B

Basic Probability Concepts

THIS appendix serves as an overview of the probability concepts that are most important in the present text's approach to estimation theory. These developments are patterned after the excellent survey provided by Bryson and Ho.[1] Still, the interested student is strongly encouraged to study probability theory formally from conventional texts such as Refs. [2]-[5].

B.1 Functions of a Single Discrete-Valued Random Variable

To appeal to the intuitive feel that we have for random variables and elementary probability concepts, attention is first directed to a simple experiment. Consider a single throw of a "true" die; the *probability* of the occurrence of each of the *events* 1, 2, 3, 4, 5, or 6 is exactly the same on a given throw. For a "loaded" die, the probability of certain of the events would be greater than others. If a given discrete-values experiment is conducted N times and N_j is the number of times that the j^{th} event $x(j)$ occurred, then it is intuitively reasonable to define the probability of the occurrence of $x(j)$ as

$$p(x(j)) \equiv \lim_{N\to\infty} \frac{N_j}{N} \tag{B.1}$$

For example, for a throw of a single die the probability of obtaining a value of 3 is given by $p(3) = 1/6$.

A *discrete-valued random variable*, x, is defined as a function having finite number of possible values $x(j)$; with the associated probability of $x(j)$ occurring being denoted by $p(x(j))$. To compact notation, $x(j)$ and $p(x(j))$ are hereafter called x and $p(x)$, whenever this substitution does not cause ambiguity.

Let us expand the die concept for the case of a single throw of two dice. We now have 36 possible outcomes over the entire set. Table B.1 shows the sum of the two dice, the number of times that sum can occur and the probability of that event. Clearly, obtaining a 7 has the highest probability. When multiple dice, $n > 2$, are used this table is much more difficult to produce. Fortunately, a simple mathematical approach known as a *generating function* can be used for this case:

$$f(x) = \left(x + x^2 + x^3 + x^4 + x^5 + x^6\right)^n \tag{B.2}$$

Table B.1: Probabilities for a Single Throw of Two Dice

Sum	Count	$p(x)$
2	1	1/36
3	2	2/36
4	3	3/36
5	4	4/36
6	5	5/36
7	6	6/36
8	5	5/36
9	4	4/36
10	3	3/36
11	2	2/36
12	1	1/36

The coefficients of the powers of x can be used to form the "count" column. The probability of each event is given by the count divided by 6^n.

Let us consider another experiment involving four flips of a coin. We want to look at the number of ways a heads appears for the 16 total number of outcomes. This is presented as a histogram in Figure B.1. Mathematically, the number of ways to obtain x heads in n flips is spoken as the "number of combinations of n things taken x at a time." The number of ways can be computed by

$$\text{Number of Ways} \equiv \binom{n}{x} = \frac{n!}{x!\,(n-x)!} \tag{B.3}$$

For example if $n = 4$ and $x = 2$, then the number of ways is computed to be 6. The probability of obtaining a heads is given by the number of ways divided by the total number of outcomes (16 in our case). This probability can be generalized by noting that the number of outcomes is given by 2^n:

$$p(x) = \frac{\binom{n}{x}}{2^n} = \frac{n!}{x!\,(n-x)!\,2^n} \tag{B.4}$$

For example if $n = 4$ and $x = 2$, then $p(2) = 0.375$.

A compound event can be defined as the occurrence of "either $x(j)$ or $x(k)$"; the probability of a compound event is defined as

$$p(x(j) \cup x(k)) = p(x(j)) + p(x(k)) - p(x(j) \cap x(k)) \tag{B.5}$$

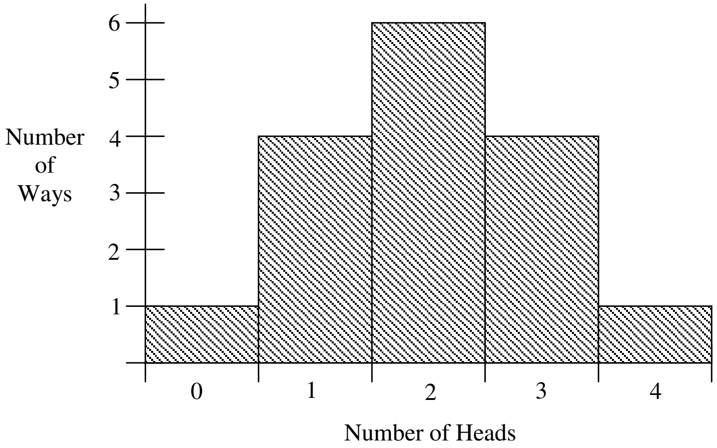

Figure B.1: Histogram of the Number of Ways a Heads Appears

where $x(j) \cup x(k)$ denotes "$x(j)$ or $x(k)$" and $x(j) \cap x(k)$ denotes "$x(j)$ and $x(k)$." The probability of obtaining one event *and* another event is known as the *joint* probability of $x(j)$ and $x(k)$. If $p(x(j) \cap x(k)) = 0$, then the individual probabilities are summed to determine the overall probability. For example, the probability of obtaining less than 3 heads in 4 flips is given by $1/16 + 4/16 + 6/16 = 0.6875$. Note that calculating the probability of obtaining 4 or less heads gives a value of 1! It is clear that a *probability mass function* $p(x(j))$ has the following properties:

$$0 \le p(x(j)) \le 1 \tag{B.6a}$$

$$\sum_j p(x(j)) = 1 \tag{B.6b}$$

If events $x(j)$ and $x(k)$ are *independent*, then we have

$$p(x(j) \cap x(k)) = p(x(j))\, p(x(k)) \tag{B.7}$$

For example, the probability of obtaining one heads in two successive trials is given by $(1/4) \times (1/4) = 1/16$.

We now define the *conditional probability* of $x(j)$ given $x(k)$, which is denoted by $p(x(j)|x(k))$. Suppose we know that an event $x(k)$ has occurred. Then $x(j)$ occurs if and only if $x(j)$ and $x(k)$ occur. Therefore, the probability of $x(j)$, given that we know $x(k)$ has occurred, should intuitively be proportional to $p(x(j) \cap x(k))$. However, the conditional probability must satisfy the properties of probability shown by eqn. (B.6). This forces a proportionality constant of $1/p(x(k))$, so that

$$p(x(j)|x(k)) = \frac{p(x(j) \cap x(k))}{p(x(k))} \tag{B.8}$$

In a similar fashion the conditional probability of $x(k)$ given $x(j)$ is

$$p(x(k)|x(j)) = \frac{p(x(k) \cap x(j))}{p(x(j))} \tag{B.9}$$

Combining eqns. (B.8) and (B.9) leads to *Bayes rule*:

$$p(x(j)|x(k)) = \frac{p(x(k)|x(j))\, p(x(j))}{p(x(k))} \tag{B.10}$$

This rule is widely used in estimation theory (e.g., see §2.6). Bayes rule can be used to show some counterintuitive results. For example, say 1 out 1,000 people have a rare disease. Tests show that 99% are positive when they have a disease and 2% are positive when they don't. Bayes rule can be used to show the probability that they actually have a disease when the test is positive is only 0.047! At first glance this seems counterintuitive, but in actuality the result is correct (note: if a 25% incidence rate is given, then the probability is 0.94, which is line with our intuition).

The random variable x is usually described in terms of its *moments*. The first two moments of x are given by the *mean* (μ) of x:

$$\mu \equiv \sum_j x(j)\, p(x(j)) \tag{B.11}$$

and the variance (σ^2) of x:

$$\sigma^2 \equiv \sum_j (x(j) - \mu)^2 p(x(j)) \tag{B.12}$$

The quantity σ is often called the *standard deviation* of x. If $p(x)$ is considered to be a function defining the mass of several discrete masses located along a straight line, then μ locates the center of mass and σ^2 is the moment of inertia of the system of masses about their centroid.

The *expected value* or "average value" of a function $f(x)$ of a discrete random variable x is defined as

$$E\{f(x)\} = \sum_j f(x(j))\, p(x(j)) \tag{B.13}$$

Clearly from eqns. (B.11) and (B.12), the mean and variance are the expected values of x and $(x-\mu)^2$, respectively. Notice that the expected value operator is linear so that

$$E\{a\, f(x) + b\, g(x)\} = a\, E\{f(x)\} + b\, E\{g(x)\} \tag{B.14}$$

for a and b arbitrary deterministic scalars, and $f(x)$ and $g(x)$ arbitrary functions of the random variable x.

B.2 Functions of Discrete-Valued Random Variables

A random vector $\mathbf{x}$ is an $n \times 1$ matrix whose elements x_i are scalar random variables as discussed in §B.1. If each scalar element x_i of $\mathbf{x}$ can take on a finite number, m_i, of discrete values $x_i(j_i)$, for $j_i = 1, 2, \ldots, m_i$, then there are $m_1 m_2 \cdots m_n$ possible vectors. For a complete probabilistic characterization of $\mathbf{x}$, its *joint probability function* $p(j_1, j_2, \ldots, j_n)$ is the probability that x_1 has its j_1^{th} value, x_2 has its j_2^{th} value, ..., x_n has its j_n^{th} value. The function $p(j_1, j_2, \ldots, j_n)$ is often written $p(x_1, x_2, \ldots, x_n)$ when no ambiguity results. On some occasions, one is interested in the *marginal probability mass function* given by

$$p(j_1) = \sum_{j_2=1}^{m_2} \sum_{j_3=1}^{m_3} \cdots \sum_{j_n=1}^{m_n} p(j_1, j_2, \ldots, j_n) \tag{B.15}$$

Note that $p(j_1)$ is the probability of a compound event; that x_1 takes on its j_1^{th} value while $x_2, x_3, \ldots, x_n$ take on arbitrary possible values. Thus, a scalar random variable may represent an elementary or compound event, depending upon the dimension of the underlying space of events.

The marginal probability functions in eqn. (B.15) are sufficient to fully probabilistically characterize the components of $\mathbf{x}$, but to fully characterize $\mathbf{x}$, it is necessary to specify $p(x_1, x_2, \ldots, x_n)$. As in the scalar case, it is customary to describe $p(x_1, x_2, \ldots, x_n)$ and $\mathbf{x}$ in terms of the moments of $\mathbf{x}$. The first two moments are the *mean* ($\boldsymbol{\mu}$) of $\mathbf{x}$:

$$\boldsymbol{\mu} \equiv E\{\mathbf{x}\} = \sum_{j_1=1}^{m_1} \cdots \sum_{j_n=1}^{m_n} \begin{bmatrix} x_1(j_1) \\ \vdots \\ x_n(j_n) \end{bmatrix} p(j_1, j_2, \ldots, j_n) \tag{B.16}$$

and the *covariance* (R) of $\mathbf{x}$:

$$\begin{aligned} R &\equiv E\left\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right\} \\ &= E\left\{\begin{bmatrix} (x_1-\mu_1)^2 & (x_1-\mu_1)(x_2-\mu_2) & \cdots & (x_1-\mu_1)(x_n-\mu_n) \\ (x_2-\mu_2)(x_1-\mu_1) & (x_2-\mu_2)^2 & \cdots & (x_2-\mu_2)(x_n-\mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ (x_n-\mu_n)(x_1-\mu_1) & (x_n-\mu_n)(x_2-\mu_2) & \cdots & (x_n-\mu_n)^2 \end{bmatrix}\right\} \end{aligned} \tag{B.17}$$

where the expectation operator $E\{\ \}$ when "operating" upon a matrix, operates upon each individual element. Notice that the covariance matrix R is symmetric. We adopt the following notations:

$$\sigma_i^2 \equiv E\left\{(x_i-\mu_i)^2\right\} = \text{ variance of } x_i \tag{B.18a}$$

$$\sigma_{ij} \equiv E\left\{(x_i-\mu_i)(x_j-\mu_j)\right\} = \text{ covariance of } x_i \text{ and } x_j \tag{B.18b}$$

The covariance matrix is commonly written as

$$R \equiv \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1}\sigma_n\sigma_1 & \rho_{n2}\sigma_n\sigma_2 & \cdots & \sigma_n^2 \end{bmatrix} \tag{B.19}$$

where ρ_{ij} is the *correlation* of x_i and x_j, defined by

$$\rho_{ij} \equiv \frac{\sigma_{ij}}{\sigma_i \sigma_j} \tag{B.20}$$

This coefficient gives a measure of the degree of linear dependence between x_i and x_j. If x_i is linear in x_j, then $\rho_{ij} = \pm 1$; however, if x_i and x_j are independent of each other, then $\rho_{ij} = 0$. If

$$p(x_1, x_2, \ldots, x_n) = p(x_1)\, p(x_2) \cdots p(x_n) \tag{B.21}$$

for all possible values of $\{x_1, x_2, \ldots, x_n\}$, then the random variables are independent, as discussed in §B.1. Note that while pairwise independence is sufficient to ensure zero correlation of $\{x_1, x_2, \ldots, x_n\}$, it is not sufficient to ensure independence of $\{x_1, x_2, \ldots, x_n\}$.[6]

Example B.1: Consider a vector with two components $\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T$. Suppose that the first component has two possible values:

$$x_1(1) = 0$$
$$x_1(2) = 10$$

Suppose that the second component has three possible values:

$$x_2(1) = -10$$
$$x_2(2) = 0$$
$$x_2(3) = 10$$

Suppose, further, that the six possible events have the following probabilities:

$$p(0,-10) = 0.1, \quad p(0,0) = 0.4, \quad p(0,10) = 0.1$$
$$p(10,-10) = 0.1, \quad p(10,0) = 0.1, \quad p(10,10) = 0.2$$

The expected value (mean) of $\mathbf{x}$ then follows from eqn. (B.16) as

$$\boldsymbol{\mu} = 0.1\begin{bmatrix} 0 \\ -10 \end{bmatrix} + 0.4\begin{bmatrix} 0 \\ 0 \end{bmatrix} + 0.1\begin{bmatrix} 0 \\ 10 \end{bmatrix} + 0.1\begin{bmatrix} 10 \\ -10 \end{bmatrix} + 0.1\begin{bmatrix} 10 \\ 0 \end{bmatrix} + 0.2\begin{bmatrix} 10 \\ 10 \end{bmatrix}$$

which reduces to

$$\boldsymbol{\mu} = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$$

Similarly, the covariance matrix follows from eqn. (B.17) as

$$
\begin{aligned}
R &= E\begin{bmatrix} (x_1-\mu_1)^2 & (x_1-\mu_1)(x_2-\mu_2) \\ (x_2-\mu_2)(x_1-\mu_1) & (x_2-\mu_2)^2 \end{bmatrix} \\
&= 0.1\begin{bmatrix} -4 \\ -11 \end{bmatrix}\begin{bmatrix} -4 & -11 \end{bmatrix} + 0.4\begin{bmatrix} -4 \\ -1 \end{bmatrix}\begin{bmatrix} -4 & -1 \end{bmatrix} + 0.1\begin{bmatrix} -4 \\ 9 \end{bmatrix}\begin{bmatrix} -4 & 9 \end{bmatrix} \\
&+ 0.1\begin{bmatrix} 6 \\ -11 \end{bmatrix}\begin{bmatrix} 6 & -11 \end{bmatrix} + 0.1\begin{bmatrix} 6 \\ -1 \end{bmatrix}\begin{bmatrix} 6 & -1 \end{bmatrix} + 0.2\begin{bmatrix} 6 \\ 9 \end{bmatrix}\begin{bmatrix} 6 & 9 \end{bmatrix}
\end{aligned}
$$

which reduces to

$$R = \begin{bmatrix} 24 & 6 \\ 6 & 49 \end{bmatrix}$$

It may be verified from the results of Appendix A that this covariance matrix is positive definite.

To investigate the definiteness of R in general, let

$$\boldsymbol{\mu} = E\{\mathbf{x}\} \tag{B.22a}$$

$$z = \mathbf{c}^T(\mathbf{x}-\boldsymbol{\mu}) \tag{B.22b}$$

where $\mathbf{c}$ is an $n \times 1$ vector of arbitrary constraints. Investigating the moments of z, we find

$$\mu_z \equiv E\{z\} = E\left\{\mathbf{c}^T(\mathbf{x}-\boldsymbol{\mu})\right\} = \mathbf{c}^T(\boldsymbol{\mu}-\boldsymbol{\mu}) = 0 \tag{B.23}$$

and

$$
\begin{aligned}
\sigma_z^2 &\equiv E\left\{(z-\mu_z)^2\right\} = E\left\{\mathbf{c}^T(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{c}\right\} \\
&= \mathbf{c}^T E\left\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right\}\mathbf{c} \\
&= \mathbf{c}^T R\,\mathbf{c}
\end{aligned}
\tag{B.24}
$$

Since $\sigma_z^2 \geq 0$ and since $\mathbf{c}$ is an arbitrary vector, then R is always *at least* positive semi-definite. For diagonal R, the positive semi-definiteness of R agrees with our intuitive interpretation of σ_i^2; since $\sigma_i^2 < 0$ implies "better than perfect knowledge" or "less than zero uncertainty" in x_i, which is impossible!

B.3 Functions of Continuous Random Variables

For our purposes, the discrete variable concepts of §B.1 and §B.2 can be extended in a natural manner.* By letting $N \rightarrow \infty$ with the probability mass function

*There are various theoretical details that must be focused in a rigorous extension of the discrete results to the continuous results (see Ref. [4], for example).

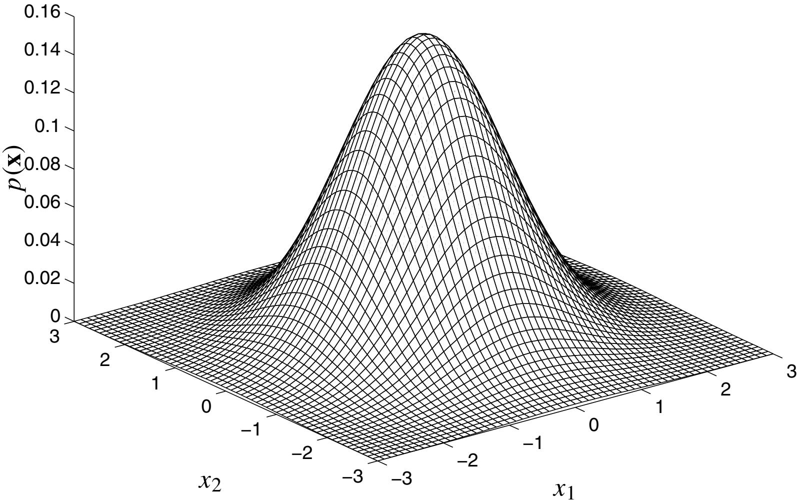

Figure B.2: Two-Dimensional Gaussian Distribution

$p(x_1(j_1), \ldots, x_n(j_n))$ being replaced by a *probability density function* $p(x_1, \ldots, x_n)$; then

$$p(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n \tag{B.25}$$

is the probability that the components of **x** lie within the differential volume given by $dx_1\, dx_2 \cdots dx_n$ centered at $x_1, x_2, \ldots, x_n$. Since all possible **x**-vectors are located in the infinite sphere, it follows that

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n = 1 \tag{B.26}$$

The expected value of an arbitrary function $g(x_1, \ldots, x_n)$ is defined in terms of the density function as

$$E\{g(x_1, \ldots, x_n)\} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n)\, p(x_1, \ldots, x_n)\, dx_1 \cdots dx_n \tag{B.27}$$

Thus the summation signs of the discrete results of §B.2 are replaced by integral signs to obtain the corresponding continuous results.

B.4 Gaussian Random Variables

The most widely used distribution for state estimation involves the Gaussian random process. Taking the limit as the number of coin flips, used to produce the histogram shown in Figure B.1, approaches infinity leads to the *Gaussian* or *normal* density function for x:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{B.28}$$

with mean given by μ and variance given by σ^2. This function can also be expanded to the multidimensional case for a vector $\mathbf{x}$:

$$p(\mathbf{x}) = \frac{1}{[\det(2\pi R)]^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T R^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{B.29}$$

A plot of this function for two variables, with $\boldsymbol{\mu} = \mathbf{0}$ and $R = I_{2\times 2}$, is shown in Figure B.2. The mean and standard deviation are sufficient enough to define this distribution. Therefore, a simple notation for this distribution is given by

$$p(\mathbf{x}) \sim N(\boldsymbol{\mu}, R) \tag{B.30}$$

The Gaussian distribution is important because of a very useful property that involves any distribution. The *central limit theorem* states that given a distribution with mean μ and variance σ^2, the sampling distribution (no matter what the shape of the original distribution) approaches a Gaussian distribution with mean μ and variance σ^2/N as N, the sample size, increases. This can be clearly seen in Figure B.1, where even for a relatively small sample size the histogram looks like the classic "bell shape" form of the Gaussian distribution. For a formal proof of the central limit theorem see Ref. [7].

A stochastic process is simply a collection of random vectors defined on the same probability space.[8] A *zero-mean Gaussian white-noise* process has the following properties:

$$E\{\mathbf{x}\} = \mathbf{0} \tag{B.31a}$$

$$E\left\{\mathbf{x}(\tau)\mathbf{x}^T(\tau')\right\} = R\,\delta(\tau'-\tau) \tag{B.31b}$$

where $\delta(\tau'-\tau)$ is the delta function. The standard deviation for this process gives a level of confidence that a particular sample lies within the distribution. Also, a process is said to be *stationary* if its random variable statistics do not vary in time (i.e., the probability statistics at time τ have the same mean and covariance as the probability statistics at time τ').

One is often interested in the probability that $\mathbf{x}$ lies inside the quadratic hypersurface

$$(\mathbf{x}-\boldsymbol{\mu})^T R^{-1}(\mathbf{x}-\boldsymbol{\mu}) < G^2 \tag{B.32}$$

where G is a constant. Using an eigenvalue/eigenvector decomposition (see Appendix A) of R, leads to the appropriate orthogonal transformation

$$\mathbf{x} = T\mathbf{y} \tag{B.33a}$$

$$S \equiv \text{diag}\left[\sigma_1^2 \; \sigma_2^2 \; \cdots \; \sigma_n^2\right] = T^T R\, T \tag{B.33b}$$

Therefore, it is always possible to transform coordinates to a principal system in which eqn. (B.32) is reduced to

$$\frac{y_1^2}{\sigma_1^2} + \frac{y_2^2}{\sigma_2^2} + \cdots + \frac{y_n^2}{\sigma_n^2} < G^2 \tag{B.34}$$

We now define another set of change of variables:

$$z_i = \frac{y_i}{\sigma_i}, \quad i = 1, 2, \ldots, n \tag{B.35}$$

so that eqn. (B.34) reduces down to

$$z_1^2 + z_2^2 + \cdots + z_n^2 < G^2 \tag{B.36}$$

The probability of finding z inside this hypersurface is obtained by integrating the Gaussian density function over the volume of the sphere in eqn. (B.36) as

$$p(g^2 \leq G^2) = \int_V p(z)\, dV \tag{B.37}$$

where

$$g^2 \equiv \sum_{i=1}^{n} z_i^2 \tag{B.38}$$

Using the element volume $dz_1 dz_2 \cdots dz_n$, eqn. (B.37) can be written as

$$p(g^2 \leq G^2) = \int \cdots \int_V \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} g^2\right) dz_1 dz_2 \cdots dz_n \tag{B.39}$$

Using an n-dimensional spherically volume element $f(g)\, dg$, eqn. (B.39) can be written as

$$p(g^2 \leq G^2) = \frac{1}{(2\pi)^{n/2}} \int_0^G \exp\left(-\frac{1}{2} g^2\right) f(g)\, dg \tag{B.40}$$

For $n = 1, 2, 3$, eqn. (B.40) is explicitly:

- $n = 1$, $f(g)\, dg = 2dg$:

$$\begin{aligned} p(g \leq G) &= \sqrt{2/\pi} \int_0^G \exp\left(-\frac{1}{2} g^2\right) dg \\ &= \text{erf}\left(\frac{G}{\sqrt{2}}\right) \end{aligned} \tag{B.41}$$

Table B.2: Probability Values for $g \le G$

	$G = 1$	$G = 2$	$G = 3$
$n = 1$	0.683	0.995	0.997
$n = 2$	0.394	0.865	0.989
$n = 3$	0.200	0.739	0.971

- $n = 2$, $f(g)\,dg = 2\pi g\,dg$:

$$\begin{aligned} p(g \le G) &= \int_0^G \exp\left(-\frac{1}{2}g^2\right) g\,dg \\ &= 1 - \exp\left(\frac{-G^2}{2}\right) \end{aligned} \tag{B.42}$$

- $n = 3$, $f(g)\,dg = 4\pi g^2\,dg$:

$$\begin{aligned} p(g \le G) &= \sqrt{2/\pi}\int_0^G \exp\left(-\frac{1}{2}g^2\right) g^2\,dg \\ &= \text{erf}\left(\frac{G}{\sqrt{2}}\right) - G\sqrt{2/\pi}\exp\left(\frac{-G^2}{2}\right) \end{aligned} \tag{B.43}$$

where erf is the error function. The numerical value of $p(g < G)$ is often of particular interest in error analysis. Table B.2 displays the “curse of dimensionality” for the probability of g being within 1, 2, and 3 “sigma ellipsoids” for 1, 2, and 3 dimensional spaces.

B.5 Chi-Square Random Variables

The chi-square distribution is often used to provide a consistency test in estimators (see §5.7.3), which is useful to determine whether or not reasonable state estimates are provided. Assuming a Gaussian distribution for the $n \times 1$ vector $\mathbf{x}$, with mean $\boldsymbol{\mu}$ and covariance R, the following variable is said to have a chi-square distribution with n degrees of freedom (DOF):

$$q = (\mathbf{x} - \boldsymbol{\mu})^T R^{-1} (\mathbf{x} - \boldsymbol{\mu}) \tag{B.44}$$

The variable q is the sum of squares of n independent zero-mean variables with variance equal to one. This can be shown by defining the following variable:[9]

$$\mathbf{u} \equiv R^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \tag{B.45}$$

Then, $\mathbf{u}$ is clearly Gaussian with $E\{\mathbf{u}\} = \mathbf{0}$ and $E\left\{\mathbf{u}\mathbf{u}^T\right\} = I$. The chi-square distribution is written as

$$q \sim \chi_n^2 \tag{B.46}$$

The mean and variance are given by

$$E\{q\} = \sum_{i=1}^{n} E\left\{u_i^2\right\} = n \tag{B.47a}$$

$$E\left\{(q-n)^2\right\} = \sum_{i=1}^{n} E\left\{(u_i^2-1)^2\right\} = \sum_{i=1}^{n}(3-2+1) = 2n \tag{B.47b}$$

where the relationship $E\left\{x^4\right\} = 3\sigma^4$ has been used for the term involving u_i^4. This relationship is given from the scalar version of[9]

$$E\left\{\mathbf{x}^T A\,\mathbf{x}\mathbf{x}^T B\,\mathbf{x}\right\} = \text{Tr}(A\,R)\,\text{Tr}(B\,R) + 2\text{Tr}(A\,R\,B\,R) \tag{B.48}$$

where A and B are $n \times n$ matrices.

The chi-square density function with n DOF is given by

$$p(q) = \frac{1}{2^{n/2}\Gamma(n/2)}\, q^{\frac{n-2}{2}} e^{-\frac{q}{2}} \tag{B.49}$$

where the gamma function Γ is defined as

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} \tag{B.50a}$$

$$\Gamma(1) = 1 \tag{B.50b}$$

$$\Gamma(m+1) = m\,\Gamma(m) \tag{B.50c}$$

Tables of points on the chi-square distribution can be found in Refs. [4] and [9]. For DOF's above 100, the following approximation can be used:[9]

$$\chi_n^2(1-Q) = \frac{1}{2}\left[\mathcal{G}(1-Q) + \sqrt{2n-1}\right]^2 \tag{B.51}$$

where $\chi_n^2(1-Q)$ indicates that to the left of a specific point, the probability mass is $1-Q$. An important quantity used in consistency tests is the 95% *two-sided probability region* for an $N(0,1)$ random variable:

$$[\mathcal{G}(0.025),\, \mathcal{G}(0.975)] = [-1.96,\, 1.96] \tag{B.52}$$

Other values for $\mathcal{G}$ can be found in Ref. [9]. Then, specific values can be calculated for $\chi_n^2(1-Q)$ using eqn. (B.51); e.g., $\chi_{400}^2(0.025) = 346$ and $\chi_{400}^2(0.975) = 457$.

B.6 Propagation of Functions through Various Models

In this section the basic concepts for the propagation of functions through linear and nonlinear models is shown. We shall see that for linear models the original assumed density function is maintained (e.g., a Gaussian input into a linear system produces a Gaussian output), but for nonlinear models this concept does not hold in general.

B.6.1 Linear Matrix Models

We consider the following linear matrix equation:

$$\mathbf{y} = A\,\mathbf{x} + \mathbf{b} \tag{B.53}$$

where A and $\mathbf{b}$ are arbitrary constant matrices with deterministic elements, and $\mathbf{x}$ is a random vector whose first two moments are assumed known:

$$\boldsymbol{\mu} = E\{\mathbf{x}\} \tag{B.54a}$$

$$R = E\left\{(\mathbf{x}-\boldsymbol{\mu})\,(\mathbf{x}-\boldsymbol{\mu})^T\right\} \tag{B.54b}$$

It is desired to determine the first and second moments of $\mathbf{y}$. The mean follows

$$\boldsymbol{\mu}_y \equiv E\{\mathbf{y}\} = E\{A\,\mathbf{x}+\mathbf{b}\} = A\,E\{\mathbf{x}\} + \mathbf{b} \tag{B.55}$$

or

$$\boldsymbol{\mu}_y = A\,\boldsymbol{\mu} + \mathbf{b} \tag{B.56}$$

The covariance matrix is then obtained from the definition

$$R_{yy} \equiv E\left\{(\mathbf{y}-\boldsymbol{\mu}_y)\,(\mathbf{y}-\boldsymbol{\mu}_y)^T\right\} \tag{B.57}$$

Substituting eqns. (B.53) and (B.56) into eqn. (B.57) gives

$$R_{yy} = E\left\{A\,(\mathbf{x}-\boldsymbol{\mu})\,(\mathbf{x}-\boldsymbol{\mu})^T A^T\right\} = A\,E\left\{(\mathbf{x}-\boldsymbol{\mu})\,(\mathbf{x}-\boldsymbol{\mu})^T\right\}\,A^T \tag{B.58}$$

or

$$R_{yy} = A\,R\,A^T \tag{B.59}$$

which is a commonly used result for “swapping” covariance matrices through linear systems.

B.6.2 Nonlinear Models

If $\mathbf{x}$ is a random vector whose density function $p(\mathbf{x})$ is known, and if $\mathbf{y} = \mathbf{f}(\mathbf{x})$ is an arbitrary (generally nonlinear) one-to-one transformation, then it can be shown that the density function of $\mathbf{y}$ is given by[1]

$$p(\mathbf{y}) = p(\mathbf{x}) \left| \det \left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right) \right|^{-1} \tag{B.60}$$

with $\mathbf{x}$ on the right-hand side of eqn. (B.60) given by

$$\mathbf{x} = \mathbf{f}^{-1}(\mathbf{y}) \tag{B.61}$$

where $\mathbf{f}^{-1}(\mathbf{y})$ denotes the "reverse" relationship. Thus to convert the density function of $\mathbf{x}$ to the density function of $\mathbf{y}$, simply write the density of $\mathbf{x}$ in terms of $\mathbf{y}$ and multiply by the inverse determinant of the Jacobian matrix.

Example B.2: We will now employ the preceding results using the linear scalar model

$$y = a\,x \tag{B.62}$$

and the following assumed Gaussian density function for x:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-x^2}{2\sigma^2} \right) \tag{B.63}$$

Then, from eqn. (B.60) we find

$$p(y) = \frac{1}{a\,\sigma\sqrt{2\pi}} \exp\left(\frac{-y^2}{2a^2\sigma^2} \right) \tag{B.64}$$

Note further

$$\begin{aligned} \mu_y \equiv E\{y\} &= \int_{-\infty}^{\infty} y\,p(y)\,dy \\ &= \frac{1}{a\,\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} y \exp\left(\frac{-y^2}{2a^2\sigma^2} \right) dy \end{aligned} \tag{B.65}$$

Integrating by parts leads to

$$\mu_y = 0 \tag{B.66}$$

which is equivalent to the expected value of x. Similarly, we find from the definition of variance that

$$\begin{aligned} \sigma_y^2 \equiv E\left\{(y-\mu_y)^2\right\} &= E\left\{y^2\right\} \\ &= \int_{-\infty}^{\infty} y^2 p(y)\,dy \\ &= \frac{1}{a\,\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 \exp\left(\frac{-y^2}{2a^2\sigma^2} \right) dy \end{aligned} \tag{B.67}$$

which integrates to

$$\sigma_y^2 = a^2 \sigma^2 \tag{B.68}$$

This mean and variance of y computed here confirms the previous results shown in eqns. (B.56) and (B.59). Also, we see that y itself is clearly a Gaussian random variable, which confirms that a transformation through a linear model does not alter the form of the distribution.

Example B.3: Assume the following quadratic model:

$$y = a\,x^2 \tag{B.69}$$

Note that for each value of y there are two x-values. Assume that x has the following Gaussian density function:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-x^2}{2\sigma^2}\right) \tag{B.70}$$

It follows from eqn. (B.60) that

$$p(y) = \frac{1}{2\sigma\sqrt{2\pi\,a\,y}} \exp\left(\frac{-y}{2a\,\sigma^2}\right), \quad \text{for } y > 0 \tag{B.71}$$

and

$$p(y) = 0, \quad \text{for } y < 0 \tag{B.72}$$

It also follows that

$$\mu_y \equiv E\{y\} = a\,\sigma^2 \tag{B.73}$$

and

$$\sigma_y^2 \equiv E\left\{(y - \mu_y)^2\right\} = 2a^2\sigma^4 \tag{B.74}$$

Note that y is no longer a Gaussian variable. Hence, unlike the linear case, a nonlinear transformation of a Gaussian variable does not necessarily produce another Gaussian variable.

References

[1] Bryson, A.E. and Ho, Y.C., *Applied Optimal Control*, Taylor & Francis, London, England, 1975.

[2] Cox, D.R. and Hinkley, D.V., *Problems and Solutions in Theoretical Statistics*, John Wiley & Sons, New York, NY, 1978.

[3] Keeping, E., *Introduction to Statistical Inference*, Dover Publications, New York, NY, 1995.

[4] Freund, J.E. and Walpole, R.E., *Mathematical Statistics*, Prentice Hall, Englewood Cliffs, NJ, 4th ed., 1987.

[5] Devore, J.L., *Probability and Statistics for Engineering and Sciences*, Duxbury Press, Pacific Grove, CA, 1995.

[6] Feller, W., *Introduction to Probability Theory and Its Applications*, John Wiley & Sons, New York, NY, 3rd ed., 1966.

[7] Kallenberg, O., *Foundations of Modern Probability*, Springer-Verlag, New York, NY, 1997.

[8] Sage, A.P. and White, C.C., *Optimum Systems Control*, Prentice Hall, Englewood Cliffs, NJ, 2nd ed., 1977.

[9] Bar-Shalom, Y., Li, X.R., and Kirubarajan, T., *Estimation with Applications to Tracking and Navigation*, John Wiley & Sons, New York, NY, 2001.