. $\dot{\mathbf{x}}=\begin{pmatrix} 0 & 2 \\ -2 & -1 \end{pmatrix}\mathbf{x}$ **8.** $\dot{\mathbf{x}}=\begin{pmatrix} 1 & -1 \\ 5 & -3 \end{pmatrix}\mathbf{x}$ **9.** $\dot{\mathbf{x}}=\begin{pmatrix} 2 & 1 \\ -5 & -2 \end{pmatrix}\mathbf{x}$

10. Show that every orbit of

$$\dot{\mathbf{x}}=\begin{pmatrix} 0 & 4 \\ -9 & 0 \end{pmatrix}\mathbf{x}$$

is an ellipse.

11. The equation of motion of a spring-mass system with damping (see Section 2.6) is $m\ddot{z}+c\dot{z}+kz=0$, where m, c, and k are positive numbers. Convert this equation to a system of first-order equations for $x=z$, $y=\dot{z}$, and draw the phase portrait of this system. Distinguish the overdamped, critically damped, and underdamped cases.

12. Suppose that a 2×2 matrix $\mathbf{A}$ has 2 linearly independent eigenvectors with eigenvalue λ. Show that $\mathbf{A}=\lambda\mathbf{I}$.

13. This problem illustrates Theorem 4. Consider the system

$$\frac{dx}{dt}=y, \qquad \frac{dy}{dt}=x+2x^3. \tag{*}$$

(a) Show that the equilibrium solution $x=0, y=0$ of the linearized system $\dot{x}=y, \dot{y}=x$ is a saddle, and draw the phase portrait of the linearized system.
(b) Find the orbits of (*), and then draw its phase portrait.
(c) Show that there are exactly two orbits of (*) (one for $x>0$ and one for $x<0$) on which $x\to 0, y\to 0$ as $t\to\infty$. Similarly, there are exactly two orbits of (*) on which $x\to 0, y\to 0$ as $t\to -\infty$. Thus, observe that the phase portraits of (*) and the linearized system look the same near the origin.

14. Verify Equation (6). *Hint*: The expression $a\cos\omega t+b\sin\omega t$ can always be written in the form $R\cos(\omega t-\delta)$ for suitable choices of R and δ.

4.8 Long time behavior of solutions; the Poincaré–Bendixson Theorem

We consider now the problem of determining the long time behavior of all solutions of the differential equation

$$\dot{\mathbf{x}}=\mathbf{f}(\mathbf{x}), \qquad \mathbf{x}=\begin{bmatrix} x_1\\ \vdots \\ x_n\end{bmatrix}, \qquad \mathbf{f}(\mathbf{x})=\begin{bmatrix} f_1(x_1,\dots,x_n)\\ \vdots \\ f_n(x_1,\dots,x_n)\end{bmatrix}. \tag{1}$$

This problem has been solved completely in the special case that $\mathbf{f}(\mathbf{x})=\mathbf{A}\mathbf{x}$. As we have seen in Sections 4.2 and 4.7, all solutions $\mathbf{x}(t)$ of $\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}$ must exhibit one of the following four types of behavior: (i) $\mathbf{x}(t)$ is constant in time; (ii) $\mathbf{x}(t)$ is a periodic function of time; (iii) $\mathbf{x}(t)$ is unbounded as t approaches infinity; and (iv) $\mathbf{x}(t)$ approaches an equilibrium point as t approaches infinity.

A partial solution to this problem, in the case of nonlinear $\mathbf{f}(\mathbf{x})$, was given in Section 4.3. In that section we provided sufficient conditions that every solution $\mathbf{x}(t)$ of (1), whose initial value $\mathbf{x}(0)$ is sufficiently close to an equilibrium point $\boldsymbol{\xi}$, must ultimately approach $\boldsymbol{\xi}$ as t approaches infinity. In many applications it is often possible to go much further and prove that every physically (biologically) realistic solution approaches a single equilibrium point as time evolves. In this context, the following two lemmas play an extremely important role.

Lemma 1. *Let $g(t)$ be a monotonic increasing (decreasing) function of time for $t\geqslant t_0$, with $g(t)\leqslant c(\geqslant c)$ for some constant c. Then, $g(t)$ has a limit as t approaches infinity.*

PROOF. Suppose that $g(t)$ is monotonic increasing for $t\geqslant t_0$, and $g(t)$ is bounded from above. Let l be the least upper bound of g; that is, l is the smallest number which is not exceeded by the values of $g(t)$, for $t\geqslant t_0$. This number must be the limit of $g(t)$ as t approaches infinity. To prove this, let $\varepsilon > 0$ be given, and observe that there exists a time $t_\varepsilon \geqslant t_0$ such that $l - g(t_\varepsilon) < \varepsilon$. (If no such time t_ε exists, then l is not the least upper bound of g.) Since $g(t)$ is monotonic, we see that $l - g(t) < \varepsilon$ for $t \geqslant t_\varepsilon$. This shows that $l = \lim_{t\to\infty} g(t)$. □

Lemma 2. *Suppose that a solution* $\mathbf{x}(t)$ *of* (1) *approaches a vector* $\boldsymbol{\xi}$ *as* t *approaches infinity. Then,* $\boldsymbol{\xi}$ *is an equilibrium point of* (1).

PROOF. Suppose that $\mathbf{x}(t)$ approaches $\boldsymbol{\xi}$ as t approaches infinity. Then, $x_j(t)$ approaches ξ_j, where ξ_j is the jth component of $\boldsymbol{\xi}$. This implies that $|x_j(t_1) - x_j(t_2)|$ approaches zero as both t_1 and t_2 approach infinity, since

$$\begin{aligned} |x_j(t_1) - x_j(t_2)| &= |(x_j(t_1) - \xi_j) + (\xi_j - x_j(t_2))| \\ &\leqslant |x_j(t_1) - \xi_j| + |x_j(t_2) - \xi_j|. \end{aligned}$$

In particular, let $t_1 = t$ and $t_2 = t_1 + h$, for some fixed positive number h. Then, $|x_j(t+h) - x_j(t)|$ approaches zero as t approaches infinity. But

$$x_j(t+h) - x_j(t) = h\frac{dx_j(\tau)}{dt} = hf_j(x_1(\tau), \dots, x_n(\tau)),$$

where τ is some number between t and $t+h$. Finally, observe that $f_j(x_1(\tau), \dots, x_n(\tau))$ must approach $f_j(\xi_1, \dots, \xi_n)$ as t approaches infinity. Hence, $f_j(\xi_1, \dots, \xi_n) = 0$, $j = 1, 2, \dots, n$, and this proves Lemma 1. □

Example 1. Consider the system of differential equations

$$\frac{dx}{dt} = ax - bxy - ex^2, \qquad \frac{dy}{dt} = -cy + dxy - fy^2 \tag{2}$$

where a, b, c, d, e, and f are positive constants. This system (see Section 4.10) describes the population growth of two species x and y, where species y is dependent upon species x for its survival. Suppose that $c/d > a/e$. Prove that every solution $x(t)$, $y(t)$ of (2), with $x(0)$ and $y(0) > 0$, approaches the equilibrium solution $x = a/e$, $y = 0$, as t approaches infinity.

Solution. Our first step is to show that every solution $x(t), y(t)$ of (2) which starts in the first quadrant $(x > 0, y > 0)$ at $t = 0$ must remain in the first quadrant for all future time. (If this were not so, then the model (2) could not correspond to reality.) To this end, recall from Section 1.5 that

$$x(t) = \frac{ax_0}{ex_0 + (a - ex_0)e^{-at}}, \qquad y(t) = 0$$

is a solution of (2) for any choice of x_0. The orbit of this solution is the point $(0,0)$ for $x_0 = 0$; the line $0 < x < a/e$ for $0 < x_0 < a/e$; the point $(a/e, 0)$ for $x_0 = a/e$; and the line $a/e < x < \infty$ for $x_0 > a/e$. Thus, the x-axis, for $x \geqslant 0$, is the union of four disjoint orbits of (2). Similarly, (see Exercise 14), the positive y-axis is a single orbit of (2). Thus, if a solution

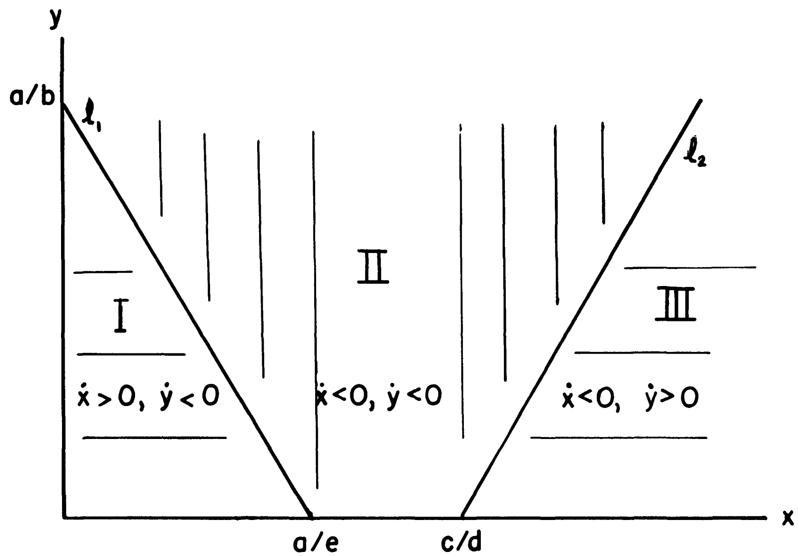


Figure 1

$x(t), y(t)$ of (2) leaves the first quadrant, its orbit must cross another orbit, and this is precluded by the uniqueness of orbits (Property 1, Section 4.6).

Our next step is to divide the first quadrant into regions where dx/dt and dy/dt have fixed signs. This is accomplished by drawing the lines $l_1: a - by - ex = 0$, and $l_2: -c + dx - fy = 0$, in the x–y plane. These lines divide the first quadrant into three regions I, II, and III as shown in Figure 1. (The lines l_1 and l_2 do not intersect in the first quadrant if $c/d > a/e$.) Now, observe that $ex + by$ is less than a in region I, while $ex + by$ is greater than a in regions II and III. Consequently, dx/dt is positive in region I and negative in regions II and III. Similarly, dy/dt is negative in regions I and II and positive in region III.

Next, we prove the following four simple lemmas.

Lemma 3. *Any solution $x(t), y(t)$ of (2) which starts in region I at time $t = t_0$ will remain in this region for all future time $t \geqslant t_0$ and ultimately approach the equilibrium solution $x = a/e, y = 0$.*

Proof. Suppose that a solution $x(t), y(t)$ of (2) leaves region I at time $t = t^*$. Then, $\dot{x}(t^*) = 0$, since the only way a solution can leave region I is by crossing the line l_1. Differentiating both sides of the first equation of (2) with respect to t and setting $t = t^*$ gives

$$\frac{d^2x(t^*)}{dt^2} = -bx(t^*)\frac{dy(t^*)}{dt}.$$

This quantity is positive. Hence, $x(t)$ has a minimum at $t = t^*$. But this is impossible, since $x(t)$ is always increasing whenever $x(t), y(t)$ is in region

I. Thus, any solution $x(t), y(t)$ of (2) which starts in region I at time $t=t_0$ will remain in region I for all future time $t \geqslant t_0$. This implies that $x(t)$ is a monotonic increasing function of time, and $y(t)$ is a monotonic decreasing function of time for $t \geqslant t_0$, with $x(t) < a/e$ and $y(t) > 0$. Consequently, by Lemma 1, both $x(t)$ and $y(t)$ have limits ξ, η respectively, as t approaches infinity. Lemma 2 implies that (ξ, η) is an equilibrium point of (2). Now, it is easily verified that the only equilibrium points of (2) in the region $x \geqslant 0$, $y \geqslant 0$ are $x=0, y=0$, and $x=a/e, y=0$. Clearly, ξ cannot equal zero since $x(t)$ is increasing in region I. Therefore, $\xi = a/e$ and $\eta = 0$. □

Lemma 4. *Any solution $x(t), y(t)$ of* (2) *which starts in region* III *at time $t = t_0$ must leave this region at some later time.*

Proof. Suppose that a solution $x(t), y(t)$ of (2) remains in region III for all time $t \geqslant t_0$. Then, $x(t)$ is a monotonic decreasing function of time, and $y(t)$ is a monotonic increasing function of time, for $t \geqslant t_0$. Moreover, $x(t)$ is greater than c/d and $y(t)$ is less than $(dx(t_0) - c)/f$. Consequently, both $x(t)$ and $y(t)$ have limits ξ, η respectively, as t approaches infinity. Lemma 2 implies that (ξ, η) is an equilibrium point of (2). But (ξ, η) cannot equal $(0,0)$ or $(a/e, 0)$ if $x(t), y(t)$ is in region III for $t \geqslant t_0$. This contradiction establishes Lemma 4. □

Lemma 5. *Any solution $x(t), y(t)$ of* (2) *which starts in region* II *at time $t = t_0$ and remains in region II for all future time $t \geqslant t_0$ must approach the equilibrium solution $x = a/e, y = 0$.*

Proof. Suppose that a solution $x(t), y(t)$ of (2) remains in region II for all time $t \geqslant t_0$. Then, both $x(t)$ and $y(t)$ are monotonic decreasing functions of time for $t \geqslant t_0$, with $x(t) > 0$ and $y(t) > 0$. Consequently, by Lemma 1, both $x(t)$ and $y(t)$ have limits ξ, η respectively, as t approaches infinity. Lemma 2 implies that (ξ, η) is an equilibrium point of (2). Now, (ξ, η) cannot equal $(0,0)$. Therefore, $\xi = a/e, \eta = 0$. □

Lemma 6. *A solution $x(t), y(t)$ of* (2) *cannot enter region* III *from region* II.

Proof. Suppose that a solution $x(t), y(t)$ of (2) leaves region II at time $t = t^*$ and enters region III. Then, $\dot{y}(t^*) = 0$. Differentiating both sides of the second equation of (2) with respect to t and setting $t = t^*$ gives

$$\frac{d^2 y(t^*)}{dt^2} = dy(t^*) \frac{dx(t^*)}{dt}.$$

This quantity is negative. Hence, $y(t)$ has a maximum at $t = t^*$. But this is impossible, since $y(t)$ is decreasing whenever $x(t), y(t)$ is in region II. □

Finally, observe that a solution $x(t), y(t)$ of (2) which starts on l_1 must immediately enter region I, and that a solution which starts on l_2 must immediately enter region II. It now follows immediately from Lemmas 3–6 that every solution $x(t), y(t)$ of (2), with $x(0)>0$ and $y(0)>0$, approaches the equilibrium solution $x=a/e, y=0$ as t approaches infinity.

Up to now, the solutions and orbits of the nonlinear equations that we have studied behaved very much like the solutions and orbits of linear equations. In actual fact, though, the situation is very different. The solutions and orbits of nonlinear equations, in general, exhibit a completely different behavior than the solutions and orbits of linear equations. A standard example is the system of equations

$$\frac{dx}{dt}=-y+x(1-x^2-y^2), \qquad \frac{dy}{dt}=x+y(1-x^2-y^2). \tag{3}$$

Since the term x^2+y^2 appears prominently in both equations, it suggests itself to introduce polar coordinates r,θ, where $x=r\cos\theta, y=r\sin\theta$, and to rewrite (3) in terms of r and θ. To this end, we compute

$$\begin{aligned}\frac{d}{dt}r^2=2r\frac{dr}{dt}&=2x\frac{dx}{dt}+2y\frac{dy}{dt}\\&=2(x^2+y^2)-2(x^2+y^2)^2=2r^2(1-r^2).\end{aligned}$$

Similarly, we compute

$$\frac{d\theta}{dt}=\frac{d}{dt}\arctan\frac{y}{x}=\frac{1}{x^2}\,\frac{x\dfrac{dy}{dt}-y\dfrac{dx}{dt}}{1+(y/x)^2}=\frac{x^2+y^2}{x^2+y^2}=1.$$

Consequently, the system of equations (3) is equivalent to the system of equations

$$\frac{dr}{dt}=r(1-r^2), \qquad \frac{d\theta}{dt}=1. \tag{4}$$

The general solution of (4) is easily seen to be

$$r(t)=\frac{r_0}{\left[r_0^2+(1-r_0^2)e^{-2t}\right]^{1/2}}, \qquad \theta=t+\theta_0 \tag{5}$$

where $r_0=r(0)$ and $\theta_0=\theta(0)$. Hence,

$$\begin{aligned}x(t)&=\frac{r_0}{\left[r_0^2+(1-r_0^2)e^{-2t}\right]^{1/2}}\cos(t+\theta_0),\\ y(t)&=\frac{r_0}{\left[r_0^2+(1-r_0^2)e^{-2t}\right]^{1/2}}\sin(t+\theta_0).\end{aligned}$$

Now, observe first that $x=0, y=0$ is the only equilibrium solution of (3). Second, observe that

$$x(t)=\cos(t+\theta_0), \qquad y(t)=\sin(t+\theta_0)$$

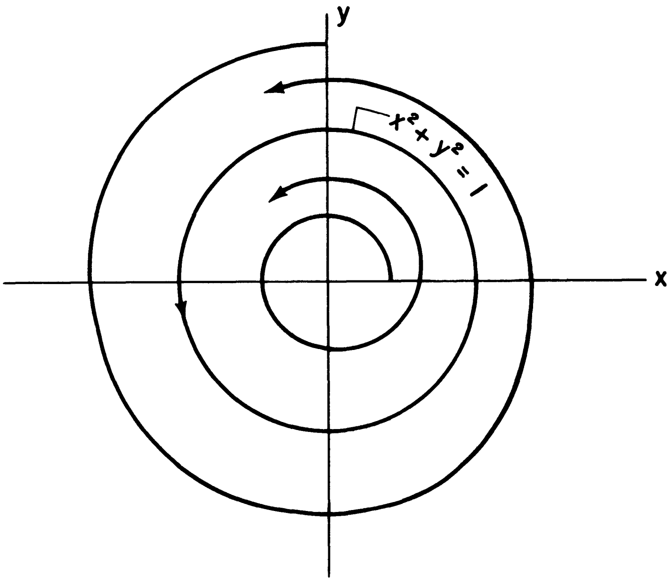


Figure 2. The phase portrait of (3)

when $r_0=1$. This solution is periodic with period 2π, and its orbit is the unit circle $x^2+y^2=1$. Finally, observe from (5) that $r(t)$ approaches one as t approaches infinity, for $r_0\neq 0$. Hence, all the orbits of (3), with the exception of the equilibrium point $x=0$, $y=0$, spiral into the unit circle. This situation is depicted in Figure 2.

The system of equations (3) shows that the orbits of a nonlinear system of equations may spiral into a simple closed curve. This, of course, is not possible for linear systems. Moreover, it is often possible to prove that orbits of a nonlinear system spiral into a closed curve even when we cannot explicitly solve the system of equations, or even find its orbits. This is the content of the following celebrated theorem.

Theorem 5. (Poincaré–Bendixson.) *Suppose that a solution $x=x(t)$, $y=y(t)$ of the system of differential equations*

$$\frac{dx}{dt}=f(x,y), \qquad \frac{dy}{dt}=g(x,y) \tag{6}$$

remains in a bounded region of the plane which contains no equilibrium points of (6). *Then, its orbit must spiral into a simple closed curve, which is itself the orbit of a periodic solution of* (6).

Example 2. Prove that the second-order differential equation

$$\ddot{z}+(z^2+2\dot{z}^2-1)\dot{z}+z=0 \tag{7}$$

has a nontrivial periodic solution.

Solution. First, we convert Equation (7) to a system of two first-order equations by setting $x=z$ and $y=\dot{z}$. Then,

$$\frac{dx}{dt}=y, \qquad \frac{dy}{dt}=-x+(1-x^2-2y^2)y. \tag{8}$$

Next, we try and find a bounded region R in the $x-y$ plane, containing no equilibrium points of (8), and having the property that every solution $x(t)$, $y(t)$ of (8) which starts in R at time $t=t_0$, remains there for all future time $t\geq t_0$. It can be shown that a simply connected region such as a square or disc will never work. Therefore, we try and take R to be an annulus surrounding the origin. To this end, compute

$$\frac{d}{dt}\left(\frac{x^2+y^2}{2}\right)=x\frac{dx}{dt}+y\frac{dy}{dt}=(1-x^2-2y^2)y^2,$$

and observe that $1-x^2-2y^2$ is positive for $x^2+y^2<\frac{1}{2}$ and negative for $x^2+y^2>1$. Hence, $x^2(t)+y^2(t)$ is increasing along any solution $x(t)$, $y(t)$ of (8) when $x^2+y^2<\frac{1}{2}$ and decreasing when $x^2+y^2>1$. This implies that any solution $x(t)$, $y(t)$ of (8) which starts in the annulus $\frac{1}{2}<x^2+y^2<1$ at time $t=t_0$ will remain in this annulus for all future time $t\geq t_0$. Now, this annulus contains no equilibrium points of (8). Consequently, by the Poincaré–Bendixson Theorem, there exists at least one periodic solution $x(t), y(t)$ of (8) lying entirely in this annulus, and then $z=x(t)$ is a nontrivial periodic solution of (7).

Exercises

1. *What Really Happened at the Paris Peace Talks*
The original plan developed by Henry Kissinger and Le Duc Tho to settle the Vietnamese war is described below. It was agreed that 1 million South Vietnamese ants and 1 million North Vietnamese ants would be placed in the backyard of the Presidential palace in Paris and be allowed to fight it out for a long period of time. If the South Vietnamese ants destroyed nearly all the North Vietnamese ants, then South Vietnam would retain control of all of its land. If the North Vietnamese ants were victorious, then North Vietnam would take over all of South Vietnam. If they appeared to be fighting to a standoff, then South Vietnam would be partitioned according to the proportion of ants remaining. Now, the South Vietnamese ants, denoted by S, and the North Vietnamese ants, denoted by N, compete against each other according to the following differential equations:

$$\begin{aligned}\frac{dS}{dt}&=\frac{1}{10}S-\frac{1}{20}S\times N\\ \frac{dN}{dt}&=\frac{1}{100}N-\frac{1}{100}N^2-\frac{1}{100}S\times N.\end{aligned} \tag{*}$$

Note that these equations correspond to reality since the South Vietnamese ants multiply much more rapidly than the North Vietnamese ants, but the North Vietnamese ants are much better fighters.

The battle began at 10:00 sharp on the morning of May 19, 1972, and was supervised by a representative of Poland and a representative of Canada. At 2:43 p.m. on the afternoon of May 21, the representative of Poland, being unhappy with the progress of the battle, slipped a bag of North Vietnamese ants into the backyard, but he was spotted by the eagle eyes of the representative of Canada. The South Vietnamese immediately claimed a foul and called off the agreement, thus setting the stage for the protracted talks that followed in Paris. The representative of Poland was hauled before a judge in Paris for sentencing. The judge, after making some remarks about the stupidity of the South Vietnamese, gave the Polish representative a very light sentence. Justify mathematically the judge's decision. *Hint*:

(a) Show that the lines $N=2$ and $N+S=1$ divide the first quadrant into three regions (see Figure 3) in which dS/dt and dN/dt have fixed signs.
(b) Show that every solution $S(t), N(t)$ of (*) which starts in either region I or region III must eventually enter region II.
(c) Show that every solution $S(t), N(t)$ of (*) which starts in region II must remain there for all future time.
(d) Conclude from (c) that $S(t)\to\infty$ for all solutions $S(t), N(t)$ of (*) with $S(t_0)$ and $N(t_0)$ positive. Conclude too that $N(t)$ has a finite limit ($\leqslant 2$) as $t\to\infty$.
(e) To prove that $N(t)\to 0$, observe that there exists t_0 such that $dN/dt \leqslant -N$ for $t \geqslant t_0$. Conclude from this inequality that $N(t)\to 0$ as $t\to\infty$.

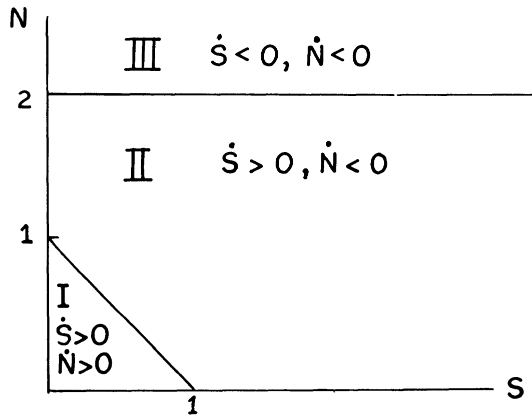


Figure 3

2. Consider the system of differential equations

$$\frac{dx}{dt} = ax - bxy, \qquad \frac{dy}{dt} = cy - dxy - ey^2 \tag{*}$$

with $a/b > c/e$. Prove that $y(t)\to 0$ as $t\to\infty$, for every solution $x(t), y(t)$ of (*) with $x(t_0)$ and $y(t_0)$ positive. *Hint*: Follow the outline in Exercise 1.

3. (a) Without computing the eigenvalues of the matrix

$$\begin{pmatrix} -3 & 1 \\ 1 & -3 \end{pmatrix},$$

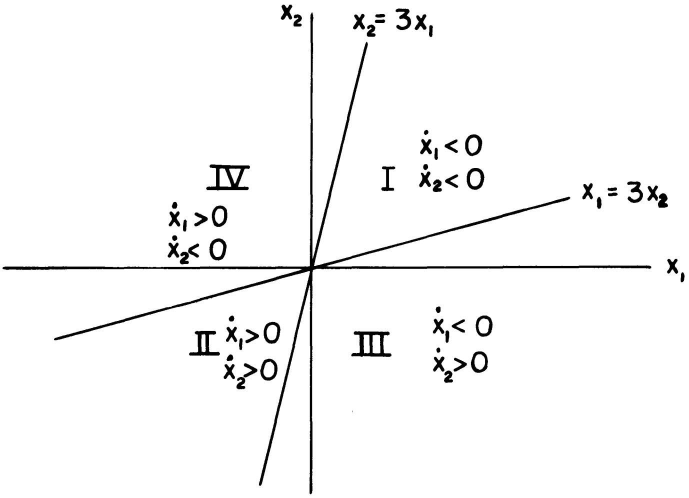


Figure 4

prove that every solution $\mathbf{x}(t)$ of

$$\dot{\mathbf{x}} = \begin{pmatrix} -3 & 1 \\ 1 & -3 \end{pmatrix}\mathbf{x}$$

approaches zero as t approaches infinity. *Hint*: (a) Show that the lines $x_2 = 3x_1$ and $x_1 = 3x_2$ divide the $x_1 - x_2$ plane into four regions (see Figure 4) in which $\dot{x}_1$ and $\dot{x}_2$ have fixed signs.

(b) Show that every solution $\mathbf{x}(t)$ which starts in either region I or II must remain there for all future time and ultimately approach the equilibrium solution $\mathbf{x} = \mathbf{0}$.

(c) Show that every solution $\mathbf{x}(t)$ which remains exclusively in region III or IV must ultimately approach the equilibrium solution $\mathbf{x} = \mathbf{0}$.

A closed curve C is said to be a limit cycle of

$$\dot{x} = f(x,y), \qquad \dot{y} = g(x,y) \tag{*}$$

if orbits of (*) spiral into it, or away from it. It is stable if all orbits of (*) passing sufficiently close to it must ultimately spiral into it, and unstable otherwise. Find all limit cycles of each of the following systems of differential equations. (*Hint*: Compute $d(x^2+y^2)/dt$. Observe too, that C must be the orbit of a periodic solution of (*) if it contains no equilibrium points of (*).)

4. $\dot{x} = -y - \dfrac{x(x^2+y^2-2)}{\sqrt{x^2+y^2}}$

$\dot{y} = x - \dfrac{y(x^2+y^2-2)}{\sqrt{x^2+y^2}}$

5. $\dot{x} = x - x^3 - xy^2$

$\dot{y} = y - y^3 - yx^2$

6. $\dot{x}=y+x(x^2+y^2-1)(x^2+y^2-2)$
$\dot{y}=-x+y(x^2+y^2-1)(x^2+y^2-2)$

7. $\dot{x}=xy+x\cos(x^2+y^2)$
$\dot{y}=-x^2+y\cos(x^2+y^2)$

8. (a) Show that the system

$$\dot{x}=y+xf(r)/r, \qquad \dot{y}=-x+yf(r)/r \qquad (r^2=x^2+y^2) \tag{*}$$

has limit cycles corresponding to the zeros of $f(r)$. What is the direction of motion on these curves?

(b) Determine all limit cycles of (*) and discuss their stability if $f(r)=(r-3)^2(r^2-5r+4)$.

Use the Poincaré–Bendixson Theorem to prove the existence of a nontrivial periodic solution of each of the following differential equations.

9. $\ddot{z}+(z^2+\dot{z}^4-2)z=0$

10. $\ddot{z}+[\ln(z^2+4\dot{z}^2)]\dot{z}+z=0$

11. (a) According to Green's theorem in the plane, if C is a closed curve which is sufficiently "smooth," and if f and g are continuous and have continuous first partial derivatives, then

$$\oint_C [f(x,y)\,dy-g(x,y)\,dx]=\iint_R [f_x(x,y)+g_y(x,y)]\,dx\,dy$$

where R is the region enclosed by C. Assume that $x(t)$, $y(t)$ is a periodic solution of $\dot{x}=f(x,y)$, $\dot{y}=g(x,y)$, and let C be the orbit of this solution. Show that for this curve, the line integral above is zero.

(b) Suppose that f_x+g_y has the same sign throughout a simply connected region D in the $x-y$ plane. Show that the system of equations $\dot{x}=f(x,y)$, $\dot{y}=g(x,y)$ can have no periodic solution which is entirely in D.

12. Show that the system of differential equations

$$\dot{x}=x+y^2+x^3, \qquad \dot{y}=-x+y+yx^2$$

has no nontrivial periodic solution.

13. Show that the system of differential equations

$$\dot{x}=x-xy^2+y^3, \qquad \dot{y}=3y-yx^2+x^3$$

has no nontrivial periodic solution which lies inside the circle $x^2+y^2=4$.

14. (a) Show that $x=0, y=\psi(t)$ is a solution of (2) for any function $\psi(t)$ satisfying $\dot{\psi}=-c\psi-f\psi^2$.

(b) Choose $\psi(t_0)>0$. Show that the orbit of $x=0, y=\psi(t)$ (for all t for which ψ exists) is the positive y axis.

4.9 Introduction to bifurcation theory

Consider the system of equations

$$\dot{\mathbf{x}}=\mathbf{f}(\mathbf{x},\varepsilon) \tag{1}$$

where

$$\mathbf{x}=\begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

and ε is a scalar. Intuitively speaking a bifurcation point of (1) is a value of ε at which the solutions of (1) change their behavior. More precisely, we say that $\varepsilon=\varepsilon_0$ is a bifurcation point of (1) if the phase portraits of (1) for $\varepsilon<\varepsilon_0$ and $\varepsilon>\varepsilon_0$ are different.

Remark. In the examples that follow we will appeal to our intuition in deciding whether two phase portraits are the same or are different. In more advanced courses we define two phase portraits to be the same, or topologically equivalent, if there exists a continuous transformation of the plane onto itself which maps one phase portrait onto the other.

Example 1. Find the bifurcation points of the system

$$\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}=\begin{pmatrix}1 & \varepsilon\\ 1 & -1\end{pmatrix}\mathbf{x} \tag{2}$$

Solution. The characteristic polynomial of the matrix $\mathbf{A}$ is

$$\begin{aligned}p(\lambda)&=\det(\mathbf{A}-\lambda\mathbf{I})\\&=\det\begin{pmatrix}1-\lambda & \varepsilon\\ 1 & -1-\lambda\end{pmatrix}\\&=(\lambda-1)(\lambda+1)-\varepsilon\\&=\lambda^2-(1+\varepsilon).\end{aligned}$$

The roots of $p(\lambda)$ are $\pm\sqrt{1+\varepsilon}$ for $\varepsilon>-1$, and $\pm\sqrt{-\varepsilon-1}\,i$ for $\varepsilon<-1$. This

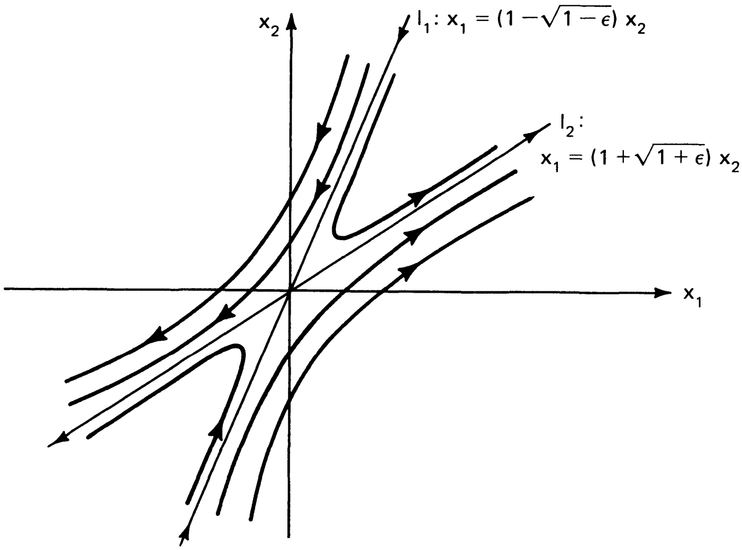


Figure 1. Phase portrait of (2) for $\varepsilon>-1$

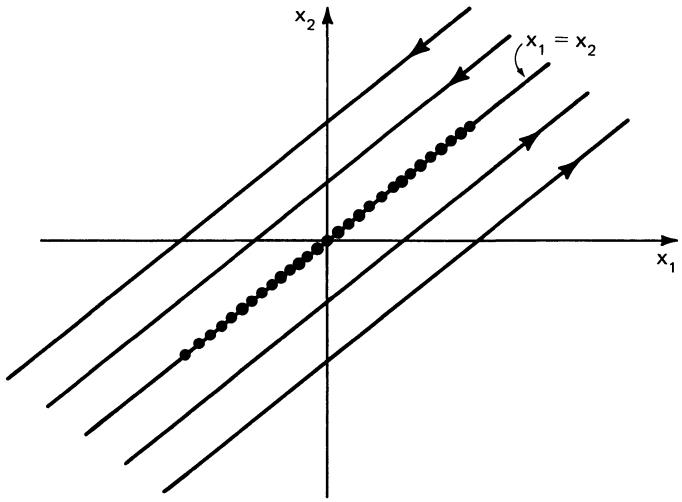


Figure 2. Phase portrait of (2) for $\varepsilon = -1$

implies that $\mathbf{x} = \mathbf{0}$ is a saddle for $\varepsilon > -1$, and a center for $\varepsilon < -1$. We conclude, therefore, that $\varepsilon = -1$ is a bifurcation point of (2). It is also clear that Eq. (2) has no other bifurcation points.

It is instructive to see how the solutions of (2) change as ε passes through the bifurcation value -1. For $\varepsilon > -1$, the eigenvalues of $\mathbf{A}$ are

$$\lambda_1 = \sqrt{1+\varepsilon}\,, \quad \lambda_2 = -\sqrt{1+\varepsilon}\,.$$

It is easily verified (see Exercise 10) that

$$\mathbf{x}^1 = \begin{pmatrix} 1+\sqrt{1+\varepsilon} \\ 1 \end{pmatrix}$$

is an eigenvector of $\mathbf{A}$ with eigenvalue $\sqrt{1+\varepsilon}$, while

$$\mathbf{x}^2 = \begin{pmatrix} 1-\sqrt{1+\varepsilon} \\ 1 \end{pmatrix}$$

is an eigenvector with eigenvalue $-\sqrt{1+\varepsilon}$. Hence, the phase portrait of (2) has the form shown in Figure 1. As $\varepsilon \to -1$ from the left, the lines l_1 and l_2 both approach the line $x_1 = x_2$. This line is a line of equilibrium points of (2) when $\varepsilon = -1$, while each line $x_1 - x_2 = c$ $(c \neq 0)$ is an orbit of (2) for $\varepsilon = -1$. The phase portrait of (2) for $\varepsilon = -1$ is given in Figure 2.

Example 2. Find the bifurcation points of the system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} = \begin{pmatrix} 0 & -1 \\ \varepsilon & -1 \end{pmatrix}\mathbf{x}. \tag{3}$$

Solution. The characteristic polynomial of the matrix **A** is

$$\begin{aligned}
p(\lambda) = \det(\mathbf{A}-\lambda\mathbf{I}) &= \det\begin{pmatrix} -\lambda & -1 \\ \varepsilon & -1-\lambda \end{pmatrix} \\
&= \lambda(1+\lambda)+\varepsilon \\
&= \lambda^2+\lambda+\varepsilon
\end{aligned}$$

and the roots of $p(\lambda)$ are

$$\lambda_1 = \frac{-1+\sqrt{1-4\varepsilon}}{2}, \lambda_2 = \frac{-1-\sqrt{1-4\varepsilon}}{2}.$$

Observe that λ_1 is positive and λ_2 is negative for $\varepsilon<0$. Hence, $\mathbf{x}=\mathbf{0}$ is a saddle for $\varepsilon<0$. For $0<\varepsilon<1/4$, both λ_1 and λ_2 are negative. Hence $\mathbf{x}=\mathbf{0}$ is a stable node for $0<\varepsilon<1/4$. Both λ_1 and λ_2 are complex, with negative real part, for $\varepsilon>1/4$. Hence $\mathbf{x}=\mathbf{0}$ is a stable focus for $\varepsilon>1/4$. Note that the phase portrait of (3) changes as ε passes through 0 and $1/4$. We conclude, therefore, that $\varepsilon=0$ and $\varepsilon=1/4$ are bifurcation points of (3).

Example 3. Find the bifurcation points of the system of equations

$$\begin{aligned}
\frac{dx_1}{dt} &= x_2 \\
\frac{dx_2}{dt} &= x_1^2 - x_2 - \varepsilon.
\end{aligned} \tag{4}$$

Solution. (i) We first find the equilibrium points of (4). Setting $dx_1/dt=0$ gives $x_2=0$, and then setting $dx_2/dt=0$ gives $x_1^2-\varepsilon=0$, so that $x_1=\pm\sqrt{\varepsilon}$, $\varepsilon>0$. Hence, $(\sqrt{\varepsilon},0)$ and $(-\sqrt{\varepsilon},0)$ are two equilibrium points of (4) for $\varepsilon>0$. The system (4) has no equilibrium points when $\varepsilon<0$. We conclude, therefore, that $\varepsilon=0$ is a bifurcation point of (4). (ii) We now analyze the behavior of the solutions of (4) near the equilibrium points $(\pm\sqrt{\varepsilon},0)$ to determine whether this system has any additional bifurcation points. Setting

$$u = x_1 \mp \sqrt{\varepsilon}, \quad v = x_2$$

gives

$$\begin{aligned}
\frac{du}{dt} &= v \\
\frac{dv}{dt} &= \left(u \pm \sqrt{\varepsilon}\right)^2 - v - \varepsilon = \pm 2\sqrt{\varepsilon}\,u - v + u^2.
\end{aligned} \tag{5}$$

The system (5) can be written in the form

$$\frac{d}{dt}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ \pm 2\sqrt{\varepsilon} & -1 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} 0 \\ u^2 \end{pmatrix}.$$

By Theorem 4, the phase portrait of (4) near the equilibrium solution

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \pm\sqrt{\varepsilon} \\ 0 \end{pmatrix}$$

is determined by the phase portrait of the linearized system

$$\frac{d}{dt}\begin{pmatrix} u \\ v \end{pmatrix} = \mathbf{A}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ \pm 2\sqrt{\varepsilon} & -1 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix}.$$

To find the eigenvalues of $\mathbf{A}$ we compute

$$\begin{aligned} p(\lambda) &= \det(\mathbf{A} - \lambda\mathbf{I}) \\ &= \det\begin{pmatrix} -\lambda & 1 \\ \pm 2\sqrt{\varepsilon} & -1-\lambda \end{pmatrix} \\ &= \lambda^2 + \lambda \mp 2\sqrt{\varepsilon}. \end{aligned}$$

Hence, the eigenvalues of $\mathbf{A}$ when $u = x_1 - \sqrt{\varepsilon}$ are

$$\lambda_1 = \frac{-1+\sqrt{1+8\sqrt{\varepsilon}}}{2}, \quad \lambda_2 = \frac{-1-\sqrt{1+8\sqrt{\varepsilon}}}{2} \tag{6}$$

while the eigenvalues of $\mathbf{A}$ when $u = x_1 + \sqrt{\varepsilon}$ are

$$\lambda_1 = \frac{-1+\sqrt{1-8\sqrt{\varepsilon}}}{2}, \quad \lambda_2 = \frac{-1-\sqrt{1-8\sqrt{\varepsilon}}}{2}. \tag{7}$$

Observe from (6) that $\lambda_1 > 0$, while $\lambda_2 < 0$. Thus, the system (4) behaves like a saddle near the equilibrium points $\begin{pmatrix} \sqrt{\varepsilon} \\ 0 \end{pmatrix}$. On the other hand, we see from (7) that both λ_1 and λ_2 are negative for $0 < \varepsilon < 1/64$, and complex for $\varepsilon > 1/64$. Consequently, the system (4) near the equilibrium solution $\begin{pmatrix} -\sqrt{\varepsilon} \\ 0 \end{pmatrix}$ behaves like a stable node for $0 < \varepsilon < 1/64$, and a stable focus for $\varepsilon > 1/64$. It can be shown that the phase portraits of a stable node and a stable focus are equivalent. Consequently, $\varepsilon = 1/64$ is *not* a bifurcation point of (4).

Another situation which is included in the context of bifurcation theory, and which is of much current research interest now, is when the system (1) has a certain number of equilibrium, or periodic, solutions for $\varepsilon=\varepsilon_0$, and a different number for $\varepsilon\neq\varepsilon_0$. Suppose, for example, that $\mathbf{x}(t)$ is an equilibrium, or periodic, solution of (1) for $\varepsilon=0$, and $\mathbf{x}^1(t),\mathbf{x}^2(t),\ldots,\mathbf{x}^k(t)$ are equilibria, or periodic solutions of (1) for $\varepsilon\neq 0$ which approach $\mathbf{x}(t)$ as $\varepsilon\to 0$. In this case we say that the solutions $\mathbf{x}^1(t),\mathbf{x}^2(t),\ldots,\mathbf{x}^k(t)$ *bifurcate* from $\mathbf{x}(t)$. We illustrate this situation with the following example.

Example 4. Find all equilibrium solutions of the system of equations

$$\begin{aligned}\frac{dx_1}{dt}&=3\varepsilon x_1-3\varepsilon x_2-x_1^2-x_2^2\\ \frac{dx}{dt}&=\varepsilon x_1-x_1x_2=x_1(\varepsilon-x_2).\end{aligned}\tag{8}$$

Solution. Let $\begin{pmatrix}x_1\\x_2\end{pmatrix}$ be an equilibrium solution of the system (8). The second equation of (8) implies that $x_1=0$ or $x_2=\varepsilon$.

$x_1=0$. In this case, the first equation of (8) implies that

$$0=3\varepsilon x_2+x_2^2=x_2(3\varepsilon+x_2)$$

so that $x_2=0$ or $x_2=-3\varepsilon$. Thus $\begin{pmatrix}0\\0\end{pmatrix}$ and $\begin{pmatrix}0\\-3\varepsilon\end{pmatrix}$ are two equilibrium points of (8).

$x_2=\varepsilon$. In this case, the first equation of (8) implies that

$$3\varepsilon x_1-3\varepsilon^2-x_1^2-\varepsilon^2=0$$

or

$$x_1^2-3\varepsilon x_1+4\varepsilon^2=0.\tag{9}$$

The solutions

$$x_1=\frac{3\varepsilon\pm\sqrt{9\varepsilon^2-16\varepsilon^2}}{2}$$

of (9) are complex. Thus,

$$\mathbf{x}^1=\begin{pmatrix}0\\0\end{pmatrix}\quad\text{and}\quad\mathbf{x}^2=\begin{pmatrix}0\\-3\varepsilon\end{pmatrix}$$

are two equilibrium points of (8), for $\varepsilon\neq 0$, which bifurcate from the single equilibrium point $\mathbf{x}=\begin{pmatrix}0\\0\end{pmatrix}$ when $\varepsilon=0$.

EXERCISES

Find the bifurcation points of each of the following systems of equations.

1. $\dot{\mathbf{x}}=\begin{pmatrix}1 & \varepsilon\\ \varepsilon & 1\end{pmatrix}\mathbf{x}$

2. $\dot{\mathbf{x}}=\begin{pmatrix}1 & 1\\ \varepsilon & 1\end{pmatrix}\mathbf{x}$

3. $\dot{\mathbf{x}}=\begin{pmatrix}0 & 2\\ -2 & \varepsilon\end{pmatrix}\mathbf{x}$

4. $\dot{\mathbf{x}}=\begin{pmatrix}0 & 2\\ 2 & \varepsilon\end{pmatrix}\mathbf{x}$

5. $\dot{\mathbf{x}}=\begin{pmatrix}0 & -\varepsilon\\ \varepsilon & 0\end{pmatrix}\mathbf{x}$

In each of Problems 6–8, show that more than one equilibrium solutions bifurcate from the equilibrium solution $\mathbf{x}=\mathbf{0}$ when $\varepsilon=0$.

6. $\dot{x}_1=\varepsilon x_1-\varepsilon x_2-x_1^2+x_2^2$
$\dot{x}_2=\varepsilon x_2+x_1x_2$

7. $\dot{x}_1=\varepsilon x_1-x_1^2-x_1x_2$
$\dot{x}_2=-2\varepsilon x_1+2\varepsilon x_2+x_1x_2-x_2^2$

8. $\dot{x}_1=\varepsilon x_2+x_1x_2$
$\dot{x}_2=-\varepsilon x_1+\varepsilon x_2+x_1^2+x_2^2$

9. Consider the system of equations

$$\begin{aligned}\dot{x}_1&=3\varepsilon x_1-5\varepsilon x_2-x_1^2+x_2^2\\ \dot{x}_2&=2\varepsilon x_1-\varepsilon x_2.\end{aligned}\qquad(*)$$

(a) Show that each point on the lines $x_2=x_1$ and $x_2=-x_1$ are equilibrium points of $(*)$ for $\varepsilon=0$.

(b) Show that

$$\begin{pmatrix}x_1\\ x_2\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix}\quad\text{and}\quad\begin{pmatrix}x_1\\ x_2\end{pmatrix}=\tfrac{7}{3}\varepsilon\begin{pmatrix}1\\ 2\end{pmatrix}.$$

are the only equilibrium points of $(*)$ for $\varepsilon\neq 0$.

10. Show that

$$\mathbf{x}^1=\begin{pmatrix}1+\sqrt{1+\varepsilon}\\ 1\end{pmatrix}\quad\text{and}\quad\mathbf{x}^2=\begin{pmatrix}1-\sqrt{1+\varepsilon}\\ 1\end{pmatrix}$$

are eigenvectors of the matrix $\begin{pmatrix}1 & \varepsilon\\ 1 & -1\end{pmatrix}$ with eigenvalues $\sqrt{1+\varepsilon}$ and $-\sqrt{1+\varepsilon}$ respectively.

4.10 Predator–prey problems; or why the percentage of sharks caught in the Mediterranean Sea rose dramatically during World War I

In the mid 1920's the Italian biologist Umberto D'Ancona was studying the population variations of various species of fish that interact with each other. In the course of his research, he came across some data on percentages-of-total-catch of several species of fish that were brought into different Mediterranean ports in the years that spanned World War I. In particular, the data gave the percentage-of-total-catch of selachians, (sharks, skates, rays, etc.) which are not very desirable as food fish. The data for the port of Fiume, Italy, during the years 1914–1923 is given below.

1914	1915	1916	1917	1918
11.9%	21.4%	22.1%	21.2%	36.4%
1919	1920	1921	1922	1923
27.3%	16.0%	15.9%	14.8%	10.7%

D'Ancona was puzzled by the very large increase in the percentage of selachians during the period of the war. Obviously, he reasoned, the increase in the percentage of selachians was due to the greatly reduced level of fishing during this period. But how does the intensity of fishing affect the fish populations? The answer to this question was of great concern to D'Ancona in his research on the struggle for existence between competing species. It was also of concern to the fishing industry, since it would have obvious implications for the way fishing should be done.

Now, what distinguishes the selachians from the food fish is that the selachians are predators, while the food fish are their prey; the selachians depend on the food fish for their survival. At first, D'Ancona thought that this accounted for the large increase of selachians during the war. Since the level of fishing was greatly reduced during this period, there were more prey available to the selachians, who therefore thrived and multiplied rapidly. However, this explanation does not hold any water since there were also more food fish during this period. D'Ancona's theory only shows that there are more selachians when the level of fishing is reduced; it does not explain why a reduced level of fishing is *more* beneficial to the predators than to their prey.

After exhausting all possible biological explanations of this phenomenon, D'Ancona turned to his colleague, the famous Italian mathematician Vito Volterra. Hopefully, Volterra would formulate a mathematical model of the growth of the selachians and their prey, the food fish, and this model would provide the answer to D'Ancona's question. Volterra began his analysis of this problem by separating all the fish into the prey population $x(t)$ and the predator population $y(t)$. Then, he reasoned that the food fish do not compete very intensively among themselves for their food supply since this is very abundant, and the fish population is not very dense. Hence, in the absence of the selachians, the food fish would grow according to the Malthusian law of population growth $\dot{x}=ax$, for some positive constant a. Next, reasoned Volterra, the number of contacts per unit time between predators and prey is bxy, for some positive constant b. Hence, $\dot{x}=ax-bxy$. Similarly, Volterra concluded that the predators have a natural rate of decrease $-cy$ proportional to their present number, and that they also increase at a rate dxy proportional to their present number y and their food supply x. Thus,

$$\frac{dx}{dt}=ax-bxy, \qquad \frac{dy}{dt}=-cy+dxy. \tag{1}$$

The system of equations (1) governs the interaction of the selachians and food fish in the absence of fishing. We will carefully analyze this system and derive several interesting properties of its solutions. Then, we will include the effect of fishing in our model, and show why a reduced level of fishing is more beneficial to the selachians than to the food fish. In fact, we will derive the surprising result that a reduced level of fishing is actually harmful to the food fish.

Observe first that (1) has two equilibrium solutions $x(t)=0, y(t)=0$ and $x(t)=c/d, y(t)=a/b$. The first equilibrium solution, of course, is of no interest to us. This system also has the family of solutions $x(t)=x_0e^{at}, y(t)=0$ and $x(t)=0, y(t)=y_0e^{-ct}$. Thus, both the x and y axes are orbits of (1). This implies that every solution $x(t)$, $y(t)$ of (1) which starts in the first quadrant $x>0, y>0$ at time $t=t_0$ will remain there for all future time $t\geqslant t_0$.

The orbits of (1), for $x,y\neq 0$ are the solution curves of the first-order equation

$$\frac{dy}{dx}=\frac{-cy+dxy}{ax-bxy}=\frac{y(-c+dx)}{x(a-by)}. \tag{2}$$

This equation is separable, since we can write it in the form

$$\frac{a-by}{y}\frac{dy}{dx}=\frac{-c+dx}{x}.$$

Consequently, $a\ln y-by+c\ln x-dx=k_1$, for some constant k_1. Taking exponentials of both sides of this equation gives

$$\frac{y^a}{e^{by}}\frac{x^c}{e^{dx}}=K \tag{3}$$

for some constant K. Thus, the orbits of (1) are the family of curves defined by (3), and these curves are *closed* as we now show.

Lemma 1. *Equation* (3) *defines a family of closed curves for* $x,y>0$.

PROOF. Our first step is to determine the behavior of the functions $f(y)=y^a/e^{by}$ and $g(x)=x^c/e^{dx}$ for x and y positive. To this end, observe that $f(0)=0, f(\infty)=0$, and $f(y)$ is positive for $y>0$. Computing

$$f'(y)=\frac{ay^{a-1}-by^a}{e^{by}}=\frac{y^{a-1}(a-by)}{e^{by}},$$

we see that $f(y)$ has a single critical point at $y=a/b$. Consequently, $f(y)$ achieves its maximum value $M_y=(a/b)^a/e^a$ at $y=a/b$, and the graph of

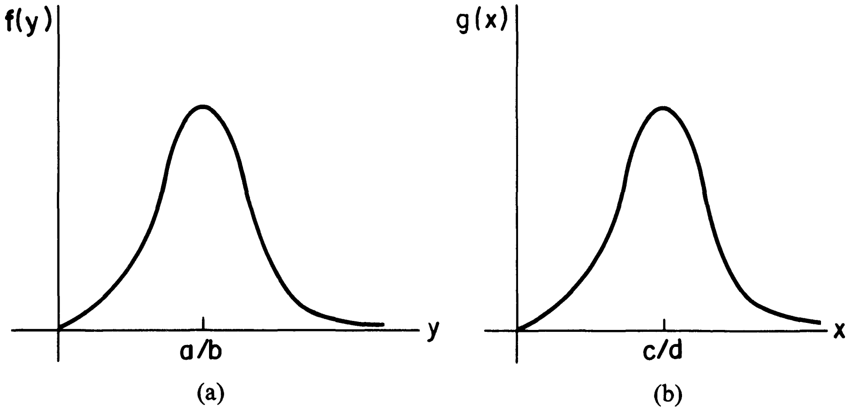


Figure 1. (a) Graph of $f(y)=y^a e^{-by}$; (b) Graph of $g(x)=x^c e^{-dx}$

$f(y)$ has the form described in Figure 1a. Similarly, $g(x)$ achieves its maximum value $M_x=(c/d)^c/e^c$ at $x=c/d$, and the graph of $g(x)$ has the form described in Figure 1b.

From the preceding analysis, we conclude that Equation (3) has no solution $x,y>0$ for $K>M_xM_y$, and the single solution $x=c/d$, $y=a/b$ for $K=M_xM_y$. Thus, we need only consider the case $K=\lambda M_y$, where λ is a positive number less than M_x. Observe first that the equation $x^c/e^{dx}=\lambda$ has one solution $x=x_m<c/d$, and one solution $x=x_M>c/d$. Hence, the equation

$$f(y)=y^a e^{-by}=\left(\frac{\lambda}{x^c e^{-dx}}\right)M_y$$

has no solution y when x is less than x_m or greater than x_M. It has the single solution $y=a/b$ when $x=x_m$ or x_M, and it has two solutions $y_1(x)$ and $y_2(x)$ for each x between x_m and x_M. The smaller solution $y_1(x)$ is always less than a/b, while the larger solution $y_2(x)$ is always greater than a/b. As x approaches either x_m or x_M, both $y_1(x)$ and $y_2(x)$ approach a/b. Consequently, the curves defined by (3) are closed for x and y positive, and have the form described in Figure 2. Moreover, none of these closed curves (with the exception of $x=c/d$, $y=a/b$) contain any equilibrium points of (1). Therefore, all solutions $x(t)$, $y(t)$ of (1), with $x(0)$ and $y(0)$ positive, are *periodic* functions of time. That is to say, each solution $x(t)$, $y(t)$ of (1), with $x(0)$ and $y(0)$ positive, has the property that $x(t+T)=x(t)$ and $y(t+T)=y(t)$ for some positive T. □

Now, the data of D'Ancona is really an *average* over each one year period of the proportion of predators. Thus, in order to compare this data with the predictions of (1), we must compute the "average values" of $x(t)$ and $y(t)$, for any solution $x(t)$, $y(t)$ of (1). Remarkably, we can find these average values even though we cannot compute $x(t)$ and $y(t)$ exactly. This is the content of Lemma 2.

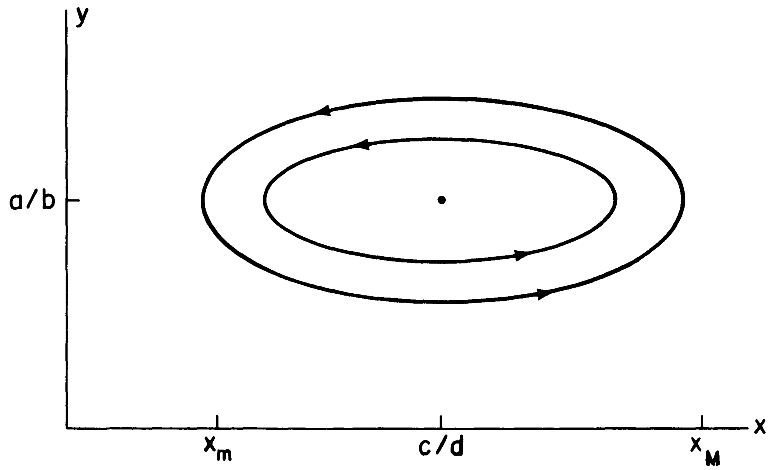


Figure 2. Orbits of (1) for x,y positive

Lemma 2. *Let $x(t), y(t)$ be a periodic solution of* (1), *with period $T>0$. Define the average values of x and y as*

$$\bar{x}=\frac{1}{T}\int_0^T x(t)\,dt, \qquad \bar{y}=\frac{1}{T}\int_0^T y(t)\,dt.$$

Then, $\bar{x}=c/d$ and $\bar{y}=a/b$. In other words, the average values of $x(t)$ and $y(t)$ are the equilibrium values.

PROOF. Dividing both sides of the first equation of (1) by x gives $\dot{x}/x=a-by$, so that

$$\frac{1}{T}\int_0^T \frac{\dot{x}(t)}{x(t)}\,dt=\frac{1}{T}\int_0^T\left[a-by(t)\right]dt.$$

Now, $\int_0^T \dot{x}(t)/x(t)\,dt=\ln x(T)-\ln x(0)$, and this equals zero since $x(T)=x(0)$. Consequently,

$$\frac{1}{T}\int_0^T by(t)\,dt=\frac{1}{T}\int_0^T a\,dt=a,$$

so that $\bar{y}=a/b$. Similarly, by dividing both sides of the second equation of (1) by $Ty(t)$ and integrating from 0 to T, we obtain that $\bar{x}=c/d$. □

We are now ready to include the effects of fishing in our model. Observe that fishing decreases the population of food fish at a rate $\varepsilon x(t)$, and decreases the population of selachians at a rate $\varepsilon y(t)$. The constant ε reflects the intensity of fishing; i.e., the number of boats at sea and the number of nets in the water. Thus, the true state of affairs is described by the modified system of differential equations

$$\begin{aligned}\frac{dx}{dt}&=ax-bxy-\varepsilon x=(a-\varepsilon)x-bxy\\ \frac{dy}{dt}&=-cy+dxy-\varepsilon y=-(c+\varepsilon)y+dxy.\end{aligned}\tag{4}$$

This system is exactly the same as (1) (for $a-\varepsilon>0$), with a replaced by $a-\varepsilon$, and c replaced by $c+\varepsilon$. Hence, the average values of $x(t)$ and $y(t)$ are now

$$\bar{x}=\frac{c+\varepsilon}{d},\qquad \bar{y}=\frac{a-\varepsilon}{b}.\tag{5}$$

Consequently, a moderate amount of fishing ($\varepsilon<a$) actually increases the number of food fish, on the average, and decreases the number of selachians. Conversely, a reduced level of fishing increases the number of selachians, on the average, and *decreases* the number of food fish. This remarkable result, which is known as Volterra's principle, explains the data of D'Ancona, and completely solves our problem.

Volterra's principle has spectacular applications to insecticide treatments, which destroy both insect predators and their insect prey. It implies that the application of insecticides will actually increase the population of those insects which are kept in control by other predatory insects. A remarkable confirmation comes from the cottony cushion scale insect (Icerya purchasi), which, when accidentally introduced from Australia in 1868, threatened to destroy the American citrus industry. Thereupon, its natural Australian predator, a ladybird beetle (Novius Cardinalis) was introduced, and the beetles reduced the scale insects to a low level. When DDT was discovered to kill scale insects, it was applied by the orchardists in the hope of further reducing the scale insects. However, in agreement with Volterra's principle, the effect was an increase of the scale insect!

Oddly enough, many ecologists and biologists refused to accept Volterra's model as accurate. They pointed to the fact that the oscillatory behavior predicted by Volterra's model is not observed in most predator–prey systems. Rather, most predator–prey systems tend to equilibrium states as time evolves. Our answer to these critics is that the system of differential equations (1) is not intended as a model of the general predator–prey interaction. This is because the food fish and selachians do not compete intensively among themselves for their available resources. A more general model of predator–prey interactions is the system of differential equations

$$\dot{x}=ax-bxy-ex^2,\qquad \dot{y}=-cy+dxy-fy^2.\tag{6}$$

Here, the term ex^2 reflects the internal competition of the prey x for their limited external resources, and the term fy^2 reflects the competition among the predators for the limited number of prey. The solutions of (6) are not, in general, periodic. Indeed, we have already shown in Example 1 of Section 4.8 that all solutions $x(t)$, $y(t)$ of (6), with $x(0)$ and $y(0)$ positive, ultimately approach the equilibrium solution $x=a/e$, $y=0$ if c/d is greater than a/e. In this situation, the predators die out, since their available food supply is inadequate for their needs.

Surprisingly, some ecologists and biologists even refuse to accept the more general model (6) as accurate. As a counterexample, they cite the experiments of the mathematical biologist G. F. Gause. In these experiments, the population was composed of two species of protozoa, one of which, Didinium nasatum, feeds on the other, Paramecium caudatum. In all of Gause's experiments, the Didinium quickly destroyed the Paramecium and then died of starvation. This situation cannot be modeled by the system of equations (6), since no solution of (6) with $x(0)y(0)\neq 0$ can reach $x=0$ or $y=0$ in finite time.

Our answer to these critics is that the Didinium are a special, and atypical type of predator. On the one hand, they are ferocious attackers and require a tremendous amount of food; a Didinium demands a fresh Paramecium every three hours. On the other hand, the Didinium don't perish from an insufficient supply of Paramecium. They continue to multiply, but give birth to smaller offspring. Thus, the system of equations (6) does not accurately model the interaction of Paramecium and Didinium. A better model, in this case, is the system of differential equations

$$\frac{dx}{dt}=ax-b\sqrt{x}\,y,\qquad \frac{dy}{dt}=\begin{cases} d\sqrt{x}\,y, & x\neq 0\\ -cy, & x=0\end{cases}. \tag{7}$$

It can be shown (see Exercise 6) that every solution $x(t)$, $y(t)$ of (7) with $x(0)$ and $y(0)$ positive reaches $x=0$ in finite time. This does not contradict the existence–uniqueness theorem, since the function

$$g(x,y)=\begin{cases} d\sqrt{x}\,y, & x\neq 0\\ -cy, & x=0\end{cases}$$

does not have a partial derivative with respect to x or y, at $x=0$.

Finally, we mention that there are several predator–prey interactions in nature which cannot be modeled by any system of ordinary differential equations. These situations occur when the prey are provided with a refuge that is inaccessible to the predators. In these situations, it is impossible to make any definitive statements about the future number of predators and prey, since we cannot predict how many prey will be stupid enough to leave their refuge. In other words, this process is now *random*, rather than *deterministic*, and therefore cannot be modeled by a system of ordinary differential equations. This was verified directly in a famous experiment of Gause. He placed five Paramecium and three Didinium in each of thirty identical test tubes, and provided the Paramecium with a refuge from the Didinium. Two days later, he found the predators dead in four tubes, and a mixed population containing from two to thirty-eight Paramecium in the remaining twenty-six tubes.

Reference

Volterra, V: "Leçons sur la théorie mathématique de la lutte pour la vie." Paris, 1931.

Exercises

1. Find all biologically realistic equilibrium points of (6) and determine their stability.

2. We showed in Section 4.8 that $y(t)$ ultimately approaches zero for all solutions $x(t), y(t)$ of (6), if $c/d > a/e$. Show that there exist solutions $x(t), y(t)$ of (6) for which $y(t)$ increases at first to a maximum value, and then decreases to zero. (To an observer who sees only the predators without noticing the prey, such a case of a population passing through a maximum to total extinction would be very difficult to explain.)

3. In many instances, it is the adult members of the prey who are chiefly attacked by the predators, while the young members are better protected, either by their smaller size, or by their living in a different station. Let x_1 be the number of adult prey, x_2 the number of young prey, and y the number of predators. Then,

$$\begin{aligned}\dot{x}_1 &= -a_1x_1 + a_2x_2 - bx_1y\\ \dot{x}_2 &= nx_1 - (a_1+a_2)x_2\\ \dot{y} &= -cy + dx_1y\end{aligned}$$

where a_2x_2 represents the number of young (per unit time) growing into adults, and n represents the birth rate proportional to the number of adults. Find all equilibrium solutions of this system.

4. There are several situations in nature where species 1 preys on species 2 which in turn preys on species 3. One case of this kind of population is the Island of Komodo in Malaya which is inhabited by giant carnivorous reptiles, and by mammals—their food—which feed on the rich vegetation of the island. We assume that the reptiles have no direct influence on the vegetation, and that only the plants compete among themselves for their available resources. A system of differential equations governing this interaction is

$$\begin{aligned}\dot{x}_1 &= -a_1x_1 - b_{12}x_1x_2 + c_{13}x_1x_3\\ \dot{x}_2 &= -a_2x_2 + b_{21}x_1x_2\\ \dot{x}_3 &= a_3x_3 - a_4x_3^2 - c_{31}x_1x_3\end{aligned}$$

Find all equilibrium solutions of this system.

5. Consider a predator–prey system where the predator has alternate means of support. This system can be modelled by the differential equations

$$\begin{aligned}\dot{x}_1 &= \alpha_1x_1(\beta_1 - x_1) + \gamma_1x_1x_2\\ \dot{x}_2 &= \alpha_2x_2(\beta_2 - x_2) - \gamma_2x_1x_2\end{aligned}$$

where $x_1(t)$ and $x_2(t)$ are the predators and prey populations, respectively, at time t.

(a) Show that the change of coordinates $\beta_i y_i(t) = x_i(t/\alpha_i\beta_i)$ reduces this system of equations to

$$\dot{y}_1 = y_1(1-y_1) + a_1 y_1 y_2, \qquad \dot{y}_2 = y_2(1-y_2) - a_2 y_1 y_2$$

where $a_1 = \gamma_1\beta_2/\alpha_1\beta_1$ and $a_2 = \gamma_2\beta_1/\alpha_2\beta_2$.

(b) What are the stable equilibrium populations when (i) $0 < a_2 < 1$, (ii) $a_2 > 1$?

(c) It is observed that $a_1 = 3a_2$ (a_2 is a measure of the aggressiveness of the predator). What is the value of a_2 if the predator's instinct is to maximize its stable equilibrium population?

6. (a) Let $x(t)$ be a solution of $\dot{x} = ax - M\sqrt{x}$, with $M > a\sqrt{x(t_0)}$. Show that

$$a\sqrt{x} = M - \left(M - a\sqrt{x(t_0)}\right)e^{a(t-t_0)/2}.$$

(b) Conclude from (a) that $x(t)$ approaches zero in finite time.

(c) Let $x(t)$, $y(t)$ be a solution of (7), with $by(t_0) > a\sqrt{x(t_0)}$. Show that $x(t)$ reaches zero in finite time. *Hint*: Observe that $y(t)$ is increasing for $t \geqslant t_0$.

(d) It can be shown that $by(t)$ will eventually exceed $a\sqrt{x(t)}$ for every solution $x(t)$, $y(t)$ of (7) with $x(t_0)$ and $y(t_0)$ positive. Conclude, therefore, that all solutions $x(t)$, $y(t)$ of (7) achieve $x = 0$ in finite time.

4.11 The principle of competitive exclusion in population biology

It is often observed, in nature, that the struggle for existence between two similar species competing for the same limited food supply and living space nearly always ends in the complete extinction of one of the species. This phenomenon is known as the "principle of competitive exclusion." It was first enunciated, in a slightly different form, by Darwin in 1859. In his paper 'The origin of species by natural selection' he writes: "As the species of the same genus usually have, though by no means invariably, much similarity in habits and constitutions and always in structure, the struggle will generally be more severe between them, if they come into competition with each other, than between the species of distinct genera."

There is a very interesting biological explanation of the principle of competitive exclusion. The cornerstone of this theory is the idea of a "niche." A niche indicates what place a given species occupies in a community; i.e., what are its habits, food and mode of life. It has been observed that as a result of competition two similar species rarely occupy the same niche. Rather, each species takes possession of those kinds of food and modes of life in which it has an advantage over its competitor. If the two species tend to occupy the same niche then the struggle for existence between them will be very intense and result in the extinction of the weaker species.

An excellent illustration of this theory is the colony of terns inhabiting the island of Jorilgatch in the Black Sea. This colony consists of four different species of terns: sandwich-tern, common-tern, blackbeak-tern, and little-tern. These four species band together to chase away predators from the colony. However, there is a sharp difference between them as regards the procuring of food. The sandwich-tern flies far out into the open sea to hunt certain species, while the blackbeak-tern feeds exclusively on land. On the other hand, common-tern and little-tern catch fish close to the shore. They sight the fish while flying and dive into the water after them. The little-tern seizes his fish in shallow swampy places, whereas the common-tern hunts somewhat further from shore. In this manner, these four similar species of tern living side by side upon a single small island differ sharply in all their modes of feeding and procuring food. Each has a niche in which it has a distinct advantage over its competitors.

In this section we present a rigorous mathematical proof of the law of competitive exclusion. This will be accomplished by deriving a system of differential equations which govern the interaction between two similar species, and then showing that every solution of the system approaches an equilibrium state in which one of the species is extinct.

In constructing a mathematical model of the struggle for existence between two competing species, it is instructive to look again at the logistic law of population growth

$$\frac{dN}{dt} = aN - bN^2. \tag{1}$$

This equation governs the growth of the population $N(t)$ of a single species whose members compete among themselves for a limited amount of food and living space. Recall (see Section 1.5) that $N(t)$ approaches the limiting population $K = a/b$, as t approaches infinity. This limiting population can be thought of as the maximum population of the species which the microcosm can support. In terms of K, the logistic law (1) can be rewritten in the form

$$\frac{dN}{dt} = aN\left(1 - \frac{b}{a}N\right) = aN\left(1 - \frac{N}{K}\right) = aN\left(\frac{K-N}{K}\right). \tag{2}$$

Equation (2) has the following interesting interpretation. When the population N is very low, it grows according to the Malthusian law $dN/dt = aN$. The term aN is called the "biotic potential" of the species. It is the potential rate of increase of the species under ideal conditions, and it is realized if there are no restrictions on food and living space, and if the individual members of the species do not excrete any toxic waste products. As the population increases though, the biotic potential is reduced by the factor $(K-N)/K$, which is the relative number of still vacant places in the microcosm. Ecologists call this factor the environmental resistance to growth.

Now, let $N_1(t)$ and $N_2(t)$ be the population at time t of species 1 and 2 respectively. Further, let K_1 and K_2 be the maximum population of species 1 and 2 which the microcosm can support, and let a_1N_1 and a_2N_2 be the biotic potentials of species 1 and 2. Then, $N_1(t)$ and $N_2(t)$ satisfy the system of differential equations

$$\frac{dN_1}{dt}=a_1N_1\left(\frac{K_1-N_1-m_2}{K_1}\right),\qquad \frac{dN_2}{dt}=a_2N_2\left(\frac{K_2-N_2-m_1}{K_2}\right),\qquad (3)$$

where m_2 is the total number of places of the first species which are taken up by members of the second species, and m_1 is the total number of places of the second species which are taken up by members of the first species. At first glance it would appear that $m_2=N_2$ and $m_1=N_1$. However, this is not generally the case, for it is highly unlikely that two species utilize the environment in identical ways. Equal numbers of individuals of species 1 and 2 do not, on the average, consume equal quantities of food, take up equal amounts of living space and excrete equal amounts of waste products of the same chemical composition. In general, we must set $m_2=\alpha N_2$ and $m_1=\beta N_1$, for some constants α and β. The constants α and β indicate the degree of influence of one species upon the other. If the interests of the two species do not clash, and they occupy separate niches, then both α and β are zero. If the two species lay claim to the same niche and are very similar, then α and β are very close to one. On the other hand, if one of the species, say species 2, utilizes the environment very unproductively; i.e., it consumes a great deal of food or excretes very poisonous waste products, then one individual of species 2 takes up the place of many individuals of species 1. In this case, then, the coefficient α is very large.

We restrict ourselves now to the case where the two species are nearly identical, and lay claim to the same niche. Then, $\alpha=\beta=1$, and $N_1(t)$ and $N_2(t)$ satisfy the system of differential equations

$$\frac{dN_1}{dt}=a_1N_1\left(\frac{K_1-N_1-N_2}{K_1}\right),\qquad \frac{dN_2}{dt}=a_2N_2\left(\frac{K_2-N_1-N_2}{K_2}\right).\qquad (4)$$

In this instance, we expect the struggle for existence between species 1 and 2 to be very intense, and to result in the extinction of one of the species. This is indeed the case as we now show.

Theorem 6 (Principle of competitive exclusion). *Suppose that K_1 is greater than K_2. Then, every solution $N_1(t)$, $N_2(t)$ of (4) approaches the equilibrium solution $N_1=K_1$, $N_2=0$ as t approaches infinity. In other words, if species 1 and 2 are very nearly identical, and the microcosm can support more members of species 1 than of species 2, then species 2 will ultimately become extinct.*

Our first step in proving Theorem 6 is to show that $N_1(t)$ and $N_2(t)$ can never become negative. To this end, recall from Section 1.5 that

$$N_1(t)=\frac{K_1N_1(0)}{N_1(0)+(K_1-N_1(0))e^{-a_1t}},\qquad N_2(t)=0$$

is a solution of (4) for any choice of $N_1(0)$. The orbit of this solution in the N_1–N_2 plane is the point $(0,0)$ for $N_1(0)=0$; the line $0<N_1<K_1$, $N_2=0$ for $0<N_1(0)<K_1$; the point $(K_1,0)$ for $N_1(0)=K_1$; and the line $K_1<N_1<\infty$, $N_2=0$ for $N_1(0)>K_1$. Thus, the N_1 axis, for $N_1\geqslant 0$, is the union of four distinct orbits. Similarly, the N_2 axis, for $N_2\geqslant 0$, is the union of four distinct orbits of (4). This implies that all solutions $N_1(t)$, $N_2(t)$ of (4) which start in the first quadrant $(N_1>0, N_2>0)$ of the N_1–N_2 plane must remain there for all future time.

Our second step in proving Theorem 6 is to split the first quadrant into regions in which both dN_1/dt and dN_2/dt have fixed signs. This is accomplished in the following manner. Let l_1 and l_2 be the lines $K_1-N_1-N_2=0$ and $K_2-N_1-N_2=0$, respectively. Observe that dN_1/dt is negative if (N_1,N_2) lies above l_1, and positive if (N_1,N_2) lies below l_1. Similarly, dN_2/dt is negative if (N_1,N_2) lies above l_2, and positive if (N_1,N_2) lies below l_2. Thus, the two parallel lines l_1 and l_2 split the first quadrant of the N_1–N_2 plane into three regions (see Figure 1) in which both dN_1/dt and dN_2/dt have fixed signs. Both $N_1(t)$ and $N_2(t)$ increase with time (along any solution of (4)) in region I; $N_1(t)$ increases, and $N_2(t)$ decreases, with time in region II; and both $N_1(t)$ and $N_2(t)$ decrease with time in region III.

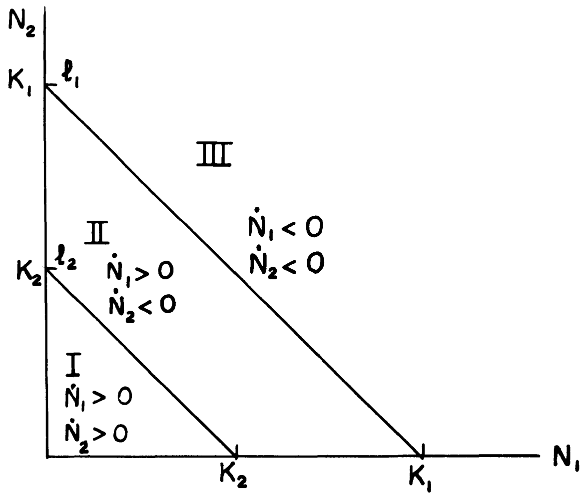


Figure 1

Lemma 1. *Any solution* $N_1(t)$, $N_2(t)$ *of* (4) *which starts in region* I *at* $t=t_0$ *must leave this region at some later time.*

Proof. Suppose that a solution $N_1(t)$, $N_2(t)$ of (4) remains in region I for all time $t\geqslant t_0$. This implies that both $N_1(t)$ and $N_2(t)$ are monotonic increasing functions of time for $t\geqslant t_0$, with $N_1(t)$ and $N_2(t)$ less than K_2. Consequently, by Lemma 1 of Section 4.8, both $N_1(t)$ and $N_2(t)$ have limits

ξ,η respectively, as t approaches infinity. Lemma 2 of Section 4.8 implies that (ξ,η) is an equilibrium point of (4). Now, the only equilibrium points of (4) are $(0,0)$, $(K_1,0)$, and $(0,K_2)$, and (ξ,η) obviously cannot equal any of these three points. We conclude therefore, that any solution $N_1(t)$, $N_2(t)$ of (4) which starts in region I must leave this region at a later time. □

Lemma 2. *Any solution $N_1(t)$, $N_2(t)$ of* (4) *which starts in region* II *at time $t=t_0$ will remain in this region for all future time $t\geqslant t_0$, and ultimately approach the equilibrium solution $N_1=K_1$, $N_2=0$.*

PROOF. Suppose that a solution $N_1(t)$, $N_2(t)$ of (4) leaves region II at time $t=t^*$. Then, either $\dot{N}_1(t^*)$ or $\dot{N}_2(t^*)$ is zero, since the only way a solution of (4) can leave region II is by crossing l_1 or l_2. Assume that $\dot{N}_1(t^*)=0$. Differentiating both sides of the first equation of (4) with respect to t and setting $t=t^*$ gives

$$\frac{d^2N_1(t^*)}{dt^2}=\frac{-a_1N_1(t^*)}{K_1}\frac{dN_2(t^*)}{dt}.$$

This quantity is positive. Hence, $N_1(t)$ has a minimum at $t=t^*$. But this is impossible, since $N_1(t)$ is increasing whenever a solution $N_1(t)$, $N_2(t)$ of (4) is in region II. Similarly, if $\dot{N}_2(t^*)=0$, then

$$\frac{d^2N_2(t^*)}{dt^2}=\frac{-a_2N_2(t^*)}{K_2}\frac{dN_1(t^*)}{dt}.$$

This quantity is negative, implying that $N_2(t)$ has a maximum at $t=t^*$. But this is impossible, since $N_2(t)$ is decreasing whenever a solution $N_1(t)$, $N_2(t)$ of (4) is in region II.

The previous argument shows that any solution $N_1(t)$, $N_2(t)$ of (4) which starts in region II at time $t=t_0$ will remain in region II for all future time $t\geqslant t_0$. This implies that $N_1(t)$ is monotonic increasing and $N_2(t)$ is monotonic decreasing for $t\geqslant t_0$, with $N_1(t)<K_1$ and $N_2(t)>K_2$. Consequently, by Lemma 1 of Section 4.8, both $N_1(t)$ and $N_2(t)$ have limits ξ,η respectively, as t approaches infinity. Lemma 2 of Section 4.8 implies that (ξ,η) is an equilibrium point of (4). Now, (ξ,η) obviously cannot equal $(0,0)$ or $(0,K_2)$. Consequently, $(\xi,\eta)=(K_1,0)$, and this proves Lemma 2. □

Lemma 3. *Any solution $N_1(t)$, $N_2(t)$ of* (4) *which starts in region* III *at time $t=t_0$ and remains there for all future time must approach the equilibrium solution $N_1(t)=K_1$, $N_2(t)=0$ as t approaches infinity.*

PROOF. If a solution $N_1(t)$, $N_2(t)$ of (4) remains in region III for $t\geqslant t_0$, then both $N_1(t)$ and $N_2(t)$ are monotonic decreasing functions of time for $t\geqslant t_0$, with $N_1(t)>0$ and $N_2(t)>0$. Consequently, by Lemma 1 of Section 4.8, both $N_1(t)$ and $N_2(t)$ have limits ξ,η respectively, as t approaches infinity. Lemma 2 of Section 4.8 implies that (ξ,η) is an equilibrium point of (4). Now, (ξ,η) obviously cannot equal $(0,0)$ or $(0,K_2)$. Consequently, $(\xi,\eta)=(K_1,0)$. □

PROOF OF THEOREM 6. Lemmas 1 and 2 above state that every solution $N_1(t)$, $N_2(t)$ of (4) which starts in regions I or II at time $t=t_0$ must approach the equilibrium solution $N_1=K_1$, $N_2=0$ as t approaches infinity. Similarly, Lemma 3 shows that every solution $N_1(t)$, $N_2(t)$ of (4) which starts in region III at time $t=t_0$ and remains there for all future time must also approach the equilibrium solution $N_1=K_1$, $N_2=0$. Next, observe that any solution $N_1(t)$, $N_2(t)$ of (4) which starts on l_1 or l_2 must immediately afterwards enter region II. Finally, if a solution $N_1(t), N_2(t)$ of (4) leaves region III, then it must cross the line l_1 and immediately afterwards enter region II. Lemma 2 then forces this solution to approach the equilibrium solution $N_1=K_1$, $N_2=0$. □

Theorem 6 deals with the case of identical species; i.e., $\alpha=\beta=1$. By a similar analysis (see Exercises 4–6) we can predict the outcome of the struggle for existence for all values of α and β.

Reference

Gause, G. F., 'The Struggle for Existence,' Dover Publications, New York, 1964.

EXERCISES

1. Rewrite the system of equations (4) in the form

$$\frac{K_1}{a_1N_1}\frac{dN_1}{dt}=K_1-N_1-N_2, \qquad \frac{K_2}{a_2N_2}\frac{dN_2}{dt}=K_2-N_1-N_2.$$

Then, subtract these two equations and integrate to obtain directly that $N_2(t)$ approaches zero for all solutions $N_1(t)$, $N_2(t)$ of (4) with $N_1(t_0)>0$.

2. The system of differential equations

$$\begin{aligned}\frac{dN_1}{dt}&=N_1[-a_1+c_1(1-b_1N_1-b_2N_2)]\\ \frac{dN_2}{dt}&=N_2[-a_2+c_2(1-b_1N_1-b_2N_2)]\end{aligned} \tag{*}$$

is a model of two species competing for the same limited resource. Suppose that $c_1>a_1$ and $c_2>a_2$. Deduce from Theorem 6 that $N_1(t)$ ultimately approaches zero if $a_1c_2>a_2c_1$, and $N_2(t)$ ultimately approaches zero if $a_1c_2<a_2c_1$.

3. In 1926, Volterra presented the following model of two species competing for the same limited food supply:

$$\begin{aligned}\frac{dN_1}{dt}&=[b_1-\lambda_1(h_1N_1+h_2N_2)]N_1\\ \frac{dN_2}{dt}&=[b_2-\lambda_2(h_1N_1+h_2N_2)]N_2.\end{aligned}$$

Suppose that $b_1/\lambda_1>b_2/\lambda_2$. (The coefficient b_i/λ_i is called the susceptibility of species i to food shortages.) Prove that species 2 will ultimately become extinct if $N_1(t_0)>0$.

Problems 4–6 are concerned with the system of equations

$$\frac{dN_1}{dt}=\frac{a_1N_1}{K_1}(K_1-N_1-\alpha N_2),\qquad \frac{dN_2}{dt}=\frac{a_2N_2}{K_2}(K_2-N_2-\beta N_1).\quad (*)$$

4. (a) Assume that $K_1/\alpha>K_2$ and $K_2/\beta<K_1$. Show that $N_2(t)$ approaches zero as t approaches infinity for every solution $N_1(t)$, $N_2(t)$ of (*) with $N_1(t_0)>0$.

(b) Assume that $K_1/\alpha<K_2$ and $K_2/\beta>K_1$. Show that $N_1(t)$ approaches zero as t approaches infinity for every solution $N_1(t)$, $N_2(t)$ of (*) with $N_1N_2(t_0)>0$. *Hint*: Draw the lines $l_1: N_1+\alpha N_2=K_1$ and $l_2: N_2+\beta N_1=K_2$, and follow the proof of Theorem 6.

5. Assume that $K_1/\alpha>K_2$ and $K_2/\beta>K_1$. Prove that all solutions $N_1(t)$, $N_2(t)$ of (*), with both $N_1(t_0)$ and $N_2(t_0)$ positive, ultimately approach the equilibrium solution

$$N_1=N_1^0=\frac{K_1-\alpha K_2}{1-\alpha\beta},\qquad N_2=N_2^0=\frac{K_2-\beta K_1}{1-\alpha\beta}.$$

Hint:

(a) Draw the lines $l_1: N_1+\alpha N_2=K_1$ and $l_2: N_2+\beta N_1=K_2$. The two lines divide the first quadrant into four regions (see Figure 2) in which both $\dot N_1$ and $\dot N_2$ have fixed signs.

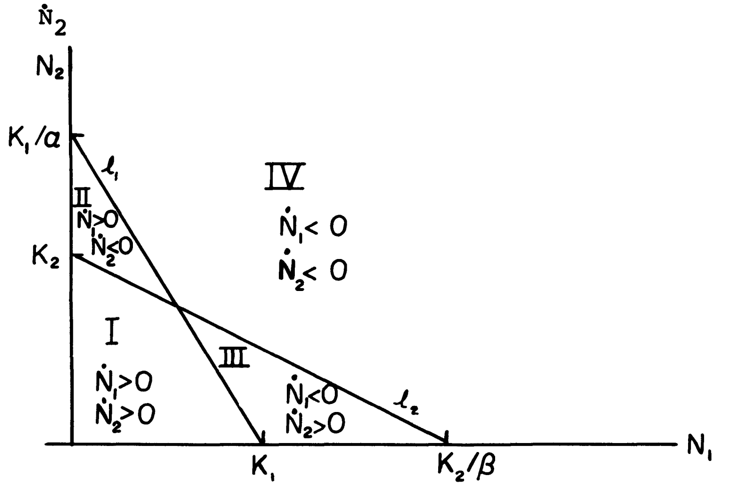


Figure 2

(b) Show that all solutions $N_1(t)$, $N_2(t)$ of (*) which start in either region II or III must remain in these regions and ultimately approach the equilibrium solution $N_1=N_1^0$, $N_2=N_2^0$.

(c) Show that all solutions $N_1(t)$, $N_2(t)$ of (*) which remain exclusively in region I or region IV for all time $t\geqslant t_0$ must ultimately approach the equilibrium solution $N_1=N_1^0$, $N_2=N_2^0$.

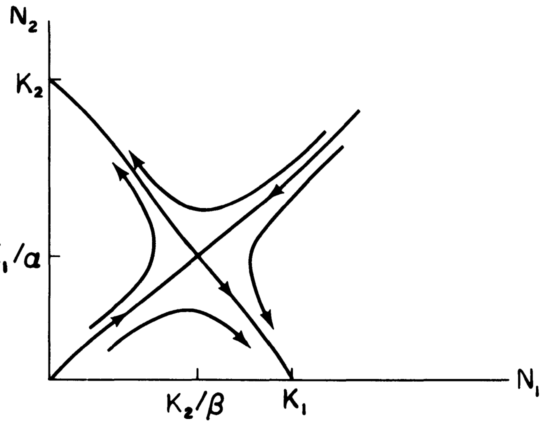


Figure 3

6. Assume that $K_1/\alpha < K_2$ and $K_2/\beta < K_1$.
(a) Show that the equilibrium solution $N_1 = 0$, $N_2 = 0$ of (*) is unstable.
(b) Show that the equilibrium solutions $N_1 = K_1$, $N_2 = 0$ and $N_1 = 0$, $N_2 = K_2$ of (*) are asymptotically stable.
(c) Show that the equilibrium solution $N_1 = N_1^0$, $N_2 = N_2^0$ (see Exercise 5) of (*) is a saddle point. (This calculation is very cumbersome.)
(d) It is not too difficult to see that the phase portrait of (*) must have the form described in Figure 3.

4.12 The Threshold Theorem of epidemiology

Consider the situation where a small group of people having an infectious disease is inserted into a large population which is capable of catching the disease. What happens as time evolves? Will the disease die out rapidly, or will an epidemic occur? How many people will ultimately catch the disease? To answer these questions we will derive a system of differential equations which govern the spread of an infectious disease within a population, and analyze the behavior of its solutions. This approach will also lead us to the famous Threshold Theorem of epidemiology which states that an epidemic will occur only if the number of people who are susceptible to the disease exceeds a certain threshold value.

We begin with the assumptions that the disease under consideration confers permanent immunity upon any individual who has completely recovered from it, and that it has a negligibly short incubation period. This latter assumption implies that an individual who contracts the disease becomes infective immediately afterwards. In this case we can divide the population into three classes of individuals: the infective class (I), the susceptible class (S) and the removed class (R). The infective class consists of those individuals who are capable of transmitting the disease to others.

The susceptible class consists of those individuals who are not infective, but who are capable of catching the disease and becoming infective. The removed class consists of those individuals who have had the disease and are dead, or have recovered and are permanently immune, or are isolated until recovery and permanent immunity occur.

The spread of the disease is presumed to be governed by the following rules.

Rule 1: The population remains at a fixed level N in the time interval under consideration. This means, of course, that we neglect births, deaths from causes unrelated to the disease under consideration, immigration and emigration.

Rule 2: The rate of change of the susceptible population is proportional to the product of the number of members of (S) and the number of members of (I).

Rule 3: Individuals are removed from the infectious class (I) at a rate proportional to the size of (I).

Let $S(t), I(t)$, and $R(t)$ denote the number of individuals in classes (S), (I), and (R), respectively, at time t. It follows immediately from Rules 1–3 that $S(t), I(t), R(t)$ satisfies the system of differential equations

$$\begin{aligned} \frac{dS}{dt} &= -rSI \\ \frac{dI}{dt} &= rSI - \gamma I \\ \frac{dR}{dt} &= \gamma I \end{aligned} \tag{1}$$

for some positive constants r and γ. The proportionality constant r is called the infection rate, and the proportionality constant γ is called the removal rate.

The first two equations of (1) do not depend on R. Thus, we need only consider the system of equations

$$\frac{dS}{dt} = -rSI, \qquad \frac{dI}{dt} = rSI - \gamma I \tag{2}$$

for the two unknown functions $S(t)$ and $I(t)$. Once $S(t)$ and $I(t)$ are known, we can solve for $R(t)$ from the third equation of (1). Alternately, observe that $d(S+I+R)/dt = 0$. Thus,

$$S(t) + I(t) + R(t) = \text{constant} = N$$

so that $R(t) = N - S(t) - I(t)$.

The orbits of (2) are the solution curves of the first-order equation

$$\frac{dI}{dS} = \frac{rSI - \gamma I}{-rSI} = -1 + \frac{\gamma}{rS}. \tag{3}$$

Integrating this differential equation gives

$$I(S) = I_0 + S_0 - S + \rho \ln \frac{S}{S_0}, \tag{4}$$

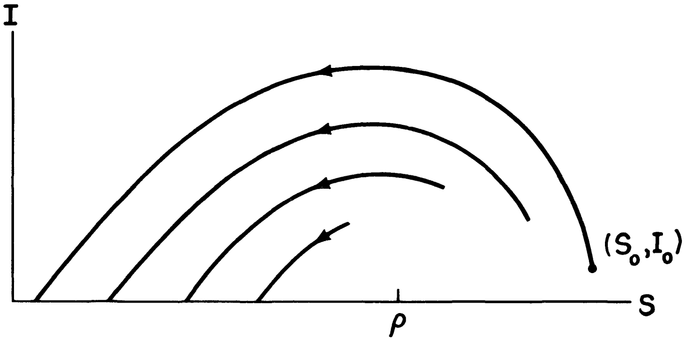


Figure 1. The orbits of (2)

where S_0 and I_0 are the number of susceptibles and infectives at the initial time $t=t_0$, and $\rho=\gamma/r$. To analyze the behavior of the curves (4), we compute $I'(S)=-1+\rho/S$. The quantity $-1+\rho/S$ is negative for $S>\rho$, and positive for $S<\rho$. Hence, $I(S)$ is an increasing function of S for $S<\rho$, and a decreasing function of S for $S>\rho$.

Next, observe that $I(0)=-\infty$ and $I(S_0)=I_0>0$. Consequently, there exists a unique point S_∞, with $0<S_\infty<S_0$, such that $I(S_\infty)=0$, and $I(S)>0$ for $S_\infty<S\leqslant S_0$. The point $(S_\infty,0)$ is an equilibrium point of (2) since both dS/dt and dI/dt vanish when $I=0$. Thus, the orbits of (2), for $t_0\leqslant t<\infty$, have the form described in Figure 1.

Let us see what all this implies about the spread of the disease within the population. As t runs from t_0 to ∞, the point $(S(t),I(t))$ travels along the curve (4), and it moves along the curve in the direction of decreasing S, since $S(t)$ decreases monotonically with time. Consequently, if S_0 is less than ρ, then $I(t)$ decreases monotonically to zero, and $S(t)$ decreases monotonically to S_∞. Thus, if a small group of infectives I_0 is inserted into a group of susceptibles S_0, with $S_0<\rho$, then the disease will die out rapidly. On the other hand, if S_0 is greater than ρ, then $I(t)$ increases as $S(t)$ decreases to ρ, and it achieves a maximum value when $S=\rho$. It only starts decreasing when the number of susceptibles falls below the threshold value ρ. From these results we may draw the following conclusions.

Conclusion 1: An epidemic will occur only if the number of susceptibles in a population exceeds the threshold value $\rho=\gamma/r$.

Conclusion 2: The spread of the disease does not stop for lack of a susceptible population; it stops only for lack of infectives. In particular, some individuals will escape the disease altogether.

Conclusion 1 corresponds to the general observation that epidemics tend to build up more rapidly when the density of susceptibles is high due to overcrowding, and the removal rate is low because of ignorance, inadequate isolation and inadequate medical care. On the other hand, outbreaks tend to be of only limited extent when good social conditions entail lower densities of susceptibles, and when removal rates are high because of good public health vigilance and control.

If the number of susceptibles S_0 is initially greater than, but close to, the threshold value ρ, then we can estimate the number of individuals who ultimately contract the disease. Specifically, if $S_0-\rho$ is small compared to ρ, then the number of individuals who ultimately contract the disease is approximately $2(S_0-\rho)$. This is the famous Threshold Theorem of epidemiology, which was first proven in 1927 by the mathematical biologists Kermack and McKendrick.

Theorem 7 (Threshold Theorem of epidemiology). *Let $S_0=\rho+\nu$ and assume that ν/ρ is very small compared to one. Assume moreover, that the number of initial infectives I_0 is very small. Then, the number of individuals who ultimately contract the disease is 2ν. In other words, the level of susceptibles is reduced to a point as far below the threshold as it originally was above it.*

PROOF. Letting t approach infinity in (4) gives

$$0=I_0+S_0-S_\infty+\rho\ln\frac{S_\infty}{S_0}.$$

If I_0 is very small compared to S_0, then we can neglect it, and write

$$\begin{aligned}0&=S_0-S_\infty+\rho\ln\frac{S_\infty}{S_0}\\ &=S_0-S_\infty+\rho\ln\left[\frac{S_0-(S_0-S_\infty)}{S_0}\right]\\ &=S_0-S_\infty+\rho\ln\left[1-\left(\frac{S_0-S_\infty}{S_0}\right)\right].\end{aligned}$$

Now, if $S_0-\rho$ is small compared to ρ, then S_0-S_∞ will be small compared to S_0. Consequently, we can truncate the Taylor series

$$\ln\left[1-\left(\frac{S_0-S_\infty}{S_0}\right)\right]=-\left(\frac{S_0-S_\infty}{S_0}\right)-\frac{1}{2}\left(\frac{S_0-S_\infty}{S_0}\right)^2+\dots$$

after two terms. Then,

$$\begin{aligned}0&=S_0-S_\infty-\rho\left(\frac{S_0-S_\infty}{S_0}\right)-\frac{\rho}{2}\left(\frac{S_0-S_\infty}{S_0}\right)^2\\ &=(S_0-S_\infty)\left[1-\frac{\rho}{S_0}-\frac{\rho}{2S_0^2}(S_0-S_\infty)\right].\end{aligned}$$

Solving for $S_0 - S_\infty$, we see that

$$S_0 - S_\infty = 2S_0\left(\frac{S_0}{\rho} - 1\right) = 2(\rho+\nu)\left[\frac{\rho+\nu}{\rho} - 1\right]$$

$$= 2(\rho+\nu)\frac{\nu}{\rho} = 2\rho\left(1+\frac{\nu}{\rho}\right)\frac{\nu}{\rho} \cong 2\nu. \qquad \square$$

During the course of an epidemic it is impossible to accurately ascertain the number of new infectives each day or week, since the only infectives who can be recognized and removed from circulation are those who seek medical aid. Public health statistics thus record only the number of new removals each day or week, not the number of new infectives. Therefore, in order to compare the results predicted by our model with data from actual epidemics, we must find the quantity dR/dt as a function of time. This is accomplished in the following manner. Observe first that

$$\frac{dR}{dt} = \gamma I = \gamma(N - R - S).$$

Second, observe that

$$\frac{dS}{dR} = \frac{dS/dt}{dR/dt} = \frac{-rSI}{\gamma I} = \frac{-S}{\rho}.$$

Hence, $S(R) = S_0 e^{-R/\rho}$ and

$$\frac{dR}{dt} = \gamma\left(N - R - S_0 e^{-R/\rho}\right). \tag{5}$$

Equation (5) is separable, but cannot be solved explicitly. However, if the epidemic is not very large, then R/ρ is small and we can truncate the Taylor series

$$e^{-R/\rho} = 1 - \frac{R}{\rho} + \frac{1}{2}\left(\frac{R}{\rho}\right)^2 + \dots$$

after three terms. With this approximation,

$$\frac{dR}{dt} = \gamma\left[N - R - S_0\left[1 - R/\rho + \frac{1}{2}(R/\rho)^2\right]\right]$$

$$= \gamma\left[N - S_0 + \left(\frac{S_0}{\rho} - 1\right)R - \frac{S_0}{2}\left(\frac{R}{\rho}\right)^2\right].$$

The solution of this equation is

$$R(t) = \frac{\rho^2}{S_0}\left[\frac{S_0}{\rho} - 1 + \alpha\tanh\left(\frac{1}{2}\alpha\gamma t - \phi\right)\right] \tag{6}$$

where

$$\alpha=\left[\left(\frac{S_0}{\rho}-1\right)^2+\frac{2S_0(N-S_0)}{\rho^2}\right]^{1/2}, \qquad \phi=\tanh^{-1}\frac{1}{\alpha}\left(\frac{S_0}{\rho}-1\right)$$

and the hyperbolic tangent function $\tanh z$ is defined by

$$\tanh z=\frac{e^z-e^{-z}}{e^z+e^{-z}}.$$

It is easily verified that

$$\frac{d}{dz}\tanh z=\operatorname{sech}^2 z=\frac{4}{(e^z+e^{-z})^2}.$$

Hence,

$$\frac{dR}{dt}=\frac{\gamma\alpha^2\rho^2}{2S_0}\operatorname{sech}^2\left(\frac{1}{2}\alpha\gamma t-\phi\right). \tag{7}$$

Equation (7) defines a symmetric bell shaped curve in the t–dR/dt plane (see Figure 2). This curve is called the epidemic curve of the disease. It illustrates very well the common observation that in many actual epidemics, the number of new cases reported each day climbs to a peak value and then dies away again.

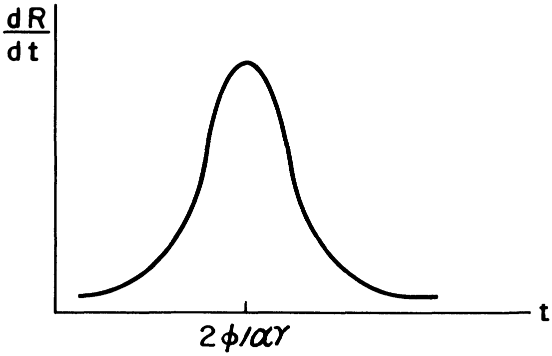


Figure 2

Kermack and McKendrick compared the values predicted for dR/dt from (7) with data from an actual plague in Bombay which spanned the last half of 1905 and the first half of 1906. They set

$$\frac{dR}{dt}=890\operatorname{sech}^2(0.2t-3.4)$$

with t measured in weeks, and compared these values with the number of deaths per week from the plague. This quantity is a very good approximation of dR/dt, since almost all cases terminated fatally. As can be seen from Figure 3, there is excellent agreement between the actual values of

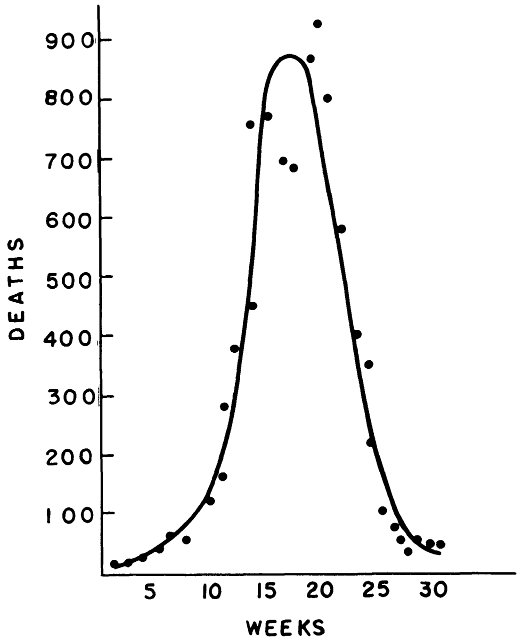


Figure 3

dR/dt, denoted by •, and the values predicted by (7). This indicates, of course, that the system of differential equations (1) is an accurate and reliable model of the spread of an infectious disease within a population of fixed size.

References

Bailey, N. T. J., 'The mathematical theory of epidemics,' 1957, New York.

Kermack, W. O. and McKendrick, A. G., Contributions to the mathematical theory of epidemics, Proceedings Roy. Stat. Soc., A, 115, 700–721, 1927.

Waltman, P., 'Deterministic threshold models in the theory of epidemics,' Springer-Verlag, New York, 1974.

Exercises

1. Derive Equation (6).

2. Suppose that the members of (S) are vaccinated against the disease at a rate λ proportional to their number. Then,

$$\frac{dS}{dt} = -rSI - \lambda S, \qquad \frac{dI}{dt} = rSI - \gamma I. \tag{*}$$

(a) Find the orbits of (*).
(b) Conclude from (a) that $S(t)$ approaches zero as t approaches infinity, for every solution $S(t)$, $I(t)$ of (*).

3. Suppose that the members of (S) are vaccinated against the disease at a rate λ proportional to the product of their numbers and the square of the members of (I). Then,

$$\frac{dS}{dt} = -rSI - \lambda SI^2, \qquad \frac{dI}{dt} = I(rS - \gamma). \tag{*}$$

(a) Find the orbits of (*).
(b) Will any susceptibles remain after the disease dies out?

4. The *intensity* i of an epidemic is the proportion of the total number of susceptibles that finally contracts the disease. Show that

$$i = \frac{I_0 + S_0 - S_\infty}{S_0}$$

where S_∞ is a root of the equation

$$S = S_0 e^{(S - S_0 - I_0)/\rho}.$$

5. Compute the intensity of the epidemic if $\rho = 1000$, $I_0 = 10$, and (a) $S_0 = 1100$, (b) $S_0 = 1200$, (c) $S_0 = 1300$, (d) $S_0 = 1500$, (e) $S_0 = 1800$, (f) $S_0 = 1900$. (This cannot be done analytically.)

6. Let R_∞ denote the total number of individuals who contract the disease.
(a) Show that $R_\infty = I_0 + S_0 - S_\infty$.
(b) Let R_1 denote the members of (R) who are removed from the population prior to the peak of the epidemic. Compute R_1/R_∞ for each of the values of S_0 in 5a–5f. Notice that most of the removals occur after the peak. This type of asymmetry is often found in actual notifications of infectious diseases.

7. It was observed in London during the early 1900's, that large outbreaks of measles epidemics recurred about once every two years. The mathematical biologist H. E. Soper tried to explain this phenomenon by assuming that the stock of susceptibles is constantly replenished by new recruits to the population. Thus, he assumed that

$$\frac{dS}{dt} = -rSI + \mu, \qquad \frac{dI}{dt} = rSI - \gamma I \tag{*}$$

for some positive constants r, γ, and μ.
(a) Show that $S = \gamma/r$, $I = \mu/\gamma$ is the only equilibrium solution of (*).
(b) Show that every solution $S(t)$, $I(t)$ of (*) which starts sufficiently close to this equilibrium point must ultimately approach it as t approaches infinity.
(c) It can be shown that every solution $S(t)$, $I(t)$ of (*) approaches the equilibrium solution $S = \gamma/r$, $I = \mu/\gamma$ as t approaches infinity. Conclude, therefore, that the system (*) does not predict recurrent outbreaks of measles epidemics. Rather, it predicts that the disease will ultimately approach a steady state.

4.13 A model for the spread of gonorrhea

Gonorrhea ranks first today among reportable communicable diseases in the United States. There are more reported cases of gonorrhea every year than the combined totals for syphilis, measles, mumps, and infectious hepatitis. Public health officials estimate that more than 2,500,000 Americans contract gonorrhea every year. This painful and dangerous disease, which is caused by the gonococcus germ, is spread from person to person by sexual contact. A few days after the infection there is usually itching and burning of the genital area, particularly while urinating. About the same time a discharge develops which males will notice, but which females may not notice. Infected women may have no easily recognizable symptoms, even while the disease does substantial internal damage. Gonorrhea can only be cured by antibiotics (usually penicillin). However, treatment must be given early if the disease is to be stopped from doing serious damage to the body. If untreated, gonorrhea can result in blindness, sterility, arthritis, heart failure, and ultimately, death.

In this section we construct a mathematical model of the spread of gonorrhea. Our work is greatly simplified by the fact that the incubation period of gonorrhea is very short (3–7 days) compared to the often quite long period of active infectiousness. Thus, we will assume in our model that an individual becomes infective immediately after contracting gonorrhea. In addition, gonorrhea does not confer even partial immunity to those individuals who have recovered from it. Immediately after recovery, an individual is again susceptible. Thus, we can split the sexually active and promiscuous portion of the population into two groups, susceptibles and infectives. Let $c_1(t)$ be the total number of promiscuous males, $c_2(t)$ the total number of promiscuous females, $x(t)$ the total number of infective males, and $y(t)$ the total number of infective females, at time t. Then, the total numbers of susceptible males and susceptible females are $c_1(t)-x(t)$ and $c_2(t)-y(t)$ respectively. The spread of gonorrhea is presumed to be governed by the following rules:

1. Male infectives are cured at a rate a_1 proportional to their total number, and female infectives are cured at a rate a_2 proportional to their total number. The constant a_1 is larger than a_2 since infective males quickly develop painful symptoms and therefore seek prompt medical attention. Female infectives, on the other hand, are usually asymptomatic, and therefore are infectious for much longer periods.

2. New infectives are added to the male population at a rate b_1 proportional to the total number of male susceptibles and female infectives. Similarly, new infectives are added to the female population at a rate b_2 proportional to the total number of female susceptibles and male infectives.

3. The total numbers of promiscuous males and promiscuous females remain at constant levels c_1 and c_2, respectively.

It follows immediately from rules 1–3 that

$$\begin{aligned}\frac{dx}{dt}&=-a_1x+b_1(c_1-x)y\\ \frac{dy}{dt}&=-a_2y+b_2(c_2-y)x.\end{aligned}\tag{1}$$

Remark. The system of equations (1) treats only those cases of gonorrhea which arise from heterosexual contacts; the case of homosexual contacts (assuming no interaction between heterosexuals and homosexuals) is treated in Exercises 5 and 6. The number of cases of gonorrhea which arise from homosexual encounters is a small percentage of the total number of incidents of gonorrhea. Interestingly enough, this situation is completely reversed in the case of syphilis. Indeed, more than 90% of all cases of syphilis reported in the state of Rhode Island during 1973 resulted from homosexual encounters. (This statistic is not as startling as it first appears. Within ten to ninety days after being infected with syphilis, an individual usually develops a chancre sore at the spot where the germs entered the body. A homosexual who contracts syphilis as a result of anal intercourse with an infective will develop a chancre sore on his rectum. This individual, naturally, will be reluctant to seek medical attention, since he will then have to reveal his identity as a homosexual. Moreover, he feels no sense of urgency, since the chancre sore is usually painless and disappears after several days. With gonorrhea, on the other hand, the symptoms are so painful and unmistakable that a homosexual will seek prompt medical attention. Moreover, he need not reveal his identity as a homosexual since the symptoms of gonorrhea appear in the genital area.)

Our first step in analyzing the system of differential equations (1) is to show that they are realistic. Specifically, we must show that $x(t)$ and $y(t)$ can never become negative, and can never exceed c_1 and c_2, respectively. This is the content of Lemmas 1 and 2.

Lemma 1. *If $x(t_0)$ and $y(t_0)$ are positive, then $x(t)$ and $y(t)$ are positive for all $t \geqslant t_0$.*

Lemma 2. *If $x(t_0)$ is less than c_1 and $y(t_0)$ is less than c_2, then $x(t)$ is less than c_1 and $y(t)$ is less than c_2 for all $t \geqslant t_0$.*

Proof of Lemma 1. Suppose that Lemma 1 is false. Let $t^* > t_0$ be the first time at which either x or y is zero. Assume that x is zero first. Then, evaluating the first equation of (1) at $t = t^*$ gives $\dot{x}(t^*) = b_1 c_1 y(t^*)$. This quantity is positive. (Note that $y(t^*)$ cannot equal zero since $x = 0$, $y = 0$ is an equilibrium solution of (1).) Hence, $x(t)$ is less than zero for t close to, and less than t^*. But this contradicts our assumption that t^* is the first time at which $x(t)$ equals zero. We run into the same contradiction if $y(t^*) = 0$. Thus, both $x(t)$ and $y(t)$ are positive for $t \geqslant t_0$. □

Proof of Lemma 2. Suppose that Lemma 2 is false. Let $t^* > t_0$ be the first time at which either $x = c_1$, or $y = c_2$. Suppose that $x(t^*) = c_1$. Evaluating the first equation of (1) at $t = t^*$ gives $\dot{x}(t^*) = -a_1 c_1$. This quantity is negative. Hence, $x(t)$ is greater than c_1 for t close to, and less than t^*. But this contradicts our assumption that t^* is the first time at which $x(t)$ equals c_1. We run into the same contradiction if $y(t^*)=c_2$. Thus, $x(t)$ is less than c_1 and $y(t)$ is less than c_2 for $t \geqslant t_0$. □

Having shown that the system of equations (1) is a realistic model of gonorrhea, we now see what predictions it makes concerning the future course of this disease. Will gonorrhea continue to spread rapidly and uncontrollably as the data in Figure 1 seems to suggest, or will it level off eventually? The following extremely important theorem of epidemiology provides the answer to this question.

Theorem 8.

(a) *Suppose that a_1a_2 is less than $b_1b_2c_1c_2$. Then, every solution $x(t)$, $y(t)$ of (1) with $0<x(t_0)<c_1$ and $0<y(t_0)<c_2$, approaches the equilibrium solution*

$$x=\frac{b_1b_2c_1c_2-a_1a_2}{a_1b_2+b_1b_2c_2}, \qquad y=\frac{b_1b_2c_1c_2-a_1a_2}{a_2b_1+b_1b_2c_1}$$

as t approaches infinity. In other words, the total numbers of infective males and infective females will ultimately level off.

(b) *Suppose that a_1a_2 is greater than $b_1b_2c_1c_2$. Then every solution $x(t)$, $y(t)$ of (1) with $0<x(t_0)<c_1$ and $0<y(t_0)<c_2$, approaches zero as t approaches infinity. In other words, gonorrhea will ultimately die out.*

Our first step in proving part (a) of Theorem 8 is to split the rectangle $0<x<c_1$, $0<y<c_2$ into regions in which both dx/dt and dy/dt have fixed signs. This is accomplished in the following manner. Setting $dx/dt=0$ in (1), and solving for y as a function of x gives

$$y=\frac{a_1x}{b_1(c_1-x)}\equiv\phi_1(x).$$

Similarly, setting $dy/dt=0$ in (1) gives

$$x=\frac{a_2y}{b_2(c_2-y)}, \quad \text{or} \quad y=\frac{b_2c_2x}{a_2+b_2x}\equiv\phi_2(x).$$

Observe first that $\phi_1(x)$ and $\phi_2(x)$ are monotonic increasing functions of x; $\phi_1(x)$ approaches infinity as x approaches c_1, and $\phi_2(x)$ approaches c_2 as x approaches infinity. Second, observe that the curves $y=\phi_1(x)$ and $y=\phi_2(x)$ intersect at $(0,0)$ and at (x_0, y_0) where

$$x_0=\frac{b_1b_2c_1c_2-a_1a_2}{a_1b_2+b_1b_2c_2}, \qquad y_0=\frac{b_1b_2c_1c_2-a_1a_2}{a_2b_1+b_1b_2c_1}.$$

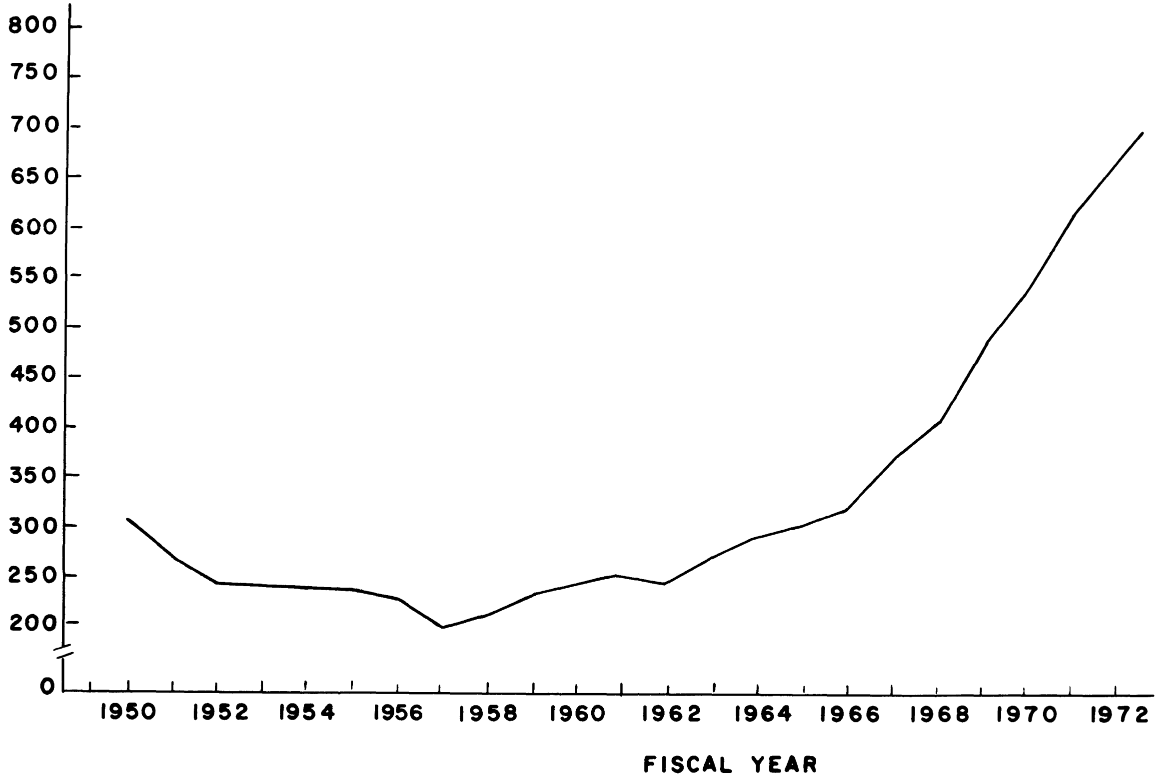


Figure 1. Reported cases of gonorrhea, in thousands, for 1950–1973.

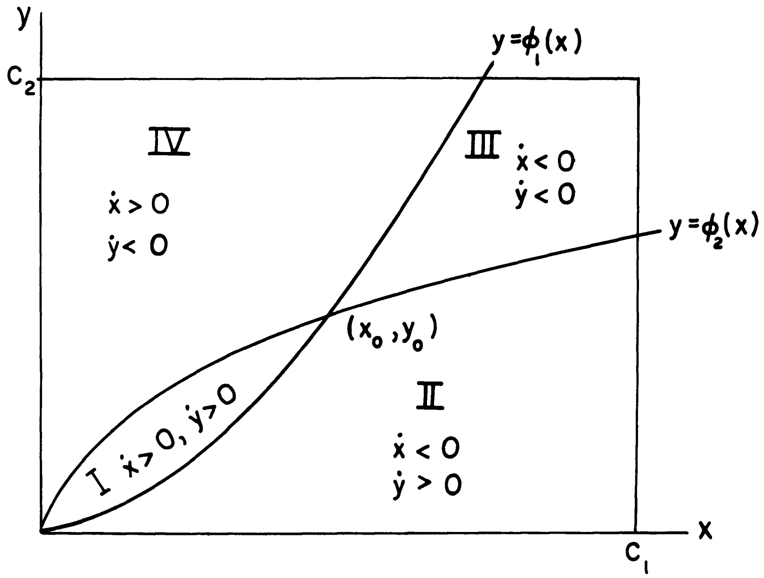


Figure 2

Third, observe that $\phi_2(x)$ is increasing faster than $\phi_1(x)$ at $x=0$, since

$$\phi_2'(0)=\frac{b_2c_2}{a_2}>\frac{a_1}{b_1c_1}=\phi_1'(0).$$

Hence, $\phi_2(x)$ lies above $\phi_1(x)$ for $0<x<x_0$, and $\phi_2(x)$ lies below $\phi_1(x)$ for $x_0<x<c_1$, as shown in Figure 2. The point (x_0,y_0) is an equilibrium point of (1) since both dx/dt and dy/dt are zero when $x=x_0$ and $y=y_0$.

Finally, observe that dx/dt is positive at any point (x,y) above the curve $y=\phi_1(x)$, and negative at any point (x,y) below this curve. Similarly, dy/dt is positive at any point (x,y) below the curve $y=\phi_2(x)$, and negative at any point (x,y) above this curve. Thus, the curves $y=\phi_1(x)$ and $y=\phi_2(x)$ split the rectangle $0<x<c_1$, $0<y<c_2$ into four regions in which dx/dt and dy/dt have fixed signs (see Figure 2).

Next, we require the following four simple lemmas.

Lemma 3. *Any solution $x(t),y(t)$ of* (1) *which starts in region* I *at time $t=t_0$ will remain in this region for all future time $t\geqslant t_0$ and approach the equilibrium solution $x=x_0$, $y=y_0$ as t approaches infinity.*

Proof. Suppose that a solution $x(t),y(t)$ of (1) leaves region I at time $t=t^*$. Then, either $\dot{x}(t^*)$ or $\dot{y}(t^*)$ is zero, since the only way a solution of (1) can leave region I is by crossing the curve $y=\phi_1(x)$ or $y=\phi_2(x)$. Assume that $\dot{x}(t^*)=0$. Differentiating both sides of the first equation of (1) with respect to t and setting $t=t^*$ gives

$$\frac{d^2x(t^*)}{dt^2}=b_1(c_1-x(t^*))\frac{dy(t^*)}{dt}.$$

This quantity is positive, since $x(t^*)$ is less than c_1, and dy/dt is positive on the curve $y=\phi_1(x)$, $0<x<x_0$. Hence, $x(t)$ has a minimum at $t=t^*$. But this is impossible, since $x(t)$ is increasing whenever the solution $x(t),y(t)$ is in region I. Similarly, if $\dot{y}(t^*)=0$, then

$$\frac{d^2y(t^*)}{dt^2}=b_2(c_2-y(t^*))\frac{dx(t^*)}{dt}.$$

This quantity is positive, since $y(t^*)$ is less than c_2, and dx/dt is positive on the curve $y=\phi_2(x)$, $0<x<x_0$. Hence, $y(t)$ has a minimum at $t=t^*$. But this is impossible, since $y(t)$ is increasing whenever the solution $x(t),y(t)$ is in region I.

The previous argument shows that any solution $x(t),y(t)$ of (1) which starts in region I at time $t=t_0$ will remain in region I for all future time $t\geqslant t_0$. This implies that $x(t)$ and $y(t)$ are monotonic increasing functions of time for $t\geqslant t_0$, with $x(t)<x_0$ and $y(t)<y_0$. Consequently, by Lemma 1 of Section 4.8, both $x(t)$ and $y(t)$ have limits ξ,η, respectively, as t approaches infinity. Lemma 2 of Section 4.8 implies that (ξ,η) is an equilibrium point of (1). Now, it is easily seen from Figure 2 that the only equilibrium points of (1) are $(0,0)$ and (x_0, y_0). But (ξ,η) cannot equal $(0,0)$ since both $x(t)$ and $y(t)$ are increasing functions of time. Hence, $(\xi,\eta)=(x_0, y_0)$, and this proves Lemma 3. □

Lemma 4. *Any solution $x(t),y(t)$ of* (1) *which starts in region* III *at time $t=t_0$ will remain in this region for all future time and ultimately approach the equilibrium solution $x=x_0, y=y_0$.*

PROOF. Exactly the same as Lemma 3 (see Exercise 1). □

Lemma 5. *Any solution $x(t),y(t)$ of* (1) *which starts in region* II *at time $t=t_0$, and remains in region* II *for all future time, must approach the equilibrium solution $x=x_0, y=y_0$ as t approaches infinity.*

PROOF. If a solution $x(t)$, $y(t)$ of (1) remains in region II for $t\geqslant t_0$, then $x(t)$ is monotonic decreasing and $y(t)$ is monotonic increasing for $t\geqslant t_0$. Moreover, $x(t)$ is positive and $y(t)$ is less than c_2, for $t\geqslant t_0$. Consequently, by Lemma 1 of Section 4.8, both $x(t)$ and $y(t)$ have limits ξ,η respectively, as t approaches infinity. Lemma 2 of Section 4.8 implies that (ξ,η) is an equilibrium point of (1). Now, (ξ,η) cannot equal $(0,0)$ since $y(t)$ is increasing for $t\geqslant t_0$. Therefore, $(\xi,\eta)=(x_0,y_0)$, and this proves Lemma 5. □

Lemma 6. *Any solution $x(t), y(t)$ of* (1) *which starts in region* IV *at time $t=t_0$ and remains in region* IV *for all future time, must approach the equilibrium solution $x=x_0, y=y_0$ as t approaches infinity.*

PROOF. Exactly the same as Lemma 5 (see Exercise 2). □

We are now in a position to prove Theorem 8.

PROOF OF THEOREM 8. (a) Lemmas 3 and 4 state that every solution $x(t)$, $y(t)$ of (1) which starts in region I or III at time $t=t_0$ must approach the equilibrium solution $x=x_0$, $y=y_0$ as t approaches infinity. Similarly, Lemmas 5 and 6 state that every solution $x(t)$, $y(t)$ of (1) which starts in region II or IV and which remains in these regions for all future time, must also approach the equilibrium solution $x=x_0, y=y_0$. Now, observe that if a solution $x(t)$, $y(t)$ of (1) leaves region II or IV, then it must cross the curve $y=\phi_1(x)$ or $y=\phi_2(x)$, and immediately afterwards enter region I or region III. Consequently, all solutions $x(t)$, $y(t)$ of (1) which start in regions II and IV or on the curves $y=\phi_1(x)$ and $y=\phi_2(x)$, must also approach the equilibrium solution $x(t)=x_0, y(t)=y_0$. □

(b) PROOF #1. If a_1a_2 is greater than $b_1b_2c_1c_2$, then the curves $y=\phi_1(x)$ and $y=\phi_2(x)$ have the form described in Figure 3 below. In region I, dx/dt is positive and dy/dt is negative; in region II, both dx/dt and dy/dt are negative; and in region III, dx/dt is negative and dy/dt is positive. It is a simple matter to show (see Exercise 3) that every solution $x(t), y(t)$ of (1) which starts in region II at time $t=t_0$ must remain in this region for all

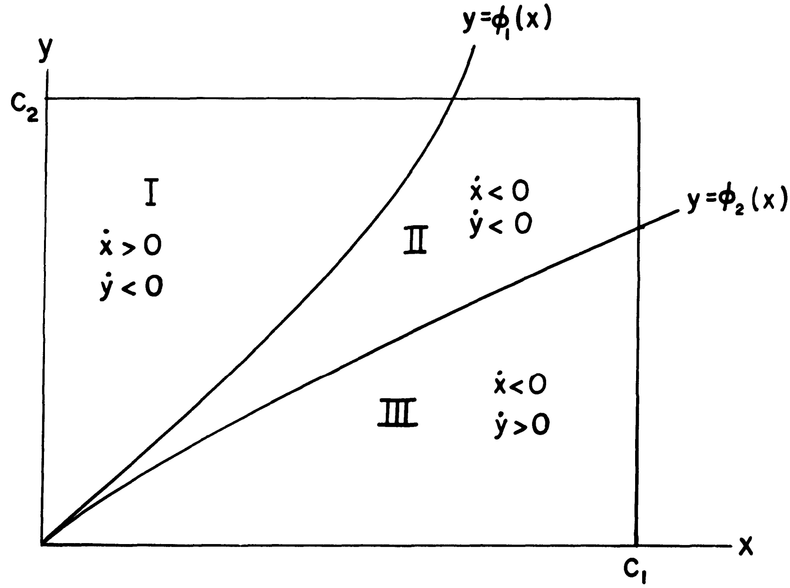


Figure 3

future time, and approach the equilibrium solution $x=0$, $y=0$ as t approaches infinity. It is also trivial to show that every solution $x(t), y(t)$ of (1) which starts in region I or region III at time $t=t_0$ must cross the curve $y=\phi_1(x)$ or $y=\phi_2(x)$, and immediately afterwards enter region II (see Exercise 4). Consequently, every solution $x(t), y(t)$ of (1), with $0<x(t_0)<c_1$ and $0<y(t_0)<c_2$, approaches the equilibrium solution $x=0, y=0$ as t approaches infinity. □

Proof #2. We would now like to show how we can use the Poincaré–Bendixson theorem to give an elegant proof of part (b) of Theorem 8. Observe that the system of differential equations (1) can be written in the form

$$\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -a_1 & b_1c_1 \\ b_2c_2 & -a_2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} b_1xy \\ b_2xy \end{pmatrix}. \tag{2}$$

Thus, by Theorem 2 of Section 4.3, the stability of the solution $x=0, y=0$ of (2) is determined by the stability of the equilibrium solution $x=0, y=0$ of the linearized system

$$\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \mathbf{A}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -a_1 & b_1c_1 \\ b_2c_2 & -a_2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}.$$

The characteristic polynomial of the matrix $\mathbf{A}$ is

$$\lambda^2 + (a_1+a_2)\lambda + a_1a_2 - b_1b_2c_1c_2$$

whose roots are

$$\lambda = \frac{-(a_1+a_2) \pm \left[(a_1+a_2)^2 - 4(a_1a_2 - b_1b_2c_1c_2)\right]^{1/2}}{2}.$$

It is easily verified that both these roots are real and negative. Hence, the equilibrium solution $x=0, y=0$ of (2) is asymptotically stable. This implies that any solution $x(t)$, $y(t)$ of (1) which starts sufficiently close to the origin $x=y=0$ will approach the origin as t approaches infinity. Now, suppose that a solution $x(t), y(t)$ of (1), with $0<x(t_0)<c_1$ and $0<y(t_0)<c_2$, does not approach the origin as t approaches infinity. By the previous remark, this solution must always remain a minimum distance from the origin. Consequently, its orbit for $t \geqslant t_0$ lies in a bounded region in the $x-y$ plane which contains no equilibrium points of (1). By the Poincaré–Bendixson Theorem, therefore, its orbit must spiral into the orbit of a periodic solution of (1). But the system of differential equations (1) has no periodic solution in the first quadrant $x \geqslant 0$, $y \geqslant 0$. This follows immediately from Exercise 11, Section 4.8, and the fact that

$$\frac{\partial}{\partial x}\left[-a_1x + b_1(c_1-x)y\right] + \frac{\partial}{\partial y}\left[-a_2y + b_2(c_2-y)x\right]$$
$$= -(a_1 + a_2 + b_1y + b_2x)$$

is strictly negative if both x and y are nonnegative. Consequently, every solution $x(t)$, $y(t)$ of (1), with $0<x(t_0)<c_1$ and $0<y(t_0)<c_2$ approaches the equilibrium solution $x=0, y=0$ as t approaches infinity. □

Now, it is quite difficult to evaluate the coefficients a_1, a_2, b_1, o_2, c_1, and c_2. Indeed, it is impossible to obtain even a crude estimate of a_2, which should be interpreted as the average amount of time that a female remains infective. (Similarly, a_1 should be interpreted as the average amount of time that a male remains infective.) This is because most females do not exhibit symptoms. Thus, a female can be infective for an amount of time varying from just one day to well over a year. Nevertheless, it is still possible to ascertain from public health data that a_1a_2 is less than $b_1b_2c_1c_2$, as we now show. Observe that the condition $a_1a_2<b_1b_2c_1c_2$ is equivalent to

$$1<\left(\frac{b_1c_1}{a_2}\right)\left(\frac{b_2c_2}{a_1}\right).$$

The quantity b_1c_1/a_2 can be interpreted as the average number of males that one female infective contacts during her infectious period, if every male is susceptible. Similarly, the quantity b_2c_2/a_1 can be interpreted as the average number of females that one male infective contacts during his infectious period, if every female is susceptible. The quantities b_1c_1/a_2 and b_2c_2/a_1 are called the maximal female and male contact rates, respectively. Theorem 8 can now be interpreted in the following manner.

(a) If the product of the maximal male and female contact rates is greater than one, then gonorrhea will approach a nonzero steady state.
(b) If the product of the maximal male and female contact rates is less than one, then gonorrhea will die out eventually.

In 1973, the average number of female contacts named by a male infective during his period of infectiousness was 0.98, while the average number of male contacts named by a female infective during her period of infectiousness was 1.15. These numbers are very good approximations of the maximal male and female contact rates, respectively, and their product does not exceed the product of the maximal male and female contact rates. (The number of contacts of a male or female infective during their period of infectiousness is slightly less than the maximal male or female contact rates. However, the *actual* number of contacts is often greater than the number of contacts named by an infective.) The product of 1.15 with 0.98 is 1.0682. Thus, gonorrhea will ultimately approach a nonzero steady state.

Remark. Our model of gonorrhea is rather crude since it lumps all promiscuous males and all promiscuous females together, regardless of age. A more accurate model can be obtained by separating the male and female populations into different age groups and then computing the rate of change of infectives in each age group. This has been done recently, but the analysis is too difficult to present here. We just mention that a result completely analogous to Theorem 8 is obtained: either gonorrhea dies out in each age group, or it approaches a constant, positive level in each age group.

Exercises

In Problems 1 and 2, we assume that $a_1a_2 < b_1b_2c_1c_2$.

1. (a) Suppose that a solution $x(t), y(t)$ of (1) leaves region III of Figure 2 at time $t = t^*$ by crossing the curve $y = \phi_1(x)$ or $y = \phi_2(x)$. Conclude that either $x(t)$ or $y(t)$ has a maximum at $t = t^*$. Then, show that this is impossible. Conclude, therefore, that any solution $x(t), y(t)$ of (1) which starts in region III at time $t = t_0$ must remain in region III for all future time $t > t_0$.
   (b) Conclude from (a) that any solution $x(t), y(t)$ of (1) which starts in region III has a limit ξ, η as t approaches infinity. Then, show that (ξ, η) must equal (x_0, y_0).

2. Suppose that a solution $x(t), y(t)$ of (1) remains in region IV of Figure 2 for all time $t \geqslant t_0$. Prove that $x(t)$ and $y(t)$ have limits ξ, η respectively, as t approaches infinity. Then conclude that (ξ, η) must equal (x_0, y_0).

In Problems 3 and 4, we assume that $a_1a_2 > b_1b_2c_1c_2$.

3. Suppose that a solution $x(t), y(t)$ of (1) leaves region II of Figure 3 at time $t = t^*$ by crossing the curve $y = \phi_1(x)$ or $y = \phi_2(x)$. Show that either $x(t)$ or $y(t)$ has a maximum at $t = t^*$. Then, show that this is impossible. Conclude, therefore, that every solution $x(t), y(t)$ of (1) which starts in region II at time $t = t_0$ must remain in region II for all future time $t \geqslant t_0$.

4. (a) Suppose that a solution $x(t), y(t)$ of (1) remains in either region I or III of Figure 3 for all time $t \geqslant t_0$. Show that $x(t)$ and $y(t)$ have limits ξ, η respectively, as t approaches infinity.
   (b) Conclude from Lemma 1 of Section 4.8 that $(\xi, \eta) = (0, 0)$.
   (c) Show that (ξ, η) cannot equal $(0, 0)$ if $x(t), y(t)$ remains in region I or region III for all time $t \geqslant t_0$.
   (d) Show that any solution $x(t), y(t)$ of (1) which starts on either $y = \phi_1(x)$ or $y = \phi_2(x)$ will immediately afterwards enter region II.

5. Assume that $a_1a_2 < b_1b_2c_1c_2$. Prove directly, using Theorem 2 of Section 4.3, that the equilibrium solution $x = x_0, y = y_0$ of (1) is asymptotically stable. *Warning*: The calculations are extremely tedious.

6. Assume that the number of homosexuals remains constant in time. Call this constant c. Let $x(t)$ denote the number of homosexuals who have gonorrhea at time t. Assume that homosexuals are cured of gonorrhea at a rate α_1, and that new infectives are added at a rate $\beta_1(c - x)x$.
   (a) Show that $\dot{x} = -\alpha_1 x + \beta_1 x(c - x)$.
   (b) What happens to $x(t)$ as t approaches infinity?

7. Suppose that the number of homosexuals $c(t)$ grows according to the logistic law $\dot{c} = c(a - bc)$, for some positive constants a and b. Let $x(t)$ denote the number of homosexuals who have gonorrhea at time t, and assume (see Problem 6) that $\dot{x} = -\alpha_1 x + \beta_1 x(c - x)$. What happens to $x(t)$ as t approaches infinity?

5 Separation of variables and Fourier series

5.1 Two point boundary-value problems

In the applications which we will study in this chapter, we will be confronted with the following problem.

Problem: For which values of λ can we find nontrivial functions $y(x)$ which satisfy

$$\frac{d^2y}{dx^2}+\lambda y=0; \qquad ay(0)+by'(0)=0, \quad cy(l)+dy'(l)=0? \tag{1}$$

Equation (1) is called a boundary-value problem, since we prescribe information about the solution $y(x)$ and its derivative $y'(x)$ at two distinct points, $x=0$ and $x=l$. In an initial-value problem, on the other hand, we prescribe the value of y and its derivative at a single point $x=x_0$.

Our intuitive feeling, at this point, is that the boundary-value problem (1) has nontrivial solutions $y(x)$ only for certain exceptional values λ. To wit, $y(x)=0$ is certainly one solution of (1), and the existence–uniqueness theorem for second-order linear equations would seem to imply that a solution $y(x)$ of $y''+\lambda y=0$ is determined uniquely once we prescribe two additional pieces of information. Let us test our intuition on the following simple, but extremely important example.

Example 1. For which values of λ does the boundary-value problem

$$\frac{d^2y}{dx^2}+\lambda y=0; \qquad y(0)=0, \quad y(l)=0 \tag{2}$$

have nontrivial solutions?

Solution.

(i) $\lambda=0$. Every solution $y(x)$ of the differential equation $y''=0$ is of the form $y(x)=c_1x+c_2$, for some choice of constants c_1 and c_2. The condition $y(0)=0$ implies that $c_2=0$, and the condition $y(l)=0$ then implies that $c_1=0$. Thus, $y(x)=0$ is the only solution of the boundary-value problem (2), for $\lambda=0$.

(ii) $\lambda<0$: In this case, every solution $y(x)$ of $y''+\lambda y=0$ is of the form $y(x)=c_1e^{\sqrt{-\lambda}\,x}+c_2e^{-\sqrt{-\lambda}\,x}$, for some choice of constants c_1 and c_2. The boundary conditions $y(0)=y(l)=0$ imply that

$$c_1+c_2=0, \qquad e^{\sqrt{-\lambda}\,l}c_1+e^{-\sqrt{-\lambda}\,l}c_2=0. \tag{3}$$

The system of equations (3) has a nonzero solution c_1,c_2 if, and only if,

$$\det\begin{pmatrix} 1 & 1 \\ e^{\sqrt{-\lambda}\,l} & e^{-\sqrt{-\lambda}\,l}\end{pmatrix}=e^{-\sqrt{-\lambda}\,l}-e^{\sqrt{-\lambda}\,l}=0.$$

This implies that $e^{\sqrt{-\lambda}\,l}=e^{-\sqrt{-\lambda}\,l}$, or $e^{2\sqrt{-\lambda}\,l}=1$. But this is impossible, since e^z is greater than one for $z>0$. Hence, $c_1=c_2=0$ and the boundary-value problem (2) has no nontrivial solutions $y(x)$ when λ is negative.

(iii) $\lambda>0$: In this case, every solution $y(x)$ of $y''+\lambda y=0$ is of the form $y(x)=c_1\cos\sqrt{\lambda}\,x+c_2\sin\sqrt{\lambda}\,x$, for some choice of constants c_1 and c_2. The condition $y(0)=0$ implies that $c_1=0$, and the condition $y(l)=0$ then implies that $c_2\sin\sqrt{\lambda}\,l=0$. This equation is satisfied, for any choice of c_2, if $\sqrt{\lambda}\,l=n\pi$, or $\lambda=n^2\pi^2/l^2$, for some positive integer n. Hence, the boundary-value problem (2) has nontrivial solutions $y(x)=c\sin n\pi x/l$ for $\lambda=n^2\pi^2/l^2$, $n=1,2,\ldots$.

Remark. Our calculations for the case $\lambda<0$ can be simplified if we write every solution $y(x)$ in the form $y=c_1\cosh\sqrt{-\lambda}\,x+c_2\sinh\sqrt{-\lambda}\,x$, where

$$\cosh\sqrt{-\lambda}\,x=\frac{e^{\sqrt{-\lambda}\,x}+e^{-\sqrt{-\lambda}\,x}}{2}$$

and

$$\sinh\sqrt{-\lambda}\,x=\frac{e^{\sqrt{-\lambda}\,x}-e^{-\sqrt{-\lambda}\,x}}{2}.$$

The condition $y(0)=0$ implies that $c_1=0$, and the condition $y(l)=0$ then implies that $c_2\sinh\sqrt{-\lambda}\,l=0$. But $\sinh z$ is positive for $z>0$. Hence, $c_2=0$, and $y(x)=0$.

Example 1 is indicative of the general boundary-value problem (1). Indeed, we have the following remarkable theorem which we state, but do not prove.

Theorem 1. *The boundary-value problem* (1) *has nontrivial solutions* $y(x)$ *only for a denumerable set of values* $\lambda_1, \lambda_2, \ldots$, *where* $\lambda_1 \leq \lambda_2 \ldots$, *and* λ_n *approaches infinity as n approaches infinity. These special values of* λ *are called eigenvalues of* (1), *and the nontrivial solutions* $y(x)$ *are called eigenfunctions of* (1). *In this terminology, the eigenvalues of* (2) *are* $\pi^2/l^2, 4\pi^2/l^2, 9\pi^2/l^2, \ldots$, *and the eigenfunctions of* (2) *are all constant multiples of* $\sin \pi x/l, \sin 2\pi x/l, \ldots$.

There is a very natural explanation of why we use the terms eigenvalue and eigenfunction in this context. Let $\mathbf{V}$ be the set of all functions $y(x)$ which have two continuous derivatives and which satisfy $ay(0)+by'(0)=0, cy(l)+dy'(l)=0$. Clearly, $\mathbf{V}$ is a vector space, of infinite dimension. Consider now the linear operator, or transformation L, defined by the equation

$$[Ly](x) = -\frac{d^2y}{dx^2}(x). \tag{4}$$

The solutions $y(x)$ of (1) are those functions y in $\mathbf{V}$ for which $Ly = \lambda y$. That is to say, the solutions $y(x)$ of (1) are exactly those functions y in $\mathbf{V}$ which are transformed by L into multiples λ of themselves.

Example 2. Find the eigenvalues and eigenfunctions of the boundary-value problem

$$\frac{d^2y}{dx^2} + \lambda y = 0; \qquad y(0)+y'(0)=0, \quad y(1)=0. \tag{5}$$

Solution.

(i) $\lambda = 0$. Every solution $y(x)$ of $y''=0$ is of the form $y(x)=c_1x+c_2$, for some choice of constants c_1 and c_2. The conditions $y(0)+y'(0)=0$ and $y(1)=0$ both imply that $c_2 = -c_1$. Hence, $y(x)=c(x-1)$, $c \neq 0$, is a nontrivial solution of (5) when $\lambda=0$; i.e., $y(x)=c(x-1)$, $c\neq 0$, is an eigenfunction of (5) with eigenvalue zero.

(ii) $\lambda<0$. In this case, every solution $y(x)$ of $y''+\lambda y=0$ is of the form $y(x)=c_1\cosh\sqrt{-\lambda}\,x + c_2\sinh\sqrt{-\lambda}\,x$, for some choice of constants c_1 and c_2. The boundary conditions $y(0)+y'(0)=0$ and $y(1)=0$ imply that

$$c_1 + \sqrt{-\lambda}\,c_2 = 0, \qquad \cosh\sqrt{-\lambda}\,c_1 + \sinh\sqrt{-\lambda}\,c_2 = 0. \tag{6}$$

(Observe that $(\cosh x)' = \sinh x$ and $(\sinh x)' = \cosh x$.) The system of equations (6) has a nontrivial solution c_1, c_2 if, and only if,

$$\det\begin{pmatrix} 1 & \sqrt{-\lambda} \\ \cosh\sqrt{-\lambda} & \sinh\sqrt{-\lambda} \end{pmatrix} = \sinh\sqrt{-\lambda} - \sqrt{-\lambda}\cosh\sqrt{-\lambda} = 0.$$

This implies that

$$\sinh\sqrt{-\lambda} = \sqrt{-\lambda}\cosh\sqrt{-\lambda}\,. \tag{7}$$

But Equation (7) has no solution $\lambda<0$. To see this, let $z=\sqrt{-\lambda}$, and consider the function $h(z)=z\cosh z-\sinh z$. This function is zero for $z=0$ and is positive for $z>0$, since its derivative

$$h'(z)=\cosh z+z\sinh z-\cosh z=z\sinh z$$

is strictly positive for $z>0$. Hence, no negative number λ can satisfy (7).

(iii) $\lambda>0$. In this case, every solution $y(x)$ of $y''+\lambda y=0$ is of the form $y(x)=c_1\cos\sqrt{\lambda}\,x+c_2\sin\sqrt{\lambda}\,x$, for some choice of constants c_1 and c_2. The boundary conditions imply that

$$c_1+\sqrt{\lambda}\,c_2=0, \qquad \cos\sqrt{\lambda}\,c_1+\sin\sqrt{\lambda}\,c_2=0. \tag{8}$$

The system of equations (8) has a nontrivial solution c_1, c_2 if, and only if,

$$\det\begin{pmatrix} 1 & \sqrt{\lambda} \\ \cos\sqrt{\lambda} & \sin\sqrt{\lambda} \end{pmatrix}=\sin\sqrt{\lambda}-\sqrt{\lambda}\cos\sqrt{\lambda}=0.$$

This implies that

$$\tan\sqrt{\lambda}=\sqrt{\lambda}. \tag{9}$$

To find those values of λ which satisfy (9), we set $\xi=\sqrt{\lambda}$ and draw the graphs of the functions $\eta=\xi$ and $\eta=\tan\xi$ in the $\xi-\eta$ plane (see Figure 1); the ξ coordinate of each point of intersection of these curves is then a root of the equation $\xi=\tan\xi$. It is clear that these curves intersect exactly once in the interval $\pi/2<\xi<3\pi/2$, and this occurs at a point $\xi_1>\pi$. Similarly, these two curves intersect exactly once in the interval $3\pi/2<\xi<5\pi/2$, and this occurs at a point $\xi_2>2\pi$. More generally, the curves $\eta=\xi$ and $\eta=\tan\xi$ intersect exactly once in the interval

$$\frac{(2n-1)\pi}{2}<\xi<\frac{(2n+1)\pi}{2}$$

and this occurs at a point $\xi_n>n\pi$.

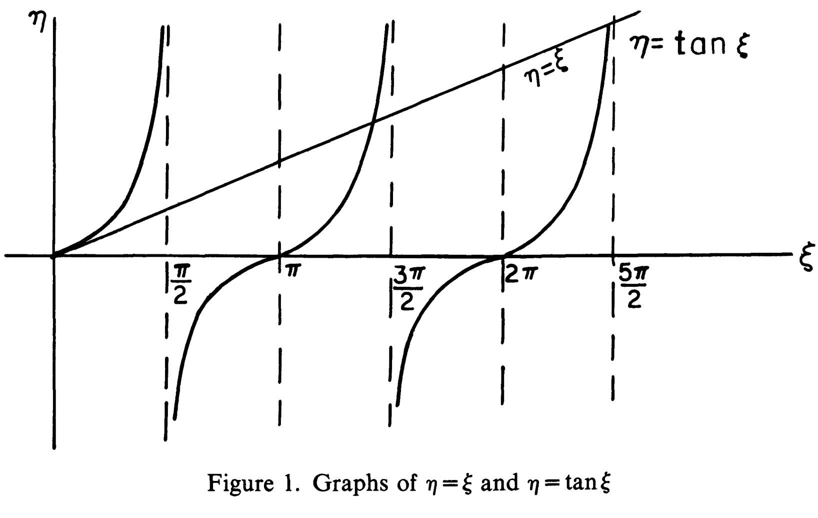


Figure 1. Graphs of $\eta=\xi$ and $\eta=\tan\xi$

Finally, the curves $\eta=\xi$ and $\eta=\tan\xi$ do not intersect in the interval $0<\xi<\pi/2$. To prove this, set $h(\xi)=\tan\xi-\xi$ and compute

$$h'(\xi)=\sec^2\xi-1=\tan^2\xi.$$

This quantity is positive for $0<\xi<\pi/2$. Consequently, the eigenvalues of (5) are $\lambda_1=\xi_1^2,\lambda_2=\xi_2^2,\dots$, and the eigenfunction of (5) are all constant multiples of the functions $-\sqrt{\lambda_1}\cos\sqrt{\lambda_1}\,x+\sin\sqrt{\lambda_1}\,x,\ -\sqrt{\lambda_2}\cos\sqrt{\lambda_2}\,x+\sin\sqrt{\lambda_2}\,x,\dots$. We cannot compute λ_n exactly. Nevertheless, we know that

$$n^2\pi^2<\lambda_n<(2n+1)^2\pi^2/4.$$

In addition, it is clear that λ_n approaches $(2n+1)^2\pi^2/4$ as n approaches infinity.

Exercises

Find the eigenvalues and eigenfunctions of each of the following boundary-value problems.

1. $y''+\lambda y=0;\quad y(0)=0,\ y'(l)=0$

2. $y''+\lambda y=0;\quad y'(0)=0,\ y'(l)=0$

3. $y''-\lambda y=0;\quad y'(0)=0,\ y'(l)=0$

4. $y''+\lambda y=0;\quad y'(0)=0,\ y(l)=0$

5. $y''+\lambda y=0;\quad y(0)=0,\ y(\pi)-y'(\pi)=0$

6. $y''+\lambda y=0;\quad y(0)-y'(0)=0,\ y(1)=0$

7. $y''+\lambda y=0;\quad y(0)-y'(0)=0,\ y(\pi)-y'(\pi)=0$

8. For which values of λ does the boundary-value problem

$$y''-2y'+(1+\lambda)y=0;\qquad y(0)=0,\ y(1)=0$$

have a nontrivial solution?

9. For which values of λ does the boundary-value problem

$$y''+\lambda y=0;\qquad y(0)=y(2\pi),\ y'(0)=y'(2\pi)$$

have a nontrivial solution?

10. Consider the boundary-value problem

$$y''+\lambda y=f(t);\qquad y(0)=0,\ y(1)=0 \tag{*}$$

(a) Show that (*) has a unique solution $y(t)$ if λ is not an eigenvalue of the homogeneous problem.
(b) Show that (*) may have no solution $y(t)$ if λ is an eigenvalue of the homogeneous problem.
(c) Let λ be an eigenvalue of the homogeneous problem. Determine conditions on f so that (*) has a solution $y(t)$. Is this solution unique?

5.2 Introduction to partial differential equations

Up to this point, the differential equations that we have studied have all been relations involving one or more functions of a single variable, and their derivatives. In this sense, these differential equations are *ordinary* differential equations. On the other hand, many important problems in applied mathematics give rise to *partial* differential equations. A partial differential equation is a relation involving one or more functions of *several* variables, and their partial derivatives. For example, the equation

$$\frac{\partial^3 u}{\partial x^3}+\left(\frac{\partial u}{\partial t}\right)^2=\frac{\partial^2 u}{\partial x^2}$$

is a partial differential equation for the function $u(x,t)$, and the equations

$$\frac{\partial u}{\partial x}=\frac{\partial v}{\partial y}, \qquad \frac{\partial u}{\partial y}=-\frac{\partial v}{\partial x}$$

are a system of partial differential equations for the two functions $u(x,y)$ and $v(x,y)$. The order of a partial differential equation is the order of the highest partial derivative that appears in the equation. For example, the order of the partial differential equation

$$\frac{\partial^2 u}{\partial t^2}=2\frac{\partial^2 u}{\partial x\partial t}+u$$

is two, since the order of the highest partial derivative that appears in this equation is two.

There are three classical partial differential equations of order two which appear quite often in applications, and which dominate the theory of partial differential equations. These equations are

$$\frac{\partial u}{\partial t}=\alpha^2\frac{\partial^2 u}{\partial x^2}, \tag{1}$$

$$\frac{\partial^2 u}{\partial t^2}=c^2\frac{\partial^2 u}{\partial x^2} \tag{2}$$

and

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=0. \tag{3}$$

Equation (1) is known as the heat equation, and it appears in the study of heat conduction and other diffusion processes. For example, consider a thin metal bar of length l whose surface is insulated. Let $u(x,t)$ denote the temperature in the bar at the point x at time t. This function satisfies the partial differential equation (1) for $0<x<l$. The constant α^2 is known as the thermal diffusivity of the bar, and it depends solely on the material from which the bar is made.

Equation (2) is known as the wave equation, and it appears in the study of acoustic waves, water waves and electromagnetic waves. Some form of this equation, or a generalization of it, almost invariably arises in any mathematical analysis of phenomena involving the propagation of waves in a continuous medium. (We will gain some insight into why this is so in Section 5.7.) The wave equation also appears in the study of mechanical vibrations. Suppose, for example, that an elastic string of length l, such as a violin string or guy wire, is set in motion so that it vibrates in a vertical plane. Let $u(x,t)$ denote the vertical displacement of the string at the point x at time t (see Figure 1). If all damping effects, such as air resistance, are negligible, and if the amplitude of the motion is not too large, then $u(x,t)$ will satisfy the partial differential equation (2) on the interval $0 \leqslant x \leqslant l$. In this case, the constant c^2 is H/ρ, where H is the horizontal component of the tension in the string, and ρ is the mass per unit length of the string.

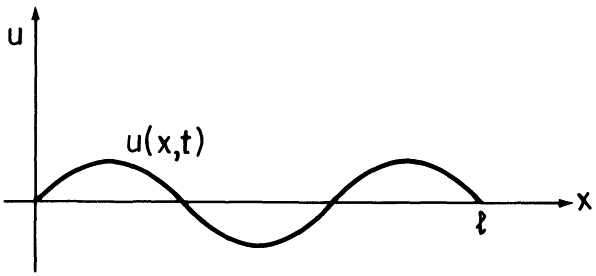


Figure 1

Equation (3) is known as Laplace's equation, and is the most famous of all partial differential equations. It arises in the study of such diverse applications as steady state heat flow, vibrating membranes, and electric and gravitational potentials. For this reason, Laplace's equation is often referred to as the potential equation.

In addition to the differential equation (1), (2), or (3), we will often impose initial and boundary conditions on the function u. These conditions will be dictated to us by the physical and biological problems themselves; they will be chosen so as to guarantee that our equation has a unique solution.

As a model case for the heat equation (1), we consider a thin metal bar of length l whose sides are insulated, and we let $u(x,t)$ denote the temperature in the bar at the point x at time t. In order to dermine the temperature in the bar at any time t we need to know (i) the initial temperature distribution in the bar, and (ii) what is happening at the ends of the bar. Are they held at constant temperatures, say 0°C, or are they insulated, so that no heat can pass through them? (This latter condition implies that $u_x(0,t) = u_x(l,t) = 0$.) Thus, a "well posed" problem for diffusion processes is the heat equation (1), together with the initial condition $u(x,0)=f(x)$, $0<x<l$, and the boundary conditions $u(0,t)=u(l,t)=0$, or $u_x(0,t)=u_x(l,t)=0$.

As a model case for the wave equation, we consider an elastic string of length l, whose ends are fixed, and which is set in motion in a vertical plane. In order to determine the position $u(x,t)$ of the string at any time t we need to know (i) the initial position of the string, and (ii) the initial velocity of the string. It is also implicit that $u(0,t)=u(l,t)=0$. Thus, a well posed problem for wave propagation is the differential equation (2) together with the initial conditions $u(x,0)=f(x)$, $u_t(x,0)=g(x)$, and the boundary conditions $u(0,t)=u(l,t)=0$.

The partial differential equation (3) does not contain the time t, so that we do not expect any "initial conditions" to be imposed here. In the problems that arise in applications, we are given u, or its normal derivative, on the boundary of a given region R, and we seek to determine $u(x,y)$ inside R. The problem of finding a solution of Laplace's equation which takes on given boundary values is known as a Dirichlet problem, while the problem of finding a solution of Laplace's equation whose normal derivative takes on given boundary values is known as a Neumann problem.

In Section 5.3 we will develop a very powerful method, known as the method of separation of variables, for solving the boundary-value problem (strictly speaking, we should say "initial boundary-value problem")

$$\frac{\partial u}{\partial t}=\alpha^2\frac{\partial^2 u}{\partial x^2}; \qquad u(x,0)=f(x), \quad 0<x<l; \quad u(0,t)=u(l,t)=0.$$

After developing the theory of Fourier series in Sections 5.4 and 5.5, we will show that the method of separation of variables can also be used to solve more general problems of heat conduction, and several important problems of wave propagation and potential theory.

5.3 The heat equation; separation of variables

Consider the boundary-value problem

$$\frac{\partial u}{\partial t}=\alpha^2\frac{\partial^2 u}{\partial x^2}; \qquad u(x,0)=f(x), \quad 0<x<l; \quad u(0,t)=u(l,t)=0. \tag{1}$$

Our goal is to find the solution $u(x,t)$ of (1). To this end, it is helpful to recall how we solved the initial-value problem

$$\frac{d^2y}{dt^2}+p(t)\frac{dy}{dt}+q(t)y=0; \qquad y(0)=y_0, \quad y'(0)=y_0'. \tag{2}$$

First we showed that the differential equation

$$\frac{d^2y}{dt^2}+p(t)\frac{dy}{dt}+q(t)y=0 \tag{3}$$

is linear; that is, any linear combination of solutions of (3) is again a solution of (3). And then, we found the solution $y(t)$ of (2) by taking an appropriate linear combination $c_1y_1(t)+c_2y_2(t)$ of two linearly independent solutions $y_1(t)$ and $y_2(t)$ of (3). Now, it is easily verified that any linear combination $c_1u_1(x,t)+\ldots+c_nu_n(x,t)$ of solutions $u_1(x,t),\ldots,u_n(x,t)$ of

$$\frac{\partial u}{\partial t}=\alpha^2\frac{\partial^2 u}{\partial x^2} \tag{4}$$

is again a solution of (4). In addition, if $u_1(x,t),\ldots,u_n(x,t)$ satisfy the boundary conditions $u(0,t)=u(l,t)=0$, then the linear combination $c_1u_1+\ldots+c_nu_n$ also satisfies these boundary conditions. This suggests the following "game plan" for solving the boundary-value problem (1):

(a) Find as many solutions $u_1(x,t),u_2(x,t),\ldots$ as we can of the boundary-value problem

$$\frac{\partial u}{\partial t}=\alpha^2\frac{\partial^2 u}{\partial x^2}; \qquad u(0,t)=u(l,t)=0. \tag{5}$$

(b) Find the solution $u(x,t)$ of (1) by taking an appropriate linear combination of the functions $u_n(x,t)$, $n=1,2,\ldots$.

(a) Since we don't know, as yet, how to solve any partial differential equations, we must reduce the problem of solving (5) to that of solving one or more ordinary differential equations. This is accomplished by setting $u(x,t)=X(x)T(t)$ (hence the name separation of variables). Computing

$$\frac{\partial u}{\partial t}=XT' \quad\text{and}\quad \frac{\partial^2 u}{\partial x^2}=X''T$$

we see that $u(x,t)=X(x)T(t)$ is a solution of the equation $u_t=\alpha^2u_{xx}$ ($u_t=\partial u/\partial t$ and $u_{xx}=\partial^2u/\partial x^2$) if

$$XT'=\alpha^2X''T. \tag{6}$$

Dividing both sides of (6) by α^2XT gives

$$\frac{X''}{X}=\frac{T'}{\alpha^2T}. \tag{7}$$

Now, observe that the left-hand side of (7) is a function of x alone, while the right-hand side of (7) is a function of t alone. This implies that

$$\frac{X''}{X}=-\lambda, \quad\text{and}\quad \frac{T'}{\alpha^2T}=-\lambda \tag{8}$$

for some constant λ. (The only way that a function of x can equal a function of t is if both are constant. To convince yourself of this, let $f(x)=g(t)$ and fix t_0. Then, $f(x)=g(t_0)$ for all x, so that $f(x)=\text{constant}=c_1$, and this immediately implies that $g(t)$ also equals c_1.) In addition, the boundary conditions

$$0=u(0,t)=X(0)T(t),$$

and

$$0=u(l,t)=X(l)T(t)$$

imply that $X(0)=0$ and $X(l)=0$ (otherwise, u must be identically zero). Thus, $u(x,t)=X(x)T(t)$ is a solution of (5) if

$$X''+\lambda X=0; \qquad X(0)=0, \quad X(l)=0 \tag{9}$$

and

$$T'+\lambda\alpha^2 T=0. \tag{10}$$

At this point, the constant λ is arbitrary. However, we know from Example 1 of Section 5.1 that the boundary-value problem (9) has a nontrivial solution $X(x)$ only if $\lambda=\lambda_n=n^2\pi^2/l^2$, $n=1,2,\ldots$; and in this case,

$$X(x)=X_n(x)=\sin\frac{n\pi x}{l}.$$

Equation (10), in turn, implies that

$$T(t)=T_n(t)=e^{-\alpha^2n^2\pi^2t/l^2}.$$

(Actually, we should multiply both $X_n(x)$ and $T_n(t)$ by constants; however, we omit these constants here since we will soon be taking linear combinations of the functions $X_n(x)T_n(t)$.) Hence,

$$u_n(x,t)=\sin\frac{n\pi x}{l}e^{-\alpha^2n^2\pi^2t/l^2}$$

is a nontrivial solution of (5) for every positive integer n.

(b) Suppose that $f(x)$ is a finite linear combination of the functions $\sin n\pi x/l$; that is,

$$f(x)=\sum_{n=1}^{N}c_n\sin\frac{n\pi x}{l}.$$

Then,

$$u(x,t)=\sum_{n=1}^{N}c_n\sin\frac{n\pi x}{l}e^{-\alpha^2n^2\pi^2t/l^2}$$

is the desired solution of (1), since it is a linear combination of solutions of (5), and it satisfies the initial condition

$$u(x,0)=\sum_{n=1}^{N}c_n\sin\frac{n\pi x}{l}=f(x), \qquad 0<x<l.$$

Unfortunately, though, most functions $f(x)$ cannot be expanded as a finite linear combination of the functions $\sin n\pi x/l$, $n=1,2,\ldots$, on the interval $0<x<l$. This leads us to ask the following question.

Question: Can an arbitrary function $f(x)$ be written as an *infinite* linear combination of the functions $\sin n\pi x/l$, $n=1,2,\ldots$, on the interval $0<x<$

l? In other words, given an arbitrary function f, can we find constants $c_1, c_2, \ldots$, such that

$$f(x) = c_1 \sin\frac{\pi x}{l} + c_2 \sin\frac{2\pi x}{l} + \ldots = \sum_{n=1}^{\infty} c_n \sin\frac{n\pi x}{l}; \qquad 0 < x < l?$$

Remarkably, the answer to this question is yes, as we show in Section 5.5.

Example 1. At time $t=0$, the temperature $u(x,0)$ in a thin copper rod ($\alpha^2 = 1.14$) of length one is $2\sin 3\pi x + 5\sin 8\pi x$, $0 \leqslant x \leqslant 1$. The ends of the rod are packed in ice, so as to maintain them at 0°C. Find the temperature $u(x,t)$ in the rod at any time $t > 0$.

Solution. The temperature $u(x,t)$ satisfies the boundary-value problem

$$\frac{\partial u}{\partial t} = 1.14\frac{\partial^2 u}{\partial x^2}; \qquad \begin{cases} u(x,0) = 2\sin 3\pi x + 5\sin 8\pi x, & 0 < x < 1 \\ u(0,t) = u(1,t) = 0 \end{cases}$$

and this implies that

$$u(x,t) = 2\sin 3\pi x\, e^{-9(1.14)\pi^2 t} + 5\sin 8\pi x\, e^{-64(1.14)\pi^2 t}.$$

EXERCISES

Find a solution $u(x,t)$ of the following problems.

1. $\dfrac{\partial u}{\partial t} = 1.71\dfrac{\partial^2 u}{\partial x^2}; \quad \begin{cases} u(x,0) = \sin \pi x/2 + 3\sin 5\pi x/2, & 0 < x < 2 \\ u(0,t) = u(2,t) = 0 \end{cases}$

2. $\dfrac{\partial u}{\partial t} = 1.14\dfrac{\partial^2 u}{\partial x^2}; \quad \begin{cases} u(x,0) = \sin \pi x/2 - 3\sin 2\pi x, & 0 < x < 2 \\ u(0,t) = u(2,t) = 0 \end{cases}$

3. Use the method of separation of variables to solve the boundary-value problem

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + u; \qquad \begin{cases} u(x,0) = 3\sin 2\pi x - 7\sin 4\pi x, & 0 < x < 10 \\ u(0,t) = u(10,t) = 0 \end{cases}$$

Use the method of separation of variables to solve each of the following boundary-value problems.

4. $\dfrac{\partial u}{\partial t} = \dfrac{\partial u}{\partial y}; \quad u(0,y) = e^{y} + e^{-2y}$

5. $\dfrac{\partial u}{\partial t} = \dfrac{\partial u}{\partial y}; \quad u(t,0) = e^{-3t} + e^{2t}$

6. $\dfrac{\partial u}{\partial t} = \dfrac{\partial u}{\partial y} + u; \quad u(0,y) = 2e^{-y} - e^{2y}$

7. $\dfrac{\partial u}{\partial t} = \dfrac{\partial u}{\partial y} - u; \quad u(t,0) = e^{-5t} + 2e^{-7t} - 14e^{13t}$

8. Determine whether the method of separation of variables can be used to replace each of the following partial differential equations by pairs of ordinary differential equations. If so, find the equations.

(a) $tu_{tt}+u_x=0$ (b) $tu_{xx}+xu_t=0$

(c) $u_{xx}+(x-y)u_{yy}=0$ (d) $u_{xx}+2u_{xt}+u_t=0$

9. The heat equation in two space dimensions is

$$u_t=\alpha^2(u_{xx}+u_{yy}). \qquad (*)$$

(a) Assuming that $u(x,y,t)=X(x)Y(y)T(t)$, find ordinary differential equations satisfied by X, Y, and T.

(b) Find solutions $u(x,y,t)$ of (*) which satisfy the boundary conditions $u(0,y,t)=0$, $u(a,y,t)=0$, $u(x,0,t)=0$, and $u(x,b,t)=0$.

10. The heat equation in two space dimensions may be expressed in terms of polar coordinates as

$$u_t=\alpha^2\left[u_{rr}+\frac{1}{r}u_r+\frac{1}{r^2}u_{\theta\theta}\right].$$

Assuming that $u(r,\theta,t)=R(r)\Theta(\theta)T(t)$, find ordinary differential equations satisfied by R, Θ, and T.

5.4 Fourier series

On December 21, 1807, an engineer named Joseph Fourier announced to the prestigious French Academy of Sciences that an arbitrary function $f(x)$ could be expanded in an infinite series of sines and cosines. Specifically, let $f(x)$ be defined on the interval $-l\leqslant x\leqslant l$, and compute the numbers

$$a_n=\frac{1}{l}\int_{-l}^{l}f(x)\cos\frac{n\pi x}{l}\,dx, \qquad n=0,1,2,\ldots \qquad (1)$$

and

$$b_n=\frac{1}{l}\int_{-l}^{l}f(x)\sin\frac{n\pi x}{l}\,dx, \qquad n=1,2,\ldots \qquad (2)$$

Then, the infinite series

$$\frac{a_0}{2}+a_1\cos\frac{\pi x}{l}+b_1\sin\frac{\pi x}{l}+\ldots=\frac{a_0}{2}+\sum_{n=1}^{\infty}\left[a_n\cos\frac{n\pi x}{l}+b_n\sin\frac{n\pi x}{l}\right] \qquad (3)$$

converges to $f(x)$. Fourier's announcement caused a loud furor in the Academy. Many of its prominent members, including the famous mathematician Lagrange, thought this result to be pure nonsense, since at that time it could not be placed on a rigorous foundation. However, mathematicians have now developed the theory of "Fourier series" to such an extent that whole volumes have been written on it. (Just recently, in fact, they have succeeded in establishing exceedingly sharp conditions for the Fourier series (3) to converge. This result ranks as one of the great mathematical theorems of the twentieth century.) The following theorem, while not the most general theorem possible, covers most of the situations that arise in applications.

Theorem 2. *Let f and f' be piecewise continuous on the interval $-l \leqslant x \leqslant l$. (This means that f and f' have only a finite number of discontinuities on this interval, and both f and f' have right- and left-hand limits at each point of discontinuity.) Compute the numbers a_n and b_n from* (1) *and* (2) *and form the infinite series* (3). *This series, which is called the Fourier series for f on the interval $-l \leqslant x \leqslant l$, converges to $f(x)$ if f is continuous at x, and to $\frac{1}{2}[f(x+0)+f(x-0)]$* if f is discontinuous at x. At $x = \pm l$, the Fourier series* (3) *converges to $\frac{1}{2}[f(l)+f(-l)]$, where $f(\pm l)$ is the limit of $f(x)$ as x approaches $\pm l$.*

Remark. The quantity $\frac{1}{2}[f(x+0)+f(x-0)]$ is the average of the right- and left-hand limits of f at the point x. If we define $f(x)$ to be the average of the right- and left-hand limits of f at any point of discontinuity x, then the Fourier series (3) converges to $f(x)$ for all points x in the interval $-l < x < l$.

Example 1. Let f be the function which is 0 for $-1 \leqslant x < 0$ and 1 for $0 \leqslant x \leqslant 1$. Compute the Fourier series for f on the interval $-1 \leqslant x \leqslant 1$.
Solution. In this problem, $l = 1$. Hence, from (1) and (2),

$$a_0 = \int_{-1}^{1} f(x)\,dx = \int_0^1 dx = 1,$$

$$a_n = \int_{-1}^{1} f(x)\cos n\pi x\,dx = \int_0^1 \cos n\pi x\,dx = 0, \qquad n \geqslant 1$$

and

$$b_n = \int_{-1}^{1} f(x)\sin n\pi x\,dx = \int_0^1 \sin n\pi x\,dx$$

$$= \frac{1}{n\pi}(1-\cos n\pi) = \frac{1-(-1)^n}{n\pi}, \qquad n \geqslant 1.$$

Notice that $b_n = 0$ for n even, and $b_n = 2/n\pi$ for n odd. Hence, the Fourier series for f on the interval $-1 \leqslant x \leqslant 1$ is

$$\frac{1}{2} + \frac{2\sin\pi x}{\pi} + \frac{2\sin 3\pi x}{3\pi} + \frac{2\sin 5\pi x}{5\pi} + \dots.$$

By Theorem 2, this series converges to 0 if $-1 < x < 0$, and to 1 if $0 < x < 1$. At $x = -1$, 0, and $+1$, this series reduces to the single number $\frac{1}{2}$, which is the value predicted for it by Theorem 2.

*The quantity $f(x+0)$ denotes the limit from the right of f at the point x. Similarly, $f(x-0)$ denotes the limit of f from the left.

Example 2. Let f be the function which is 1 for $-2\leqslant x<0$ and x for $0\leqslant x\leqslant 2$. Compute the Fourier series for f on the interval $-2\leqslant x\leqslant 2$.
Solution. In this problem $l=2$. Hence from (1) and (2),

$$a_0=\frac{1}{2}\int_{-2}^{2}f(x)\,dx=\frac{1}{2}\int_{-2}^{0}dx+\frac{1}{2}\int_{0}^{2}x\,dx=2$$

$$a_n=\frac{1}{2}\int_{-2}^{2}f(x)\cos\frac{n\pi x}{2}\,dx=\frac{1}{2}\int_{-2}^{0}\cos\frac{n\pi x}{2}\,dx+\frac{1}{2}\int_{0}^{2}x\cos\frac{n\pi x}{2}\,dx$$

$$=\frac{2}{(n\pi)^2}(\cos n\pi-1),\qquad n\geqslant 1$$

and

$$b_n=\frac{1}{2}\int_{-2}^{2}f(x)\sin\frac{n\pi x}{2}\,dx=\frac{1}{2}\int_{-2}^{0}\sin\frac{n\pi x}{2}\,dx+\frac{1}{2}\int_{0}^{2}x\sin\frac{n\pi x}{2}\,dx$$

$$=-\frac{1}{n\pi}(1+\cos n\pi),\qquad n\geqslant 1.$$

Notice that $a_n=0$ if n is even; $a_n=-4/n^2\pi^2$ if n is odd; $b_n=0$ if n is odd; and $b_n=-2/n\pi$ if n is even. Hence, the Fourier series for f on the interval $-2\leqslant x\leqslant 2$ is

$$1-\frac{4}{\pi^2}\cos\frac{\pi x}{2}-\frac{1}{\pi}\sin\pi x-\frac{4}{9\pi^2}\cos\frac{3\pi x}{2}-\frac{1}{2\pi}\sin 2\pi x+\ldots$$

$$=1-\frac{4}{\pi^2}\sum_{n=0}^{\infty}\frac{\cos(2n+1)\pi x/2}{(2n+1)^2}-\frac{1}{\pi}\sum_{n=1}^{\infty}\frac{\sin n\pi x}{n}.\tag{4}$$

By Theorem 2, this series converges to 1 if $-2<x<0$; to x, if $0<x<2$; to $\frac{1}{2}$ if $x=0$; and to $\frac{3}{2}$ if $x\pm 2$. Now, at $x=0$, the Fourier series (4) is

$$1-\frac{4}{\pi^2}\left[\frac{1}{1^2}+\frac{1}{3^2}+\frac{1}{5^2}+\frac{1}{7^2}+\ldots\right].$$

Thus, we deduce the remarkable identity

$$\frac{1}{2}=1-\frac{4}{\pi^2}\left[\frac{1}{1^2}+\frac{1}{3^2}+\frac{1}{5^2}+\frac{1}{7^2}+\ldots\right]$$

or

$$\frac{1}{1^2}+\frac{1}{3^2}+\frac{1}{5^2}+\frac{1}{7^2}+\ldots=\frac{\pi^2}{8}.$$

The Fourier coefficients a_n and b_n defined by (1) and (2) can be derived in a simple manner. Indeed, if a piecewise continuous function f can be expanded in a series of sines and cosines on the interval $-l\leqslant x\leqslant l$, then, of necessity, this series must be the Fourier series (3). We prove this in the following manner. Suppose that f is piecewise continuous, and that

$$f(x)=\frac{c_0}{2}+\sum_{k=1}^{\infty}\left[c_k\cos\frac{k\pi x}{l}+d_k\sin\frac{k\pi x}{l}\right] \tag{5}$$

for some numbers c_k and d_k. Equation (5) is assumed to hold at all but a finite number of points in the interval $-l\leqslant x\leqslant l$. Integrating both sides of (5) between $-l$ and l gives $c_0 l=\int_{-l}^{l}f(x)dx$, since

$$\int_{-l}^{l}\cos\frac{k\pi x}{l}\,dx=\int_{-l}^{l}\sin\frac{k\pi x}{l}\,dx=0;\qquad k=1,2,\ldots^*$$

Similarly, multiplying both sides of (5) by $\cos n\pi x/l$ and integrating between $-l$ and l gives

$$lc_n=\int_{-l}^{l}f(x)\cos\frac{n\pi x}{l}\,dx$$

while multiplying both sides of (5) by $\sin n\pi x/l$ and integrating between $-l$ and l gives

$$ld_n=\int_{-l}^{l}f(x)\sin\frac{n\pi x}{l}\,dx.$$

This follows immediately from the relations (see Exercise 19)

$$\int_{-l}^{l}\cos\frac{n\pi x}{l}\cos\frac{k\pi x}{l}\,dx=\begin{cases}0, & k\neq n\\ l, & k=n\end{cases} \tag{6}$$

$$\int_{-l}^{l}\cos\frac{n\pi x}{l}\sin\frac{k\pi x}{l}\,dx=0 \tag{7}$$

and

$$\int_{-l}^{l}\sin\frac{n\pi x}{l}\sin\frac{k\pi x}{l}\,dx=\begin{cases}0, & k\neq n\\ l, & k=n\end{cases}. \tag{8}$$

Hence, the coefficients c_n and d_n must equal the Fourier coefficients a_n and b_n. In particular, therefore, a function f can be expanded in one, and only one, Fourier series on the interval $-l\leqslant x\leqslant l$.

Example 3. Find the Fourier series for the function $f(x)=\cos^2 x$ on the interval $-\pi\leqslant x\leqslant\pi$.

Solution. By the preceding remark, the function $f(x)=\cos^2 x$ has a unique Fourier series

$$\frac{a_0}{2}+\sum_{n=1}^{\infty}\left[a_n\cos\frac{n\pi x}{\pi}+b_n\sin\frac{n\pi x}{\pi}\right]=\frac{a_0}{2}+\sum_{n=1}^{\infty}\left[a_n\cos nx+b_n\sin nx\right]$$

*It can be shown that it is permissible to integrate the series (5) term by term.

on the interval $-\pi \leqslant x \leqslant \pi$. But we already know that

$$\cos^2 x = \frac{1}{2} + \frac{\cos 2x}{2}.$$

Hence, the Fourier series for $\cos^2 x$ on the interval $-\pi \leqslant x \leqslant \pi$ must be $\frac{1}{2} + \frac{1}{2}\cos 2x$.

The functions $\cos n\pi x/l$ and $\sin n\pi x/l$, $n = 1,2,\ldots$ all have the interesting property that they are periodic with period $2l$; that is, they repeat themselves over every interval of length $2l$. This follows trivially from the identities

$$\cos\frac{n\pi}{l}(x+2l) = \cos\left(\frac{n\pi x}{l} + 2n\pi\right) = \cos\frac{n\pi x}{l}$$

and

$$\sin\frac{n\pi}{l}(x+2l) = \sin\left(\frac{n\pi x}{l} + 2n\pi\right) = \sin\frac{n\pi x}{l}.$$

Hence, the Fourier series (3) converges for all x to a periodic function $F(x)$. This function is called the periodic extension of $f(x)$. It is defined by the equations

$$\begin{cases} F(x) = f(x), & -l < x < l \\ F(x) = \frac{1}{2}[f(l) + f(-l)], & x = \pm l \\ F(x+2l) = F(x). & \end{cases}$$

For example, the periodic extension of the function $f(x) = x$ is described in Figure 1, and the periodic extension of the function $f(x) = |x|$ is the sawtoothed function described in Figure 2.

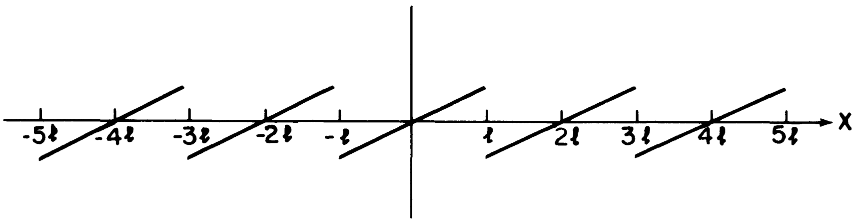


Figure 1. Periodic extension of $f(x) = x$

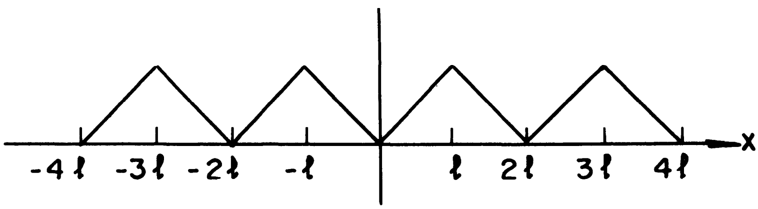


Figure 2. Periodic extension of $f(x) = |x|$

Exercises

In each of Problems 1–13, find the Fourier series for the given function f on the prescribed interval.

1. $f(x)=\begin{cases} -1, & -1\leqslant x<0 \\ 1, & 0\leqslant x\leqslant 1 \end{cases};\quad |x|\leqslant 1$

2. $f(x)=\begin{cases} x, & -2\leqslant x<0 \\ 0, & 0\leqslant x\leqslant 2 \end{cases};\quad |x|\leqslant 2$

3. $f(x)=x;\quad -1\leqslant x\leqslant 1$

4. $f(x)=\begin{cases} -x, & -1\leqslant x<0 \\ x, & 0\leqslant x\leqslant 1 \end{cases};\quad |x|\leqslant 1$

5. $f(x)=\begin{cases} 1, & -2\leqslant x<0 \\ 0, & 0\leqslant x<1 \\ 1, & 1\leqslant x\leqslant 2 \end{cases};\quad |x|\leqslant 2$

6. $f(x)=\begin{cases} 0, & -2\leqslant x<1 \\ 3, & 1\leqslant x\leqslant 2 \end{cases};\quad |x|\leqslant 2$

7. $f(x)=\begin{cases} 0, & -l\leqslant x<0 \\ e^{x}, & 0\leqslant x\leqslant l \end{cases};\quad |x|\leqslant l$

8. $f(x)=\begin{cases} e^{x}, & -l\leqslant x<0 \\ 0, & 0\leqslant x\leqslant l \end{cases};\quad |x|\leqslant l$

9. $f(x)=\begin{cases} e^{-x}, & -l\leqslant x<0 \\ e^{x}, & 0\leqslant x\leqslant l \end{cases};\quad -l\leqslant x\leqslant l$

10. $f(x)=e^{x};\quad |x|\leqslant l$

11. $f(x)=e^{-x};\quad |x|\leqslant l$

12. $f(x)=\sin^2 x;\quad |x|\leqslant \pi$

13. $f(x)=\sin^3 x;\quad |x|\leqslant \pi$

14. Let $f(x)=(\pi\cos ax)/2a\sin a\pi$, a not an integer.
(a) Find the Fourier series for f on the interval $-\pi\leqslant x\leqslant \pi$.
(b) Show that this series converges at $x=\pi$ to the value $(\pi/2a)\cot\pi a$.
(c) Use this result to sum the series

$$\frac{1}{1^2-a^2}+\frac{1}{2^2-a^2}+\frac{1}{3^2-a^2}+\dots.$$

15. Suppose that f and f' are piecewise continuous on the interval $-l\leqslant x\leqslant l$. Show that the Fourier coefficients a_n and b_n approach zero as n approaches infinity.

16. Let

$$f(x)=\frac{a_0}{2}+\sum_{n=1}^{\infty}\left[a_n\cos\frac{n\pi x}{l}+b_n\sin\frac{n\pi x}{l}\right].$$

Show that

$$\frac{1}{l}\int_{-l}^{l}f^2(x)\,dx=\frac{a_0^2}{2}+\sum_{n=1}^{\infty}(a_n^2+b_n^2).$$

This relation is known as Parseval's identity. *Hint*: Square the Fourier series for f and integrate term by term.

17. (a) Find the Fourier series for the function $f(x)=x^2$ on the interval $-\pi \leqslant x \leqslant \pi$.
(b) Use Parseval's identity to show that
$$\frac{1}{1^4}+\frac{1}{2^4}+\frac{1}{3^4}+\dots=\frac{\pi^4}{90}.$$

18. If the Dirac delta function $\delta(x)$ had a Fourier series on the interval $-l \leqslant x \leqslant l$, what would it be?

19. Derive Equations (6)–(8). *Hint*: Use the trigonometric identities
$$\sin A \cos B = \tfrac{1}{2}[\sin(A+B)+\sin(A-B)]$$
$$\sin A \sin B = \tfrac{1}{2}[\cos(A-B)-\cos(A+B)]$$
$$\cos A \cos B = \tfrac{1}{2}[\cos(A+B)+\cos(A-B)].$$

5.5 Even and odd functions

There are certain special cases when the Fourier series of a function f reduces to a pure cosine or a pure sine series. These special cases occur when f is even or odd.

Definition. A function f is said to be *even* if $f(-x)=f(x)$.

Example 1. The function $f(x)=x^2$ is even since
$$f(-x)=(-x)^2=x^2=f(x).$$

Example 2. The function $f(x)=\cos n\pi x/l$ is even since
$$f(-x)=\cos\frac{-n\pi x}{l}=\cos\frac{n\pi x}{l}=f(x).$$

Definition. A function f is said to be *odd* if $f(-x)=-f(x)$.

Example 3. The function $f(x)=x$ is odd since
$$f(-x)=-x=-f(x).$$

Example 4. The function $f(x)=\sin n\pi x/l$ is odd since
$$f(-x)=\sin\frac{-n\pi x}{l}=-\sin\frac{n\pi x}{l}=-f(x).$$

Even and odd functions satisfy the following elementary properties.

1. The product of two even functions is even.
2. The product of two odd functions is even.
3. The product of an odd function with an even function is odd.

The proofs of these assertions are trivial and follow immediately from the definitions. For example, let f and g be odd and let $h(x)=f(x)g(x)$. This function h is even since

$$h(-x)=f(-x)g(-x)=[-f(x)][-g(x)]=f(x)g(x)=h(x).$$

In addition to the multiplicative properties 1–3, even and odd functions satisfy the following integral properties.

4. The integral of an odd function f over a symmetric interval $[-l,l]$ is zero; that is, $\int_{-l}^{l} f(x)\,dx=0$ if f is odd.
5. The integral of an even function f over the interval $[-l,l]$ is twice the integral of f over the interval $[0,l]$; that is,

$$\int_{-l}^{l} f(x)\,dx=2\int_0^l f(x)\,dx$$

if f is even.

PROOF OF PROPERTY 4. If f is odd, then the area under the curve of f between $-l$ and 0 is the negative of the area under the curve of f between 0 and l. Hence, $\int_{-l}^{l} f(x)\,dx=0$ if f is odd. □

PROOF OF PROPERTY 5. If f is even, then the area under the curve of f between $-l$ and 0 equals the area under the curve of f between 0 and l. Hence,

$$\int_{-l}^{l} f(x)\,dx=\int_{-l}^{0} f(x)\,dx+\int_0^l f(x)\,dx=2\int_0^l f(x)\,dx$$

if f is even. □

Concerning even and odd functions, we have the following important lemma.

Lemma 1.

(a) *The Fourier series for an even function is a pure cosine series; that is, it contains no terms of the form* $\sin n\pi x/l$.

(b) *The Fourier series for an odd function is a pure sine series; that is, it contains no terms of the form* $\cos n\pi x/l$.

PROOF. (a) If f is even, then the function $f(x)\sin n\pi x/l$ is odd. Thus, by Property 4, the coefficients

$$b_n=\frac{1}{l}\int_{-l}^{l} f(x)\sin\frac{n\pi x}{l}\,dx, \qquad n=1,2,3,\ldots$$

in the Fourier series for f are all zero.

(b) If f is odd, then the function $f(x)\cos n\pi x/l$ is also odd. Consequently, by Property 4, the coefficients

$$a_n=\frac{1}{l}\int_{-l}^{l} f(x)\cos\frac{n\pi x}{l}\,dx, \qquad n=0,1,2,\ldots$$

in the Fourier series for f are all zero. □

We are now in a position to prove the following extremely important extension of Theorem 2. This theorem will enable us to solve the heat conduction problem of Section 5.3 and many other boundary-value problems that arise in applications.

Theorem 3. *Let f and f' be piecewise continuous on the interval $0 \leqslant x \leqslant l$. Then, on this interval, $f(x)$ can be expanded in either a pure cosine series*

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\frac{n\pi x}{l},$$

or a pure sine series

$$\sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{l}.$$

In the former case, the coefficients a_n are given by the formula

$$a_n = \frac{2}{l}\int_0^l f(x)\cos\frac{n\pi x}{l}\,dx, \qquad n=0,1,2,\dots \tag{1}$$

while in the latter case, the coefficients b_n are given by the formula

$$b_n = \frac{2}{l}\int_0^l f(x)\sin\frac{n\pi x}{l}\,dx. \tag{2}$$

PROOF. Consider first the function

$$F(x) = \begin{cases} f(x), & 0 \leqslant x \leqslant l \\ f(-x), & -l \leqslant x < 0 \end{cases}.$$

The graph of $F(x)$ is described in Figure 1, and it is easily seen that F is even. (For this reason, F is called the even extension of f.) Hence, by

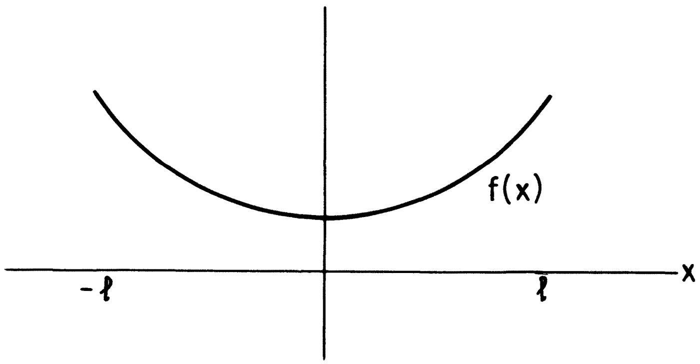


Figure 1. Graph of $F(x)$

Lemma 1, the Fourier series for F on the interval $-l \leqslant x \leqslant l$ is

$$F(x)=\frac{a_0}{2}+\sum_{n=1}^{\infty} a_n \cos\frac{n\pi x}{l}; \qquad a_n=\frac{1}{l}\int_{-l}^{l} F(x)\cos\frac{n\pi x}{l}\,dx. \tag{3}$$

Now, observe that the function $F(x)\cos n\pi x/l$ is even. Thus, by Property 5

$$a_n=\frac{2}{l}\int_0^l F(x)\cos\frac{n\pi x}{l}\,dx=\frac{2}{l}\int_0^l f(x)\cos\frac{n\pi x}{l}\,dx.$$

Finally, since $F(x)=f(x)$, $0\leqslant x\leqslant l$, we conclude from (3) that

$$f(x)=\frac{a_0}{2}+\sum_{n=1}^{\infty} a_n \cos\frac{n\pi x}{l}, \qquad 0\leqslant x\leqslant l.$$

Observe too, that the series (3) converges to $f(x)$ for $x=0$ and $x=l$.

To show that $f(x)$ can also be expanded in a pure sine series, we consider the function

$$G(x)=\begin{cases} f(x), & 0<x<l \\ -f(-x), & -l<x<0 \\ 0, & x=0, \pm l. \end{cases}$$

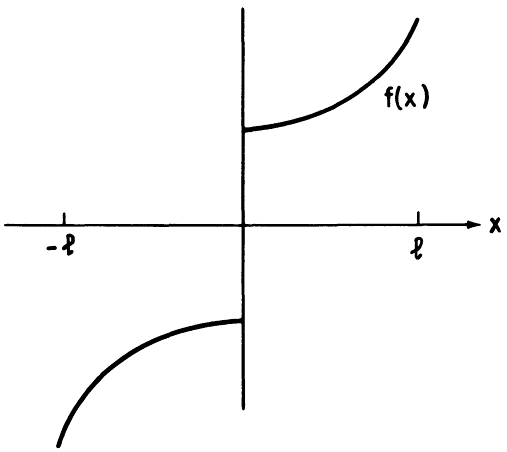


Figure 2. Graph of $G(x)$

The graph of $G(x)$ is described in Figure 2, and it is easily seen that G is odd. (For this reason, G is called the odd extension of f.) Hence, by Lemma 1, the Fourier series for G on the interval $-l\leqslant x\leqslant l$ is

$$G(x)=\sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{l}; \qquad b_n=\frac{1}{l}\int_{-l}^{l} G(x)\sin\frac{n\pi x}{l}\,dx. \tag{4}$$

Now, observe that the function $G(x)\sin n\pi x/l$ is even. Thus, by Property 5,

$$b_n=\frac{2}{l}\int_0^l G(x)\sin\frac{n\pi x}{l}\,dx=\frac{2}{l}\int_0^l f(x)\sin\frac{n\pi x}{l}\,dx.$$

Finally, since $G(x)=f(x), 0<x<l$, we conclude from (4) that

$$f(x)=\sum_{n=1}^{\infty} b_n\sin\frac{n\pi x}{l},\qquad 0<x<l.$$

Observe too, that the series (4) is zero for $x=0$ and $x=l$. □

Example 5. Expand the function $f(x)=1$ in a pure sine series on the interval $0<x<\pi$.

Solution. By Theorem 3, $f(x)=\sum_{n=1}^{\infty}b_n\sin nx$, where

$$b_n=\frac{2}{\pi}\int_0^{\pi}\sin nx\,dx=\frac{2}{n\pi}(1-\cos n\pi)=\begin{cases}0, & n \text{ even}\\ \dfrac{4}{n\pi}, & n \text{ odd}.\end{cases}$$

Hence,

$$1=\frac{4}{\pi}\left[\frac{\sin x}{1}+\frac{\sin 3x}{3}+\frac{\sin 5x}{5}+\dots\right],\qquad 0<x<\pi.$$

Example 6. Expand the function $f(x)=e^x$ in a pure cosine series on the interval $0\leqslant x\leqslant 1$.

Solution. By Theorem 3, $f(x)=a_0/2+\sum_{n=1}^{\infty}a_n\cos n\pi x$, where

$$a_0=2\int_0^1 e^x\,dx=2(e-1)$$

and

$$a_n=2\int_0^1 e^x\cos n\pi x\,dx=2\,\mathrm{Re}\int_0^1 e^x e^{in\pi x}\,dx$$

$$=2\,\mathrm{Re}\int_0^1 e^{(1+in\pi)x}\,dx=2\,\mathrm{Re}\left\{\frac{e^{1+in\pi}-1}{1+in\pi}\right\}$$

$$=2\,\mathrm{Re}\left\{\frac{(e\cos n\pi-1)(1-in\pi)}{1+n^2\pi^2}\right\}=\frac{2(e\cos n\pi-1)}{1+n^2\pi^2}.$$

Hence,

$$e^x=e-1+2\sum_{n=1}^{\infty}\frac{(e\cos n\pi-1)}{1+n^2\pi^2}\cos n\pi x,\qquad 0\leqslant x\leqslant 1.$$

EXERCISES

Expand each of the following functions in a Fourier cosine series on the prescribed interval.

1. $f(x)=e^{-x}; \qquad 0\leqslant x\leqslant 1$

2. $f(x)=\begin{cases} 0, & 0\leqslant x\leqslant 1 \\ 1, & 1<x\leqslant 2\end{cases}; \qquad 0\leqslant x\leqslant 2$

3. $f(x)=\begin{cases} x, & 0\leqslant x<a \\ a, & a\leqslant x\leqslant 2a\end{cases}; \qquad 0\leqslant x\leqslant 2a$

4. $f(x)=\cos^2 x; \qquad 0\leqslant x\leqslant \pi$

5. $f(x)=\begin{cases} x, & 0\leqslant x\leqslant l/2 \\ l-x, & l/2\leqslant x\leqslant l\end{cases}; \qquad 0\leqslant x\leqslant l$

Expand each of the following functions in a Fourier sine series on the prescribed interval.

6. $f(x)=e^{-x}; \qquad 0<x<1$

7. $f(x)=\begin{cases} 0, & 0<x\leqslant 1 \\ 1, & 1<x<2\end{cases}; \qquad 0<x<2$

8. $f(x)=\begin{cases} x, & 0<x<a \\ a, & a\leqslant x<2a\end{cases}; \qquad 0<x<2a$

9. $f(x)=2\sin x\cos x; \qquad 0<x<\pi$

10. $f(x)=\begin{cases} x, & 0<x<l/2 \\ l-x, & l/2\leqslant x<l\end{cases}; \qquad 0<x<l$

11. (a) Expand the function $f(x)=\sin x$ in a Fourier cosine series on the interval $0\leqslant x\leqslant\pi$.
(b) Expand the function $f(x)=\cos x$ in a Fourier sine series on the interval $0<x<\pi$.
(c) Can you expand the function $f(x)=\sin x$ in a Fourier cosine series on the interval $-\pi\leqslant x\leqslant\pi$? Explain.

5.6 Return to the heat equation

We return now to the boundary-value problem

$$\frac{\partial u}{\partial t}=\alpha^2\frac{\partial^2 u}{\partial x^2}; \qquad \begin{cases} u(x,0)=f(x) \\ u(0,t)=u(l,t)=0\end{cases}. \tag{1}$$

We showed in Section 5.3 that the function

$$u(x,t)=\sum_{n=1}^{\infty} c_n \sin\frac{n\pi x}{l}\, e^{-\alpha^2 n^2\pi^2 t/l^2}$$

is (formally) a solution of the boundary-value problem

$$\frac{\partial u}{\partial t}=\alpha^2\frac{\partial^2 u}{\partial x^2}; \qquad u(0,t)=u(l,t)=0 \tag{2}$$

for any choice of constants $c_1, c_2, \ldots$. This led us to ask whether we can find constants $c_1, c_2, \ldots$ such that

$$u(x,0)=\sum_{n=1}^{\infty} c_n \sin\frac{n\pi x}{l}=f(x), \qquad 0<x<l. \tag{3}$$

As we showed in Section 5.5, the answer to this question is yes; if we choose

$$c_n=\frac{2}{l}\int_0^l f(x)\sin\frac{n\pi x}{l}\,dx,$$

then the Fourier series $\sum_{n=1}^{\infty} c_n \sin n\pi x/l$ converges to $f(x)$ if f is continuous at the point x. Thus,

$$u(x,t)=\frac{2}{l}\sum_{n=1}^{\infty}\left[\int_0^l f(x)\,\sin\frac{n\pi x}{l}\,dx\right]\sin\frac{n\pi x}{l}e^{-\alpha^2 n^2\pi^2 t/l^2} \tag{4}$$

is the desired solution of (1).

Remark. Strictly speaking, the solution (4) cannot be regarded as the solution of (1) until we rigorously justify all the limiting processes involved. Specifically, we must verify that the function $u(x,t)$ defined by (4) actually has partial derivatives with respect to x and t, and that $u(x,t)$ satisfies the heat equation $u_t=\alpha^2 u_{xx}$. (It is not true, necessarily, that an infinite sum of solutions of a linear differential equation is again a solution. Indeed, an infinite sum of solutions of a given differential equation need not even be differentiable.) However, in the case of (4) it is possible to show (see Exercise 3) that $u(x,t)$ has partial derivatives with respect to x and t of all orders, and that $u(x,t)$ satisfies the boundary-value problem (1). The argument rests heavily upon the fact that the infinite series (4) converges very rapidly, due to the presence of the factor $e^{-\alpha^2 n^2\pi^2 t/l^2}$. Indeed, the function $u(x,t)$, for fixed $t>0$, is even analytic for $0<x<l$. Thus, heat conduction is a diffusive process which instantly smooths out any discontinuities that may be present in the initial temperature distribution in the rod. Finally, we observe that $\lim_{t\to\infty} u(x,t)=0$, for all x, regardless of the initial temperature in the rod. This is in accord with our physical intuition that the heat distribution in the rod should ultimately approach a "steady state"; that is, a state in which the temperature does not change with time.

Example 1. A thin aluminum bar ($\alpha^2=0.86\ \text{cm}^2/\text{s}$) 10 cm long is heated to a uniform temperature of 100°C. At time $t=0$, the ends of the bar are plunged into an ice bath at 0°C, and thereafter they are maintained at this temperature. No heat is allowed to escape through the lateral surface of the bar. Find an expression for the temperature at any point in the bar at any later time t.

Solution. Let $u(x,t)$ denote the temperature in the bar at the point x at time t. This function satisfies the boundary-value problem

$$\frac{\partial u}{\partial t}=0.86\frac{\partial^2 u}{\partial x^2}; \qquad \begin{cases} u(x,0)=100, & 0<x<10 \\ u(0,t)=u(10,t)=0 \end{cases} \tag{5}$$

The solution of (5) is

$$u(x,t)=\sum_{n=1}^{\infty} c_n \sin\frac{n\pi x}{10} e^{-0.86n^2\pi^2 t/100}$$

where

$$c_n=\frac{1}{5}\int_0^{10} 100\sin\frac{n\pi x}{10}\,dx=\frac{200}{n\pi}(1-\cos n\pi).$$

Notice that $c_n=0$ if n is even, and $c_n=400/n\pi$ if n is odd. Hence,

$$u(x,t)=\frac{400}{\pi}\sum_{n=1}^{\infty}\frac{\sin(2n+1)\frac{\pi x}{10}}{(2n+1)} e^{-0.86(2n+1)^2\pi^2 t/100}.$$

There are several other problems of heat conduction which can be solved by the method of separation of variables. Example 2 below treats the case where the ends of the bar are also insulated, and Exercise 4 treats the case where the ends of the bar are kept at constant, but nonzero temperatures T_1 and T_2.

Example 2. Consider a thin metal rod of length l and thermal diffusivity α^2, whose sides and ends are insulated so that there is no passage of heat through them. Let the initial temperature distribution in the rod be $f(x)$. Find the temperature distribution in the rod at any later time t.
Solution. Let $u(x,t)$ denote the temperature in the rod at the point x at time t. This function satisfies the boundary-value problem

$$\frac{\partial u}{\partial t}=\alpha^2\frac{\partial^2 u}{\partial x^2}; \qquad \begin{cases} u(x,0)=f(x), & 0<x<l \\ u_x(0,t)=u_x(l,t)=0 \end{cases} \tag{6}$$

We solve this problem in two steps. First, we will find infinitely many solutions $u_n(x,t)=X_n(x)T_n(t)$ of the boundary-value problem

$$u_t=\alpha^2 u_{xx}; \qquad u_x(0,t)=u_x(l,t)=0, \tag{7}$$

and then we will find constants $c_0, c_1, c_2, \ldots$ such that

$$u(x,t)=\sum_{n=0}^{\infty} c_n u_n(x,t)$$

satisfies the initial condition $u(x,0)=f(x)$.

Step 1: Let $u(x,t)=X(x)T(t)$. Computing

$$\frac{\partial u}{\partial t}=XT' \quad \text{and} \quad \frac{\partial^2 u}{\partial x^2}=X''T$$

we see that $u(x,t)$ is a solution of $u_t=\alpha^2 u_{xx}$ if

$$XT'=\alpha^2 X''T, \quad \text{or} \quad \frac{X''}{X}=\frac{T'}{\alpha^2 T}. \tag{8}$$

As we showed in Section 5.3, Equation (8) implies that

$$X''+\lambda X=0 \quad \text{and} \quad T'+\lambda\alpha^2 T=0$$

for some constant λ. In addition, the boundary conditions

$$0=u_x(0,t)=X'(0)T(t) \quad \text{and} \quad 0=u_x(l,t)=X'(l)T(t)$$

imply that $X'(0)=0$ and $X'(l)=0$. Hence $u(x,t)=X(x)T(t)$ is a solution of (7) if

$$X''+\lambda X=0; \qquad X'(0)=0, \quad X'(l)=0 \tag{9}$$

and

$$T'+\lambda\alpha^2 T=0. \tag{10}$$

At this point, the constant λ is arbitrary. However, the boundary-value problem (9) has a nontrivial solution $X(x)$ (see Exercise 1, Section 5.1) only if $\lambda=n^2\pi^2/l^2$, $n=0,1,2,\ldots$, and in this case

$$X(x)=X_n(x)=\cos\frac{n\pi x}{l}.$$

Equation (10), in turn, implies that $T(t)=e^{-\alpha^2 n^2\pi^2 t/l^2}$. Hence,

$$u_n(x,t)=\cos\frac{n\pi x}{l}\, e^{-\alpha^2 n^2\pi^2 t/l^2}$$

is a solution of (7) for every nonnegative integer n.

Step 2: Observe that the linear combination

$$u(x,t)=\frac{c_0}{2}+\sum_{n=1}^{\infty} c_n \cos\frac{n\pi x}{l}\, e^{-\alpha^2 n^2\pi^2 t/l^2}$$

is a solution (formally) of (7) for every choice of constants $c_0, c_1, c_2, \ldots$. Its initial value is

$$u(x,0)=\frac{c_0}{2}+\sum_{n=1}^{\infty} c_n \cos\frac{n\pi x}{l}.$$

Thus, in order to satisfy the initial condition $u(x,0)=f(x)$ we must choose constants $c_0, c_1, c_2, \ldots$ such that

$$f(x)=\frac{c_0}{2}+\sum_{n=1}^{\infty} c_n \cos\frac{n\pi x}{l}, \qquad 0\leqslant x\leqslant l.$$

In other words, we must expand f in a Fourier cosine series on the interval

$0 \leqslant x \leqslant l$. This is precisely the situation in Theorem 3 of Section 5.5, and we conclude, therefore, that

$$c_n = \frac{2}{l}\int_0^l f(x)\cos\frac{n\pi x}{l}\,dx.$$

Hence,

$$u(x,t) = \frac{1}{l}\int_0^l f(x)\,dx + \frac{2}{l}\sum_{n=1}^{\infty}\left[\int_0^l f(x)\cos\frac{n\pi x}{l}\,dx\right]\cos\frac{n\pi x}{l}\,e^{-\alpha^2 n^2\pi^2 t/l^2} \quad (11)$$

is the desired solution of (6).

Remark. Observe from (11) that the temperature in the rod ultimately approaches the steady state temperature

$$\frac{1}{l}\int_0^l f(x)\,dx.$$

This steady state temperature can be interpreted as the "average" of the initial temperature distribution in the rod.

Exercises

1. The ends $x=0$ and $x=10$ of a thin aluminum bar ($\alpha^2=0.86$) are kept at $0°$C, while the surface of the bar is insulated. Find an expression for the temperature $u(x,t)$ in the bar if initially

(a) $u(x,0)=70, \quad 0<x<10$

(b) $u(x,0)=70\cos x, \quad 0<x<10$

(c) $u(x,0)=\begin{cases} 10x, & 0<x<5 \\ 10(10-x), & 5\leqslant x<10 \end{cases}$

(d) $u(x,0)=\begin{cases} 0, & 0<x<3 \\ 65, & 3\leqslant x<10 \end{cases}$

2. The ends and sides of a thin copper bar ($\alpha^2=1.14$) of length 2 are insulated so that no heat can pass through them. Find the temperature $u(x,t)$ in the bar if initially

(a) $u(x,0)=65\cos^2\pi x, \quad 0\leqslant x\leqslant 2$

(b) $u(x,0)=70\sin x, \quad 0\leqslant x\leqslant 2$

(c) $u(x,0)=\begin{cases} 60x, & 0\leqslant x<1 \\ 60(2-x), & 1\leqslant x\leqslant 2 \end{cases}$

(d) $u(x,0)=\begin{cases} 0, & 0\leqslant x<1 \\ 75, & 1\leqslant x\leqslant 2 \end{cases}$

3. Verify that the function $u(x,t)$ defined by (4) satisfies the heat equation. *Hint*: Use the Cauchy ratio test to show that the infinite series (4) can be differentiated term by term with respect to x and t.

4. A steady state solution $u(x,t)$ of the heat equation $u_t=\alpha^2 u_{xx}$ is a solution $u(x,t)$ which does not change with time.

(a) Show that all steady state solutions of the heat equation are linear functions of x; i.e., $u(x)=Ax+B$.

(b) Find a steady state solution of the boundary-value problem

$$u_t = \alpha^2 u_{xx}; \qquad u(0,t) = T_1, \quad u(l,t) = T_2.$$

(c) Solve the heat conduction problem

$$u_t = \alpha^2 u_{xx}; \qquad \begin{cases} u(x,0) = 75, & 0 < x < 1 \\ u(0,t) = 20, & u(1,t) = 60 \end{cases}$$

Hint: Let $u(x,t) = v(x) + w(x,t)$ where $v(x)$ is the steady state solution of the boundary-value problem $u_t = \alpha^2 u_{xx}$; $u(0,t) = 20$, $u(1,t) = 60$.

5. (a) The ends of a copper rod ($\alpha^2 = 1.14$) 10 cm long are maintained at 0°C, while the center of the rod is maintained at 100°C by an external heat source. Show that the temperature in the rod will ultimately approach a steady state distribution regardless of the initial temperature in the rod. *Hint*: Split this problem into two boundary-value problems.
(b) Assume that the temperature in the rod is at its steady state distribution. At time $t = 0$, the external heat source is removed from the center of the rod, and placed at the left end of the rod. Find the temperature in the rod at any later time t.

6. Solve the boundary-value problem

$$u_t = u_{xx} + u; \qquad \begin{cases} u(x,0) = \cos x, & 0 < x < 1 \\ u(0,t) = 0, & u(1,t) = 0. \end{cases}$$

5.7 The wave equation

We consider now the boundary-value problem

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}; \qquad \begin{cases} u(x,0) = f(x), & u_t(x,0) = g(x) \\ u(0,t) = u(l,t) = 0 \end{cases} \tag{1}$$

which characterizes the propagation of waves in various media, and the mechanical vibrations of an elastic string. This problem, too, can be solved by the method of separation of variables. Specifically, we will (a) find solutions $u_n(x,t) = X_n(x)T_n(t)$ of the boundary-value problem

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}; \qquad u(0,t) = u(l,t) = 0 \tag{2}$$

and (b) find the solution $u(x,t)$ of (1) by taking a suitable linear combination of the functions $u_n(x,t)$.

(a) Let $u(x,t) = X(x)T(t)$. Computing

$$\frac{\partial^2 u}{\partial t^2} = XT'' \quad \text{and} \quad \frac{\partial^2 u}{\partial x^2} = X''T$$

we see that $u(x,t) = X(x)T(t)$ is a solution of the wave equation $u_{tt} = c^2 u_{xx}$

if $XT''=c^2X''T$, or

$$\frac{T''}{c^2T}=\frac{X''}{X}. \tag{3}$$

Next, we observe that the left-hand side of (3) is a function of t alone, while the right-hand side is a function of x alone. This implies that

$$\frac{T''}{c^2T}=-\lambda=\frac{X''}{\Lambda}$$

for some constant λ. In addition, the boundary conditions

$$0=u(0,t)=X(0)T(t), \quad \text{and} \quad 0=u(l,t)=X(l)T(t)$$

imply that $X(0)=0$ and $X(l)=0$. Hence $u(x,t)=X(x)T(t)$ is a solution of (2) if

$$X''+\lambda X=0; \qquad X(0)=X(l)=0 \tag{4}$$

and

$$T''+\lambda c^2T=0. \tag{5}$$

At this point, the constant λ is arbitrary. However, the boundary-value problem (4) has a nontrivial solution $X(x)$ only if $\lambda=\lambda_n=n^2\pi^2/l^2$, and in this case,

$$X(x)=X_n(x)=\sin\frac{n\pi x}{l}.$$

Equation (5), in turn, implies that

$$T(t)=T_n(t)=a_n\cos\frac{n\pi ct}{l}+b_n\sin\frac{n\pi ct}{l}.$$

Hence,

$$u_n(x,t)=\sin\frac{n\pi x}{l}\left[a_n\cos\frac{n\pi ct}{l}+b_n\sin\frac{n\pi ct}{l}\right]$$

is a nontrivial solution of (2) for every positive integer n, and every pair of constants a_n, b_n.

(b) The linear combination

$$u(x,t)=\sum_{n=1}^{\infty}\sin\frac{n\pi x}{l}\left[a_n\cos\frac{n\pi ct}{l}+b_n\sin\frac{n\pi ct}{l}\right]$$

formally satisfies the boundary-value problem (2) and the initial conditions

$$u(x,0)=\sum_{n=1}^{\infty}a_n\sin\frac{n\pi x}{l} \quad \text{and} \quad u_t(x,0)=\sum_{n=1}^{\infty}\frac{n\pi c}{l}b_n\sin\frac{n\pi x}{l}.$$

Thus, to satisfy the initial conditions $u(x,0)=f(x)$ and $u_t(x,0)=g(x)$, we must choose the constants a_n and b_n such that

$$f(x)=\sum_{n=1}^{\infty}a_n\sin\frac{n\pi x}{l} \quad \text{and} \quad g(x)=\sum_{n=1}^{\infty}\frac{n\pi c}{l}b_n\sin\frac{n\pi x}{l}$$

on the interval $0<x<l$. In other words, we must expand the functions $f(x)$ and $g(x)$ in Fourier sine series on the interval $0<x<l$. This is precisely the situation in Theorem 3 of Section 5.5, and we conclude, therefore, that

$$a_n=\frac{2}{l}\int_0^l f(x)\sin\frac{n\pi x}{l}\,dx \quad\text{and}\quad b_n=\frac{2}{n\pi c}\int_0^l g(x)\sin\frac{n\pi x}{l}\,dx.$$

For simplicity, we now restrict ourselves to the case where $g(x)$ is zero; that is, the string is released with zero initial velocity. In this case the displacement $u(x,t)$ of the string at any time $t>0$ is given by the formula

$$u(x,t)=\sum_{n=1}^{\infty} a_n\sin\frac{n\pi x}{l}\cos\frac{n\pi ct}{l}; \qquad a_n=\frac{2}{l}\int_0^l f(x)\sin\frac{n\pi x}{l}\,dx. \tag{6}$$

There is a physical significance to the various terms in (6). Each term represents a particular mode in which the string vibrates. The first term ($n=1$) represents the first mode of vibration in which the string oscillates about its equilibrium position with frequency

$$\omega_1=\frac{1}{2\pi}\frac{\pi c}{l}=\frac{c}{2l}\text{ cycles per second.}$$

This lowest frequency is called the fundamental frequency, or first harmonic of the string. Similarly, the nth mode has a frequency

$$\omega_n=\frac{1}{2\pi}\frac{n\pi c}{l}=n\omega_1\text{ cycles per second}$$

which is called the nth harmonic of the string.

In the case of the vibrating string, all the harmonic frequencies are integer multiples of the fundamental frequency ω_1. Thus, we have music in this case. Of course, if the tension in the string is not large enough, then the sound produced will be of such very low frequency that it is not in the audible range. As we increase the tension in the string, we increase the frequency, and the result is a musical note that can be heard by the human ear.

Justification of solution. We cannot prove directly, as in the case of the heat equation, that the function $u(x,t)$ defined in (6) is a solution of the wave equation. Indeed, we cannot even prove directly that the infinite series (6) has a partial derivative with respect to t and x. For example, on formally computing u_t, we obtain that

$$u_t=-\sum_{n=1}^{\infty}\frac{n\pi c}{l}a_n\sin\frac{n\pi x}{l}\sin\frac{n\pi ct}{l}$$

and due to the presence of the factor n, this series may not converge. However, there is an alternate way to establish the validity of the solution (6). At the same time, we will gain additional insight into the structure of

$u(x,t)$. Observe first that

$$\sin\frac{n\pi x}{l}\cos\frac{n\pi ct}{l}=\frac{1}{2}\left[\sin\frac{n\pi}{l}(x-ct)+\sin\frac{n\pi}{l}(x+ct)\right].$$

Next, let F be the odd periodic extension of f on the interval $-l<x<l$; that is,

$$F(x)=\begin{cases} f(x), & 0<x<l \\ -f(-x), & -l<x<0 \end{cases} \quad \text{and} \quad F(x+2l)=F(x).$$

It is easily verified (see Exercise 6) that the Fourier series for F is

$$F(x)=\sum_{n=1}^{\infty} c_n\sin\frac{n\pi x}{l}, \qquad c_n=\frac{2}{l}\int_0^l f(x)\sin\frac{n\pi x}{l}\,dx.$$

Therefore, we can write $u(x,t)$ in the form

$$u(x,t)=\tfrac{1}{2}\left[F(x-ct)+F(x+ct)\right] \tag{7}$$

and it is now a trivial matter to show that $u(x,t)$ satisfies the wave equation if $f(x)$ has two continuous derivatives.

Equation (7) has the following interpretation. If we plot the graph of the function $y=F(x-ct)$ for any fixed t, we see that it is the same as the graph of $y=F(x)$, except that it is displaced a distance ct in the positive x direction, as shown in Figures 1a and 1b. Thus, $F(x-ct)$ is a *wave* which travels with velocity c in the positive x direction. Similarly, $F(x+ct)$ is a wave which travels with velocity c in the negative x direction. The number c represents the velocity with which a disturbance propagates along the string. If a disturbance occurs at the point x_0, then it will be felt at the point x after a time $t=(x-x_0)/c$ has elapsed. Thus, the wave equation, or some form of it, characterizes the propagation of waves in a medium where disturbances (or signals) travel with a finite, rather than infinite, velocity.

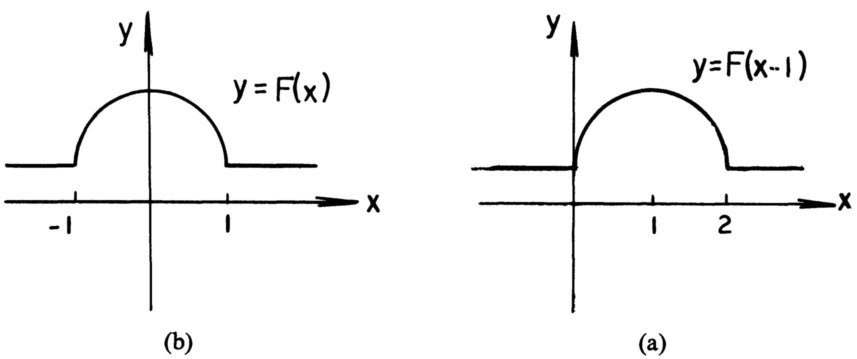


Figure 1

Exercises

Solve each of the following boundary-value problems.

1. $u_{tt}=c^2u_{xx};\quad \begin{cases} u(x,0)=\cos x-1, & u_t(x,0)=0, \quad 0\leqslant x\leqslant 2\pi \\ u(0,t)=0, & u(2\pi,t)=0 \end{cases}$

2. $u_{tt}=c^2u_{xx};\quad \begin{cases} u(x,0)=0, & u_t(x,0)=1, \quad 0\leqslant x\leqslant 1 \\ u(0,t)=0, & u(1,t)=0 \end{cases}$

3. $\begin{matrix} u_{tt}=c^2u_{xx} \\ u(0,t)=u(3,t)=0 \end{matrix};\quad u(x,0)=\begin{cases} x, & 0\leqslant x\leqslant 1, \quad u_t(x,0)=0 \\ 1, & 1\leqslant x\leqslant 2 \\ 3-x, & 2\leqslant x\leqslant 3 \end{cases}$

4. $u_{tt}=c^2u_{xx};\quad \begin{cases} u(x,0)=x\cos \pi x/2, & u_t(x,0)=0, \quad 0\leqslant x\leqslant 1 \\ u(0,t)=0, & u(1,t)=0 \end{cases}$

5. A string of length 10 ft is raised at the middle to a distance of 1 ft, and then released. Describe the motion of the string, assuming that $c^2=1$.

6. Let F be the odd periodic extension of f on the interval $-l<x<l$. Show that the Fourier series

$$\frac{2}{l}\sum_{n=1}^{\infty}\left[\int_0^l f(x)\sin\frac{n\pi x}{l}\,dx\right]\sin\frac{n\pi x}{l}$$

converges to $F(x)$ if F is continuous at x.

7. Show that the transformation $\xi=x-ct$, $\eta=x+ct$ reduces the wave equation to the equation $u_{\xi\eta}=0$. Conclude, therefore, that every solution $u(x,t)$ of the wave equation is of the form $u(x,t)=F(x-ct)+G(x+ct)$ for some functions F and G.

8. Show that the solution of the boundary-value problem

$$u_{tt}=c^2u_{xx};\qquad \begin{cases} u(x,0)=f(x), & u_t(x,0)=g(x), \quad -l<x<l \\ u(0,t)=u(l,t)=0 \end{cases}$$

is

$$u(x,t)=\frac{1}{2}[F(x-ct)+F(x+ct)]+\frac{1}{2c}\int_{x-ct}^{x+ct}g(s)\,ds$$

where F is the odd periodic extension of f.

9. The wave equation in two dimensions is $u_{tt}=c^2(u_{xx}+u_{yy})$. Find solutions of this equation by the method of separation of variables.

10. Solve the boundary-value problem

$$u_{tt}=c^2u_{xx}+u;\qquad \begin{cases} u(x,0)=f(x), & u_t(x,0)=0, \quad 0<x<l \\ u(0,t)=0, & u(l,t)=0 \end{cases}.$$

5.8 Laplace's equation

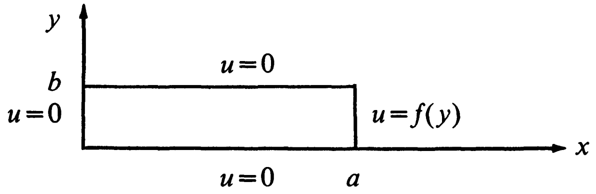
We consider now Laplace's equation

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=0. \tag{1}$$

As we mentioned in Section 5.2, two important boundary-value problems that arise in connection with (1) are the Dirichlet problem and the Neumann problem. In a Dirichlet problem we seek a function $u(x,y)$ which satisfies Laplace's equation inside a region R, and which assumes prescribed values on the boundary of R. In a Neumann problem, we seek a function $u(x,y)$ which satisfies Laplace's equation inside a region R, and whose derivative in the direction normal to the boundary of R takes on prescribed values. Both of these problems can be solved by the method of separation of variables if R is a rectangle.

Example 1. Find a function $u(x,y)$ which satisfies Laplace's equation in the rectangle $0<x<a$, $0<y<b$ and which also satisfies the boundary conditions

$$\begin{aligned} u(x,0)&=0, \qquad & u(x,b)&=0\\ u(0,y)&=0, \qquad & u(a,y)&=f(y) \end{aligned}$$



(2)

Solution. We solve this problem in two steps. First, we will find functions $u_n(x,y)=X_n(x)Y_n(y)$ which satisfy the boundary-value problem

$$u_{xx}+u_{yy}=0; \qquad u(x,0)=0, \quad u(x,b)=0, \quad u(0,y)=0. \tag{3}$$

Then, we will find constants c_n such that the linear combination

$$u(x,y)=\sum_{n=1}^{\infty} c_n u_n(x,y)$$

satisfies the boundary condition $u(a,y)=f(y)$.

Step 1: Let $u(x,y)=X(x)Y(y)$. Computing $u_{xx}=X''Y$ and $u_{yy}=XY''$, we see that $u(x,y)=X(x)Y(y)$ is a solution of Laplace's equation if $X''Y+XY''=0$, or

$$\frac{Y''}{Y}=-\frac{X''}{X}. \tag{4}$$

Next, we observe that the left-hand side of (4) is a function of y alone, while the right-hand side is a function of x alone. This implies that

$$\frac{Y''}{Y} = -\frac{X''}{X} = -\lambda.$$

for some constant λ. In addition, the boundary conditions

$$0 = u(x,0) = X(x)Y(0), \qquad 0 = u(x,b) = X(x)Y(b),$$
$$0 = u(0,y) = X(0)Y(y)$$

imply that $Y(0)=0$, $Y(b)=0$, and $X(0)=0$. Hence $u(x,y)=XY$ is a solution of (3) if

$$Y'' + \lambda Y = 0; \qquad Y(0)=0, \quad Y(b)=0 \tag{5}$$

and

$$X'' - \lambda X = 0, \qquad X(0) = 0. \tag{6}$$

At this point the constant λ is arbitrary. However, the boundary-value problem (5) has a nontrivial solution $Y(y)$ only if $\lambda = \lambda_n = n^2\pi^2/b^2$, and in this case,

$$Y(y) = Y_n(y) = \sin n\pi y/b.$$

Equation (6), in turn, implies that $X_n(x)$ is proportional to $\sinh n\pi x/b$. (The differential equation $X'' - (n^2\pi^2/b^2)X = 0$ implies that $X(x) = c_1\cosh n\pi x/b + c_2 \sinh n\pi x/b$ for some choice of constants c_1, c_2, and the initial condition $X(0)=0$ forces c_1 to be zero.) We conclude, therefore, that

$$u_n(x,y) = \sinh\frac{n\pi x}{b}\sin\frac{n\pi y}{b}$$

is a solution of (3) for every positive integer n.

Step 2: The function

$$u(x,y) = \sum_{n=1}^{\infty} c_n \sinh\frac{n\pi x}{b}\sin\frac{n\pi y}{b}$$

is a solution (formally) of (3) for every choice of constants $c_1, c_2, \ldots$. Its value at $x=a$ is

$$u(a,y) = \sum_{n=1}^{\infty} c_n \sinh\frac{n\pi a}{b}\sin\frac{n\pi y}{b}.$$

Therefore, we must choose the constants c_n such that

$$f(y) = \sum_{n=1}^{\infty} c_n \sinh\frac{n\pi a}{b}\sin\frac{n\pi y}{b}, \qquad 0<y<b.$$

In other words, we must expand f in a Fourier sine series on the interval $0<y<b$. This is precisely the situation described in Theorem 3 of Section

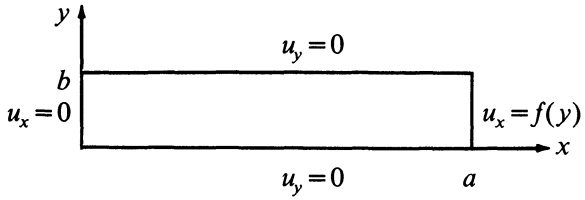
5.5, and we conclude, therefore, that

$$c_n = \frac{2}{b\sinh\frac{n\pi a}{b}} \int_0^b f(y)\sin\frac{n\pi y}{b}\,dy, \qquad n=1,2,\ldots.$$

Remark. The method of separation of variables can always be used to solve the Dirichlet problem for a rectangle R if u is zero on three sides of R. We can solve an arbitrary Dirichlet problem for a rectangle R by splitting it up into four problems where u is zero on three sides of R (see Exercises 1–4).

Example 2. Find a function $u(x,y)$ which satisfies Laplace's equation in the rectangle $0<x<a$, $0<y<b$, and which also satisfies the boundary conditions

$$\begin{aligned} &u_y(x,0)=0, \quad u_y(x,b)=0 \\ &u_x(0,y)=0, \quad u_x(a,y)=f(y) \end{aligned} \tag{7}$$



Solution. We attempt to solve this problem in two steps. First, we will find functions $u_n(x,y)=X_n(x)Y_n(y)$ which satisfy the boundary-value problem

$$u_{xx}+u_{yy}=0; \qquad u_y(x,0)=0, \quad u_y(x,b)=0, \quad \text{and} \quad u_x(0,y)=0. \tag{8}$$

Then, we will try and find constants c_n such that the linear combination $u(x,y)=\sum_{n=0}^{\infty}c_n u_n(x,y)$ satisfies the boundary condition $u_x(a,y)=f(y)$.
Step 1: Set $u(x,y)=X(x)Y(y)$. Then, as in Example 1,

$$\frac{Y''}{Y} = -\frac{X''}{X} = -\lambda$$

for some constant λ. The boundary conditions

$$\begin{aligned} &0=u_y(x,0)=X(x)Y'(0), \qquad 0=u_y(x,b)=X(x)Y'(b), \\ &0=u_x(0,y)=X'(0)Y(y) \end{aligned}$$

imply that

$$Y'(0)=0, \quad Y'(b)=0 \quad \text{and} \quad X'(0)=0.$$

Hence, $u(x,y)=X(x)Y(y)$ is a solution of (8) if

$$Y''+\lambda Y=0; \qquad Y'(0)=0, \quad Y'(b)=0 \tag{9}$$

and

$$X'' - \lambda X = 0; \qquad X'(0) = 0. \tag{10}$$

At this point, the constant λ is arbitrary. However, the boundary-value problem (9) has a nontrivial solution $Y(y)$ only if $\lambda = \lambda_n = n^2\pi^2/b^2, n = 0, 1, 2, \ldots$, and in this case

$$Y(y) = Y_n(y) = \cos n\pi y/b.$$

Equation (10), in turn, implies that $X(x)$ is proportional to $\cosh n\pi x/b$. (The differential equation $X'' - n^2\pi^2 X/b^2 = 0$ implies that $X(x) = c_1 \cosh n\pi x/b + c_2 \sinh n\pi x/b$ for some choice of constants c_1, c_2, and the boundary condition $X'(0) = 0$ forces c_2 to be zero.) We conclude, therefore, that

$$u_n(x,y) = \cosh\frac{n\pi x}{b}\cos\frac{n\pi y}{b}$$

is a solution of (8) for every nonnegative integer n.

Step 2: The function

$$u(x,y) = \frac{c_0}{2} + \sum_{n=1}^{\infty} c_n \cosh\frac{n\pi x}{b}\cos\frac{n\pi y}{b}$$

is a solution (formally) of (8) for every choice of constants $c_0, c_1, c_2, \ldots$. The value of u_x at $x = a$ is

$$u_x(a,y) = \sum_{n=1}^{\infty} \frac{n\pi}{b} c_n \sinh\frac{n\pi a}{b}\cos\frac{n\pi y}{b}.$$

Therefore, we must choose the constants $c_1, c_2, \ldots$, such that

$$f(y) = \sum_{n=1}^{\infty} \frac{n\pi}{b} c_n \sinh\frac{n\pi a}{b}\cos\frac{n\pi y}{b}, \qquad 0 < y < b. \tag{11}$$

Now, Theorem 3 of Section 5.5 states that we can expand $f(y)$ in the cosine series

$$f(y) = \frac{1}{b}\int_0^b f(y)\,dy + \frac{2}{b}\sum_{n=1}^{\infty}\left[\int_0^b f(y)\cos\frac{n\pi y}{b}\,dy\right]\cos\frac{n\pi y}{b} \tag{12}$$

on the interval $0 \leqslant y \leqslant b$. However, we cannot equate coefficients in (11) and (12) since the series (11) has no constant term. Therefore, the condition

$$\int_0^b f(y)\,dy = 0$$

is necessary for this Neumann problem to have a solution. If this is the case, then

$$c_n=\frac{2}{n\pi\sinh\frac{n\pi a}{b}}\int_0^b f(y)\cos\frac{n\pi y}{b}\,dy, \qquad n\geq 1.$$

Finally, note that c_0 remains arbitrary, and thus the solution $u(x,y)$ is only determined up to an additive constant. This is a property of all Neumann problems.

EXERCISES

Solve each of the following Dirichlet problems.

1. $\begin{matrix} u_{xx}+u_{yy}=0 \\ 0<x<a,\quad 0<y<b \end{matrix};\quad \begin{matrix} u(x,0)=0, & u(x,b)=0 \\ u(a,y)=0, & u(0,y)=f(y) \end{matrix}$

2. $\begin{matrix} u_{xx}+u_{yy}=0 \\ 0<x<a,\quad 0<y<b \end{matrix};\quad \begin{matrix} u(0,y)=0, & u(a,y)=0 \\ u(x,0)=0, & u(x,b)=f(x) \end{matrix}$

Remark. You can do this problem the long way, by separation of variables, or you can try something smart, like interchanging x with y and using the result of Example 1 in the text.

3. $\begin{matrix} u_{xx}+u_{yy}=0 \\ 0<x<a,\quad 0<y<b \end{matrix};\quad \begin{matrix} u(0,y)=0, & u(a,y)=0 \\ u(x,b)=0, & u(x,0)=f(x) \end{matrix}$

4. $\begin{matrix} u_{xx}+u_{yy}=0 \\ 0<x<a,\quad 0<y<b \end{matrix};\quad \begin{matrix} u(x,0)=f(x), & u(x,b)=g(x) \\ u(0,y)=h(y), & u(a,y)=k(y) \end{matrix}$

Hint: Write $u(x,y)$ as the sum of 4 functions, each of which is zero on three sides of the rectangle.

5. $\begin{matrix} u_{xx}+u_{yy}=0 \\ 0<x<a,\quad 0<y<b \end{matrix};\quad \begin{matrix} u(x,0)=0, & u(x,b)=1 \\ u(0,y)=0, & u(a,y)=1 \end{matrix}$

6. $\begin{matrix} u_{xx}+u_{yy}=0 \\ 0<x<a,\quad 0<y<b \end{matrix};\quad \begin{matrix} u(x,b)=0, & u(x,0)=1 \\ u(0,y)=0, & u(a,y)=1 \end{matrix}$

7. $\begin{matrix} u_{xx}+u_{yy}=0 \\ 0<x<a,\quad 0<y<b \end{matrix};\quad \begin{matrix} u(x,0)=1, & u(x,b)=1 \\ u(0,y)=0, & u(a,y)=1 \end{matrix}$

8. $\begin{matrix} u_{xx}+u_{yy}=0 \\ 0<x<a,\quad 0<y<b \end{matrix};\quad \begin{matrix} u(x,0)=1, & u(x,b)=1 \\ u(0,y)=1, & u(a,y)=1 \end{matrix}$

Remark. *Think!*

9. Solve the boundary-value problem

$$\begin{matrix} u_{xx}+u_{yy}=u \\ 0<x<1,\quad 0<y<1 \end{matrix};\quad \begin{matrix} u(x,0)=0, & u(x,1)=0 \\ u(0,y)=0, & u(1,y)=y \end{matrix}$$

10. (a) For which functions $f(y)$ can we find a solution $u(x,y)$ of the Neumann problem

$$\begin{array}{ll} u_{xx}+u_{yy}=0 & ; \quad u_x(1,y)=0, \quad u_x(0,y)=f(y) \\ 0<x<1, \quad 0<y<1 & \quad\; u_y(x,0)=0, \quad u_y(x,1)=0 \end{array}$$

(b) Solve this problem if $f(y)=\sin 2\pi y$.

11. Laplace's equation in three dimensions is

$$u_{xx}+u_{yy}+u_{zz}=0.$$

Assuming that $u=X(x)Y(y)Z(z)$, find 3 ordinary differential equations satisfied by X, Y, and Z.

6 Sturm-Liouville boundary value problems

6.1 Introduction

In Section 5.5 we described the remarkable result that an arbitrary piecewise differentiable function $f(x)$ could be expanded in either a pure sine series of the form

$$f(x) = \sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{l} \tag{1}$$

or a pure cosine series of the form

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\frac{n\pi x}{l} \tag{2}$$

on the interval $0 < x < l$. We were led to the trigonometric functions appearing in the series (1) and (2) by considering the 2 point boundary value problems

$$y'' + \lambda y = 0, \qquad y(0) = 0,\, y(l) = 0, \tag{3}$$

and

$$y'' + \lambda y = 0, \qquad y'(0) = 0,\, y'(l) = 0. \tag{4}$$

Recall that Equations (3) and (4) have nontrivial solutions

$$y_n(x) = c\sin\frac{n\pi x}{l} \quad \text{and} \quad y_n(x) = c\cos\frac{n\pi x}{l},$$

respectively, only if $\lambda = \lambda_n = \dfrac{n^2\pi^2}{l^2}$. These special values of λ were called eigenvalues, and the corresponding solutions were called eigenfunctions.

There was another instance in our study of differential equations where something interesting happened for special values of a parameter λ. To wit, in Section 2.8 we studied, either in the text or in the exercises, the four differential equations

$$y'' - 2xy + \lambda y = 0, \tag{5}$$

$$(1 - x^2)y'' - 2xy' + \lambda(\lambda + 1)y = 0, \tag{6}$$

$$(1 - x^2)y'' - xy' + \lambda^2 y = 0, \tag{7}$$

and

$$xy'' + (1 - x)y' + \lambda y = 0. \tag{8}$$

Equations (5)–(8) are the famous Hermite, Legendre, Tchebycheff, and Laguerre differential equations, respectively. The Legendre, Tchebycheff, and Laguerre equations each have a polynomial solution of degree n if $\lambda = n$, while the Hermite equation has a polynomial solution of degree n if $\lambda = 2n$. These polynomials, when properly normalized, i.e., when multiplied by a suitable constant, are known as the Hermite, Legendre, Tchebycheff, and Laguerre polynomials. It turns out, remarkably, that any piecewise differentiable function $f(x)$ can also be expanded in a series of Hermite, Legendre, Tchebycheff, and Laguerre polynomials, on an appropriate interval.

There is a very pretty theory that ties together not only the trigonometric functions and the Hermite, Legendre, Tchebycheff, and Laguerre polynomials, but also many of the other famous functions of mathematical physics, such as the various Bessel functions. This theory is commonly called Sturm-Liouville Theory; it has its roots, essentially, in an area of linear algebra known as inner product spaces, and it is to this area that we now turn our attention.

6.2 Inner product spaces

Up to this point, our study of linear algebra in general and linear vector spaces in particular was algebraic in nature. We were able to add two vectors together and multiply a vector by a constant. By means of these operations we can define the geometric concepts of dimension, line, plane, and even parallelism of lines. Recall that the dimension of a space V is the number of elements in a basis, i.e., the fewest number of linearly independent vectors that span V. Once we have the concept of dimension, we can define a line in V as a subspace of dimension 1, a plane as a subspace of dimension 2, etc. Finally, two vectors are parallel if one is a constant multiple of the other.

Many important geometric concepts of so-called Euclidean geometry, however, still cannot be defined for arbitrary linear vector spaces. Specifically, we have no way of formulating, as yet, the definition of *length* of a vector and the *angle* between two vectors. To accomplish this, we need to superimpose on V some additional structure. This structure is known as an *inner product.*

Our definition of an inner product is modelled after the traditional *dot product* of two vectors in R^2 and R^3, which we studied in calculus. Recall that if $\mathbf{x}$ and $\mathbf{y}$ are vectors in R^2 or R^3, then

$$\mathbf{x}\cdot\mathbf{y} = x_1y_1 + x_2y_2 \quad \text{in } R^2 \tag{1}$$

and

$$\mathbf{x}\cdot\mathbf{y} = x_1y_1 + x_2y_2 + x_3y_3 \quad \text{in } R^3 \tag{2}$$

where $x_1, x_2, \ldots$, and $y_1, y_2, \ldots$ are the components of $\mathbf{x}$ and $\mathbf{y}$, respectively. Recall too, the famous identity, proven in most calculus courses, that

$$\mathbf{x}\cdot\mathbf{y} = |\mathbf{x}|\,|\mathbf{y}|\cos\theta \tag{3}$$

where θ is the angle between $\mathbf{x}$ and $\mathbf{y}$ and

$$|\mathbf{x}| = (x_1^2 + x_2^2)^{1/2} \quad \text{or} \quad (x_1^2 + x_2^2 + x_3^2)^{1/2}$$

in R^2 and R^3, respectively, is the Euclidean length of $\mathbf{x}$. Rewriting Equation (3) in the form

$$\cos\theta = \frac{\mathbf{x}\cdot\mathbf{y}}{|\mathbf{x}|\,|\mathbf{y}|} \tag{4}$$

enables us to compute the angle θ between two vectors. Finally, if we observe that

$$|\mathbf{x}| = (\mathbf{x}\cdot\mathbf{x})^{1/2} \tag{5}$$

then both the length of a vector and the angle between two vectors can be computed from the dot product alone.

We are now ready to generalize this concept of dot product. This generalization will be called an *inner product.* We will first define a *real* inner product, and then toward the end of this section, a *complex* inner product.

Definition. Let V be a real vector space. A *real inner product* on V is a real-valued function that associates with each pair of vectors $\mathbf{x}$, $\mathbf{y}$ a real number, denoted by $\langle\mathbf{x},\mathbf{y}\rangle$, that satisfies the following properties:

(i) $\langle\mathbf{x},\mathbf{y}\rangle = \langle\mathbf{y},\mathbf{x}\rangle$ for all $\mathbf{x}$ and $\mathbf{y}$ in V,
(ii) $\langle k\mathbf{x},\mathbf{y}\rangle = k\langle\mathbf{x},\mathbf{y}\rangle$ for all scalars k and vectors $\mathbf{x}$, $\mathbf{y}$,
(iii) $\langle\mathbf{x}+\mathbf{y},\mathbf{z}\rangle = \langle\mathbf{x},\mathbf{z}\rangle + \langle\mathbf{y},\mathbf{z}\rangle$ for all $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ in V,
(iv) $\langle\mathbf{x},\mathbf{x}\rangle \geqslant 0$ and $\langle\mathbf{x},\mathbf{x}\rangle = 0$ only if $\mathbf{x} = \mathbf{0}$.

It is customary to refer to the vector space V, together with some inner product $\langle\ ,\ \rangle$ as a real inner product space, and a finite-dimensional real inner product space is often called a Euclidean space.

Remark. Observe from (i) that properties (ii) and (iii) hold equally well on opposite sides of the inner product bracket, i.e.,

$$\langle \mathbf{x}, k\mathbf{y}\rangle = k\langle \mathbf{x},\mathbf{y}\rangle \quad \text{and} \quad \langle \mathbf{x},\mathbf{y}+\mathbf{z}\rangle = \langle \mathbf{x},\mathbf{y}\rangle + \langle \mathbf{x},\mathbf{z}\rangle.$$

The classic example of an inner product on R^n, of course, is the dot product

$$\langle \mathbf{x},\mathbf{y}\rangle = \mathbf{x}\cdot\mathbf{y} = x_1y_1 + x_2y_2 + \cdots + x_ny_n.$$

It is easily verified that the dot product satisfies properties (i)–(iv). Here are some additional examples, which have proven extremely useful in applications.

Example 1. Let V be the space of all continuous functions on the interval $[a,b]$, and define the inner product of two functions $f(x)$ and $g(x)$ in V as

$$\langle f,g\rangle = \int_a^b f(x)g(x)\,dx. \tag{6}$$

It is easily verified that the definition (6) satisfies properties (i)–(iv).

Example 2. Let V be as in Example 1, and define

$$\langle f,g\rangle = \int_a^b r(x)f(x)g(x)\,dx \tag{7}$$

where $r(x)$ is positive on the open interval (a,b). Again, it is trivial to verify that the inner product (7) satisfies properties (i)–(iv).

Example 3. Let $V = R^3$ and set

$$\langle \mathbf{x},\mathbf{y}\rangle = x_1y_1 + 2x_2y_2 + 2x_3y_3 - x_1y_2 - x_2y_1 - x_2y_3 - x_3y_2.$$

Properties (ii) and (iii) are trivial to check, and property (i) is simple to verify. To check property (iv), we write

$$\begin{aligned}\langle \mathbf{x},\mathbf{x}\rangle &= x_1^2 + 2x_2^2 + 2x_3^2 - x_1x_2 - x_2x_1 - x_2x_3 - x_3x_2\\ &= x_1^2 + 2x_2^2 + 2x_3^2 - 2x_1x_2 - 2x_2x_3\\ &= x_1^2 - 2x_1x_2 + x_2^2 + x_2^2 - 2x_2x_3 + x_3^2 + x_3^2\\ &= (x_1-x_2)^2 + (x_2-x_3)^2 + x_3^2,\end{aligned}$$

and it is now clear that $\langle \mathbf{x},\mathbf{x}\rangle = 0$ if and only if $\mathbf{x} = \mathbf{0}$.

Let $\mathbf{x}$ and $\mathbf{y}$ be vectors in an arbitrary inner product space V, and suppose, following Equation (5), that we define the lengths of $\mathbf{x}$ and $\mathbf{y}$, denoted by $\|\mathbf{x}\|$ and $\|\mathbf{y}\|$, respectively, as

$$\|\mathbf{x}\| = \langle \mathbf{x},\mathbf{x}\rangle^{1/2}, \qquad \|\mathbf{y}\| = \langle \mathbf{y},\mathbf{y}\rangle^{1/2}, \tag{8}$$

where the double vertical lines are used to denote length. It is certainly very plausible to try to mimic Equation (4) and define the angle θ between $\mathbf{x}$ and y via the equation

$$\cos\theta = \frac{\langle \mathbf{x}, \mathbf{y}\rangle}{\|\mathbf{x}\|\,\|\mathbf{y}\|} = \frac{\langle \mathbf{x}, \mathbf{y}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle^{1/2}\langle \mathbf{y}, \mathbf{y}\rangle^{1/2}}. \tag{9}$$

Clearly, the right-hand side of Equation (9) must be $\leqslant 1$ in order for this equation to have any meaning at all. A second and more subtle requirement is that the quantity $\|\mathbf{x}\|$, defined as the *length* of $\mathbf{x}$, really is a length; that is, it satisfies the geometric properties usually associated with length. Let us therefore take one last digression and discuss the concept of length of a vector, or as it is traditionally called in linear algebra, the *norm* of a vector.

Definition. A *norm* on a real or complex vector space V is a real-valued function, usually denoted by $\|\ \ \|$, which satisfies

(i) $\|\mathbf{x}\| \geqslant 0$ for all $\mathbf{x}$ in V and $\|\mathbf{x}\| = 0$ only if $\mathbf{x} = \mathbf{0}$,
(ii) $\|k\mathbf{x}\| = |k|\,\|\mathbf{x}\|$ for all scalars k,
(iii) $\|\mathbf{x} + \mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|$.

Properties (i) and (ii) are fairly obvious properties of length, while property (iii) is simply the triangle law, which states that the length of one side of a triangle is always less than (or equal to) the sum of the lengths of the other two sides (see Figure 1).

The classical example of a norm is ordinary length in R^n; i.e., if

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

then

$$\|\mathbf{x}\| = (x_1^2 + x_2^2 + \cdots + x_n^2)^{1/2}$$

obviously satisfies (i)–(iii). Here are some additional examples, which have proven quite useful in applications.

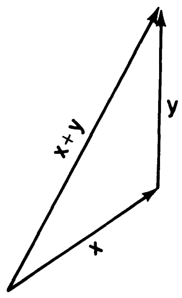


Figure 1

Example 4. Let $V = R^n$ and define

$$\|\mathbf{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|.$$

It is extremely easy to verify that $\| \quad \|_1$ is a norm on R^n.

Example 5. Let $V = R^n$ and define

$$\|\mathbf{x}\|_\infty = \max[|x_1|, |x_2|, \ldots, |x_n|].$$

Again, it is extremely simple to verify that $\| \quad \|_\infty$ defines a norm on R^n.

Example 6. Let $V = R^n$ and define

$$\|\mathbf{x}\|_p = [|x_1|^p + |x_2|^p + \cdots + |x_n|^p]^{1/p}. \tag{10}$$

It is extremely simple to verify that $\| \quad \|_p$, referred to as the p norm, satisfies properties (i) and (ii). Property (iii) is also true but quite difficult to verify, hence we will not do so here. We wish to point out (see Exercise 6), however, that

$$\|\mathbf{x}\|_\infty = \lim_{p\to\infty} \|\mathbf{x}\|_p,$$

and this is the motivation for the strange notation $\| \quad \|_\infty$.

Example 7. Let V be the space of all continuous functions $f(x)$ on the interval $[0,1]$, and define

$$\|f\| = \int_0^1 |f(x)|\,dx. \tag{11}$$

It is easily verified that Equation (11) defines a norm on V.

Let us return now to the questions raised via the definitions (8) and (9). As mentioned previously, we must show that the right-hand side of Equation (9) is always between -1 and $+1$ in order for this equation to make sense. But this is equivalent to the inequality

$$|\langle \mathbf{x}, \mathbf{y}\rangle| \leqslant \langle \mathbf{x}, \mathbf{x}\rangle^{1/2}\langle \mathbf{y}, \mathbf{y}\rangle^{1/2}$$

or, equivalently,

$$|\langle \mathbf{x}, \mathbf{y}\rangle| \leqslant \|\mathbf{x}\|\,\|\mathbf{y}\|. \tag{12}$$

Equation (12) is the precise statement of a very famous theorem, known as Schwarz's Inequality, which we now prove.

Theorem 1 (Schwarz's Inequality). *Let V be a real inner product space with inner product $\langle\ ,\ \rangle$. Then*

$$|\langle \mathbf{x}, \mathbf{y}\rangle| \leqslant \|\mathbf{x}\|\,\|\mathbf{y}\| \tag{13}$$

for all vectors $\mathbf{x}$, $\mathbf{y}$, *where* $\|\mathbf{x}\|$ *is defined by Equation* (8). *Furthermore, equality holds in Equation* (13) *only if the vectors* $\mathbf{x}$ *and* $\mathbf{y}$ *are linearly dependent, i.e.,* $\mathbf{y} = k\mathbf{x}$ *for some scalar* k.

PROOF. There are several ways of proving Theorem 1, but we will choose here the proof that generalizes most easily to the complex case. Observe that

$$\left\langle \mathbf{x} - \frac{\langle \mathbf{x},\mathbf{y}\rangle}{\langle \mathbf{y},\mathbf{y}\rangle}\mathbf{y},\, \mathbf{x} - \frac{\langle \mathbf{x},\mathbf{y}\rangle}{\langle \mathbf{y},\mathbf{y}\rangle}\mathbf{y} \right\rangle \geqslant 0 \tag{14}$$

for all nonzero vectors $\mathbf{x}$ and $\mathbf{y}$ in V. This follows immediately from property (iv) of inner products. Next, using properties (i)–(iii) of inner products, we can rewrite Equation (14) in the form

$$\begin{aligned}
\langle \mathbf{x},\mathbf{x}\rangle - \frac{\langle \mathbf{x},\mathbf{y}\rangle}{\langle \mathbf{y},\mathbf{y}\rangle}\langle \mathbf{y},\mathbf{x}\rangle - \frac{\langle \mathbf{x},\mathbf{y}\rangle}{\langle \mathbf{y},\mathbf{y}\rangle}\langle \mathbf{x},\mathbf{y}\rangle + \frac{\langle \mathbf{x},\mathbf{y}\rangle^2}{\langle \mathbf{y},\mathbf{y}\rangle^2}\langle \mathbf{y},\mathbf{y}\rangle \\
&= \langle \mathbf{x},\mathbf{x}\rangle - 2\frac{\langle \mathbf{x},\mathbf{y}\rangle^2}{\langle \mathbf{y},\mathbf{y}\rangle} + \frac{\langle \mathbf{x},\mathbf{y}\rangle^2}{\langle \mathbf{y},\mathbf{y}\rangle} \\
&= \langle \mathbf{x},\mathbf{x}\rangle - \frac{\langle \mathbf{x},\mathbf{y}\rangle^2}{\langle \mathbf{y},\mathbf{y}\rangle} \\
&= \|\mathbf{x}\|^2 - \frac{\langle \mathbf{x},\mathbf{y}\rangle^2}{\|\mathbf{y}\|^2}.
\end{aligned}$$

Hence,

$$\frac{\langle \mathbf{x},\mathbf{y}\rangle^2}{\|\mathbf{y}\|^2} \leqslant \|\mathbf{x}\|^2$$

or, equivalently,

$$|\langle \mathbf{x},\mathbf{y}\rangle| \leqslant \|\mathbf{x}\|\,\|\mathbf{y}\|.$$

Finally, again from property (iv), equality holds in Equation (14) only if

$$\mathbf{x} - \frac{\langle \mathbf{x},\mathbf{y}\rangle}{\langle \mathbf{y},\mathbf{y}\rangle}\mathbf{y} = \mathbf{0}$$

or

$$\mathbf{x} = k\mathbf{y}, \qquad k = \frac{\langle \mathbf{x},\mathbf{y}\rangle}{\langle \mathbf{y},\mathbf{y}\rangle}.$$

Remark 1. The proof we just gave assumes that $\mathbf{y} \neq \mathbf{0}$. We leave it to the reader to verify (see Exercise 13) Schwarz's Inequality in the trivial case that $\mathbf{y} = \mathbf{0}$.

Remark 2. The vector $(\langle \mathbf{x},\mathbf{y}\rangle/\langle \mathbf{y},\mathbf{y}\rangle)\mathbf{y}$ introduced in Equation (14) was not a luckly guess. This vector is actually the *projection* of the vector $\mathbf{x}$ onto the vector $\mathbf{y}$, as shown in Figure 2.

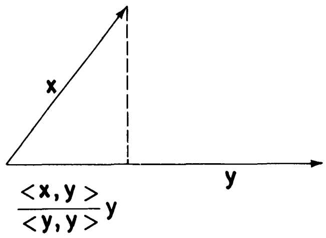


Figure 2

The second question raised was whether

$$\|\mathbf{x}\| = \langle \mathbf{x}, \mathbf{x} \rangle^{1/2} \tag{15}$$

satisfies properties (i)–(iii) of norms. Property (i) follows immediately from property (iv) of inner products. To verify property (ii), we compute

$$\|k\mathbf{x}\| = \langle k\mathbf{x}, k\mathbf{x} \rangle^{1/2} = [k^2\langle \mathbf{x}, \mathbf{x} \rangle]^{1/2} = |k|\langle \mathbf{x}, \mathbf{x} \rangle^{1/2} = |k|\,\|\mathbf{x}\|.$$

Property (iii) is a bit more difficult to verify. Observe first that

$$\begin{aligned}
\|\mathbf{x} + \mathbf{y}\|^2 &= \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle \\
&= \langle \mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle \\
&= \|\mathbf{x}\|^2 + 2\langle \mathbf{x}, \mathbf{y} \rangle + \|\mathbf{y}\|^2 \\
&\leqslant \|\mathbf{x}\|^2 + 2\|\mathbf{x}\|\,\|\mathbf{y}\| + \|\mathbf{y}\|^2 \\
&= [\|\mathbf{x}\| + \|\mathbf{y}\|]^2,
\end{aligned}$$

where the inequality above follows directly from Schwarz's Inequality. Taking square roots gives the triangle inequality, and this completes our proof.

Remark. Schwarz's Inequality is an extremely powerful tool, as witnessed by the preceding proof. Here is another illustration of its strength. Let V be the space of all continuous functions on the interval $[a, b]$, and define

$$\langle f, g \rangle = \int_a^b f(x)g(x)\,dx.$$

Then, from Schwarz's Inequality,

$$|\langle f, g \rangle| = \left| \int_a^b f(x)g(x)\,dx \right| \leqslant \|f(x)\|\,\|g(x)\|.$$

Hence,

$$\left| \int_a^b f(x)g(x)\,dx \right| \leqslant \left[\int_a^b f^2(x)\,dx \right]^{1/2} \left[\int_a^b g^2(x)\,dx \right]^{1/2}. \tag{16}$$

We defy the reader to give a pure calculus proof of the inequality (16).

Once we have an inner product space V, we can define two vectors in V to be *orthogonal*, or perpendicular to each other, if their inner product is zero; i.e., $\mathbf{x}$ and $\mathbf{y}$ are orthogonal if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$.

Example 8. Let V be the space of all continuous functions on the interval $[-\pi, \pi]$, and define an inner product on V via the relation

$$\langle f, g \rangle = \int_{-\pi}^{\pi} f(x)g(x)\,dx.$$

The set of functions

$$f_k(x) = \sin kx, \qquad k = 1, 2, \ldots,$$

are all mutually orthogonal; that is,

$$\langle f_j(x), f_k(x) \rangle = 0, \qquad j \neq k,$$

since

$$\langle f_j(x), f_k(x) \rangle = \int_{-\pi}^{\pi} \sin jx \sin kx\, dx = 0, \qquad j \neq k,$$

following the discussion in Section 5.4.

Example 9. Let V be the space of all continuous functions on the interval $[-1, 1]$, and define

$$\langle f, g \rangle = \int_{-1}^{1} \frac{f(x)g(x)}{\sqrt{1-x^2}}\,dx$$

for f and g in V. The set of functions

$$f_k(x) = \cos(k \cos^{-1} x), \qquad k = 0, 1, 2, \ldots,$$

are mutually orthogonal with respect to $\langle\ ,\ \rangle$, i.e.,

$$\int_{-1}^{1} \frac{\cos(j\cos^{-1}x)\cos(k\cos^{-1}x)}{\sqrt{1-x^2}}\,dx = 0, \qquad j \neq k. \tag{17}$$

To establish Equation (17) we make the substitution

$$u = \cos^{-1}x \qquad du = \frac{-1}{\sqrt{1-x^2}}\,dx.$$

Then the left-hand side of Equation (17) becomes

$$\begin{aligned} -\int_{\pi}^{0} \cos ju \cos ku\, du &= \int_{0}^{\pi} \cos ju \cos ku\, du \\ &= \frac{1}{2}\int_{0}^{\pi} [\cos(j+k)u + \cos(j-k)u]\, du = 0, \qquad j \neq k. \end{aligned}$$

Remark. It can be shown that the functions $f_k(x)$ defined in Example 9 are the Tchebycheff polynomials $T_k(x)$.

We conclude this section by extending the notion and definition of inner product to complex vector spaces. The main problem we have to overcome is the requirement that $\langle \mathbf{x},\mathbf{x}\rangle$ be positive for all nonzero vectors $\mathbf{x}$. For example, if V is the space of all complex-valued functions

$$h(t) = f(t) + ig(t), \qquad a \leqslant t \leqslant b,$$

then we cannot define

$$\langle h_1, h_2\rangle = \int_a^b h_1(t)h_2(t)\,dt$$

as in Example 1, since

$$\langle h, h\rangle = \int_a^b [f(t) + ig(t)][f(t) + ig(t)]\,dt$$

is not even real [for $\langle f, g\rangle \neq 0$], let alone positive.

To motivate the proper extension, we return to the familiar dot product

$$\langle \mathbf{x},\mathbf{y}\rangle = \mathbf{x}\cdot\mathbf{y} = x_1y_1 + x_2y_2 + \cdots + x_ny_n \tag{18}$$

where $\mathbf{x}$ and $\mathbf{y}$ are *real* vectors, i.e., their components are real numbers. We can extend this definition to complex vector spaces by making the simple adjustment

$$\mathbf{x}\cdot\mathbf{y} = x_1\bar{y}_1 + x_2\bar{y}_2 + \cdots + x_n\bar{y}_n. \tag{19}$$

Then, following Equation (19),

$$\langle \mathbf{x},\mathbf{x}\rangle = x_1\bar{x}_1 + x_2\bar{x}_2 + \cdots + x_n\bar{x}_n$$

is always real and nonnegative. The only thing we have to be extra careful about now is that

$$\mathbf{y}\cdot\mathbf{x} = y_1\bar{x}_1 + \cdots + y_n\bar{x}_n = \overline{\langle \mathbf{x},\mathbf{y}\rangle}$$

instead of $\langle \mathbf{x},\mathbf{y}\rangle$ and

$$\mathbf{x}\cdot k\mathbf{y} = x_1\bar{k}\bar{y}_1 + \cdots + x_n\bar{k}\bar{y}_n = \bar{k}\langle \mathbf{x},\mathbf{y}\rangle$$

instead of $k\,\langle \mathbf{x},\mathbf{y}\rangle$. This leads us to the following definition.

Definition. Let V be a complex vector space. A *complex*, or Hermitian, *inner product* on V is a complex-valued function that associates with each pair of vectors $\mathbf{x}$, $\mathbf{y}$ a complex number, denoted by $\langle \mathbf{x},\mathbf{y}\rangle$, that satisfies the following properties:

(i) $\langle \mathbf{x},\mathbf{y}\rangle = \overline{\langle \mathbf{y},\mathbf{x}\rangle}$,
(ii) $\langle k\mathbf{x},\mathbf{y}\rangle = k\langle \mathbf{x},\mathbf{y}\rangle$,

(iii) $\langle \mathbf{x}+\mathbf{y},\mathbf{z}\rangle = \langle \mathbf{x},\mathbf{z}\rangle + \langle \mathbf{y},\mathbf{z}\rangle$,
(iv) $\langle \mathbf{x},\mathbf{x}\rangle \geqslant 0$ and $\langle \mathbf{x},\mathbf{x}\rangle = 0$ only if $\mathbf{x} = \mathbf{0}$

for all vectors $\mathbf{x}$ and $\mathbf{y}$ and complex numbers k.

Property (i) is known as *conjugate symmetry* in contrast to the ordinary symmetry of the real case. It implies that

$$\langle \mathbf{x},k\mathbf{y}\rangle = \overline{\langle k\mathbf{y},\mathbf{x}\rangle} = \overline{k\langle \mathbf{y},\mathbf{x}\rangle} = \bar{k}\overline{\langle \mathbf{y},\mathbf{x}\rangle} = \bar{k}\langle \mathbf{x},\mathbf{y}\rangle.$$

Thus, scalars pulled out from the right-hand side of a complex inner product must be conjugated. Finally, we leave it to the reader to verify (see Exercise 14) that

$$\langle \mathbf{x},\mathbf{y}+\mathbf{z}\rangle = \langle \mathbf{x},\mathbf{y}\rangle + \langle \mathbf{x},\mathbf{z}\rangle.$$

Example 10. Let V be the space of all complex-valued functions $h(t) = f(t) + ig(t)$ on the interval $[0,1]$, and define

$$\begin{aligned}\langle h_1,h_2\rangle &= \int_0^1 h_1(t)\overline{h_2(t)}\,dt\\ &= \int_0^1 [f_1(t)+ig_1(t)][f_2(t)-ig_2(t)]\,dt \qquad (20)\\ &= \int_0^1 [f_1(t)f_2(t)+g_1(t)g_2(t)]\,dt + i\int_0^1 [f_2(t)g_1(t)-f_1(t)g_2(t)]\,dt\end{aligned}$$

It is easily verified that Equation (20) satisfies properties (i)–(iv) of our definition.

Complex inner product spaces are the same as real inner product spaces, essentially, except for the conjugate symmetry property, as noted above. For example, we define

$$\|\mathbf{x}\| = \langle \mathbf{x},\mathbf{x}\rangle^{1/2} \qquad (21)$$

to be the length of a vector $\mathbf{x}$ in a complex vector space, just as in a real vector space. The only thing that we must do, though, is extend the proof of Schwarz's Inequality to the complex case, since our proof made use of the symmetry property of real inner products.

Theorem 2 (Schwarz's Inequality). *Let V be a complex inner product space with inner product $\langle\ ,\ \rangle$. Then*

$$|\langle \mathbf{x},\mathbf{y}\rangle| \leqslant \|\mathbf{x}\|\,\|\mathbf{y}\| \qquad (22)$$

for all $\mathbf{x}$, $\mathbf{y}$ *in* V, *where* $\|\mathbf{x}\|$ *is defined by Equation* (21). *Furthermore, equality holds in Equation* (22) *if and only if the vectors* $\mathbf{x}$ *and* $\mathbf{y}$ *are linearly dependent.*

PROOF. Observe from property (iv) that

$$\left\langle \mathbf{x} - \frac{\langle \mathbf{x},\mathbf{y}\rangle}{\langle \mathbf{y},\mathbf{y}\rangle}\mathbf{y}, \mathbf{x} - \frac{\langle \mathbf{x},\mathbf{y}\rangle}{\langle \mathbf{y},\mathbf{y}\rangle}\mathbf{y} \right\rangle \geqslant 0 \tag{23}$$

for all vectors $\mathbf{x}$ and nonzero vectors $\mathbf{y}$ in V. Using properties (i)–(iii), we can rewrite the left-hand side of Equation (23) in the form

$$\begin{aligned}
&\langle \mathbf{x},\mathbf{x}\rangle - \frac{\langle \mathbf{x},\mathbf{y}\rangle}{\langle \mathbf{y},\mathbf{y}\rangle}\langle \mathbf{y},\mathbf{x}\rangle - \frac{\overline{\langle \mathbf{x},\mathbf{y}\rangle}}{\langle \mathbf{y},\mathbf{y}\rangle}\langle \mathbf{x},\mathbf{y}\rangle + \frac{\langle \mathbf{x},\mathbf{y}\rangle}{\langle \mathbf{y},\mathbf{y}\rangle}\frac{\overline{\langle \mathbf{x},\mathbf{y}\rangle}}{\langle \mathbf{y},\mathbf{y}\rangle}\langle \mathbf{y},\mathbf{y}\rangle \\
&\qquad = \langle \mathbf{x},\mathbf{x}\rangle - \frac{\langle \mathbf{x},\mathbf{y}\rangle}{\langle \mathbf{y},\mathbf{y}\rangle}\overline{\langle \mathbf{x},\mathbf{y}\rangle} - \frac{\overline{\langle \mathbf{x},\mathbf{y}\rangle}}{\langle \mathbf{y},\mathbf{y}\rangle}\langle \mathbf{x},\mathbf{y}\rangle + \frac{\langle \mathbf{x},\mathbf{y}\rangle}{\langle \mathbf{y},\mathbf{y}\rangle}\overline{\langle \mathbf{x},\mathbf{y}\rangle} \\
&\qquad = \|\mathbf{x}\|^2 - \frac{|\langle \mathbf{x},\mathbf{y}\rangle|^2}{\|\mathbf{y}\|^2} \\
&\qquad = \frac{\|\mathbf{x}\|^2\|\mathbf{y}\|^2 - |\langle \mathbf{x},\mathbf{y}\rangle|^2}{\|\mathbf{y}\|^2}.
\end{aligned}$$

Hence,

$$|\langle \mathbf{x},\mathbf{y}\rangle|^2 \leqslant \|\mathbf{x}\|^2\|\mathbf{y}\|^2, \tag{24}$$

and Schwarz's Inequality follows immediately by taking square roots in Equation (24).

Remark. As in the real case, our proof assumes that $\mathbf{y} \neq \mathbf{0}$. We leave it to the reader (see Exercise 13) to complete the proof for $\mathbf{y} = \mathbf{0}$ and to verify that equality holds if and only if $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent.

EXERCISES

1. Show that Equation (6) defines an inner product on $C[a,b]$, the space of all continuous functions on the interval $[a,b]$.
2. Show that Equation (7) defines an inner product on $C[a,b]$.
3. Show that $\|\mathbf{x}\|_1$ defined in Example 4 is a norm.
4. Show that $\|\mathbf{x}\|_\infty$ defined in Example 5 is a norm.
5. Show that $\|\mathbf{x}\|_p$ defined in Example 6 is a norm.
6. Show that $\|\mathbf{x}\|_\infty = \lim_{p\to\infty} \|\mathbf{x}\|_p$.
7. Show that $\|\mathbf{x}\| = |x_1| + 2|x_2| + 4|x_3|$ is a norm on R^3.
8. Show that $|x_1| + |x_2| + |x_3|^2$ does not define a norm on R^3.
9. Show that $e^{|x_1|+|x_2|}$ does not define a norm on R^2.
10. Show that Equation (11) defines a norm on $C[a,b]$.

11. Let V be a real inner product space. Prove that

$$\langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{4}\|\mathbf{x}+\mathbf{y}\|^2 - \frac{1}{4}\|\mathbf{x}-\mathbf{y}\|^2.$$

12. Let $\mathbf{x} \times \mathbf{y}$ be the vector with components

$$x_2y_3 - x_3y_2, x_3y_1 - x_1y_3, x_1y_2 - x_2y_1,$$

respectively. Show that $\mathbf{x} \times \mathbf{y}$ is perpendicular to both $\mathbf{x}$ and $\mathbf{y}$ under the inner product $(\mathbf{x}, \mathbf{y})$.

13. Complete the proof of Schwarz's Inequality in both R^n and C^n for the case $\mathbf{y} = \mathbf{0}$.

14. Show that

$$\langle \mathbf{x}, \mathbf{y}+\mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{x}, \mathbf{z} \rangle$$

for all vectors $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ in C^n.

6.3 Orthogonal bases, Hermitian operators

It is often the case that a crucial step in the solution of many problems in a linear vector space revolves around a judicious choice of basis. A type of basis that is especially useful is one that is *orthogonal.*

Definition. A set of vectors is said to be *orthogonal* if the inner product of any two distinct vectors in the set is zero.

One of the nice things about an orthogonal set of vectors is that they are automatically linearly independent. This is the content of Lemma 1.

Lemma 1. *Let* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ *be mutually orthogonal, that is,*

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle = 0, \qquad i \neq j.$$

Then, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ *are linearly independent.*

PROOF. Suppose that

$$c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_N\mathbf{x}_N = \mathbf{0}. \tag{1}$$

Taking inner products of both sides of Equation (1) with $\mathbf{x}_j$ gives

$$c_1\langle \mathbf{x}_1, \mathbf{x}_j \rangle + c_2\langle \mathbf{x}_2, \mathbf{x}_j \rangle + \cdots + c_N\langle \mathbf{x}_N, \mathbf{x}_j \rangle = 0. \tag{2}$$

The left-hand side of Equation (2) reduces to $c_j\langle \mathbf{x}_j, \mathbf{x}_j \rangle$. Since $\langle \mathbf{x}_j, \mathbf{x}_j \rangle > 0$, we see that $c_j = 0$, $j = 1, 2, \ldots, N$, which proves Lemma 1.

Another benefit of working with orthogonal bases is that it is relatively easy to find the coordinates of a vector with respect to a given orthogonal basis. To wit, let $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ be a mutually orthogonal set of vectors in a real n-dimensional vector space V. By Lemma 1, this set of vectors is also a basis for V, and every vector $\mathbf{x}$ in V can be expanded in the form

$$\mathbf{x} = c_1\mathbf{u}_1 + c_2\mathbf{u}_2 + \cdots + c_n\mathbf{u}_n. \tag{3}$$

Taking inner products of both sides of Equation (3) with $\mathbf{u}_j$ gives $\langle \mathbf{x}, \mathbf{u}_j \rangle = c_j\langle \mathbf{u}_j, \mathbf{u}_j \rangle$ so that

$$c_j = \frac{\langle \mathbf{x}, \mathbf{u}_j \rangle}{\langle \mathbf{u}_j, \mathbf{u}_j \rangle}, \qquad j = 1, 2, \ldots, n. \tag{4}$$

Example 1. Let $V = R^2$ and define $\langle \mathbf{x}, \mathbf{y} \rangle = x_1y_1 + x_2y_2$. The vectors

$$\mathbf{u}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \mathbf{u}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

are orthogonal and, thus, form a basis for R^2. Consequently, from Equations (3) and (4), any vector $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ can also be written in the form

$$\begin{aligned}\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &= c_1\begin{pmatrix} 1 \\ 1 \end{pmatrix} + c_2\begin{pmatrix} 1 \\ -1 \end{pmatrix} \\ &= \frac{\langle \mathbf{x}, \mathbf{u}_1 \rangle}{\langle \mathbf{u}_1, \mathbf{u}_1 \rangle}\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{\langle \mathbf{x}, \mathbf{u}_2 \rangle}{\langle \mathbf{u}_2, \mathbf{u}_2 \rangle}\begin{pmatrix} 1 \\ -1 \end{pmatrix} \\ &= \frac{x_1 + x_2}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{x_1 - x_2}{2}\begin{pmatrix} 1 \\ -1 \end{pmatrix}.\end{aligned}$$

Given the benefit of working with orthogonal bases, the following question naturally arises: Does every n-dimensional Euclidean space have an orthogonal basis? The answer to this question is a resounding yes and is given by the following theorem.

Theorem 3 (Gram-Schmidt). *Every n-dimensional Euclidean space V has an orthogonal basis.*

PROOF. Choose a basis $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ for V. We will inductively construct an orthogonal basis $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ by taking suitable combinations of the vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$. To this end, let $\mathbf{v}_1 = \mathbf{u}_1$ and set

$$\mathbf{v}_2 = \mathbf{u}_2 + \lambda\mathbf{v}_1. \tag{5}$$

Taking the inner product of $\mathbf{v}_2$ with $\mathbf{v}_1$ gives

$$\langle \mathbf{v}_2, \mathbf{v}_1 \rangle = \langle \mathbf{u}_2, \mathbf{v}_1 \rangle + \lambda\langle \mathbf{v}_1, \mathbf{v}_1 \rangle$$

so that $\mathbf{v}_2$ will be orthogonal to $\mathbf{v}_1$ if $\lambda = -\langle \mathbf{u}_2, \mathbf{v}_1 \rangle / \langle \mathbf{v}_1, \mathbf{v}_1 \rangle$. Note that $\mathbf{v}_2$ cannot equal $\mathbf{0}$ by virtue of the linear independence of the vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$. Proceeding inductively, let us assume that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ are mutually orthogonal, and set

$$\mathbf{v}_{k+1} = \mathbf{u}_{k+1} + \lambda_1 \mathbf{v}_1 + \cdots + \lambda_k \mathbf{v}_k. \tag{6}$$

The requirement that $\mathbf{v}_{k+1}$ be orthogonal to $\mathbf{v}_1, \ldots, \mathbf{v}_k$ yields

$$0 = \langle \mathbf{v}_{k+1}, \mathbf{v}_j \rangle = \langle \mathbf{u}_{k+1}, \mathbf{v}_j \rangle + \lambda_j \langle \mathbf{v}_j, \mathbf{v}_j \rangle,$$

which gives the desired relation

$$\lambda_j = -\frac{\langle \mathbf{u}_{k+1}, \mathbf{v}_j \rangle}{\langle \mathbf{v}_j, \mathbf{v}_j \rangle}, \qquad j = 1, \ldots, k.$$

With this choice of $\lambda_1, \ldots, \lambda_k$, the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$, $\mathbf{v}_{k+1}$ are mutually orthogonal. Moreover, $\mathbf{v}_{k+1}$ cannot be $\mathbf{0}$ by virtue of the linear independence of $\mathbf{u}_1, \ldots, \mathbf{u}_{k+1}$. Proceeding inductively until $k = n$, we obtain n mutually orthogonal nonzero vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$.

Remark 1. The procedure outlined in Theorem 3 is usually referred to as the Gram-Schmidt orthogonalization procedure.

Remark 2. If $\mathbf{u}_1, \ldots, \mathbf{u}_n$ is an orthogonal set of vectors, and, in addition, each of the vectors $\mathbf{u}_j$ has length 1, i.e.,

$$\|\mathbf{u}_j\| = \langle \mathbf{u}_j, \mathbf{u}_j \rangle^{1/2} = 1, \qquad j = 1, \ldots, n,$$

then $\mathbf{u}_1, \ldots, \mathbf{u}_n$ is called an *orthonormal* set of vectors. It is a simple matter to construct an orthonormal basis from an orthogonal one: if $\mathbf{u}_1, \ldots, \mathbf{u}_n$ is an orthogonal basis, then

$$\mathbf{v}_1 = \frac{\mathbf{u}_1}{\|\mathbf{u}_1\|}, \qquad \mathbf{v}_2 = \frac{\mathbf{u}_2}{\|\mathbf{u}_2\|}, \qquad \ldots, \qquad \mathbf{v}_n = \frac{\mathbf{u}_n}{\|\mathbf{u}_n\|}$$

are orthonormal, since

$$\|\mathbf{v}_j\| = \left\| \frac{\mathbf{u}_j}{\|\mathbf{u}_j\|} \right\| = \frac{1}{\|\mathbf{u}_j\|} \|\mathbf{u}_j\| = 1.$$

Example 2. Let V be the space of all polynomials of degree $n - 1$, and define

$$\langle f, g \rangle = \int_{-1}^{1} f(x) g(x)\, dx$$

for all functions f and g in V. It is easily verified that V is an n-dimensional Euclidean space and that

$$f_0(x) = 1, \qquad f_1(x) = x, \qquad \ldots, \qquad f_{n-1}(x) = x^{n-1}$$

form a basis for V. Applying the Gram-Schmidt orthogonalization procedure to the functions $f_0(x), \ldots, f_{n-1}(x)$ gives

$$p_0(x) = 1;$$

$$p_1(x) = x + \lambda = x - \frac{\langle f_1, p_0 \rangle}{\langle p_0, p_0 \rangle} = x - \frac{\int_{-1}^{1} x\,dx}{\int_{-1}^{1} 1\,dx} = x;$$

$$p_2(x) = x^2 + \lambda_2 x + \lambda_1$$

where

$$\lambda_2 = -\frac{\int_{-1}^{1} x^2 \cdot x\,dx}{\int_{-1}^{1} x^2\,dx} = 0, \qquad \lambda_1 = -\frac{\int_{-1}^{1} x^2\,dx}{\int_{-1}^{1} 1\,dx} = -\frac{1}{3}$$

so that

$$p_2(x) = x^2 - \frac{1}{3},$$

$$p_3(x) = x^3 + \lambda_3\left(x^2 - \frac{1}{3}\right) + \lambda_2 x + \lambda_1$$

with

$$\lambda_3 = \frac{-\int_{-1}^{1} x^3(x^2 - \frac{1}{3})\,dx}{\int_{-1}^{1} (x^2 - \frac{1}{3})^2\,dx} = 0, \qquad \lambda_2 = \frac{-\int_{-1}^{1} x^3 \cdot x\,dx}{\int_{-1}^{1} x^2\,dx} = -\frac{3}{5},$$

and

$$\lambda_1 = \frac{-\int_{-1}^{1} x^3\,dx}{\int_{-1}^{1} 1\,dx} = 0$$

so that

$$p_3(x) = x^3 - \frac{3}{5}x.$$

These polynomials $p_0(x)$, $p_1(x)$, $p_2(x)$, $p_3(x), \ldots$, when properly normalized, i.e., when multiplied by a suitable constant, are the Legendre Polynomials $P_k(x)$ discussed in Chapter 2.

We now digress for a moment to discuss the following seemingly totally unrelated problem. Let $V = C^n$ with inner product

$$\langle \mathbf{x}, \mathbf{y} \rangle = (\mathbf{x}, \mathbf{y}) = x_1\bar{y}_1 + x_2\bar{y}_2 + \cdots + x_n\bar{y}_n$$

where $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ are the coordinates of $\mathbf{x}$ and $\mathbf{y}$, respectively, and let A be a given $n \times n$ matrix. Does there exist a matrix B such that

$$(A\mathbf{x}, \mathbf{y}) = (\mathbf{x}, B\mathbf{y}) \tag{7}$$

for all vectors $\mathbf{x}$ and $\mathbf{y}$?

To answer this question, we let $A = (a_{ij})$, $B = (b_{ij})$ and observe that

$$(A\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{n} \left(\sum_{i=1}^{n} a_{ij} x_i \right) \bar{y}_j = \sum_{i,j=1}^{n} a_{ji} x_i \bar{y}_j$$

and

$$(\mathbf{x}, B\mathbf{y}) = \sum_{i=1}^{n} x_i \overline{\left(\sum_{j=1}^{n} b_{ij} y_j \right)} = \sum_{i,j=1}^{n} \overline{b_{ij}} x_i \bar{y}_j.$$

Hence, Equation (7) is satisfied if and only if $\overline{b_{ij}} = a_{ji}$ or, equivalently,

$$b_{ij} = \overline{a_{ji}}. \tag{8}$$

Definition. The matrix B whose elements are given by Equation (8) is called the *adjoint* of A and is denoted by A^*. If $A^* = A$, then A is said to be *selfadjoint* or *Hermitian*, and

$$(A\mathbf{x}, \mathbf{y}) = (\mathbf{x}, A\mathbf{y}) \quad \text{for all } \mathbf{x} \text{ and } \mathbf{y}.$$

Remark 1. If A is real, then A^* is simply A^{T}, the transpose of A. Thus, if A is real and symmetric, i.e., $A = A^{\mathrm{T}}$, then A is Hermitian.

Remark 2. It is a simple matter to show (see Exercise 6) that $(A^*)^* = A$ and that

$$(\mathbf{x}, A\mathbf{y}) = (A^*\mathbf{x}, \mathbf{y}) \quad \text{for all } \mathbf{x} \text{ and } \mathbf{y}.$$

Example 3. Let

$$A = \begin{bmatrix} 1+i & 2 & i \\ 3 & -i & 1-i \\ i & 2i & 3i \end{bmatrix}.$$

Then,

$$A^* = \begin{bmatrix} 1-i & 3 & -i \\ 2 & i & -2i \\ -i & 1+i & -3i \end{bmatrix}.$$

Example 4. Let

$$A = \begin{bmatrix} 2 & 3 & 4 \\ 3 & 4 & 5 \\ 4 & 5 & 6 \end{bmatrix}.$$

Then $A^* = A^{\mathrm{T}} = A$.

Selfadjoint matrices have several very important properties, which we now wish to describe.

Property 1. Let A be selfadjoint and let $\mathbf{x}$ be an eigenvector of A with eigenvalue λ. Since A is selfadjoint, $(A\mathbf{x},\mathbf{x}) = (\mathbf{x}, A\mathbf{x})$, but

$$(A\mathbf{x},\mathbf{x}) = (\lambda\mathbf{x},\mathbf{x}) = \lambda(\mathbf{x},\mathbf{x}),$$

while

$$(\mathbf{x},A\mathbf{x}) = (\mathbf{x},\lambda\mathbf{x}) = \bar{\lambda}(\mathbf{x},\mathbf{x}).$$

Hence, $\lambda = \bar{\lambda}$.

Corollary. *The eigenvalues of a real symmetric matrix are real.*

Property 2. Let A be selfadjoint, and let $\mathbf{u}_1$ and $\mathbf{u}_2$ be eigenvectors of A with eigenvalues λ_1 and λ_2, respectively, with $\lambda_1 \neq \lambda_2$. Then, $\mathbf{u}_1$ and $\mathbf{u}_2$ are orthogonal, i.e., $(\mathbf{u}_1,\mathbf{u}_2) = 0$.

PROOF. Observe first that

$$(A\mathbf{u}_1,\mathbf{u}_2) = (\lambda_1\mathbf{u}_1,\mathbf{u}_2) = \lambda_1(\mathbf{u}_1,\mathbf{u}_2).$$

Since A is selfadjoint,

$$(A\mathbf{u}_1,\mathbf{u}_2) = (\mathbf{u}_1,A\mathbf{u}_2) = (\mathbf{u}_1,\lambda_2\mathbf{u}_2) = \lambda_2(\mathbf{u}_1,\mathbf{u}_2)$$

since λ_2 is real. Hence,

$$\lambda_1(\mathbf{u}_1,\mathbf{u}_2) = \lambda_2(\mathbf{u}_1,\mathbf{u}_2), \tag{9}$$

and since $\lambda_1 \neq \lambda_2$, Equation (9) now forces $(\mathbf{u}_1,\mathbf{u}_2) = 0$, which proves Property 2.

In actuality, we can state a much stronger property than Property 2: Not only are the eigenvectors belonging to distinct eigenvalues of a selfadjoint matrix orthogonal, but every $n \times n$ selfadjoint matrix posesses a full complement of n mutually orthogonal eigenvectors. This is the content of Property 3.

Property 3. Every $n \times n$ selfadjoint matrix A has n mutually orthogonal eigenvectors.

PROOF. Every matrix has at least one eigenvector. Let $\mathbf{v}_1$ be an eigenvector of A with eigenvalue λ_1, and let W be the subspace of V, which consists of all vectors in V that are orthogonal to $\mathbf{v}_1$. We leave it to the reader to verify (see Exercise 7) that W is a linear subspace of V of dimension $n-1$. The crucial observation now is that A takes W into itself; i.e., if $\mathbf{x}$ is orthogonal to $\mathbf{v}_1$, then so is $A\mathbf{x}$. But this follows immediately from the observation that

$$(A\mathbf{x},\mathbf{v}_1) = (\mathbf{x},A^*\mathbf{v}_1) = (\mathbf{x},A\mathbf{v}_1) = (\mathbf{x},\lambda_1\mathbf{v}_1) = \lambda_1(\mathbf{x},\mathbf{v}_1).$$

Thus, $(\mathbf{x},\mathbf{v}_1)=0 \Rightarrow (A\mathbf{x},\mathbf{v}_1)=0$.

The next part of our proof is a little tricky. Recall from Section 3.7 that every $n \times n$ matrix A induces a linear transformation $\mathcal{A}$ on $C^n(R^n)$ with

$$\mathcal{A}(\mathbf{x}) = A\mathbf{x}. \tag{10}$$

Conversely, every linear transformation $\mathcal{A}$ on $C^n(R^n)$ can be written in the form (10) for some appropriate matrix A. Let $\mathcal{A}$ be the linear transformation induced by A^*. From the remarks above, $\mathcal{A}$ takes W into itself. Thus, assuming that (see Exercise 8) every linear transformation that takes a vector space into itself has at least one eigenvector, we see that $\mathcal{A}$ has at least one eigenvector $\mathbf{v}_2$ in W; i.e.,

$$\mathcal{A}(\mathbf{v}_2) = A\mathbf{v}_2 = \lambda_2\mathbf{v}_2$$

for some number λ_2. Proceeding inductively, that is, by considering the subspace of W perpendicular to $\mathbf{v}_2$, we can produce n eigenvectors $\mathbf{v}_1, \mathbf{v}_2,\ldots, \mathbf{v}_n$ of A that are mutually orthogonal, and this completes the proof of Property 3.

Corollary. *Every symmetric $n \times n$ matrix A has n mutually orthogonal eigenvectors.*

At the beginning of this section we mentioned that orthogonal bases were often very useful in applications. In Chapter 3, we observed that a basis that consists of eigenvectors of a matrix A is also very useful since in this basis the matrix is diagonal. The ideal situation in practice is a basis of eigenvectors of a selfadjoint matrix A, for then we have the best of both worlds, i.e., the eigenvectors form an orthogonal basis. This is the real secret behind the famous Sturm-Liouville Theory to which we now turn our attention.

Exercises

1. Let V be the space of all polynomials of degree $\leqslant 2$ on the interval $[-2,2]$, with inner product

$$\langle f,g\rangle = \int_{-2}^{2} x^2 f(x)g(x)\,dx.$$

Starting with the basis 1, x, and x^2, use the Gram-Schmidt procedure to construct an orthonormal basis for V.

2. Let V be the space of all polynomials of degree $< n$, on the interval $[a,b]$, and define

$$\langle f,g\rangle = \int_a^b r(x)f(x)g(x)\,dx.$$

Use the Gram-Schmidt procedure to obtain an orthogonal basis for V in the case
(a) $(a,b)=(-1,1)$, $r(x)=|x|^{1/2}$;
(b) $(a,b)=(-\infty,\infty)$, $r(x)=e^{-x}$.

3. Obtain the first four Hermite polynomials by applying the Gram-Schmidt procedure to the monomials 1, x, x^2, and x^3, with weight function $r(x) = e^{-x^2/2}$.

4. Let V be the n-dimensional space spanned by the n functions e^{-kx}, $k = 0, 1, \ldots, n-1$, and let

$$(f,g) = \int_0^\infty f(x)g(x)\,dx$$

for f, g in V. Obtain an orthogonal basis for V.

5. Obtain the first four Laguerre polynomials by applying the Gram-Schmidt procedure to the monomials 1, x, x^2, and x^3, with weight function $r(x) = e^{-x}$.

6. Show that $(A^*)^* = A$.

7. Show that the subspace W introduced in the proof of Property 3 has dimension $n-1$.

8. Show that every linear transformation that takes a vector space into itself has at least one eigenvector.

6.4 Sturm-Liouville theory

In Chapter 5 we studied equations of the form

$$L[y](x) = y''(x) + \lambda y(x) = 0 \tag{1}$$

subject to the boundary conditions

$$a_1 y(0) + b_1 y'(0) = 0, \qquad a_2 y(l) + b_2 y'(l) = 0. \tag{1'}$$

Equation (1) together with the boundary conditions (1′) was referred to as a boundary value problem. We then stated a theorem to the effect that this boundary value problem has a nontrivial solution only for a special set of numbers

$$\lambda_1 < \lambda_2 < \lambda_3 < \cdots < \lambda_n < \cdots$$

called *eigenvalues* and every differentiable function $y(x)$ can be expanded in a series of these eigenfunctions, i.e.,

$$y(x) = \sum_{n=1}^{\infty} c_n y_n(x), \qquad 0 < x < l.$$

For example, the boundary value problem

$$y''(x) + \lambda y(x) = 0, \qquad y(0) = 0,\ y(l) = 0, \tag{2}$$

has eigenvalues $\lambda_n = n^2\pi^2/l^2$, with corresponding eigenfunctions

$$y_n(x) = \sin\frac{n\pi x}{l}, \qquad n = 1, 2, \ldots,$$

and every differentiable function $f(x)$ can be expanded in a Fourier series of the form

$$y(x) = \sum_{n=1}^{\infty} c_n \sin\frac{n\pi x}{l}, \qquad 0 < x < l.$$

The theorem we just stated is a special case of a much more general theorem, which applies to the more general equation

$$L[y](x) = a(x)y''(x) + b(x)y'(x) + c(x)y(x) = -\lambda r(x)y(x) \tag{3}$$

and the boundary conditions

$$a_1 y(\alpha) + b_1 y'(\alpha) = 0, \qquad a_2 y(\beta) + b_2 y'(\beta) = 0. \tag{3$'$}$$

We can gain a lot of insight into this problem by studying it from a linear algebra point of view. To wit, let V be the space of all twice continuously differentiable functions. V is a linear vector space but is infinite dimensional, rather than finite dimensional. Next, define an inner product on V via the relation

$$(f,g) = \int_{\alpha}^{\beta} f(x)g(x)\,dx. \tag{4}$$

The *operator* L defined by

$$L[y](x) = a(x)y''(x) + b(x)y'(x) + c(x)y(x) \tag{5}$$

is a *linear* operator defined on V (see Section 2.1) and it is natural to ask whether L is *selfadjoint* in the sense that

$$(Lu, v) = (u, Lv) \tag{6}$$

for all functions u, v in V. The answer to this question is given in Theorem 4.

Theorem 4. *Let V be the space of all twice continuously differentiable functions on the interval $[\alpha,\beta]$. Then, the operator L defined by Equation* (5) *is selfadjoint on V, in the sense of Equation* (6), *if and only if*

$$b(x) = a'(x). \tag{7}$$

PROOF. Computing

$$(Lu, v) = \int_{\alpha}^{\beta} [a(x)u''(x) + b(x)u'(x) + c(x)u(x)]v(x)\,dx$$

and

$$(u, Lv) = \int_{\alpha}^{\beta} [a(x)v''(x) + b(x)v'(x) + c(x)v(x)]u(x)\,dx$$

we see that

$$(Lu,v)-(u,Lv)=\int_\alpha^\beta [a(x)v(x)u''(x)+b(x)v(x)u'(x)]\,dx$$
$$-\int_\alpha^\beta [a(x)u(x)v''(x)+b(x)u(x)v'(x)]\,dx.$$

Integrating the terms containing $u''(x)$ and $v''(x)$ by parts gives

$$\begin{aligned}(Lu,v)-(u,Lv) &= -\int_\alpha^\beta a(x)u'(x)v'(x)\,dx-\int_\alpha^\beta a'(x)v(x)u'(x)\,dx\\ &\quad+\int_\alpha^\beta b(x)v(x)u'(x)\,dx+\int_\alpha^\beta a(x)u'(x)v'(x)\,dx\\ &\quad+\int_\alpha^\beta a'(x)u(x)v'(x)\,dx-\int_\alpha^\beta b(x)u(x)v'(x)\,dx\\ &\quad+a(x)v(x)u'(x)|_\alpha^\beta-a(x)u(x)v'(x)|_\alpha^\beta\\ &=\int_\alpha^\beta [b(x)-a'(x)][v(x)u'(x)-u(x)v'(x)]\,dx\\ &\quad+a(x)[v(x)u'(x)-u(x)v'(x)]_\alpha^\beta.\end{aligned}\tag{8}$$

The boundary terms in Equation (8) are zero. To see this, assume that the constants b_1, b_2 in Equation (3′) are not zero (the proof is trivial if they are—see Exercise 1), and set $c_1=a_1/b_1$, $c_2=a_2/b_2$. Then

$$\begin{aligned}&a(x)[v(x)u'(x)-u(x)v'(x)]_\alpha^\beta\\ &\qquad=a(\beta)[v(\beta)u'(\beta)-u(\beta)v'(\beta)]-a(\alpha)[v(\alpha)u'(\alpha)-u(\alpha)v'(\alpha)];\end{aligned}\tag{9}$$

but

$$u'(\beta)=c_2u(\beta),\quad v'(\beta)=c_2v(\beta),\qquad u'(\alpha)=c_1u(\alpha),\quad v'(\alpha)=c_1v(\alpha)$$

since $u(x)$ and $v(x)$ satisfy the boundary conditions (3′). Hence, the right-hand side of Equation (9) becomes

$$a(\beta)[v(\beta)c_2u(\beta)-u(\beta)c_2v(\beta)]-a(\alpha)[v(\alpha)c_1u(\alpha)-u(\alpha)c_1v(\alpha)]=0.$$

Thus, we have shown that

$$(Lu,v)-(u,Lv)=\int_\alpha^\beta [b(x)-a'(x)][v(x)u'(x)-u(x)v'(x)]\,dx\tag{10}$$

for all u and v in V. The right-hand side of Equation (10) is zero if $b(x)=a'(x)$. Thus, L is certainly selfadjoint in the sense of Equation (6) if $b(x)=a'(x)$. In addition, it is intuitively clear that $b(x)=a'(x)$ is also a necessary condition for L to be selfadjoint; the right-hand side of Equation (10) cannot vanish for all u and v in V if $b(x)\neq a'(x)$. A proof of this assertion is outlined in the exercises.

In summary, the boundary value problem (3), (3′) is self-adjoint if and only if $b(x) = a'(x)$. Setting $a(x) = p(x)$ and $c(x) = -q(x)$, we can now write all selfadjoint boundary value problems in the canonical form

$$L[y](x) = [p(x)y'(x)]' - q(x)y(x) = -\lambda r(x)y(x) \tag{11}$$

$$a_1 y(\alpha) + b_1 y'(\alpha) = 0, \qquad a_2 y(\beta) + b_2 y'(\beta) = 0. \tag{11$'$}$$

Equation (11) together with the boundary conditions (11′) is often referred to as a Sturm-Liouville boundary value problem.

Remark 1. In the notation of Equation (11), Equation (8) becomes

$$(Lu, v) - (u, Lv) = -p(x)[u'(x)v(x) - u(x)v'(x)]_\alpha^\beta. \tag{12}$$

Equation (12) is known as Lagrange's identity, and if u and v satisfy the boundary conditions (11′), then Lagrange's identity reduces to

$$(Lu, v) - (u, Lv) = 0. \tag{13}$$

Remark 2. We have derived Equation (13) in the case that u and v are both real, because that is how Lagrange's identity is usually expressed. It is a simple matter, however, to verify (see Exercise 3) that Equation (13) is also true for u and v complex, provided we define

$$(f, g) = \int_\alpha^\beta f(x)\overline{g(x)}\,dx. \tag{14}$$

Definition. A Sturm-Liouville boundary value problem is said to be *regular* if each of the following conditions hold:

(i) $r(x) > 0$ and $p(x) > 0$ for $x \in [\alpha, \beta]$.

(ii) $p(x)$, $p'(x)$, $q(x)$, and $r(x)$ are continuous on $[\alpha, \beta]$.

(iii) (α, β) is finite.

We are now ready to state and prove the main theorem of this chapter.

Theorem 5. *For any regular Sturm-Liouville boundary value problem*:

(1) *All the eigenvalues (and consequently all the eigenfunctions) are real.*

(2) *Eigenfunctions belonging to different eigenvalues are orthogonal under the inner product*

$$\langle f, g \rangle = \int_\alpha^\beta r(x)f(x)g(x)\,dx. \tag{15}$$

(3) *To each eigenvalue, there corresponds one, and only one, eigenfunction.*

(4) *There are a countably infinite number of eigenvalues $\lambda_0, \lambda_1, \lambda_2, \ldots$ with corresponding eigenfunctions $y_0(x)$, $y_1(x)$, $y_2(x), \ldots$. The eigenvalues λ_n can be ordered so that n refers to the number of zeros of y_n in the interval $[\alpha, \beta]$.*

In addition, the zeros of y_{n+1} interlace those of y_n. Finally, $\lambda_0 < \lambda_1 < \lambda_2 < \lambda_3 \cdots$ with $\lambda_n \to \infty$ as $n \to \infty$.

(5) *Let $f(x)$ be any continuously differentiable function on the interval $[\alpha, \beta]$. Then, $f(x)$ can be expanded in a convergent series of the eigenfunctions of L; i.e.,*

$$f(x) = \sum_{n=0}^{\infty} a_n y_n(x)$$

where

$$a_n = \frac{\int_\alpha^\beta r(x) f(x) y_n(x)\, dx}{\int_\alpha^\beta r(x) y_n^2(x)\, dx}. \tag{16}$$

PROOF. (1) The proof of (1) follows immediately from the selfadjointness of L: if $Ly = -\lambda y$ then

$$(Ly, y) = (-\lambda r y, y) = -\lambda \int_\alpha^\beta r(x)|y^2(x)|\, dx \tag{17}$$

and

$$(y, Ly) = (y, -\lambda r y) = -\bar{\lambda} \int_\alpha^\beta r(x)|y^2(x)|\, dx \tag{18}$$

since $r(x)$ is real. Equating Equations (17) and (18), we see that λ must be real.

(2) Let $y_n(x)$ and $y_m(x)$ be eigenfunctions of L with eigenvalues λ_n and λ_m, respectively, with $\lambda_n \neq \lambda_m$. Then

$$\begin{aligned}(Ly_n, y_m) &= (-\lambda_n r y_n, y_m)\\ &= -\lambda_n \int_\alpha^\beta r(x) y_n(x) y_m(x)\, dx;\end{aligned}$$

however, since L is selfadjoint,

$$\begin{aligned}(Ly_n, y_m) &= (y_n, Ly_m)\\ &= (y_n, -\lambda_m r y_m)\\ &= -\lambda_m \int_\alpha^\beta r(x) y_n(x) y_m(x)\, dx,\end{aligned}$$

and since $\lambda_n \neq \lambda_m$, we see that

$$\int_\alpha^\beta r(x) y_n(x) y_m(x)\, dx = \langle y_n, y_m \rangle = 0.$$

(3) Suppose that $u_1(x)$ and $u_2(x)$ are two independent eigenfunctions with eigenvalue λ. Then both u_1 and u_2 satisfy the second-order differential

equation

$$(p(x)y')' - q(x)y = -\lambda r(x)y \tag{19}$$

in addition to the boundary conditions

$$a_1 y(\alpha) + b_1 y'(\alpha) = 0, \qquad a_2 y(\beta) + b_2 y'(\beta) = 0. \tag{20}$$

Since u_1 and u_2 are independent, they form a basis for all the solutions of Equation (19), that is, every solution $y(x)$ of Equation (19) can be written in the form

$$y(x) = c_1 u_1(x) + c_2 u_2(x) \tag{21}$$

for some choice of constants c_1, c_2. Finally, observe that if $u_1(x)$ and $u_2(x)$ satisfy the boundary conditions (20), then so does the linear combination (21). Hence, every solution of Equation (19) must satisfy the boundary conditions (20), which is clearly absurd. Thus, to each eigenvalue λ of L, there corresponds only one eigenfunction.

(4) The proof of (4), though not very difficult, is quite long, so it will be omitted here.

(5) The convergence part of (5) is too difficult to present here. We wish to point out though, that if $f(x)$ can be expanded in a series of the eigenfunctions $y_n(x)$, i.e.,

$$f(x) = \sum_{n=0}^{\infty} a_n y_n(x) \tag{22}$$

then a_n must, of necessity, be given by Equation (16). To see this, assume that the series (22) is convergent and that the inner product of f with any function $y(x)$ can be distributed, that is,

$$\langle f, y \rangle = \sum_{n=0}^{\infty} a_n \langle y_n, y \rangle. \tag{23}$$

If $y(x) = y_j(x)$, then Equation (23) reduces to

$$\langle f, y_j \rangle = a_j \langle y_j, y_j \rangle \tag{24}$$

by virtue of the orthogonality of the eigenfunctions $y_0(x)$, $y_1(x)$, $y_2(x), \ldots$, and it now follows immediately from Equation (24) that

$$a_j = \frac{\langle f, y_j \rangle}{\langle y_j, y_j \rangle} = \frac{\displaystyle\int_\alpha^\beta r(x) f(x) y_j(x)\,dx}{\displaystyle\int_\alpha^\beta r(x) y_j^2(x)\,dx}. \tag{25}$$

Part (5) of Theorem 5 deals with *pointwise* convergence, i.e., the series (22) converges for each x in (α, β) if $f(x)$ is continuously differentiable. A different type of convergence, which has proven quite useful in applications, is known as convergence in the mean.

Definition. A sequence of functions $f_n(x)$, $\alpha < x < \beta$, is said to converge in the mean to $f(x)$ if

$$\|f - f_n\| = \left[\int_\alpha^\beta r(x)(f(x) - f_n(x))^2\,dx\right]^{1/2}$$

approaches zero as $n \to 0$.

Part (5) of Theorem 5 now has the following analogue in terms of *mean convergence.*

Theorem 6. *Let*

$$f_n(x) = \sum_{j=0}^{n} a_j y_j(x) \tag{26}$$

where a_j is given by Equation (25). *Then, $f_n(x)$ converges to $f(x)$ in the mean for all functions $f(x)$ with*

$$\int_\alpha^\beta r(x)f^2(x)\,dx < \infty,$$

i.e., for all functions $f(x)$ with finite norm.

Example 1. Consider the Sturm-Liouville boundary value problem

$$L[y] = y'' = -\lambda y, \qquad y'(0) = 0, y'(1) = 0. \tag{27}$$

Here $p(x) = r(x) = 1$ and $q(x) = 0$. It is easily verified that

$$y = y_n(x) = \cos n\pi x, \qquad n = 0, 1, 2, \ldots,$$

are the nontrivial solutions, i.e., eigenfunctions, of Equation (27), with corresponding eigenvalues

$$\lambda_n = n^2\pi^2, \qquad n = 0, 1, 2, \ldots.$$

Parts (1)–(3) of Theorem 5 are trivial to verify and, in fact, have been done already in Chapter 5. The zeros of $y_n(x)$ are located at the n points

$$\frac{1}{2n}, \frac{3}{2n}, \frac{5}{2n}, \ldots, \frac{2n-1}{2n},$$

and it is easily verified that the zeros of y_{n+1} interlace those of y_n. For example, the four zeros $\frac{1}{8}, \frac{3}{8}, \frac{5}{8}$, and $\frac{7}{8}$ of $y_4(x)$ interlace the three zeros $\frac{1}{6}, \frac{3}{6}, \frac{5}{6}$ of $y_3(x)$ since

$$\frac{1}{8} < \frac{1}{6} < \frac{3}{8} < \frac{3}{6} < \frac{5}{8} < \frac{5}{6} < \frac{7}{8}.$$

Finally, part (5) of Theorem 5 is simply Fourier's Theorem of Section 5.4.

Let us return now to the Hermite, Legendre, Tchebycheff, and Laguerre differential equations

$$y'' - xy' + \lambda y = 0, \qquad -\infty < x < \infty, \tag{28}$$

$$(1 - x^2)y'' - 2xy' + \lambda(\lambda + 1)y = 0, \qquad -1 < x < 1, \tag{29}$$

$$(1 - x^2)y'' - xy' + \lambda^2 y = 0, \qquad -1 < x < 1, \tag{30}$$

$$xy'' + (1 - x)y' + \lambda y = 0, \qquad 0 < x < \infty, \tag{31}$$

respectively. Except for the Legendre equation (29), the coefficient of y' in each of these equations is *not* the derivative of the coefficient of y''. We remedy this situation by multiplying through by an appropriate integrating factor (see Section 1.2). Then we can rewrite Equations (28)–(31) in the equivalent form

$$L[y](x) = \frac{d}{dx}e^{-x^2/2}y'(x) = -\lambda e^{-x^2/2}y(x), \qquad -\infty < x < \infty, \tag{28'}$$

$$L[y](x) = \frac{d}{dx}(1 - x^2)y'(x) = -\lambda(\lambda + 1)y(x), \qquad -1 < x < 1, \tag{29'}$$

$$L[y](x) = \frac{d}{dx}(1 - x^2)^{1/2}y'(x) = \frac{-\lambda^2}{(1 - x^2)^{1/2}}y(x), \qquad -1 < x < 1, \tag{30'}$$

$$L[y](x) = \frac{d}{dx}xe^{-x}y'(x) = -\lambda e^{-x}y(x), \qquad 0 < x < \infty. \tag{31'}$$

Indeed, we can now begin to understand the presence of the factor $r(x)$ in the Sturm-Liouville problem

$$L[y](x) = -\lambda r(x)y(x).$$

Our next step is to determine what boundary conditions should be imposed on the solutions of Equations (28)–(31). To this end, observe that none of the Equations (28′)–(31′) are regular in the sense of the definition preceding Theorem 5. To wit, the functions

$$p(x) = 1 - x^2 \quad \text{and} \quad p(x) = (1 - x^2)^{1/2}$$

for the Legendre and Tchebycheff equations vanish at the end points -1 and $+1$ (which in a limiting sense is also true for the functions $e^{-x^2/2}$ and xe^{-x}, which occur in the Hermite and Laguerre equations), and the interval (α, β) for both the Hermite and Laguerre equations is not finite.

Let us return now to our proof of Theorem 4 where we derived the relation

$$(Lu, v) - (u, Lv) = -p(x)[u'(x)v(x) - u(x)v'(x)]_\alpha^\beta. \tag{32}$$

For the operators L appearing in Equations (28)–(31), the integrals appearing in the left-hand sides of Equation (32) may now have to be considered as improper integrals. In the case of the Legendre and Tchebycheff equations, the function $p(x)$ vanishes at the end points. Thus, L will certainly be self-adjoint if we impose the boundary conditions that $y'(x)$ [and hence $y(x)$, see Exercise 12] is bounded as $x \to \pm 1$. In the case of the Hermite equation, we require that

$$\lim_{x\to\pm\infty} e^{-x^2/2}[u(x)v'(x) - u'(x)v(x)] = 0. \tag{33}$$

In the case of the Laguerre equation, $p(x)$ vanishes at $x = 0$. Thus, one boundary condition is that $y'(x)$ [and hence $y(x)$] is bounded as $x \to 0$. The second boundary condition is determined from the requirement that

$$\lim_{x\to\infty} xe^{-x}[u(x)v'(x) - u'(x)v(x)] = 0. \tag{34}$$

Finally, we note that the conditions (33) and (34) translate into the boundary conditions

$$\lim_{x\to\pm\infty} e^{-x^2/2}y(x) = 0 \quad \text{and} \quad \lim_{x\to\infty} xe^{-x}y(x) = 0$$

for the Hermite and Laguerre equations, respectively.

It is interesting to note (but difficult to prove) that the boundary conditions imposed on the solutions of Equations (28)–(31) are exactly sufficient to force the power series solutions of these equations to terminate at a finite point. In other words, the only solutions of Equations (28)–(31) that satisfy our boundary conditions are the Hermite, Legendre, Tchebycheff, and Laguerre polynomials, respectively; or, in the language of linear algebra, the only nontrivial solutions of the equation

$$L[y](x) = -\lambda r(x)y(x)$$

for the operators L in Equations (28′)–(31′) and the boundary conditions discussed above are the eigenfunctions $y_n(x)$ with eigenvalues λ_n, and these eigenfunctions, when properly normalized, are the Hermite, Legendre, Tchebycheff, and Laguerre polynomials, respectively.

Remark. We have imposed boundary conditions on Equations (28)–(31) to ensure that the respective operators L are selfadjoint. This will guarantee that parts (1) and (2) of Theorem 5 are true. More generally, parts (1) and (2) of Theorem 5 are true even for singular Sturm-Liouville problems, but parts (3)–(5) may sometimes fail to be true, even though they are true for the Hermite, Legendre, Tchebycheff, and Laguerre equations.

We conclude this section by proving that the only nontrivial solutions of the Hermite equation

$$y'' - 2xy' + \lambda y = 0 \quad \text{or} \quad (e^{-x^2/2}y')' = -\lambda e^{-x^2/2}y \tag{35}$$

satisfying the boundary conditions

$$\lim_{x\to\pm\infty} e^{-x^2/2}y(x) = 0 \tag{36}$$

are the Hermite polynomials $H_n(x)$, which are the polynomial solutions of the equation

$$y'' - 2xy' + 2ny = 0.$$

In other words, all nonpolynomial solutions of Equation (35) fail to satisfy the boundary conditions (36). To this end we set

$$y(x) = \sum_{n=0}^{\infty} a_n x^n.$$

Computing

$$y'(x) = \sum_{n=0}^{\infty} n a_n x^{n-1} \quad \text{and} \quad y''(x) = \sum_{n=0}^{\infty} n(n-1) a_n x^{n-2}$$

and plugging into Equation (35) yields

$$\sum_{n=0}^{\infty} (n+2)(n+1) a_{n+2} x^n - \sum_{n=0}^{\infty} 2n a_n x^n + \sum_{n=0}^{\infty} \lambda a_n x^n = 0.$$

This, in turn, implies that

$$a_{n+2} = \frac{(2n-\lambda)}{(n+2)(n+1)} a_n. \tag{37}$$

For the moment, let us set $a_1 = 0$ and $a_0 = 1$. Then Equation (37) forces all the odd coefficients $a_3, a_5, \ldots$, to be zero. Next, observe that

$$a_{n+2} = \frac{2}{(n+2)} \frac{(n-\lambda/2)}{(n+1)} a_n. \tag{38}$$

For large n,

$$\frac{(n-\lambda/2)}{(n+1)} \approx 1.$$

Thus, for $n > N$ sufficiently large,

$$a_{n+2} \geqslant \frac{2}{n+2}(.9) a_n, \qquad n > N. \tag{39}$$

Next, look at

$$z(x) = \sum_{n=0}^{\infty} b_n x^n$$

where

$$b_{n+2} = \frac{2}{n+2}(.9) b_n \quad \text{and} \quad b_0 = 1. \tag{40}$$

Then,

$$b_2 = \frac{2}{2}(.9) = \frac{1}{1!}(.9)$$

$$b_4 = \frac{2^2}{2\cdot 4}(.9)^2 = \frac{1}{2!}(.9)^2$$

$$b_6 = \frac{2^3}{2\cdot 4\cdot 6}(.9)^3 = \frac{1}{3!}(.9)^3$$

$$\vdots$$

and

$$\begin{aligned} z(x) &= 1 + \frac{(.9)}{1!}x^2 + \frac{(.9)^2}{2!}x^4 + \frac{(.9)^3}{3!}x^6 + \cdots \\ &= 1 + \frac{(.9x^2)}{1!} + \frac{(.9x^2)^2}{2!} + \frac{(.9x^2)^3}{3!} + \cdots \\ &= e^{.9x^2}. \end{aligned}$$

It is clear from Equations (39) and (40) that for $\lambda \neq 2n$ (so that a_{n+2} is never 0)

$$y(x) \geqslant p_N(x) + e^{.9x^2}$$

for some appropriate polynomial $p_N(x)$ of degree N. But then

$$\lim_{x\to\pm\infty} e^{-x^2/2}y(x) = \lim_{x\to\pm\infty} [e^{-x^2/2}p_N(x) + e^{.4x^2}] = \infty.$$

Hence, the only way $y(x)$ can satisfy the boundary condition (36) is if $\lambda = 2n$, so that its series terminates.

The case $a_0 = 0$, $a_1 = 1$ is similar, and we leave it as an exercise for the reader (see Exercise 13).

Exercises

1. Show that the boundary terms in Equation (8) are trivially zero if either b_1 or b_2 is zero.

2. Here is the outline of a proof that L [see Equation (3)] is not selfadjoint if $b(x) \neq a'(x)$. Observe first that if $b(x) \neq a'(x)$ and we can choose u and v so that

$$v(x)u'(x) - u(x)v'(x) = b(x) - a'(x),$$

then the integral in Equation (10) reduces to

$$\int_\alpha^\beta [b(x) - a'(x)]^2\,dx \neq 0.$$

The only problem is that u and v have to satisfy the boundary conditions (3′). To modify the proof,

(a) Observe that

$$v(x)u'(x) - u(x)v'(x) = v^2(x)\left(\frac{u}{v}\right)'.$$

(b) Choose u, v, and constants $c_1, c_3, \ldots, c_{k+1}$ (with k to be determined) so that

$$\left(\frac{u}{v}\right)' = c_1[b(x)-a'(x)] + c_3[b(x)-a'(x)]^3 + c_5[b(x)-a'(x)]^5 + \cdots,$$

u and v satisfy the boundary conditions (3′), and the integral in Equation (10) is $\neq 0$.

3. Show that Equation (13) is true even for u, v complex.

4. Show that Liouville's transformation

$$y = \frac{z}{(p(x)r(x))^{1/4}}, \qquad t'(x) = \sqrt{\frac{r(x)}{p(x)}}$$

reduces Equation (11) to

$$z'' - f(t)z = -\lambda z.$$

What is $f(t)$? Note that under this transformation, the eigenvalues remain fixed and the weight function becomes unity.

5. Let $P_n(x)$ be the Legendre polynomial of degree n.
(a) Show that $P_n'(x)$ satisfies a selfadjoint equation with $\lambda = n(n+1) - 2$.
(b) Show that $\int_{-1}^{1} P_n'(x)P_m'(x)(1 - x^2)\,dx = 0$, $m \neq n$.

6. Find the eigenvalues and eigenfunctions of the boundary value problem

$$x^2y'' = -\lambda y, \qquad y(1) = 0,\ y(2) = 0.$$

(*Hint*: Try $y = x^p$.)

In each of Problems 7–9, find the eigenvalues and eigenfunctions of the given boundary value problem.

7. $y'' + \lambda y = 0; \qquad y(0) = 0,\ y(\pi) - y'(\pi) = 0$

8. $y'' + \lambda y = 0; \qquad y(0) - y'(0) = 0,\ y(1) = 0$

9. $y'' + \lambda y = 0; \qquad y(0) - y'(0) = 0,\ y(\pi) - y'(\pi) = 0$

10. For which values of λ does the boundary value problem

$$y'' - 2y' + (1 + \lambda)y = 0, \qquad y(0) = 0,\ y(1) = 0,$$

have a nontrivial solution?

11. Show that the singular boundary value problem

$$y'' + \lambda y = 0, \qquad y(0) = y(2\pi),\ y'(0) = y'(2\pi),$$

has a continuum of eigenvalues.

12. Suppose that $y'(0)$ is bounded as $x \to \pm 1$. Show that $y(x)$ must also be bounded as $x \to \pm 1$. [*Hint*: Write $y(x)$ as an integral involving $y'(x)$.]

13. Let $a_0 = 0$ and $a_1 = 1$. Show that the power series solution $\sum_{n=0}^{\infty} a_n x^n$ of Equation (35), with a_n determined from Equation (37), will not satisfy the boundary conditions (36) unless the series terminates, i.e., $\lambda = 2n$ for some integer n.

Appendix A

Some simple facts concerning functions of several variables

1. A function $f(x,y)$ is continuous at the point (x_0,y_0) if for every $\varepsilon>0$ there exists $\delta(\varepsilon)$ such that

$$|f(x,y)-f(x_0,y_0)|<\varepsilon \quad \text{if} \quad |x-x_0|+|y-y_0|<\delta(\varepsilon).$$

2. The partial derivative of $f(x,y)$ with respect to x is the ordinary derivative of f with respect to x, assuming that y is constant. In other words

$$\frac{\partial f(x_0,y_0)}{\partial x}=\lim_{h\to 0}\frac{f(x_0+h,y_0)-f(x_0,y_0)}{h}.$$

3. (a) A function $f(x_1,\ldots,x_n)$ is said to be differentiable if

$$f(x_1+\Delta x_1,\ldots,x_n+\Delta x_n)-f(x_1,\ldots,x_n)=\frac{\partial f}{\partial x_1}\Delta x_1+\ldots+\frac{\partial f}{\partial x_n}\Delta x_n+e$$

where $e/[|\Delta x_1|+\ldots+|\Delta x_n|]$ approaches zero as $|\Delta x_1|+\ldots+|\Delta x_n|$ approaches zero. (b) A function $f(x_1,\ldots,x_n)$ is differentiable in a region R if $\partial f/\partial x_1,\ldots,\partial f/\partial x_n$ are continuous in R.

4. Let $f=f(x_1,\ldots,x_n)$ and $x_j=g_j(y_1,\ldots,y_m)$, $j=1,\ldots,n$. If f and g are differentiable, then

$$\frac{\partial f}{\partial y_k}=\sum_{j=1}^{n}\frac{\partial f}{\partial x_j}\frac{\partial g_j}{\partial y_k}.$$

This is the chain rule of partial differentiation.

5. If all the partial derivatives of order two of f are continuous in a region R, then

$$\frac{\partial^2 f}{\partial x_j \partial x_k} = \frac{\partial^2 f}{\partial x_k \partial x_j}; \qquad j,k = 1,\ldots,n.$$

6. The general term in the Taylor series expansion of f about $x_1 = x_1^0,\ldots,x_n = x_n^0$ is

$$\frac{1}{j_1!\ldots j_n!}\frac{\partial^{j_1+\cdots+j_n} f(x_1^0,\ldots,x_n^0)}{\partial x_1^{j_1}\ldots\partial x_n^{j_n}}(x_1 - x_1^0)^{j_1}\ldots(x_n - x_n^0)^{j_n}.$$

Appendix B

Sequences and series

1. A sequence of numbers a_n, $n=1,2,\dots$ is said to converge to the limit a if the numbers a_n get closer and closer to a as n approaches infinity. More precisely, the sequence (a_n) converges to a if for every $\varepsilon>0$ there exists $N(\varepsilon)$ such that

$$|a-a_n|<\varepsilon \quad \text{if } n\geqslant N(\varepsilon).$$

2. **Theorem 1.** *If $a_n\to a$ and $b_n\to b$, then $a_n\pm b_n\to a\pm b$.*

PROOF. Let $\varepsilon>0$ be given. Choose $N_1(\varepsilon), N_2(\varepsilon)$ such that

$$|a-a_n|<\frac{\varepsilon}{2} \quad \text{for } n\geqslant N_1(\varepsilon), \quad \text{and} \quad |b-b_n|<\frac{\varepsilon}{2} \quad \text{for } n\geqslant N_2(\varepsilon).$$

Let $N(\varepsilon)=\max\{N_1(\varepsilon),N_2(\varepsilon)\}$. Then, for $n\geqslant N(\varepsilon)$,

$$|a_n\pm b_n-(a\pm b)|\leqslant|a_n-a|+|b_n-b|<\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon. \qquad \square$$

3. **Theorem 2.** *Suppose that $a_{n+1}\geqslant a_n$, and there exists a number K such that $|a_n|\leqslant K$, for all n. Then, the sequence (a_n) has a limit.*

PROOF. Exactly the same as the proof of Lemma 1, Section 4.8. $\square$

4. The infinite series $a_1+a_2+\dots=\sum a_n$ is said to converge if the sequence of partial sums

$$s_1=a_1, \quad s_2=a_1+a_2, \quad \dots, \quad s_n=a_1+a_2+\dots+a_n,\dots$$

has a limit.

5. The sum and difference of two convergent series are also convergent. This follows immediately from Theorem 1.

6. **Theorem 3.** *Let $a_n \geqslant 0$. The series $\sum a_n$ converges if there exists a number K such that $a_1 + \ldots + a_n \leqslant K$ for all n.*

PROOF. The sequence of partial sums $s_n = a_1 + \ldots + a_n$ satisfies $s_{n+1} \geqslant s_n$. Since $s_n \leqslant K$, we conclude from Theorem 2 that $\sum a_n$ converges. □

7. **Theorem 4.** *The series $\sum a_n$ converges if there exists a number K such that $|a_1| + \ldots + |a_n| \leqslant K$ for all n.*

PROOF. From Theorem 3, $\sum |a_n|$ converges. Let $b_n = a_n + |a_n|$. Clearly, $0 \leqslant b_n \leqslant 2|a_n|$. Thus, $\sum b_n$ also converges. This immediately implies that the series

$$\sum a_n = \sum [b_n - |a_n|]$$

also converges. □

8. **Theorem 5** (Cauchy ratio test). *Suppose that the sequence $|a_{n+1}/a_n|$ has a limit λ. Then, the series $\sum a_n$ converges if $\lambda < 1$ and diverges if $\lambda > 1$.*

PROOF. Suppose that $\lambda < 1$. Then, there exists $\rho < 1$, and an index N, such that $|a_{n+1}| \leqslant \rho |a_n|$ for $n \geqslant N$. This implies that $|a_{N+p}| \leqslant \rho^p |a_N|$. Hence,

$$\sum_{n=N}^{N+K} |a_n| \leqslant (1 + \rho + \ldots + \rho^K)|a_N| \leqslant \frac{|a_N|}{1-\rho},$$

and $\sum a_n$ converges.

If $\lambda > 1$, then $|a_{N+p}| \geqslant \rho^p |a_N|$, with $\rho > 1$. Thus $|a_n| \to \infty$ as $n \to \infty$. But $|a_n|$ must approach zero if $\sum a_n$ converges to s, since

$$|a_{n+1}| = |s_{n+1} - s_n| \leqslant |s_{n+1} - s| + |s_n - s|$$

and both these quantities approach zero as n approaches infinity. Thus, $\sum a_n$ diverges. □

Appendix C

C Programs

C Program 1

```
#include <stdio.h>
#include <math.h>

int main( )
{
  double X[200]={1};
  int k, N=15;

  printf("       N%10sX[N]\n", " ");
  for (k=0; k<=N; k++)
    {
      printf("%4d%4s%17.9lf\n", k, "  ", X[k]);
      X[k+1]=0.25+sin(X[k]);
    }
}
```

C Program 2

```
#include <stdio.h>

int main( )
{
  double X[200]={1.4};
  int k, N=5;

  printf("     N%10sX[N]\n", " ");
  for (k=0; k<=N; k++)
    {
      printf("%4d%4s%17.91f\n", k, " ", X[k]);
      X[k+1]=X[k]/2+1/X[k];
    }
}
```

C Program 3

```
#include <stdio.h>
#include <math.h>

int main( )
{
  double a, d, V[200]={45.7};
  double c=0.08, g=32.2, W=527.436, B=470.327;
  int k, N=15;

  a=(W-B)/c;
  d=300*c*g/W;
  printf("     N%10sX[N]\n", " ");
  for (k=0; k<=N; k++)
    {
      printf("%4d%4s%17.91f\n", k, " ", V[k]);
      V[k+1]=V[k]+((a-V[k])/V[k])
             *(V[k]+d+a*log(1-(V[k]/a)));
    }
}
```

C Program 4

```
#include <stdio.h>
#include <math.h>

int main( )
{
  double T[1000]={0}, Y[1000]={0.5}, a=1, h;
  int k, N=10;

  h=a/N;
  for (k=0; k<N; k++)
    {
      T[k+1]=T[k]+h;
      Y[k+1]=Y[k]+h*(1+pow(Y[k]-T[k], 2.0));
    }
  printf("%4s%16s\n", "T", "Y");
  for (k=0; k<=N; k++)
    printf("%10.71f%2s%16.91f\n", T[k], "  ", Y[k]);
}
```

C Program 5

```
#include <stdio.h>
#include <math.h>

int main( )
{
  double T[1000]={0}, Y[1000]={0.5}, a=1, h;
  int k, N=10;

  h=a/N;
  for (k=0; k<N; k++)
    {
      T[k+1]=T[k]+h;
      Y[k+1]=Y[k]+h*(1+pow(Y[k]-T[k], 2.0))
                 +h*h*pow(Y[k]-T[k], 3.0);
    }
  printf("%4s%16s\n", "T", "Y");
  for (k=0; k<=N; k++)
    printf("%10.71f%2s%16.91f\n", T[k], "  ", Y[k]);
}
```

C Program 6

```
#include <stdio.h>
#include <math.h>

int main( )
{
   double T[1000]={0}, Y[1000]={0.5}, a=1, h, R;
   int k, N=10;

   h=a/N;
   for (k=0; k<N; k++)
      {
         R=1+pow(Y[k]-T[k], 2.0);
         T[k+1]=T[k]+h;
         Y[k+1]=Y[k]+(h/2)*(R+1
                     +pow(Y[k]+h*R-T[k+1], 2.0));
      }
   printf("%4s%16s\n", "T", "Y");
   for (k=0; k<=N; k++)
      printf("%10.71f%2s%16.91f\n", T[k], "  ", Y[k]);
}
```

C Program 7

```
#include <stdio.h>
#include <math.h>

int main( )
{
   double T[1000]={0}, Y[1000]={0.5}, a=1, h;
   double LK1, LK2, LK3, LK4;
   int k, N=10;

   h=a/N;
   for (k=0; k<N; k++)
      {
         T[k+1]=T[k]+h;
         LK1=1+pow(Y[k]-T[k], 2.0);
         LK2=1+pow((Y[k]+(h/2)*LK1)-(T[k]+h/2), 2.0);
         LK3=1+pow((Y[k]+(h/2)*LK2)-(T[k]+h/2), 2.0);
         LK4=1+pow((Y[k]+h*LK3)-(T[k]+h), 2.0);
         Y[k+1]=Y[k]+(h/6)*(LK1+LK4+2*(LK2+LK3));
      }
   printf("%4s%16s\n", "T", "Y");
   for (k=0; k<=N; k++)
      printf("%10.71f%2s%16.91f\n", T[k], "  ", Y[k]);
}
```

C Program 8

```
#include <stdio.h>
#include <math.h>

int main( )
{
  double T[1000]={0}, Y[1000]={0}, a=1, h;
  int k, N=10;

  h=a/N;
  for (k=0; k<N; k++)
    {
      T[k+1]=T[k]+h;
      Y[k+1]=Y[k]+h*(Y[k]*(1+exp(-Y[k]))+exp(T[k]));
    }
  printf("%4s%10s%20s\n", "N", "h", "Y[N]");
  printf("%4d%2s%10.7lf%20.10lf\n", N, "  ", h, Y[N]);
}
```

C Program 9

```
#include <stdio.h>
#include <math.h>

int main( )
{
  double T[1000]={0}, Y[1000]={0}, a=1, h;
  double DY1, DY2;
  int k, N=10;

  h=a/N;
  for (k=0; k<N; k++)
    {
      T[k+1]=T[k]+h;
      DY1=1+(1-Y[k])*exp(-Y[k]);
      DY2=Y[k]*(1+exp(-Y[k]))+exp(T[k]);
      Y[k+1]=Y[k]+h*DY2+(h*h/2)*(exp(T[k])+DY1*DY2);
    }
  printf("%4s%10s%20s\n", "N", "h", "Y[N]");
  printf("%4d%2s%10.7lf%20.10lf\n", N, "  ", h, Y[N]);
}
```

C Program 10

```
#include <stdio.h>
#include <math.h>

int main( )
{
  double T[1000]={0}, Y[1000]={0}, a=1, h;
  double R1, R2;
  int k, N=10;

  h=a/N;
  for (k=0; k<N; k++)
    {
      T[k+1]=T[k]+h;
      R1=Y[k]*(1+exp(-Y[k]))+exp(T[k]);
      R2=(Y[k]+h*R1)*(1+exp(-(Y[k]+h*R1)))+exp(T[k+1]);
      Y[k+1]=Y[k]+(h/2)*(R1+R2);
    }
  printf("%4s%10s%20s\n", "N", "h", "Y[N]");
  printf("%4d%2s%10.7lf%2s%18.10lf\n", N, "  ", h, "  ", Y[N]);
}
```

C Program 11

```
#include <stdio.h>
#include <math.h>

int main( )
{
  double T[1000]={0}, Y[1000]={0}, a=1, h;
  double LK1, LK2, LK3, LK4;
  int k, N=10;

  h=a/N;
  for (k=0; k<N; k++)
    {
      T[k+1]=T[k]+h;
      LK1=Y[k]*(1+exp(-Y[k]))+exp(T[k]);
      LK2=(Y[k]+(h/2)*LK1)*(1+exp(-(Y[k]+(h/2)*LK1)))
            +exp(T[k]+(h/2));
      LK3=(Y[k]+(h/2)*LK2)*(1+exp(-(Y[k]+(h/2)*LK2)))
            +exp(T[k]+(h/2));
      LK4=(Y[k]+h*LK3)*(1+exp(-(Y[k]+h*LK3)))
            +exp(T[k+1]);
```

```
        Y[k+1]=Y[k]+(h/6)*(LK1+2*LK2+2*LK3+LK4);
    }
  printf("%4s%10s%20s\n", "N", "h", "Y[N]");
  printf("%4d%2s%10.7lf%2s%18.10lf\n", N, "  ", h, "  ", Y[N]);
}
```

C Program 12

```
#include <stdio.h>
#include <math.h>

int main( )
{
  double T[1000]={0}, Y[1000]={1}, a=1, h;
  double LK1, LK2, LK3, LK4;
  int k, N=10;

  h=a/N;
  for (k=0; k<N; k++)
    {
        T[k+1]=T[k]+h;
        LK1=T[k]*T[k]+Y[k]*Y[k];
        LK2=pow(T[k]+h/2, 2.0)+pow(Y[k]+(h/2)*LK1, 2.0);
        LK3=pow(T[k]+h/2, 2.0)+pow(Y[k]+(h/2)*LK2, 2.0);
        LK4=pow(T[k]+h, 2.0)+pow(Y[k]+h*LK3, 2.0);
        Y[k+1]=Y[k]+(h/6)*(LK1+2*LK2+2*LK3+LK4);
    }
  printf("%4s%15s\n", "T", "Y");
  for (k=0; k<=N; k++)
    printf("%10.7lf%2s%16.9lf\n", T[k], "  ", Y[k]);
}
```

C Program 13

```
#include <stdio.h>
#include <math.h>

#define PI 3.141592654

int main( )
{
  double T[1000], Y[1000], h;
  int k, N=25;

  h=2.0/(double)N;
  T[1]=h;
```

```
  Y[1]=0;
  for (k=1; k<N; k++)
    {
      T[k+1]=T[k]+h;
      if (Y[k]==0) Y[k+1]=h*T[k]*sin(PI/T[k]); else
      Y[k+1]=Y[k]+h*(Y[k]*pow(fabs(Y[k]), -0.75)
                  +T[k]*sin(PI/T[k]));
    }
  printf("%4s%15s\n", "T", "Y");
  for (k=1; k<=N; k++)
    printf("%10.7lf%2s%16.9lf\n", T[k], "  ", Y[k]);
}
```

C Program 14

```
#include <stdio.h>
#include <math.h>

int main( )
{
  double A[200]={1, 1}, T=0.5, sum;
  int k, N=20;

  A[2]= -0.5*(A[1]+A[0]);
  sum=A[0]+A[1]*T+A[2]*T*T;
  for (k=1; k<=N-2; k++)
    {
      A[k+2]=((k+1)*(k-1)*A[k+1]-A[k]+A[k-1])
              /((k+1)*(k+2));
      sum=sum+A[k+2]*pow(T, (double)(k+2));
    }
  printf("For N= %d and T= %6.4lf\n", N, T);
  printf("the sum is: %11.9lf\n", sum);
}
```

Answers to odd-numbered exercises

Chapter 1

Section 1.2

1. $y(t)=ce^{-\sin t}$ **3.** $y(t)=\dfrac{t+c}{1+t^2}$ **5.** $y(t)=\exp(-\frac{1}{3}t^3)[\int\exp(\frac{1}{3}t^3)\,dt+c]$

7. $y(t)=\dfrac{c+\int(1+t^2)^{1/2}(1+t^4)^{1/4}\,dt}{(1+t^2)^{1/2}(1+t^4)^{1/4}}$ **9.** $y(t)=\exp\left(-\int_0^t\sqrt{1+s^2}\,e^{-s}ds\right)$

11. $y(t)=\dfrac{3e^{t^2}-1}{2}$ **13.** $y(t)=e^{-t}\left[2e+\int_1^t\dfrac{e^s}{1+s^2}ds\right]$

15. $y(t)=\left(\dfrac{t^5}{5}+\dfrac{2t^3}{3}+t+c\right)(1+t^2)^{-1/2}$ **17.** $y(t)=\begin{cases}2(1-e^{-t}), & 0\leqslant t<1\\ 2(e-1)e^{-t}, & t>1\end{cases}$

21. Each solution approaches a distinct limit as $t\to 0$.

23. All solutions approach zero as $t\to\pi/2$.

Section 1.3

3. 127,328 **5.** (a) $N_{238}(t)=N_{238}(0)2^{-10^{-9}t/4.5}$; (b) $N_{235}(t)=N_{235}(0)2^{-10^{-9}t/.707}$

7. About 13,550 B.C.

Section 1.4

1. $y(t)=\dfrac{t+c}{1-ct}$ **3.** $y(t)=\tan(t-\frac{1}{2}t^2+c)$ **5.** $y(t)=\arcsin(c\sin t)$

7. $y(t)=\left[9+2\ln\left(\dfrac{1+t^2}{5}\right)\right]^{1\ 2}$, $-\infty<t<\infty$

9. $y(t)=1-[4+2t+2t^2+t^3]^{1/2}, \quad -2<t<\infty$

11. $y(t)=\dfrac{a^2kt}{1+akt}, \quad \dfrac{-1}{ak}<t<\infty$, if $a=b$

$y(t)=\dfrac{ab[1-e^{k(b-a)t}]}{a-be^{k(b-a)t}}, \quad \dfrac{1}{k(b-a)}\ln\dfrac{a}{b}<t<\infty; \quad a\neq b$

13. (b) $y(t)=-t$ and $y(t)=\dfrac{ct^2}{1-ct}$ **15.** $y(t)=\dfrac{t^2-1}{2}$

17. $y=c^2e^{-2\sqrt{t/y}}, \quad t>0; \; y=-c^2e^{2\sqrt{t/y}}, \quad t<0$ **19.** $t+ye^{t/y}=c$

21. (b) $|(b+c)(at+by)+an+bm|=ke^{(b+c)(at-cy)}$

23. $(t+2y)^2-(t+2y)=c-7t$

Section 1.5

3. $a>0.4685$

7. (a) $\dfrac{dp}{dt}=0.003p-0.001p^2-0.002$; (b) $p(t)=\dfrac{1{,}999{,}998-999{,}998e^{-0.001t}}{999{,}999-999{,}998e^{-0.001t}}$, $p(t)\to 2$ as $t\to\infty$

Section 1.6

1. $p(t)=\dfrac{Ne^{cNt}}{N-1+e^{cNt}}$

Section 1.7

1. $V(t)=\dfrac{W-B}{c}[1-e^{(-cg/W)t}]$

7. $V(t)=\sqrt{322}\left[\dfrac{K_0e^{(64.4)(t-5)/\sqrt{322}}+1}{K_0e^{(64.4)(t-5)/\sqrt{322}}-1}\right], \quad K_0=\dfrac{\sqrt{322}\left(1-\dfrac{1}{\sqrt{e}}\right)+1}{\sqrt{322}\left(1-\dfrac{1}{\sqrt{e}}\right)-1}$

9. (a) $\sqrt{V}-\sqrt{V_0}+\dfrac{mg}{c}\ln\dfrac{mg-c\sqrt{V}}{mg-c\sqrt{V_0}}=\dfrac{-ct}{2m}$; (b) $V_T=\dfrac{m^2g^2}{c^2}$

11. (a) $v\dfrac{dv}{dy}=\dfrac{-gR^3}{(y+R)^2}$; (b) $V_0=\sqrt{2gR}$

Section 1.8

3. $\dfrac{2\ln 5}{\ln 2}$ yrs. **5.** 10^4e^{10} **7.** (a) $c(t)=1-e^{-0.001t}$; (b) $1000\ln\frac{5}{4}$ min.

9. $c(t)=\frac{1}{2}(1-e^{-0.02t})+\dfrac{Q_0}{150}e^{-0.02t}$ **11.** 48% **13.** $xy=c$

15. $x^2+y^2=cy$ **17.** $y^2(\ln y^2-1)=2(c-x^2)$

SECTION 1.9

3. $t^2 \sin y + y^3 e^t = c$ **5.** $y \tan t + \sec t + y^2 = c$ **7.** $y(t) = t^{-2/3}$

9. $y(t) = -t^2 + \sqrt{t^4 - (t^3 - 1)}$ **11.** $t^3 y + \frac{1}{2}(t^2 y^2) = 10$

13. $a = -2;\ y(t) = \pm\left[\dfrac{t(2t-1)}{2(1+ct)}\right]^{1/2}$ **19.** $a = 1;\ y(t) = \arcsin[(c-t)e^t]$

SECTION 1.10

1. $y_n(t) = t^2 + \dfrac{t^4}{2!} + \dfrac{t^6}{3!} + \ldots + \dfrac{t^{2n}}{n!}$

3. $y_1(t) = e^t - 1;\quad y_2(t) = t - e^t + \dfrac{1+e^{2t}}{2}$

$$y_3(t) = -\frac{107}{48} + \frac{t}{4} + \frac{t^2}{2} + \frac{t^3}{3} + 2(1-t)e^t + \frac{(1+t)e^{2t}}{2} - \frac{e^{3t}}{3} + \frac{e^{4t}}{16}$$

19. $y(t) = \sin\dfrac{t^2}{2}$

SECTION 1.11

3. (a) $0 < x_0 < 10$ **5.** 1.48982 **7.** (c) 0.73909

Section 1.11.1

5. 0.61547 **7.** 0.22105 **9.** 1.2237

SECTION 1.12

1. $y_n = (-7)^n - \frac{1}{4}[(-7)^n - 1]$ **3.** (a) $E_n \leqslant \frac{1}{2}(3^n - 1)$; (b) $E_n \leqslant 2(2^n - 1)$

5. $y_n = a_1 \ldots a_{n-1}\left[y_1 + \displaystyle\sum_{j=1}^{n-1} \frac{b_j}{a_1 \ldots a_j}\right]$ **7.** $y_{25} = \dfrac{1}{25}\left[1 + \displaystyle\sum_{j=1}^{24} 2^j\right]$

9. (a) $x = \$251.75$; (b) $x = \$289.50$

SECTION 1.13

1. $h = 0.1, y_{10} = 1;\quad h = 0.025, y_{40} = 1$

3. $h = 0.1, y_{10} = 1;\quad h = 0.025, y_{40} = 1$

5. $h = 0.1, y_{10} = 1.80516901;\quad h = 0.025, y_{40} = 1.94633036$

Section 1.13.1

1. $E_k \leqslant \dfrac{3h}{2}(e^{4/5} - 1)$ **3.** $E_k \leqslant \dfrac{h(1+e+e^2)}{2e}[e^{e/(e+1)} - 1]$ **5.** $h \leqslant \dfrac{4(10^{-5})}{3(e^2 - 1)}$

SECTION 1.14

1. $h = 0.1, y_{10} = 1;\quad h = 0.025, y_{40} = 1$ **3.** $h = 0.1, y_{10} = 1;\quad h = 0.025, y_{40} = 1$

5. $h = 0.1, y_{10} = 2;\quad h = 0.025, y_{40} = 2$

SECTION 1.15

1. $h=0.1, y_{10}=1$; $h=0.025, y_{40}=1$ **3.** $h=0.1, y_{10}=1$; $h=0.025, y_{40}=1$

5. $h=0.1, y_{10}=1.98324929$; $h=0.025, y_{40}=1.99884368$

SECTION 1.16

1. $h=0.1, y_{10}=1$; $h=0.025, y_{40}=1$ **3.** $h=0.1, y_{10}=1$; $h=0.025, y_{40}=1$

5. $h=0.1, y_{10}=1.99997769$; $h=0.025, y_{40}=1.9999999$

SECTION 1.17

1. 2.4103 **3.** 0.0506 **5.** 0.4388

Chapter 2

SECTION 2.1

1. (a) $(4-3t)e^t$; (b) $3\sqrt{3}\ t\sin\sqrt{3}\ t$; (c) $2(4-3t)e^t+12\sqrt{3}\ t\sin\sqrt{3}\ t$; (d) $2-3t^2$; (e) $5(2-3t^2)$; (f) 0; (g) $2-3t^2$

5. (b) $W=\dfrac{-3}{2t^{3/2}}$; (d) $y(t)=2\sqrt{t}$

7. (a) $-(b\sin at\sin bt+a\cos at\cos bt)$; (b) 0; (c) $(b-a)e^{(a+b)t}$; (d) e^{2at}; (e) t; (f) $-be^{2at}$

13. $W=\dfrac{1}{t}$

SECTION 2.2

1. $y(t)=c_1e^t+c_2e^{-t}$ **3.** $y(t)=e^{3t/2}\left[c_1e^{\sqrt{5}\,t/2}+c_2e^{-\sqrt{5}\,t/2}\right]$

5. $y(t)=\frac{1}{5}(e^{4t}+4e^{-t})$ **7.** $y(t)=\dfrac{\sqrt{5}}{3}e^{-t/2}\left[e^{3\sqrt{5}\,t/10}-e^{-3\sqrt{5}\,t/10}\right]$

9. $V\geqslant -3$ **11.** $y(t)=c_1t+c_2/t^5$

Section 2.2.1

1. $y(t)=e^{-t/2}\left[c_1\cos\dfrac{\sqrt{3}\,t}{2}+c_2\sin\dfrac{\sqrt{3}\,t}{2}\right]$

3. $y(t)=e^{-t}\left(c_1\cos\sqrt{2}\,t+c_2\sin\sqrt{2}\,t\right)$

5. $y(t)=e^{-t/2}\left[\cos\dfrac{\sqrt{7}\,t}{2}-\dfrac{3}{\sqrt{7}}\sin\dfrac{\sqrt{7}\,t}{2}\right]$

9. $y(t)=e^{(t-2)/3}\left[\cos\dfrac{\sqrt{11}\,(t-2)}{3}-\dfrac{4}{\sqrt{11}}\sin\dfrac{\sqrt{11}\,(t-2)}{3}\right]$

11. $y_1(t)=\cos\omega t$, $y_2(t)=\sin\omega t$

15. $\sqrt{i} = \dfrac{\pm(1+i)}{\sqrt{2}}$; $\sqrt{1+i} = \dfrac{\pm 1}{\sqrt{2}}\left[\sqrt{\sqrt{2}+1} + i\sqrt{\sqrt{2}-1}\,\right]$

$\sqrt{-i} = \dfrac{\pm(1-i)}{\sqrt{2}}$; $\sqrt{\sqrt{i}} = \pm\left[\sqrt{\sqrt{2}-1} + i\sqrt{\sqrt{2}+1}\,\right]$, $(\sqrt{i} = e^{i\pi/4})$

19. $y(t) = \dfrac{1}{\sqrt{t}}\left[c_1\cos\dfrac{\sqrt{7}}{2}\ln t + c_2\sin\dfrac{\sqrt{7}}{2}\ln t\right]$

Section 2.2.2

1. $y(t)=(c_1+c_2t)e^{3t}$ **3.** $y(t)=(1+\frac{1}{3}t)e^{-t/3}$ **7.** $y(t)=2(t-\pi)e^{2(t-\pi)/3}$

11. $y(t) = (c_1 + c_2t)e^{t^2}$ **13.** $y(t) = c_1t + c_2(t^2 - 1)$ **15.** $y(t) = c_1(t + 1) + c_2e^{2t}$

17. $y(t)=c_1(1+3t)+c_2e^{3t}$ **19.** $y(t)=\dfrac{c_1+c_2\ln t}{t}$

Section 2.3

1. $y(t)=c_1+c_2e^{2t}+t^2$ **3.** $y(t)=e^{t^2}+2e^{t}-2e^{-t^3}$

Section 2.4

1. $y(t)=(c_1+\ln\cos t)\cos t+(c_2+t)\sin t$

3. $y(t)=c_1e^{t/2}+[c_2+9t-2t^2+\frac{1}{3}t^3]e^{t}$

5. $y(t)=\frac{1}{13}(2\cos t-3\sin t)e^{-t}-e^{-t}+\frac{24}{13}e^{-t/3}$

7. $y(t)=\displaystyle\int_0^t\sqrt{s+1}\,[e^{2(t-s)}-e^{(t-s)}]\,ds$ **9.** $y(t)=c_1t^2+\dfrac{c_2}{t}+\dfrac{t^2\ln t}{3}$

11. $y(t)=\sqrt{t}\left[c_1+c_2\ln t+\displaystyle\int_0^t f\sqrt{s}\,\cos s[\ln t-\ln s]\,ds\right]$

Section 2.5

1. $\psi(t)=\dfrac{t^3-2t-1}{3}$ **3.** $\psi(t)=t(\frac{1}{4}-\frac{1}{4}t+\frac{1}{6}t^2)e^{t}$ **5.** $\psi(t)=\dfrac{t^2}{2}e^{-t}$

7. $\psi(t)=\frac{1}{16}t[\sin 2t-2t\cos 2t]$ **9.** $\psi(t)=\frac{1}{5}+\frac{1}{17}(\cos 2t-4\sin 2t)$

11. $\psi(t)=\dfrac{-1}{50}(\cos t+7\sin t)+\left(\dfrac{t}{2}-\dfrac{1}{5}\right)\dfrac{te^{2t}}{5}$ **13.** $\psi(t)=t(e^{2t}-e^{t})$

15. $\psi(t)=-\frac{1}{16}\cos 3t+\frac{1}{4}t\sin t$

17. (b) $\psi(t)=0.005+\dfrac{1}{32{,}000{,}018}(15\sin 30t-20{,}000\cos 30t)+\frac{15}{2}\sin 10t$

Section 2.6

1. Amplitude $=\frac{1}{4}$, period $=\frac{1}{4}\pi$, frequency $=8$

7. $\alpha\geqslant\beta$, where $\left[1+(1-\beta)^2\right]e^{-2\beta}=10^{-6}$

9. $y(t)=\dfrac{e^{-(t-\pi)}}{2}\left[\left(\dfrac{1}{2}+\dfrac{(\pi+1)}{2}e^{-\pi}\right)\cos(t-\pi)+\dfrac{\pi}{2}e^{-\pi}\sin(t-\pi)\right]$

11. $\pi/2$ seconds

Section 2.6.2

3. $Q(1)=\dfrac{12}{10^6}\left[1-\left(\cos 500\sqrt{3}+\dfrac{1}{\sqrt{3}}\sin 500\sqrt{3}\right)e^{-500}\right]$

Steady state charge $=\dfrac{3}{250{,}000}$

7. $\omega=\left(\dfrac{1}{LC}-\dfrac{R^2}{2L^2}\right)^{1/2}$

Section 2.8

1. $y(t)=a_0e^{-t^2/2}+a_1\sum_{n=0}^{\infty}\dfrac{(-1)^n t^{2n+1}}{3\cdot 5\ldots(2n+1)}$

3. $y(t)=a_0\left[1+\dfrac{3t^2}{2^2}+\dfrac{3t^4}{2^4\cdot 2!}-\dfrac{3t^6}{2^6\cdot 3!}+\dfrac{3\cdot 3t^8}{2^8\cdot 4!}-\dfrac{3\cdot 3\cdot 5t^{10}}{2^{10}\cdot 5!}+\ldots\right]+a_1\left(t+\dfrac{t^3}{3}\right)$

5. $y(t)=\sum_{n=0}^{\infty}(n+1)(t-1)^{2n}$

7. $y(t)=-2[t+\dfrac{t^6}{5\cdot 6}+\dfrac{t^{11}}{5\cdot 6\cdot 10\cdot 11}+\dfrac{t^{16}}{5\cdot 6\cdot 10\cdot 11\cdot 15\cdot 16}+\ldots]$

9. (a) $y_1(t)=1-\dfrac{\lambda t^2}{2!}-\dfrac{\lambda(4-\lambda)t^4}{4!}-\dfrac{\lambda(4-\lambda)(8-\lambda)t^6}{6!}+\ldots$

(b) $y_2(t)=t+(2-\lambda)\dfrac{t^3}{3!}+(2-\lambda)(6-\lambda)\dfrac{t^5}{5!}+(2-\lambda)(6-\lambda)(10-\lambda)\dfrac{t^7}{7!}+\ldots$

11. (a) $y_1(t)=1-\alpha^2\dfrac{t^2}{2!}-\alpha^2(2^2-\alpha^2)\dfrac{t^4}{4!}$

$-\alpha^2(2^2-\alpha^2)(4^2-\alpha^2)\dfrac{t^6}{6!}+\ldots$

$y_2(t)=t+(1^2-\alpha^2)\dfrac{t^3}{3!}+(1^2-\alpha^2)(3^2-\alpha^2)\dfrac{t^5}{5!}+\ldots$

13. (a) $y(t)=1-\frac{1}{2}t^2-\frac{1}{6}t^3+\frac{1}{24}t^4+\ldots$ (b) $y(\frac{1}{2})\approx 0.8592436$

15. (a) $y(t)=1-\frac{1}{2}t^2-\frac{1}{6}t^3+\frac{1}{12}t^4+\ldots$ (b) $y(\frac{1}{2})\approx 0.86087198$

17. (a) $y(t)=3+5t-4t^2+t^3-\frac{3}{8}t^4+\ldots$ (b) $y(\frac{1}{2})\approx 5.3409005$

Section 2.8.1

1. $y(t)=c_1t+c_2/t^5$

3. $y(t)=c_1(t-1)+c_2(t-1)^2$

5. $y(t)=t(c_1+c_2\ln t)$

7. $y(t)=c_1\cos(\ln t)+c_2\sin(\ln t)$

9. $y(t)=\frac{t}{2\sqrt{3}}\left[t^{\sqrt{3}}-\frac{1}{t^{\sqrt{3}}}\right]$

SECTION 2.8.2

1. Yes **3.** Yes **5.** No

7. $$y(t)=\frac{c_1}{t}\left(1-t-\frac{t^2}{2!}-\frac{t^3}{3\cdot 3!}-\frac{t^4}{3\cdot 5\cdot 4!}-\cdots\right)$$
$$+c_2t^{1/2}\left(1+\frac{t}{5}+\frac{t^2}{5\cdot 7\cdot 2!}+\frac{t^3}{5\cdot 7\cdot 9\cdot 3!}+\cdots\right)$$

9. $$y(t)=c_1\left(1+2t+\frac{t^2}{3}\right)+c_2t^{1/2}\left(1+\frac{t}{2}+\frac{t^2}{2^2\cdot 2!\cdot 5}-\frac{t^3}{2^3\cdot 3!\cdot 5\cdot 7}+\cdots\right)$$

11. $$y(t)=c_1\left(1+\frac{3t}{1\cdot 3}+\frac{3^2t^2}{3\cdot 7\cdot 2!}+\frac{3^3t^3}{3\cdot 7\cdot 11\cdot 3!}+\cdots\right)$$
$$+c_2t^{1/4}\left(1+\frac{3t}{5}+\frac{3^2t^2}{5\cdot 9\cdot 2!}+\frac{3^3t^3}{5\cdot 9\cdot 13\cdot 3!}+\cdots\right)$$

13. $$y_1(t)=t^{5/2}\left(1+\frac{t^2}{2\cdot 5}+\frac{t^4}{2\cdot 4\cdot 5\cdot 7}+\cdots\right)$$
$$y_2(t)=t^{-1/2}\left(1-\frac{t^2}{2}-\frac{t^4}{2\cdot 4}-\frac{t^6}{2\cdot 4\cdot 6\cdot 8}-\cdots\right)$$

15. $y_1(t)=e^{t^2}/2,\ y_2(t)=t^3+\frac{t^5}{5}+\frac{t^7}{5\cdot 7}+\frac{t^9}{5\cdot 7\cdot 9}+\cdots$

17. $y_1(t)=\frac{1}{t}-1,\ y_2(t)=\frac{1}{t}[e^{-t}-(1-t)]$

19. (e) $y_2(t)=t^3e^{-t}\int\frac{e^t}{t^3}\,dt$

21. (b) $y_1(t)=t\left(2-3t+\frac{4t^2}{2!}-\frac{5t^3}{3!}+\frac{6t^4}{4!}\cdots\right)$

(c) $y_2(t)=y_1(t)\int\frac{e^{-t}}{y_1^2(t)}\,dt$

23. (c) $y_2(t)=J_0(t)\int\frac{dt}{tJ_0^2(t)}$

25. (b) $$y(t)=1-\frac{\lambda t}{(1!)^2}-\lambda\frac{(1-\lambda)t^2}{(1!)^2(2!)^2}+\cdots$$
$$+\frac{(-\lambda)(1-\lambda)\cdots(n-1-\lambda)}{(n!)^2}t^n+\cdots$$

27. (b) $y_1(t)=1+2t+t^2+\dfrac{4t^3}{15}+\dfrac{t^4}{14}+\cdots$

$$y_2(t)=t^{1/2}\left[1+\frac{5t}{6}+\frac{17t^2}{60}+\frac{89t^3}{(60)(21)}+\frac{941t^4}{(36)(21)(60)}+\cdots\right]$$

Section 2.8.3

1. $y_1(t)=1+4t+4t^2+\cdots$

$y_2(t)=y_1(t)\ln t-\left[8t+12t^2+\dfrac{176}{27}t^3+\cdots\right]$

3. (b) $J_1(t)=\frac{1}{2}t\displaystyle\sum_{n=0}^{\infty}\frac{(-1)^n t^{2n}}{2^{2n}(n+1)!n!}$

(c) $y_2(t)=-J_1(t)\ln t+\dfrac{1}{t}\left[1-\displaystyle\sum_{n=1}^{\infty}\frac{(-1)^n\left(H_n+H_n^{-1}\right)}{2^{2n}n!(n-1)!}t^{2n}\right]$

Section 2.9

1. $\dfrac{1}{s^2}$ **3.** $\dfrac{s-a}{(s-a)^2+b^2}$ **5.** $\dfrac{1}{2}\left[\dfrac{1}{s}+\dfrac{s}{s^2+4a^2}\right]$

7. $\dfrac{1}{2}\left[\dfrac{a+b}{s^2+(a+b)^2}+\dfrac{a-b}{s^2+(a-b)^2}\right]$ **9.** $\sqrt{\dfrac{\pi}{s}}$

15. $y(t)=2e^t-\frac{1}{2}e^{4t}-\frac{1}{2}e^{2t}$ **17.** $Y(s)=\dfrac{s^2+6s+6}{(s+1)^3}$

19. $Y(s)=\dfrac{2s^2+s+2}{(s^2+3s+7)(s^2+1)}$ **23.** $y(t)=-\frac{1}{6}e^t+\frac{1}{2}e^{2t}-\frac{1}{2}e^{3t}+\frac{1}{6}e^{4t}$

Section 2.10

1. $\dfrac{n!}{s^{n+1}}$ **3.** $\dfrac{2as}{(s^2+a^2)^2}$ **5.** $\dfrac{15}{8s^3}\sqrt{\dfrac{\pi}{s}}$

7. (a) $\dfrac{\pi}{2}-\arctan s$; (b) $\ln\dfrac{s}{\sqrt{s^2+a^2}}$; (c) $\ln\dfrac{s-b}{s-a}$ **9.** $-\frac{5}{3}+2e^{-t}+\frac{2}{3}e^{-3t}$

11. $\left[\cosh\dfrac{\sqrt{57}\,t}{2}+\dfrac{3}{\sqrt{57}}\sinh\dfrac{\sqrt{57}\,t}{2}\right]e^{3t/2}$ **13.** $\frac{1}{2}(3-t)t^2e^{-t}$

15. $\frac{1}{2}(\cos t-te^{-t})$ **19.** $y(t)=\cos t+\frac{5}{2}\sin t-\frac{1}{2}t\cos t$ **21.** $y(t)=\frac{1}{6}t^3e^t$

23. $y(t)=1+e^{-t}+\left[\cos\dfrac{\sqrt{3}\,t}{2}-\dfrac{7}{\sqrt{3}}\sin\dfrac{\sqrt{3}\,t}{2}\right]e^{-t/2}$

Section 2.11

1. $y(t)=(2+3t)e^{-t}+2H_3(t)\left[(t-5)+(t-1)e^{-(t-3)}\right]$

3. $y(t)=3\cos 2t-\sin 2t+\frac{1}{4}(1-\cos 2t)-\frac{1}{4}(1-\cos 2(t-4))H_4(t)$

5. $y(t)=3\cos t-\sin t+\frac{1}{2}t\sin t+\frac{1}{2}H_{\pi/2}(t)\left[\left(t-\frac{1}{2}\pi\right)\sin t-\cos t\right]$

7. $$y(t)=\frac{1}{49}\left[(7t-1)+\left(\cos\frac{\sqrt{27}\,t}{2}-\frac{13}{\sqrt{27}}\sin\frac{\sqrt{27}\,t}{2}\right)e^{-t/2}\right.$$
$$-H_2(t)(7t-1)-H_2(t)\left[13\cos\sqrt{27}\,\frac{(t-2)}{2}\right.$$
$$\left.\left.+\sqrt{27}\sin\frac{\sqrt{27}\,(t-2)}{2}\right]e^{-(t-2)/2}\right]$$

9. $y(t)=te^{t}+H_1(t)\left[2+t+(2t-5)e^{t-1}\right]-H_2(t)\left[1+t+(2t-7)e^{t-2}\right]$

Section 2.12

3. (a) $y(t)=(\cosh\frac{1}{2}t-3\sinh\frac{1}{2}t)e^{3t/2}-2H_2(t)\sinh\frac{1}{2}(t-2)e^{3(t-2)/2}$

5. $y(t)=\frac{1}{2}(\sin t-t\cos t)-H_\pi(t)\sin t$

7. $y(t)=3te^{-t}+\frac{1}{2}t^2e^{-t}+3H_3(t)(t-3)e^{-(t-3)}$

Section 2.13

1. $\dfrac{e^{bt}-e^{at}}{b-a}$ 3. $\dfrac{a\sin at-b\sin bt}{a^2-b^2}$ 5. $\dfrac{\sin at-at\cos at}{2a}$ 7. $t-\sin t$

9. $\frac{1}{2}t\sin t$ 11. $(t-2)+(t+2)e^{-t}$ 13. $y(t)=t+\frac{3}{2}\sin 2t$ 15. $y(t)=\frac{1}{2}t^2$

17. $y(t)=\frac{1}{2}\sin t+(1-\frac{3}{2}t)e^{-t}$

Section 2.14

1. $x(t)=c_1e^{3t}+3c_2e^{4t},\quad y(t)=c_1e^{3t}+2c_2e^{4t}$

3. $x(t)=[(c_1+c_2)\sin t+(c_1-c_2)\cos t]e^{-2t}, y(t)=[c_1\cos t+c_2\sin t]e^{-2t}$

5. $x(t)=\frac{7}{4}e^{3t}+\frac{1}{4}e^{-t},\quad y(t)=\frac{7}{8}e^{3t}-\frac{1}{8}e^{-t}$

7. $x(t)=\cos t\,e^{-t},\quad y(t)=(2\cos t+\sin t)e^{-t}$

9. $x(t)=2(t\cos t+3t\sin t+\sin t)e^{t},\quad y(t)=-2t\sin t\,e^{t}$

11. $x(t)=-4\sin t-\frac{1}{2}t\sin t-t\cos t+5\cos t\ln(\sec t+\tan t)$,
$y(t)=-t\sin t-t\cos t-\frac{1}{2}\sin^2 t\cos t-5\sin t\cos t+5\sin^2 t$
$-\frac{1}{2}\sin^3 t+5(\cos t-\sin t)\ln(\sec t+\tan t)$

Section 2.15

1. $y(t)=c_1e^{t}+c_2e^{-t}+c_3e^{2t}$ 3. $y(t)=(c_1+c_2t+c_3t^2)e^{2t}+c_4e^{-t}$ 5. $y(t)=0$

7. $y(t)=-3-2t-\frac{1}{2}t^2+(3-t)e^{t}$ 9. $\psi(t)=1-\cos t-\ln\cos t-\sin t\ln(\sec t+\tan t)$

11. $\psi(t)=\dfrac{1}{\sqrt{2}}\displaystyle\int_0^t\Big[\sin(t-s)/\sqrt{2}\,\cosh(t-s)/\sqrt{2}$
$-\cos(t-s)/\sqrt{2}\,\sinh(t-s)/\sqrt{2}\,\Big]g(s)\,ds$

13. $\psi(t)=\frac{1}{4}t(e^{-2t}-1)-\frac{1}{5}\sin t$ **15.** $\psi(t)=\frac{1}{4}t^2\left[\left(\frac{1}{2}-\frac{1}{3}t\right)\cos t+\left(\frac{3}{4}+\frac{1}{6}t-\frac{1}{12}t^2\right)\sin t\right]$

17. $\psi(t)=t-1+\frac{1}{2}te^{-t}$

Chapter 3

Section 3.1

1. $\dot{x}_1=x_2$, $\dot{x}_2=x_3$, $\dot{x}_3=-x_2^2$ **3.** $\dot{x}_1=x_2$, $\dot{x}_2=x_3$, $\dot{x}_3=x_4$, $\dot{x}_4=1-x_3$ **7.** $\dot{\mathbf{x}}=\begin{pmatrix}5 & 5\\ -1 & 7\end{pmatrix}\mathbf{x}$, $\mathbf{x}(3)=\begin{pmatrix}0\\6\end{pmatrix}$

9. $\dot{\mathbf{x}}=\begin{pmatrix}0 & 0 & -1\\ 1 & 0 & 0\\ 0 & -1 & 0\end{pmatrix}\mathbf{x}$, $\mathbf{x}(-1)=\begin{pmatrix}2\\3\\4\end{pmatrix}$ **11.** (a) $\begin{pmatrix}1\\3\\-1\end{pmatrix}$; (b) $\begin{pmatrix}2\\0\\-1\end{pmatrix}$; (c) $\begin{pmatrix}-1\\4\\2\end{pmatrix}$

13. (a) $\begin{pmatrix}11\\16\\7\end{pmatrix}$; (b) $\begin{pmatrix}-7\\2\\-3\end{pmatrix}$; (c) $\begin{pmatrix}5\\10\\1\end{pmatrix}$; (d) $\begin{pmatrix}-1\\9\\1\end{pmatrix}$ **15.** $\mathbf{A}=\begin{pmatrix}\frac{1}{2} & \frac{7}{2}\\ -2 & 4\end{pmatrix}$

Section 3.2

1. Yes **3.** No **5.** No **7.** Yes **9.** No **11.** Yes

Section 3.3

1. Linearly dependent **3.** Linearly independent

5. (a) Linearly dependent; (b) Linearly dependent

7. (b) $y_1(t)=e^t$, $y_2(t)=e^{-t}$ **9.** $p_1(t)=t-1$, $p_2(t)=t^2-6$

11. $\mathbf{x}^1=\begin{pmatrix}-1\\-1\\1\end{pmatrix}$ **17.** $x_1=y_1$
$$x_2=\frac{1}{\sqrt{2}}(y_2-y_3)$$
$$x_3=\frac{1}{\sqrt{2}}(y_2+y_3)$$

Section 3.4

1. $\mathbf{x}^1(t)=\begin{bmatrix}\cos\sqrt{3}\,t/2\;e^{-t/2}\\ \left(-\frac{1}{2}\cos\sqrt{3}\,t/2-\frac{1}{2}\sqrt{3}\,\sin\sqrt{3}\,t/2\right)e^{-t/2}\end{bmatrix}$

$\mathbf{x}^2(t)=\begin{bmatrix}\sin\sqrt{3}\,t/2\;e^{-t/2}\\ \left(\frac{3}{2}\cos\sqrt{3}\,t/2-\frac{1}{2}\sin\sqrt{3}\,t/2\right)e^{-t/2}\end{bmatrix}$

3. $\mathbf{x}^1(t)=\begin{pmatrix}0\\ e^t\end{pmatrix}$, $\mathbf{x}^2(t)=\begin{pmatrix}e^t\\ 2te^t\end{pmatrix}$ **5.** Yes **7.** Yes **9.** No **11.** (b) No

Section 3.5

3. -48 **5.** 0 **7.** -97 **11.** No solutions

13. $\begin{bmatrix}x_1\\x_2\\x_3\\x_4\end{bmatrix}=\begin{bmatrix}1\\0\\0\\0\end{bmatrix}$ **15.** $\begin{bmatrix}x_1\\x_2\\x_3\\x_4\end{bmatrix}=c\begin{bmatrix}-1\\-2\\1\\4\end{bmatrix}$

Section 3.6

1. $\mathbf{AB}=\begin{pmatrix}1&2&25\\1&7&10\\10&6&12\end{pmatrix}$, $\mathbf{BA}=\begin{pmatrix}15&52&0\\5&-1&1\\10&18&6\end{pmatrix}$

3. $\mathbf{AB}=\begin{pmatrix}1&1&2\\1&2&1\\2&1&1\end{pmatrix}$, $\mathbf{BA}=\begin{pmatrix}1&2&1\\2&1&1\\1&1&2\end{pmatrix}$

9. $\frac{1}{72}\begin{pmatrix}-3&5&9\\18&-6&18\\6&14&-18\end{pmatrix}$ **11.** $\frac{i}{2}\begin{pmatrix}2i&0&-2i\\1-2i&-1&-1+i\\1-2i&1&1+i\end{pmatrix}$

13. $\frac{i}{2}\begin{pmatrix}0&-2i&0\\-1&1&1-i\\1&-1&-1-i\end{pmatrix}$

Section 3.7

1. $\mathbf{b}=\begin{bmatrix}b_1\\b_2\\b_1+2b_2\end{bmatrix}$ **3.** $\mathbf{b}=\begin{bmatrix}2b_2+3b_3\\b_2\\b_3\end{bmatrix}$ **5.** $\mathbf{x}=c\begin{pmatrix}1\\-1\\1\end{pmatrix}+\begin{bmatrix}-4\\ \frac{7}{2}\\0\end{bmatrix}$

7. No solutions **9.** $\mathbf{x}=\begin{pmatrix}1\\0\\0\end{pmatrix}$ **11.** $\lambda=1,-1$

13. (a) $\lambda=-1$; (b) $\mathbf{x}=\frac{1}{6}\begin{pmatrix}5\\-1\\0\end{pmatrix}+c\begin{pmatrix}-1\\-7\\6\end{pmatrix}$

Section 3.8

1. $\mathbf{x}(t)=c_1\begin{pmatrix}1\\1\end{pmatrix}e^{3t}+c_2\begin{pmatrix}3\\2\end{pmatrix}e^{4t}$ **3.** $\mathbf{x}(t)=c_1\begin{pmatrix}1\\0\\-1\end{pmatrix}e^{-t}+c_2\begin{pmatrix}0\\2\\-1\end{pmatrix}e^{-t}+c_3\begin{pmatrix}2\\1\\2\end{pmatrix}e^{8t}$

5. $\mathbf{x}(t)=c_1\begin{pmatrix}2\\0\\-1\end{pmatrix}e^{-10t}+\left[c_2\begin{pmatrix}0\\1\\0\end{pmatrix}+c_3\begin{pmatrix}1\\0\\2\end{pmatrix}\right]e^{5t}$

7. $\mathbf{x}(t)=\frac{7}{4}\begin{pmatrix}1\\2\end{pmatrix}e^{3t}+\frac{1}{4}\begin{pmatrix}1\\-2\end{pmatrix}e^{-t}$

9. $\mathbf{x}(t)=\begin{pmatrix}1\\-2\\-1\end{pmatrix}e^{2t}$ **11.** $\mathbf{x}(t)=\begin{pmatrix}-2\\0\\3\end{pmatrix}e^{-2t}$

13. (b) $\mathbf{x}(t)=9\begin{pmatrix}1\\0\\1\end{pmatrix}e^{t-1}+\frac{4}{3}\begin{pmatrix}-7\\2\\-13\end{pmatrix}e^{-(t-1)}+\frac{4}{3}\begin{pmatrix}1\\1\\1\end{pmatrix}e^{2(t-1)}$

Section 3.9

1. $\mathbf{x}(t)=\left[c_1\begin{pmatrix}2\\1-\sin t\end{pmatrix}+c_2\begin{pmatrix}2\sin t\\\cos t+\sin t\end{pmatrix}\right]e^{-2t}$

3. $\mathbf{x}(t)=e^t\left[c_1\begin{pmatrix}2\\-2\\3\end{pmatrix}+c_2\begin{pmatrix}0\\\cos 2t\\\sin 2t\end{pmatrix}+c_3\begin{pmatrix}0\\\sin 2t\\-\cos 2t\end{pmatrix}\right]$ **5.** $\mathbf{x}(t)=\begin{pmatrix}\cos t\\2\cos t+\sin t\end{pmatrix}e^{-t}$

7. $\mathbf{x}(t)=\begin{pmatrix}2\\-2\\1\end{pmatrix}e^{-2t}+\begin{bmatrix}-\sqrt{2}\,\sin\sqrt{2}\,t-\sqrt{2}\,\cos\sqrt{2}\,t\\\cos\sqrt{2}\,t-\sqrt{2}\,\sin\sqrt{2}\,t\\-3\cos\sqrt{2}\,t\end{bmatrix}$ **9.** $\mathbf{x}(0)=\begin{pmatrix}x_1\\0\\x_3\end{pmatrix}$

Section 3.10

1. $\mathbf{x}(t)=c_1\begin{pmatrix}0\\1\\1\end{pmatrix}e^{-2t}+\left[c_2\begin{pmatrix}1\\1\\0\end{pmatrix}+c_3\begin{pmatrix}t\\t\\1\end{pmatrix}\right]e^{-t}$

3. $\mathbf{x}(t)=c_1\begin{pmatrix}0\\0\\1\end{pmatrix}e^{-2t}+\left[c_2\begin{pmatrix}1\\0\\0\end{pmatrix}+c_3\begin{pmatrix}-t\\1\\0\end{pmatrix}\right]e^{-t}$

5. $\mathbf{x}(t)=\begin{pmatrix}1\\0\\1\end{pmatrix}e^{t}$ **7.** $\mathbf{x}(t)=\begin{pmatrix}1-t\\t\\t\end{pmatrix}e^{2t}$

13. (b) $\frac{1}{10}\begin{bmatrix}7e^t-5e^{2t}+8e^{-t} & -e^t+5e^{2t}-4e^{-t} & 3e^t-5e^{2t}-2e^{-t}\\21e^t-5e^{2t}-16e^{-t} & -3e^t+5e^{2t}+8e^{-t} & 9e^t-5e^{2t}-4e^{-t}\\14e^t+10e^{2t}-24e^{-t} & -2e^t-10e^{2t}+12e^{-t} & 6e^t+10e^{2t}-6e^{-t}\end{bmatrix}$

15. $\mathbf{I}+\frac{\mathbf{A}}{\alpha}(e^{\alpha t}-1)$

Section 3.11

1. $\frac{1}{5}\begin{bmatrix}e^{-2t}+5e^{2t}-e^{3t} & 5e^{2t}-5e^{3t} & e^{-2t}-e^{3t}\\-e^{-2t}+e^{3t} & 5e^{3t} & -e^{-2t}+e^{3t}\\4e^{-2t}-5e^{2t}+e^{3t} & -5e^{2t}+5e^{3t} & 4e^{-2t}+e^{3t}\end{bmatrix}$

3. $\frac{1}{5}\begin{bmatrix} 6e^{2t}-\cos t-2\sin t & -2e^{2t}+2\cos t+4\sin t & 2e^{2t}-2\cos t+\sin t \\ 3e^{2t}-3\cos t-\sin t & -e^{2t}+6\cos t+2\sin t & e^{2t}-\cos t+3\sin t \\ 5\sin t & -10\sin t & 5\cos t \end{bmatrix}$

5. $\begin{bmatrix} (t+1)e^{-t} & (t+1)e^{-t}-e^{-2t} & e^{-t}-e^{-2t} \\ -te^{-t} & -te^{-t}+e^{-2t} & e^{-2t}-e^{-t} \\ te^{-t} & te^{-t} & e^{-t} \end{bmatrix}$

7. $\mathbf{A}=\begin{pmatrix} 3 & 1 & -1 \\ 1 & 3 & -1 \\ 3 & 3 & -1 \end{pmatrix}$ 9. $\mathbf{A}=\frac{1}{13}\begin{pmatrix} 16 & -25 & 30 \\ 8 & -6 & -24 \\ 0 & 13 & 26 \end{pmatrix}$ 11. No

Section 3.12

1. $\mathbf{x}(t)=2e^{t}\begin{pmatrix} t\cos t+3t\sin t+\sin t \\ -t\sin t \end{pmatrix}$

3. $\mathbf{x}(t)=\begin{bmatrix} -4\sin t-\frac{1}{2}t\sin t-t\cos t+5\cos t\ln(\sec t+\tan t) \\ -t\sin t-t\cos t+\sin t\cos^2 t-\frac{1}{2}\sin^2 t\cos t \\ -5\sin t\cos t+5\sin^2 t-\frac{1}{2}\sin^3 t \\ +5(\cos t-\sin t)\ln(\sec t+\tan t) \end{bmatrix}$

5. $\mathbf{x}(t)=\begin{bmatrix} 3e^{3t}-2e^{2t}-te^{2t} \\ e^{2t} \\ 3e^{3t}-2e^{2t} \end{bmatrix}$ 11. $\boldsymbol{\psi}(t)=\begin{pmatrix} \frac{3}{4}t^2+\frac{1}{2}t+\frac{1}{8} \\ -\frac{1}{4}t^2-t+\frac{1}{8} \end{pmatrix}$

13. $\boldsymbol{\psi}(t)=\begin{pmatrix} 2 \\ -2 \\ -1 \end{pmatrix}e^{t}$ 17. $\boldsymbol{\psi}(t)=\begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}te^{3t}$

Section 3.13

1. $\mathbf{x}(t)=e^{-t}\begin{pmatrix} 3 \\ 2 \end{pmatrix}+e^{4t}\begin{pmatrix} -3 \\ 3 \end{pmatrix}$

3. $\mathbf{x}(t)=-\frac{4}{3}\begin{pmatrix} 1 \\ 2 \end{pmatrix}e^{-t}+\frac{1}{6}\begin{pmatrix} 2 \\ 1 \end{pmatrix}e^{2t}+\frac{1}{2}\begin{pmatrix} 0 \\ 1 \end{pmatrix}-t\begin{pmatrix} 1 \\ 1 \end{pmatrix}+3\begin{pmatrix} 1 \\ 1 \end{pmatrix}e^{t}$

5. $\mathbf{x}(t)=e^{t}\begin{pmatrix} 1-t-t^2 \\ 1-\frac{3}{2}t^2 \end{pmatrix}$ 7. $\mathbf{x}(t)=2e^{t}\begin{pmatrix} t\cos t+3t\sin t+\sin t \\ -t\sin t \end{pmatrix}$

9. $\mathbf{x}(t)=\begin{cases} \begin{pmatrix} \cos 2t+\sin 2t \\ 2\sin 2t \end{pmatrix}, & t<\pi \\ \begin{pmatrix} \cos 2t \\ \sin 2t+\cos 2t \end{pmatrix}, & t>\pi \end{cases}$ 11. $\mathbf{x}(t)=\begin{pmatrix} 1-t \\ t \\ t \end{pmatrix}e^{2t}$

13. $\mathbf{x}(t)=\begin{bmatrix} t-t^2-\frac{1}{6}t^3-\frac{1}{2}t^4+\frac{1}{12}t^5 \\ -\frac{1}{2}t^2 \\ t^2+\frac{1}{6}t^3 \end{bmatrix}$ 15. $\mathbf{x}(t)=\begin{bmatrix} 1 \\ 1+t \\ 1 \\ 1+2t \end{bmatrix}e^{3t}$

Chapter 4

Section 4.1

1. $x=0, y=0$; $x=0, y=1$; $x=1, y=0$; $x=\frac{1}{2}, y=\frac{1}{4}$

3. $x=0, y=0, z=0$; $x=\frac{c}{d}, y=\frac{a}{b}, z=-\left(\frac{c^2}{d^2}+\frac{a^2}{b^2}\right)$

5. $x=0, y=y_0$; y_0 arb.
$x=x_0, y=1$; x_0 arb.
$x=x_0, y=-1$; x_0 arb.

7. $x=0, y=-2, z=n\pi$

Section 4.2

1. Stable **3.** Unstable **5.** Asymptotically stable

7. Asymptotically stable **9.** Stable

11. $x(t)=1$ is stable; $x(t)=0$ is unstable.

Section 4.3

1. $x=0, y=0$ is unstable; $x=1, y=0$ is unstable;
$x=-1, y=0$ is unstable; $x=0, y=2^{1/4}$ is stable;
$x=0, y=-2^{1/4}$ is stable.

3. $x=0, y=1$ is unstable; $x=0, y=-1$ is unstable;
$x=1, y=0$ is unstable; $x=-1, y=0$ is stable.

5. $x=0, y=n\pi$ is unstable for all integers n.

7. Unstable **9.** Unstable **11.** Stable **13.** Stable

15. Unstable **17.** Unstable

Section 4.4

1. $y=\cos(x-1)^2$ **3.** $y=\tan x$ **5.** $x^2+y^2=c^2$

7. The points $x=x_0, y=y_0$ with $x_0+y_0=-1$, and the circles $x^2+y^2=c^2$, minus these points.

9. The points $x=0, y=y_0$; $x=x_0, y=0$, and the curves $y=ce^{-(2/3)e^{3x}}$, $x>0$; $y=ce^{-(2/3)e^{3x}}$, $x<0$.

11. The points $x=0, y=y_0$, and the curves

$$(by-dx)-\frac{(ad-bc)}{d}\ln|c-dy|=k, \quad x>0$$

$$(by-dx)-\frac{(ad-bc)}{d}\ln|c-dy|=k, \quad x<0$$

13. The point $x=0, y=0$; and the curves $xy^2-\frac{1}{3}x^3=c$.

Section 4.5.1

3. $\frac{y}{c}-\frac{b}{c^2}\ln(b+cy)=\frac{x^2}{2a}+k$ **5.** $\frac{y}{d}-\frac{x}{c}=\frac{b}{d^2}\ln(b+dy)-\frac{a}{c^2}\ln(a+cx)+k$

SECTION 4.7

13. (b) $y^2 = x^2 + x^4 + c^2$

SECTION 4.8

5. The circle $x^2 + y^2 = 1$. Note that every point on this circle is an equilibrium point.

7. The circles $x^2 + y^2 = (2n+1)\pi/2$

SECTION 4.9

1. $\varepsilon = -1, 1$

3. $\varepsilon = -2, 2$

5. No bifurcation points

SECTION 4.10

1. $x=0, y=0$ is unstable; $x=(a/e), y=0$ is stable if $ad<ec$, and unstable if $ad>ec$; $x=(af+bc)/(ef+bd), y=(ad-ec)/(bd+ef)$ exists only for $ad>ec$ and is stable.

3. $x_1=0,\ x_2=0,\ y=0$; $\quad x_1=c/d,\ x_2=nc/[d(a_1+a_2)],\ y=na_2-a_1^2-a_1a_2$; assuming that $na_2>a_1^2+a_1a_2$.

5. (b) (i) $y_1 = \dfrac{1+a_1}{1+a_1a_2}, y_2 = \dfrac{1-a_2}{1-a_1a_2}$; (ii) $y_1 = 1, y_2 = 0$; (c) $a_2 = \frac{1}{3}$

SECTION 4.12

3. (a) $rI + \lambda I^2/2 = \gamma \ln S - rS + c$; (b) Yes

5. (a) 0.24705; (b) 0.356; (c) 0.45212; (d) 0.60025; (e) 0.74305; (f) 0.77661

SECTION 4.13

7. Either $x(t)\to 0$ or $x(t)\to \dfrac{\beta_1 c - \alpha_1}{\beta_1}$ (for $\beta_1 c - \alpha_1 > 0$)

Chapter 5

SECTION 5.1

1. $\lambda_n = \dfrac{(2n+1)^2\pi^2}{4l^2}, y(x) = c\sin\dfrac{(2n+1)\pi x}{2l}$

3. $\lambda = 0, y = c$; $\quad \lambda = \dfrac{-n^2\pi^2}{l^2}$, $y(x) = c\cos\dfrac{n\pi x}{l}$

5. $y(x) = c\sinh\sqrt{-\lambda_0}\, x$, where $\sinh\sqrt{-\lambda_0}\,\pi = \sqrt{-\lambda_0}\,\cosh\sqrt{-\lambda_0}\,\pi$; $y(x) = c\sin\sqrt{\lambda_n}\, x$, where $\tan\sqrt{\lambda_n}\,\pi = \sqrt{\lambda_n}$.

7. $\lambda = -1, y(x) = ce^x$; $\quad \lambda = n^2, y(x) = c[n\cos nx + \sin nx]$

Section 5.3

1. $u(x,t)=\sin\frac{1}{2}\pi x\, e^{-1.71\pi^2 t/4}+3\sin\frac{5}{2}\pi x\, e^{-(1.71)25\pi^2 t/4}$

3. $u(x,t)=3\sin 2\pi x\, e^{(1-4\pi^2)t}-7\sin 4\pi x\, e^{(1-16\pi^2)t}$

5. $u(t,y)=e^{2(t+y)}+e^{-3(t+y)}$

7. $u(t,y)=e^{-5t}e^{-4y}+2e^{-7t}e^{-6y}-14e^{13t}e^{14y}$

9. (a) $X''-\mu X=0;\quad Y''-(\mu+\lambda)Y=0;\quad T'-\lambda\alpha^2 T=0;$

(b) $u(x,y,t)=\sin\dfrac{n\pi x}{a}\sin\dfrac{n\pi y}{b}\, e^{\alpha^2 n^2\pi^2(b^2-a^2)t/a^2b^2}$

Section 5.4

1. $f(x)=\dfrac{4}{\pi}\left[\dfrac{\sin\pi x}{1}+\dfrac{\sin 3\pi x}{3}+\dfrac{\sin 5\pi x}{5}+\dots\right]$

3. $f(x)=\dfrac{2}{\pi}\left[\dfrac{\sin\pi x}{1}-\dfrac{\sin 2\pi x}{2}+\dfrac{\sin 3\pi x}{3}\pm\dots\right]$

5. $f(x)=\dfrac{3}{4}-\dfrac{1}{\pi}\sum_{n=1}^{\infty}\dfrac{1}{n}\left[\sin\dfrac{n\pi}{2}\cos\dfrac{n\pi x}{2}+\left(1-\cos\dfrac{n\pi}{2}\right)\sin\dfrac{n\pi x}{2}\right]$

7. $f(x)=\dfrac{e^l-1}{2l}+\sum_{n=1}^{\infty}\dfrac{e^l(-1)^n-1}{l^2+n^2\pi^2}\left[l\cos\dfrac{n\pi x}{l}-n\pi\sin\dfrac{n\pi x}{l}\right]$

9. $f(x)=\dfrac{e^l}{l}+2\sum_{n=1}^{\infty}\dfrac{l\left(e^l(-1)^n-1\right)}{l^2+n^2\pi^2}\cos\dfrac{n\pi x}{l}$

11. $f(x)=\dfrac{e^l-1}{l}+\sum_{n=1}^{\infty}\dfrac{(-1)^n(e^l-e^{-l})}{l^2+n^2\pi^2}\left[l\cos\dfrac{n\pi x}{l}+n\pi\sin\dfrac{n\pi x}{l}\right]$

13. $f(x)=\frac{3}{4}\sin x-\frac{1}{4}\sin 3x$

17. (a) $f(x)=\dfrac{\pi^2}{3}+4\sum_{n=1}^{\infty}\dfrac{(-1)^n\cos nx}{n^2}$

Section 5.5

1. $f(x)=\dfrac{e-1}{e}+\sum_{n=1}^{\infty}\dfrac{2(1-\cos n\pi/e)}{1+n^2\pi^2}\cos n\pi x$

3. $f(x)=\dfrac{3a}{4}+\sum_{n=1}^{\infty}\dfrac{4a(\cos n\pi/2\ -1)}{n^2\pi^2}\cos\dfrac{n\pi x}{2a}$

5. $f(x)=\dfrac{l}{4}+\dfrac{2l}{\pi^2}\sum_{n=1}^{\infty}\dfrac{1}{n^2}\left[2\cos n\pi/2\ -1-(-1)^n\right]\cos\dfrac{n\pi x}{l}$

7. $f(x)=\dfrac{2}{\pi}\sum_{n=1}^{\infty}\dfrac{1}{n}\left(\cos n\pi/2\ -(-1)^n\right)\sin\dfrac{n\pi x}{2}$

9. $f(x)=\sin 2x$

11. (a) $\sin x=\dfrac{2}{\pi}+\dfrac{4}{\pi}\left[\dfrac{\cos 2x}{1-2^2}+\dfrac{\cos 4x}{1-4^2}+\dots\right];\quad 0<x<1;$

(b) $\cos x=\dfrac{4}{\pi}\left[\dfrac{2\sin 2x}{2^2-1}+\dfrac{4\sin 4x}{4^2-1}+\dots\right];\quad 0<x<1;$ (c) No

Section 5.6

1. (a) $u(x,t)=\dfrac{280}{\pi}\sum_{n=0}^{\infty}\dfrac{\sin(2n+1)\pi x/10}{2n+1}e^{-0.86(2n+1)^2\pi^2 t/100}$;

(b) $u(x,t)=\sum_{n=1}^{\infty}\dfrac{2n\left(1-(-1)^n\cos 10\right)}{n^2\pi^2-100}\sin\dfrac{n\pi x}{10}e^{-0.86n^2\pi^2 t/100}$;

(c) $u(x,t)=100\sum_{n=1}^{\infty}\left[\dfrac{4\sin n\pi/2}{n^2\pi^2}-\dfrac{(-1)^n}{n\pi}\right]\sin\dfrac{n\pi x}{100}e^{-0.86n^2\pi^2 t/100}$;

(d) $u(x,t)=\dfrac{-130}{\pi}\sum_{n=1}^{\infty}\dfrac{\left((-1)^n-\cos 3\pi/10\right)}{n}\sin\dfrac{n\pi x}{10}e^{-0.86n^2\pi^2 t/100}$

5. (b) $u(x,t)=10x+800\sum_{n=1}^{\infty}\dfrac{\sin n\pi/2}{n^2\pi^2}\sin\dfrac{n\pi x}{10}e^{(1-n^2\pi^2)t}$

Section 5.7

1. $u(x,t)=\dfrac{4}{\pi}\sum_{n=0}^{\infty}\left[\dfrac{2n+1}{(2n+1)^2-4}-\dfrac{1}{2n+1}\right]\sin(2n+1)\dfrac{x}{2}\cos(2n+1)\dfrac{ct}{2}$

3. $u(x,t)=\dfrac{12}{\pi^2}\sum_{n=1}^{\infty}\dfrac{\sin 2n\pi/3}{n^2}\sin\dfrac{n\pi x}{3}\cos\dfrac{n\pi ct}{3}$

5. $u(x,t)=\dfrac{8}{\pi^2}\sum_{n=1}^{\infty}\dfrac{1}{n^2}\sin\dfrac{n\pi}{2}\sin\dfrac{n\pi x}{2}\cos\dfrac{n\pi t}{10}$

9. $u(x,y,t)=X(x)Y(y)T(t)$; $\quad Y(y)=a_2\cos\sqrt{\lambda^2-\mu^2}\,y+b_2\sin\sqrt{\lambda^2-\mu^2}\,y$

$X(x)=a_1\cos\mu x+b_1\sin\mu x \qquad T(t)=a_3\cos\lambda ct+b_3\sin\lambda ct$

Section 5.8

1. $u(x,y)=\sum_{n=1}^{\infty}c_n\sinh\dfrac{n\pi(x-a)}{b}\sin\dfrac{n\pi y}{b}$; $\quad c_n=\dfrac{-2}{b\sinh n\pi a/b}\int_0^b f(y)\sin\dfrac{n\pi y}{b}\,dy$

3. $u(x,y)=\sum_{n=1}^{\infty}c_n\sin\dfrac{n\pi x}{a}\sinh\dfrac{n\pi(y-b)}{a}$; $\quad c_n=\dfrac{-2}{a\sinh n\pi b/a}\int_0^a f(x)\sin\dfrac{n\pi x}{a}\,dx$

5. $$u(x,y)=\frac{4}{\pi}\sum_{n=1}^{\infty}\frac{\sin(2n-1)\frac{\pi x}{a}\sinh(2n-1)\frac{\pi y}{b}}{(2n-1)\sinh(2n-1)\pi b/a}+\frac{4}{\pi}\sum_{n=1}^{\infty}\frac{\sinh(2n-1)\frac{\pi x}{b}\sin(2n-1)\frac{\pi y}{b}}{(2n-1)\sinh(2n-1)\pi a/b}$$

7. $$u(x,y)=1+\frac{4}{\pi}\sum_{n=1}^{\infty}\frac{\sinh(2n-1)\pi\frac{x-a}{b}\sin(2n-1)\frac{\pi y}{b}}{(2n-1)\sinh(2n-1)\pi a/b}$$

9. $$u(x,y)=\frac{-2}{\pi}\sum_{n=1}^{\infty}\frac{(-1)^n\sinh\sqrt{1+n^2\pi^2}\,x\sin n\pi y}{n\sinh\sqrt{1+n^2\pi^2}}$$

11. $X''+\lambda X=0$; $\quad Y''+\mu Y=0$; $\quad Z''-(\lambda+\mu)Z=0$

SECTION 6.3

1. $p_0(x) = \dfrac{\sqrt{3}}{4}$, $p_1(x) = \dfrac{\sqrt{5}}{8}x$, $p_2(x) = \dfrac{5\sqrt{7}}{32}\left(x^2 - \dfrac{12}{5}\right)$

SECTION 6.4

7. $y(x) = c\sinh\sqrt{-\lambda_0}x$, where $\sinh\sqrt{-\lambda_0}\pi = \sqrt{-\lambda_0}\cosh\sqrt{-\lambda_0}\pi$;
$y(x) = c\sin\sqrt{\lambda_n}x$, where $\tan\sqrt{\lambda_n}\pi = \sqrt{\lambda_n}$

9. $\lambda = -1$, $y(x) = ce^x$; $\lambda = n^2$, $y(x) = c[n\cos nx + \sin nx]$

Index

Texts in Applied Mathematics

(continued from page ii)

31. *Brémaud:* Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues.
32. *Durran:* Numerical Methods for Wave Equations in Geophysical Fluid Dynamics.
33. *Thomas:* Numerical Partial Differential Equations: Conservation Laws and Elliptic Equations.
34. *Chicone:* Ordinary Differential Equations with Applications.
35. *Kevorkian:* Partial Differential Equations: Analytical Solution Techniques, 2nd ed.
36. *Dullerud/Paganini:* A Course in Robust Control Theory: A Convex Approach.
37. *Quarteroni/Sacco/Saleri:* Numerical Mathematics.
38. *Gallier:* Geometric Methods and Applications: For Computer Science and Engineering.
39. *Atkinson/Han:* Theoretical Numerical Analysis: A Functional Analysis Framework.
40. *Brauer/Castillo-Chávez:* Mathematical Models in Population Biology and Epidemiology.
41. *Davies:* Integral Transforms and Their Applications, 3rd ed.
42. *Deuflhard/Bornemann:* Scientific Computing with Ordinary Differential Equations.
43. *Deuflhard/Hohmann:* Numerical Analysis in Modern Scientific Computing: An Introduction, 2nd ed.
44. *Knabner/Angermann:* Numerical Methods for Elliptic and Parabolic Partial Differential Equations.
45. *Larsson/Thomée:* Partial Differential Equations with Numerical Methods.
46. *Pedregal:* Introduction to Optimization.
47. *Ockendon/Ockendon:* Waves and Compressible Flow.
48. *Hinrichsen:* Mathematical Systems Theory I.
49. *Bullo/Lewis:* Geometric Control of Mechanical Systems: Modeling, Analysis, and Design for Simple Mechanical Control Systems.