STOCHASTIC PROCESSES, DETECTION AND ESTIMATION

6.432 Course Notes

Alan S. Willsky, Gregory W. Wornell, and Jeffrey H. Shapiro

Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139

Fall 2003

1

Probability, Random Vectors, and Vector Spaces

1.1 INTRODUCTION

In these notes, we explore a powerful and remarkably broad framework for generating, modeling, and processing signals characterized by some degree of randomness or uncertainty. As we'll see, this framework will be useful not only conceptually, but also practically, allowing us to develop a wide range of efficient algorithms for various kinds of applications. In this first chapter of the course notes, we develop important foundations for this framework, which we'll build on in subsequent chapters.

We build this foundation by combining the tools of probability theory with concepts from the theory of vector spaces. We assume you've had a good deal of exposure to basic probability concepts in your undergraduate curriculum. And we assume you've also developed considerable experience with Euclidean vector spaces and the associated tools of linear algebra from your undergraduate curriculum. One role this chapter serves is to collect together and summarize those concepts and techniques from this material that we will exploit extensively throughout the course. However, the larger and more important purpose of this chapter is to introduce and develop new ideas that arise from exploiting these ideas jointly. As an example, we'll develop the concept of a random vector, and explore some important ways for characterizing such quantities. And we'll introduce the notion of abstract (non-Euclidean) vector spaces, which we'll use in turn, to explore, e.g., the notion of vector spaces of random variables. Some of these ideas will undoubtedly seem quite unusual at first, and will take some time and effort to digest.

However, as we'll see they lead to some powerful geometric perspectives that will play a key role in the course.

A detailed outline of the chapter is as follows. We begin with a compact summary of those probability concepts that will be of most use to us. Building on this foundation, we then introduce random vectors using a vector-matrix notation, and develop key concepts and properties. Finally, we introduce the concept of abstract vector space, and develop several important examples of such spaces, including those involving random variables. The accompanying appendices summarize important concepts and results from linear algebra and vector calculus that we rely on in this and future chapters. Additional results from linear algebra and vector space theory will be developed as we need them in subsequent chapters.

1.2 AXIOMS OF PROBABILITY AND BASIC CONCEPTS

A probability space, $(\Omega, \Pr[\cdot])$, for an experiment consists of a sample space $\Omega = \{\omega\}$ containing all the elementary outcomes of the experiment, and a probability measure $\Pr[\cdot]$ which assigns probabilities to subsets of Ω (called events).[1] The measure $\Pr[\cdot]$ has the following properties:

$$0 \leq \Pr[\mathcal{A}] \leq 1 \qquad \text{for all valid } \mathcal{A} \subset \Omega \tag{1.1}$$

$$\Pr[\Omega] = 1 \tag{1.2}$$

$$\Pr[\mathcal{A} \cup \mathcal{B}] = \Pr[\mathcal{A}] + \Pr[\mathcal{B}] \qquad \text{if } \mathcal{A} \cap \mathcal{B} = \emptyset. \tag{1.3}$$

Two of the many consequences of these axioms are

$$\Pr[\emptyset] = 0 \tag{1.4}$$

and

$$\Pr[\mathcal{A} \cup \mathcal{B}] = \Pr[\mathcal{A}] + \Pr[\mathcal{B}] - \Pr[\mathcal{A} \cap \mathcal{B}]. \tag{1.5}$$

Finally, (1.5) can be used with induction to establish the *union bound:* if the $\mathcal{A}_i$, $i = 1, 2, \ldots, n$ are an arbitrary collection of events, then

$$\Pr\left[\bigcup_{i=1}^{n} \mathcal{A}_i\right] \leq \sum_{i=1}^{n} \Pr[\mathcal{A}_i], \tag{1.6}$$

where equality in (1.6) holds if and only if the $\mathcal{A}_i$ are a collection of *mutually exclusive* events, i.e., if $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$ for $i \neq j$.

[1] In fact we cannot compute the probability of *every* subset of Ω. Those that we can we will term *valid* subsets. In formal mathematical treatments a probability space is specified in terms of a sample space, a probability measure, and a collection of valid sets. At our level of treatment, however, you can assume that any subset we mention or construct—either explicitly or implicitly—is valid.

1.2.1 Conditional Probabilities

The *conditional probability* of event $\mathcal{A}$ given event $\mathcal{B}$ is defined by

$$\Pr[\mathcal{A} \mid \mathcal{B}] = \frac{\Pr[\mathcal{A} \cap \mathcal{B}]}{\Pr[\mathcal{B}]}. \tag{1.7}$$

Exploiting, in turn, the definition of $\Pr[\mathcal{B} \mid \mathcal{A}]$ in the numerator of (1.7) yields

$$\Pr[\mathcal{A} \mid \mathcal{B}] = \frac{\Pr[\mathcal{B} \mid \mathcal{A}]\ \Pr[\mathcal{A}]}{\Pr[\mathcal{B}]}. \tag{1.8}$$

A straightforward extention of (1.8) is *Bayes' Rule*: let $\mathcal{A}_i$, $i = 1, 2, \ldots, n$ be a set of mutually exclusive events that are also *exhaustive*, i.e.,

$$\bigcup_{i=1}^{n} \mathcal{A}_i = \Omega;$$

then for any event $\mathcal{B}$,

$$\Pr[\mathcal{A}_j \mid \mathcal{B}] = \frac{\Pr[\mathcal{B} \mid \mathcal{A}_j]\ \Pr[\mathcal{A}_j]}{\sum_{i=1}^{n} \Pr[\mathcal{B} \mid \mathcal{A}_i]\ \Pr[\mathcal{A}_i]}. \tag{1.9}$$

1.2.2 Independence

Two (nontrivial) events are *independent* if knowledge of one event's occurrence provides no information about the other event's occurrence, i.e., if

$$\Pr[\mathcal{A} \mid \mathcal{B}] = \Pr[\mathcal{A}]. \tag{1.10}$$

Using (1.7) we see that (1.10) is equivalent to

$$\Pr[\mathcal{A} \cap \mathcal{B}] = \Pr[\mathcal{A}]\ \Pr[\mathcal{B}]. \tag{1.11}$$

More generally, a collection of events $\{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_N\}$ are said to be *mutually independent* if for every i we have

$$\Pr[\mathcal{A}_i \mid \{\mathcal{A}_j, j \in \mathcal{J}\}] = \Pr[\mathcal{A}_i], \tag{1.12}$$

where $\mathcal{J}$ is any subset of indices between 1 through N but excluding i. The condition (1.12) is equivalent to the requirement that for *every* subset of distinct indices $i_1, i_2, \ldots, i_K$, drawn from $1, 2, \ldots, N$ and corresponding to some $K \leq N$ we have

$$\Pr\left[\bigcap_{k=1}^{K} \mathcal{A}_{i_k}\right] = \prod_{k=1}^{K} \Pr[\mathcal{A}_{i_k}]. \tag{1.13}$$

For three events $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ to be mutually independent, for example, this means that we require that *all* the following hold:

$$\Pr[\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}] = \Pr[\mathcal{A}] \Pr[\mathcal{B}] \Pr[\mathcal{C}] \tag{1.14}$$

$$\Pr[\mathcal{A} \cap \mathcal{B}] = \Pr[\mathcal{A}] \Pr[\mathcal{B}] \tag{1.15}$$

$$\Pr[\mathcal{A} \cap \mathcal{C}] = \Pr[\mathcal{A}] \Pr[\mathcal{C}] \tag{1.16}$$

$$\Pr[\mathcal{B} \cap \mathcal{C}] = \Pr[\mathcal{B}] \Pr[\mathcal{C}]. \tag{1.17}$$

In particular, (1.14) alone is not sufficient; (1.15)–(1.17) are also required.

1.3 RANDOM VARIABLES

In these notes, we adopt the useful convention of using fonts without serifs for random variables, and the corresponding fonts with serifs for sample values and dummy arguments. For example, x, y, z, and Θ will denote random variables, and x, y, z, and Θ corresponding generic sample values.

Formally, a random variable x is a real-valued function on the sample space Ω. The *probability distribution function* for x is defined by

$$P_{\mathsf{x}}(x) = \Pr[\omega \mid \mathsf{x}(\omega) \leq x] = \Pr[\mathsf{x} \leq x], \tag{1.18}$$

where the last expression we use for notational convenience. This distribution is a complete characterization of the random variable. Likewise, the *probability density function* (pdf) $p_{\mathsf{x}}(x)$, which is related to the distribution by

$$p_{\mathsf{x}}(x) = \frac{dP_{\mathsf{x}}(x)}{dx}, \tag{1.19}$$

is also a complete characterization.[2] This follows from the fact that for any valid $\mathcal{A}$, we can write

$$\Pr[\mathsf{x} \in \mathcal{A}] = \int_{\mathcal{A}} p_{\mathsf{x}}(x)\, dx. \tag{1.20}$$

If x takes on particular values with nonzero probability, then $P_{\mathsf{x}}(x)$ will contain step-discontinuities and $p_{\mathsf{x}}(x)$ will contain impulses. For example,

$$p_{\mathsf{x}}(x) = \frac{1}{2}\delta(x+1) + \frac{1}{2}\delta(x-1) \tag{1.21}$$

is the density of a random variable taking on the values ± 1 each with the probability $1/2$. To accommodate the possibility of $p_{\mathsf{x}}(x)$ having impulses and remain consistent with (1.18), we write the inverse of (1.19) as

$$P_{\mathsf{x}}(x) = \int_{-\infty}^{x+} p_{\mathsf{x}}(u)\, du, \tag{1.22}$$

[2] We'll assume in our treatment that densities always exist for the quantities of interest, at least in this generalized sense (i.e., allowing impulses). However, it is worth keeping in mind that there exist random variables whose probability distributions are not differentiable even in a generalized sense.

using $x+$ in the upper limit of (1.22) to indicate that the endpoint x is included in the interval. Also, since [cf. (1.19)]

$$p_x(x_0) = \lim_{\delta x \to 0} \frac{\Pr[x_0 < x \leq x_0 + \delta x]}{\delta x},$$

we have the frequently useful approximation valid for suitably small δx:

$$\Pr[x_0 < x \leq x_0 + \delta x] \approx p_x(x_0)\, \delta x. \tag{1.23}$$

1.3.1 Expectations

Often we are interested in partial characterizations of a random variable in the form of certain *expectations*. The *expected value* of a function $g(x)$ of a random variable is given by

$$\overline{g(x)} \triangleq E[g(x)] = \int_{-\infty}^{+\infty} g(x)\, p_x(x)\, dx. \tag{1.24}$$

Remember that since $z = g(x)$ is itself a random variable we may also write (1.24) in the form

$$E[g(x)] = E[z] = \int_{-\infty}^{+\infty} z\, p_z(z)\, dz, \tag{1.25}$$

where $p_z(z)$ is the probability density for $z = g(x)$. If $g(\cdot)$ is a one-to-one and differentiable function, a simple expression for $p_z(z)$ can be derived, viz.,

$$p_z(z) = \frac{p_x\left(g^{-1}(z)\right)}{\left|g'\left(g^{-1}(z)\right)\right|} \tag{1.26}$$

where $g'(\cdot)$ denotes the first derivative of $g(\cdot)$. If $g(\cdot)$ is not invertible, the more general method-of-events approach for deriving densities, which we briefly review later in the multivariate case, can be employed to obtain $p_z(z)$. In terms of the ultimate goal of evaluating $E[g(x)]$, whether (1.24) or (1.25) turns out to be more convenient depends on the problem at hand.

Several expectations that are important partial characterizations of a random variable are the *mean value* (or first moment)

$$E[x] = \overline{x} \triangleq m_x, \tag{1.27}$$

the *mean-squared value* (or second moment)

$$E\left[x^2\right] = \overline{x^2}, \tag{1.28}$$

and the *variance* (or second central-moment)

$$E\left[(x - m_x)^2\right] = \overline{x^2} - m_x^2 \triangleq \operatorname{var} x \triangleq \sigma_x^2 \triangleq \lambda_x. \tag{1.29}$$

In (1.27)–(1.29) we have introduced a variety of notation that will be convenient to use in subsequent sections of these notes. The *standard deviation* is σ_x, the square

root of the variance. One important bound provided by these moments is the *Chebyshev inequality*

$$\Pr\left[|x - m_x| \geq \varepsilon\right] \leq \frac{\sigma_x^2}{\varepsilon^2} \tag{1.30}$$

Observe that (1.30) implies that x is a constant (i.e., $\Pr[x = \alpha] = 1$ for some constant α) if $\sigma_x^2 = 0$. Note that the corresponding "only if" statement follows immediately from (1.29).

The Chebyshev bound is a particularly convenient bound to use in practice because its calculation involves only the mean and variance of the random variable, i.e., it doesn't depend on the detailed form of the density. However, for this same reason the Chebyshev bound is not a particularly tight bound.[3]

1.3.2 Characteristic Functions

The *characteristic function* of a random variable is defined as

$$M_x(jv) = E\left[e^{jvx}\right] = \int_{-\infty}^{+\infty} e^{jvx}\, p_x(x)\, dx \tag{1.31}$$

and, as is apparent from the integral in (1.31), corresponds to the Fourier transform of the density (to within a minor sign change). As a Fourier transform, we can recover $p_x(x)$ from $M_x(jv)$ via the inverse formula

$$p_x(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-jvx}\, M_x(jv)\, dv. \tag{1.32}$$

and hence the characteristic function is an equivalent complete characterization of a random variable.

Characteristic functions are particularly useful in computing certain expectations involving the random variable. For example, the moments of x can all be efficiently recovered from $M_x(jv)$ by differentiation, i.e.,

$$E\left[x^n\right] = \left[\frac{1}{j^n} \frac{d^n}{dv^n} M_x(jv)\right]\Bigg|_{v=0} \tag{1.33}$$

Observe that (1.33) implies that the characteristic function can be expanded in terms of the power series

$$M_x(jv) = \sum_{k=0}^{+\infty} E\left[x^k\right] \frac{(jv)^k}{k!} \tag{1.34}$$

when all the moments of the form (1.33) exist. This result implies, in turn, that knowledge of all moments is an equivalent characterization for such random variables: given these moments we can reconstruct $M_x(jv)$ via (1.34).

[3] As an aside, an alternative bound that is typically much tighter but which requires access to more information about the random variable is the *Chernoff bound*.

The characteristic function is also frequently useful in deriving the density of the sum of independent random variables. In particular, if $\{x_1, x_2, \ldots, x_N\}$ is a set of mutually independent random variables then the characteristic function for their sum

$$z = x_1 + x_2 + \cdots + x_N$$

is simply given by the product of the characteristic functions of the constituents, i.e.,

$$\begin{aligned} M_z(jv) &= E\left[e^{jvz}\right] = E\left[e^{jv(x_1+x_2+\cdots+x_N)}\right] \\ &= E\left[e^{jvx_1}\right]\, E\left[e^{jvx_2}\right]\, \cdots\, E\left[e^{jvx_N}\right] \\ &= M_{x_1}(jv)\, M_{x_2}(jv)\, \cdots\, M_{x_N}(jv). \end{aligned} \tag{1.35}$$

Thus, after computing $M_z(jv)$ via (1.35), we can determine $p_z(z)$ by the Fourier transform inverse formula (1.32). Note that inverting (1.35) directly yields, via the convolution property of Fourier transforms, the familiar result

$$p_z(z) = p_{x_1}(z) * p_{x_2}(z) * \cdots * p_{x_N}(z), \tag{1.36}$$

where $*$ denotes the convolution operator.

1.3.3 Discrete Random Variables

Random variables that take on only integer values can be fully developed within the framework we've been describing. In particular, their probability densities consist entirely of uniformly-spaced impulses with suitable weights. However, to make manipulation of these quantities less cumbersome, it is sometimes convenient to adopt some special notation for specifically discrete random variables. In particular, we define the *probability mass function* (pmf) of an integer-valued random variable k as

$$p_k[k] = \Pr[k = k] \tag{1.37}$$

using square brackets to distinguish masses from densities, and to remind us that the argument is integer-valued. The density can, of course, be derived from the mass function via

$$p_k(k) = \sum_{i=-\infty}^{+\infty} p_k[k]\, \delta(i-k) \tag{1.38}$$

and related to the distribution function via

$$P_k(k) = \sum_{i=-\infty}^{k} p_k[i]. \tag{1.39}$$

With this notation, expectations can be expressed in terms of sums rather than integrals, e.g.,

$$E[f[k]] = \sum_{k=-\infty}^{+\infty} f[k]\, p_k[k].$$

Furthermore, the characteristic function can be viewed as the *discrete-time* Fourier transform (again to within a minor sign change) of the mass function, i.e.,

$$M_k(jv) = E\left[e^{jvk}\right] = \sum_{k=-\infty}^{+\infty} e^{jvk} p_k[k].$$

1.4 PAIRS OF RANDOM VARIABLES

We will frequently deal with several random variables, and we will find it convenient to use vector notation in this case. Before we do that, however, let us recall some complete joint characterizations of a pair of random variables x and y. One such characterization is the joint distribution function for x and y, which is defined by

$$P_{x,y}(x, y) = \Pr\left[x \leq x \text{ and } y \leq y\right]. \tag{1.40}$$

A second complete joint characterization is the joint density of x and y, i.e.,

$$p_{x,y}(x, y) = \frac{\partial^2 P_{x,y}(x, y)}{\partial x\, \partial y}. \tag{1.41}$$

If $\mathcal{A} \subset \mathbb{R}^2$ is a valid set, where $\mathbb{R}^2$ denotes[4] the plane of all pairs (x, y), then

$$\Pr\left[(x, y) \in \mathcal{A}\right] = \int\int_{\mathcal{A}} p_{x,y}(x, y)\, dx\, dy. \tag{1.42}$$

Evidently, a special case of (1.42) is the inverse to (1.41), i.e.,

$$P_{x,y}(x, y) = \int_{-\infty}^{x+} \int_{-\infty}^{y+} p_{x,y}(u, v)\, du\, dv.$$

Again, from (1.41) we also have the following approximation valid for suitably small δx and δy:

$$\Pr\left[x_0 < x \leq x_0 + \delta x \text{ and } y_0 < y \leq y_0 + \delta y\right] \approx p_{x,y}(x_0, y_0)\, \delta x\, \delta y. \tag{1.43}$$

1.4.1 Marginal and Conditional Densities

Recall that the marginal densities of either x or y can be recovered by integrating out the other variable:

$$p_x(x) = \int_{-\infty}^{+\infty} p_{x,y}(x, y)\, dy \tag{1.44}$$

$$p_y(y) = \int_{-\infty}^{+\infty} p_{x,y}(x, y)\, dx. \tag{1.45}$$

[4]See the Appendix 1.A for a discussion of such spaces.

Geometrically, it is useful to visualize marginal densities $p_x(x)$ and $p_y(y)$ as (integrated) projections of the joint density $p_{x,y}(x,y)$ onto the x and y axes, respectively, of the (x,y) plane.

The conditional density for x given that $y = y$ is defined by

$$p_{x|y}(x|y) = \frac{p_{x,y}(x,y)}{p_y(y)}, \tag{1.46}$$

which, as a function of x and for a particular $y = y_0$, corresponds to a slice into the plane through the joint density along the $y = y_0$ line that is normalized to have unit integral. In particular, the denominator in (1.46), i.e., $p_y(y)$, is precisely this normalization factor. Note too that since $p_{y|x}(y|x)$ is defined analogously, we have, e.g.,

$$p_{x|y}(x|y) = \frac{p_{y|x}(y|x)\, p_x(x)}{p_y(y)}.$$

1.4.2 Independence

A pair of random variables x and y are independent if knowledge of the value of one does not affect the density of the other, i.e., if

$$p_{x|y}(x|y) = p_x(x). \tag{1.47}$$

Using (1.47) with (1.46), we get the equivalent condition for independence

$$p_{x,y}(x,y) = p_x(x)\, p_y(y). \tag{1.48}$$

1.4.3 Expectations and Correlation

Expectations also provide partial characterizations for pairs of random variables. The expected value of a function of x and y is given by

$$E\left[f(x,y)\right] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\, p_{x,y}(x,y)\, dx\, dy = \overline{f(x,y)}. \tag{1.49}$$

Note that (1.49) gives (1.24) as a special case:

$$E\left[g(x)\right] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x)\, p_{x,y}(x,y)\, dx\, dy = \int_{-\infty}^{+\infty} g(x)\, p_x(x)\, dx. \tag{1.50}$$

On many occasions we will exploit the fact that expectations are *linear operations*. For example, for arbitrary constants α and β we have

$$E\left[\alpha x + \beta y\right] = \alpha E\left[x\right] + \beta E\left[y\right].$$

This means that in computations we can typically interchange expectations with summations, integrations, and other linear operations.

In addition to (1.27)–(1.29) for x and their counterparts for y, some additional expectations that constitute useful partial characterizations of the statistical relationship between x and y are

Correlation:

$$E[xy] \tag{1.51}$$

Covariance:

$$\begin{aligned} E[(x - m_x)(y - m_y)] &= E[xy] - m_x m_y \\ &= \lambda_{xy} \triangleq \operatorname{cov}(x, y). \end{aligned} \tag{1.52}$$

Note that $\operatorname{var} x = \operatorname{cov}(x, x)$.

A pair of variables x and y are said to be *uncorrelated* if $\lambda_{xy} = 0$, i.e., if

$$E[xy] = E[x]\, E[y].$$

If random variables x and y are independent, they are also uncorrelated:

$$\begin{aligned} E[xy] &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy\, p_{x,y}(x, y)\, dx\, dy \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy\, p_x(x) p_y(y)\, dx\, dy \\ &= E[x]\ E[y]. \end{aligned}$$

However, the converse is not true—uncorrelated random variables are generally not independent. A pair of random variables is said to be *orthogonal* when

$$E[xy] = 0.$$

For reasons that will become apparent in Section 1.7, we sometimes express this condition using the notation $x \perp y$.

The *correlation coefficient* ρ_{xy} is a normalized measure of the correlation between two random variables x and y, and is defined by

$$\rho_{xy} = \frac{\lambda_{xy}}{\sigma_x \sigma_y}. \tag{1.53}$$

Later in Section 1.7, we will establish that $-1 \leq \rho_{xy} \leq 1$, with

$$\begin{aligned} \rho_{xy} = 0 &\Leftrightarrow x \text{ and } y \text{ uncorrelated} \\ \rho_{xy} = +1 &\Leftrightarrow x \text{ is a positive multiple of } y \text{ plus a constant} \\ \rho_{xy} = -1 &\Leftrightarrow x \text{ is a negative multiple of } y \text{ plus a constant} \end{aligned} \tag{1.54}$$

We can also define *conditional expectations* involving x and y. The conditional expectation of x given $y = y$, for example, is given by

$$E[x|y = y] = \int_{-\infty}^{+\infty} x\, p_{x|y}(x|y)\, dx. \tag{1.55}$$

Note that $E[x|y=y]$ is a function of y and consequently $E[x|y]$ can be viewed as a random variable. Its expectation is then

$$\begin{aligned} E[E[x|y]] &= \int_{-\infty}^{+\infty} E[x|y=y]\ p_y(y)\,dy \\ &= \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} x\, p_{x|y}(x|y)\,dx \right) p_y(y)\,dy \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x\, p_{x,y}(x,y)\,dx\,dy = E[x] \end{aligned} \tag{1.56}$$

The identity (1.56), which we'll use on many occasions in this course, is called the law of "iterated expectation."

As a final remark, we point out that the condition

$$E[x|y] = E[x] \tag{1.57}$$

is equivalent to neither independence nor uncorrelatedness. While it is true that (1.57) holds if x and y are independent since

$$E[x|y=y] = \int_{-\infty}^{+\infty} x\, p_{x|y}(x|y)\,dx = \int_{-\infty}^{+\infty} x\, p_x(x)\,dx = E[x],$$

the converse is not true: x and y are not necessarily independent if (1.57) holds. As a simple counterexample we have the joint density

$$p_{x,y}(x,y) = \frac{1}{3}\delta(x,y-1) + \frac{1}{3}\delta(x+1,y) + \frac{1}{3}\delta(x-1,y). \tag{1.58}$$

Likewise, it is true that x and y are uncorrelated if (1.57) holds since, using iterated expectation, we have

$$E[xy] = E[E[xy|y]] = E[yE[x|y]] = E[yE[x]] = E[x]\,E[y];$$

however, the converse is again not true: if x and y are uncorrelated, we cannot deduce that (1.57) holds. A simple counterexample is the density (1.58) with x and y interchanged:

$$p_{x,y}(x,y) = \frac{1}{3}\delta(x-1,y) + \frac{1}{3}\delta(x,y+1) + \frac{1}{3}\delta(x,y-1).$$

1.5 RANDOM VECTORS

It is generally more convenient to represent collections of two or more random variables in terms of a vector of random variables.[5] If, for example, we let

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \tag{1.59}$$

denote a vector of N random variables, then the joint distribution function can be expressed in the form

$$P_{\mathbf{x}}(\mathbf{x}) = \Pr\left[x_1 \leq x_1, x_2 \leq x_2, \ldots, x_N \leq x_N\right]. \tag{1.60}$$

Provided this distribution function is differentiable, the corresponding joint density function is

$$p_{\mathbf{x}}(\mathbf{x}) = \frac{\partial^N P_{\mathbf{x}}(\mathbf{x})}{\partial x_1 \, \partial x_2 \cdots \partial x_N}. \tag{1.61}$$

Analogous to the scalar case, we have $p_{\mathbf{x}}(\mathbf{x}) \geq 0$ and

$$\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} p_{\mathbf{x}}(\mathbf{x}) \, d\mathbf{x} = 1. \tag{1.62}$$

The joint density function is a complete characterization of the random variables that comprise the vector. In particular, for any valid set $\mathcal{A} \subset \mathbb{R}^N$

$$\Pr\left[\mathbf{x} \in \mathcal{A}\right] = \int \cdots \int_{\mathcal{A}} p_{\mathbf{x}}(\mathbf{x}) \, dx_1 \cdots dx_N = \int_{\mathcal{A}} p_{\mathbf{x}}(\mathbf{x}) \, d\mathbf{x} \tag{1.63}$$

where the last expression is a notational convenience. Furthermore, we can reconstruct the original distribution function from the joint density via

$$P_{\mathbf{x}}(\mathbf{x}) = \int_{-\infty}^{x_1+} \int_{-\infty}^{x_2+} \cdots \int_{-\infty}^{x_N+} p_{\mathbf{x}}(\mathbf{u}) \, d\mathbf{u}. \tag{1.64}$$

Rather than collecting all the random variables of interest into a single vector $\mathbf{x}$, in many problems it is often more natural and more convenient to divide them among several random vectors of possibly different sizes.

In the case where we divide our random variables into two random vectors $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^M$, we can define the joint distribution

$$P_{\mathbf{x},\mathbf{y}}(\mathbf{x}, \mathbf{y}) = \Pr\left[x_1 \leq x_1, x_2 \leq x_2, \ldots, x_N \leq x_N, y_1 \leq y_1, y_2 \leq y_2, \ldots, y_M \leq y_M\right] \tag{1.65}$$

[5] Since we use bold face fonts for vectors, random vectors will be denoted using bold face fonts without serifs, and sample values will be denoted using bold face fonts with serifs. For example, $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, and $\boldsymbol{\Theta}$ will be random vectors, and $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, and $\boldsymbol{\Theta}$ will be associated sample values.

and the joint density

$$p_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y}) = \frac{\partial^{N+M} P_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y})}{\partial x_1\, \partial x_2 \cdots \partial x_N\, \partial y_1\, \partial y_2 \cdots \partial y_M}, \tag{1.66}$$

each of which fully characterizes the statistical relationship among all the elements of $\mathbf{x}$ and $\mathbf{y}$.

In turn, using the joint density (1.66), we can recover marginal densities, e.g.,

$$p_{\mathbf{x}}(\mathbf{x}) = \int_{-\infty}^{+\infty} p_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y})\, d\mathbf{y}. \tag{1.67}$$

1.5.1 Conditional Densities and Independence

The conditional density for $\mathbf{x}$ given $\mathbf{y} = \mathbf{y}$ is given by

$$p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) = \frac{p_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y})}{p_{\mathbf{y}}(\mathbf{y})}, \tag{1.68}$$

and hence we have

$$p_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y}) = p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\, p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\, p_{\mathbf{x}}(\mathbf{x}). \tag{1.69}$$

Thus, using (1.69) we have, in turn,

$$p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) = \frac{p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x})\, p_{\mathbf{x}}(\mathbf{x})}{p_{\mathbf{y}}(\mathbf{y})}. \tag{1.70}$$

Two random vectors $\mathbf{x}$ and $\mathbf{y}$ are independent (meaning that the two corresponding collections of random variables are mutually independent of one another) if knowledge of any of the elements of $\mathbf{y}$ provides no information about any of the elements of $\mathbf{x}$ (or vice versa), i.e., if

$$p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) = p_{\mathbf{x}}(\mathbf{x}). \tag{1.71}$$

Analogous to our earlier results, using (1.68) we find that (1.71) is equivalent to the condition

$$p_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y}) = p_{\mathbf{x}}(\mathbf{x})\, p_{\mathbf{y}}(\mathbf{y}). \tag{1.72}$$

All of these formulas extend to more than two random vectors. For instance, a collection of K random vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K$ are mutually independent if for every i we have

$$p_{\mathbf{x}_i|\{\mathbf{x}_j, j\in\mathcal{J}\}}(\mathbf{x}_i \mid \{\mathbf{x}_j, j \in \mathcal{J}\}) = p_{\mathbf{x}_i}(\mathbf{x}_i), \tag{1.73}$$

where $\mathcal{J}$ is any subset of indices between 1 through K but excluding i. The condition (1.73) is equivalent to the requirement that

$$p_{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = \prod_{i=1}^{n} p_{\mathbf{x}_i}(\mathbf{x}_i). \tag{1.74}$$

Note that by integrating out any subset of vectors in (1.74) we obtain lower-order independence relations among arbitrary subsets of the random vectors as well, i.e., if $\mathcal{I}$ is an arbitrary subset of distinct indices selected from 1 to K, then

$$p_{\{\mathbf{x}_i, i\in\mathcal{I}\}}(\{\mathbf{x}_i, i \in \mathcal{I}\}) = \prod_{i\in\mathcal{I}} p_{\mathbf{x}_i}(\mathbf{x}_i). \tag{1.75}$$

To show that (1.74) implies (1.73) involves a straightforward application of the definition of conditional probabilities:

$$\begin{aligned} & p_{\mathbf{x}_i|\{\mathbf{x}_j, j\in\mathcal{J}\}}(\mathbf{x}_i \mid \{\mathbf{x}_j, j \in \mathcal{J}\}) \\ &= \frac{p_{\mathbf{x}_i,\{\mathbf{x}_j,j\in\mathcal{J}\}}(\mathbf{x}_i, \{\mathbf{x}_j, j \in \mathcal{J}\})}{p_{\{\mathbf{x}_j,j\in\mathcal{J}\}}(\{\mathbf{x}_j, j \in \mathcal{J}\})} \\ &= \frac{\prod_{k\in\{i\}\cup\mathcal{J}} p_{\mathbf{x}_k}(\mathbf{x}_k)}{\prod_{k\in\mathcal{J}} p_{\mathbf{x}_k}(\mathbf{x}_k)} \\ &= p_{\mathbf{x}_i}(\mathbf{x}_i). \end{aligned}$$

To show the converse—that (1.73) implies (1.74)—requires rewriting the joint density as the product of conditionals of the form of the left-hand side of (1.73). As a special case, if $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are three random vectors then their joint density can be expressed as

$$\begin{aligned} p_{\mathbf{x},\mathbf{y},\mathbf{z}}(\mathbf{x},\mathbf{y},\mathbf{z}) &= p_{\mathbf{x}|\mathbf{y},\mathbf{z}}(\mathbf{x}|\mathbf{y},\mathbf{z})\, p_{\mathbf{y},\mathbf{z}}(\mathbf{y},\mathbf{z}) \\ &= p_{\mathbf{x}|\mathbf{y},\mathbf{z}}(\mathbf{x}|\mathbf{y},\mathbf{z})\, p_{\mathbf{y}|\mathbf{z}}(\mathbf{y}|\mathbf{z})\, p_{\mathbf{z}}(\mathbf{z}). \end{aligned} \tag{1.76}$$

Applying (1.73) to each of the right-hand side terms of (1.76), we get that $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are mutually independent if

$$p_{\mathbf{x},\mathbf{y},\mathbf{z}}(\mathbf{x},\mathbf{y},\mathbf{z}) = p_{\mathbf{x}}(\mathbf{x})\, p_{\mathbf{y}}(\mathbf{y})\, p_{\mathbf{z}}(\mathbf{z}). \tag{1.77}$$

We emphasize that from integrations of (1.77) we get that mutual independence implies pairwise independence, i.e.,

$$\begin{aligned} p_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y}) &= p_{\mathbf{x}}(\mathbf{x})\, p_{\mathbf{y}}(\mathbf{y}) \\ p_{\mathbf{x},\mathbf{z}}(\mathbf{x},\mathbf{z}) &= p_{\mathbf{x}}(\mathbf{x})\, p_{\mathbf{z}}(\mathbf{z}) \\ p_{\mathbf{y},\mathbf{z}}(\mathbf{y},\mathbf{z}) &= p_{\mathbf{y}}(\mathbf{y})\, p_{\mathbf{z}}(\mathbf{z}) \end{aligned}$$

However, the converse is not true—pairwise independence alone does not ensure mutual independence.

Finally, we note that all of the results in this section can be used in the special case in which each of the random vectors has only a single element and are, hence, scalar random variables. As an example, we have that the random variables $\{x_1, x_2, \ldots, x_N\}$ are mutually independent if and only if

$$p_{x_1,x_2,\ldots,x_N}(x_1, x_2, \ldots, x_N) = p_{x_1}(x_1)\, p_{x_2}(x_2) \cdots p_{x_N}(x_N). \tag{1.78}$$

1.5.2 Derived Distributions and Jacobians

Suppose

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = \mathbf{g}(\mathbf{x}) = \begin{bmatrix} g_1(\mathbf{x}) \\ g_2(\mathbf{x}) \\ \vdots \\ g_M(\mathbf{x}) \end{bmatrix}$$

is an M-dimensional random vector obtained as a function of the N-dimensional random vector $\mathbf{x}$. We can always in principle calculate the distribution for $\mathbf{y}$ from the method of events:

$$\begin{aligned} P_{\mathbf{y}}(\mathbf{y}) &= \Pr\left[g_1(\mathbf{x}) \le y_1,\, g_2(\mathbf{x}) \le y_2, \ldots,\, g_M(\mathbf{x}) \le y_M\right] \\ &= \int_{\mathcal{A}(\mathbf{y})} p_{\mathbf{x}}(\mathbf{x})\, d\mathbf{x} \end{aligned} \tag{1.79}$$

where

$$\mathcal{A}(\mathbf{y}) = \{\mathbf{x} \mid g_1(\mathbf{x}) \le y_1,\, g_2(\mathbf{x}) \le y_2, \ldots,\, g_M(\mathbf{x}) \le y_M\} \tag{1.80}$$

We can then obtain the density via

$$p_{\mathbf{y}}(\mathbf{y}) = \frac{\partial^M P_{\mathbf{y}}(\mathbf{y})}{\partial y_1\, \partial y_2 \cdots \partial y_M}. \tag{1.81}$$

If $M = N$ and $\mathbf{g}(\mathbf{x})$ is one-to-one, this approach leads to the expression

$$p_{\mathbf{y}}(\mathbf{y}) = \frac{p_x\left(\mathbf{g}^{-1}(\mathbf{y})\right)}{\left|\frac{d\mathbf{g}}{d\mathbf{x}}\left(\mathbf{g}^{-1}(\mathbf{y})\right)\right|} \tag{1.82}$$

where, as is discussed in Appendix 1.B, $d\mathbf{g}/d\mathbf{x}$ is the Jacobian matrix corresponding to $\mathbf{g}$ and $|\cdot| = \det(\cdot)$ denotes the determinant of its matrix argument.

1.5.3 Expectations and Covariance Matrices

The expectation of a scalar-valued function of $\mathbf{x}$ is given by

$$E\left[f(\mathbf{x})\right] = \int_{-\infty}^{+\infty} f(\mathbf{x})\, p_{\mathbf{x}}(\mathbf{x})\, d\mathbf{x} \tag{1.83}$$

The expectations of vector-valued (or even matrix-valued) functions of $\mathbf{x}$ are defined component-wise. For example, if

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ \vdots \\ f_M(\mathbf{x}) \end{bmatrix}, \tag{1.84}$$

then

$$E\left[\mathbf{f}(\mathbf{x})\right] = \begin{bmatrix} E\left[f_1(\mathbf{x})\right] \\ E\left[f_2(\mathbf{x})\right] \\ \vdots \\ E\left[f_M(\mathbf{x})\right] \end{bmatrix}. \tag{1.85}$$

Some important expectations are:

Mean Vector:

$$E\left[\mathbf{x}\right] = \mathbf{m}_{\mathbf{x}} \tag{1.86}$$

Correlation Matrix:

$$E\left[\mathbf{x}\mathbf{x}^{\mathrm{T}}\right] \tag{1.87}$$

Covariance Matrix:

$$\mathrm{cov}\left(\mathbf{x},\mathbf{x}\right) = \mathbf{\Lambda}_{\mathbf{x}\mathbf{x}} = E\left[(\mathbf{x}-\mathbf{m}_{\mathbf{x}})(\mathbf{x}-\mathbf{m}_{\mathbf{x}})^{\mathrm{T}}\right] = E\left[\mathbf{x}\mathbf{x}^{\mathrm{T}}\right] - \mathbf{m}_{\mathbf{x}}\mathbf{m}_{\mathbf{x}}^{\mathrm{T}} \tag{1.88}$$

Cross-Covariance Matrix:

$$\mathrm{cov}\left(\mathbf{x},\mathbf{y}\right) = \mathbf{\Lambda}_{\mathbf{x}\mathbf{y}} = E\left[(\mathbf{x}-\mathbf{m}_{\mathbf{x}})(\mathbf{y}-\mathbf{m}_{\mathbf{y}})^{\mathrm{T}}\right] = E\left[\mathbf{x}\mathbf{y}^{\mathrm{T}}\right] - \mathbf{m}_{\mathbf{x}}\mathbf{m}_{\mathbf{y}}^{\mathrm{T}} \tag{1.89}$$

Conditional Mean:

$$\mathbf{m}_{\mathbf{x}|\mathbf{y}}(\mathbf{y}) = \mathbf{m}_{\mathbf{x}|\mathbf{y}=\mathbf{y}} = E\left[\mathbf{x}|\mathbf{y}=\mathbf{y}\right] = \int_{-\infty}^{+\infty} \mathbf{x}\, p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\, d\mathbf{x} \tag{1.90}$$

Conditional Covariance:

$$\begin{aligned} \mathbf{\Lambda}_{\mathbf{x}|\mathbf{y}}(\mathbf{y}) &= \mathbf{\Lambda}_{\mathbf{x}|\mathbf{y}=\mathbf{y}} \\ &= \int_{-\infty}^{+\infty} \left(\mathbf{x} - E\left[\mathbf{x}|\mathbf{y}=\mathbf{y}\right]\right)\left(\mathbf{x} - E\left[\mathbf{x}|\mathbf{y}=\mathbf{y}\right]\right)^{\mathrm{T}} p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\, d\mathbf{x} \end{aligned} \tag{1.91}$$

As before we can think of the conditional statistics $\mathbf{m}_{\mathbf{x}|\mathbf{y}}$ and $\mathbf{\Lambda}_{\mathbf{x}|\mathbf{y}}$ in (1.90) and (1.91), respectively, as deterministic quantities that are functions of a particular value $\mathbf{y} = \mathbf{y}$. Alternatively, $\mathbf{m}_{\mathbf{x}|\mathbf{y}}$ and $\mathbf{\Lambda}_{\mathbf{x}|\mathbf{y}}$ can be viewed as functions of $\mathbf{y}$ and therefore random variables in their own right. As before, the law of iterated expectation applies, i.e.,

$$E\left[E\left[\mathbf{x}|\mathbf{y}\right]\right] = E\left[\mathbf{x}\right].$$

For notational convenience, we will often drop one of the subscripts in dealing with the covariance of a random vector $\mathbf{x}$, i.e., we will often write $\mathbf{\Lambda}_{\mathbf{x}}$ instead of $\mathbf{\Lambda}_{\mathbf{x}\mathbf{x}}$. In terms of dimensions, note that if $\mathbf{x}$ is N-dimensional and $\mathbf{y}$ is M-dimensional, then $\mathbf{\Lambda}_{\mathbf{x}}$ is $N \times N$, $\mathbf{\Lambda}_{\mathbf{x}\mathbf{y}}$ is $N \times M$, and $\mathbf{\Lambda}_{\mathbf{y}\mathbf{x}}$ is $M \times N$. Furthermore the (i,j)th and (i,i)th elements of $\mathbf{\Lambda}_{\mathbf{x}}$ are

$$\left[\mathbf{\Lambda}_{\mathbf{x}}\right]_{ij} = \mathrm{cov}\left(x_i, x_j\right) \tag{1.92}$$

$$\left[\mathbf{\Lambda}_{\mathbf{x}}\right]_{ii} = \sigma_{x_i}^2 \tag{1.93}$$

while the (i, j)th element of $\mathbf{\Lambda_{xy}}$ is

$$[\mathbf{\Lambda_{xy}}]_{ij} = \mathrm{cov}\,(x_i, y_j)\,. \tag{1.94}$$

Furthermore

$$\mathbf{\Lambda_{yx}} = \mathbf{\Lambda_{xy}}^{\mathrm{T}} \tag{1.95}$$

and $\mathbf{\Lambda_x}$ is a symmetric matrix, i.e.,

$$\mathbf{\Lambda_x} = \mathbf{\Lambda_x}^{\mathrm{T}}. \tag{1.96}$$

The random vectors $\mathbf{x}$ and $\mathbf{y}$ are *uncorrelated* if every element of $\mathbf{x}$ is uncorrelated with every element of $\mathbf{y}$, i.e.,

$$\mathbf{\Lambda_{xy}} = \mathbf{0} \tag{1.97}$$

or

$$E\left[\mathbf{x}\mathbf{y}^{\mathrm{T}}\right] = E\left[\mathbf{x}\right]\left[E\left[\mathbf{y}\right]\right]^{\mathrm{T}}. \tag{1.98}$$

And we say two random vectors $\mathbf{x}$ and $\mathbf{y}$ are *orthogonal* if

$$E\left[\mathbf{x}\mathbf{y}^{\mathrm{T}}\right] = \mathbf{0},$$

i.e., if every element of $\mathbf{x}$ is orthogonal to every element of $\mathbf{y}$.

1.5.4 Characteristic Functions of Random Vectors

The characteristic function of a random vector is given by

$$M_{\mathbf{x}}(j\mathbf{v}) = E\left[e^{j\mathbf{v}^{\mathrm{T}}\mathbf{x}}\right] = \int_{-\infty}^{+\infty} e^{j\mathbf{v}^{\mathrm{T}}\mathbf{x}} p_{\mathbf{x}}(\mathbf{x})\, d\mathbf{x}, \tag{1.99}$$

and corresponds to the (sign-reversed) N-dimensional Fourier transform of the joint density. Analogous to the scalar case, when it exists, the characteristic function corresponds to an alternative complete statistical characterization of the random vector, and in particular the inverse formula for reconstructing $p_{\mathbf{x}}(\mathbf{x})$ from $M_{\mathbf{x}}(j\mathbf{v})$ is

$$p_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^N} \int_{-\infty}^{+\infty} e^{-j\mathbf{v}^{\mathrm{T}}\mathbf{x}} M_{\mathbf{x}}(j\mathbf{v})\, d\mathbf{v}. \tag{1.100}$$

Among its uses, all mixed moments of $\mathbf{x}$ can be efficiently recovered from the characteristic function via differentiation; specifically, with $K = k_1 + k_2 + \cdots + k_N$, we have

$$E\left[x_1^{k_1} x_2^{k_2} \cdots x_N^{k_N}\right] = \left[\frac{1}{j^K} \frac{\partial^K M_{\mathbf{x}}(j\mathbf{v})}{\partial v_1^{k_1}\, \partial v_2^{k_2} \cdots \partial v_N^{k_N}}\right]\Bigg|_{\mathbf{v}=\mathbf{0}}. \tag{1.101}$$

In turn, (1.101) implies that $M_{\mathbf{x}}(j\mathbf{v})$ can be expanded in a power series of the form

$$M_{\mathbf{x}}(j\mathbf{v}) = \sum_{k_1=0}^{+\infty} \sum_{k_2=0}^{+\infty} \cdots \sum_{k_N=0}^{+\infty} E\left[x_1^{k_1} x_2^{k_2} \cdots x_N^{k_N}\right] \frac{(jv_1)^{k_1}}{k_1!} \frac{(jv_2)^{k_2}}{k_2!} \cdots \frac{(jv_N)^{k_N}}{k_N!}, \tag{1.102}$$

provided all the constituent moments exist. Hence, many classes of random vectors are completely characterized by the complete set of moments of the form (1.101).

In addition, note that the collection of random variables $x_1, x_2, \ldots, x_N$ are mutually independent if and only if

$$M_{\mathbf{x}}(j\mathbf{v}) = M_{x_1}(jv_1)\, M_{x_2}(jv_2) \cdots M_{x_N}(jv_N). \tag{1.103}$$

To establish the "only if" part, it suffices to note that if the $x_1, x_2, \ldots, x_N$ are mutually independent then

$$\begin{aligned} M_{\mathbf{x}}(j\mathbf{v}) &= E\left[e^{j\mathbf{v}^{\mathrm{T}}\mathbf{x}}\right] = E\left[e^{j(v_1x_1+v_2x_2+\cdots+v_Nx_N)}\right] \\ &= E\left[e^{jv_1x_1}\right]\, E\left[e^{jv_2x_2}\right] \cdots E\left[e^{jv_Nx_N}\right] \\ &= M_{x_1}(jv_1)\, M_{x_2}(jv_2) \cdots M_{x_N}(jv_N). \end{aligned} \tag{1.104}$$

To establish the "if" part, we note that if (1.103) holds then by (1.100) and (1.32) we have

$$\begin{aligned} p_{\mathbf{x}}(\mathbf{x}) &= \frac{1}{(2\pi)^N} \int_{-\infty}^{+\infty} e^{-j\mathbf{v}^{\mathrm{T}}\mathbf{x}} M_{\mathbf{x}}(j\mathbf{v})\, d\mathbf{v} \\ &= \frac{1}{(2\pi)^N} \int_{-\infty}^{+\infty} d\mathbf{v} \prod_{i=1}^{N} e^{-jv_ix_i} M_{x_i}(jv_i) \\ &= \prod_{i=1}^{N} \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-jv_ix_i} M_{x_i}(jv_i)\, dv_i \\ &= \prod_{i=1}^{N} p_{x_i}(x_i). \end{aligned}$$

Among several implications of (1.103) is the following conceptually useful alternative condition for mutual independence: the collection of random variables $\{x_1, x_2, \ldots, x_N\}$ is mutually independent if and only if for all choices of functions $f_1(\cdot), f_2(\cdot), \ldots, f_N(\cdot)$ we have

$$E\left[f_1(x_1)\, f_2(x_2) \cdots f_N(x_N)\right] = E\left[f_1(x_1)\right]\, E\left[f_2(x_2)\right] \cdots E\left[f_N(x_N)\right]. \tag{1.105}$$

The "only if" part requires a straightforward application of (1.78). To establish the "if" part, it suffices to choose the $f_i(x_i) = e^{jv_ix_i}$ and then exploit (1.103).

Finally, by combining (1.103) with the power series expansions (1.102) and (1.34) we get a related but milder equivalent condition for mutual independence: the collection of random variables $\{x_1, x_2, \ldots, x_N\}$ is mutually independent if and only if for every set of nonnegative integers $k_1, k_2, \ldots, k_N$ we have

$$E\left[x_1^{k_1}\, x_2^{k_2} \cdots x_N^{k_N}\right] = E\left[x_1^{k_1}\right]\, E\left[x_2^{k_2}\right] \cdots E\left[x_N^{k_N}\right]. \tag{1.106}$$

Properties and Geometry of the Covariance Matrix

Let $\mathbf{x}$ and $\mathbf{y}$ be random vectors and define $\mathbf{z}$ as follows:

$$\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y} + \mathbf{b} \tag{1.107}$$

where $\mathbf{A}$ and $\mathbf{B}$ are matrices of appropriate dimensions and $\mathbf{b}$ is a deterministic vector. Then straightforward calculations yield the following:

$$\mathbf{m}_\mathbf{z} = \mathbf{A}\mathbf{m}_\mathbf{x} + \mathbf{B}\mathbf{m}_\mathbf{y} + \mathbf{b} \tag{1.108}$$

$$\mathbf{\Lambda}_\mathbf{z} = \mathbf{A}\mathbf{\Lambda}_\mathbf{x}\mathbf{A}^{\mathrm{T}} + \mathbf{A}\mathbf{\Lambda}_{\mathbf{xy}}\mathbf{B}^{\mathrm{T}} + \mathbf{B}\mathbf{\Lambda}_{\mathbf{yx}}\mathbf{A}^{\mathrm{T}} + \mathbf{B}\mathbf{\Lambda}_\mathbf{y}\mathbf{B}^{\mathrm{T}}. \tag{1.109}$$

Note that if $\mathbf{x}$ and $\mathbf{y}$ are uncorrelated (1.109) simplifies to

$$\mathbf{\Lambda}_\mathbf{z} = \mathbf{A}\mathbf{\Lambda}_\mathbf{x}\mathbf{A}^{\mathrm{T}} + \mathbf{B}\mathbf{\Lambda}_\mathbf{y}\mathbf{B}^{\mathrm{T}}. \tag{1.110}$$

As a special case, let $\mathbf{a}$ be a vector of numbers and consider the scalar random variable

$$z = \mathbf{a}^{\mathrm{T}}\mathbf{x} = \sum_{i=1}^{N} a_i x_i. \tag{1.111}$$

Then from (1.109)

$$\sigma_z^2 = \lambda_z = \mathbf{a}^{\mathrm{T}}\mathbf{\Lambda}_\mathbf{x}\mathbf{a}. \tag{1.112}$$

Since σ_z^2 must be a non-negative number we see that $\mathbf{\Lambda}_\mathbf{x}$ must be a positive semidefinite matrix.[6] If $\mathbf{\Lambda}_\mathbf{x}$ is invertible, i.e., if it is positive definite, then $\sigma_z^2 > 0$ for any vector $\mathbf{a} \neq 0$. However, if $\mathbf{\Lambda}_\mathbf{x}$ is singular, so that $\mathbf{\Lambda}_\mathbf{x}$ is not positive definite, then there is some vector $\mathbf{a} \neq 0$ so that $\sigma_z^2 = \mathbf{a}^{\mathrm{T}}\mathbf{\Lambda}_\mathbf{x}\mathbf{a} = 0$. Consequently, z in this case is a known constant and therefore one of the x_i equals a constant plus a deterministic linear combination of the other components of $\mathbf{x}$.

Example 1.1

Let x and y be two scalar random variables, and consider the random vector

$$\mathbf{w} = \begin{bmatrix} x \\ y \end{bmatrix}. \tag{1.113}$$

Then

$$\mathbf{\Lambda}_\mathbf{w} = \begin{bmatrix} \sigma_x^2 & \lambda_{xy} \\ \lambda_{xy} & \sigma_y^2 \end{bmatrix} = \begin{bmatrix} \sigma_x^2 & \rho_{xy}\sigma_x\sigma_y \\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}. \tag{1.114}$$

For $\mathbf{\Lambda}_\mathbf{w}$ to be positive definite, it must be true that the determinant of $\mathbf{\Lambda}_\mathbf{w}$ is positive:

$$\det(\mathbf{\Lambda}_\mathbf{w}) = (1 - \rho_{xy}^2)\sigma_x^2\sigma_y^2 > 0. \tag{1.115}$$

From this equation we can see that $\mathbf{\Lambda}_\mathbf{w}$ will not be positive definite if and only if the correlation coefficient ρ_{xy} equals ± 1. In either of these cases we can conclude that x must equal a multiple of y plus a constant, i.e., $x = cy + d$ for some constants c and d. It is straightforward to check that the sign of c is the same as that of ρ_{xy}.

[6] See Appendix 1.A for a discussion of positive definite and semidefinite matrices.

To develop the geometrical properties of the covariance matrix, suppose $\mathbf{x}$ is an N-dimensional random vector. As discussed in Appendix 1.A, since $\mathbf{\Lambda_x}$ is symmetric, there exists an orthogonal matrix $\mathbf{P}$ such that if we define

$$\mathbf{z} = \mathbf{P}\mathbf{x} \tag{1.116}$$

then

$$\mathbf{\Lambda_z} = \mathbf{P}\mathbf{\Lambda_x}\mathbf{P}^{\mathrm{T}} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N) \tag{1.117}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_N$ are the eigenvalues of $\mathbf{\Lambda_x}$.

The interpretation of this result is very important. Specifically, we see that we can perform a change of coordinates (1.116) on our random vector so that the components of the transformed vector $\mathbf{z}$ are uncorrelated. The columns of $\mathbf{P}^{\mathrm{T}}$, which are the eigenvectors of $\mathbf{\Lambda_x}$, are the "principal directions" of $\mathbf{\Lambda_x}$, i.e., they specify the linear combinations of $\mathbf{x}$ that make up the uncorrelated components of $\mathbf{z}$. Moreover, the eigenvalues of $\mathbf{\Lambda_x}$ are the variances of the corresponding components of $\mathbf{z}$.

Example 1.2

Continuing Example 1.1, suppose that

$$\mathbf{\Lambda_w} = \begin{bmatrix} 3/2 & 1/2 \\ 1/2 & 3/2 \end{bmatrix}.$$

Then the eigenvalues are $\lambda_1 = 2$ and $\lambda_2 = 1$, and the corresponding normalized eigenvectors are

$$\mathbf{p}_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \qquad \mathbf{p}_2 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}.$$

Hence, we can conclude that the pair of random variables

$$u = \sqrt{2}\,\mathbf{p}_1^{\mathrm{T}}\mathbf{w} = x + y$$
$$v = \sqrt{2}\,\mathbf{p}_2^{\mathrm{T}}\mathbf{w} = x - y$$

are uncorrelated and have variances

$$\mathrm{var}\, u = \mathrm{var}\,[x + y] = 2\ \mathrm{var}\left[\frac{x+y}{\sqrt{2}}\right] = 2\lambda_1 = 4$$
$$\mathrm{var}\, v = \mathrm{var}\,[x - y] = 2\ \mathrm{var}\left[\frac{x-y}{\sqrt{2}}\right] = 2\lambda_2 = 2.$$

1.6 GAUSSIAN RANDOM VARIABLES

In this section we define and develop the basic properties of jointly Gaussian or normal random variables (or, equivalently, Gaussian or normal random vectors). Gaussian random variables are important for at least two reasons. First, Gaussian random vectors are good models in many physical scenarios. For example,

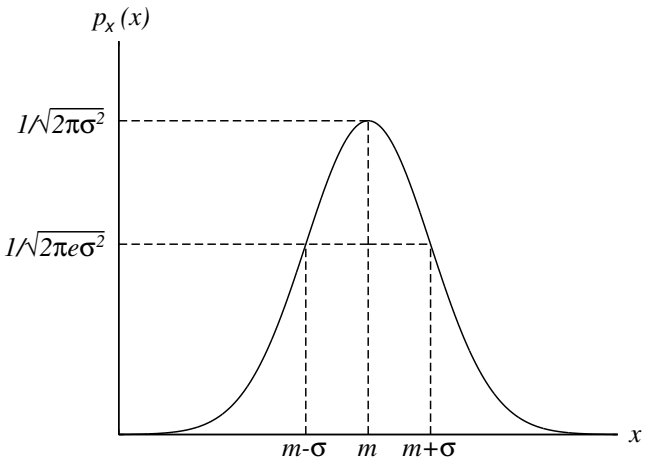

Figure 1.1. The probability density function of a scalar Gaussian random variable.

Gaussian distributions arise in practice when the quantity observed is composed of a superposition of a large number of small, independent, random contributions. This behavior is captured by the *Central Limit Theorem*, which we describe shortly. Second, jointly Gaussian random variables are highly tractable, having convenient mathematical properties that greatly simplify a variety of calculations involving, e.g., linear transformations.

A Gaussian random variable x has a probability density of the form

$$p_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right] \tag{1.118}$$

for some parameters m and $\sigma^2 > 0$. To emphasize the fact that this density is parametrized by these two numbers, we will use the notation $x \sim N(m, \sigma^2)$ as shorthand for "x is Gaussian with mean m and variance σ^2" and we will also write

$$p_x(x) = N(x; m, \sigma^2) = N(x - m; 0, \sigma^2) = \frac{1}{\sigma} N\left(\frac{x-m}{\sigma}; 0, 1\right). \tag{1.119}$$

The density (1.118) is the familiar bell-shaped curve depicted in Fig. 1.1. It is centered at $x = m$ and σ_x is a measure of its width. In fact, the first and second (central) moments of x are related to the parameters in the manner one would expect based on our choice of notation, i.e.,

$$E[x] = m \tag{1.120}$$

$$\operatorname{var} x = \sigma^2. \tag{1.121}$$

Eqs. (1.120) and (1.121) can be verified by direct computation of the expectation integrals.

The characteristic function of a Gaussian random variable x is frequently useful in computations. It takes the form

$$M_x(jv) = \exp\left[jvm - \frac{1}{2}v^2\sigma^2\right], \tag{1.122}$$

which, again, can be verified by direct computation of the corresponding Fourier transform (1.31).

As an example application, we can use the characteristic function to establish that if a and b are arbitrary constants, the random variable $z = ax + b$ is $N(am + b, a^2\sigma^2)$. To verify this, we note that

$$M_z(jv) = E\left[e^{jvz}\right] = E\left[e^{jvax}\right] e^{jvb} = M_x(jva)e^{jvb} = \exp\left[jv(am+b) - \frac{1}{2}v^2(a^2\sigma^2)\right], \tag{1.123}$$

where we have used (1.122) to obtain the last equality in (1.123).

1.6.1 Central Limit Theorem

One fairly general form of the Central Limit Theorem is stated formally as follows. Let

$$x_1, x_2, x_3, \ldots$$

be a sequence of mutually independent zero-mean random variables with distributions

$$P_{x_1}(x_1),\ P_{x_2}(x_2),\ P_{x_3}(x_3),\ \ldots$$

and variances

$$\sigma_1^2,\ \sigma_2^2,\ \sigma_3^2,\ \ldots,$$

respectively. If for any $\epsilon > 0$ there exists a k (depending on ϵ) sufficiently large that

$$\sigma_i < \epsilon\, S_k \qquad \text{for } i = 1, 2, \ldots, k \tag{1.124}$$

with

$$S_k = \sqrt{\sum_{i=1}^{k} \sigma_i^2},$$

then the distribution function of the normalized sum

$$z_n = \frac{1}{S_n} \sum_{i=1}^{n} x_i \tag{1.125}$$

converges to the distribution function of a Gaussian random variable with zero-mean and unit-variance, i.e.,

$$P_{z_n}(z) \to \int_{-\infty}^{z} N(x; 0, 1)\, dx \qquad \text{as } n \to \infty.$$

A couple of points are worth emphasizing. First, the somewhat exotic constraint (1.124) essentially ensures that no one term dominates the sum (1.125).[7] In

[7] To see that this constraint is critical, it suffices to consider the sequence of independent Bernoulli random variables x_i, each of which is $\pm 1/2^i$ with equal probability. Note that this sequence does not satisfy (1.124). For this sequence, it is straightforward to verify using (1.36) that the distribution of the normalized sum (1.125) converges to a uniform rather than Gaussian distribution.

fact, a simpler special case of this theorem corresponds to the x_i being identically-distributed and having a finite common variance. Second, it is important to emphasize that the theorem guarantees convergence in distribution but not in density. In fact, when the random variables in the sum are discrete, it is impossible to have convergence in density since arbitrary partial sums will be discrete!

1.6.2 Error Functions

In the context of many engineering applications of Gaussian random variables, we need to compute the area under the tail of a Gaussian density. In general, there is no closed-form expression for such quantities. However, the corresponding quantities for normalized Gaussian densities are often available numerically via tables or computer software packages.

In particular, if $x \sim N(0,1)$, then the standard $\mathcal{Q}$-function is defined according to $\mathcal{Q}(\alpha) = \Pr[x > \alpha]$, i.e.,

$$\mathcal{Q}(\alpha) \triangleq \frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\infty} e^{-x^2/2}\, dx. \tag{1.126}$$

This function, the area under the tail of the unit Gaussian (normal) density, is closely related to the so-called "complementary error function" $\operatorname{erfc}(\cdot)$ via

$$\mathcal{Q}(\alpha) = \frac{1}{2} \operatorname{erfc}\left(\frac{\alpha}{\sqrt{2}}\right). \tag{1.127}$$

This function is also well-tabulated, and can be evaluated, e.g., using the MATLAB function `erfc`. In calculations, it is often convenient to exploit the symmetry property

$$\mathcal{Q}(\alpha) = 1 - \mathcal{Q}(-\alpha) \tag{1.128}$$

and the bound

$$\mathcal{Q}(\alpha) \leq \frac{1}{2} e^{-\alpha^2/2} \tag{1.129}$$

valid for $\alpha > 0$. A variety of tighter bounds that are useful in a number of applications can also be developed.

Via a change of variables, the area under tails of other nonnormalized Gaussian densities is readily expressed in terms of the $\mathcal{Q}$-function. For example, if $x \sim N(m, \sigma^2)$, then $\tilde{x} = (x - m)/\sigma \sim N(0,1)$, so

$$\Pr[x > \alpha] = \Pr\left[\frac{x - m}{\sigma} > \frac{\alpha - m}{\sigma}\right] = \mathcal{Q}\left(\frac{\alpha - m}{\sigma}\right).$$

1.6.3 Gaussian Random Vectors

The notion of a Gaussian random vector is a powerful and important one, and builds on our notion of a Gaussian random variable. Specifically, an N-dimensional

random vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$$

is defined to be a Gaussian random vector, or equivalently $\{x_1, x_2, \ldots, x_N\}$ is defined to be a set of jointly Gaussian random variables when for all choices of the constant vector

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} \tag{1.130}$$

the scalar $y = \mathbf{a}^{\mathrm{T}}\mathbf{x}$ is a Gaussian random variable.

Gaussian random vectors have several important properties. In what follows, suppose $\mathbf{x}$ is a Gaussian random vector whose mean is $\mathbf{m}_{\mathbf{x}}$ and whose covariance matrix is $\mathbf{\Lambda}_{\mathbf{x}}$.

First, all subsets of $\{x_1, x_2, \ldots, x_N\}$ are jointly Gaussian. Deriving this result simply requires setting some of the a_i's in (1.130) to zero. As a special case of this result—corresponding to having only one nonzero component in (1.130)—we have that all the constituents must be individually Gaussian random variables, i.e.,

$$x_i \sim N(m_i, \lambda_{ii}) \qquad \text{for } i = 1, 2, \ldots, N$$

where

$$\lambda_{ii} = [\mathbf{\Lambda}_{\mathbf{x}}]_{ii}\,.$$

The characteristic function for a Gaussian random vector takes the form

$$M_{\mathbf{x}}(j\mathbf{v}) = \exp\left[j\mathbf{v}^{\mathrm{T}}\mathbf{m}_{\mathbf{x}} - \frac{1}{2}\mathbf{v}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{x}}\mathbf{v}\right]. \tag{1.131}$$

To prove (1.131), first note that

$$M_{\mathbf{x}}(j\mathbf{v}) = E\left[e^{j\mathbf{v}^{\mathrm{T}}\mathbf{x}}\right] = E\left[e^{j(v_1x_1+v_2x_2+\cdots+v_Nx_N)}\right] \tag{1.132}$$

Now for an arbitrary $\mathbf{a}$ let

$$y = \mathbf{a}^{\mathrm{T}}\mathbf{x}, \tag{1.133}$$

which from the definition of a Gaussian random vector means that y is a Gaussian random variable; specifically, $y \sim N(\mathbf{a}^{\mathrm{T}}\mathbf{m}_{\mathbf{x}}, \mathbf{a}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{x}}\mathbf{a})$. Then

$$M_y(jv) = E\left[e^{jvy}\right] = E\left[e^{j(va_1x_1+va_2x_2+\cdots+va_Nx_N)}\right] \tag{1.134}$$

Combining (1.132) with (1.134) we obtain

$$M_y(jv) = M_{\mathbf{x}}(jv\mathbf{a}), \tag{1.135}$$

while combining (1.134) with (1.122) we obtain

$$M_y(jv) = \exp\left[j(v\mathbf{a}^{\mathrm{T}})\mathbf{m}_{\mathbf{x}} - \frac{1}{2}(v\mathbf{a}^{\mathrm{T}})\mathbf{\Lambda}_{\mathbf{x}}(v\mathbf{a})\right]. \tag{1.136}$$

Finally, equating (1.135) and (1.136), and choosing $\mathbf{a} = \mathbf{v}/v$ we obtain our desired result (1.131).

Having derived the characteristic function of a Gaussian random vector, we now turn our attention to the corresponding density. When $|\mathbf{\Lambda}_{\mathbf{x}}| = \det(\mathbf{\Lambda}_{\mathbf{x}}) > 0$ (i.e., the nondegenerate case), the density function for a Gaussian random vector takes the following form

$$\begin{aligned} p_{\mathbf{x}}(\mathbf{x}) &= \frac{\exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{m}_{\mathbf{x}})^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{x}}^{-1}(\mathbf{x}-\mathbf{m}_{\mathbf{x}})\right]}{(2\pi)^{N/2}|\mathbf{\Lambda}_{\mathbf{x}}|^{1/2}} && (1.137)\\ &\triangleq N(\mathbf{x};\mathbf{m}_{\mathbf{x}},\mathbf{\Lambda}_{\mathbf{x}}) \\ &= |\mathbf{\Lambda}_{\mathbf{x}}|^{-1/2}N(\mathbf{\Lambda}_{\mathbf{x}}^{-1/2}(\mathbf{x}-\mathbf{m}_{\mathbf{x}});\mathbf{0},\mathbf{I}) && (1.138)\end{aligned}$$

where $\mathbf{\Lambda}_{\mathbf{x}}^{-1}$ is the inverse matrix associated with $\mathbf{\Lambda}_{\mathbf{x}}$, and where $\mathbf{\Lambda}_{\mathbf{x}}^{1/2}$ is the positive definite square root matrix of $\mathbf{\Lambda}_{\mathbf{x}}$, i.e., as discussed in Appendix 1.A, the (unique) matrix satisfying[8]

$$\mathbf{\Lambda}_{\mathbf{x}}^{1/2} = \left[\mathbf{\Lambda}_{\mathbf{x}}^{1/2}\right]^{\mathrm{T}} > 0 \tag{1.139a}$$

$$\mathbf{\Lambda}_{\mathbf{x}}^{1/2}\mathbf{\Lambda}_{\mathbf{x}}^{1/2} = \mathbf{\Lambda}_{\mathbf{x}}. \tag{1.139b}$$

The density (1.137) can be obtained via direct computation of the inverse Fourier transform of (1.131); a derivation is as follows. First, observe that

$$\begin{aligned} p_{\mathbf{x}}(\mathbf{x}) &= \frac{1}{(2\pi)^N}\int_{-\infty}^{+\infty} M_{\mathbf{x}}(j\mathbf{v})e^{-j\mathbf{v}^{\mathrm{T}}\mathbf{x}}\,d\mathbf{v} \\ &= \frac{1}{(2\pi)^N}\int_{-\infty}^{+\infty} \exp\left[-j\mathbf{v}^{\mathrm{T}}(\mathbf{x}-\mathbf{m}_{\mathbf{x}}) - \frac{1}{2}\mathbf{v}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{x}}\mathbf{v}\right]\,d\mathbf{v}. && (1.140)\end{aligned}$$

Then, using the change of variables $\mathbf{u} = \mathbf{\Lambda}_{\mathbf{x}}^{1/2}\mathbf{v}$ with $\mathbf{\Lambda}_{\mathbf{x}}^{1/2}$ as defined in (1.139), and noting that the Jacobian of the transformation is, using (1.139b), $|d\mathbf{u}/d\mathbf{v}| = |\mathbf{\Lambda}_{\mathbf{x}}^{1/2}| = |\mathbf{\Lambda}_{\mathbf{x}}|^{1/2}$, we can rewrite (1.140) as

$$p_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^N}\int_{-\infty}^{+\infty} \exp\left[-j\mathbf{u}^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{x}}^{-1/2}(\mathbf{x}-\mathbf{m}_{\mathbf{x}}) - \frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{u}\right]\,\frac{d\mathbf{u}}{|\mathbf{\Lambda}_{\mathbf{x}}|^{1/2}},$$

which when we adopt the convenient notation

$$\tilde{\mathbf{x}} = \mathbf{\Lambda}_{\mathbf{x}}^{-1/2}(\mathbf{x}-\mathbf{m}_{\mathbf{x}}) \tag{1.141}$$

[8]We have used $\mathbf{\Lambda}_{\mathbf{x}}^{-1/2}$ to denote the inverse of this square root matrix. Incidently, it is straightforward to verify that this matrix is also the positive definite square root matrix of $\mathbf{\Lambda}_{\mathbf{x}}^{-1}$.

and complete the square in the exponential yields

$$p_{\mathsf{x}}(\mathbf{\Lambda}_{\mathsf{x}}^{1/2}\tilde{\mathbf{x}} + \mathbf{m}_{\mathsf{x}}) = \frac{1}{(2\pi)^N |\mathbf{\Lambda}_{\mathsf{x}}|^{1/2}} \exp\left[-\frac{1}{2}\tilde{\mathbf{x}}^{\mathrm{T}}\tilde{\mathbf{x}}\right] \int_{-\infty}^{+\infty} \exp\left[-\frac{1}{2}(\mathbf{u} + j\tilde{\mathbf{x}})^{\mathrm{T}}(\mathbf{u} + j\tilde{\mathbf{x}})\right] d\mathbf{u}$$
$$= \frac{1}{(2\pi)^N |\mathbf{\Lambda}_{\mathsf{x}}|^{1/2}} \exp\left[-\frac{1}{2}\tilde{\mathbf{x}}^{\mathrm{T}}\tilde{\mathbf{x}}\right] \prod_{i=1}^{N} \sqrt{2\pi} \int_{-\infty}^{+\infty} N(u_i; -j\tilde{x}_i; 1)\, du_i \tag{1.142}$$

Finally, recognizing that each of the integrals in (1.142) is unity, and replacing $\tilde{\mathbf{x}}$ with its definition (1.141) we obtain, after some simple manipulations, our desired result (1.137).

Other Properties of Gaussian Random Vectors

Several additional important properties of Gaussian random vectors are worth developing. First, a pair of jointly Gaussian random vectors x and y are independent if and only if they are uncorrelated. We established the "only if" part for any pair of random vectors earlier. To establish the "if" part, let

$$\mathsf{z} = \begin{bmatrix} \mathsf{x} \\ \mathsf{y} \end{bmatrix} \tag{1.143}$$

so that $\mathsf{z} \sim N(\mathbf{m}_{\mathsf{z}}, \mathbf{\Lambda}_{\mathsf{z}})$, with

$$\mathbf{m}_{\mathsf{z}} = \begin{bmatrix} \mathbf{m}_{\mathsf{x}} \\ \mathbf{m}_{\mathsf{y}} \end{bmatrix} \tag{1.144}$$

$$\mathbf{\Lambda}_{\mathsf{z}} = \begin{bmatrix} \mathbf{\Lambda}_{\mathsf{x}} & \mathbf{\Lambda}_{\mathsf{xy}} \\ \mathbf{\Lambda}_{\mathsf{yx}} & \mathbf{\Lambda}_{\mathsf{y}} \end{bmatrix}. \tag{1.145}$$

Then when x and y are uncorrelated, i.e., when $\mathbf{\Lambda}_{\mathsf{xy}} = \mathbf{0}$, we have

$$\det \mathbf{\Lambda}_{\mathsf{z}} = \det \mathbf{\Lambda}_{\mathsf{x}} \det \mathbf{\Lambda}_{\mathsf{y}} \tag{1.146}$$

and

$$\mathbf{\Lambda}_{\mathsf{z}}^{-1} = \begin{bmatrix} \mathbf{\Lambda}_{\mathsf{x}}^{-1} & 0 \\ 0 & \mathbf{\Lambda}_{\mathsf{y}}^{-1} \end{bmatrix}. \tag{1.147}$$

Using the expression (1.137) for the Gaussian density with these results, one can easily check that in this case

$$p_{\mathsf{z}}(\mathbf{z}) = p_{\mathsf{x},\mathsf{y}}(\mathbf{x}, \mathbf{y}) = p_{\mathsf{x}}(\mathbf{x})\, p_{\mathsf{y}}(\mathbf{y}). \tag{1.148}$$

Second, linear transformations of Gaussian random vectors always produce Gaussian random vectors. To see this, let $\mathbf{A}$ be an arbitrary $m \times n$ matrix and let $\mathsf{y} = \mathbf{A}\mathsf{x}$ where x is a Gaussian random vector. Then y is a Gaussian random vector provided $z = \mathbf{b}^{\mathrm{T}}\mathsf{y}$ is a Gaussian random variable for every $\mathbf{b}$. But $z = \tilde{\mathbf{b}}^{\mathrm{T}}\mathsf{x}$ where

$\tilde{\mathbf{b}} = \mathbf{A}^{\mathrm{T}}\mathbf{b}$. Hence, since x is a Gaussian random vector, z is indeed a Gaussian random variable.

Note that as an immediate corollary to the last result we have that x and y are also jointly Gaussian random vectors. To see this, it suffices to observe that

$$\begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix} = \hat{\mathbf{A}}\mathsf{x} = \begin{bmatrix}\mathbf{I}\\ \mathbf{A}\end{bmatrix}\mathsf{x}.$$

Third, for jointly Gaussian random vectors x and y the conditional distribution for x given $\mathsf{y} = \mathbf{y}$ is also Gaussian with mean

$$\mathbf{m}_{\mathsf{x}|\mathsf{y}}(\mathbf{y}) = \mathbf{m}_{\mathsf{x}} + \mathbf{\Lambda}_{\mathsf{xy}}\mathbf{\Lambda}_{\mathsf{y}}^{-1}(\mathbf{y} - \mathbf{m}_{\mathsf{y}}) \tag{1.149}$$

and covariance[9]

$$\mathbf{\Lambda}_{\mathsf{x}|\mathsf{y}}(\mathbf{y}) = \mathbf{\Lambda}_{\mathsf{x}} - \mathbf{\Lambda}_{\mathsf{xy}}\mathbf{\Lambda}_{\mathsf{y}}^{-1}\mathbf{\Lambda}_{\mathsf{xy}}^{\mathrm{T}}. \tag{1.150}$$

A particularly straightforward proof of this result will appear later in our discussion of optimal estimation of random vectors in Chapter 3. Also, note that consistent with our preceding discussions, if $\mathbf{\Lambda}_{\mathsf{xy}} = \mathbf{0}$ then $\mathbf{m}_{\mathsf{x}|\mathsf{y}}(\mathbf{y}) = \mathbf{m}_{\mathsf{x}}$ and $\mathbf{\Lambda}_{\mathsf{x}|\mathsf{y}}(\mathbf{y}) = \mathbf{\Lambda}_{\mathsf{x}}$.

Finally, we stress that a Gaussian random vector is completely determined by its mean and covariance. For example, suppose that x and y are independent Gaussian random vectors and consider

$$\mathsf{z} = \mathbf{A}\mathsf{x} + \mathbf{B}\mathsf{y} + \mathbf{b}. \tag{1.151}$$

Then, since as we've shown Gaussianity is preserved under linear operations, we know that z is Gaussian. Consequently, in order to specify its density completely, we need only calculate its mean and covariance. As we saw in (1.108) and (1.110), these computations are also straightforward; we repeat them here for convenience:

$$\begin{aligned}\mathbf{m}_{\mathsf{z}} &= \mathbf{A}\mathbf{m}_{\mathsf{x}} + \mathbf{B}\mathbf{m}_{\mathsf{y}} + \mathbf{b}\\ \mathbf{\Lambda}_{\mathsf{z}} &= \mathbf{A}\mathbf{\Lambda}_{\mathsf{x}}\mathbf{A}^{\mathrm{T}} + \mathbf{B}\mathbf{\Lambda}_{\mathsf{y}}\mathbf{B}^{\mathrm{T}}.\end{aligned}$$

Since the mean vector and covariance matrix fully parameterize the density of a collection of jointly Gaussian random variables, this means that all moments can be expressed as functions of the mean and covariance. Moreover, in the Gaussian case, these moments can be computed extremely efficiently. To see this, let $\{x_1, x_2, \ldots, x_N\}$ be a set of jointly Gaussian random variables with mean values $\bar{x}_i$ and covariances $\lambda_{ij} = \mathrm{cov}\,(x_i, x_j)$, $1 \le i, j \le N$. In addition, for convenience define

$$\tilde{x}_i = x_i - \bar{x}_i.$$

Then for any set of integers $i_1, i_2, \ldots, i_L$ selected from $\{1, 2, \ldots, N\}$—with repetition allowed—it follows that

$$E\left[\tilde{x}_{i_1}\tilde{x}_{i_2}\cdots\tilde{x}_{i_L}\right] = \begin{cases} 0 & L \text{ odd}\\ \sum \lambda_{j_1 j_2}\lambda_{j_3 j_4}\ldots\lambda_{j_{L-1}j_L} & L \text{ even}\end{cases} \tag{1.152}$$

[9]Note from (1.150) that $\mathbf{\Lambda}_{\mathsf{x}|\mathsf{y}}(\mathbf{y})$ is a constant matrix, i.e., independent of the value of $\mathbf{y}$.

where the summation in (1.152) is over all distinct pairings $\{j_1, j_2\}$, $\{j_3, j_4\}, \ldots,$ $\{j_{L-1}, j_L\}$ of the set of symbols $\{i_1, i_2, \ldots, i_L\}$. Although we won't develop it here, this result may be derived in a relatively straightforward manner using, e.g., a Taylor series expansion of $M_{\mathbf{x}}(j\mathbf{v})$. As an example application of (1.152) we have

$$E\left[\tilde{x}_{i_1}\tilde{x}_{i_2}\tilde{x}_{i_3}\tilde{x}_{i_4}\right] = \lambda_{i_1 i_2}\lambda_{i_3 i_4} + \lambda_{i_1 i_3}\lambda_{i_2 i_4} + \lambda_{i_1 i_4}\lambda_{i_2 i_3} \tag{1.153}$$

so that

$$E\left[\tilde{x}_1\tilde{x}_2\tilde{x}_3\tilde{x}_4\right] = \lambda_{12}\lambda_{34} + \lambda_{13}\lambda_{24} + \lambda_{14}\lambda_{23} \tag{1.154}$$

$$E\left[\tilde{x}_1^2\tilde{x}_2^2\right] = \lambda_{11}\lambda_{22} + 2\lambda_{12}^2 \tag{1.155}$$

$$E\left[\tilde{x}_1^4\right] = 3\lambda_{11}^2. \tag{1.156}$$

As a final remark, we point out that the detailed shape of the contours of equiprobability for the multidimensional Gaussian density can be directly deduced from geometry of the covariance matrix as developed in Section 1.5.4. In particular, from (1.137) we see that the contours of equiprobability are the N-dimensional ellipsoids defined by

$$(\mathbf{x} - \mathbf{m}_{\mathbf{x}})^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{x}}^{-1}(\mathbf{x} - \mathbf{m}_{\mathbf{x}}) = \text{constant}. \tag{1.157}$$

From this perspective the transformation (1.116) from $\mathbf{x}$ to $\mathbf{z}$ corresponds to a generalized coordinate rotation (i.e., length preserving transformation) such that the components of $\mathbf{z}$ represent the principal or major axes of this ellipsoid, i.e.,

$$\frac{(z_1 - m_{z_1})^2}{\lambda_1} + \frac{(z_2 - m_{z_2})^2}{\lambda_2} + \cdots + \frac{(z_N - m_{z_N})^2}{\lambda_N} = \text{constant}.$$

Note that the $\lambda_i = \operatorname{var} z_i$ describe the proportions of the ellipsoid: they correspond to (squares of) the relative lengths along the principal axes. Note too that since $\mathbf{z}$ is Gaussian, its components are not only uncorrelated but mutually independent random variables.

We conclude this section by specializing our results to the case of two-dimensional Gaussian random vectors, where we let

$$\mathbf{m}_{\mathbf{x}} = \begin{bmatrix} m_1 \\ m_2 \end{bmatrix} \qquad \mathbf{\Lambda}_{\mathbf{x}} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}. \tag{1.158}$$

Here

$$p_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^{N/2}|\mathbf{\Lambda}_{\mathbf{x}}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{m}_{\mathbf{x}})^{\mathrm{T}}\mathbf{\Lambda}_{\mathbf{x}}^{-1}(\mathbf{x} - \mathbf{m}_{\mathbf{x}})\right] \tag{1.159}$$

$$= \frac{\exp\left[-\frac{(x_1-m_1)^2\sigma_2^2 - 2(x_1-m_1)(x_2-m_2)\rho\sigma_1\sigma_2 + (x_2-m_2)^2\sigma_1^2}{2\sigma_1^2\sigma_2^2(1-\rho^2)}\right]}{2\pi\sigma_1\sigma_2(1-\rho^2)^{1/2}} \tag{1.160}$$

Fig. 1.2 depicts the joint density of a pair of Gaussian random variables. In Fig. 1.3 we have plotted the associated contours of constant values of $p_{\mathbf{x}}(\mathbf{x})$ which are the ellipses

$$(x_1 - m_1)^2\sigma_2^2 - 2(x_1 - m_1)(x_2 - m_2)\rho\sigma_1\sigma_2 + (x_2 - m_2)^2\sigma_1^2 = \text{constant}. \tag{1.161}$$

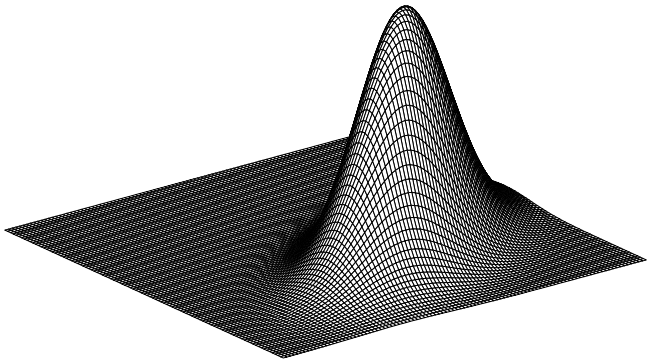

Figure 1.2. The two-dimensional probability density function of a pair of jointly Gaussian random variables.

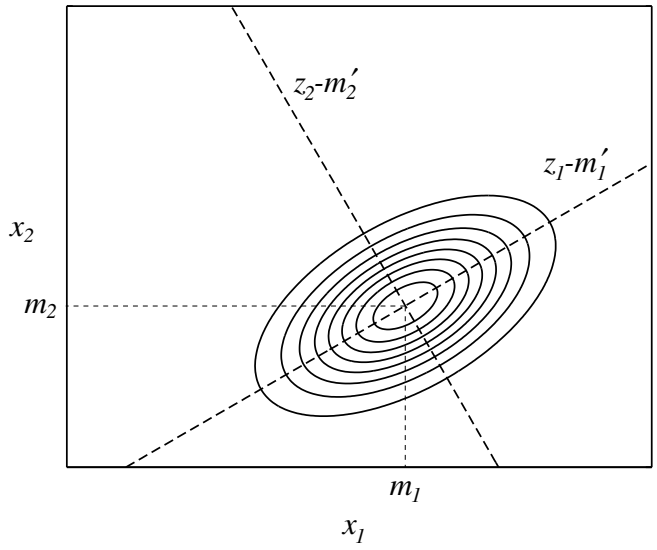

Figure 1.3. The contours of equiprobability corresponding to the density of Fig. 1.2.

As indicated in the figure, the components of $\mathbf{z}$ define the principal axes of the ellipses in (1.161), i.e., this equation in the transformed coordinates becomes

$$\frac{(z_1 - m_1')^2}{\lambda_1} + \frac{(z_2 - m_2')^2}{\lambda_2} = \text{constant}$$

where λ_1 and λ_2 are the eigenvalues of $\mathbf{\Lambda_x}$ and where

$$\mathbf{m_z} = \begin{bmatrix} m_1' \\ m_2' \end{bmatrix} = \mathbf{P}\mathbf{m_x}.$$

The larger $|\rho|$ is, the more eccentric these ellipses become, degenerating to lines when $|\rho| = 1$.

1.7 ABSTRACT VECTOR SPACE, AND SPACES OF RANDOM VARIABLES

The notion of a vector space is very powerful, and one that we will exploit on numerous occasions throughout the course. Clearly, we've already used certain

vector space ideas in preceding sections exploiting results from Appendix 1.A. In particular, we've exploited properties of the Euclidean space $\mathbb{R}^N$ consisting of N-dimensional vectors. However, while Euclidean space is an important example of a vector space, there are in fact many other somewhat more abstract vector spaces that turn out to be at least as important to us in this course. Although more abstract, many properties carry over from the Euclidean case, and you will often be able to rely on the geometric picture and intuition you have developed for this case.

Most generally, a vector space is a collection of elements or objects satisfying certain properties. This collection of elements may indeed consist of vectors $\mathbf{x}$ as we usually think of them, or they may be other kinds of objects like whole sequences $x[n]$ or functions $x(t)$, or even random variables $x(\omega)$. To avoid a conceptual bias, we'll just use the generic notation x for one such element.

For our purposes, vector spaces are special classes of *metric spaces*—i.e., spaces in which there is some notion of distance between the various elements in the collection.[10] A metric space is described by the pair $(\mathcal{S}, d(\cdot,\cdot))$ where $\mathcal{S}$ is the collection of elements and $d(\cdot,\cdot)$ is referred to as the metric. It is a measure of distance between an arbitrary pair of elements in the set; in particular $d(x,y)$ is the distance between elements x and y in $\mathcal{S}$.

For a metric to be useful, it must satisfy certain key properties that are consistent with our intuition about what distance is. In particular, we must have, for any elements x, y, and z in $\mathcal{S}$,

$$d(x,y) \geq 0 \tag{1.162}$$

$$d(x,y) = 0 \Leftrightarrow x = y \tag{1.163}$$

$$d(x,y) = d(y,x) \tag{1.164}$$

$$d(x,y) \leq d(x,z) + d(z,y) \tag{1.165}$$

The last of these, i.e., (1.165), is referred to as the triangle inequality.

An obvious (but not unique) example of a metric in $\mathbb{R}^N$ is the usual Euclidean distance

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{n=1}^{N}(x_n - y_n)^2}. \tag{1.166}$$

where x_n and y_n are the nth elements of $\mathbf{x}$ and $\mathbf{y}$, respectively. You can verify that (1.166) satisfies (1.162)–(1.165).

The metric spaces we're usually interested in have additional structure.

First, we want to work with spaces that are *complete*. While the technical definition is beyond the scope of our treatment here, in essence completeness means

[10] As a note of caution, our use of the term "vector space" is not universal. Some references consider the term to be equivalent to the term "linear space." However, as will become apparent, we will find it convenient to define vector spaces as linear spaces that are also metric spaces.

the metric space has no "holes." An example of a metric space that *isn't* complete is $\mathcal{S} = (0, 1] \subset \mathbb{R}$ with $d(x, y) = |x-y|$. Note that the sequence of elements $x_n = 1/n$ for $n = 1, 2, \ldots$, are all in the space $\mathcal{S}$, but $\lim_{n\to\infty} x_n = 0$ is not. The sequence x_n is an example of what is called a *Cauchy sequence*, and for a metric space to be complete, *all* such Cauchy sequences must converge to an element of $\mathcal{S}$.

A vector space $\mathcal{V}$ is a metric space that is *linear*. In order to talk about linearity, we'll need to define addition and scalar multiplication operators for objects in $\mathcal{V}$. For the cases of interest to us, we'll be using the usual definitions of these operators. We say $\mathcal{V}$ is a vector space if the following two properties hold:[11]

$$x, y \in \mathcal{V} \Rightarrow x + y \in \mathcal{V} \tag{1.167}$$

$$x \in \mathcal{V}, \alpha \in \mathbb{R} \Rightarrow \alpha x \in \mathcal{V}. \tag{1.168}$$

There are lots of important examples of vector spaces. First, there is the usual Euclidean space $\mathbb{R}^N$ composed of N-dimensional vectors $\mathbf{x}$. There is also the space of sequences $x[n]$ with finite energy

$$\sum_{n=-\infty}^{\infty} x^2[n] < \infty$$

which is usually denoted $\ell^2(\mathbb{Z})$, and the space of (integrable) functions $x(t)$ with finite energy

$$\int_{-\infty}^{+\infty} x^2(t)\, dt < \infty$$

which is usually denoted $L^2(\mathbb{R})$. And there is the space of random variables $\mathsf{x}(\omega)$ with finite mean-square

$$\operatorname{var} \mathsf{x} = E\left[\mathsf{x}^2\right] = \int_{-\infty}^{+\infty} x^2\, p_{\mathsf{x}}(x)\, dx < \infty,$$

which is usually denoted $L^2(\Omega)$.[12]

1.7.1 Linear Subspaces

Subspaces are an important concept associated with vector space. A subspace is a vector space that lies within another vector space, i.e., a subset $\mathcal{W} \subset \mathcal{V}$ is a subspace

[11] When the scalar α is restricted to be a real number as (1.168) indicates, the result is referred to as a real vector space; when it can be a complex number, i.e., $\alpha \in \mathbb{C}$, the result is a complex vector space. Although we will largely focus on the former class in this course to simplify our development, we remark in advance that we will sometimes need to work with complex vector spaces. Fortunately, however, there are no significant conceptual differences between the two.

[12] Incidently, for every probability space, there is an associated vector space of such random variables.

if $\mathcal{W}$ is itself a vector space.[13] As an example if $\mathcal{V}$ is the plane $\mathbb{R}^2$, then the line

$$\mathcal{W} = \{(x, y) \in \mathbb{R}^2 \mid y = x\}$$

is a subspace.

1.7.2 Linear Transformations

A linear transformation $L(\cdot)$ is a linear mapping from one vector space $\mathcal{V}$ to another vector space $\mathcal{U}$. This means that the powerful *principle of superposition* is satisfied, i.e., if x_k for $k = 1, 2, \ldots, K$ are each elements of $\mathcal{V}$, and if α_k for $k = 1, 2, \ldots, K$ are scalars, then

$$L(\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_K x_K) = \alpha_1 y_1 + \alpha_2 y_2 + \cdots + \alpha_K y_K$$

where $y_k = L(x_k)$.

When the vector spaces are the familiar Euclidean spaces, e.g., $\mathcal{V} = \mathbb{R}^N$ and $\mathcal{U} = \mathbb{R}^M$, then $L(\cdot)$ is represented by a matrix, i.e.,

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

where $\mathbf{A}$ is an $M \times N$-dimensional matrix. Several properties of matrices are developed in Appendix 1.A.

1.7.3 Linear Independence

A set of elements $x_1, x_2, \ldots, x_K$ in a vector space $\mathcal{V}$ are said to be linearly independent when

$$\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_K x_K = 0 \Leftrightarrow \alpha_1 = \alpha_2 = \cdots = \alpha_K = 0. \tag{1.169}$$

From (1.169) we see that linear dependency implies that the set of elements is redundant, i.e., that some of the elements can be expressed as a linear combination of the others. For example, if (1.169) does not hold, then assuming $\alpha_1 \neq 0$ we have

$$x_1 = \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_K x_K$$

where $\beta_k = -\alpha_k/\alpha_1$.

1.7.4 Bases

A basis for $\mathcal{V}$ is a linearly independent set of elements in $\mathcal{V}$ that *span* $\mathcal{V}$. A set of elements $x_1, x_2, \ldots, x_K$ is said to *span* $\mathcal{V}$ if *any* element $x \in \mathcal{V}$ can be represented as a linear combination of of the elements in this set, i.e., there exist α_k for $k = 1, 2, \ldots, K$ such that

$$x = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_K x_K.$$

[13] Technically, the mathematical notion of a subspace is more general. The definition we provide is of a specifically *linear* subspace, which is the only type of interest to us.

Note that when this set is linearly independent, the α_k must be unique.

All bases for a vector space have the same cardinality. This cardinality is referred to as the *dimension* of the vector space. As you've seen, the Euclidean space $\mathcal{V} = \mathbb{R}^N$ has dimension N. Hence, these spaces are finite-dimensional. Other spaces, like the space of finite-energy sequences $\ell^2(\mathbb{Z})$ and the space of finite-energy functions $L^2(\mathbb{R})$ are infinite-dimensional. The space of finite mean-square random variables $L^2(\Omega)$ is not only infinite-dimensional, but its dimension is uncountable (unless the probability space is discrete)! Although infinite-dimensional spaces are difficult to visualize, much intuition from finite-dimensional Euclidean space carries over. Furthermore, in many problems involving these vector spaces, we will often work with finite-dimensional subspaces, for which our geometric pictures are well-developed.

Let us continue to add more geometric structure to our notion of vector space.

1.7.5 Normed Vector Spaces

A *normed* vector space is a special kind of vector space for which the concept of *length* is defined for elements of the space. Let us use $\|x\|$ to denote the length or *norm* of each $x \in \mathcal{V}$, so a normed vector space is defined by specifying the pair $(\mathcal{V}, \|\cdot\|)$. In order for a function $\|\cdot\|$ to make sense as a norm on $\mathcal{V}$ it must satisfy certain properties. In particular, for $x \in \mathcal{V}$ and α an arbitrary scalar, it must satisfy:

$$\|x\| \geq 0 \tag{1.170}$$

$$\|x + y\| \leq \|x\| + \|y\| \tag{1.171}$$

$$\|\alpha x\| = |\alpha| \|x\| \tag{1.172}$$

$$\|x\| = 0 \Leftrightarrow x = 0 \tag{1.173}$$

Note that (1.171) is referred to as the triangle inequality for norms.

For normed vector spaces, the following rather natural metric can be defined: for x and y in $\mathcal{V}$,

$$d(x, y) = \|x - y\|. \tag{1.174}$$

It should be a straightforward exercise to verify that (1.174) satisfies the necessary properties of a metric, i.e., (1.162)–(1.165). Normed vector spaces that are complete in the sense we discussed earlier have a special and somewhat arcane name—they are referred to as *Banach* spaces.

As examples, $\mathbb{R}^N$, $\ell^2(\mathbb{Z})$, $L^2(\mathbb{R})$, and $L^2(\Omega)$ are all normed vector spaces. The corresponding norms are defined by, respectively,

$$\|\mathbf{x}\|^2 = \sum_{n=1}^{N} x_n^2$$
$$\|x[\cdot]\|^2 = \sum_{n} x^2[n]$$
$$\|x(\cdot)\|^2 = \int x^2(t)\,dt$$
$$\|\mathsf{x}(\cdot)\|^2 = E\left[\mathsf{x}^2\right].$$

Note however that there are many other norms one can define even for vectors in $\mathbb{R}^N$; for example,

$$\|\mathbf{x}\| = \max_{1\le n\le N} |x_n|.$$

Likewise, for functions $x(t)$, for any positive integer p,

$$\|x(\cdot)\| = \left(\int |x(t)|^p\,dt\right)^{1/p}$$

is a valid norm, and defines a whole family of normed vector spaces $L^p(\mathbb{R})$ parameterized by p. We emphasize that to fully specify a normed vector space we need both a collection of elements and a norm.

Ultimately, we're interested in normed vector spaces with even more structure, as we'll now develop.

1.7.6 Inner Product Spaces

An inner product space is a normed vector space where there is a notion of relative orientation or "angle" between elements. We use the notation $\langle x, y\rangle$ to denote the inner product between two elements x and y in $\mathcal{V}$. An inner product space is therefore defined by the pair $(\mathcal{V}, \langle\cdot,\cdot\rangle)$. An inner product defines the operation of *projection* of one element onto another. A valid inner product must satisfy the following properties[14]:

$$\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle \tag{1.175}$$
$$\langle \alpha x, y\rangle = \alpha\,\langle x, y\rangle \tag{1.176}$$
$$\langle x, y\rangle = \langle y, x\rangle \tag{1.177}$$
$$\langle x, x\rangle > 0 \Leftrightarrow x \neq 0. \tag{1.178}$$

[14]For simplicity, we'll restrict our attention to real-valued inner products even though many important examples are complex-valued.

For each inner product space there is a natural notion of norm. We call this the *induced* norm, and it is defined in terms of the inner product as follows:

$$\|x\| = \sqrt{\langle x, x \rangle}. \tag{1.179}$$

In turn, from the induced norm we get the associated metric

$$d(x, y) = \sqrt{\langle x - y, x - y \rangle}.$$

From the inner product and the induced norm, we arrive at a definition of the angle θ between two elements $x, y \in \mathcal{V}$. In particular, we have

$$\cos\theta = \frac{\langle x, y \rangle}{\|x\|\|y\|}. \tag{1.180}$$

One enormously useful inequality that applies to inner product spaces is the *Cauchy-Schwarz* inequality: for any x and y in $\mathcal{V}$,

$$|\langle x, y \rangle| \leq \|x\|\|y\| \tag{1.181}$$

A proof is as follows. For any α, we have, from the properties of a norm and (1.179),

$$\langle x - \alpha y, x - \alpha y \rangle = \|x - \alpha y\|^2 \geq 0. \tag{1.182}$$

Exploiting (1.175)–(1.178), we can rewrite the left hand side of (1.182) to get

$$\langle x - \alpha y, x - \alpha y \rangle = \|x\|^2 - 2\alpha \langle x, y \rangle + \alpha^2 \|y\|^2 \geq 0 \tag{1.183}$$

Then if we let $\alpha = \langle x, y \rangle / \langle y, y \rangle$, (1.183) becomes

$$\|x\|^2 - \frac{\langle x, y \rangle^2}{\|y\|^2} \geq 0 \tag{1.184}$$

which can be rewritten in the form (1.181). As a final comment, note from (1.182) and (1.178) that equality in (1.181) holds if and only if $x - \alpha y = 0$ for an arbitrary α.

Inner product spaces that are complete also have a special and arcane name—they are referred to as *Hilbert* spaces.

As examples, $\mathbb{R}^N$, $\ell^2(\mathbb{Z})$, $L^2(\mathbb{R})$, and $L^2(\Omega)$ are also all complete inner product (i.e., Hilbert) spaces. The corresponding inner products are defined by, respectively,

$$\begin{aligned}
\langle \mathbf{x}, \mathbf{y} \rangle &= \mathbf{x}^{\mathrm{T}} \mathbf{y} = \sum_{n=1}^{N} x_n y_n \\
\langle x[\cdot], y[\cdot] \rangle &= \sum_{n} x[n] y[n] \\
\langle x(\cdot), y(\cdot) \rangle &= \int x(t)\, y(t)\, dt \\
\langle \mathsf{x}(\cdot), \mathsf{y}(\cdot) \rangle &= E\left[\mathsf{x}\mathsf{y}\right].
\end{aligned}$$

Note that the Cauchy-Schwarz inequality for $L^2(\Omega)$ implies that

$$(E[xy])^2 \leq E[x^2]\ E[y^2] \tag{1.185}$$

with equality if and only if $x = \alpha y$ for some α, i.e., if and only if x and y are scaled versions of the same random variable. Likewise, for the subspace of $L^2(\Omega)$ consisting of zero-mean, finite-variance random variables, the specialization of (1.185) yields the following property of the correlation coefficient mentioned earlier in the chapter:

$$|\rho_{xy}| = \frac{|\mathrm{cov}(x,y)|}{\sqrt{\mathrm{var}\, x\, \mathrm{var}\, y}} \leq 1,$$

again with equality if and only if $x = \alpha y$ for some α.

1.7.7 Orthonormal Bases

With inner product spaces we have enough structure that we can finally talk about the concept of orthogonality. Specifically, we say that elements x and y in $\mathcal{V}$ are orthogonal, denoted $x \perp y$, when their inner product is zero, i.e.,

$$x \perp y \Leftrightarrow \langle x, y\rangle = 0. \tag{1.186}$$

In turn, we can talk about *orthogonal complements* of a subspace. In particular, if $\mathcal{W} \subset \mathcal{V}$ is a subspace of $\mathcal{V}$, then its orthogonal complement, denoted $\mathcal{W}^\perp$, is defined as follows:

$$\mathcal{W}^\perp = \{x \in \mathcal{V} : \langle x, y\rangle = 0, \text{ for all } y \in \mathcal{W}\}.$$

Note that $\mathcal{V}$ is the *direct sum* of $\mathcal{W}$ and $\mathcal{W}^\perp$, which we write as $\mathcal{V} = \mathcal{W} \oplus \mathcal{W}^\perp$. This means that every $x \in \mathcal{V}$ can be uniquely expressed as the sum $x = u + v$ where $u \in \mathcal{W}$ and $v \in \mathcal{W}^\perp$.

A collection of elements

$$\{x_1, x_2, \ldots, x_K\}$$

in $\mathcal{V}$ (with K possibly infinite) is said to be an *orthonormal* set if

$$\langle x_i, x_j\rangle = \delta[i-j] = \begin{cases} 1 & i = j \\ 0 & \text{otherwise} \end{cases}.$$

If this collection of elements is an orthonormal *basis* for $\mathcal{V}$ (equivalently referred to as a "complete orthonormal set"), then expansions in this basis are especially easy to compute. In particular (and consistent with our geometric intuition), if $x \in \mathcal{V}$ then we can write

$$x = \sum_{k=1}^{K} \alpha_k x_k$$

where the α_k are projections of x onto the basis functions x_k, i.e., $\alpha_k = \langle x, x_k\rangle$.

An important identity that applies to orthonormal bases is *Parseval's relation*. In particular, if $\{x_1, x_2, \ldots, x_K\}$ is an orthonormal basis for $\mathcal{V}$ and if x and y are arbitrary elements of $\mathcal{V}$, then

$$\langle x, y \rangle = \sum_{k=1}^{K} \alpha_k \beta_k$$

where $\alpha_k = \langle x, x_k \rangle$ and $\beta_k = \langle y, x_k \rangle$. A special case of Parseval's relation that corresponds to choosing $x = y$ is the *Plancherel formula*:

$$\|x\|^2 = \sum_{k=1}^{K} |\alpha_k|^2$$

where again $\alpha_k = \langle x, x_k \rangle$.

If you are interested in exploring the concept of an abstract vector space in more detail, a good starting point is, e.g., A. W. Naylor and G. R. Sell, *Linear Operator Theory in Engineering and Science*, Springer-Verlag, New York, 1982.

1.A LINEAR ALGEBRA AND EUCLIDEAN VECTOR SPACE

1.A.1 Vectors and Matrices

In this course vectors will be matrices that are specifically columns, and as such will be denoted by boldface lowercase characters; for example,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{1.187}$$

where $x_1, x_2, \ldots, x_n$ are either real or complex numbers. The set of all such n-dimensional vectors of real numbers is denoted by $\mathbb{R}^n$. The corresponding set of all n-dimensional vectors of complex numbers is denoted by $\mathbb{C}^n$. The *transpose* of a column vector $\mathbf{x}$ is the row vector

$$\mathbf{x}^{\mathrm{T}} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}. \tag{1.188}$$

Vector addition and scalar multiplication are defined componentwise, i.e.,

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{bmatrix} \tag{1.189}$$

and

$$\alpha \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \alpha x_1 \\ \alpha x_2 \\ \vdots \\ \alpha x_n \end{bmatrix} \tag{1.190}$$

where α is a real or complex number.

A set of vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_r$ in $\mathbb{R}^n$ is *linearly independent* if[15]

$$\alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2 + \cdots + \alpha_r \mathbf{x}_r = \mathbf{0} \tag{1.191}$$

implies that

$$\alpha_1 = \alpha_2 = \cdots = \alpha_r = 0. \tag{1.192}$$

Otherwise the set of vectors is said to be *linearly dependent*, and in this case one of the $\mathbf{x}_i$ can be written as a *linear combination* of the others. For example, if $\alpha_1 \neq 0$ in (1.191)

$$\mathbf{x}_1 = \beta_2 \mathbf{x}_2 + \cdots + \beta_r \mathbf{x}_r \tag{1.193}$$

with

$$\beta_2 = -\alpha_2/\alpha_1, \ldots, \beta_r = -\alpha_r/\alpha_1. \tag{1.194}$$

In $\mathbb{R}^n$ there exist sets of at most n linearly independent vectors. Any such set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ forms a *basis* for $\mathbb{R}^n$. That is, any $\mathbf{x} \in \mathbb{R}^n$ can be written as a linear combination of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$.

Matrices will in general be denoted by boldface uppercase characters. The element in the ith row and jth column of $\mathbf{A}$ will be denoted by a_{ij} or, alternatively, by $[\mathbf{A}]_{ij}$. If $\mathbf{A}$ is $m \times n$, i.e., if $\mathbf{A}$ has m rows and n columns, then

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}. \tag{1.195}$$

The set of all $m \times n$ real-valued matrices is denoted $\mathbb{R}^{m \times n}$. As with vectors, we define matrix addition and scalar multiplication componentwise. If $m = n$, $\mathbf{A}$ is a *square* matrix. The *transpose* of an $m \times n$ matrix $\mathbf{A}$ is the $n \times m$ matrix

$$\mathbf{A}^{\mathrm{T}} = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}. \tag{1.196}$$

[15]The symbol $\mathbf{0}$ denotes the matrix or vector of appropriate dimension, all of whose components are zero.

A square matrix is said to be *symmetric* if $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$. A *diagonal* square matrix is one of the form

$$\mathbf{A} = \begin{bmatrix} \mu_1 & 0 & \cdots & 0 \\ 0 & \mu_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mu_n \end{bmatrix} = \operatorname{diag}(\mu_1, \mu_2, \ldots, \mu_n) \tag{1.197}$$

where the last expression in (1.197) introduces notation that is sometimes convenient. The *identity matrix* is defined as

$$\mathbf{I} = \operatorname{diag}(1, 1, \ldots, 1). \tag{1.198}$$

On (rare) occasions when there is risk of ambiguity, we will write $\mathbf{I}_n$ to make explicit the size (i.e., $n \times n$) of the identity matrix. The *trace* of a square matrix $\mathbf{A}$ is the sum of its diagonal elements:

$$\operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}. \tag{1.199}$$

Let $\mathbf{A}$ be an $m \times n$ matrix and $\mathbf{B}$ an $n \times p$ matrix. Then we can define the product

$$\mathbf{C} = \mathbf{AB}. \tag{1.200}$$

Here $\mathbf{C}$ is an $m \times p$ matrix whose elements are given by

$$c_{ij} = \sum_{k=1}^{n} a_{ik}\, b_{kj}. \tag{1.201}$$

Note the required compatibility condition—the number of columns of $\mathbf{A}$ must equal the number of rows of $\mathbf{B}$ for $\mathbf{AB}$ to be defined. Note too that $\mathbf{BA}$ may not be defined even if $\mathbf{AB}$ is (e.g., let $m = 7$, $n = 4$, $p = 3$). Even if $\mathbf{BA}$ is defined it is generally not the same size as $\mathbf{AB}$. For example, if $\mathbf{A}$ is 2×4 and $\mathbf{B}$ is 4×2, then $\mathbf{AB}$ is 2×2, but $\mathbf{BA}$ is 4×4. If $\mathbf{A}$ and $\mathbf{B}$ are square and of the same size, then $\mathbf{AB}$ and $\mathbf{BA}$ are as well. In general, however, $\mathbf{AB} \neq \mathbf{BA}$. Note also that

$$\mathbf{AI} = \mathbf{IA} = \mathbf{A} \tag{1.202}$$

$$(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\, \mathbf{A}^{\mathrm{T}} \tag{1.203}$$

Also, if $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{x} \in \mathbb{R}^n$, then $\mathbf{Ax} \in \mathbb{R}^m$. In addition, if both $\mathbf{AB}$ and $\mathbf{BA}$ are defined,

$$\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA}). \tag{1.204}$$

Let $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$. Then the *dyadic* or *outer product* of $\mathbf{x}$ and $\mathbf{y}$ is the $n \times m$ matrix

$$\mathbf{xy}^{\mathrm{T}} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_m \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_m \\ \vdots & \vdots & \ddots & \vdots \\ x_n y_1 & x_n y_2 & \cdots & x_n y_m \end{bmatrix} \in \mathbb{R}^{n \times m}. \tag{1.205}$$

If $n = m$ we can also define the *dot* or *inner* product

$$\mathbf{x}^{\mathrm{T}}\mathbf{y} = \sum_{i=1}^{n} x_i y_i = \mathbf{y}^{\mathrm{T}}\mathbf{x} \in \mathbb{R} \tag{1.206}$$

Two n-vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal, denoted $\mathbf{x} \perp \mathbf{y}$, if

$$\mathbf{x}^{\mathrm{T}}\mathbf{y} = 0. \tag{1.207}$$

Note that a set of nonzero, mutually orthogonal vectors is linearly independent. The *length* or *norm* of $\mathbf{x} \in \mathbb{R}^n$ is

$$\|\mathbf{x}\| = (\mathbf{x}^{\mathrm{T}}\mathbf{x})^{1/2} = \left(x_1^2 + x_2^2 + \cdots + x_n^2\right)^{1/2} \tag{1.208}$$

Note, too, that

$$\|\mathbf{x}\|^2 = \mathbf{x}^{\mathrm{T}}\mathbf{x} = \mathrm{tr}(\mathbf{x}^{\mathrm{T}}\mathbf{x}) = \mathrm{tr}(\mathbf{x}\mathbf{x}^{\mathrm{T}}) \tag{1.209}$$

On occasion we will find it useful to deal with matrices written in block form, such as

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \tag{1.210}$$

where $\mathbf{A}_{11}$ and $\mathbf{A}_{12}$ have the same number of rows, and $\mathbf{A}_{11}$ and $\mathbf{A}_{21}$ have the same number of columns. The product of two matrices in block form is computed in a manner analogous to usual matrix multiplication. For example,

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{bmatrix} \tag{1.211}$$

where the blocks on the left side of (1.211) must be partitioned in a compatible fashion, and where the order of multiplication of the various terms on the right-hand side is important.

1.A.2 Matrix Inverses and Determinants

An $n \times n$ matrix $\mathbf{A}$ is *invertible* or *nonsingular* if there exists another $n \times n$ matrix $\mathbf{A}^{-1}$, called the *inverse* of $\mathbf{A}$, so that

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}. \tag{1.212}$$

If no such matrix exists $\mathbf{A}$ is said to be singular or, equivalently, noninvertible. Consider the set of equations

$$\mathbf{A}\mathbf{x} = \mathbf{y} \tag{1.213}$$

where $\mathbf{A}$ is $n \times n$. This equation has a unique solution $\mathbf{x}$ for any $\mathbf{y}$ if and only if $\mathbf{A}$ is invertible (in which case the solution is $\mathbf{A}^{-1}\mathbf{y}$). Consequently, the equation $\mathbf{A}\mathbf{x} = \mathbf{0}$ has a nonzero solution if and only if $\mathbf{A}$ is *not* invertible.

The determinant of a square matrix $\mathbf{A}$, denoted by $|\mathbf{A}|$ or $\det(\mathbf{A})$, can be computed recursively. If $\mathbf{A}$ is 1×1, then $|\mathbf{A}| = \mathbf{A}$. If $\mathbf{A}$ is $n \times n$, then we can

compute $|\mathbf{A}|$ by "expanding by minors" using any row or column. For example, using the ith row,

$$|\mathbf{A}| = a_{i1}A_{i1} + a_{i2}A_{i2} + \cdots + a_{in}A_{in}, \tag{1.214}$$

or, using the jth column,

$$|\mathbf{A}| = a_{1j}A_{1j} + a_{2j}A_{2j} + \cdots + a_{nj}A_{nj}, \tag{1.215}$$

where the *cofactors* A_{ij} are given by

$$A_{ij} = (-1)^{i+j} \det(\mathbf{M}_{ij}) \tag{1.216}$$

and where $\mathbf{M}_{ij}$ is the $(n-1) \times (n-1)$ matrix obtained from $\mathbf{A}$ by deleting the ith row and jth column.

As a simple example, we have

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}. \tag{1.217}$$

As a more complex example we have

$$\begin{aligned}
&\begin{vmatrix} 2 & 0 & 0 & 3 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 5 & 1 & 1 & 9 \end{vmatrix} \\
&= 2(-1)^{1+1}\begin{vmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 9 \end{vmatrix} + 0(-1)^{1+2}\begin{vmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 5 & 1 & 9 \end{vmatrix} \\
&\quad + 0(-1)^{1+3}\begin{vmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 5 & 1 & 9 \end{vmatrix} + 3(-1)^{1+4}\begin{vmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 5 & 1 & 1 \end{vmatrix} \\
&= 2 \cdot 1 \cdot (-1)^{1+1}\begin{vmatrix} 1 & 0 \\ 1 & 9 \end{vmatrix} - 3 \cdot 1 \cdot (-1)^{1+1}\begin{vmatrix} 1 & 1 \\ 1 & 1 \end{vmatrix} - 3 \cdot 1 \cdot (-1)^{1+2}\begin{vmatrix} 1 & 1 \\ 5 & 1 \end{vmatrix} \\
&= 2 \cdot 9 - 3 \cdot 0 + 3 \cdot (-4) = 6
\end{aligned}$$

Several useful properties of determinants are

$$|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}| \tag{1.218}$$

$$|\alpha\mathbf{A}| = \alpha^n|\mathbf{A}| \tag{1.219}$$

$$|\mathbf{A}^{\mathrm{T}}| = |\mathbf{A}| \tag{1.220}$$

$$|\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|} \tag{1.221}$$

The invertibility of a matrix $\mathbf{A}$ is equivalent to each of the following statements:

1. $|\mathbf{A}| \neq 0$

2. All of the columns of $\mathbf{A}$ are linearly independent.
3. All of the rows of $\mathbf{A}$ are linearly independent.

The inverse of $\mathbf{A}$ can be expressed as

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|} \operatorname{adj} \mathbf{A} \tag{1.222}$$

where $[\operatorname{adj} \mathbf{A}]_{ij} = A_{ji}$ is referred to as the *adjugate* or *adjoint* matrix, with A_{ij} as defined in (1.216). As a simple example, we have

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} \tag{1.223}$$

Some useful properties of inverses are

$$(\mathbf{A}^{\mathrm{T}})^{-1} = (\mathbf{A}^{-1})^{\mathrm{T}} \tag{1.224}$$

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1} \tag{1.225}$$

and

$$\mathbf{A} = \operatorname{diag}(\mu_1, \mu_2, \ldots, \mu_n) \Leftrightarrow \mathbf{A}^{-1} = \operatorname{diag}\left(\frac{1}{\mu_1}, \frac{1}{\mu_2}, \ldots, \frac{1}{\mu_n}\right). \tag{1.226}$$

A matrix $\mathbf{P}$ is said to be *orthogonal* if

$$\mathbf{P}^{-1} = \mathbf{P}^{\mathrm{T}}. \tag{1.227}$$

If we think of $\mathbf{P}$ as consisting of a set of columns, i.e.,

$$\mathbf{P} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{bmatrix} \tag{1.228}$$

then, in general,

$$\mathbf{P}^{\mathrm{T}}\mathbf{P} = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}}\mathbf{x}_1 & \mathbf{x}_1^{\mathrm{T}}\mathbf{x}_2 & \cdots & \mathbf{x}_1^{\mathrm{T}}\mathbf{x}_n \\ \mathbf{x}_2^{\mathrm{T}}\mathbf{x}_1 & \mathbf{x}_2^{\mathrm{T}}\mathbf{x}_2 & \cdots & \mathbf{x}_2^{\mathrm{T}}\mathbf{x}_n \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n^{\mathrm{T}}\mathbf{x}_1 & \mathbf{x}_n^{\mathrm{T}}\mathbf{x}_2 & \vdots & \mathbf{x}_n^{\mathrm{T}}\mathbf{x}_n \end{bmatrix}. \tag{1.229}$$

Consequently, we see that $\mathbf{P}$ is *orthogonal* if and only if its columns are *orthonormal*, i.e., if $\mathbf{x}_i \perp \mathbf{x}_j$ for $i \neq j$, and if $\|\mathbf{x}_i\| = 1$.

There are also some useful results for block matrices. For example, for a block diagonal matrix

$$\mathbf{A} = \operatorname{diag}(\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_r) \Leftrightarrow \mathbf{A}^{-1} = \operatorname{diag}(\mathbf{F}_1^{-1}, \mathbf{F}_2^{-1}, \ldots, \mathbf{F}_r^{-1}). \tag{1.230}$$

Also, we have the formulas

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1} =$$

$$\begin{bmatrix} (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} & -(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} & \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \end{bmatrix} \tag{1.231}$$

and

$$\det \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} = \left|\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right| |\mathbf{A}_{22}|, \tag{1.232}$$

which are valid if $\mathbf{A}_{22}$ is nonsingular. These formulas can be verified by exploiting the identity

$$\begin{bmatrix} \mathbf{I} & -\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix}. \tag{1.233}$$

If on the other hand $\mathbf{A}_{11}$ is nonsingular, then as an alternative to (1.231) we have

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \\ -(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \end{bmatrix}. \tag{1.234}$$

Other useful results are obtained by comparing (1.231) and (1.234). For example, equating the upper left blocks in these two expressions yields the useful identity

$$(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} = \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}. \tag{1.235}$$

1.A.3 Eigenvalues and Eigenvectors

Let $\mathbf{A}$ be an $n \times n$ real matrix. A scalar λ is called an *eigenvalue* of $\mathbf{A}$ with associated nonzero *eigenvector* $\mathbf{x}$ if

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}. \tag{1.236}$$

The above equation can be rewritten as

$$(\lambda\mathbf{I} - \mathbf{A})\mathbf{x} = \mathbf{0}. \tag{1.237}$$

Thus λ is an eigenvalue of $\mathbf{A}$ if and only if (1.237) has a solution $\mathbf{x} \neq \mathbf{0}$. This will be the case if and only if $\lambda\mathbf{I} - \mathbf{A}$ is singular, i.e., if and only if λ is a solution of the *characteristic equation*

$$\phi_{\mathbf{A}}(\lambda) = |\lambda\mathbf{I} - \mathbf{A}| = 0. \tag{1.238}$$

Here $\phi_{\mathbf{A}}(\lambda)$ is called the *characteristic polynomial* of $\mathbf{A}$ and is of the form

$$\phi_{\mathbf{A}}(\lambda) = \lambda^n + \alpha_{n-1}\lambda^{n-1} + \cdots + \alpha_1\lambda + \alpha_0 \tag{1.239}$$

$$= (\lambda - \lambda_1) \cdot (\lambda - \lambda_2) \cdots (\lambda - \lambda_n). \tag{1.240}$$

The $\lambda_1, \lambda_2, \ldots, \lambda_n$ in (1.240) are the n eigenvalues, which may or may not be distinct. Some of the λ_i may in general be complex, in which case they occur in

complex conjugate pairs. However, if $\mathbf{A}$ is symmetric, the λ_i are always real. Also note that

$$|\mathbf{A}| = (-1)^n \phi_{\mathbf{A}}(0) = (-1)^n \alpha_0 = \prod_{i=1}^{n} \lambda_i \tag{1.241}$$

so that $\mathbf{A}$ is invertible if and only if all of the eigenvalues of $\mathbf{A}$ are nonzero. In addition, one can show that

$$\mathrm{tr}(\mathbf{A}) = -\alpha_{n-1} = \sum_{i=1}^{n} \lambda_i. \tag{1.242}$$

If λ_i is an eigenvalue of $\mathbf{A}$, then we can determine an associated eigenvector $\mathbf{x}_i$ by solving the set of linear equations

$$\mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_i. \tag{1.243}$$

Note that if $\mathbf{x}_i$ is an eigenvector, so is $\alpha \mathbf{x}_i$ for any scalar α. Consequently, we can always adjust the length of the eigenvectors arbitrarily, and, in particular, we can normalize them to have unit length. It is also possible to show that each distinct λ_i has a linearly independent $\mathbf{x}_i$ corresponding to it. If, on the other hand, λ_i has multiplicity $k > 1$, i.e., if λ_i is a kth-order root of $\phi_{\mathbf{A}}(\lambda)$, then in general there may be anywhere from 1 to k linearly independent eigenvectors associated with λ_i. Note that we can always combine (1.243) for different values of i into one equation

$$\mathbf{A}\begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n\end{bmatrix} = \begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n\end{bmatrix}\begin{bmatrix}\lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n\end{bmatrix}. \tag{1.244}$$

If $\mathbf{A}$ is symmetric, some special properties result. First, eigenvectors corresponding to distinct eigenvalues are not only linearly independent, but *orthogonal*. Second, eigenvalues with multiplicity k have a full set of (i.e., k) linearly independent eigenvectors, which can also be chosen to be orthogonal to one another. Hence, symmetric matrices always have a full set of linearly independent eigenvectors that can be chosen so as to be *orthonormal* as well.

In general, an $n \times n$ matrix $\mathbf{A}$ that has a full set of n linearly independent eigenvectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ is called *diagonalizable*. For a diagonalizable matrix, the matrix of eigenvectors in (1.244), viz.,

$$\begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n\end{bmatrix} \triangleq \mathbf{P}^{-1}, \tag{1.245}$$

is nonsingular. Hence, we can write

$$\mathbf{P}\mathbf{A}\mathbf{P}^{-1} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n). \tag{1.246}$$

We emphasize that if $\mathbf{A}$ is a symmetric matrix we can choose the $\mathbf{x}_i$ to be orthonormal so that $\mathbf{P}^{-1} = \mathbf{P}^{\mathrm{T}}$, further simplifying manipulations. The matrix representation (1.246) will prove very useful. It is an example of a similarity transformation, which we briefly describe next.

1.A.4 Similarity Transformation

Let $\mathbf{A}$ be an $n \times n$ matrix, and let $\mathbf{P}$ be an invertible matrix of the same size. We can then define a *similarity transformation* of $\mathbf{A}$ as

$$\mathbf{B} = \mathbf{P}\mathbf{A}\mathbf{P}^{-1}. \tag{1.247}$$

We sometimes say that "$\mathbf{B}$ is similar to $\mathbf{A}$". A similarity transformation can be interpreted as arising out of a change of coordinates. To see this, suppose

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{1.248}$$

and consider the change of coordinates

$$\mathbf{u} = \mathbf{P}\mathbf{x} \tag{1.249}$$

$$\mathbf{v} = \mathbf{P}\mathbf{y}, \tag{1.250}$$

so that (since $\mathbf{x} = \mathbf{P}^{-1}\mathbf{u}$) each component of $\mathbf{u}$, for example, is a weighted sum of components of $\mathbf{x}$ and vice versa. Then

$$\mathbf{v} = \mathbf{B}\mathbf{u} \tag{1.251}$$

with $\mathbf{B}$ as given in (1.247). Furthermore,

$$\begin{aligned}\phi_{\mathbf{B}}(\lambda) &= |\lambda\mathbf{I} - \mathbf{B}| = |\lambda\mathbf{P}\mathbf{P}^{-1} - \mathbf{P}\mathbf{A}\mathbf{P}^{-1}| = |\mathbf{P}^{-1}(\lambda\mathbf{I} - \mathbf{A})\mathbf{P}| \\ &= |\mathbf{P}^{-1}||\lambda\mathbf{I} - \mathbf{A}||\mathbf{P}| = |\lambda\mathbf{I} - \mathbf{A}| = \phi_{\mathbf{A}}(\lambda)\end{aligned}$$

so the eigenvalues of $\mathbf{B}$ and $\mathbf{A}$ are the same. Thus by (1.241) and (1.242), $\mathbf{A}$ and $\mathbf{B}$ have the same determinant and trace, respectively.

1.A.5 Positive Definite Matrices

A symmetric square matrix $\mathbf{A}$ is *positive semidefinite*, written $\mathbf{A} \geq 0$, if and only if

$$\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} \geq 0 \tag{1.252}$$

for all vectors $\mathbf{x}$. This matrix $\mathbf{A}$ is *positive definite*, written $\mathbf{A} > 0$, if and only if

$$\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} > 0 \qquad \text{for any } \mathbf{x} \neq \mathbf{0}. \tag{1.253}$$

It is not difficult to see that a positive semidefinite matrix is positive definite if and only if it is invertible.

Some basic facts about positive semidefinite matrices are the following:

1. If $\mathbf{A} \geq 0$ and $\mathbf{B} \geq 0$, then $\mathbf{A} + \mathbf{B} \geq 0$, since
$$\mathbf{x}^{\mathrm{T}}(\mathbf{A} + \mathbf{B})\mathbf{x} = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{B}\mathbf{x} \tag{1.254}$$
2. If either $\mathbf{A}$ or $\mathbf{B}$ in (1.254) is positive definite, then so is $\mathbf{A} + \mathbf{B}$. This again follows from (1.254).

3. If $\mathbf{A} > 0$, then $\mathbf{A}^{-1} > 0$, since

$$\mathbf{x}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{x} = (\mathbf{A}^{-1}\mathbf{x})^{\mathrm{T}}\mathbf{A}(\mathbf{A}^{-1}\mathbf{x}) > 0 \qquad \text{if } \mathbf{x} \neq \mathbf{0} \tag{1.255}$$

4. If $\mathbf{Q} \geq 0$ then $\mathbf{F}^{\mathrm{T}}\mathbf{Q}\mathbf{F} \geq 0$ for *any* (not necessarily square) matrix $\mathbf{F}$ for which $\mathbf{F}^{\mathrm{T}}\mathbf{Q}\mathbf{F}$ is defined. This follows from

$$\mathbf{x}^{\mathrm{T}}(\mathbf{F}^{\mathrm{T}}\mathbf{Q}\mathbf{F})\mathbf{x} = (\mathbf{F}\mathbf{x})^{\mathrm{T}}\mathbf{Q}(\mathbf{F}\mathbf{x}) \geq 0 \tag{1.256}$$

5. If $\mathbf{Q} > 0$ and $\mathbf{F}$ is invertible, $\mathbf{F}^{\mathrm{T}}\mathbf{Q}\mathbf{F} > 0$. This also follows from (1.256).

One test for positive definiteness is *Sylvester's Test*. Let

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{12} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{bmatrix}. \tag{1.257}$$

Then $\mathbf{A}$ is positive definite if and only if the determinant of *every* upper left submatrix of $\mathbf{A}$ is positive, i.e.,[16]

$$\begin{aligned} a_{11} &> 0 \\ \begin{vmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{vmatrix} &> 0 \\ \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{vmatrix} &> 0 \\ &\text{etc.} \end{aligned} \tag{1.258}$$

Let $\mathbf{A}$ be symmetric and let $\mathbf{P}$ be the orthogonal matrix of eigenvectors so that [cf. (1.246)]

$$\mathbf{P}\mathbf{A}\mathbf{P}^{\mathrm{T}} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \triangleq \mathbf{\Lambda}. \tag{1.259}$$

Then, letting $\mathbf{z} = \mathbf{P}\mathbf{x}$, we have

$$\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} = \mathbf{x}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}(\mathbf{P}\mathbf{A}\mathbf{P}^{\mathrm{T}})\mathbf{P}\mathbf{x} = \mathbf{z}^{\mathrm{T}}\mathbf{\Lambda}\mathbf{z} = \lambda_1 z_1^2 + \lambda_2 z_2^2 + \cdots + \lambda_n z_n^2 \tag{1.260}$$

From this we can conclude that a symmetric matrix $\mathbf{A}$ is positive semidefinite (positive definite) if and only if all its eigenvalues are nonnegative (positive).

Another characterization of positive semidefinite matrices is in terms of their square root matrices. In particular, any $\mathbf{A} \geq 0$ has a *square root matrix* $\mathbf{F}$ such that

$$\mathbf{A} = \mathbf{F}^{\mathrm{T}}\mathbf{F}. \tag{1.261}$$

[16]Beware, however—there is no corresponding test for positive semidefiniteness that involves examining upper submatrices for nonnegative determinants. Consider, e.g., the matrix $\left[\begin{smallmatrix} 0 & 0 \\ 0 & -1 \end{smallmatrix}\right]$ which is not positive semidefinite.

Specifically, from (1.259) we see that we can take

$$\mathbf{F} = \sqrt{\mathbf{\Lambda}}\mathbf{P}. \tag{1.262}$$

where

$$\sqrt{\mathbf{\Lambda}} \triangleq \operatorname{diag}\left(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}\right). \tag{1.263}$$

Note that the $\mathbf{F}$ we choose in (1.262) is invertible if and only if $\mathbf{A} > 0$.

In general, the square root matrix as defined in (1.261) is not unique. For example, let $\mathbf{Q}$ be any orthogonal matrix, and let

$$\hat{\mathbf{F}} = \mathbf{Q}\mathbf{F} \tag{1.264}$$

Then $\hat{\mathbf{F}}$ is also a valid square root matrix for $\mathbf{A}$, i.e.,

$$\hat{\mathbf{F}}^{\mathrm{T}}\hat{\mathbf{F}} = \mathbf{F}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{Q}\mathbf{F} = \mathbf{F}^{\mathrm{T}}\mathbf{I}\mathbf{F} = \mathbf{F}^{\mathrm{T}}\mathbf{F} = \mathbf{A}. \tag{1.265}$$

However, choosing $\mathbf{Q} = \mathbf{P}^{\mathrm{T}}$ in (1.264) gives the positive semidefinite square root matrix

$$\hat{\mathbf{F}} = \mathbf{P}^{\mathrm{T}}\sqrt{\mathbf{\Lambda}}\mathbf{P}. \tag{1.266}$$

In fact, (1.266) is the *unique* positive semidefinite square root matrix associated with $\mathbf{A}$, and hence we will reserve the notation $\mathbf{A}^{1/2}$ for this particular matrix.

As a final important remark, it is often convenient to make use of matrix inequalities of the form

$$\mathbf{A} \geq \mathbf{B}, \tag{1.267}$$

which are interpreted in the sense of positive definiteness. In particular, (1.267) means that $\mathbf{A} - \mathbf{B} \geq 0$, i.e., that the difference matrix $\mathbf{A} - \mathbf{B}$ is positive semidefinite. Similarly, the notation $\mathbf{A} > \mathbf{B}$ means that $\mathbf{A} - \mathbf{B}$ is positive definite, and the notation $\mathbf{A} < \mathbf{B}$ means that $\mathbf{B} - \mathbf{A}$ is positive definite. Also, it is occasionally convenient to use the terminology *negative definite* to refer to a matrix $\mathbf{A}$ satisfying $\mathbf{A} < 0$, and *negative semidefinite* to refer to a matrix $\mathbf{A}$ satisfying $\mathbf{A} \leq 0$. Using these conventions, we have, for example, that $\mathbf{A}$ is negative definite whenever $-\mathbf{A}$ is positive definite, etc. A matrix that is neither positive semidefinite nor negative semidefinite is termed *indefinite*.

We emphasize that (1.267) does *not* mean that every entry of $\mathbf{A}$ is at least as big as the corresponding entry of $\mathbf{B}$. However, if we choose, for any j,

$$[\mathbf{x}]_i = \begin{cases} 1 & i = j \\ 0 & \text{otherwise} \end{cases},$$

then the definition of positive semidefiniteness, i.e., (1.252), implies that

$$[\mathbf{A}]_{jj} \geq 0 \qquad \text{for all } j. \tag{1.268}$$

Hence, using (1.268) we can conclude that (1.267) implies, among other relationships, that every *diagonal* entry of $\mathbf{A}$ is not less than the corresponding entry of $\mathbf{B}$.

1.A.6 Subspaces

A subset $\mathcal{S} \subset \mathbb{R}^n$ is a subspace if $\mathcal{S}$ is closed under vector addition and scalar multiplication. Examples of subspaces of $\mathbb{R}^2$ are[17]

$$\mathcal{S}_1 = \left\{ \begin{bmatrix} a \\ 0 \end{bmatrix} \;\middle|\; a \in \mathbb{R} \right\} \tag{1.269}$$

$$\mathcal{S}_2 = \left\{ \begin{bmatrix} a \\ 2a \end{bmatrix} \;\middle|\; a \in \mathbb{R} \right\} \tag{1.270}$$

The dimension of a subspace equals the maximum number of vectors in $\mathcal{S}$ that can form a linearly independent set.

Let $\mathcal{K}$ be any subset of $\mathbb{R}^n$. The *orthogonal complement* of $\mathcal{K}$ in $\mathbb{R}^n$ is defined as follows:

$$\mathcal{K}^\perp = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} \perp \mathbf{y} \text{ for all } \mathbf{y} \in \mathcal{K}\}. \tag{1.271}$$

Note that $\mathcal{K}^\perp$ is a subspace whether or not $\mathcal{K}$ is, since if $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{K}^\perp$ and $\mathbf{y} \in \mathcal{K}$,

$$(\mathbf{x}_1 + \mathbf{x}_2)^{\mathrm{T}}\mathbf{y} = \mathbf{x}_1^{\mathrm{T}}\mathbf{y} + \mathbf{x}_2^{\mathrm{T}}\mathbf{y} = 0 \tag{1.272}$$

$$(\alpha\mathbf{x}_1)^{\mathrm{T}}\mathbf{y} = \alpha\mathbf{x}_1^{\mathrm{T}}\mathbf{y} = 0 \tag{1.273}$$

so $\mathbf{x}_1 + \mathbf{x}_2 \in \mathcal{K}^\perp$ and $\alpha\mathbf{x}_1 \in \mathcal{K}^\perp$.

Let $\mathbf{d}$ be a single nonzero vector in $\mathbb{R}^n$ (so $\{\mathbf{d}\}$ is not a subspace), and consider $\{\mathbf{d}\}^\perp$. This is a subspace of dimension $n-1$. For example, as illustrated in Fig. 1.4, when $n = 2$ the set of $\mathbf{x}$ such that $\mathbf{d}^{\mathrm{T}}\mathbf{x} = 0$ is a line through the origin perpendicular to $\mathbf{d}$. In 3-dimensions this set is a plane through the origin, again perpendicular to $\mathbf{d}$. Note that the subspace $\{\mathbf{d}\}^\perp$ splits $\mathbb{R}^n$ into two *half-spaces*, one corresponding to those $\mathbf{x}$ for which $\mathbf{d}^{\mathrm{T}}\mathbf{x} > 0$, the other to $\mathbf{d}^{\mathrm{T}}\mathbf{x} < 0$.

For additional insights into the concepts and results summarized in this section, see, e.g., G. S. Strang, *Linear Algebra and its Applications*, 3rd ed., Academic Press, New York, 1988.

1.B VECTOR CALCULUS

Several results from vector calculus, which we briefly summarize here, will prove useful. First, consider a scalar function of a vector of n real variables

$$f(\mathbf{x}) = f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \right) = f(x_1, x_2, \ldots, x_n). \tag{1.274}$$

[17] Here $\mathbb{R}$ equals the set of real numbers.

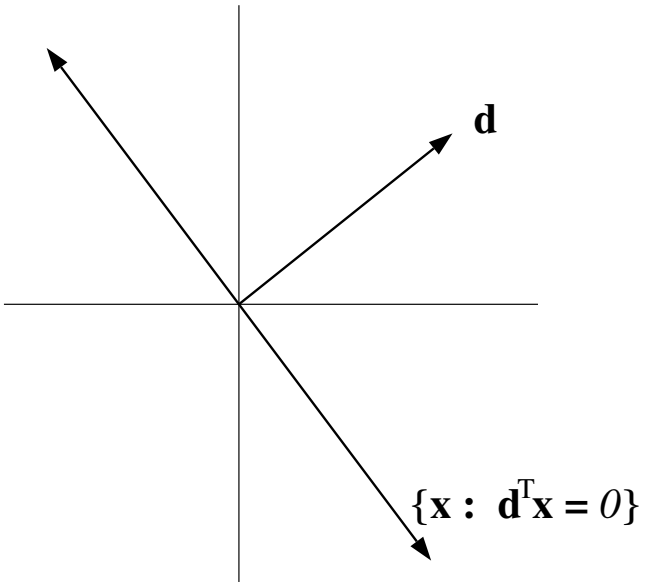

Figure 1.4. An example of a one-dimensional orthogonal complement subspace.

Partial derivatives, integrals, etc., can all be defined in a useful manner. For example, it is convenient to define a *Jacobian* row vector, which consists of first partial derivatives:

$$\frac{df}{d\mathbf{x}}(\mathbf{x}) = \nabla_{\mathbf{x}} f(\mathbf{x}) = \left[\frac{\partial f}{\partial x_1}(\mathbf{x}) \quad \frac{\partial f}{\partial x_2}(\mathbf{x}) \quad \cdots \quad \frac{\partial f}{\partial x_n}(\mathbf{x}) \right]. \tag{1.275}$$

It is also convenient to define a *Hessian* matrix, which consists of second-order partial derivatives:

$$\frac{d^2 f}{d\mathbf{x}^2}(\mathbf{x}) = \nabla_{\mathbf{x}}^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{x}) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_2^2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_n \partial x_2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(\mathbf{x}) \end{bmatrix}. \tag{1.276}$$

Note that the Hessian is a symmetric matrix. Furthermore, the Hessian matrix at $\mathbf{x} = \mathbf{x}_0$ is positive semidefinite, i.e., $d^2 f / d\mathbf{x}^2(\mathbf{x}_0) \geq 0$ whenever $\mathbf{x}_0$ corresponds to a local minimum of $f(\cdot)$. Similarly, if $\mathbf{x} = \mathbf{x}_0$ is the location of a local *maximum* of $f(\cdot)$, then the Hessian satisfies $d^2 f / d\mathbf{x}^2(\mathbf{x}_0) \leq 0$ which means that $-d^2 f / d\mathbf{x}^2(\mathbf{x}_0)$ is positive semidefinite.

Using the notation (1.275) and (1.276) we can conveniently express the multivariable Taylor's series expansion as

$$f(\mathbf{x} + \delta\mathbf{x}) = f(\mathbf{x}) + \frac{df}{d\mathbf{x}}(\mathbf{x})\delta\mathbf{x} + \frac{1}{2!}(\delta\mathbf{x})^{\mathrm{T}} \frac{d^2 f}{d\mathbf{x}^2}(\mathbf{x})\delta\mathbf{x} + \cdots \tag{1.277}$$

where $\cdots$ in (1.277) denotes higher order terms.

Finally, we briefly discuss vector-valued functions $\mathbf{f}(\cdot)$. Derivatives, integrals, limits, etc., for functions of this type are defined component-wise, e.g., for a

vector-valued function with a scalar argument, we have

$$\frac{d}{dx}\mathbf{f}(x) = \begin{bmatrix} \frac{d}{dx}f_1(x) \\ \frac{d}{dx}f_2(x) \\ \vdots \\ \frac{d}{dx}f_m(x) \end{bmatrix}. \tag{1.278}$$

More generally, for a vector-valued function of a vector argument, we define the Jacobian matrix[18]

$$\frac{d\mathbf{f}}{d\mathbf{x}}(\mathbf{x}) = \nabla_{\mathbf{x}}\mathbf{f}(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \frac{\partial f_1}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}) \\ \frac{\partial f_2}{\partial x_1}(\mathbf{x}) & \frac{\partial f_2}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_2}{\partial x_n}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{x}) & \frac{\partial f_m}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_m}{\partial x_n}(\mathbf{x}) \end{bmatrix}. \tag{1.279}$$

Defining second-order derivatives for vector-valued functions of vector-valued arguments is possible but generally less useful (because of the need for three-dimensional matrices). In any case, a multidimensional Taylor series expansion can be obtained from (1.277) through componentwise operations on $\mathbf{f}(\cdot)$, yielding

$$\mathbf{f}(\mathbf{x}+\delta\mathbf{x}) = \mathbf{f}(\mathbf{x}) + \frac{d\mathbf{f}}{d\mathbf{x}}(\mathbf{x})\delta\mathbf{x} + \cdots \tag{1.280}$$

where, again, $\cdots$ denotes higher order terms.

Some simple (but useful) examples of the calculations described in this section are the following:

$$\frac{d\mathbf{x}}{d\mathbf{x}} = \mathbf{I} \tag{1.281}$$

$$\frac{d}{d\mathbf{x}}\mathbf{A}\mathbf{x} = \mathbf{A} \tag{1.282}$$

$$\frac{d}{d\mathbf{x}}\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} = \mathbf{x}^{\mathrm{T}}(\mathbf{A}+\mathbf{A}^{\mathrm{T}}) \tag{1.283}$$

$$\frac{d^2}{d\mathbf{x}^2}\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} = \mathbf{A}+\mathbf{A}^{\mathrm{T}}. \tag{1.284}$$

[18]Note that (1.279) is consistent both with (1.275) when $m = 1$ and with (1.278) when $n = 1$.