Optical Propagation, Detection, and Communication

Jeffrey H. Shapiro
Massachusetts Institute of Technology

Chapter 3

Probability Review

The quantitative treatment of our generic photodetector model will require the mathematics of probability and random processes. Although the reader is *assumed* to have prior acquaintance with the former, it is nevertheless worthwhile to furnish a high-level review, both to refresh memories and to establish notation.

3.1 Probability Space

Probability is a mathematical theory for modeling and analyzing real-world situations, often called *experiments,* which exhibit the following attributes.

- The outcome of a particular trial of the experiment appears to be random.[1]

- In a long sequence of independent, macroscopically identical trials of the experiment, the outcomes exhibit statistical regularity.

- Statements about the average behavior of the experimental outcomes are useful.

To make these abstractions more explicit, consider the ubiquitous introductory example of coin flipping. On a particular coin flip, the outcome—either heads (H) or tails (T)—cannot be predicted with certainty. However, in a long

[1] We use the phrase *appears to be random* to emphasize that the indeterminacy need not be fundamental, i.e., that it may arise from our inability—or unwillingness—to specify microscopic initial conditions for the experiment with sufficient detail to determine the outcome precisely.

sequence of N independent, macroscopically identical coin flips, the relative frequency of getting heads, i.e., the fraction of flips which come up heads,

$$f_N(H) \equiv \frac{N(H)}{N}, \tag{3.1}$$

where $N(H)$ is the number of times H occurs, stabilizes to a constant value as $N \to \infty$. Finally, based on this relative-frequency behavior, we are willing to accept statements of the form, "For a *fair* coin, the *probability* of getting $|f_N(H) - \frac{1}{2}| \leq 0.1$ exceeds 99% when $N \geq 250$.", where we have injected our notion that $f_N(H)$ should stabilize at something called a probability, and that this probability should be $\frac{1}{2}$ for a fair coin.

The coin-flip example suggests that probability theory should be developed as an empirical science. It is much better, however, to develop probability theory axiomatically, and then show that its consequences are in accord with empirical relative-frequency behavior. The basic unit in the probabilistic treatment of a random experiment is its *probability space*, $\mathcal{P} = \{\Omega, \Pr(\cdot)\}$, which consists of a *sample space,* Ω, and a *probability measure,* $\Pr(\cdot)$.[2] The sample space Ω is the set of all elementary outcomes, or *sample points* $\{\omega\}$, of the experiment. In order that these $\{\omega\}$ be elementary outcomes, they must be *mutually exclusive*—if $\omega_1 \in \Omega$ occurred when the experiment was performed, then $\omega_2 \in \Omega$ cannot also have occurred on that trial, for all $\omega_2 \neq \omega_1$. In order that these $\{\omega\}$ be elementary outcomes, they must also be *finest grained*—if $\omega_1 \in \Omega$ is known to have occurred when the experiment was performed, no deeper level of information about the experiment's outcome is of interest. Finally, in order that the sample space $\Omega = \{\omega\}$ comprise *all* the elementary outcomes of the experiment, the $\{\omega\}$ must be *collectively exhaustive*—when the experiment is performed, the resulting outcome is always a member of the sample space.

In terms of a simple experiment in which a coin is flipped twice, the natural choice for the sample space is

$$\Omega = \{HH, HT, TH, TT\}, \tag{3.2}$$

where HT denotes heads occurred on the first flip and tails on the second flip, etc. Ignoring strange effects, like a coin's landing stably on its side, it is clear that these sample points are mutually exclusive and collectively exhaustive. Whether or not they are finest grained is a little more subjective—one might

[2] Purists will know that a probability space must also include a field of events, i.e., a collection, $\mathcal{F}$, of subsets of Ω whose probabilities can be meaningfully assigned by $\Pr(\cdot)$. We shall not require that level of rigor in our development.

be interested in the orientation of a fixed reference axis on the coin relative to local magnetic north, in which case the sample space would have to be enlarged. Usually, trivialities such as the preceding example can be disposed of easily. There are cases, however, in which defining the sample space should be done with care to ensure that all the effects of interest are included.

Now let us turn to the probability measure component of $\mathcal{P}$. A probability measure, $\Pr(\cdot)$, assigns *probabilities* to subsets, called *events,* of the sample space Ω. If $A \subseteq \Omega$ is an event,[3] we say that A has occurred on a trial of the experiment whenever the ω that has occurred on that trial is a member of A. The probability that A will occur when the experiment is performed is the number $\Pr(A)$. Because we want $\Pr(A)$ to represent the limit approached by the relative frequency of A in a long sequence of independent trials of the real-world version of the experiment being modeled probabilistically, we impose the following constraints on the probability measure.

- Probabilities are proper fractions, i.e.,

$$0 \le \Pr(A) \le 1, \qquad \text{for all } A \subseteq \Omega. \tag{3.3}$$

- The probability that *something* happens when the experiment is performed is unity, i.e.,

$$\Pr(\Omega) = 1. \tag{3.4}$$

- If A and B are disjoint events, i.e., if they have no sample points in common, then the probability of either A or B occurring equals the sum of their probabilities, viz.

$$\Pr(A \cup B) = \Pr(A) + \Pr(B), \qquad \text{if } A \cap B = \emptyset. \tag{3.5}$$

These properties are obvious features of relative-frequency behavior. For example, consider N trials of the coin-flip-twice experiment whose sample space is given by Eq. 3.2. Let us define events $A \equiv \{HT\}$ and $B \equiv \{TH\}$, and use $N(\cdot)$ to denote the number of times a particular event occurs in the sequence of outcomes. It is then apparent that relative frequencies, $f_N(\cdot) \equiv N(\cdot)/N$, obey

$$0 \le f_N(A) \le 1, \qquad 0 \le f_N(B) \le 1, \tag{3.6}$$

[3] For curious non-purists, here is where a set of events, $\mathcal{F}$, enters probability theory—many probability measures cannot meaningfully assign probabilities to *all* subsets of Ω. The problem arises because of uncountable infinities, and will not be cited further in what follows—we shall allow all subsets of the sample space as events.

$$f_N(\Omega) = 1, \tag{3.7}$$

and

$$f_N(A \cup B) = f_N(\{HT, TH\}) = f_N(A) + f_N(B), \tag{3.8}$$

where the last equality can be justified by Venn diagrams. To complete this coin-flip-twice example, we note that the assignment

$$\Pr(\omega) = \frac{1}{4}, \qquad \text{for all } \omega \in \Omega \tag{3.9}$$

satisfies all the constraints specified for a probability measure, and is the obvious model for two independent flips of a fair coin.

There is one final notion from the basic theory of probability spaces that we shall need—conditional probability. The probability space, $\{\Omega, \Pr(\cdot)\}$, is an *a priori* description of the experiment. For an event A, $\Pr(A)$ measures the likelihood that A will occur when the experiment is performed, given our prior knowledge of the experimental configuration. If the experiment is performed and we are told that event B has occurred, we have additional information, and the likelihood—given this new data—that A has occurred may differ dramatically from $\Pr(A)$. For example, if $A \cap B = \emptyset$, i.e., if A and B are disjoint, then B's having occurred guarantees than A cannot have occurred, even though A's occurrence may be exceedingly likely a priori, e.g., $\Pr(A) = 0.9999$. When we are given the additional information that B has occurred on performance of the experiment, we must replace the a priori probability space, $\{\Omega, \Pr(\cdot)\}$, with the *a posteriori,* or *conditional,* probability space, $\{B, \Pr(\cdot \mid B\,)\}$, in which B takes the role of sample space, and

$$\Pr(\,\cdot \mid B\,) \equiv \frac{\Pr(\cdot \cap B)}{\Pr(B)}, \tag{3.10}$$

is the conditional probability measure.

The structure of a conditional probability space is fairly easy to understand. When we know that B has occurred, all events $A \subseteq \Omega$ which have no sample points in common with B *cannot* have occurred, therefore the sample points that comprise B form a mutually exclusive, collectively exhaustive, finest grained description of all the possible outcomes, given the information that we now have about the experiment's outcome. The *relative* likelihood of occurrence for the sample points in B should not be affected by our knowledge that B has occurred. However, these elementary probabilities need to be scaled—through division by $\Pr(B)$—in order that the conditional probability measure yield its version of the "something always happens" condition, namely

$$\Pr(\,B \mid B\,) = 1. \tag{3.11}$$

Closely related to conditional probability is the concept of statistical independence. Events A and B in the probability space $\mathcal{P} = \{\Omega, \Pr(\cdot)\}$ are said to be *statistically independent* if

$$\Pr(\, A \mid B \,) = \Pr(A), \tag{3.12}$$

i.e., if the likelihood of A's occurring is unaffected by the knowledge that B has occurred. Via Bayes' rule,

$$\Pr(\, B \mid A \,) = \frac{\Pr(\, A \mid B)\Pr(B)}{\Pr(A)}, \tag{3.13}$$

which is a simple consequence of Eq. 3.10, we see that statistically independent events A and B will also satisfy

$$\Pr(\, B \mid A \,) = \Pr(B), \tag{3.14}$$

as well as

$$\Pr(A \cap B) = \Pr(A)\Pr(B). \tag{3.15}$$

Of course, if two events are not statistically independent, they must be statistically dependent—knowledge that B has occurred will then modify our assessment of the likelihood that A has also occurred. Note that disjoint events, of non-zero a priori probability, *must* be dependent.

3.2 Random Variables

Were this chapter to be the text for a first course in probability theory, considerably more time would be spent with the basic probability space structure that was established in the previous section.[4] We, however, have the luxury of assuming prior acquaintance probability theory. Thus, we shall immediately press on to random variables—numerically-valued random occurrences. This material, especially basic results for first and second moments, will comprise the foundation for a great deal of what follows in the theory of optical communications.

[4] The reader who wants to brush up on statistical dependence, for example, might consider the coin-flip-twice experiment with $A = \{HT\}$ and $C = \{HT, TH, HH\}$—try to understand why $\Pr(\, A \mid C\,) = \frac{4}{3}\Pr(A)$ prevails for a fair coin.

Probability Density Functions

Suppose we have a probability space, $\{\Omega, \Pr(\cdot)\}$, for some experiment, and suppose that $x(\cdot)$ is a *deterministic* function that maps sample points, $\omega \in \Omega$, into real numbers, $-\infty < x(\omega) < \infty$. We then say that $x(\cdot)$ is a *random variable,* because our uncertainty as to which sample point in Ω will occur—as quantified by the probability measure $\Pr(\cdot)$—implies, in general, a corresponding uncertainty in the value of the number $x(\omega)$. We have many places in which random variables arise in our generic semiclassical photodetector model, e.g., the light and dark event times, $\{\tau_n\}$ and $\{\tau'_n\}$, and the light and dark event gains, $\{g_n\}$ and $\{g'_n\}$, etc. Our principal task in this subsection is to understand how to specify the statistics—the probability measure—for a random variable.

If x is a random variable, there is no loss of generality in treating its possible values—its *sample values*—as the sample points in the probability space. We can then write $\Omega = \mathcal{R}^1 \equiv \{\, X : -\infty < X < \infty \,\}$ for the sample space associated with x, and the events associated with x are thus subsets of the real line, e.g., $A = \{\, X : 2 \leq |X| < 4 \,\}$, etc. The probability measure for x can then be specified as a *probability density function,* $p_x(X)$, such that

$$\Pr(A) \equiv \Pr(x \in A) = \int_{X \in A} p_x(X)\, dX, \qquad \text{for all } A \subseteq \Omega. \tag{3.16}$$

Because Eq. 3.16 must represent a properly-behaved probability measure, it must obey the constraints laid out in Eqs. 3.3–3.5. The linear, additive nature of integration ensures that Eq. 3.16 satisfies Eq. 3.5. In order that Eq. 3.4 hold, we require that $p_x(X)$ satisfy

$$\int_{-\infty}^{\infty} p_x(X)\, dX = 1; \tag{3.17}$$

in order that Eq. 3.3 be satisfied we also need

$$p_x(X) \geq 0, \qquad \text{for all } X. \tag{3.18}$$

Basically, any non-negative function of a single parameter that integrates to one over the real line can be used as a probability density function, i.e., it generates a meaningful probability measure via Eq. 3.16.

We now introduce two probability density functions which we will encounter frequently in our study of optical communications.

Gaussian random variable The probability density for a Gaussian random variable x is

$$p_x(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-m)^2}{2\sigma^2}}, \tag{3.19}$$

where m is a real-valued constant, and σ^2 is a non-negative constant.

Poisson random variable The probability density for a Poisson random variable x is

$$p_x(X) = \sum_{n=0}^{\infty} \frac{m^n}{n!} e^{-m} \delta(X - n), \tag{3.20}$$

where m is a non-negative constant, and $\delta(\cdot)$ is the unit impulse function.

The Gaussian random variable, whose probability density Eq. 3.19 has been sketched in Fig. 3.1, is an example of a *continuous* random variable, in that its $p_x(X)$ has no singularities. Thus, for an arbitrary sample value X_0, we find that

$$\Pr(x = X_0) = \lim_{\epsilon \to 0} \int_{X_0 - \epsilon}^{X_0} p_x(X)\, dX = 0, \tag{3.21}$$

i.e., there is zero probability that x equals any particular X_0. However, because $p_x(X)$ is non-zero for a continuum of X values, Eq. 3.16 does assign positive probabilities to events which are *intervals* of non-zero width on the real line. Indeed, if $p_x(X)$ is continuous at $X = X_0$, we find—from Eq. 3.16 and the mean value theorem—that

$$\frac{\Pr(\, X_0 - \epsilon < x \leq X_0\,)}{\epsilon} \approx p_x(X_0), \qquad \text{for } \epsilon \text{ sufficiently small}, \tag{3.22}$$

thus justifying our terming $p_x(X)$ a probability *density* function.

The Poisson random variable is an example of a *discrete* random variable, in that its $p_x(X)$ consists of a collection of impulses, which assign non-zero occurrence probabilities to a discrete set of sample values. We shall do general random-variable analyses by means of probability density functions. However, when we are specifically dealing with a discrete random variable, it will be more convenient for us to use its *probability mass function,* which gives the non-zero occurrence probabilities for the various sample values, i.e., the areas of the impulses in the probability density function. In particular, the Poisson random variable's probability mass function, which has been sketched in Fig. 3.2, is

$$P_x(n) \equiv \Pr(x = n) = \frac{m^n}{n!} e^{-m}, \qquad \text{for } n = 0, 1, 2, \ldots; \tag{3.23}$$

for $\Omega \equiv \{\, n : n = 0, 1, 2, \ldots \}$; this mass function constitutes a proper probability measure from which we can calculate event probabilities via

$$\Pr(x \in A) = \sum_{n \in A} P_x(n). \tag{3.24}$$

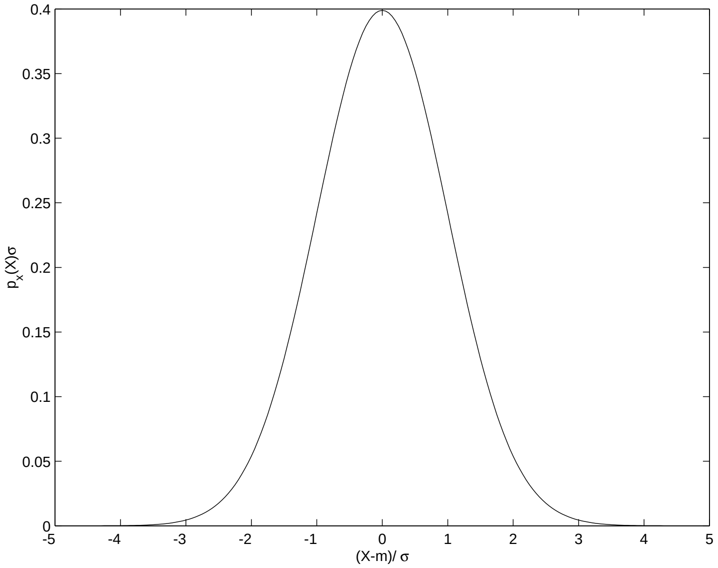

Figure 3.1: Gaussian probability density function

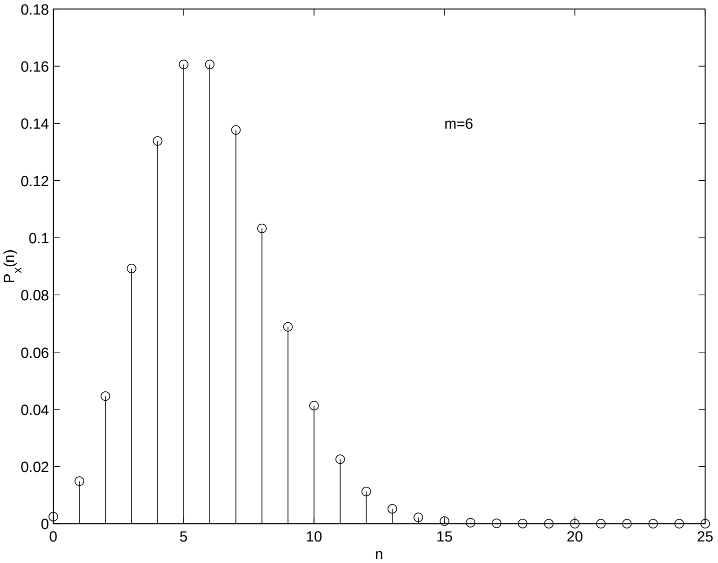

Figure 3.2: Poisson probability mass function; $m = 6$

We shall encounter additional random-variable probability densities as we proceed. The Gaussian and the Poisson cases are of overwhelming importance, however, because they have a variety of useful mathematical properties, to be developed below, *and* because they are good models for a variety of real-world random phenomena. Indeed, the Gaussian variable will be used to model thermal noise, and the Poisson variable will be used to quantify light and dark event times, etc.

Expectation Values

Knowledge of a random variable's probability density function provides complete statistical information about that variable—in principle, we can use the density to calculate $\Pr(x \in A)$ for any event A. Often, however, we focus our attention on simpler ensemble averages, or expectation values, of the random variable. Specifically, if x is a random variable with probability density $p_x(X)$, and $f(\cdot)$ is a deterministic function, we say that

$$E[f(x)] \equiv \int_{-\infty}^{\infty} f(X) p_x(X)\, dX \tag{3.25}$$

is the *ensemble average* of $f(x)$.[5]

Equation 3.25, which is known as the fundamental theorem of expectation, has a simple interpretation. Because x is a random variable, $f(x)$—a deterministic function of a random variable—is also a random variable, i.e., our uncertainty as to which sample value of x will occur translates into an uncertainty as to which sample value of $f(x)$ will occur. Thus, the integral on the right in Eq. 3.25 is summing, over all X, the sample value, $f(X)$, that $f(x)$ takes on when $x = X$ occurs, times its incremental occurrence probability $\Pr(\, X - dX < x \leq X\,) = p_x(X)\, dX$. Of particular interest to us are a variety of special ensemble averages—$f(\cdot)$ choices—described below.

mean value The mean value of the random variable x, denoted m_x or $\bar{x}$, is

$$m_x = \int_{-\infty}^{\infty} X p_x(X)\, dX. \tag{3.26}$$

It represents the *deterministic* part of the random variable x, in that it is not random and

$$\Delta x \equiv x - \bar{x} \tag{3.27}$$

[5] The terms ensemble average, mean value, expectation value, and probability average are interchangeable.

is, via the linearity of expectation,[6] a zero-mean random variable satisfying

$$x = \bar{x} + \Delta x, \tag{3.28}$$

by construction. We call Δx the *noise* part of x.

mean square It is easy to have a non-zero random variable whose mean value is zero—this will be the case for any $p_x(X)$ that is an even function of X. The mean-square value of x,

$$\overline{x^2} \equiv \int_{-\infty}^{\infty} X^2 p_x(X)\, dX, \tag{3.29}$$

is a useful measure of the strength of the random variable x, regardless of the symmetry of its probability density function. Because densities are non-negative, we see that $\overline{x^2} = 0$ implies that $x = 0$ with probability one.

variance The variance of a random variable x, denoted σ_x^2 or $\mathrm{var}(x)$, is its mean-square noise strength, namely

$$\begin{aligned} \sigma_x^2 &\equiv E(\Delta x^2) = E[(x-\bar{x})^2] \\ &= \int_{-\infty}^{\infty} (X-\bar{x})^2 p_x(X)\, dX. \end{aligned} \tag{3.30}$$

By squaring out the integrand in Eq. 3.30 and using the linearity of integration, we can show that

$$\sigma_x^2 = \overline{x^2} - \bar{x}^2. \tag{3.31}$$

Via the remarks made for mean squares, we see that $x = \bar{x}$ prevails with probability one if $\mathrm{var}(x) = 0$, i.e., random variables with zero variance are not really random.

characteristic function Knowing the mean and variance of a random variable x is very useful information, but it does *not* determine the probability density function. The characteristic function of x,

$$M_x(jv) \equiv E(e^{jvx}) = \int_{-\infty}^{\infty} e^{jvX} p_x(X)\, dX, \tag{3.32}$$

[6] Three features of Eq. 3.25 recur endlessly—the average of the sum of random variables is the sum of their averages; the average of a constant times a random variable is the constant times the average of the random variable; and the average of a constant is that constant.

does provide complete information about the density—$M_x(jv)$ is basically the Fourier transform of $p_x(X)$,[7] and Fourier transform pairs form a one-to-one correspondence, viz.

$$p_x(X) = \frac{1}{2\pi}\int_{-\infty}^{\infty} M_x(jv)e^{-jvX}\,dv. \tag{3.33}$$

The well known properties of Fourier transformation thus confer special properties on the characteristic function. Most notable of these is the moment relationship

$$E(x^n) = \left(\frac{\partial^n}{\partial(jv)^n}M_x(jv)\right)\bigg|_{jv=j0}. \tag{3.34}$$

As an illustration of the utility of the mean and variance of a random variable x, let us exploit their respective interpretations as the deterministic part and mean-square noise strength of x by defining a *signal-to-noise ratio,* SNR, according to

$$\text{SNR} \equiv \frac{m_x^2}{\sigma_x^2}, \tag{3.35}$$

i.e., the SNR is the ratio of the squared signal strength—the squared mean of x—to the mean-squared noise strength.[8] With this definition, the well known Chebyschev inequality,

$$\Pr(\,|x-\bar{x}| \ge \epsilon\,) \le \frac{\sigma_x^2}{\epsilon^2}, \tag{3.36}$$

can be rewritten in the form

$$\Pr\left(\frac{|x-\bar{x}|}{|\bar{x}|} \ge \delta\right) \le \frac{1}{\delta^2\text{SNR}}, \quad \text{for } \bar{x} \ne 0. \tag{3.37}$$

Thus, random variables with high SNRs will be close to their mean values with high probability, and single samples of such variables will then yield high-quality measurements of their signal components.

We have assembled in Table 3.1 the means, variances, and characteristic functions of the Gaussian and Poisson variables whose densities were presented earlier. These results can be derived without great difficulty, a task left as

[7] We shall also make use of the bilateral Laplace transform of $p_x(X)$, $M_x(s) \equiv E(e^{sx})$, in some future calculations. This quantity is usually called the *moment generating function* of x.

[8] SNR evaluations will abound in our future efforts. It is important to note that what constitutes *signal,* and what constitutes *noise,* are generally context dependent. It is also worth mentioning that we are taking the theorist's squared-strength ratios, rather than the experimentalist's root-mean-square (rms) ratios, for our SNR formulas.

distribution	mean	variance	characteristic function
Gaussian [Eq. 3.19]	m	σ^2	$\exp\left(jvm - \frac{v^2\sigma^2}{2}\right)$
Poisson [Eq. 3.20]	m	m	$\exp\left[m\left(e^{jv} - 1\right)\right]$

Table 3.1: Moments of the Gaussian and Poisson distributions

an exercise for the reader. We will close our review of the moments of a single random variable x by quoting a probability bound for the Gaussian distribution—whose derivation will be given in a later chapter—for comparison with the Chebyschev inequality Eq. 3.37:

$$\Pr\left(\frac{|x - \bar{x}|}{|\bar{x}|} \geq \delta\right) \leq e^{-\delta^2 \mathrm{SNR}/2}, \qquad \text{for } x \text{ Gaussian distributed.} \tag{3.38}$$

Here we see an *exponential* decrease with increasing SNR of the probability that $|x - \bar{x}|$ exceeds a threshold; the Chebyschev inequality predicts a much weaker *algebraic* decay with increasing SNR. The Chebyschev inequality is a very general result, which only requires knowledge of the mean and variance of the random variable for its evaluation. The cost of this generality is the weakness of the resulting bound, i.e., when we use the probability density of x we are apt to find a much smaller probability that $|x - \bar{x}|$ exceeds the prescribed threshold than the upper limit set by the Chebyschev inequality.

Transformations

If x is a random variable and $f(\cdot)$ is a deterministic function, we know that $y \equiv f(x)$ is a random variable, and we know how to calculate its moments given knowledge of the probability density of x. Sometimes, it is important or convenient to use $p_x(X)$ and the known *transformation* $f(\cdot)$ to obtain the probability density of y. A simple example is the following. Suppose x is a temperature Fahrenheit obtained from a sensor monitoring some fluctuating chemical reaction. We might well prefer to analyze this experiment in Celsius temperature units, i.e., by dealing with

$$y = \frac{5}{9}(x - 32). \tag{3.39}$$

There are a variety of techniques for deriving the density, $p_y(Y)$, for a deterministic transformation, $y = f(x)$, of a random variable, x, of known density, $p_x(X)$. The most general of these various procedures, and the only one that we shall discuss at present, is the *method of events.*

The method of events for computing $p_y(Y)$ proceeds as follows. The probability density function can always be found as the derivative of the probability distribution function,

$$p_y(Y) = \frac{dF_y(Y)}{dY}, \tag{3.40}$$

where

$$F_y(Y) \equiv \Pr(y \leq Y). \tag{3.41}$$

The distribution function, in turn, can be computed, via the known transformation and the given $p_x(X)$, from the following development

$$\Pr(y \leq Y) = \Pr(f(x) \leq Y) = \int_{\{\, X : f(X) \leq Y\,\}} p_x(X)\, dX. \tag{3.42}$$

All the hard work then revolves around performing the necessary setting of limits, integration, and differentiation.

For the transformation $y = ax + b$, where $a > 0$ and b are constants, the method of events yields

$$\begin{aligned} F_y(Y) &= \int_{\{\, X : aX+b \leq Y\,\}} p_x(X)\, dX \\ &= \int_{-\infty}^{(Y-b)/a} p_x(X)\, dX, \end{aligned} \tag{3.43}$$

which may be differentiated via the Leibniz rule

$$\begin{aligned} \frac{d}{dY}\int_{l(Y)}^{u(Y)} g(X,Y)\, dX &= g(u(Y),Y)\frac{du(Y)}{dY} - g(l(Y),Y)\frac{dl(Y)}{dY} + \\ & \int_{l(Y)}^{u(Y)} \frac{\partial g(X,Y)}{\partial Y}\, dX, \end{aligned} \tag{3.44}$$

with the result

$$p_y(Y) = \frac{1}{a} p_x\left(\frac{Y-b}{a}\right). \tag{3.45}$$

For $a < 0$ a similar calculation produces the above result with a replaced by $|a|$ in the coefficient multiplying the p_x term.

We could complete the $ax + b$ transformation example by substituting in the a and b values associated with the Fahrenheit-to-Celsius conversion. It is much more interesting to leave a and b arbitrary, and see that $p_x(X)$ Gaussian with mean m_x and variance σ_x^2 results in

$$p_y(Y) = \frac{1}{\sqrt{2\pi a^2\sigma_x^2}} \exp\left\{-\left[\frac{Y-(am_x+b)}{2a^2\sigma_x^2}\right]^2\right\}, \tag{3.46}$$

which shows that y is also Gaussian, with

$$m_y = am_x + b, \tag{3.47}$$

and

$$\sigma_y^2 = a^2\sigma_x^2. \tag{3.48}$$

Equations 3.47 and 3.48 do *not* depend on the density function of x being Gaussian; they are immediate consequences of the $ax+b$ transformation and the linearity of expectation. Direct derivations of them, namely

$$m_y = E(y) = E(ax+b) = aE(x) + b, \tag{3.49}$$

and

$$\begin{aligned} \sigma_y^2 &= E(\Delta y^2) \\ &= E\{[(ax+b)-(a\bar{x}+b)]^2\} \\ &= E[(a\Delta x)^2] = a^2\sigma_x^2, \end{aligned} \tag{3.50}$$

are of interest for future use. Note, from these derivations, that this linear—strictly-speaking affine—transformation has the following properties.

- The mean output of a deterministic linear transformation driven by a random variable is the mean input passed through the transformation.

- The noise part of the output of a deterministic linear transformation driven by a random variable is the noise part of the input passed through the transformation.

One final comment on transformations of a single random variable. The fact that $y = ax+b$ is Gaussian, for all constants a and b, when x is Gaussian is no accident. In fact, if $y = ax+b$ is Gaussian for all choice of the constants, then x *must* be Gaussian. This can be shown via characteristic functions, and forms the basis for the important case of jointly Gaussian random variables to be seen below.

3.3 Two Joint Random Variables

A single random variable is insufficient for the probabilistic analyses we shall pursue—the photodetection systems we want to model produce collections of *joint* random variables, i.e., a multiplicity of numerically valued outcomes. We begin, in this section, with two joint random variables—the 2-D case. In the next section we generalize to N-D case, namely, N joint random variables. In the chapter that follows, we shall see how the latter case extends naturally into the study of random waveforms.

2-D Joint Probability Densities

Suppose we have a probability space, $\mathcal{P} = \{\Omega, \Pr(\cdot)\}$, and two deterministic functions, $x(\cdot)$ and $y(\cdot)$, which map sample points $\omega \in \Omega$ into real numbers, $-\infty < x(\omega) < \infty$ and $-\infty < y(\omega) < \infty$. This is the natural two-dimensional extension of the case addressed in the last section. We say that x and y are joint random variables on $\mathcal{P}$. They are random variables because our uncertainty as to which ω will occur when the experiment is performed translates into corresponding uncertainties in the values of $x(\omega)$ and $y(\omega)$. They are *joint* random variables because they are defined on the *same* probability space, thus it is meaningful—and, for completeness, it is necessary—to address the joint probability that x falls into some particular interval *and* y falls into some other specific interval on the same trial of the experiment. For these joint random variables, we will briefly indicate how the key single random variable notions of probability density functions, expectations, and transformations extend. We shall also translate the basic probability space concepts of conditional probability and statistical independence into joint random variable terms.

Without loss of generality, we can regard the sample space for two joint random variables x and y as the X–Y plane, viz. $\Omega = \mathcal{R}^2 \equiv \{\,(X,Y) : -\infty < X, Y < \infty\,\}$, and we can specify the probability measure for events A via a joint probability density function $p_{x,y}(X,Y)$ and the relation

$$\Pr(A) = \iint_{(X,Y)\in A} dX\,dY\,p_{x,y}(X,Y), \qquad \text{for all } A \subseteq \Omega. \tag{3.51}$$

Equation 3.51 is the 2-D version of Eq. 3.16. The fundamental probability measure constraints, Eqs. 3.3–3.5, led to the restrictions embodied in Eqs. 3.17 and 3.18 on functions of a single variable, $p_x(\cdot)$, which can be probability densities. These same probability measure constraints imply that 2-D joint probability density functions must obey

$$\int_{-\infty}^{\infty} dX \int_{-\infty}^{\infty} dY\,p_{x,y}(X,Y) = 1, \tag{3.52}$$

and

$$p_{x,y}(X,Y) \geq 0, \qquad \text{for all } X,\, Y; \tag{3.53}$$

i.e., any deterministic non-negative function of two variables that integrates to one over the X–Y plane is a valid $p_{x,y}(X,Y)$. Taking $A = \{\, (X,Y) : X_0 - \epsilon < X \leq X_0, Y_0 - \epsilon < Y \leq Y_0 \,\}$ with $\epsilon \to 0$ in Eq. 3.51 shows that $p_{x,y}(X_0, Y_0)$ is the *joint* probability per unit area of obtaining $x = X_0$ *and* $y = Y_0$ when the experiment is performed.[9]

An immediate issue that arises in conjunction with two-dimensional joint probability densities is their relationship with the one-dimensional, *marginal,* densities encountered previously. Suppose that x and y are joint random variables and that B is a region on the real line. Then, we can compute $\Pr(x \in B)$ either directly—via Eq. 3.16 using the marginal density $p_x(X)$—or indirectly—via Eq. 3.51 and the two-dimensional event

$$A \equiv \{\, (X,Y) : X \in B, -\infty < Y < \infty \,\}. \tag{3.54}$$

In either case, we must arrive at the same number $\Pr(x \in B)$. Thus, since B was arbitrary, it must be that

$$p_x(X) = \int_{-\infty}^{\infty} p_{x,y}(X,Y)\, dY \quad \text{for all } X; \tag{3.55}$$

interchanging the roles of x and y in this derivation shows that the marginal density for y can be found from the x–y joint density by integrating out over X from $-\infty$ to ∞.

We have shown that *marginal* statistics can be found from *joint* statistics by integrating out the unwanted variables—if we know the function $p_{x,y}(X,Y)$ we can, in principle, calculate the functions $p_x(X)$ and $p_y(Y)$. The converse is not generally true, i.e., knowing the functions $p_x(X)$ and $p_y(Y)$ places constraints on the permissible joint distribution $p_{x,y}(X,Y)$, but does *not* determine the joint density.[10] There is one case in which the marginal densities determine the joint density; it is when the two variables are statistically independent.

For x and y joint random variables, with joint density $p_{x,y}(X,Y)$, and $B \subseteq \mathcal{R}^1$ a region on the real line, the *a priori* probability of having $x \in B$ occur is computed from the marginal density of x, as described above. If, upon performance of the experiment, we observe that $y = Y$ has occurred, then we

[9] Here, (X_0, Y_0) is assumed to be a point of continuity of $p_{x,y}(X,Y)$, cf. Eq. 3.22.

[10] One of the home problems for this chapter exhibits an infinite family of $p_{x,y}$ functions associated with a given pair of p_x and p_y functions.

need to find the *conditional* probability, $\Pr(\, x \in B \mid y = Y \,)$, to assess the likelihood of $x \in B$. With the aid of Eq. 3.10 it can be shown that

$$\Pr(\, x \in B \mid y = Y \,) = \int_{X \in B} p_{x|y}(\, X \mid Y \,)\, dX, \tag{3.56}$$

where

$$p_{x|y}(\, X \mid Y \,) \equiv \frac{p_{x,y}(X,Y)}{p_y(Y)} \tag{3.57}$$

is the conditional probability density for x given $y = Y$ has occurred. The random variables x and y are said to be statistically independent if and only if

$$\Pr(\, x \in B \mid y = Y \,) = \Pr(x \in B), \qquad \text{for all } B \text{ and } Y. \tag{3.58}$$

Equations 3.16 and 3.57 then imply that x and y are statistically independent if and only if

$$p_{x|y}(\, X \mid Y \,) = p_x(X), \qquad \text{for all } X,\, Y, \tag{3.59}$$

which is equivalent to the *functional* factorization of the joint density into the product of its marginals, viz.

$$p_{x,y}(X,Y) = p_x(X)p_y(Y), \qquad \text{for all } X,\, Y. \tag{3.60}$$

In doing calculations, it is a great convenience to deal with statistically independent random variables. In doing communication theory analyses we are often confronted with statistically dependent random variables, i.e., the random variable that is received may be a noise-corrupted version of the random variable that was sent.

We shall complete our brief examination of 2-D joint random variables by augmenting our knowledge of expectations, transformations, and Gaussian random variables—all topics that will be of use in succeeding chapters.

2-D Expectations

The fundamental theorem of expectation, Eq. 3.25, has the following extension to two dimensions

$$E[f(x,y)] = \int_{-\infty}^{\infty} dX \int_{-\infty}^{\infty} dY\, f(X,Y)p_{x,y}(X,Y), \tag{3.61}$$

for any deterministic function of two variables, $f(\cdot,\cdot)$. The relation between joint and marginal densities guarantees that Eqs. 3.25 and 3.61 yield identical results for the mean values, mean-square values, variances, and characteristic functions of x and y, because they only involve marginal statistics. Our present interest is in the following two expectations which involve joint statistics.

covariance The covariance of x and y, denoted λ_{xy} or $\text{cov}(x, y)$, is the average of the product of their noise parts, i.e.,

$$\lambda_{xy} \equiv E(\Delta x \Delta y) = E(xy) - \bar{x}\bar{y}, \tag{3.62}$$

where the linearity of expectation has been used to obtain the second equality, and $\Delta x \equiv x - \bar{x}$ as given earlier in our discussion of variance.

joint characteristic function The joint characteristic function of x and y, denoted $M_{x,y}(jv_x, jv_y)$, is the two-dimensional Fourier transform of the joint density, i.e.,

$$M_{x,y}(jv_x, jv_y) \equiv E[\exp(jv_x x + jv_y y)], \tag{3.63}$$

cf. Eq. 3.32.

The covariance plays a key role in second-moment calculations involving linear transformations of x and y, and it provides an imperfect but simple measure of the statistical dependence between x and y. Suppose a, b, and c are constants, and we construct a new random variable z from x and y via the linear transformation

$$z = ax + by + c. \tag{3.64}$$

Following the precepts established in deriving Eqs. 3.47 and 3.48, we have that

$$m_z = E(ax + by + c) = E(ax) + E(by) + E(c) = am_x + bm_y + c, \tag{3.65}$$

and

$$\begin{aligned} \sigma_z^2 &= E(\Delta z^2) \\ &= E\{[(ax + by + c) - (a\bar{x} + b\bar{y} + c)]^2\} \\ &= E[(a\Delta x + b\Delta y)^2] = a^2\sigma_x^2 + 2ab\lambda_{xy} + b^2\sigma_y^2. \end{aligned} \tag{3.66}$$

The last equality in Eq. 3.66 is obtained by squaring out inside the expectation and using manipulations similar to those employed in Eq. 3.65—the average of the sum is the sum of the averages, etc.

Because the variance is a *second* moment, the variance of a sum of random variables is *not* usually the sum of their variances, as can be seen from the above calculations with $a = b = 1$ and $c = 0$. If, however, the two random variables are *uncorrelated,* i.e., their covariance is zero, then the variance of their sum *is* the sum of their variances. It turns out that statistically independent random variables are always uncorrelated, as the following argument shows:[11]

$$\begin{aligned} E(\Delta x \Delta y) &= \int_{-\infty}^{\infty} dX \int_{-\infty}^{\infty} dY\,(X-\bar{x})(Y-\bar{y})p_{x,y}(X,Y) \\ &= \int_{-\infty}^{\infty} dX \int_{-\infty}^{\infty} dY\,(X-\bar{x})(Y-\bar{y})p_x(X)p_y(Y) \\ &= E(\Delta x)E(\Delta y) = 0, \end{aligned} \tag{3.67}$$

where the second equality used the statistical independence of x and y, and the last equality used the zero-mean nature of the noise-parts of x and y. The converse result is *not* true—there are uncorrelated random variables that are statistically *dependent.*

The last paragraph addressed the *minimum* possible value of $|\lambda_{xy}|$; there is also useful information to be mined from the *maximum* possible value of $|\lambda_{xy}|$. For all a, b, and c, the random variable z generated by the transformation Eq. 3.64 must have a non-negative variance. A little differential calculus applied to Eq. 3.66 then reveals that

$$|\lambda_{xy}| \le \sigma_x \sigma_y, \tag{3.68}$$

with equality if and only if

$$\frac{y-m_y}{\sigma_y} = \text{sgn}(\lambda_{xy})\frac{x-m_x}{\sigma_x}, \qquad \text{with probability one,} \tag{3.69}$$

where $\text{sgn}(\cdot)$ is the signum function,

$$\text{sgn}(t) = \begin{cases} 1, & \text{for } t \ge 0, \\ -1, & \text{for } t < 0. \end{cases} \tag{3.70}$$

In view of the form of the upper limit on $|\lambda_{xy}|$, Eq. 3.68, it is useful to introduce the *correlation coefficient,* ρ_{xy}, according to[12]

$$\rho_{xy} \equiv \frac{\lambda_{xy}}{\sigma_x \sigma_y}. \tag{3.71}$$

Uncorrelated random variables x and y have $\rho_{xy} = 0$. Random variables x and y with $|\rho_{xy}| = 1$ are said to be *completely* correlated, because knowledge of the value of one determines the value of the other through Eq. 3.69—two

[11] This proof can easily be extended to demonstrate that $E[f(x)g(y)] = E[f(x)]E[g(y)]$ for statistically independent x and y, where $f(\cdot)$ and $g(\cdot)$ are arbitrary deterministic functions.

[12] Really, ρ_{xy} should be the *covariance* coefficient; the term correlation coefficient is standard, however.

completely correlated random variables are as *dependent* as a pair of joint random variables can be.

Before passing on to transformations, a few comments about the joint characteristic function merit inclusion. The lengthy discussion of covariance, in conjunction with our earlier development of the utilities of mean values and variances, argues eloquently for the use of first and second moments in doing probabilistic studies. This is doubly true when linear transformations are involved, for, as we have seen, the first and second moments of the transformed variable can easily be found from the first and second moments of the input variables and the transformation coefficients. Our enthusiasm for the low-order moments $\{m_x, m_y, \sigma_x^2, \sigma_y^2, \lambda_{xy}\}$ must be tempered by the knowledge that they do *not,* in general, provide the same information as the joint probability density function $p_{x,y}(X, Y)$. The joint characteristic function, like its one-dimensional cousin, *does* provide a complete statistical characterization—$p_{x,y}(X, Y)$ can be found from $M_{x,y}(jv_x, jv_y)$ via 2-D inverse Fourier transformation. Moreover, the 2-D versions of the standard Fourier transform properties can be used to prove the 2-D moment relation

$$E[x^n y^m] = \left(\frac{\partial^{n+m}}{\partial(jv_x)^n \partial(jv_y)^m} M_{x,y}(jv_x, jv_y) \right) \Bigg|_{jv_x = jv_y = j0}, \tag{3.72}$$

which can be reduced to our earlier result for a single random variable, Eq. 3.34, by setting $m = 0$ and recognizing that $M_x(jv) = M_{x,y}(jv, j0)$.

2-D Transformations

Let x and y be joint random variables, and let $f(\cdot, \cdot)$ and $g(\cdot, \cdot)$ be two deterministic functions of two variables. Then $z \equiv f(x, y)$ and $u \equiv g(x, y)$ comprise a 2-D to 2-D transformation of x and y into new joint random variables z and u. If the joint density of z and u is sought, given knowledge of the transformation and the joint input-variable density $p_{x,y}(X, Y)$, we can again turn to the method of events, now in 2-D form. We find the joint probability density function as the second mixed partial derivative of the joint distribution function,

$$p_{z,u}(Z, U) = \frac{\partial^2 F_{z,u}(Z, U)}{\partial Z \partial U}, \tag{3.73}$$

where

$$F_{z,u}(Z, U) \equiv \Pr(z \leq Z, u \leq U). \tag{3.74}$$

The distribution function is found by means of

$$\Pr(z \leq Z, u \leq U) \;=\; \Pr(f(x, y) \leq Z, g(x, y) \leq U)$$

$$\begin{aligned} &= \iint_{(X,Y)\in A} dX\, dY\, p_{x,y}(X,Y), && (3.75)\\ \text{for } A &\equiv \{\, (X,Y) : f(X,Y) \leq Z, g(X,Y) \leq U \,\}, && (3.76) \end{aligned}$$

and calculational elbow grease—the latter sometimes involves relentless application of the Leibniz rule.

In the case of a linear transformation, there is a useful alternative procedure. As a prelude to our treatment of 2-D Gaussian random variables, said procedure is worth exploring. Suppose that z and u are obtained from x and y via

$$\begin{aligned} z &= ax + by + c, && (3.77)\\ u &= dx + ey + f, && (3.78) \end{aligned}$$

where a through f are constants. Here we can directly compute the joint characteristic function of z and u:

$$\begin{aligned} M_{z,u}(jv_z, jv_u) &= E[e^{j(av_z+dv_u)x+j(bv_z+ev_u)y+j(cv_z+fv_u)}] && (3.79)\\ &= M_{x,y}[j(av_z+dv_u), j(bv_z+ev_u)]e^{j(cv_z+fv_u)}. && (3.80) \end{aligned}$$

This form conveniently yields marginal statistics, because

$$M_z(jv) = M_{z,u}(jv, j0), \qquad (3.81)$$

etc. Let us take this approach to determine the probability density for $z = x+y$, when x and y are statistically independent random variables with known marginal densities p_x and p_y. Setting $a = b = 1$ and $c = 0$ in $M_{z,u}(jv, j0)$, we find that

$$M_z(jv) = M_{x,y}(jv, jv) = M_x(jv)M_y(jv), \qquad (3.82)$$

where the last equality makes use of statistical independence. The convolution multiplication theorem of Fourier analysis now can be used to inverse transform this characteristic function equation, with the following result

$$p_z(Z) = \int_{-\infty}^{\infty} p_x(X)p_y(Z-X)\, dX = p_x * p_y, \qquad (3.83)$$

where $*$ denotes convolution.

Two special applications of Eq. 3.82 are of particular note—when x and y are statistically independent and Poisson distributed, and when x and y are statistically independent and Gaussian distributed. In the Poisson situation, Eq. 3.82 in conjunction with Table 3.1 prove that $z = x + y$ will be Poisson distributed; in the Gaussian case, these results imply that z will be Gaussian distributed. The low-order moments needed to complete specification of $p_z(Z)$ in either of these circumstances are given by limiting forms of our linear-transformation formulas, Eqs. 3.65 and 3.66.

2-D Gaussian Random Variables

Let x and y be joint random variables. They are said to be *jointly Gaussian* if

$$p_{x,y}(X,Y) = \frac{\exp\left[-\frac{\sigma_y^2(X-m_x)^2-2\lambda_{xy}(X-m_x)(Y-m_y)+\sigma_x^2(Y-m_y)^2}{2(\sigma_x^2\sigma_y^2-\lambda_{xy}^2)}\right]}{2\pi\sqrt{\sigma_x^2\sigma_y^2-\lambda_{xy}^2}} \tag{3.84}$$

is their joint probability density function,[13] where, as the notation suggests, m_x, m_y, σ_x^2, σ_y^2, and λ_{xy} are the means, variances, and covariance, respectively, of x and y. This joint probability density has been sketched in Fig. 3.3; it is a 2-D generalization of the bell-shaped curve, Fig. 3.1, of the 1-D Gaussian density, Eq. 3.19. With some analytic geometry, it can be shown that equal-height contours of the jointly Gaussian density are ellipses. With some Fourier integration, it can be shown that the joint characteristic function associated with Eq. 3.84 is (cf. Table 3.1 for the 1-D Gaussian case)

$$M_{x,y}(jv_x, jv_y) = \exp\left[j(v_xm_x+v_ym_y) - \frac{v_x^2\sigma_x^2+2v_xv_y\lambda_{xy}+v_y^2\sigma_y^2}{2}\right]. \tag{3.85}$$

These 2-D formulas embody a variety of important properties, as spelled out below:

joint vs. marginal statistics If x and y are *jointly* Gaussian, then they are also *marginally* Gaussian—this is an immediate consequence of Eq. 3.85 with v_x or v_y set equal to 0. The converse is not generally true—a pair of marginally Gaussian random variables need *not* be jointly Gaussian.

conditional statistics If x and y are jointly Gaussian, then Eqs. 3.57,3.84, 3.19, plus the customary algebraic elbow grease yield

$$p_{x|y}(\,X \mid Y\,) = \frac{\exp\left\{-\frac{\left[X-m_x-\rho_{xy}\frac{\sigma_x}{\sigma_y}(Y-m_y)\right]^2}{2\sigma_x^2(1-\rho_{xy}^2)}\right\}}{\sqrt{2\pi\sigma_x^2(1-\rho_{xy}^2)}}, \tag{3.86}$$

for the conditional density of $x = X$ given $y = Y$ has occurred. Comparing Eqs 3.19 and 3.86 shows that x is *still* Gaussian distributed, given

[13] This density is well behaved so long as x and y are *not* completely correlated. For x and y completely correlated, Eq. 3.84 is impulsive along the line dictated by Eq. 3.69, as it should be.

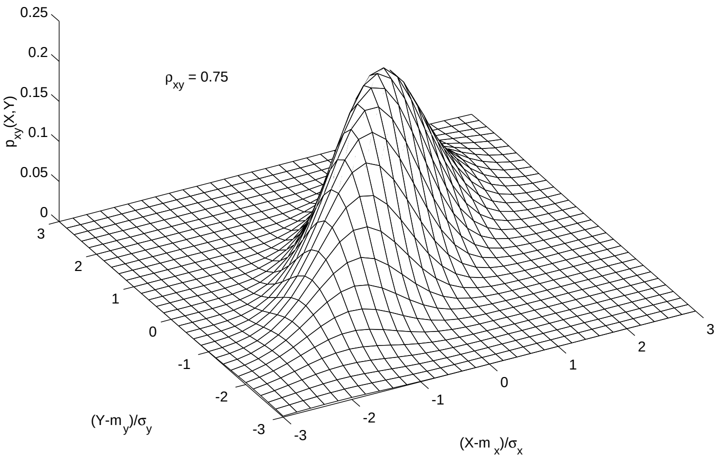

Figure 3.3: Jointly Gaussian probability density function; $\rho = 0.75$

$y = Y$, with conditional mean

$$E(\, x \mid y = Y \,) = m_x + \rho_{xy} \frac{\sigma_x}{\sigma_y} (Y - m_y), \tag{3.87}$$

and conditional variance

$$\mathrm{var}(\, x \mid y = Y \,) = \sigma_x^2 (1 - \rho_{xy}^2). \tag{3.88}$$

linear transformations If x and y are jointly Gaussian, and z is generated from them via the linear transformation

$$z = ax + by + c, \tag{3.89}$$

where a, b, and c are constants, then z is a 1-D Gaussian random variable, with mean

$$m_z = am_x + bm_y + c, \tag{3.90}$$

and variance

$$\sigma_z^2 = a^2 \sigma_x^2 + 2ab\lambda_{xy} + b^2 \sigma_y^2. \tag{3.91}$$

These moment equations are general consequences of the linearity of the transformation, recall Eqs. 3.64–3.66. That p_z will be Gaussian is easily shown from Eq. 3.85 and the characteristic function approach to linear-transformation calculations.

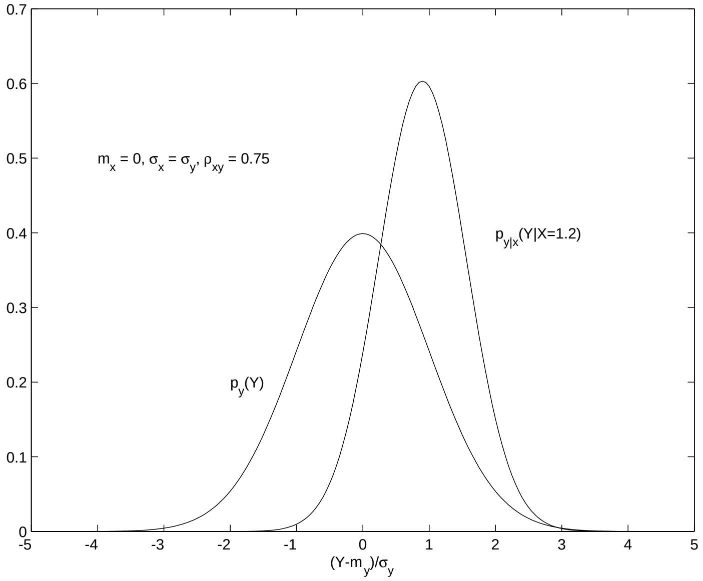

Figure 3.4: Conditional density, $p_{x|y}$, and marginal density, p_x, when x and y are jointly Gaussian; $m_x = 0, \sigma_x = \sigma_y, \rho_{xy} = 0.75, X = 1.2$

We will soon generalize these properties to N-D Gaussian random variables, and later we will see them again in the guise of Gaussian random processes. There are several points worth noting now, so that we may re-emphasize them in what follows. Many probability calculations involving jointly Gaussian random variables are appreciably simpler than general-case results. If x and y are known to be jointly Gaussian, then specification of their joint probability density function can be completed by giving values for their first and second moments; for arbitrary joint random variables, these low-order moments do *not* determine the joint density. For jointly Gaussian x and y, we do not have to integrate to find the marginal statistics, as is generally the case. If x and y are jointly Gaussian, then they will be statistically independent if and only if they are uncorrelated, as can be seen from comparing Eq. 3.19 and Eq. 3.86 with $\rho_{xy} = 0$, or from Fig. 3.4;[14] arbitrary uncorrelated random variables need *not* be statistically independent.

The linear transformation property deserves special attention. We defined x and y to be jointly Gaussian if their joint probability density function had

[14] Figure 3.4 also illustrates how $p_{x|y}$ collapses to a unit-area impulse at $X = m_x + \mathrm{sgn}(\rho_{xy})\frac{\sigma_x}{\sigma_y}(Y - m_y)$ when x and y become completely correlated, cf. Eq. 3.69.

the form Eq. 3.84.[15] From this definition we concluded that any linear transformation of this jointly Gaussian pair yields a 1-D Gaussian random variable. This *closure* under linear transformations is powerful enough to *determine* the jointly Gaussian density, i.e., the *only* joint probability density for which $z = ax + by + c$ will be a 1-D Gaussian random variable regardless of the choice of the constants a through c is Eq. 3.84.[16] Indeed, we will use this linear-closure approach, in the next section, for our definition of N jointly Gaussian random variables.

One final comment, of a *physical* nature, regarding Gaussian random variables will serve as a useful cap to the present development. Gaussian statistics are good models for random experiments in which a *macroscopic* observation is comprised of a large number of more-or-less small, more-or-less independent, *microcopic* contributions—thermal noise and high-density shot noise are two examples of such circumstances. Mathematically, the preceding statement, made rigorous, constitutes the *Central Limit Theorem* of basic probability theory.

3.4 Random Vectors

The transition from two joint random variables to N joint random variables is primarily one of notation—no new probabilistic concepts need to be introduced.[17] Consider a random experiment whose outcomes comprise N real numbers—an ordered N-tuple—$(x_1, x_2, \ldots, x_N)$. A probabilistic model for this experiment will represent these values as N joint random variables in a probability space, $\mathcal{P}$, whose sample space is

$$\Omega = \mathcal{R}^N \equiv \{\, (X_1, X_2, \ldots, X_N) : -\infty < X_n < \infty, \quad 1 \leq n \leq N \,\}, \qquad (3.92)$$

and whose probability measure, $\Pr(\cdot)$, is given by the joint probability density $p_{x_1, x_2, \ldots, x_N}(X_1, X_2, \ldots, X_N)$.[18] The probability that the N-tuple that occurs will fall in $A \subseteq \mathcal{R}^N$ is found by integrating the joint density over A. We won't

[15] Equivalently, x and y are jointly Gaussian if their joint characteristic function obeys Eq. 3.85.

[16] Note that the *moments* of z will depend on a through c, as given by Eqs. 3.65 and 3.66, but the density of z must obey Eq. 3.19 if x and y are jointly Gaussian.

[17] However, unlike the material on one and two random variables, we shall neither assume great prior familiarity with the N-D case, nor shall need to do many complicated N-D calculations in the chapters that follow.

[18] Here, we have dispensed with the formality of initially defining N *deterministic* functions, $x_1(\omega), x_2(\omega), \ldots, x_N(\omega)$, which map sample points $\omega \in \Omega$ into real numbers for some abstract $\mathcal{P} = \{\Omega, \Pr(\cdot)\}$.

exhibit the equation, because it is notationally cumbersome. Instead, we shall deal with general issues of N joint random variables in vector notation.

Vector Notation

The N joint random variables $x_1, x_2, \ldots, x_N$ are equivalent to a random N-vector,

$$\mathbf{x} \equiv \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}, \tag{3.93}$$

with probability density function[19]

$$p_{\mathbf{x}}(\mathbf{X}) \equiv p_{x_1,x_2,\ldots,x_N}(X_1, X_2, \ldots, X_N), \tag{3.94}$$

where the dummy-variable N-vector $\mathbf{X}$ is

$$\mathbf{X} \equiv \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}. \tag{3.95}$$

Thus, for $A \subseteq \mathcal{R}^N$, we have that

$$\Pr(\mathbf{x} \in A) = \int_{\mathbf{X} \in A} p_{\mathbf{x}}(\mathbf{X})\, \mathbf{dX}. \tag{3.96}$$

The $\mathbf{dX}$ appearing on the right in Eq. 3.96 is the N-D differential volume element, i.e., the notation implies integration over a region in $\mathcal{R}^N$. Although calculating *numerical* results from Eq. 3.96 may be quite tedious, developing *conceptual* results from this formula is relatively easy. The forms of Eq. 3.16—for a single random variable—and Eq. 3.96—for N joint random variables—are so similar that we can immediately draw the following conclusions. Any deterministic function, $p_{\mathbf{x}}(\mathbf{X})$ for $\mathbf{X} \in \mathcal{R}^N$, that is non-negative and integrates to one over $\mathcal{R}^N$ can serve as a probability density for a random N-vector $\mathbf{x}$. At its points of continuity, $p_{\mathbf{x}}(\mathbf{X})$ is the probability per unit N-D volume that $\mathbf{x}$=$\mathbf{X}$ will occur.

Several vector versions of concepts we have seen earlier can be developed by considering a random $(N+M)$-vector $\mathbf{z}$, whose first N dimensions comprise

[19] In words, the probability density function for a random vector, $\mathbf{x}$, is the *joint* probability density function for its components, $\{x_1, x_2, \ldots, x_N\}$.

a random N-vector $\mathbf{x}$, and whose last M dimensions form a random M-vector $\mathbf{y}$. In partitioned notation, we can then write

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ --- \\ \mathbf{y} \end{bmatrix}, \tag{3.97}$$

and use

$$p_{\mathbf{x},\mathbf{y}}(\mathbf{X},\mathbf{Y}) \equiv p_{\mathbf{z}}(\mathbf{Z}), \tag{3.98}$$

where

$$\mathbf{Z} = \begin{bmatrix} \mathbf{X} \\ --- \\ \mathbf{Y} \end{bmatrix}, \tag{3.99}$$

for the joint probability density function of the *vectors* $\mathbf{x}$ and $\mathbf{y}$. We now have, via analogy with the 2-D theory, that the *marginal* statistics of the vector $\mathbf{x}$ are found by integrating out the $\mathbf{Y}$ dependence of $p_{\mathbf{x},\mathbf{y}}$, viz.

$$p_{\mathbf{x}}(\mathbf{X}) = \int_{\mathbf{Y}\in\mathcal{R}^M} p_{\mathbf{x},\mathbf{y}}(\mathbf{X},\mathbf{Y})\,\mathbf{dY}. \tag{3.100}$$

Similarly, we have that the conditional probability density for $\mathbf{x}$, given $\mathbf{y} = \mathbf{Y}$ has occurred, is

$$p_{\mathbf{x}|\mathbf{y}}(\mathbf{X} \mid \mathbf{Y}) = \frac{p_{\mathbf{x},\mathbf{y}}(\mathbf{X},\mathbf{Y})}{p_{\mathbf{y}}(\mathbf{Y})}. \tag{3.101}$$

The random vector $\mathbf{x}$ is said to be statistically independent of the random vector $\mathbf{y}$ if and only its conditional density, from Eq. 3.101, is equal to its a priori density, $p_{\mathbf{x}}(\mathbf{X})$, for all $\mathbf{X}$ and $\mathbf{Y}$.[20]

Expectations

Suppose $\mathbf{x}$ is a random N-vector with probability density function $p_{\mathbf{x}}(\mathbf{X})$. If $f(\mathbf{X})$, for $\mathbf{X} \in \mathcal{R}^N$, is a deterministic *scalar* function of an N-vector argument, then $f(\mathbf{x})$ is a random variable whose mean value can be found from

$$E[f(\mathbf{x})] = \int_{\mathbf{X}\in\mathcal{R}^N} f(\mathbf{X})p_{\mathbf{x}}(\mathbf{X})\,\mathbf{dX}. \tag{3.102}$$

[20] Note that $\mathbf{x}$ and $\mathbf{y}$ being statistically independent implies that x_n and y_m are statistically independent for all $1 \le n \le N$ and $1 \le m \le M$. It does *not* imply that x_n and $x_{n'}$ are statistically independent for $n \neq n'$.

If $\mathbf{f}(\mathbf{X})$ is an M-vector deterministic function of an N-vector argument,

$$\mathbf{f}(\mathbf{X}) \equiv \begin{bmatrix} f_1(\mathbf{X}) \\ f_2(\mathbf{X}) \\ \vdots \\ f_M(\mathbf{X}) \end{bmatrix}, \tag{3.103}$$

then $\mathbf{f}(\mathbf{x})$ is a random M-vector, whose expectation value—its *mean vector*—is defined as follows

$$E[\mathbf{f}(\mathbf{x})] \equiv \begin{bmatrix} E[f_1(\mathbf{x})] \\ E[f_2(\mathbf{x})] \\ \vdots \\ E[f_M(\mathbf{x})] \end{bmatrix}, \tag{3.104}$$

where the components on the right in Eq. 3.104 can be found from Eq. 3.102. Finally, if $\mathbf{F}(\mathbf{X})$ is a deterministic $(M \times K)$-matrix function of an N-vector argument, then its expectation value—also a matrix—is found either by ensemble averaging its column vectors using Eq. 3.104, or, equivalently, by ensemble averaging each entry using Eq. 3.102.

Some of the benefits that accrue from the foregoing notational machinery can be gleaned from a simple example, which compares the component and vector/matrix approaches to calculating the first and second moments of a linear transformation. Let $x_1, x_2, \ldots, x_N$ be a collection of joint random variables with known first and second moments—means, variances, and covariances. Let us consider the first and second moments of the random variables

$$y_m \equiv \sum_{n=1}^{N} a_{mn} x_n + b_m, \qquad \text{for } 1 \leq m \leq M, \tag{3.105}$$

where the $\{a_{mn}\}$ and the $\{b_m\}$ are constants. Straightforward application of the basic properties cited earlier for expectation gives us

$$m_{y_m} = \sum_{n=1}^{N} a_{mn} m_{x_n} + b_m, \qquad \text{for } 1 \leq m \leq M, \tag{3.106}$$

and

$$\begin{aligned} \lambda_{y_m y_{m'}} &= E[\Delta y_m \Delta y_{m'}] \\ &= E\left[\sum_{n=1}^{N} a_{mn} \Delta x_n \sum_{n'=1}^{N} a_{m'n'} \Delta x_{n'}\right] \\ &= E\left[\sum_{n=1}^{N} \sum_{n'=1}^{N} a_{mn} a_{m'n'} \Delta x_n \Delta x_{n'}\right] \end{aligned}$$

$$= \sum_{n=1}^{N} \sum_{n'=1}^{N} a_{mn} a_{m'n'} \lambda_{x_n x_{n'}}, \qquad \text{for } 1 \leq m, m' \leq M. \quad (3.107)$$

The one novel calculational device encountered in the above development is the use of *different* dummy summing indices in the second equality. This device permits the product of two summations to be written as a double sum. The latter form is amenable to the interchange of expectation and summation in that the average of the sum is always the sum of the averages—even for multiple sums.

Throughout most of this book, the component-wise calculation just performed will be reprised in a variety of settings. Sometimes, however, it will be more efficient to employ the fully-equivalent and notationally more compact vector/matrix approach, which we will illustrate now. Let $\mathbf{x}$, $\mathbf{b}$, and $\mathbf{y}$ be vectors—of the appropriate dimensions—constructed as columns from the coefficients $\{x_n\}$, $\{b_m\}$, and $\{y_m\}$, respectively, and let $\mathbf{A}$ be the $M \times N$ matrix whose mnth element is a_{mn}. Equation 3.105 can then be rewritten as follows,

$$\mathbf{y} \equiv \mathbf{A}\mathbf{x} + \mathbf{b}, \quad (3.108)$$

which can easily be verified by component expansion, using the definition of matrix multiplication. Even for vectors and matrices, the linearity of expectation implies that: the average of a sum is the sum of the averages; the average of a constant times a random quantity is the constant times the average of the random quantity; and the average of a constant is that constant. Thus, we readily obtain that

$$\mathbf{m}_{\mathbf{y}} \equiv E[\mathbf{y}] = \mathbf{A}\mathbf{m}_{\mathbf{x}} + \mathbf{b} \quad (3.109)$$

is the mean vector of $\mathbf{y}$, which, in component form, is identical to Eq. 3.106.

Equation 3.109 shows, once again, that the *mean output* of a linear transformation is the transformation's response to the *mean input.* The *noise* in the output is thus the response to the *noise* in the input, viz.

$$\boldsymbol{\Delta}\mathbf{y} \equiv \mathbf{y} - \mathbf{m}_{\mathbf{y}} = \mathbf{A}\boldsymbol{\Delta}\mathbf{x}. \quad (3.110)$$

Multipliying Eq. 3.110 on its right by its *transpose* and taking expectations produces a single matrix equation which specifies *all* the $\{\lambda_{y_m y_{m'}}\}$, namely

$$\begin{aligned} \boldsymbol{\Lambda}_{\mathbf{y}} &\equiv E[\boldsymbol{\Delta}\mathbf{y}\boldsymbol{\Delta}\mathbf{y}^T] \\ &= E[\mathbf{A}\boldsymbol{\Delta}\mathbf{x}\boldsymbol{\Delta}\mathbf{x}^T\mathbf{A}^T] \\ &= \mathbf{A}E[\boldsymbol{\Delta}\mathbf{x}\boldsymbol{\Delta}\mathbf{x}^T]\mathbf{A}^T \\ &= \mathbf{A}\boldsymbol{\Lambda}_{\mathbf{x}}\mathbf{A}^T, \end{aligned} \quad (3.111)$$

where ^{T} denotes transpose. The $\mathbf{\Lambda}$ matrices are *covariance matrices.* In particular, because the transpose of a column vector is a row vector, the mm'th element of $\mathbf{\Lambda_y}$ is $\lambda_{y_m y_{m'}}$, the covariance between y_m and $y_{m'}$. Matrix multiplication then verifies that the mm'th element of Eq. 3.111 is Eq. 3.107.[21]

The last point to be made in our brief rundown of random-vector expectations concerns the joint characteristic function. For $\mathbf{x}$ a random N-vector with probability density $p_{\mathbf{x}}(\mathbf{X})$, its characteristic function is defined to be

$$\begin{aligned} M_{\mathbf{x}}(j\mathbf{v}) &= E[\exp(j\mathbf{v}^T\mathbf{x})] &\quad (3.112)\\ &= E\left[\exp\left(j\sum_{n=1}^{N} v_n x_n\right)\right], &\quad (3.113)\end{aligned}$$

an obvious extension of the 2-D case. We shall not exhibit the formulas, but it is worth noting that: knowledge of $M_{\mathbf{x}}$ is equivalent to knowledge of $p_{\mathbf{x}}$, because these functions comprise an N-D Fourier transform pair; and moments of products of the components of $\mathbf{x}$ can be found by differentiating $M_{\mathbf{x}}$.

Gaussian Random Vectors

We will close our whirlwind tour of probability theory by describing the extension of jointly Gaussian random variables to the N-D case, i.e., to Gaussian random vectors. Let $\mathbf{x}$ be a random N-vector. Then $\mathbf{x}$ is a Gaussian random vector if, for all deterministic N-vectors $\mathbf{a}$ and all deterministic scalars b, the random variable

$$z \equiv \mathbf{a}^T\mathbf{x} + b = \sum_{n=1}^{N} a_n x_n + b \qquad (3.114)$$

is a 1-D Gaussian. This linear-closure definition is in fact sufficient to show that

$$M_{\mathbf{x}}(j\mathbf{v}) = \exp\left(j\mathbf{v}^T\mathbf{m_x} - \frac{\mathbf{v}^T\mathbf{\Lambda_x}\mathbf{v}}{2}\right) \qquad (3.115)$$

is the characteristic function for a Gaussian random vector $\mathbf{x}$ with mean vector $\mathbf{m_x}$ and covariance matrix $\mathbf{\Lambda_x}$. The associated probability density function is

$$p_{\mathbf{x}}(\mathbf{X}) = \frac{\exp\left[-\frac{(\mathbf{X}-\mathbf{m_x})^T\mathbf{\Lambda_x}^{-1}(\mathbf{X}-\mathbf{m_x})}{2}\right]}{\sqrt{(2\pi)^N \det\mathbf{\Lambda_x}}}, \qquad (3.116)$$

[21] In conjunction with Eq. 3.111, we note that the expression $\mathbf{y}\mathbf{y}^T$, for $\mathbf{y}$ an M-vector, is an *outer* product—it is an $M \times M$ matrix. The more familiar $\mathbf{y}^T\mathbf{y}$ expression is a scalar quantity called the *inner* product.

where $\det\mathbf{\Lambda_x}$ is the determinant and $\mathbf{\Lambda_x}^{-1}$ is the inverse of the covariance matrix.

Equation 3.116 would be hopelessly more complicated were it written in component notation. It is the N-D form of the bell-shaped curve. Its equal-probability contours are N-D hyperellipsoids. Thankfully, we will seldom, if ever, have to confront Eq. 3.116 in its full generality. For now, it suffices to note that, as in the 2-D case, the *marginal* statistics of a Gaussian random vector $\mathbf{x}$ are all Gaussian, i.e., any subset of a collection of jointly Gaussian random variables $x_1, x_2, \ldots, x_N$ are also jointly Gaussian.[22] Thus, because the jointly Gaussian density of any dimensionality is completely determined by first and second moments, we *never* have to perform integrations to find the marginal statistics of a Gaussian random vector. As in the 2-D case, the converse result does not hold—there are $\{x_1, x_2, \ldots, x_N\}$ whose 1-D densities are all Gaussian but whose joint density is not Gaussian.

Two final properties and we shall be done. First, let $\mathbf{z}$ be an $(N+M)$-D Gaussian random vector partitioned into an N-D vector $\mathbf{x}$ and an M-D vector $\mathbf{y}$, as in Eq. 3.97; because $\mathbf{z}$ is a Gaussian random vector, we say that $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian random vectors. It follows that the conditional probability density for $\mathbf{x}$, given that $\mathbf{y} = \mathbf{Y}$ has occurred, is still Gaussian. Moreover, if every component of $\mathbf{x}$ is uncorrelated with every component of $\mathbf{y}$, i.e., if $\lambda_{x_n y_m} = 0$ for $1 \leq n \leq N$ and $1 \leq m \leq M$, then $\mathbf{x}$ and $\mathbf{y}$ are statistically independent random vectors.

Finally, for $\mathbf{x}$ a Gaussian random N-vector, the random M-vector $\mathbf{y}$ obtained from the linear transformation Eq. 3.108 is also Gaussian; its density is then fully specified by the simple first and second moment results Eqs. 3.109 and 3.111.

[22] This is easily proven by means of linear closure, e.g., x_1 and x_2 can be shown to be jointly Gaussian by using the linear-closure definition with $a_n = 0$ for $n \geq 3$.