PART I

INTRODUCTORY MATERIAL

Optimal State Estimation, First Edition. By Dan J. Simon
ISBN 0471708585

CHAPTER 1

Linear systems theory

> Finally, we make some remarks on why *linear* systems are so important. The answer is simple: because we can solve them!
>
> —Richard Feynman [Fey63, p. 25-4]

This chapter reviews some essentials of linear systems theory. This material is typically covered in a linear systems course, which is a first-semester graduate level course in electrical engineering. The theory of optimal state estimation heavily relies on matrix theory, including matrix calculus, so matrix theory is reviewed in Section 1.1. Optimal state estimation can be applied to both linear and nonlinear systems, although state estimation is much more straightforward for linear systems. Linear systems are briefly reviewed in Section 1.2 and nonlinear systems are discussed in Section 1.3. State-space systems can be represented in the continuous-time domain or the discrete-time domain. Physical systems are typically described in continuous time, but control and state estimation algorithms are typically implemented on digital computers. Section 1.4 discusses some standard methods for obtaining a discrete-time representation of a continuous-time system. Section 1.5 discusses how to simulate continuous-time systems on a digital computer. Sections 1.6 and 1.7 discuss the standard concepts of stability, controllability, and observability of linear systems. These concepts are necessary to understand some of the optimal state estimation material later in the book. Students with a strong

Optimal State Estimation, First Edition. By Dan J. Simon
ISBN 0471708585

background in linear systems theory can skip the material in this chapter. However, it would still help to at least review this chapter to solidify the foundational concepts of state estimation before moving on to the later chapters of this book.

1.1 MATRIX ALGEBRA AND MATRIX CALCULUS

In this section, we review matrices, matrix algebra, and matrix calculus. This is necessary in order to understand the rest of the book because optimal state estimation algorithms are usually formulated with matrices.

A scalar is a single quantity. For example, the number 2 is a scalar. The number $1 + 3j$ is a scalar (we use j in this book to denote the square root of -1). The number π is a scalar.

A vector consists of scalars that are arranged in a row or column. For example, the vector

$$\begin{bmatrix} 1 & 3 & \pi \end{bmatrix} \tag{1.1}$$

is a 3-element vector. This vector is a called a 1×3 vector because it has 1 row and 3 columns. This vector is also called a row vector because it is arranged as a single row. The vector

$$\begin{bmatrix} -2 \\ \pi^2 \\ j \\ 0 \end{bmatrix} \tag{1.2}$$

is a 4-element vector. This vector is a called a 4×1 vector because it has 4 rows and 1 column. This vector is also called a column vector because it is arranged as a single column. Note that a scalar can be viewed as a 1-element vector; a scalar is a degenerate vector. (This is just like a plane can be viewed as a 3-dimensional shape; a plane is a degenerate 3-dimensional shape.)

A matrix consists of scalars that are arranged in a rectangle. For example, the matrix

$$\begin{bmatrix} -2 & 3 \\ 0 & \pi^2 \\ j & 0 \end{bmatrix} \tag{1.3}$$

is a 3×2 matrix because it has 3 rows and 2 columns. The number of rows and columns in a matrix can be collectively referred to as the dimension of the matrix. For example, the dimension of the matrix in the preceding equation is 3×2. Note that a vector can be viewed as a degenerate matrix. For example, Equation (1.1) is a 1×3 matrix. A scalar can also be viewed as a degenerate matrix. For example, the scalar 6 is a 1×1 matrix.

The rank of a matrix is defined as the number of linearly independent rows. This is also equal to the number of linearly independent columns. The rank of a matrix A is often indicated with the notation $\rho(A)$. The rank of a matrix is always less than or equal to the number of rows, and it is also less than or equal to the number of columns. For example, the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \tag{1.4}$$

has a rank of one because it has only one linearly independent row; the two rows are multiples of each other. It also has only one linearly independent column; the two columns are multiples of each other. On the other hand, the matrix

$$A = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix} \tag{1.5}$$

has a rank of two because it has two linearly independent rows. That is, there are no nonzero scalars c_1 and c_2 such that

$$c_1 \begin{bmatrix} 1 & 3 \end{bmatrix} + c_2 \begin{bmatrix} 2 & 4 \end{bmatrix} = \begin{bmatrix} 0 & 0 \end{bmatrix} \tag{1.6}$$

so the two rows are linearly independent. It also has two linearly independent columns. That is, there are no nonzero scalars c_1 and c_2 such that

$$c_1 \begin{bmatrix} 1 \\ 2 \end{bmatrix} + c_2 \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{1.7}$$

so the two columns are linearly independent. A matrix whose elements are comprised entirely of zeros has a rank of zero. An $n \times m$ matrix whose rank is equal to $\min(n,m)$ is called full rank. The nullity of an $n \times m$ matrix A is equal to $[m - \rho(A)]$.

The transpose of a matrix (or vector) can be taken by changing all the rows to columns, and all the columns to rows. The transpose of a matrix is indicated with a T superscript, as in A^T.[1] For example, if A is the $r \times n$ matrix

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & & \vdots \\ A_{r1} & \cdots & A_{rn} \end{bmatrix} \tag{1.8}$$

then A^T is the $n \times r$ matrix

$$A^T = \begin{bmatrix} A_{11} & \cdots & A_{r1} \\ \vdots & & \vdots \\ A_{1n} & \cdots & A_{rn} \end{bmatrix} \tag{1.9}$$

Note that we use the notation A_{ij} to indicate the scalar in the ith row and jth column of the matrix A. A symmetric matrix is one for which $A = A^T$.

The hermitian transpose of a matrix (or vector) is the complex conjugate of the transpose, and is indicated with an H superscript, as in A^H. For example, if

$$A = \begin{bmatrix} 1 & 2j & 3-j \\ 4j & 5+j & 1-3j \end{bmatrix} \tag{1.10}$$

then

$$A = \begin{bmatrix} 1 & -4j \\ -2j & 5-j \\ 3+j & \\ 1+3j & \end{bmatrix} \tag{1.11}$$

A hermitian matrix is one for which $A = A^H$.

[1]Many papers or books indicate transpose with a prime, as in A', or with a lower case t, as in A^t.

1.1.1 Matrix algebra

Matrix addition and subtraction is simply defined as element-by-element addition and subtraction. For example,

$$\begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 4 & 1 \\ 1 & -1 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 6 & 4 \\ 4 & 1 & -1 \end{bmatrix} \tag{1.12}$$

The sum $(A + B)$ and the difference $(A - B)$ is defined only if the dimension of A is equal to the dimension of B.

Suppose that A is an $n \times r$ matrix and B is an $r \times p$ matrix. Then the product of A and B is written as $C = AB$. Each element in the matrix product C is computed as

$$C_{ij} = \sum_{k=1}^{r} A_{ik} B_{kj} \qquad i = 1, \ldots, n \qquad j = 1, \ldots, p \tag{1.13}$$

The matrix product AB is defined only if the number of columns in A is equal to the number of rows in B. It is important to note that matrix multiplication does not commute. In general, $AB \neq BA$.

Suppose we have an $n \times 1$ vector x. We can compute the 1×1 product $x^T x$, and the $n \times n$ product xx^T as follows:

$$\begin{aligned} x^T x &= \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \\ &= x_1^2 + \cdots + x_n^2 \\ xx^T &= \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \\ &= \begin{bmatrix} x_1^2 & \cdots & x_1 x_n \\ \vdots & \ddots & \vdots \\ x_n x_1 & \cdots & x_n^2 \end{bmatrix} \end{aligned} \tag{1.14}$$

Suppose that we have a $p \times n$ matrix H and an $n \times n$ matrix P. Then H^T is a $n \times p$ matrix, and we can compute the $p \times p$ matrix product HPH^T.

$$\begin{aligned} HPH^T &= \begin{bmatrix} H_{11} & \cdots & H_{1n} \\ \vdots & \ddots & \vdots \\ H_{p1} & \cdots & H_{pn} \end{bmatrix} \begin{bmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & \ddots & \vdots \\ P_{n1} & \cdots & P_{nn} \end{bmatrix} \begin{bmatrix} H_{11} & \cdots & H_{p1} \\ \vdots & \ddots & \vdots \\ H_{1n} & \cdots & H_{pn} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{j,k} H_{1j} P_{jk} H_{1k} & \cdots & \sum_{j,k} H_{1j} P_{jk} H_{pk} \\ \vdots & \ddots & \vdots \\ \sum_{j,k} H_{pj} P_{jk} H_{1k} & \cdots & \sum_{j,k} H_{pj} P_{jk} H_{pk} \end{bmatrix} \end{aligned} \tag{1.15}$$

This matrix of sums can be written as the following sum of matrices:

$$
\begin{aligned}
HPH^T &= \begin{bmatrix} H_{11}P_{11}H_{11} & \cdots & H_{11}P_{11}H_{p1} \\ \vdots & \ddots & \vdots \\ H_{p1}P_{11}H_{11} & \cdots & H_{p1}P_{11}H_{p1} \end{bmatrix} + \cdots + \\
&\quad \begin{bmatrix} H_{1n}P_{nn}H_{1n} & \cdots & H_{1n}P_{nn}H_{pn} \\ \vdots & \ddots & \vdots \\ H_{pn}P_{nn}H_{1n} & \cdots & H_{pn}P_{nn}H_{pn} \end{bmatrix} \\
&= H_1P_{11}H_1^T + \cdots + H_nP_{nn}H_n^T \\
&= \sum_{j,k} H_jP_{jk}H_k^T
\end{aligned} \tag{1.16}
$$

where we have used the notation that H_k is the kth column of H.

Matrix division is not defined; we cannot divide a matrix by another matrix (unless, of course, the denominator matrix is a scalar).

An identity matrix I is defined as a square matrix with ones on the diagonal and zeros everywhere else. For example, the 3×3 identity matrix is equal to

$$
I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{1.17}
$$

The identity matrix has the property that $AI = A$ for any matrix A, and $IA = A$ (as long the dimensions of the identity matrices are compatible with those of A). The 1×1 identity matrix is equal to the scalar 1.

The determinant of a matrix is defined inductively for square matrices. The determinant of a scalar (i.e., a 1×1 matrix) is equal to the scalar. Now consider an $n \times n$ matrix A. Use the notation $A^{(i,j)}$ to denote the matrix that is formed by deleting the ith row and jth column of A. The determinant of A is defined as

$$
|A| = \sum_{j=1}^{n} (-1)^{i+j} A_{ij} |A^{(i,j)}| \tag{1.18}
$$

for any value of $i \in [1, n]$. This is called the Laplace expansion of A along its ith row. We see that the determinant of the $n \times n$ matrix A is defined in terms of the determinants of $(n-1) \times (n-1)$ matrices. Similarly, the determinants of $(n-1) \times (n-1)$ matrices are defined in terms of the determinants of $(n-2) \times (n-2)$ matrices. This continues until the determinants of 2×2 matrices are defined in terms of the determinants of 1×1 matrices, which are scalars. The determinant of A can also be defined as

$$
|A| = \sum_{i=1}^{n} (-1)^{i+j} A_{ij} |A^{(i,j)}| \tag{1.19}
$$

for any value of $j \in [1, n]$. This is called the Laplace expansion of A along its jth column. Interestingly, Equation (1.18) (for any value of i) and Equation (1.19) (for any value of j) both give identical results. From the definition of the determinant

we see that

$$\begin{aligned}
\det[A_{11}] &= A_{11} \\
\det\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} &= A_{11}A_{22} - A_{12}A_{21} \\
\det\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} &= A_{11}(A_{22}A_{33} - A_{23}A_{32}) - \\
&\quad A_{12}(A_{21}A_{33} - A_{23}A_{31}) + \\
&\quad A_{13}(A_{21}A_{32} - A_{22}A_{31}) \qquad (1.20)
\end{aligned}$$

Some interesting properties of determinants are

$$|AB| = |A||B| \qquad (1.21)$$

assuming that A and B are square and have the same dimensions. Also,

$$|A| = \prod_{i=1}^{n} \lambda_i \qquad (1.22)$$

where λ_i (the eigenvalues of A) are defined below.

The inverse of a matrix A is defined as the matrix A^{-1} such that $AA^{-1} = A^{-1}A = I$. A matrix cannot have an inverse unless it is square. Some square matrices do not have an inverse. A square matrix that does not have an inverse is called singular or invertible. In the scalar case, the only number that does not have an inverse is the number 0. But in the matrix case, there are many matrices that are singular. A matrix that does have an inverse is called nonsingular or invertible. For example, notice that

$$\begin{bmatrix} 1 & 0 \\ 2 & 3 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ -2/3 & 1/3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad (1.23)$$

Therefore, the two matrices on the left side of the equation are inverses of each other. The nonsingularity of an $n \times n$ matrix A can be stated in many equivalent ways, some of which are the following [Hor85]:

- A is nonsingular.
- A^{-1} exists.
- The rank of A is equal to n.
- The rows of A are linearly independent.
- The columns of A are linearly independent.
- $|A| \neq 0$.
- $Ax = b$ has a unique solution x for all b.
- 0 is not an eigenvalue of A.

The trace of a square matrix is defined as the sum of its diagonal elements:

$$\mathrm{Tr}(A) = \sum_i A_{ii} \tag{1.24}$$

The trace of a matrix is defined only if the matrix is square. The trace of a 1×1 matrix is equal to the trace of a scalar, which is equal to the value of the scalar. One interesting property of the trace of a square matrix is

$$\mathrm{Tr}(A) = \sum_i \lambda_i \tag{1.25}$$

That is, the trace of a square matrix is equal to the sum of its eigenvalues.

Some interesting and useful characteristics of matrix products are the following:

$$\begin{aligned} (AB)^T &= B^T A^T \\ (AB)^{-1} &= B^{-1} A^{-1} \\ \mathrm{Tr}(AB) &= \mathrm{Tr}(BA) \end{aligned} \tag{1.26}$$

This assumes that the inverses exist for the inverse equation, and that the matrix dimensions are compatible so that matrix multiplication is defined. The transpose of a matrix product is equal to the product of the transposes in the opposite order. The inverse of a matrix product is equal to the product of the inverses in the opposite order. The trace of a matrix product is independent of the order in which the matrices are multiplied.

The two-norm of a column vector of real numbers, also called the Euclidean norm, is defined as follows:

$$\begin{aligned} ||x||_2 &= \sqrt{x^T x} \\ &= \sqrt{x_1^2 + \cdots + x_n^2} \end{aligned} \tag{1.27}$$

From (1.14) we see that

$$xx^T = \begin{bmatrix} x_1^2 & \cdots & x_1 x_n \\ \vdots & \ddots & \vdots \\ x_n x_1 & \cdots & x_n^2 \end{bmatrix} \tag{1.28}$$

Taking the trace of this matrix is

$$\begin{aligned} \mathrm{Tr}(xx^T) &= x_1^2 + \cdots + x_n^2 \\ &= ||x||_2^2 \end{aligned} \tag{1.29}$$

An $n \times n$ matrix A has n eigenvalues and n eigenvectors. The scalar λ is an eigenvalue of A, and the $n \times 1$ vector x is an eigenvector of A, if the following equation holds:

$$Ax = \lambda x \tag{1.30}$$

The eigenvalues and eigenvectors of a matrix are collectively referred to as the eigendata of the matrix.[2] An $n \times n$ matrix has exactly n eigenvalues, although

[2]Eigendata have also been referred to by many other terms over the years, including characteristic roots, latent roots and vectors, and proper numbers and vectors [Fad59].

some may be repeated. This is like saying that an nth order polynomial equation has exactly n roots, although some may be repeated. From the above definitions of eigenvalues and eigenvectors we can see that

$$\begin{aligned} Ax &= \lambda x \\ A^2x &= A\lambda x \\ &= \lambda(Ax) \\ &= \lambda(\lambda x) \\ &= \lambda^2 x \end{aligned} \tag{1.31}$$

So if A has eigendata (λ, x), then A^2 has eigendata (λ^2, x). It can be shown that A^{-1} exists if and only if none of the eigenvalues of A are equal to 0. If A is symmetric then all of its eigenvalues are real numbers.

A symmetric $n \times n$ matrix A can be characterized as either positive definite, positive semidefinite, negative definite, negative semidefinite, or indefinite. Matrix A is:

- *Positive definite* if $x^T Ax > 0$ for all nonzero $n \times 1$ vectors x. This is equivalent to saying that all of the eigenvalues of A are positive real numbers. If A is positive definite, then A^{-1} is also positive definite.
- *Positive semidefinite* if $x^T Ax \geq 0$ for all $n \times 1$ vectors x. This is equivalent to saying that all of the eigenvalues of A are nonnegative real numbers. Positive semidefinite matrices are sometimes called nonnegative definite.
- *Negative definite* if $x^T Ax < 0$ for all nonzero $n \times 1$ vectors x. This is equivalent to saying that all of the eigenvalues of A are negative real numbers. If A is negative definite, then A^{-1} is also negative definite.
- *Negative semidefinite* if $x^T Ax \leq 0$ for all $n \times 1$ vectors x. This is equivalent to saying that all of the eigenvalues of A are nonpositive real numbers. Negative semidefinite matrices are sometimes called nonpositive definite.
- *Indefinite* if it does not fit into any of the above four categories. This is equivalent to saying that some of its eigenvalues are positive and some are negative.

Some books generalize the idea of positive definiteness and negative definiteness to include nonsymmetric matrices.

The weighted two-norm of an $n \times 1$ vector x is defined as

$$||x||_Q^2 = \sqrt{x^T Q x} \tag{1.32}$$

where Q is required to be an $n \times n$ positive definite matrix. The above norm is also called the Q-weighted two-norm of x. A quantity of the form $x^T Qx$ is called a quadratic in analogy to a quadratic term in a scalar equation.

The singular values σ of a matrix A are defined as

$$\begin{aligned} \sigma^2(A) &= \lambda(A^T A) \\ &= \lambda(AA^T) \end{aligned} \tag{1.33}$$

If A is an $n \times m$ matrix, then it has $\min(n, m)$ singular values. AA^T will have n eigenvalues, and A^TA will have m eigenvalues. If $n > m$ then AA^T will have the same eigenvalues as A^TA plus an additional $(n - m)$ zeros. These additional zeros are not considered to be singular values of A, because A always has $\min(n, m)$ singular values. This knowledge can help reduce effort during the computation of singular values. For example, if A is a 13×3 matrix, then it is much easier to compute the eigenvalues of the 3×3 matrix A^TA rather than the 13×13 matrix AA^T. Either computation will result in the same three singular values.

1.1.2 The matrix inversion lemma

In this section, we will derive the matrix inversion lemma, which is a tool that we will use many times in this book. It is also a tool that is frequently useful in other areas of control, estimation theory, and signal processing.

Suppose we have the partitioned matrix $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ where A and D are invertible square matrices, and the B and C matrices may or may not be square. We define E and F matrices as follows:

$$\begin{aligned} E &= D - CA^{-1}B \\ F &= A - BD^{-1}C \end{aligned} \tag{1.34}$$

Assume that E is invertible. Then we can show that

$$\begin{aligned} &\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} A^{-1} + A^{-1}BE^{-1}CA^{-1} & -A^{-1}BE^{-1} \\ -E^{-1}CA^{-1} & E^{-1} \end{bmatrix} \\ &\quad = \begin{bmatrix} I + BE^{-1}CA^{-1} - BE^{-1}CA^{-1} & -BE^{-1} + BE^{-1} \\ CA^{-1} + CA^{-1}BE^{-1}CA^{-1} - DE^{-1}CA^{-1} & -CA^{-1}BE^{-1} + DE^{-1} \end{bmatrix} \\ &\quad = \begin{bmatrix} I & 0 \\ CA^{-1} - (D - CA^{-1}B)E^{-1}CA^{-1} & (D - CA^{-1}B)E^{-1} \end{bmatrix} \\ &\quad = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \end{aligned} \tag{1.35}$$

Now assume that F is invertible. Then we can show that

$$\begin{aligned} &\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} F^{-1} & -A^{-1}BE^{-1} \\ -D^{-1}CF^{-1} & E^{-1} \end{bmatrix} \\ &\quad = \begin{bmatrix} AF^{-1} - BD^{-1}CF^{-1} & -BE^{-1} + BE^{-1} \\ CF^{-1} - CF^{-1} & -CA^{-1}BE^{-1} + DE^{-1} \end{bmatrix} \\ &\quad = \begin{bmatrix} (A - BD^{-1}C)F^{-1} & 0 \\ 0 & (D - CA^{-1}B)E^{-1} \end{bmatrix} \\ &\quad = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \end{aligned} \tag{1.36}$$

Equations (1.35) and (1.36) are two expressions for the inverse of $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$. Since these two expressions are inverses of the same matrix, they must be equal. We therefore conclude that the upper-left partitions of the matrices are equal, which gives

$$F^{-1} = A^{-1} + A^{-1}BE^{-1}CA^{-1} \tag{1.37}$$

Now we can use the definition of F to obtain

$$(A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} \tag{1.38}$$

This is called the matrix inversion lemma. It is also referred to by other terms, such as the Sherman–Morrison formula, Woodbury's identity, and the modified matrices formula. One of its earliest presentations was in 1944 by William Duncan [Dun44], and similar identities were developed by Alston Householder [Hou53]. An account of its origins and variations (e.g., singular A) is given in [Hen81]. The matrix inversion lemma is often stated in slightly different but equivalent ways. For example,

$$(A + BD^{-1}C)^{-1} = A^{-1} - A^{-1}B(D + CA^{-1}B)^{-1}CA^{-1} \tag{1.39}$$

The matrix inversion lemma can sometimes be used to reduce the computational effort of matrix inversion. For instance, suppose that A is $n \times n$, B is $n \times p$, C is $p \times n$, D is $p \times p$, and $p < n$. Suppose further that we already know A^{-1}, and we want to add some quantity to A and then compute the new inverse. A straightforward computation of the new inverse would be an $n \times n$ inversion. But if the new matrix to invert can be written in the form of the left side of Equation (1.39), then we can use the right side of Equation (1.39) to compute the new inverse, and the right side of Equation (1.39) requires a $p \times p$ inversion instead of an $n \times n$ inversion (since we already know the inverse of the old A matrix).

■ EXAMPLE 1.1

At your investment firm, you notice that in January the New York Stock Exchange index decreased by 2%, the American Stock Exchange index increased by 1%, and the NASDAQ stock exchange index increased by 2%. As a result, investors increased their deposits by 1%. The next month, the stock exchange indices changed by −4%, 3%, and 2%, respectively, and investor deposits increased by 2%. The following month, the stock exchange indices changed by −5%, 1%, and 5%, respectively, and investor deposits increased by 2%. You suspect that investment changes y can be modeled as $y = g_1x_1 + g_2x_2 + g_3x_3$, where the x_i variables are the stock exchange index changes, and the g_i are unknown constants. In order to determine the g_i constants you need to invert the matrix

$$A = \begin{bmatrix} -2 & 1 & 2 \\ -4 & 3 & 2 \\ -5 & 1 & 5 \end{bmatrix} \tag{1.40}$$

The result is

$$\begin{aligned} A^{-1} &= \frac{1}{6}\begin{bmatrix} 13 & -3 & -4 \\ 10 & 0 & -4 \\ 11 & -3 & -2 \end{bmatrix} \\ g &= A^{-1}\begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \\ &= \frac{1}{6}\begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix} \end{aligned} \tag{1.41}$$

This allows you to use stock exchange index changes to predict investment changes in the following month, which allows you to better schedule personnel and computer resources. However, soon afterward you find out that the NASDAQ change in the third month was actually 6% rather than 5%. This means that in order to find the g_i constants you need to invert the matrix

$$A' = \begin{bmatrix} -2 & 1 & 2 \\ -4 & 3 & 2 \\ -5 & 1 & 6 \end{bmatrix} \tag{1.42}$$

You are tired of inverting matrices and so you wonder if you can somehow use the inverse of A (which you have already calculated) to find the inverse of A'. Remembering the matrix inversion lemma, you realize that $A' = A + BD^{-1}C$, where

$$\begin{aligned} B &= \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^T \\ C &= \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \\ D &= 1 \end{aligned} \tag{1.43}$$

You therefore use the matrix inversion lemma to compute

$$\begin{aligned} (A')^{-1} &= (A + BD^{-1}C)^{-1} \\ &= A^{-1} - A^{-1}B(D + CA^{-1}B)^{-1}CA^{-1} \end{aligned} \tag{1.44}$$

The $(D + CA^{-1}B)$ term that needs to be inverted in the above equation is a scalar, so its inversion is simple. This gives

$$\begin{aligned} (A')^{-1} &= \begin{bmatrix} 4.00 & 1.00 & -1.00 \\ 3.50 & -0.50 & -1.00 \\ 2.75 & -0.75 & -0.50 \end{bmatrix} \\ g &= (A')^{-1} \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ 0.5 \\ 0.25 \end{bmatrix} \end{aligned} \tag{1.45}$$

In this example, the use of the matrix inversion lemma is not really necessary because A' (the new matrix to invert) is only 3×3. However, with larger matrices, such as 1000×1000 matrices, the computational savings that is realized by using the matrix inversion lemma could be significant.

$\nabla\nabla\nabla$

Now suppose that A, B, C, and D are matrices, with A and D being square. Then it can be seen that

$$\begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} I & -A^{-1}B \\ 0 & I \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & D - CA^{-1}B \end{bmatrix} \tag{1.46}$$

This means that

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |A||D - CA^{-1}B| \tag{1.47}$$

Similarly, it can be shown that

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |D||A - BD^{-1}C| \tag{1.48}$$

These formulas are called product rules for determinants. They were first given by the Russian-born mathematician Issai Schur in a German paper [Sch17] that was reprinted in English in [Sch86].

1.1.3 Matrix calculus

In our first calculus course, we learned the mathematics of derivatives and integrals and how to apply those concepts to scalars. We can also apply the mathematics of calculus to vectors and matrices. Some aspects of matrix calculus are identical to scalar calculus, but some scalar calculus concepts need to be extended in order to derive formulas for matrix calculus.

As intuition would lead us to believe, the time derivative of a matrix is simply equal to the matrix of the time derivatives of the individual matrix elements. Also, the integral of a matrix is equal to the matrix of the integrals of the individual matrix elements. In other words, assuming that A is an $m \times n$ matrix, we have

$$\begin{aligned} \dot{A}(t) &= \begin{bmatrix} \dot{A}_{11}(t) & \cdots & \dot{A}_{1n}(t) \\ \vdots & \ddots & \vdots \\ \dot{A}_{n1}(t) & \cdots & \dot{A}_{nn}(t) \end{bmatrix} \\ \int A(t)\,dt &= \begin{bmatrix} \int A_{11}(t)\,dt & \cdots & \int A_{1n}(t)\,dt \\ \vdots & \ddots & \vdots \\ \int A_{n1}(t)\,dt & \cdots & \int A_{nn}(t)\,dt \end{bmatrix} \end{aligned} \tag{1.49}$$

Next we will compute the time derivative of the inverse of a matrix. Suppose that matrix $A(t)$, which we will denote as A, has elements that are functions of time. We know that $AA^{-1} = I$; that is, AA^{-1} is a constant matrix and therefore has a time derivative of zero. But the time derivative of AA^{-1} can be computed as

$$\frac{d}{dt}(AA^{-1}) = \dot{A}A^{-1} + A\frac{d}{dt}(A^{-1}) \tag{1.50}$$

Since this is zero, we can solve for $d(A^{-1})/dt$ as

$$\frac{d}{dt}(A^{-1}) = -A^{-1}\dot{A}A^{-1} \tag{1.51}$$

Note that for the special case of a scalar A, this reduces to the familiar equation

$$\begin{aligned} \frac{d}{dt}(1/A) &= \frac{\partial(1/A)}{\partial A}\frac{dA}{dt} \\ &= -\dot{A}/A^2 \end{aligned} \tag{1.52}$$

Now suppose that x is an $n \times 1$ vector and $f(x)$ is a scalar function of the elements of x. Then

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \partial f/\partial x_1 & \cdots & \partial f/\partial x_n \end{bmatrix} \tag{1.53}$$

Even though x is a column vector, $\partial f/\partial x$ is a row vector. The converse is also true – if x is a row vector, then $\partial f/\partial x$ is a column vector. Note that some authors define this the other way around. That is, they say that if x is a column vector then $\partial f/\partial x$ is also a column vector. There is no accepted convention for the definition of the partial derivative of a scalar with respect to a vector. It does not really matter which definition we use as long as we are consistent. In this book, we will use the convention described by Equation (1.53).

Now suppose that A is an $m \times n$ matrix and $f(A)$ is a scalar. Then the partial derivative of a scalar with respect to a matrix can be computed as follows:

$$\frac{\partial f}{\partial A} = \begin{bmatrix} \partial f/\partial A_{11} & \cdots & \partial f/\partial A_{1n} \\ \vdots & \ddots & \vdots \\ \partial f/\partial A_{m1} & \cdots & \partial f/\partial A_{mn} \end{bmatrix} \tag{1.54}$$

With these definitions we can compute the partial derivative of the dot product of two vectors. Suppose x and y are n-element column vectors. Then

$$\begin{aligned} x^T y &= x_1 y_1 + \cdots + x_n y_n \\ \frac{\partial (x^T y)}{\partial x} &= \begin{bmatrix} \partial (x^T y)/\partial x_1 & \cdots & \partial (x^T y)/\partial x_n \end{bmatrix} \\ &= \begin{bmatrix} y_1 & \cdots & y_n \end{bmatrix} \\ &= y^T \end{aligned} \tag{1.55}$$

Likewise, we can obtain

$$\frac{\partial (x^T y)}{\partial y} = x^T \tag{1.56}$$

Now we will compute the partial derivative of a quadratic with respect to a vector. First write the quadratic as follows:

$$\begin{aligned} x^T A x &= \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{n1} & \cdots & A_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \\ &= \begin{bmatrix} \sum_i x_i A_{i1} & \cdots & \sum_i x_i A_{in} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \\ &= \sum_{i,j} x_i x_j A_{ij} \end{aligned} \tag{1.57}$$

Now take the partial derivative of the quadratic as follows:

$$\begin{aligned} \frac{\partial (x^T A x)}{\partial x} &= \begin{bmatrix} \partial (x^T A x)/\partial x_1 & \cdots & \partial (x^T A x)/\partial x_n \end{bmatrix} \\ &= \begin{bmatrix} \sum_j x_j A_{1j} + \sum_i x_i A_{i1} & \cdots & \sum_j x_j A_{nj} + \sum_i x_i A_{in} \end{bmatrix} \\ &= \begin{bmatrix} \sum_j x_j A_{1j} & \cdots & \sum_j x_j A_{nj} \end{bmatrix} + \begin{bmatrix} \sum_i x_i A_{i1} & \cdots & \sum_i x_i A_{in} \end{bmatrix} \\ &= x^T A^T + x^T A \end{aligned} \tag{1.58}$$

If A is symmetric, as it often is in quadratic expressions, then $A = A^T$ and the above expression simplifies to

$$\frac{\partial(x^T A x)}{\partial x} = 2x^T A \qquad \text{if } A = A^T \tag{1.59}$$

Next we define the partial derivative of a vector with respect to another vector. Suppose $g(x) = \begin{bmatrix} g_1(x) \\ \vdots \\ g_m(x) \end{bmatrix}$ and $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$. Then

$$\frac{\partial g}{\partial x} = \begin{bmatrix} \partial g_1/\partial x_1 & \cdots & \partial g_1/\partial x_n \\ \vdots & & \vdots \\ \partial g_m/\partial x_1 & \cdots & \partial g_m \partial x_n \end{bmatrix} \tag{1.60}$$

If either $g(x)$ or x is transposed, then the partial derivative is also transposed.

$$\begin{aligned} \frac{\partial g^T}{\partial x} &= \left(\frac{\partial g}{\partial x}\right)^T \\ \frac{\partial g}{\partial x^T} &= \left(\frac{\partial g}{\partial x}\right)^T \\ \frac{\partial g^T}{\partial x^T} &= \frac{\partial g}{\partial x} \end{aligned} \tag{1.61}$$

With these definitions, the following important equalities can be derived. Suppose A is an $m \times n$ matrix and x is an $n \times 1$ vector. Then

$$\begin{aligned} \frac{\partial(Ax)}{\partial x} &= A \\ \frac{\partial(x^T A)}{\partial x} &= A \end{aligned} \tag{1.62}$$

Now we suppose that A is an $m \times n$ matrix, B is an $n \times n$ matrix, and we want to compute the partial derivative of $\mathrm{Tr}(ABA^T)$ with respect to A. First compute ABA^T as follows:

$$\begin{aligned} ABA^T &= \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & & \vdots \\ A_{m1} & \cdots & A_{mn} \end{bmatrix} \begin{bmatrix} B_{11} & \cdots & B_{1n} \\ \vdots & \ddots & \vdots \\ B_{n1} & \cdots & B_{nn} \end{bmatrix} \begin{bmatrix} A_{11} & \cdots & A_{m1} \\ \vdots & \ddots & \vdots \\ A_{1n} & \cdots & A_{nn} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{j,k} A_{1k} B_{kj} A_{1j} & \cdots & \sum_{j,k} A_{1k} B_{kj} A_{mj} \\ \vdots & & \vdots \\ \sum_{j,k} A_{mk} B_{kj} A_{1j} & \cdots & \sum_{j,k} A_{mk} B_{kj} A_{mj} \end{bmatrix} \end{aligned} \tag{1.63}$$

From this we see that the trace of ABA^T is given as

$$\mathrm{Tr}(ABA^T) = \sum_{i,j,k} A_{ik} B_{kj} A_{ij} \tag{1.64}$$

Its partial derivative with respect to A can be computed as

$$\begin{aligned}
\frac{\partial \mathrm{Tr}(ABA^T)}{\partial A} &= \begin{bmatrix} \partial \mathrm{Tr}(ABA^T)/\partial A_{11} & \cdots & \partial \mathrm{Tr}(ABA^T)/\partial A_{1n} \\ \vdots & & \vdots \\ \partial \mathrm{Tr}(ABA^T)/\partial A_{m1} & \cdots & \partial \mathrm{Tr}(ABA^T)/\partial A_{mn} \end{bmatrix} \\
&= \begin{bmatrix} \sum_j A_{1j}B_{1j} + \sum_k A_{1k}B_{k1} & \cdots & \sum_j A_{1j}B_{nj} + \sum_k A_{1k}B_{kn} \\ \vdots & \ddots & \vdots \\ \sum_j A_{mj}B_{1j} + \sum_k A_{mk}B_{k1} & \cdots & \sum_j A_{mj}B_{nj} + \sum_k A_{mk}B_{kn} \end{bmatrix} \\
&= \begin{bmatrix} \sum_j A_{1j}B_{1j} & \cdots & \sum_j A_{1j}B_{nj} \\ \vdots & & \vdots \\ \sum_j A_{mj}B_{1j} & \cdots & \sum_j A_{mj}B_{nj} \end{bmatrix} + \\
&\quad \begin{bmatrix} \sum_k A_{1k}B_{k1} & \cdots & \sum_k A_{1k}B_{kn} \\ \vdots & & \vdots \\ \sum_k A_{mk}B_{k1} & \cdots & \sum_k A_{mk}B_{kn} \end{bmatrix} \\
&= AB^T + AB
\end{aligned} \tag{1.65}$$

If B is symmetric, as it often is in partial derivatives of the form above, then this can be simplified to

$$\frac{\partial \mathrm{Tr}(ABA^T)}{\partial A} = 2AB \qquad \text{if } B = B^T \tag{1.66}$$

A number of additional interesting results related to matrix calculus can be found in [Ske98, Appendix B].

1.1.4 The history of matrices

This section is a brief diversion to present some of the history of matrix theory. Much of the information in this section is taken from [OCo96].

The use of matrices can be found as far back as the fourth century BC. We see in ancient clay tablets that the Babylonians studied problems that led to simultaneous linear equations. For example, a tablet dating from about 300 BC contains the following problem: "There are two fields whose total area is 1800 units. One produces grain at the rate of 2/3 of a bushel per unit while the other produces grain at the rate of 1/2 a bushel per unit. If the total yield is 1100 bushels, what is the size of each field?"

Later, the Chinese came even closer to the use of matrices. In [She99] (originally published between 200 BC and 100 AD) we see the following problem: "There are three types of corn, of which three bundles of the first, two of the second, and one of the third make 39 measures. Two of the first, three of the second, and one of the third make 34 measures. And one of the first, two of the second and three of the third make 26 measures. How many measures of corn are contained in one bundle of each type?" At that point, the ancient Chinese essentially use Gaussian elimination (which was not well known until the 19th century) to solve the problem.

In spite of this very early beginning, it was not until the end of the 17th century that serious investigation of matrix algebra began. In 1683, the Japanese

mathematician Takakazu Seki Kowa wrote a book called "Method of Solving the Dissimulated Problems." This book gives general methods for calculating determinants and presents examples for matrices as large as 5×5. Coincidentally, in the same year (1683) Gottfried Leibniz in Europe also first used determinants to solve systems of linear equations. Leibniz also discovered that a determinant could be expanded using any of the matrix columns.

In the middle of the 1700s, Colin Maclaurin and Gabriel Cramer published some major contributions to matrix theory. After that point, work on matrices became rather regular, with significant contributions by Etienne Bezout, Alexandre Vandermonde, Pierre Laplace, Joseph Lagrange, and Carl Gauss. The term "determinant" was first used in the modern sense by Augustin Cauchy in 1812 (although the word was used earlier by Gauss in a different sense). Cauchy also discovered matrix eigenvalues and diagonalization, and introduced the idea of similar matrices. He was the first to prove that every real symmetric matrix is diagonalizable.

James Sylvester (in 1850) was the first to use the term "matrix." Sylvester moved to England in 1851 to became a lawyer and met Arthur Cayley, a fellow lawyer who was also interested in mathematics. Cayley saw the importance of the idea of matrices and in 1853 he invented matrix inversion. Cayley also proved that 2×2 and 3×3 matrices satisfy their own characteristic equations. The fact that a matrix satisfies its own characteristic equation is now called the Cayley–Hamilton theorem (see Problem 1.5). The theorem has William Hamilton's name associated with it because he proved the theorem for 4×4 matrices during the course of his work on quaternions.

Camille Jordan invented the Jordan canonical form of a matrix in 1870. Georg Frobenius proved in 1878 that all matrices satisfy their own characteristic equation (the Cayley Hamilton theorem). He also introduced the definition of the rank of a matrix. The nullity of a square matrix was defined by Sylvester in 1884. Karl Weierstrass's and Leopold Kronecker's publications in 1903 were instrumental in establishing matrix theory as an important branch of mathematics. Leon Mirsky's book in 1955 [Mir90] helped solidify matrix theory as a fundamentally important topic in university mathematics.

1.2 LINEAR SYSTEMS

Many processes in our world can be described by state-space systems. These include processes in engineering, economics, physics, chemistry, biology, and many other areas. If we can derive a mathematical model for a process, then we can use the tools of mathematics to control the process and obtain information about the process. This is why state-space systems are so important to engineers. If we know the state of a system at the present time, and we know all of the present and future inputs, then we can deduce the values of all future outputs of the system.

State-space models can be generally divided into linear models and nonlinear models. Although most real processes are nonlinear, the mathematical tools that are available for estimation and control are much more accessible and well understood for linear systems. That is why nonlinear systems are often approximated as linear systems. That way we can use the tools that have been developed for linear systems to derive estimation or control algorithms.

A continuous-time, deterministic linear system can be described by the equations

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx \end{aligned} \tag{1.67}$$

where x is the state vector, u is the control vector, and y is the output vector. Matrices A, B, and C are appropriately dimensioned matrices. The A matrix is often called the system matrix, B is often called the input matrix, and C is often called the output matrix. In general, A, B, and C can be time-varying matrices and the system will still be linear. If A, B, and C are constant then the solution to Equation (1.67) is given by

$$\begin{aligned} x(t) &= e^{A(t-t_0)}x(t_0) + \int_{t_0}^{t} e^{A(t-\tau)}Bu(\tau)\,d\tau \\ y(t) &= Cx(t) \end{aligned} \tag{1.68}$$

where t_0 is the initial time of the system and is often taken to be 0. This is easy to verify when all of the quantities in Equation (1.67) are scalar, but it happens to be true in the vector case also. Note that in the zero input case, $x(t)$ is given as

$$x(t) = e^{A(t-t_0)}x(t_0), \qquad \text{zero input case} \tag{1.69}$$

For this reason, e^{At} is called the state-transition matrix of the system.[3] It is the matrix that describes how the state changes from its initial condition in the absence of external inputs. We can evaluate the above equation at $t = t_0$ to see that

$$e^{A0} = I \tag{1.70}$$

in analogy with the scalar exponential of zero.

As stated above, even if x is an n-element vector, then Equation (1.68) still describes the solution of Equation (1.67). However, a fundamental question arises in this case: How can we take the exponential of the matrix A in Equation (1.68)? What does it mean to raise the scalar e to the power of a matrix? There are many different ways to compute this quantity [Mol03]. Three of the most useful are the following:

$$\begin{aligned} e^{At} &= \sum_{j=0}^{\infty} \frac{(At)^j}{j!} \\ &= \mathcal{L}^{-1}[(sI - A)^{-1}] \\ &= Qe^{\hat{A}t}Q^{-1} \end{aligned} \tag{1.71}$$

The first expression above is the definition of e^{At}, and is analogous to the definition of the exponential of a scalar. This definition shows that A must be square in order for e^{At} to exist. From Equation (1.67), we see that a system matrix is always square. The definition of e^{At} can also be used to derive the following properties.

$$\begin{aligned} \frac{d}{dt}e^{At} &= Ae^{At} \\ &= e^{At}A \end{aligned} \tag{1.72}$$

[3]The MATLAB function EXPM computes the matrix exponential. Note that the MATLAB function EXP computes the element-by-element exponential of a matrix, which is generally not the same as the matrix exponential.

In general, matrices do not commute under multiplication but, interestingly, a matrix always commutes with its exponential.

The first expression in Equation (1.71) is not usually practical for computational purposes since it is an infinite sum (although the latter terms in the sum often decrease rapidly in magnitude, and may even become zero). The second expression in Equation (1.71) uses the inverse Laplace transform to compute e^{At}. In the third expression of Equation (1.71), Q is a matrix whose columns comprise the eigenvectors of A, and $\hat{A}$ is the Jordan form[4] of A. Note that Q and $\hat{A}$ are well defined for any square matrix A, so the matrix exponential e^{At} exists for all square matrices A and all finite t. The matrix $\hat{A}$ is often diagonal, in which case $e^{\hat{A}t}$ is easy to compute:

$$\hat{A} = \begin{bmatrix} \hat{A}_{11} & 0 & \cdots & 0 \\ 0 & \hat{A}_{22} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \hat{A}_{nn} \end{bmatrix}$$

$$e^{\hat{A}t} = \begin{bmatrix} e^{\hat{A}_{11}} & 0 & \cdots & 0 \\ 0 & e^{\hat{A}_{22}} & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & e^{\hat{A}_{nn}} \end{bmatrix} \tag{1.73}$$

This can be computed from the definition of $e^{\hat{A}t}$ in Equation (1.71). Even if the Jordan form matrix $\hat{A}$ is not diagonal, $e^{\hat{A}t}$ is easy to compute [Bay99, Che99, Kai80]. We can also note from the third expression in Equation (1.71) that

$$\begin{aligned} \left[e^{At}\right]^{-1} &= e^{-At} \\ &= Qe^{-\hat{A}t}Q^{-1} \end{aligned} \tag{1.74}$$

(Recall that A and $-A$ have the same eigenvectors, and their eigenvalues are negatives of each other. See Problem 1.10.) We see from this that e^{At} is always invertible. This is analogous to the scalar situation in which the exponential of a scalar is always nonzero.

Another interesting fact about the matrix exponential is that all of the individual elements of the matrix exponential e^{A} are nonnegative if and only if all of the individual elements of A are nonnegative [Bel60, Bel80].

■ EXAMPLE 1.2

As an example of a linear system, suppose that we are controlling the angular acceleration of a motor (for example, with some applied voltage across the motor windings). The derivative of the position is the velocity. A simplified motor model can then be written as

[4]In fact, Equation (1.71) can be used to define the Jordan form of a matrix. That is, if e^{At} can be written as shown in Equation (1.71), where Q is a matrix whose columns comprise the eigenvectors of A, then $\hat{A}$ is the Jordan form of A. More discussion about Jordan forms and their computation can be found in most linear systems books [Kai80, Bay99, Che99].

$$\begin{aligned} \dot{\theta} &= \omega \\ \dot{\omega} &= u + w_1 \end{aligned} \tag{1.75}$$

The scalar w_1 is the acceleration noise and could consist of such factors as uncertainty in the applied acceleration, motor shaft eccentricity, and load disturbances. If our measurement consists of the angular position of the motor then a state space description of this system can be written as

$$\begin{aligned} \begin{bmatrix} \dot{\theta} \\ \dot{\omega} \end{bmatrix} &= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \theta \\ \omega \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u + \begin{bmatrix} 0 \\ w_1 \end{bmatrix} \\ y &= \begin{bmatrix} 1 & 0 \end{bmatrix} x + v \end{aligned} \tag{1.76}$$

The scalar v consists of measurement noise. Comparing with Equation (1.67), we see that the state vector x is a 2×1 vector containing the scalars θ and ω.

$\nabla\nabla\nabla$

■ **EXAMPLE 1.3**

In this example, we will use the three expressions in Equation (1.71) to compute the state-transition matrix of the system described in Example 1.2. From the first expression in Equation (1.71) we obtain

$$\begin{aligned} e^{At} &= \sum_{j=0}^{\infty} \frac{(At)^j}{j!} \\ &= (At)^0 + (At)^1 + \frac{(At)^2}{2!} + \frac{(At)^3}{3!} + \cdots \\ &= I + At \end{aligned} \tag{1.77}$$

where the last equality comes from the fact that $A^k = 0$ when $k > 1$ for the A matrix given in Example 1.2. We therefore obtain

$$\begin{aligned} e^{At} &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & t \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \end{aligned} \tag{1.78}$$

From the second expression in Equation (1.71) we obtain

$$\begin{aligned} e^{At} &= \mathcal{L}^{-1}[(sI - A)^{-1}] \\ &= \mathcal{L}^{-1}\left(\begin{bmatrix} s & -1 \\ 0 & s \end{bmatrix}^{-1} \right) \\ &= \mathcal{L}^{-1} \begin{bmatrix} 1/s & 1/s^2 \\ 0 & 1/s \end{bmatrix} \\ &= \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \end{aligned} \tag{1.79}$$

In order to use the third expression in Equation (1.71) we first need to obtain the eigendata (i.e., the eigenvalues and eigenvectors) of the A matrix. These are found as

$$\begin{aligned} \lambda(A) &= \{0, 0\} \\ v(A) &= \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\} \end{aligned} \tag{1.80}$$

This shows that

$$\begin{aligned} \hat{A} &= \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \\ Q &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{aligned} \tag{1.81}$$

Note that in this simple example A is already in Jordan form, so $\hat{A} = A$ and $Q = I$. The third expression in Equation (1.71) therefore gives

$$\begin{aligned} e^{At} &= Q e^{\hat{A}t} Q^{-1} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}^{-1} \\ &= \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \end{aligned} \tag{1.82}$$

$\nabla\nabla\nabla$

1.3 NONLINEAR SYSTEMS

The discussion of linear systems in the preceding section is a bit optimistic, because in reality linear systems do not exist. Real systems always have some nonlinearities. Even a simple resistor is ultimately nonlinear if we apply a large enough voltage across it. However, we often model a resistor with the simple linear equation $V = IR$ because this equation accurately describes the operation of the resistor over a wide operating range. So even though linear systems do not exist in the real world, linear systems theory is still a valuable tool for dealing with nonlinear systems.

The general form of a continuous-time nonlinear system can be written as

$$\begin{aligned} \dot{x} &= f(x, u, w) \\ y &= h(x, v) \end{aligned} \tag{1.83}$$

where $f(\cdot)$ and $h(\cdot)$ are arbitrary vector-valued functions. We use w to indicate process noise, and v to indicate measurement noise. If $f(\cdot)$ and $h(\cdot)$ are explicit functions of t then the system is time-varying. Otherwise, the system is time-invariant. If $f(x, u, w) = Ax + Bu + w$, and $h(x, v) = Hx + v$, then the system is linear [compare with Equation (1.67)]. Otherwise, the system is nonlinear.

In order to apply tools from linear systems theory to nonlinear systems, we need to linearize the nonlinear system. In other words, we need to find a linear system

that is approximately equal to the nonlinear system. To see how this is done, let us start with a nonlinear vector function $f(\cdot)$ of a scalar x. We expand $f(x)$ in a Taylor series around some nominal operating point (also called a linearization point) $x = \bar{x}$, defining $\tilde{x} = x - \bar{x}$:

$$f(x) = f(\bar{x}) + \left.\frac{\partial f}{\partial x}\right|_{\bar{x}} \tilde{x} + \frac{1}{2!}\left.\frac{\partial^2 f}{\partial x^2}\right|_{\bar{x}} \tilde{x}^2 + \frac{1}{3!}\left.\frac{\partial^3 f}{\partial x^3}\right|_{\bar{x}} \tilde{x}^3 + \cdots \tag{1.84}$$

Now suppose that x is a 2×1 vector. This implies that $f(x)$ is a nonlinear function of two independent variables x_1 and x_2. The Taylor series expansion of $f(x)$ becomes

$$\begin{aligned} f(x) \;=\; & f(\bar{x}) + \left.\frac{\partial f}{\partial x_1}\right|_{\bar{x}} \tilde{x}_1 + \left.\frac{\partial f}{\partial x_2}\right|_{\bar{x}} \tilde{x}_2 + \\ & \frac{1}{2!}\left(\left.\frac{\partial^2 f}{\partial x_1^2}\right|_{\bar{x}} \tilde{x}_1^2 + \left.\frac{\partial^2 f}{\partial x_2^2}\right|_{\bar{x}} \tilde{x}_2^2 + 2\left.\frac{\partial^2 f}{\partial x_1 x_2}\right|_{\bar{x}} \tilde{x}_1\tilde{x}_2\right) + \\ & \frac{1}{3!}\left(\left.\frac{\partial^3 f}{\partial x_1^3}\right|_{\bar{x}} \tilde{x}_1^3 + \left.\frac{\partial^3 f}{\partial x_2^3}\right|_{\bar{x}} \tilde{x}_2^3 + 3\left.\frac{\partial^3 f}{\partial x_1^2 x_2}\right|_{\bar{x}} \tilde{x}_1^2\tilde{x}_2 + 3\left.\frac{\partial^3 f}{\partial x_1 x_2^2}\right|_{\bar{x}} \tilde{x}_1\tilde{x}_2^2\right) + \cdots \end{aligned} \tag{1.85}$$

This can be written more compactly as

$$\begin{aligned} f(x) \;=\; & f(\bar{x}) + \left(\tilde{x}_1\frac{\partial}{\partial x_1} + \tilde{x}_2\frac{\partial}{\partial x_2}\right) f\Big|_{\bar{x}} + \frac{1}{2!}\left(\tilde{x}_1\frac{\partial}{\partial x_1} + \tilde{x}_2\frac{\partial}{\partial x_2}\right)^2 f\Big|_{\bar{x}} + \\ & \frac{1}{3!}\left(\tilde{x}_1\frac{\partial}{\partial x_1} + \tilde{x}_2\frac{\partial}{\partial x_2}\right)^3 f\Big|_{\bar{x}} + \cdots \end{aligned} \tag{1.86}$$

Extending this to the general case in which x is an $n \times 1$ vector, we see that any continuous vector-valued function $f(x)$ can be expanded in a Taylor series as

$$\begin{aligned} f(x) \;=\; & f(\bar{x}) + \left(\tilde{x}_1\frac{\partial}{\partial x_1} + \cdots + \tilde{x}_2\frac{\partial}{\partial x_n}\right) f\Big|_{\bar{x}} + \\ & \frac{1}{2!}\left(\tilde{x}_1\frac{\partial}{\partial x_1} + \cdots + \tilde{x}_n\frac{\partial}{\partial x_n}\right)^2 f\Big|_{\bar{x}} + \\ & \frac{1}{3!}\left(\tilde{x}_1\frac{\partial}{\partial x_1} + \cdots + \tilde{x}_n\frac{\partial}{\partial x_n}\right)^3 f\Big|_{\bar{x}} + \cdots \end{aligned} \tag{1.87}$$

Now we define the operation $D_{\tilde{x}}^k f$ as

$$D_{\tilde{x}}^k f = \left(\sum_{i=1}^{n} \tilde{x}_i \frac{\partial}{\partial x_i}\right)^k f(x)\Big|_{\bar{x}} \tag{1.88}$$

Using this definition we write the Taylor series expansion of $f(x)$ as

$$f(x) = f(\bar{x}) + D_{\tilde{x}} f + \frac{1}{2!} D_{\tilde{x}}^2 f + \frac{1}{3!} D_{\tilde{x}}^3 f + \cdots \tag{1.89}$$

If the nonlinear function $f(x)$ is "sufficiently smooth," then high-order derivatives of $f(x)$ should be "somewhat small." Also, if $f(x)$ is expanded around a point such

that x is "close" to $\bar{x}$, then $\tilde{x}$ will be "small" and the higher powers of $\tilde{x}$ in Equation (1.89) will be "small." Finally, the higher-order derivatives in the Taylor series expansion of Equation (1.89) are divided by increasingly large factorials, which further diminishes the magnitude of the higher-order terms in Equation (1.89). This justifies the approximation

$$\begin{aligned} f(x) &\approx f(\bar{x}) + D_{\tilde{x}} f \\ &\approx f(\bar{x}) + \left.\frac{\partial f}{\partial x}\right|_{\bar{x}} \tilde{x} \\ &\approx f(\bar{x}) + A\tilde{x} \end{aligned} \tag{1.90}$$

where A is the matrix defined by the above equation.

Returning to our nonlinear system equations in Equation (1.83), we can expand the nonlinear system equation $f(x, u, w)$ around the nominal operating point $(\bar{x}, \bar{u}, \bar{w})$. We then obtain a linear system approximation as follows.

$$\begin{aligned} \dot{x} &= f(x, u, w) \\ &\approx f(\bar{x}, \bar{u}, \bar{w}) + \left.\frac{\partial f}{\partial x}\right|_0 (x - \bar{x}) + \left.\frac{\partial f}{\partial u}\right|_0 (u - \bar{u}) + \left.\frac{\partial f}{\partial w}\right|_0 (w - \bar{w}) \\ &= \dot{\bar{x}} + A\tilde{x} + B\tilde{u} + L\tilde{w} \end{aligned} \tag{1.91}$$

where the 0 subscript means that the function is evaluated at the nominal point $(\bar{x}, \bar{u}, \bar{w})$, and A, B, and L are defined by the above equations. Subtracting $\dot{\bar{x}}$ from both sides of Equation (1.91) gives

$$\dot{\tilde{x}} = A\tilde{x} + B\tilde{u} + L\tilde{w} \tag{1.92}$$

Since w is noise, we will set $\bar{w} = 0$ so that $\tilde{w} = w$ and we obtain

$$\dot{\tilde{x}} = A\tilde{x} + B\tilde{u} + Lw \tag{1.93}$$

We see that we have a linear equation for $\dot{\tilde{x}}$ in terms of $\tilde{x}$, $\tilde{u}$, and w. We have a linear equation for the deviations of the state and control from their nominal values. As long as the deviations remain small, the linearization will be accurate and the linear equation will accurately describe deviations of x from its nominal value $\bar{x}$.

In a similar manner we can expand the nonlinear measurement equation given by Equation (1.83) around a nominal operating point $x = \bar{x}$ and $v = \bar{v} = 0$. This results in the linearized measurement equation

$$\begin{aligned} \tilde{y} &= \left.\frac{\partial h}{\partial x}\right|_0 \tilde{x} + \left.\frac{\partial h}{\partial v}\right|_0 \tilde{v} \\ &= C\tilde{x} + Dv \end{aligned} \tag{1.94}$$

where C and D are defined by the above equation. Equations (1.93) and (1.94) comprise a linear system that describes the deviations of the state and output from their nominal values. Recall that the tilde quantities in Equations (1.93) and (1.94) are defined as

$$\begin{aligned} \tilde{x} &= x - \bar{x} \\ \tilde{u} &= u - \bar{u} \\ \tilde{y} &= y - \bar{y} \end{aligned} \tag{1.95}$$

EXAMPLE 1.4

Consider the following model for a two-phase permanent magnet synchronous motor:

$$\begin{aligned}
\dot{i_a} &= \frac{-R}{L} i_a + \frac{\omega\lambda}{L}\sin\theta + \frac{u_a}{L} \\
\dot{i_b} &= \frac{-R}{L} i_b - \frac{\omega\lambda}{L}\cos\theta + \frac{u_b}{L} \\
\dot{\omega} &= \frac{-3\lambda}{2J} i_a \sin\theta + \frac{3\lambda}{2J} i_b \cos\theta - \frac{F\omega}{J} - \frac{T_l}{J} \\
\dot{\theta} &= \omega
\end{aligned} \tag{1.96}$$

where i_a and i_b are the currents through the two windings, R and L are the resistance and inductance of the windings, θ and ω are the angular position and velocity of the rotor, λ is the flux constant of the motor, u_a and u_b are the voltages applied across the two windings, J is the moment of inertia of the rotor and its load, F is the viscous friction of the rotor, and T_l is the load torque. The time variable does not explicitly appear on the right side of the above equation, so this is a time-invariant system. However, the system is highly nonlinear and we therefore cannot directly use any linear systems tools for control or estimation. However, if we linearize the system around a nominal (possibly time-varying) operating point then we can use linear system tools for control and estimation. We start by defining a state vector as $x = \begin{bmatrix} i_a & i_b & \omega & \theta \end{bmatrix}^T$. With this definition we write

$$\begin{aligned}
\dot{x} &= \begin{bmatrix} \dot{x}_1 & \dot{x}_2 & \dot{x}_3 & \dot{x}_4 \end{bmatrix}^T \\
&= f(x,u) \\
&= \begin{bmatrix} \frac{-R}{L}x_1 + \frac{x_3\lambda}{L}\sin x_4 + \frac{u_a}{L} \\ \frac{-R}{L}x_2 - \frac{x_3\lambda}{L}\cos x_4 + \frac{u_b}{L} \\ \frac{-3\lambda}{2J}x_1 \sin x_4 + \frac{3\lambda}{2J}x_2 \cos x_4 - \frac{Fx_3}{J} - \frac{T_l}{J} \\ x_3 \end{bmatrix}
\end{aligned} \tag{1.97}$$

We linearize the system equation by taking the partial derivative of $f(x,u)$ with respect to x and u to obtain

$$\begin{aligned}
A &= \frac{\partial f}{\partial x} \\
&= \begin{bmatrix} -R/L & 0 & \lambda s_4/L & x_3\lambda c_4/L \\ 0 & -R/L & -\lambda c_4/L & x_3\lambda s_4/L \\ -3\lambda s_4/2/J & 3\lambda c_4/2/J & -F/J & -3\lambda(x_1c_4 + x_2s_4)/2/J \\ 0 & 0 & 1 & 0 \end{bmatrix} \\
B &= \frac{\partial f}{\partial u} \\
&= \begin{bmatrix} 1/L & 0 \\ 0 & 1/L \\ 0 & 0 \\ 0 & 0 \end{bmatrix}
\end{aligned} \tag{1.98}$$

where $s_4 = \sin x_4$ and $c_4 = \cos x_4$. The linear system

$$\dot{\tilde{x}} = A\tilde{x} + B\tilde{u} \tag{1.99}$$

approximately describes the deviation of x from its nominal value $\bar{x}$. The nonlinear system was simulated with the nominal control values $\bar{u}_a(t) = \sin 2\pi t$ and $\bar{u}_b(t) = \cos 2\pi t$. This resulted in a nominal state trajectory $\bar{x}(t)$. The linear and nonlinear systems were then simulated with nonnominal control values. Figure 1.1 shows the results of the linear and nonlinear simulations when the control magnitude deviation from nominal is a small positive number. It can be seen that the simulations result in similar state-space trajectories, although they do not match exactly. If the deviation is zero, then the linear and nonlinear simulations will match exactly. As the deviation from nominal increases, the difference between the linear and nonlinear simulations will increase.

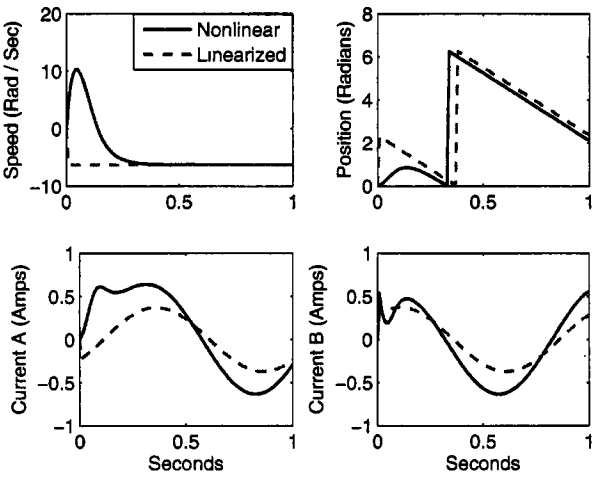

Figure 1.1 Example 1.4 comparison of nonlinear and linearized motor simulations.

▽▽▽

1.4 DISCRETIZATION

Most systems in the real world are described with continuous-time dynamics of the type shown in Equations (1.67) or (1.83). However, state estimation and control algorithms are almost always implemented in digital electronics. This often requires a transformation of continuous-time dynamics to discrete-time dynamics. This section discusses how a continuous-time linear system can be transformed into a discrete-time linear system.

Recall from Equation (1.68) that the solution of a continuous-time linear system is given by

$$x(t) = e^{A(t-t_0)}x(t_0) + \int_{t_0}^{t} e^{A(t-\tau)}Bu(\tau)\,d\tau \tag{1.100}$$

Let $t = t_k$ (some discrete time point) and let the initial time $t_0 = t_{k-1}$ (the previous discrete time point). Assume that $A(\tau)$, $B(\tau)$, and $u(\tau)$ are approximately constant in the interval of integration. We then obtain

$$x(t_k) = e^{A(t_k - t_{k-1})} x(t_{k-1}) + \int_{t_{k-1}}^{t_k} e^{A(t_k - \tau)}\, d\tau\, B u(t_{k-1}) \tag{1.101}$$

Now define $\Delta t = t_k - t_{k-1}$, define $\alpha = \tau - t_{k-1}$, and substitute for τ in the above equation to obtain

$$\begin{aligned} x(t_k) &= e^{A\Delta t} x(t_{k-1}) + \int_0^{\Delta t} e^{A(\Delta t - \alpha)}\, d\alpha\, B u(t_{k-1}) \\ &= e^{A\Delta t} x(t_{k-1}) + e^{A\Delta t} \int_0^{\Delta t} e^{-A\alpha}\, d\alpha\, B u(t_{k-1}) \\ x_k &= F_{k-1} x_{k-1} + G_{k-1} u_{k-1} \end{aligned} \tag{1.102}$$

where x_k, F_k, G_k, and u_k are defined by the above equation. This is a linear discrete-time approximation to the continuous-time dynamics given in Equation (1.67). Note that this discrete-time system defines x_k only at the discrete time points $\{t_k\}$; it does not say anything about what happens to the continuous-time signal $x(t)$ in between the discrete time points.

The difficulty with the above discrete-time system is the computation of the integral of the matrix exponential, which is necessary in order to compute the G matrix. This computation can be simplified if A is invertible:

$$\begin{aligned} \int_0^{\Delta t} e^{-A\tau} d\tau &= \int_0^{\Delta t} \sum_{j=0}^{\infty} \frac{(-A\tau)^j}{j!}\, d\tau \\ &= \int_0^{\Delta t} \left[I - A\tau + A^2\tau^2/2! - \cdots \right] d\tau \\ &= \left[I\tau - A\tau^2/2! + A^2\tau^3/3! - \cdots \right]_0^{\Delta t} \\ &= \left[I\Delta t - A(\Delta t)^2/2! + A^2(\Delta t)^3/3! - \cdots \right] \\ &= \left[A\Delta t - (A\Delta t)^2/2! + (A\Delta t)^3/3! - \cdots \right] A^{-1} \\ &= \left[I - e^{-A\Delta t} \right] A^{-1} \end{aligned} \tag{1.103}$$

The conversion from continuous-time system matrices A and B to discrete-time system matrices F and G can be summarized as follows:

$$\begin{aligned} F &= e^{A\Delta t} \\ G &= F \int_0^{\Delta t} e^{-A\tau}\, d\tau\, B \\ &= F \left[I - e^{-A\Delta t} \right] A^{-1} B \end{aligned} \tag{1.104}$$

where Δt is the discretization step size.

1.5 SIMULATION

In this section, we discuss how to simulate continuous-time systems (either linear or nonlinear) on a digital computer. We consider the following form of the general

system equation from Equation (1.83):

$$\dot{x} = f(x, u, t) \tag{1.105}$$

where $u(t)$ is a known control input. In order to simulate this system on a computer, we need to program a computer to solve for $x(t_f)$ at some user-specified value of t_f. In other words, we want to compute

$$x(t_f) = x(t_0) + \int_{t_0}^{t_f} f[x(t), u(t), t]\, dt \tag{1.106}$$

Often, the initial time is taken as $t_0 = 0$, in which case we have the slightly simpler looking equation

$$x(t_f) = x(0) + \int_{0}^{t_f} f[x(t), u(t), t]\, dt \tag{1.107}$$

We see that in order to find the solution $x(t_f)$ to the differential equation $\dot{x} = f(x, u, t)$, we need to compute an integral. The problem of finding the solution $x(t_f)$ is therefore commonly referred to as an integration problem.

Now suppose that we divide the time interval $[0, t_f]$ into L equally spaced intervals so that $t_k = kT$ for $k = 0, \cdots, L$, and the time interval $T = t_f/L$. From this we note that $t_f = t_L$. With this division of the time interval, we can write the solution of Equation (1.107) as

$$\begin{aligned} x(t_f) &= x(t_L) \\ &= x(0) + \sum_{k=0}^{L} \int_{t_k}^{t_{k+1}} f[x(t), u(t), t]\, dt \end{aligned} \tag{1.108}$$

More generally, for some $n \in [0, L-1]$, we can write $x(t_n)$ and $x(t_{n+1})$ as

$$\begin{aligned} x(t_n) &= x(0) + \sum_{k=0}^{n} \int_{t_k}^{t_{k+1}} f[x(t), u(t), t]\, dt \\ x(t_{n+1}) &= x(0) + \sum_{k=0}^{n+1} \int_{t_k}^{t_{k+1}} f[x(t), u(t), t]\, dt \end{aligned} \tag{1.109}$$

which means that

$$x(t_{n+1}) = x(t_n) + \int_{t_n}^{t_{n+1}} f[x(t), u(t), t]\, dt \tag{1.110}$$

If we can find a way to approximate the integral on the right side of the above equation, we can repeatedly propagate our $x(t)$ approximation from time t_n to time t_{n+1}, thus obtaining an approximation for $x(t)$ at any desired time t. The algorithm could look something like the following.

Differential equation solution

Assume that $x(0)$ is given
for t $= 0 : T : t_f - T$
  Find an approximation $I(t) \approx \int_t^{t+T} f[x(t), u(t), t]\, dt$
  $x(t+T) = x(t) + TI(t)$
end

In the following sections, we present three different ways to approximate this integral. The approximations, in order of increasing computational effort and increasing accuracy, are rectangular integration, trapezoidal integration, and fourth-order Runge–Kutta integration.

1.5.1 Rectangular integration

If the time interval $(t_{n+1} - t_n)$ is small, then $f[x(t), u(t), t]$ is approximately constant in this interval. Equation (1.110) can therefore be approximated as

$$\begin{aligned} x(t_{n+1}) &\approx x(t_n) + \int_{t_n}^{t_{n+1}} f[x(t_n), u(t_n), t_n]\, dt \\ &\approx x(t_n) + f[x(t_n), u(t_n), t_n]T \end{aligned} \tag{1.111}$$

Equation (1.109) can therefore be approximated as

$$\begin{aligned} x(t_n) &\approx x(0) + \sum_{k=0}^{n} \int_{t_k}^{t_{k+1}} f[x(t_k), u(t_k), t_k]\, dt \\ &= x(0) + \sum_{k=0}^{n} f[x(t_k), u(t_k), t_k]T \end{aligned} \tag{1.112}$$

This is called Euler integration, or rectangular integration, and is illustrated in Figure 1.2. As long as T is sufficiently small, this gives a good approximation for $x(t_n)$.

This gives the following algorithm for integrating continuous-time dynamics using rectangular integration. The time loop in the algorithm is executed for $t = 0, T, 2T, \cdots, t_f - T$.

Rectangular integration

Assume that $x(0)$ is given
for t $= 0 : T : t_f - T$
  Compute $f[x(t), u(t), t]$
  $x(t+T) = x(t) + f[x(t), u(t), t]T$
end

1.5.2 Trapezoidal integration

An inspection of Figure 1.2 suggests an idea for improving the approximation for $x(t)$. Instead of approximating each area as a rectangle, what if we approximate each area as a trapezoid? Figure 1.3 shows how an improved integration algorithm can be implemented. This is called modified Euler integration, or trapezoidal integration. A comparison of Figures 1.2 and 1.3 shows that trapezoidal integration

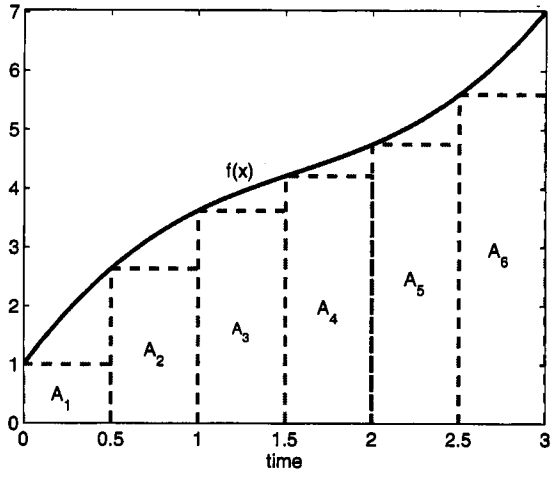

Figure 1.2 An illustration of rectangular integration. We have $\dot{x} = f(x)$, so $x(t)$ is the area under the $f(x)$ curve. This area can be approximated as the sum of the rectangular areas A_i. That is, $x(0.5) \approx A_1$, $x(1) \approx A_1 + A_2$, $\cdots$.

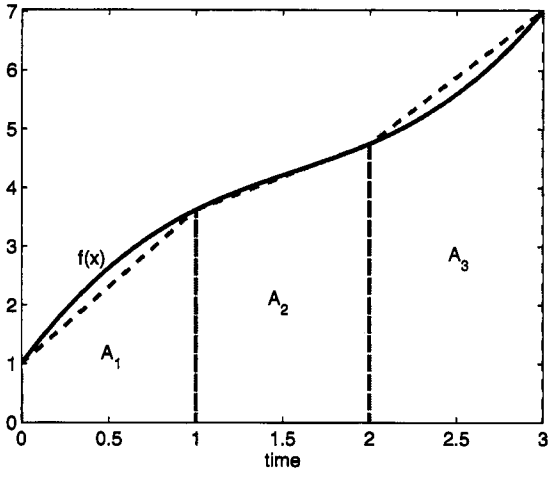

Figure 1.3 An illustration of trapezoidal integration. We have $\dot{x} = f(x)$, so $x(t)$ is the area under the $f(x)$ curve. This area can be approximated as the sum of trapezoidal areas A_i. That is, $x(1) \approx A_1$, $x(2) \approx A_1 + A_2$, and $x(3) \approx A_1 + A_2 + A_3$.

appears to give a better approximation than rectangular integration, even though the time axis is only divided into half as many intervals in trapezoidal integration.

With rectangular integration we approximated $f[x(t), u(t), t]$ as a constant in the interval $t \in [t_n, t_{n+1}]$. With trapezoidal integration, we instead approximate $f[x(t), u(t), t]$ as a linear function in the interval $t \in [t_n, t_{n+1}]$. That is,

$$f[x(t)] \approx f[x(t_n), u(t_n), t_n] + \left(\frac{f[x(t_{n+1}), u(t_{n+1}), t_{n+1}] - f[x(t_n), u(t_n), t_n]}{T}\right)(t - t_n)$$
$$\text{for } t \in [t_n, t_{n+1}] \tag{1.113}$$

Equation (1.110) can therefore be approximated as

$$\begin{aligned} x(t_{n+1}) &\approx x(t_n) + \int_{t_n}^{t_{n+1}} \left\{ f[x(t_n), u(t_n), t_n] + \left(\frac{f[x(t_{n+1}), u(t_{n+1}), t_{n+1}] - f[x(t_n), u(t_n), t_n]}{T}\right)(t - t_n) \right\} dt \\ &= x(t_n) + \left(\frac{f[x(t_n), u(t_n), t_n] + f[x(t_{n+1}), u(t_{n+1}), t_{n+1}]}{2}\right) T \\ &= x(t_n) + \frac{1}{2}\left(f[x(t_n), u(t_n), t_n]T + f[x(t_{n+1}), u(t_{n+1}), t_{n+1}]T\right) \end{aligned} \tag{1.114}$$

This equation to approximate $x(t_{n+1})$, however, has $x(t_{n+1})$ on the right side of the equation. How can we plug $x(t_{n+1})$ into the right side of the equation if we do not yet know $x(t_{n+1})$? The answer is that we can use the rectangular integration approximation from the previous section for $x(t_{n+1})$ on the right side of the equation. The above equation can therefore be written as

$$\begin{aligned} \Delta x_1 &= f[x(t_n), u(t_n), t_n]T \\ \Delta x_2 &= f[x(t_{n+1}), u(t_{n+1}), t_{n+1}]T \\ &\approx f[x(t_n) + \Delta x_1, u(t_{n+1}), t_{n+1}]T \\ x(t_{n+1}) &\approx x(t_n) + \frac{1}{2}(\Delta x_1 + \Delta x_2) \end{aligned} \tag{1.115}$$

This gives the following algorithm for integrating continuous-time dynamics using trapezoidal integration. The time loop in the algorithm is executed for $t = 0, T, 2T, \cdots, t_f - T$.

Trapezoidal integration

Assume that $x(0)$ is given
for t = $0 : T : t_f - T$
  $\Delta x_1 = f[x(t), u(t), t]T$
  $\Delta x_2 = f[x(t) + \Delta x_1, u(t + T), t + T]T$
  $x(t + T) = x(t) + (\Delta x_1 + \Delta x_2)/2$
end

1.5.3 Runge–Kutta integration

From the previous sections, we see that rectangular integration involves the calculation of one function value at each time step, and trapezoidal integration involves the calculation of two function values at each time step. In order to further improve the integral approximation, we can perform additional function calculations at each time step. nth-order Runge–Kutta integration is the approximation of an integral

by performing n function calculations at each time step. Rectangular integration is therefore equivalent to first-order Runge–Kutta integration, and trapezoidal integration is equivalent to second-order Runge–Kutta integration.

The most commonly used integration scheme of this type is fourth-order Runge–Kutta integration. We present the fourth-order Runge–Kutta integration algorithm (without derivation) as follows:

$$\begin{aligned} \Delta x_1 &= f[x(t_k), u(t_k), t_k]T \\ \Delta x_2 &= f[x(t_k) + \Delta x_1/2, u(t_{k+1/2}), t_{k+1/2}]T \\ \Delta x_3 &= f[x(t_k) + \Delta x_2/2, u(t_{k+1/2}), t_{k+1/2}]T \\ \Delta x_4 &= f[x(t_k) + \Delta x_3, u(t_{k+1}), t_{k+1}]T \\ x(t_{k+1}) &\approx x(t_k) + (\Delta x_1 + 2\Delta x_2 + 2\Delta x_3 + \Delta x_4)/6 \end{aligned} \tag{1.116}$$

where $t_{k+1/2} = t_k + T/2$. Fourth-order Runge–Kutta integration is more computationally demanding than rectangular or trapezoidal integration, but it also provides far greater accuracy. This gives the following algorithm for integrating continuous-time dynamics using fourth-order Runge–Kutta integration. The time loop in the algorithm is executed for $t = 0, T, 2T, \cdots, t_f - T$.

Fourth-order Runge–Kutta integration

Assume that $x(0)$ is given
for t = 0 : T : $t_f - T$
  $t_1 = t + T/2$
  $\Delta x_1 = f[x(t), u(t), t]T$
  $\Delta x_2 = f[x(t) + \Delta x_1/2, u(t_1), t_1]T$
  $\Delta x_3 = f[x(t) + \Delta x_2/2, u(t_1), t_1]T$
  $\Delta x_4 = f[x(t) + \Delta x_3, u(t+T), t+T]T$
  $x(t+T) = x(t) + (\Delta x_1 + 2\Delta x_2 + 2\Delta x_3 + \Delta x_4)/6$
end

Runge–Kutta integration was invented by Carl Runge, a German mathematician and physicist, in 1895. It was independently invented and generalized by Wilhelm Kutta, a German mathematician and aerodynamicist, in 1901. More accurate integration algorithms have also been derived and are sometimes used, but fourth-order Runge–Kutta integration is generally considered a good trade-off between accuracy and computational effort. Further information and derivations of numerical integration algorithms can be found in many numerical analysis texts, including [Atk89].

■ EXAMPLE 1.5

Suppose we want to numerically compute $x(t)$ at $t = 1$ based on the differential equation

$$\dot{x} = \cos t \tag{1.117}$$

with the initial condition $x(0) = 0$. We can analytically integrate the equation to find out that $x(1) = \sin 1 \approx 0.8415$. If we use a numerical integration scheme, we have to choose the step size T. Table 1.1 shows the error of the rectangular, trapezoidal, and fourth-order Runge–Kutta integration methods for this example for various values of T. As expected, Runge–Kutta is more accurate than trapezoidal, and trapezoidal is more accurate than rectangular.

Also as expected, the error for given method decreases as T decreases. However, perhaps the most noteworthy feature of Table 1.1 is *how* the integration error decreases with T. We can see that with rectangular integration, when T is halved, the integration error is also halved. With trapezoidal integration, when T is halved, the integration error decreases by a factor of four. With Runge–Kutta integration, when T is halved, the integration error decreases by a factor of 16. We conclude that (in general) the error of rectangular integration is proportional to T, the error of trapezoidal integration is proportional to T^2, and the error of Runge–Kutta integration is proportional to T^4.

Table 1.1 Example 1.5 results. Percent errors when numerically integrating $\dot{x} = \cos t$ from $t = 0$ to $t = 1$, for various integration algorithms, and for various time step sizes T.

	$T = 0.1$	$T = 0.05$	$T = 0.025$
Rectangular	2.6	1.3	0.68
Trapezoidal	0.083	0.021	0.0052
Fourth-order Runge–Kutta	3.5×10^{-6}	2.2×10^{-7}	1.4×10^{-8}

$\nabla\nabla\nabla$

1.6 STABILITY

In this section, we review the concept of stability for linear time-invariant systems. We first deal with continuous-time systems in Section 1.6.1, and then discrete-time systems in Section 1.6.2. We state the important results here without proof. The interested reader can refer to standard books on linear systems for more details and additional results [Kai80, Bay99, Che99].

1.6.1 Continuous-time systems

Consider the zero-input, linear, continuous-time system

$$\begin{aligned} \dot{x} &= Ax \\ y &= Cx \end{aligned} \tag{1.118}$$

The definitions of marginal stability and asymptotic stability are as follows.

Definition 1 *A linear continuous-time, time-invariant system is marginally stable if the state $x(t)$ is bounded for all t and for all bounded initial states $x(0)$.*

Marginal stability is also called Lyapunov stability.

Definition 2 *A linear continuous-time, time-invariant system is asymptotically stable if, for all bounded initial states $x(0)$,*

$$\lim_{t \to \infty} x(t) = 0 \tag{1.119}$$

The above two definitions show that a system is marginally stable if it is asymptotically stable. That is, asymptotic stability is a subset of marginal stability. Marginal stability and asymptotic stability are types of internal stability. This is because they deal with only the state of the system (i.e., the internal condition of the system) and do not consider the output of the system. More specific categories of internal stability (e.g., uniform stability and exponential stability) are given in some books on linear systems.

Since the solution of Equation (1.118) is given as

$$x(t) = \exp(At)x(0) \tag{1.120}$$

we can state the following theorem.

Theorem 1 *A linear continuous-time, time-invariant system is marginally stable if and only if*

$$\lim_{t\to\infty} \exp(At) \leq M < \infty \tag{1.121}$$

for some constant matrix M. This is just a way of saying that the matrix exponential does not increase without bound.

The "less than or equal to" relation in the above theorem raises some questions, because the quantities on either side of this mathematical symbol are matrices. What does it mean for a matrix to be less than another matrix? It can be interpreted several ways. For example, to say that $A < B$ is usually interpreted to mean that $(B - A)$ is positive definite.[5] In the above theorem we can use any reasonable definition for the matrix inequality and the theorem still holds.

A similar theorem can be stated by combining Definition (2) with Equation (1.120).

Theorem 2 *A linear continuous-time, time-invariant system is asymptotically stable if and only if*

$$\lim_{t\to\infty} \exp(At) = 0 \tag{1.122}$$

Now recall that $\exp(At) = Q\exp(\hat{A}t)Q^{-1}$, where Q is a constant matrix containing the eigenvectors of A, and $\hat{A}$ is the Jordan form of A. The exponential $\exp(\hat{A}t)$ therefore contains terms like $\exp(\lambda_i t)$, $t\exp(\lambda_i t)$, $t^2\exp(\lambda_i t)$, and so on, where λ_i is an eigenvalue of A. The boundedness of $\exp(At)$ is therefore related to the eigenvalues of A as stated by the following theorems.

Theorem 3 *A linear continuous-time, time-invariant system is marginally stable if and only if one of the following conditions holds.*

1. *All of the eigenvalues of A have negative real parts.*
2. *All of the eigenvalues of A have negative or zero real parts, and those with real parts equal to zero have a geometric multiplicity equal to their algebraic multiplicity. That is, the Jordan blocks that are associated with the eigenvalues that have real parts equal to zero are first order.*

Theorem 4 *A linear continuous-time, time-invariant system is asymptotically stable if and only if all of the eigenvalues of A have negative real parts.*

[5] Sometimes the statement $A < B$ means that every element of A is less than the corresponding element of B. However, we will not use that definition in this book.

■ **EXAMPLE 1.6**

Consider the system

$$\dot{x} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} x \tag{1.123}$$

Since the A matrix is upper triangular, we know that its eigenvalues are on the diagonal; that is, the eigenvalues of A are equal to 0, 0, and -1. We see that the system is asymptotically unstable since some of the eigenvalues are nonnegative. We also note that the A matrix is already in Jordan form, and we see that the Jordan block corresponding to the 0 eigenvalue is second order. Therefore, the system is also marginally unstable. The solution of this system is

$$\begin{aligned} x(t) &= \exp(At)x(0) \\ &= \begin{bmatrix} 1 & t & 0 \\ 0 & 1 & 0 \\ 0 & 0 & e^{-t} \end{bmatrix} x(0) \end{aligned} \tag{1.124}$$

The element in the first row and second column of $\exp(At)$ increases without bound as t increases, so there are some initial states $x(0)$ that will result in unbounded $x(t)$. However, there are also some initial states $x(0)$ that will result in bounded $x(t)$. For example, if $x(0) = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}^T$, then

$$\begin{aligned} x(t) &= \begin{bmatrix} 1 & t & 0 \\ 0 & 1 & 0 \\ 0 & 0 & e^{-t} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 \\ 0 \\ e^{-t} \end{bmatrix} \end{aligned} \tag{1.125}$$

and $x(t)$ will be bounded for all t. However, this does not say anything about the stability of the system; it only says that there exists some $x(0)$ that results in a bounded $x(t)$. If we instead choose $x(0) = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^T$, then

$$\begin{aligned} x(t) &= \begin{bmatrix} 1 & t & 0 \\ 0 & 1 & 0 \\ 0 & 0 & e^{-t} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} t \\ 1 \\ 0 \end{bmatrix} \end{aligned} \tag{1.126}$$

and $x(t)$ increases without bound. This proves that the system is asymptotically unstable and marginally unstable.

$\nabla\nabla\nabla$

■ EXAMPLE 1.7

Consider the system

$$\dot{x} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} x \tag{1.127}$$

The eigenvalues of A are equal to 0, 0, and -1. We see that the system is asymptotically unstable since some of the eigenvalues are nonnegative. In order to see if the system is marginally stable, we need to compute the geometric multiplicity of the 0 eigenvalue. (This can be done by noticing that A is already in Jordan form, but we will go through the exercise more completely for the sake of illustration.) Solving the equation

$$(\lambda I - A)v = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{1.128}$$

(where $\lambda = 0$) for nonzero vectors v, we see that there are two linearly independent solutions given as

$$v = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \tag{1.129}$$

This shows that the geometric multiplicity of the 0 eigenvalue is equal to 2, which means that the system is marginally stable. The solution of this system is

$$\begin{aligned} x(t) &= \exp(At)x(0) \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & e^{-t} \end{bmatrix} x(0) \end{aligned} \tag{1.130}$$

Regardless of $x(0)$, we see that $x(t)$ will always be bounded, which means that the system is marginally stable. Note that $x(t)$ may approach 0 as t increases, depending on the value of $x(0)$. For example, if $x(0) = \begin{bmatrix} 0 & 0 & -1 \end{bmatrix}^T$, then

$$x(t) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & e^{-t} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -e^{-t} \end{bmatrix} \tag{1.131}$$

and $x(t)$ approaches 0 as t increases. However, this does not say anything about the asymptotic stability of the system; it only says that there exists some $x(0)$ that results in state $x(t)$ that asymptotically approaches 0. If we instead choose $x(0) = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}^T$, then

$$x(t) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & e^{-t} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \tag{1.132}$$

and $x(t)$ does not approach 0. This proves that the system is asymptotically unstable.

$\nabla\nabla\nabla$

1.6.2 Discrete-time systems

Consider the zero-input, linear, discrete-time, time-invariant system

$$\begin{aligned} x_{k+1} &= Fx_k \\ y_k &= Hx_k \end{aligned} \tag{1.133}$$

The definitions of marginal stability (also called Lyapunov stability) and asymptotic stability are analogous to the definitions for continuous-time systems that were given in Section 1.6.1.

Definition 3 *A linear discrete-time, time-invariant system is marginally stable if the state x_k is bounded for all k and for all bounded initial states x_0.*

Definition 4 *A linear discrete-time, time-invariant system is asymptotically stable if*

$$\lim_{k\to\infty} x_k = 0 \tag{1.134}$$

for all bounded initial states x_0.

Marginal stability and asymptotic stability are types of internal stability. This is because they deal with only the state of the system (i.e., the internal condition of the system) and do not consider the output of the system. More specific categories of internal stability (e.g., uniform stability and exponential stability) are given in some books on linear systems.

Since the solution of Equation (1.133) is given as

$$x_k = A^k x_0 \tag{1.135}$$

we can state the following theorems.

Theorem 5 *A linear discrete-time, time-invariant system is marginally stable if and only if*

$$\lim_{k\to\infty} A^k \leq M < \infty \tag{1.136}$$

for some constant matrix M. This is just a way of saying that the powers of A do not increase without bound.

Theorem 6 *A linear discrete-time, time-invariant system is asymptotically stable if and only if*

$$\lim_{k\to\infty} A^k = 0 \tag{1.137}$$

Now recall that $A^k = Q\hat{A}^k Q^{-1}$, where Q is a constant matrix containing the eigenvectors of A, and $\hat{A}$ is the Jordan form of A. The matrix $\hat{A}^k$ therefore contains terms like λ_i^k, $k\lambda_i^k$, $k^2\lambda_i^k$, and so on, where λ_i is an eigenvalue of A. The boundedness of A^k is therefore related to the eigenvalues of A as stated by the following theorems.

Theorem 7 *A linear discrete-time, time-invariant system is marginally stable if and only if one of the following conditions holds.*

1. *All of the eigenvalues of A have magnitude less than one.*

2. *All of the eigenvalues of A have magnitude less than or equal to one, and those with magnitude equal to one have a geometric multiplicity equal to their algebraic multiplicity. That is, the Jordan blocks that are associated with the eigenvalues that have magnitude equal to one are first order.*

Theorem 8 *A linear discrete-time, time-invariant system is asymptotically stable if and only if all of the eigenvalues of A have magnitude less than one.*

1.7 CONTROLLABILITY AND OBSERVABILITY

The concepts of controllability and observability are fundamental to modern control theory. These concepts define how well we can control a system (i.e., drive the state to a desired value) and how well we can observe a system (i.e., determine the initial conditions after measuring the outputs). These concepts are also important to some of the theoretical results related to optimal state estimation that we will encounter later in this book.

1.7.1 Controllability

The following definitions and theorems give rigorous definitions for controllability for linear systems in the both the continuous-time and discrete-time cases.

Definition 5 *A continuous-time system is controllable if for any initial state $x(0)$ and any final time $t > 0$ there exists a control that transfers the state to any desired value at time t.*

Definition 6 *A discrete-time system is controllable if for any initial state x_0 and some final time k there exists a control that transfers the state to any desired value at time k.*

Note the controllability definition in the continuous-time case is much more demanding than the definition in the discrete-time case. In the continuous-time case, the existence of a control is required for *any* final time. In the discrete-time case, the existence of a control is required for *some* final time. In both cases, controllability is independent of the output equation.

There are several tests for controllability. The following equivalent theorems can be used to test for the controllability of continuous linear time-invariant systems.

Theorem 9 *The n-state*[6] *continuous linear time-invariant system $\dot{x} = Ax + Bu$ has the controllability matrix P defined by*

$$P = \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} \tag{1.138}$$

The system is controllable if and only if $\rho(P) = n$.

Theorem 10 *The n-state continuous linear time-invariant system $\dot{x} = Ax + Bu$ is controllable if and only if the controllability grammian defined by*

$$\int_0^t e^{A\tau} B B^T e^{A^T \tau} \, d\tau \tag{1.139}$$

[6]The notation *n-state system* indicates a system that has n elements in its state variable x.

is positive definite for some $t \in (0, \infty)$.

Theorem 11 *The n-state continuous linear time-invariant system* $\dot{x} = Ax + Bu$ *is controllable if and only if the differential Lyapunov equation*

$$\begin{aligned} W(0) &= 0 \\ \dot{W} &= WA^T + AW + BB^T \end{aligned} \tag{1.140}$$

has a positive definite solution $W(t)$ *for some* $t \in (0, \infty)$. *This is also called a Sylvester equation.*

Similar to the continuous-time case, the following equivalent theorems can be used to test for the controllability of discrete linear time-invariant systems.

Theorem 12 *The n-state discrete linear time-invariant system* $x_k = Fx_{k-1} + Gu_{k-1}$ *has the controllability matrix* P *defined by*

$$P = \begin{bmatrix} G & FG & \cdots & F^{n-1}G \end{bmatrix} \tag{1.141}$$

The system is controllable if and only if $\rho(P) = n$.

Theorem 13 *The n-state discrete linear time-invariant system* $x_k = Fx_{k-1} + Gu_{k-1}$ *is controllable if and only if the controllability grammian defined by*

$$\sum_{i=0}^{k} A^{k-i} BB^T (A^T)^{k-i} \tag{1.142}$$

is positive definite for some $k \in (0, \infty)$.

Theorem 14 *The n-state discrete linear time-invariant system* $x_k = Fx_{k-1} + Gu_{k-1}$ *is controllable if and only if the difference Lyapunov equation*

$$\begin{aligned} W_0 &= 0 \\ W_{i+1} &= FW_iF^T + GG^T \end{aligned} \tag{1.143}$$

has a positive definite solution W_k *for some* $k \in (0, \infty)$. *This is also called a Stein equation.*

Note that Theorems 9 and 12 give identical tests for controllability for both continuous-time and discrete-time systems. In general, these are the simplest controllability tests. Controllability tests for time-varying linear systems can be obtained by generalizing the above theorems. Controllability for nonlinear systems is much more difficult to formalize.

■ **EXAMPLE 1.8**

The RLC circuit of Figure 1.4 has the system description

$$\begin{bmatrix} \dot{v}_C \\ \dot{i}_L \end{bmatrix} = \begin{bmatrix} -2/RC & 1/C \\ -1/L & 0 \end{bmatrix} \begin{bmatrix} v_C \\ i_L \end{bmatrix} + \begin{bmatrix} 1/RC \\ 1/L \end{bmatrix} u \tag{1.144}$$

where v_C is the voltage across the capacitor, i_L is the current through the inductor, and u is the applied voltage. We will use Theorem 9 to determine the conditions under which this system is controllable. The controllability matrix is computed as

$$\begin{aligned} P &= \begin{bmatrix} B & AB \end{bmatrix} \\ &= \begin{bmatrix} 1/RC & 1/LC - 2/R^2C^2 \\ 1/L & -1/RLC \end{bmatrix} \end{aligned} \tag{1.145}$$

From this we can compute the determinant of P as

$$|P| = 1/R^2LC^2 - 1/L^2C \tag{1.146}$$

The determinant of P is 0 only if $R = \sqrt{L/C}$. So the system is controllable unless $R = \sqrt{L/C}$. It would be very difficult to obtain this result from Theorems 10 and 11.

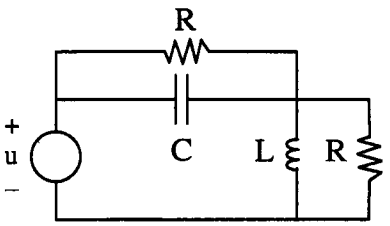

Figure 1.4 RLC circuit for Example 1.8.

$\nabla\nabla\nabla$

1.7.2 Observability

The following definitions and theorems give rigorous definitions for observability for linear systems in both the continuous-time and discrete-time cases.

Definition 7 *A continuous-time system is observable if for any initial state $x(0)$ and any final time $t > 0$ the initial state $x(0)$ can be uniquely determined by knowledge of the input $u(\tau)$ and output $y(\tau)$ for all $\tau \in [0, t]$.*

Definition 8 *A discrete-time system is observable if for any initial state x_0 and some final time k the initial state x_0 can be uniquely determined by knowledge of the input u_i and output y_i for all $i \in [0, k]$.*

Note the observability definition in the continuous-time case is much more demanding than the definition in the discrete-time case. In the continuous-time case, the initial state must be able to be determined at *any* final time. In the discrete-time case, the initial state must be able to be determined at *some* final time. If a system is observable then the initial state can be determined, and if the initial state can be determined then all states between the initial and final times can be determined.

There are several tests for controllability. The following equivalent theorems can be used to test for the controllability of continuous linear time-invariant systems.

Theorem 15 *The n-state continuous linear time-invariant system*

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx \end{aligned} \tag{1.147}$$

has the observability matrix Q defined by

$$Q = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \tag{1.148}$$

The system is observable if and only if $\rho(Q) = n$.

Theorem 16 *The n-state continuous linear time-invariant system*

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx \end{aligned} \tag{1.149}$$

is observable if and only if the observability grammian defined by

$$\int_0^t e^{A^T\tau} C^T C e^{A\tau}\, d\tau \tag{1.150}$$

is positive definite for some $t \in (0, \infty)$.

Theorem 17 *The n-state continuous linear time-invariant system*

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx \end{aligned} \tag{1.151}$$

is observable if and only if the differential Lyapunov equation

$$\begin{aligned} W(t) &= 0 \\ -\dot{W} &= WA + A^T W + C^T C \end{aligned} \tag{1.152}$$

has a positive definite solution $W(\tau)$ *for some* $\tau \in (0, t)$. *This is also called a Sylvester equation.*

Similar to the continuous-time case, the following equivalent theorems can be used to test for the observability of discrete linear time-invariant systems.

Theorem 18 *The n-state discrete linear time-invariant system*

$$\begin{aligned} x_k &= Fx_{k-1} + Gu_{k-1} \\ y_k &= Hx_k \end{aligned} \tag{1.153}$$

has the observability matrix Q defined by

$$Q = \begin{bmatrix} H \\ HF \\ \vdots \\ HF^{n-1} \end{bmatrix} \tag{1.154}$$

The system is observable if and only if $\rho(Q) = n$.

Theorem 19 *The n-state discrete linear time-invariant system*

$$\begin{aligned} x_k &= Fx_{k-1} + Gu_{k-1} \\ y_k &= Hx_k \end{aligned} \tag{1.155}$$

is observable if and only if the observability grammian defined by

$$\sum_{i=0}^{k} (F^T)^i H^T H F^i \tag{1.156}$$

is positive definite for some $k \in (0, \infty)$.

Theorem 20 *The n-state discrete linear time-invariant system*

$$\begin{aligned} x_k &= Fx_{k-1} + Gu_{k-1} \\ y_k &= Hx_k \end{aligned} \tag{1.157}$$

is observable if and only if the difference Lyapunov equation

$$\begin{aligned} W_k &= 0 \\ W_i &= F^T W_{i+1} F + H^T H \end{aligned} \tag{1.158}$$

has a positive definite solution W_0 for some $k \in (0, \infty)$. This is also called a Stein equation.

Note that Theorems 15 and 18 give identical tests for observability for both continuous-time and discrete-time systems. In general, these are the simplest observability tests. Observability tests for time-varying linear systems can be obtained by generalizing the above theorems. Observability for nonlinear systems is much more difficult to formalize.

■ **EXAMPLE 1.9**

The RLC circuit of Example 1.8 has the system description

$$\begin{aligned} \begin{bmatrix} \dot{v}_C \\ \dot{i}_L \end{bmatrix} &= \begin{bmatrix} -2/RC & 1/C \\ -1/L & 0 \end{bmatrix} \begin{bmatrix} v_C \\ i_L \end{bmatrix} + \begin{bmatrix} 1/RC \\ 1/L \end{bmatrix} u \\ y &= \begin{bmatrix} -1 & 0 \end{bmatrix} \begin{bmatrix} v_C \\ i_L \end{bmatrix} \end{aligned} \tag{1.159}$$

where v_C is the voltage across the capacitor, i_L is the current through the inductor, and u is the applied voltage. We will use Theorem 15 to determine

the conditions under which this system is observable. The observability matrix is computed as

$$Q = \begin{bmatrix} C \\ CA \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 2/RC & -1/C \end{bmatrix} \tag{1.160}$$

The determinant of the observability matrix can be computed as

$$|Q| = 1/C \tag{1.161}$$

The determinant of Q is nonzero, so the system is observable. On the other hand, suppose that $R = L = C = 1$ and the output equation is

$$y = \begin{bmatrix} -1 & 1 \end{bmatrix} \begin{bmatrix} v_C \\ i_L \end{bmatrix} \tag{1.162}$$

Then the observability matrix can be computed as

$$\begin{aligned} Q &= \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \\ |Q| &= 0 \end{aligned} \tag{1.163}$$

So the system is unobservable. It would be very difficult to obtain this result from Theorems 16 and 17.

$\nabla\nabla\nabla$

1.7.3 Stabilizability and detectability

The concepts of stabilizability and detectability are closely related to controllability and observability, respectively. These concepts are also related to the modes of a system. The modes of a system are all of the decoupled states after the system is transformed into Jordan form. A system can be transformed into Jordan form as follows. Consider the system

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx + Du \end{aligned} \tag{1.164}$$

First find the eigendata of the system matrix A. Suppose the eigenvectors are denoted as $v_1, \cdots, v_n$. Create an $n \times n$ matrix M by augmenting the eigenvectors as follows.

$$M = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} \tag{1.165}$$

Define a new system as

$$\begin{aligned} \dot{\bar{x}} &= M^{-1}AM\bar{x} + M^{-1}B \\ &= \bar{A}\bar{x} + \bar{B}u \\ y &= CM\bar{x} + Du \\ &= \bar{C}\bar{x} + Du \end{aligned} \tag{1.166}$$

The new system is called the Jordan form representation of the original system. Note that the matrix M will always be invertible because the eigenvectors of a matrix can always be chosen to be linearly independent. The two systems of Equations (1.164) and (1.166) are called algebraically equivalent systems. This is because they have the same input and the same output (and therefore they have the same transfer function) but they have different states.

■ **EXAMPLE 1.10**

Consider the system

$$\begin{aligned} \dot{x} &= Ax + Bu \\ &= \begin{bmatrix} 1 & 1 & 2 \\ 0 & 1 & 3 \\ 0 & 0 & -2 \end{bmatrix} x + \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} u \\ y &= Cx + Du \\ &= \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} + 2u \end{aligned} \tag{1.167}$$

This system has the same transfer function as

$$\begin{aligned} \dot{\bar{x}} &= \bar{A}\bar{x} + \bar{B}u \\ &= \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix} x + \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} u \\ y &= \bar{C}\bar{x} + Du \\ &= \begin{bmatrix} 1 & 0 & 1 \end{bmatrix} \bar{x} + 2u \end{aligned} \tag{1.168}$$

The eigenvector matrix of A is

$$\begin{aligned} M &= \begin{bmatrix} v_1 & v_2 & v_n \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & -3 \end{bmatrix} \end{aligned} \tag{1.169}$$

Note the equivalences

$$\begin{aligned} \bar{A} &= M^{-1}AM \\ \bar{B} &= M^{-1}B \\ \bar{C} &= CM \end{aligned} \tag{1.170}$$

The Jordan form system has two decoupled modes. The first mode is

$$\begin{aligned} \dot{\bar{x}}_1 &= \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \bar{x}_1 + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u \\ y_1 &= \begin{bmatrix} 1 & 0 \end{bmatrix} \bar{x}_1 \end{aligned} \tag{1.171}$$

The second mode is

$$\begin{aligned} \dot{\bar{x}}_2 &= -2\bar{x}_2 + 0\,u \\ y_2 &= \bar{x}_2 \end{aligned} \tag{1.172}$$

$\nabla\nabla\nabla$

Definition 9 *If a system is controllable or stable, then it is also stabilizable. If a system is uncontrollable or unstable, then it is stabilizable if its uncontrollable modes are stable.*

In Example 1.10, the first mode is unstable (both eigenvalues at $+1$) but controllable. The second mode is stable (eigenvalue at -2) but uncontrollable. Therefore, the system is stabilizable.

Definition 10 *If a system is observable or stable, then it is also detectable. If a system is unobservable or unstable, then it is detectable if its unobservable modes are stable.*

In Example 1.10, the first mode is unstable but observable. The second mode is both stable and observable. Therefore, the system is detectable.

Controllability and observability were introduced by Rudolph Kalman at a conference in 1959 whose proceedings were published in an obscure Mexican journal in 1960 [Kal60b]. The material was also presented at an IFAC conference in 1960 [Kal60c], and finally published in a more widely accessible format in 1963 [Kal63].

1.8 SUMMARY

In this chapter we have reviewed some of the basic concepts of linear systems theory that are fundamental to many approaches to optimal state estimation. We began with a review of matrix algebra and matrix calculus, which proves to be indispensable in much of the theory of state estimation techniques. For additional information on matrix theory, the reader can refer to several excellent texts [Hor85, Gol89, Ber05]. We continued in this chapter with a review of linear and nonlinear systems, in both continuous time and discrete time. We regard time as continuous for physical systems, but our simulations and estimation algorithms operate in discrete time because of the popularity of digital computing. We discussed the discretization of continuous-time systems, which is a way of obtaining a discrete-time mathematical representation of a continuous-time system. The concept of stability can be used to tell us if a system's states will always remain bounded. Controllability tells us if it is possible to find a control input to force system states to our desired values, and observability tells us if it is possible to determine the initial state of a system on the basis of output measurements. State-space theory in general, and linear systems theory in particular, is a wide-ranging discipline that is typically covered in a one-semester graduate course, but there is easily enough material to fill up a two-semester course. Many excellent textbooks have been written on the subject, including [Bay99, Che99, Kai00] and others. A solid understanding of linear systems will provide a firm foundation for further studies in areas such as control theory, estimation theory, and signal processing.

PROBLEMS

Written exercises

1.1 Find the rank of the matrix $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$.

1.2 Find two 2×2 matrices A and B such that $A \neq B$, neither A nor B are diagonal, $A \neq cB$ for any scalar c, and $AB = BA$. Find the eigenvectors of A and

B. Note that they share an eigenvector. Interestingly, every pair of commuting matrices shares at least one eigenvector [Hor85, p. 51].

1.3 Prove the three identities of Equation (1.26).

1.4 Find the partial derivative of the trace of AB with respect to A.

1.5 Consider the matrix

$$A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

Recall that the eigenvalues of A are found by find the roots of the polynomial $P(\lambda) = |\lambda I - A|$. Show that $P(A) = 0$. (This is an illustration of the Cayley–Hamilton theorem [Bay99, Che99, Kai00].)

1.6 Suppose that A is invertible and

$$\begin{bmatrix} A & A \\ B & A \end{bmatrix} \begin{bmatrix} A \\ C \end{bmatrix} = \begin{bmatrix} 0 \\ I \end{bmatrix}$$

Find B and C in terms of A [Lie67].

1.7 Show that AB may not be symmetric even though both A and B are symmetric.

1.8 Consider the matrix

$$A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

where a, b, and c are real, and a and c are nonnegative.

a) Compute the solutions of the characteristic polynomial of A to prove that the eigenvalues of A are real.

b) For what values of b is A positive semidefinite?

1.9 Derive the properties of the state transition matrix given in Equation (1.72).

1.10 Suppose that the matrix A has eigenvalues λ_i and eigenvectors v_i $(i = 1, \cdots, n)$. What are the eigenvalues and eigenvectors of $-A$?

1.11 Show that $|e^{At}| = e^{|A|t}$ for any square matrix A.

1.12 Show that if $\dot{A} = BA$, then

$$\frac{d|A|}{dt} = \mathrm{Tr}(B)|A|$$

1.13 The linear position p of an object under constant acceleration is

$$p = p_0 + \dot{p}t + \frac{1}{2}\ddot{p}t^2$$

where p_0 is the initial position of the object.

a) Define a state vector as $x = \begin{bmatrix} p & \dot{p} & \ddot{p} \end{bmatrix}^T$ and write the state space equation $\dot{x} = Ax$ for this system.

b) Use all three expressions in Equation (1.71) to find the state transition matrix e^{At} for the system.

c) Prove for the state transition matrix found above that $e^{A0} = I$.

1.14 Consider the following system matrix.

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

Show that the matrix

$$S(t) = \begin{bmatrix} e^t & 0 \\ 0 & 2e^{-t} \end{bmatrix}$$

satisfies the relation $\dot{S}(t) = AS(t)$, but $S(t)$ is not the state transition matrix of the system.

1.15 Give an example of a discrete-time system that is marginally stable but not asymptotically stable.

1.16 Show (H, F) is an observable matrix pair if and only if (H, F^{-1}) is observable (assuming that F is nonsingular).

Computer exercises

1.17 The dynamics of a DC motor can be described as

$$J\ddot{\theta} + F\dot{\theta} = T$$

where θ is the angular position of the motor, J is the moment of inertia, F is the coefficient of viscous friction, and T is the torque applied to the motor.

a) Generate a two-state linear system equation for this motor in the form

$$\dot{x} = Ax + Bu$$

b) Simulate the system for 5 s and plot the angular position and velocity. Use $J = 10$ kg m^2, $F = 100$ kg m^2/s, $x(0) = \begin{bmatrix} 0 & 0 \end{bmatrix}^T$, and $T = 10$ N m. Use rectangular integration with a step size of 0.05 s. Do the output plots look correct? What happens when you change the step size Δt to 0.2 s? What happens when you change the step size to 0.5 s? What are the eigenvalues of the A matrix, and how can you relate their magnitudes to the step size that is required for a correct simulation?

1.18 The dynamic equations for a series RLC circuit can be written as

$$\begin{aligned} u &= IR + L\dot{I} + V_c \\ I &= C\dot{V}_c \end{aligned}$$

where u is the applied voltage, I is the current through the circuit, and V_c is the voltage across the capacitor.

a) Write a state-space equation in matrix form for this system with x_1 as the capacitor voltage and x_2 as the current.

b) Suppose that $R = 3$, $L = 1$, and $C = 0.5$. Find an analytical expression for the capacitor voltage for $t \geq 0$, assuming that the initial state is zero, and the input voltage is $u(t) = e^{-2t}$.

c) Simulate the system using rectangular, trapezoidal, and fourth-order Runge–Kutta integration to obtain a numerical solution for the capacitor voltage. Simulate from $t = 0$ to $t = 5$ using step sizes of 0.1 and 0.2. Tabulate the RMS value of the error between the numerical and analytical solutions for the capacitor voltage for each of your six simulations.

1.19 The vertical dimension of a hovering rocket can be modeled as

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= \frac{Ku - gx_2}{x_3} - \frac{GM}{(R + x_1)^2} \\ \dot{x}_3 &= -u \end{aligned}$$

where x_1 is the vertical position of the rocket, x_2 is the vertical velocity, x_3 is the mass of the rocket, u is the control input (the flow rate of rocket propulsion), $K = 1000$ is the thrust constant of proportionality, $g = 50$ is the drag constant, $G = 6.673E - 11$ m^3/kg/s^2 is the universal gravitational constant, $M = 5.98E24$ kg is the mass of the earth, and $R = 6.37E6$ m is the radius of the earth radius.

a) Find $u(t) = u_0(t)$ such that the system is in equilibrium at $x_1(t) = 0$ and $x_2(t) = 0$.

b) Find $x_3(t)$ when $x_1(t) = 0$ and $x_2(t) = 0$.

c) Linearize the system around the state trajectory found above.

d) Simulate the nonlinear system for five seconds and the linearized system for five seconds with $u(t) = u_0(t) + \Delta u \cos(t)$. Plot the altitude of the rocket for the nonlinear simulation and the linear simulation (on the same plot) when $\Delta u = 10$. Repeat for $\Delta u = 100$ and $\Delta u = 300$. Hand in your source code and your three plots. What do you conclude about the accuracy of your linearization?