4

Further Topics on Random Variables and Expectations

Contents

In this chapter, we develop a number of more advanced topics. We introduce methods that are useful in:

(a) dealing with the sum of independent random variables, including the case where the number of random variables is itself random;

(b) addressing problems of estimation or prediction of an unknown random variable on the basis of observed values of other random variables.

With these goals in mind, we introduce a number of tools, including transforms and convolutions, and refine our understanding of the concept of conditional expectation.

4.1 TRANSFORMS

In this section, we introduce the transform associated with a random variable. The transform provides us with an alternative representation of its probability law (PMF or PDF). It is not particularly intuitive, but it is often convenient for certain types of mathematical manipulations.

The **transform** of the distribution of a random variable X (also referred to as the **moment generating function** of X) is a function $M_X(s)$ of a free parameter s, defined by

$$M_X(s) = \mathbf{E}[e^{sX}].$$

The simpler notation $M(s)$ can also be used whenever the underlying random variable X is clear from the context. In more detail, when X is a discrete random variable, the corresponding transform is given by

$$M(s) = \sum_x e^{sx} p_X(x),$$

while in the continuous case, we have†

$$M(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x)\, dx.$$

Example 4.1. Let

$$p_X(x) = \begin{cases} 1/2, & \text{if } x = 2, \\ 1/6, & \text{if } x = 3, \\ 1/3, & \text{if } x = 5. \end{cases}$$

† The reader who is familiar with Laplace transforms may recognize that the transform associated with a continuous random variable is essentially the same as the Laplace transform of its PDF, the only difference being that Laplace transforms usually involve e^{-sx} rather than e^{sx}. For the discrete case, a variable z is sometimes used in place of e^s and the resulting transform $M(z) = \sum_x z^x p_X(x)$ is known as the *z-transform*. However, we will not be using z-transforms in this book.

Then, the corresponding transform is

$$M(s) = \frac{1}{2}e^{2s} + \frac{1}{6}e^{3s} + \frac{1}{3}e^{5s}$$

(see Fig. 4.1).

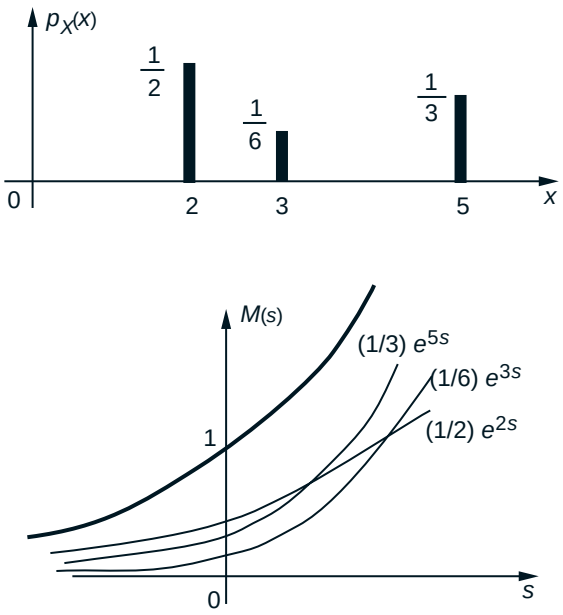

Figure 4.1: The PMF and the corresponding transform for Example 4.1. The transform $M(s)$ consists of the weighted sum of the three exponentials shown. Note that at $s = 0$, the transform takes the value 1. This is generically true since

$$M(0) = \sum_x e^{0\cdot x} p_X(x) = \sum_x p_X(x) = 1.$$

Example 4.2. The Transform of a Poisson Random Variable. Consider a Poisson random variable X with parameter λ:

$$p_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \qquad x = 0, 1, 2, \ldots$$

The corresponding transform is given by

$$M(s) = \sum_{x=0}^{\infty} e^{sx} \frac{\lambda^x e^{-\lambda}}{x!}.$$

We let $a = e^s\lambda$ and obtain

$$M(s) = e^{-\lambda}\sum_{x=0}^{\infty}\frac{a^x}{x!} = e^{-\lambda}e^a = e^{a-\lambda} = e^{\lambda(e^s-1)}.$$

Example 4.3. The Transform of an Exponential Random Variable. Let X be an exponential random variable with parameter λ:

$$f_X(x) = \lambda e^{-\lambda x}, \qquad x \geq 0.$$

Then,

$$\begin{aligned} M(s) =& \lambda\int_0^{\infty} e^{sx}e^{-\lambda x}\,dx \\ =& \lambda\int_0^{\infty} e^{(s-\lambda)x}\,dx \\ =& \lambda\frac{e^{(s-\lambda)x}}{s-\lambda}\bigg|_0^{\infty} \qquad \text{(if } s<\lambda\text{)} \\ =& \frac{\lambda}{\lambda-s}. \end{aligned}$$

The above calculation and the formula for $M(s)$ is correct only if the integrand $e^{(s-\lambda)x}$ decays as x increases, which is the case if and only if $s < \lambda$; otherwise, the integral is infinite.

It is important to realize that the transform is not a number but rather a *function* of a free variable or parameter s. Thus, we are dealing with a transformation that starts with a function, e.g., a PDF $f_X(x)$ (which is a function of a free variable x) and results in a new function, this time of a real parameter s. Strictly speaking, $M(s)$ is only defined for those values of s for which $\mathbf{E}[e^{sX}]$ is finite, as noted in the preceding example.

Example 4.4. The Transform of a Linear Function of a Random Variable. Let $M_X(s)$ be the transform associated with a random variable X. Consider a new random variable $Y = aX + b$. We then have

$$M_Y(s) = \mathbf{E}[e^{s(aX+b)}] = e^{sb}\mathbf{E}[e^{saX}] = e^{sb}M_X(sa).$$

For example, if X is exponential with parameter $\lambda = 1$, so that $M_X(s) = 1/(1-s)$, and if $Y = 2X + 3$, then

$$M_Y(s) = e^{3s}\frac{1}{1-2s}.$$

Example 4.5. The Transform of a Normal Random Variable. Let X be a normal random variable with mean μ and variance σ^2. To calculate the corresponding transform, we first consider the special case of the standard normal random variable Y, where $\mu = 0$ and $\sigma^2 = 1$, and then use the formula of the preceding example. The PDF of the standard normal is

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2},$$

and its transform is

$$\begin{aligned} M_Y(s) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, e^{sy}\, dy \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(y^2/2)+sy} dy \\ &= e^{s^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(y^2/2)+sy-(s^2/2)} dy \\ &= e^{s^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(y-s)^2/2} dy \\ &= e^{s^2/2}, \end{aligned}$$

where the last equality follows by using the normalization property of a normal PDF with mean s and unit variance.

A general normal random variable with mean μ and variance σ^2 is obtained from the standard normal via the linear transformation

$$X = \sigma Y + \mu.$$

The transform of the standard normal is $M_Y(s) = e^{s^2/2}$, as verified above. By applying the formula of Example 4.4, we obtain

$$M_X(s) = e^{s\mu} M_Y(s\sigma) = e^{\frac{\sigma^2 s^2}{2} + \mu s}.$$

From Transforms to Moments

The reason behind the alternative name "moment generating function" is that the moments of a random variable are easily computed once a formula for the associated transform is available. To see this, let us take the derivative of both sides of the definition

$$M(s) = \int_{-\infty}^{\infty} e^{sx} f_X(x)\, dx,$$

with respect to s. We obtain

$$\begin{aligned}\frac{d}{ds}M(s) &= \frac{d}{ds}\int_{-\infty}^{\infty} e^{sx} f_X(x)\,dx \\ &= \int_{-\infty}^{\infty} \frac{d}{ds} e^{sx} f_X(x)\,dx \\ &= \int_{-\infty}^{\infty} x e^{sx} f_X(x)\,dx.\end{aligned}$$

This equality holds for all values of s. By considering the special case where $s = 0$, we obtain†

$$\frac{d}{ds}M(s)\bigg|_{s=0} = \int_{-\infty}^{\infty} x f_X(x)\,dx = \mathbf{E}[X].$$

More generally, if we differentiate n times the function $M(s)$ with respect to s, a similar calculation yields

$$\frac{d^n}{ds^n}M(s)\bigg|_{s=0} = \int_{-\infty}^{\infty} x^n f_X(x)\,dx = \mathbf{E}[X^n].$$

Example 4.6. We saw earlier (Example 4.1) that the PMF

$$p_X(x) = \begin{cases} 1/2, & \text{if } x = 2, \\ 1/6, & \text{if } x = 3, \\ 1/3, & \text{if } x = 5, \end{cases}$$

has the transform

$$M(s) = \frac{1}{2}e^{2s} + \frac{1}{6}e^{3s} + \frac{1}{3}e^{5s}.$$

Thus,

$$\begin{aligned}\mathbf{E}[X] &= \frac{d}{ds}M(s)\bigg|_{s=0} \\ &= \frac{1}{2}2e^{2s} + \frac{1}{6}3e^{3s} + \frac{1}{3}5e^{5s}\bigg|_{s=0} \\ &= \frac{1}{2}\cdot 2 + \frac{1}{6}\cdot 3 + \frac{1}{3}\cdot 5 \\ &= \frac{19}{6}.\end{aligned}$$

† This derivation involves an interchange of differentiation and integration. The interchange turns out to be justified for all of the applications to be considered in this book. Furthermore, the derivation remains valid for general random variables, including discrete ones. In fact, it could be carried out more abstractly, in the form

$$\frac{d}{ds}M(s) = \frac{d}{ds}\mathbf{E}[e^{sX}] = \mathbf{E}\left[\frac{d}{ds}e^{sX}\right] = \mathbf{E}[Xe^{sX}],$$

leading to the same conclusion.

Also,

$$\begin{aligned}\mathbf{E}[X^2] &= \frac{d^2}{ds^2}M(s)\bigg|_{s=0} \\ &= \frac{1}{2}4e^{2s} + \frac{1}{6}9e^{3s} + \frac{1}{3}25e^{5s}\bigg|_{s=0} \\ &= \frac{1}{2}\cdot 4 + \frac{1}{6}\cdot 9 + \frac{1}{3}\cdot 25 \\ &= \frac{71}{6}.\end{aligned}$$

For an exponential random variable with PDF

$$f_X(x) = \lambda e^{-\lambda x}, \qquad x \geq 0,$$

we found earlier that

$$M(s) = \frac{\lambda}{\lambda - s}.$$

Thus,

$$\frac{d}{ds}M(s) = \frac{\lambda}{(\lambda - s)^2}, \qquad \frac{d^2}{ds^2}M(s) = \frac{2\lambda}{(\lambda - s)^3}.$$

By setting $s = 0$, we obtain

$$\mathbf{E}[X] = \frac{1}{\lambda}, \qquad \mathbf{E}[X^2] = \frac{2}{\lambda^2},$$

which agrees with the formulas derived in Chapter 3.

Inversion of Transforms

A very important property of transforms is the following.

Inversion Property

The transform $M_X(s)$ completely determines the probability law of the random variable X. In particular, if $M_X(s) = M_Y(s)$ for all s, then the random variables X and Y have the same probability law.

This property is a rather deep mathematical fact that we will use frequently.† There exist explicit formulas that allow us to recover the PMF or PDF of a random variable starting from the associated transform, but they are quite difficult to use. In practice, transforms are usually inverted by "pattern matching," based on tables of known distribution-transform pairs. We will see a number of such examples shortly.

† In fact, the probability law of a random variable is completely determined even if we only know the transform $M(s)$ for values of s in some interval of positive length.

Example 4.7. We are told that the transform associated with a random variable X is

$$M(s) = \frac{1}{4}e^{-s} + \frac{1}{2} + \frac{1}{8}e^{4s} + \frac{1}{8}e^{5s}.$$

Since $M(s)$ is a sum of terms of the form e^{sx}, we can compare with the general formula

$$M(s) = \sum_x e^{sx} p_X(x),$$

and infer that X is a discrete random variable. The different values that X can take can be read from the corresponding exponents and are -1, 0, 4, and 5. The probability of each value x is given by the coefficient multiplying the corresponding e^{sx} term. In our case, $\mathbf{P}(X = -1) = 1/4$, $\mathbf{P}(X = 0) = 1/2$, $\mathbf{P}(X = 4) = 1/8$, $\mathbf{P}(X = 5) = 1/8$.

Generalizing from the last example, the distribution of a finite-valued discrete random variable can be always found by inspection of the corresponding transform. The same procedure also works for discrete random variables with an infinite range, as in the example that follows.

Example 4.8. The Transform of a Geometric Random Variable. We are told that the transform associated with random variable X is of the form

$$M(s) = \frac{pe^s}{1 - (1-p)e^s},$$

where p is a constant in the range $0 < p < 1$. We wish to find the distribution of X. We recall the formula for the geometric series:

$$\frac{1}{1-\alpha} = 1 + \alpha + \alpha^2 + \cdots,$$

which is valid whenever $|\alpha| < 1$. We use this formula with $\alpha = (1-p)e^s$, and for s sufficiently close to zero so that $(1-p)e^s < 1$. We obtain

$$M(s) = pe^s\Big(1 + (1-p)e^s + (1-p)^2e^{2s} + (1-p)^3e^{3s} + \cdots\Big).$$

As in the previous example, we infer that this is a discrete random variable that takes positive integer values. The probability $\mathbf{P}(X = k)$ is found by reading the coefficient of the term e^{ks}. In particular, $\mathbf{P}(X = 1) = p$, $\mathbf{P}(X = 2) = p(1-p)$, etc., and

$$\mathbf{P}(X = k) = p(1-p)^{k-1}, \qquad k = 1, 2, \ldots$$

We recognize this as the geometric distribution with parameter p.

Note that

$$\frac{d}{ds}M(s) = \frac{pe^s}{1-(1-p)e^s} + \frac{(1-p)pe^s}{(1-(1-p)e^s)^2}.$$

If we set $s = 0$, the above expression evaluates to $1/p$, which agrees with the formula for $\mathbf{E}[X]$ derived in Chapter 2.

Example 4.9. The Transform of a Mixture of Two Distributions. The neighborhood bank has three tellers, two of them fast, one slow. The time to assist a customer is exponentially distributed with parameter $\lambda = 6$ at the fast tellers, and $\lambda = 4$ at the slow teller. Jane enters the bank and chooses a teller at random, each one with probability 1/3. Find the PDF of the time it takes to assist Jane and its transform.

We have

$$f_X(x) = \frac{2}{3} \cdot 6e^{-6x} + \frac{1}{3} \cdot 4e^{-4x}, \qquad x \geq 0.$$

Then,

$$\begin{aligned} M(s) &= \int_0^\infty e^{sx} \left(\frac{2}{3} 6e^{-6x} + \frac{1}{3} 4e^{-4x} \right) dx \\ &= \frac{2}{3} \int_0^\infty e^{sx} 6e^{-6x}\, dx + \frac{1}{3} \int_0^\infty e^{sx} 4e^{-4x}\, dx \\ &= \frac{2}{3} \cdot \frac{6}{6-s} + \frac{1}{3} \cdot \frac{4}{4-s} \qquad \text{(for } s < 4\text{)}. \end{aligned}$$

More generally, let $X_1, \ldots, X_n$ be continuous random variables with PDFs $f_{X_1}, \ldots f_{X_n}$, and let Y be a random variable, which is equal to X_i with probability p_i. Then,

$$f_Y(y) = p_1 f_{X_1}(y) + \cdots + p_n f_{X_n}(y),$$

and

$$M_Y(s) = p_1 M_{X_1}(s) + \cdots + p_n M_{X_n}(s).$$

The steps in this problem can be reversed. For example, we may be told that the transform associated with a random variable Y is of the form

$$\frac{1}{2} \cdot \frac{1}{2-s} + \frac{3}{4} \cdot \frac{1}{1-s}.$$

We can then rewrite it as

$$\frac{1}{4} \cdot \frac{2}{2-s} + \frac{3}{4} \cdot \frac{1}{1-s},$$

and recognize that Y is the mixture of two exponential random variables with parameters 2 and 1, which are selected with probabilities 1/4 and 3/4, respectively.

Sums of Independent Random Variables

Transform methods are particularly convenient when dealing with a sum of random variables. This is because it turns out that *addition of independent random variables corresponds to multiplication of transforms*, as we now show.

Let X and Y be independent random variables, and let $W = X + Y$. The transform associated with W is, by definition,

$$M_W(s) = \mathbf{E}[e^{sW}] = \mathbf{E}[e^{s(X+Y)}] = \mathbf{E}[e^{sX}e^{sY}].$$

Consider a fixed value of the parameter s. Since X and Y are independent, e^{sX} and e^{sY} are independent random variables. Hence, the expectation of their product is the product of the expectations, and

$$M_W(s) = \mathbf{E}[e^{sX}]\mathbf{E}[e^{sY}] = M_X(s)M_Y(s).$$

By the same argument, if $X_1, \ldots, X_n$ is a collection of independent random variables, and

$$W = X_1 + \cdots + X_n,$$

then

$$M_W(s) = M_{X_1}(s) \cdots M_{X_n}(s).$$

Example 4.10. The Transform of the Binomial. Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with a common parameter p. Then,

$$M_{X_i}(s) = (1-p)e^{0s} + pe^{1s} = 1 - p + pe^s, \qquad \text{for all } i.$$

The random variable $Y = X_1 + \cdots + X_n$ is binomial with parameters n and p. Its transform is given by

$$M_Y(s) = \left(1 - p + pe^s\right)^n.$$

Example 4.11. The Sum of Independent Poisson Random Variables is Poisson. Let X and Y be independent Poisson random variables with means λ and μ, respectively, and let $W = X + Y$. Then,

$$M_X(s) = e^{\lambda(e^s-1)}, \qquad M_Y(s) = e^{\mu(e^s-1)},$$

and

$$M_W(s) = M_X(s)M_Y(s) = e^{\lambda(e^s-1)}e^{\mu(e^s-1)} = e^{(\lambda+\mu)(e^s-1)}.$$

Thus, W has the same transform as a Poisson random variable with mean $\lambda + \mu$. By the uniqueness property of transforms, W is Poisson with mean $\lambda + \mu$.

Example 4.12. The Sum of Independent Normal Random Variables is Normal. Let X and Y be independent normal random variables with means μ_x, μ_y, and variances σ_x^2, σ_y^2, respectively. Let $W = X + Y$. Then,

$$M_X(s) = e^{\frac{\sigma_x^2 s^2}{2} + \mu_x s}, \qquad M_Y(s) = e^{\frac{\sigma_y^2 s^2}{2} + \mu_y s},$$

and

$$M_W(s) = e^{\frac{(\sigma_x^2 + \sigma_y^2) s^2}{2} + (\mu_x + \mu_y) s}.$$

Thus, W has the same transform as a normal random variable with mean $\mu_x + \mu_y$ and variance $\sigma_x^2 + \sigma_y^2$. By the uniqueness property of transforms, W is normal with these parameters.

Summary of Transforms and their Properties

- The transform associated with the distribution of a random variable X is given by

$$M_X(s) = \mathbf{E}[e^{sX}] = \begin{cases} \displaystyle\sum_x e^{sx} p_X(x), & x \text{ discrete}, \\ \displaystyle\int_{-\infty}^{\infty} e^{sx} f_X(x)\, dx, & x \text{ continuous}. \end{cases}$$

- The distribution of a random variable is completely determined by the corresponding transform.
- Moment generating properties:

$$M_X(0) = 1, \qquad \frac{d}{ds} M_X(s)\bigg|_{s=0} = \mathbf{E}[X], \qquad \frac{d^n}{ds^n} M_X(s)\bigg|_{s=0} = \mathbf{E}[X^n].$$

- If $Y = aX + b$, then $M_Y(s) = e^{sb} M_X(as)$.
- If X and Y are independent, then $M_{X+Y}(s) = M_X(s) M_Y(s)$.

We have derived formulas for the transforms of a few common random variables. Such formulas can be derived with a moderate amount of algebra for many other distributions. Some of the most useful ones are summarized in the tables that follow.

Transforms of Joint Distributions

If two random variables X and Y are described by some joint distribution (e.g., a joint PDF), then each one is associated with a transform $M_X(s)$ or $M_Y(s)$. These

Transforms for Common Discrete Random Variables

Bernoulli(p)

$$p_X(k) = \begin{cases} p, & \text{if } k = 1, \\ 1-p, & \text{if } k = 0. \end{cases} \qquad M_X(s) = 1 - p + pe^s.$$

Binomial(n, p)

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

$$M_X(s) = (1 - p + pe^s)^n.$$

Geometric(p)

$$p_X(k) = p(1-p)^{k-1}, \quad k = 1, 2, \ldots \qquad M_X(s) = \frac{pe^s}{1-(1-p)e^s}.$$

Poisson(λ)

$$p_X(k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, \ldots \qquad M_X(s) = e^{\lambda(e^s - 1)}.$$

Uniform(a, b)

$$p_X(k) = \frac{1}{b-a+1}, \quad k = a, a+1, \ldots, b.$$

$$M_X(s) = \frac{e^{as}}{b-a+1} \frac{e^{(b-a+1)s} - 1}{e^s - 1}.$$

are the transforms of the marginal distributions and do not convey information on the dependence between the two random variables. Such information is contained in a multivariate transform, which we now define.

Consider n random variables $X_1, \ldots, X_n$ related to the same experiment. Let $s_1, \ldots, s_n$ be scalar free parameters. The associated multivariate transform is a function of these n parameters and is defined by

$$M_{X_1,\ldots,X_n}(s_1, \ldots, s_n) = \mathbf{E}\big[e^{sX_1 + \cdots + s_n X_n}\big].$$

The inversion property of transforms discussed earlier extends to the multivariate case. That is, if $Y_1, \ldots, Y_n$ is another set of random variables and $M_{X_1,\ldots,X_n}(s_1, \ldots, s_n)$, $M_{Y_1,\ldots,Y_n}(s_1, \ldots, s_n)$ are the same functions of $s_1, \ldots, s_n$,

Transforms for Common Continuous Random Variables

Uniform(a, b)

$$f_X(x) = \frac{1}{b-a}, \quad a \le x \le b. \qquad M_X(s) = \frac{1}{b-a}\,\frac{e^{sb} - e^{sa}}{s}.$$

Exponential(λ)

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \ge 0. \qquad M_X(s) = \frac{\lambda}{\lambda - s}, \qquad (s > \lambda).$$

Normal(μ, σ^2)

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty. \qquad M_X(s) = e^{\frac{\sigma^2 s^2}{2} + \mu s}.$$

then the joint distribution of $X_1, \ldots, X_n$ is the same as the joint distribution of $Y_1, \ldots, Y_n$.

4.2 SUMS OF INDEPENDENT RANDOM VARIABLES — CONVOLUTIONS

If X and Y are independent random variables, the distribution of their sum $W = X + Y$ can be obtained by computing and then inverting the transform $M_W(s) = M_X(s)M_Y(s)$. But it can also be obtained directly, using the method developed in this section.

The Discrete Case

Let $W = X+Y$, where X and Y are independent integer-valued random variables with PMFs $p_X(x)$ and $p_Y(y)$. Then, for any integer w,

$$\begin{aligned} p_W(w) =& \mathbf{P}(X + Y = w) \\ =& \sum_{(x,y):\ x+y=w} \mathbf{P}(X = x \text{ and } Y = y) \\ =& \sum_x \mathbf{P}(X = x \text{ and } Y = w - x) \\ =& \sum_x p_X(x) p_Y(w - x). \end{aligned}$$

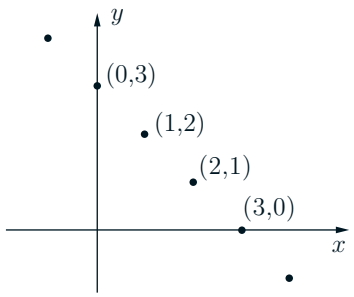

Figure 4.2: The probability $p_W(3)$ that $X+Y=3$ is the sum of the probabilities of all pairs (x,y) such that $x+y=3$, which are the points indicated in the figure. The probability of a generic such point is of the form $p_{X,Y}(x,3-x) = p_X(x)p_Y(3-x)$.

The resulting PMF $p_W(w)$ is called the **convolution** of the PMFs of X and Y. See Fig. 4.2 for an illustration.

Example 4.13. Let X and Y be independent and have PMFs given by

$$p_X(x) = \begin{cases} \frac{1}{3} & \text{if } x = 1,2,3, \\ 0 & \text{otherwise,} \end{cases} \qquad p_Y(y) = \begin{cases} \frac{1}{2} & \text{if } x = 0, \\ \frac{1}{3} & \text{if } x = 1, \\ \frac{1}{6} & \text{if } x = 2, \\ 0 & \text{otherwise.} \end{cases}$$

To calculate the PMF of $W = X + Y$ by convolution, we first note that the range of possible values of w are the integers from the range $[1,5]$. Thus we have

$$p_W(w) = 0 \quad \text{if } w \neq 1,2,3,4,5.$$

We calculate $p_W(w)$ for each of the values $w = 1,2,3,4,5$ using the convolution formula. We have

$$p_W(1) = \sum_x p_X(x)p_Y(1-x) = p_X(1) \cdot p_Y(0) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6},$$

where the second equality above is based on the fact that for $x \neq 1$ either $p_X(x)$ or $p_Y(1-x)$ (or both) is zero. Similarly, we obtain

$$p_W(2) = p_X(1) \cdot p_Y(1) + p_X(2) \cdot p_Y(0) = \frac{1}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{2} = \frac{5}{18},$$

$$p_W(3) = p_X(1) \cdot p_Y(2) + p_X(2) \cdot p_Y(1) + p_X(3) \cdot p_Y(0) = \frac{1}{3} \cdot \frac{1}{6} + \frac{1}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{3},$$

$$p_W(4) = p_X(2) \cdot p_Y(2) + p_X(3) \cdot p_Y(1) = \frac{1}{3} \cdot \frac{1}{6} + \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{6},$$

$$p_W(5) = p_X(3) \cdot p_Y(2) = \frac{1}{3} \cdot \frac{1}{6} = \frac{1}{18}.$$

The Continuous Case

Let X and Y be independent continuous random variables with PDFs $f_X(x)$ and $f_Y(y)$. We wish to find the PDF of $W = X + Y$. Since W is a function of two random variables X and Y, we can follow the method of Chapter 3, and start by deriving the CDF $F_W(w)$ of W. We have

$$\begin{aligned} F_W(w) &= \mathbf{P}(W \le w) \\ &= \mathbf{P}(X + Y \le w) \\ &= \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{w-x} f_X(x) f_Y(y)\, dy\, dx \\ &= \int_{x=-\infty}^{\infty} f_X(x) \left[\int_{y=-\infty}^{w-x} f_Y(y)\, dy \right] dx \\ &= \int_{x=-\infty}^{\infty} f_X(x) F_Y(w-x)\, dx. \end{aligned}$$

The PDF of W is then obtained by differentiating the CDF:

$$\begin{aligned} f_W(w) &= \frac{dF_W}{dw}(w) \\ &= \frac{d}{dw} \int_{x=-\infty}^{\infty} f_X(x) F_Y(w-x)\, dx \\ &= \int_{x=-\infty}^{\infty} f_X(x) \frac{dF_Y}{dw}(w-x)\, dx \\ &= \int_{x=-\infty}^{\infty} f_X(x) f_Y(w-x)\, dx. \end{aligned}$$

This formula is entirely analogous to the formula for the discrete case, except that the summation is replaced by an integral and the PMFs are replaced by PDFs. For an intuitive understanding of this formula, see Fig. 4.3.

Example 4.14. The random variables X and Y are independent and uniformly distributed in the interval $[0, 1]$. The PDF of $W = X + Y$ is

$$f_W(w) = \int_{-\infty}^{\infty} f_X(x) f_Y(w-x)\, dx.$$

The integrand $f_X(x) f_Y(w-x)$ is nonzero (and equal to 1) for $0 \le x \le 1$ and $0 \le w - x \le 1$. Combining these two inequalities, the integrand is nonzero for $\max\{0, w-1\} \le x \le \min\{1, w\}$. Thus,

$$f_W(w) = \begin{cases} \min\{1, w\} - \max\{0, w-1\}, & 0 \le w \le 2, \\ 0, & \text{otherwise}, \end{cases}$$

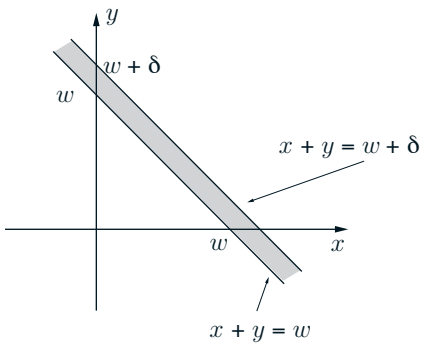

Figure 4.3: Illustration of the convolution formula for the case of continuous random variables (compare with Fig. 4.2). For small δ, the probability of the strip indicated in the figure is $\mathbf{P}(w \le X + Y \le w + \delta) \approx f_W(w) \cdot \delta$. Thus,

$$\begin{aligned} f_W(w) \cdot \delta =& \mathbf{P}(w \le X + Y \le w + \delta) \\ &= \int_{x=-\infty}^{\infty} \int_{y=w-x}^{w-x+\delta} f_X(x) f_Y(y)\, dy\, dx \\ &\approx \int_{x=-\infty}^{\infty} f_X(x) f_Y(w-x) \delta\, dx. \end{aligned}$$

The desired formula follows by canceling δ from both sides.

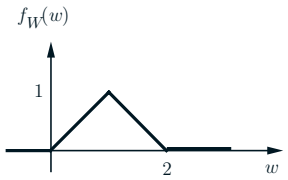

Figure 4.4: The PDF of the sum of two independent uniform random variables in $[0, 1]$.

which has the triangular shape shown in Fig. 4.4.

The calculation in the last example was based on a literal application of the convolution formula. The most delicate step was to determine the correct limits for the integration. This is often tedious and error prone, but can be bypassed using a graphical method described next.

Graphical Calculation of Convolutions

We will use a dummy variable t as the argument of the different functions involved in this discussion; see also Fig. 4.5. Consider a PDF $f_X(t)$ which is zero outside the range $a \le t \le b$ and a PDF $f_Y(t)$ which is zero outside the range $c \le t \le d$. Let us fix a value w, and plot $f_Y(w-t)$ as a function of t. This plot has the same shape as the plot of $f_Y(t)$ except that it is first "flipped" and then shifted by an amount w. (If $w > 0$, this is a shift to the right, if $w < 0$, this is a shift to the left.) We then place the plots of $f_X(t)$ and $f_Y(w-t)$ on top of each other. The value of $f_W(w)$ is equal to the integral of the product of these two plots. By varying the amount w by which we are shifting, we obtain $f_W(w)$ for any w.

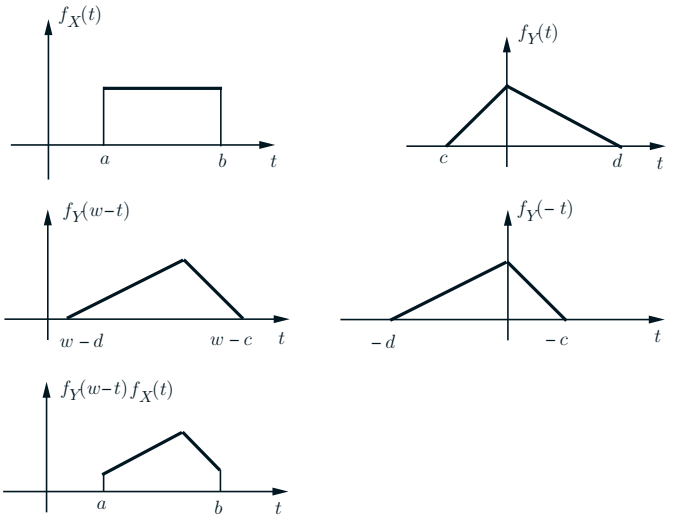

Figure 4.5: Illustration of the convolution calculation. For the value of w under consideration, $f_W(w)$ is equal to the integral of the function shown in the last plot.

4.3 CONDITIONAL EXPECTATION AS A RANDOM VARIABLE

The value of the conditional expectation $\mathbf{E}[X \,|\, Y = y]$ of a random variable X given another random variable Y depends on the realized experimental value y of Y. This makes $\mathbf{E}[X \,|\, Y]$ a function of Y, and therefore a random variable. In this section, we study the expectation and variance of $\mathbf{E}[X \,|\, Y]$. In the process,

we obtain some useful formulas (the **law of iterated expectations** and the **law of conditional variances**) that are often convenient for the calculation of expected values and variances.

Recall that the conditional expectation $\mathbf{E}[X \mid Y = y]$ is defined by

$$\mathbf{E}[X \mid Y = y] = \sum_x x p_{X|Y}(x \mid y), \qquad \text{(discrete case)},$$

and

$$\mathbf{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\, dx, \qquad \text{(continuous case)}.$$

Once a value of y is given, the above summation or integration yields a numerical value for $\mathbf{E}[X \mid Y = y]$.

Example 4.15. Let the random variables X and Y have a joint PDF which is equal to 2 for (x, y) belonging to the triangle indicated in Fig. 4.6(a), and zero everywhere else. In order to compute $\mathbf{E}[X \mid Y = y]$, we first need to obtain the conditional density of X given $Y = y$.

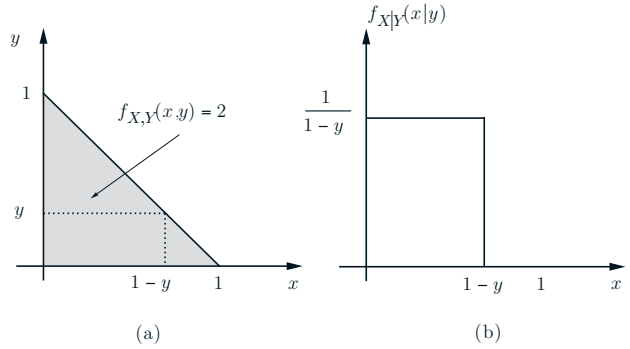

Figure 4.6: (a) The joint PDF in Example 4.15. (b) The conditional density of X.

We have

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx = \int_0^{1-y} 2\, dx = 2(1 - y), \qquad 0 \le y \le 1,$$

and

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{1}{1 - y}, \qquad 0 \le x \le 1 - y.$$

The conditional density is shown in Fig. 4.6(b).

Intuitively, since the joint PDF is constant, the conditional PDF (which is a "slice" of the joint, at some fixed y) is also a constant. Therefore, the conditional PDF must be a uniform distribution. Given that $Y = y$, X ranges from 0 to $1 - y$. Therefore, for the PDF to integrate to 1, its height must be equal to $1/(1 - y)$, in agreement with Fig. 4.6(b).

For $y > 1$ or $y < 0$, the conditional PDF is undefined, since these values of y are impossible. For $y = 1$, X must be equal to 0, with certainty, and $\mathbf{E}[X \,|\, Y = 1] = 0$.

For $0 \leq y < 1$, the conditional mean $\mathbf{E}[X \,|\, Y = y]$ is the expectation of the uniform PDF in Fig. 4.6(b), and we have

$$\mathbf{E}[X \,|\, Y = y] = \frac{1 - y}{2}, \qquad 0 \leq y < 1.$$

Since $\mathbf{E}[X \,|\, Y = 1] = 0$, the above formula is also valid when $y = 1$. The conditional expectation is undefined when y is outside $[0, 1]$.

For any number y, $\mathbf{E}[X \,|\, Y = y]$ is also a number. As y varies, so does $\mathbf{E}[X \,|\, Y = y]$, and we can therefore view $\mathbf{E}[X \,|\, Y = y]$ as a function of y. Since y is the experimental value of the random variable Y, we are dealing with a function of a random variable, hence a new random variable. More precisely, we **define** $\mathbf{E}[X \,|\, Y]$ to be the random variable whose value is $\mathbf{E}[X \,|\, Y = y]$ when the outcome of Y is y.

Example 4.15. (continued) We saw that $\mathbf{E}[X \,|\, Y = y] = (1 - y)/2$. Hence, $\mathbf{E}[X \,|\, Y]$ is the random variable $(1 - Y)/2$:

$$\mathbf{E}[X \,|\, Y] = \frac{1 - Y}{2}.$$

Since $\mathbf{E}[X \,|\, Y]$ is a random variable, it has an expectation $\mathbf{E}\big[\mathbf{E}[X \,|\, Y]\big]$ of its own. Applying the expected value rule, this is given by

$$\mathbf{E}\big[\mathbf{E}[X \,|\, Y]\big] = \begin{cases} \displaystyle\sum_y \mathbf{E}[X \,|\, Y = y] p_Y(y), & Y \text{ discrete}, \\ \displaystyle\int_{-\infty}^{\infty} \mathbf{E}[X \,|\, Y = y] f_Y(y)\, dy, & Y \text{ continuous}. \end{cases}$$

Both expressions in the right-hand side should be familiar from Chapters 2 and 3, respectively. By the corresponding versions of the total expectation theorem, they are equal to $\mathbf{E}[X]$. This brings us to the following conclusion, which is actually valid for every type of random variable Y (discrete, continuous, mixed, etc.), as long as X has a well-defined and finite expectation $\mathbf{E}[X]$.

Law of iterated expectations: $\mathbf{E}\big[\mathbf{E}[X \,|\, Y]\big] = \mathbf{E}[X]$.

Example 4.15 (continued) In Example 4.15, we found $\mathbf{E}[X \,|\, Y] = (1 - Y)/2$ [see Fig. 4.6(b)]. Taking expectations of both sides, and using the law of iterated expectations to evaluate the left-hand side, we obtain $\mathbf{E}[X] = \big(1-\mathbf{E}[Y]\big)/2$. Because of symmetry, we must have $\mathbf{E}[X] = \mathbf{E}[Y]$. Therefore, $\mathbf{E}[X] = \big(1 - \mathbf{E}[X]\big)/2$, which yields $\mathbf{E}[X] = 1/3$. In a slightly different version of this example, where there is no symmetry between X and Y, we would use a similar argument to express $\mathbf{E}[Y]$.

Example 4.16. We start with a stick of length ℓ. We break it at a point which is chosen randomly and uniformly over its length, and keep the piece that contains the left end of the stick. We then repeat the same process on the stick that we were left with. What is the expected length of the stick that we are left with, after breaking twice?

Let Y be the length of the stick after we break for the first time. Let X be the length after the second time. We have $\mathbf{E}[X \,|\, Y] = Y/2$, since the breakpoint is chosen uniformly over the length Y of the remaining stick. For a similar reason, we also have $\mathbf{E}[Y] = \ell/2$. Thus,

$$\mathbf{E}[X] = \mathbf{E}\big[\mathbf{E}[X \,|\, Y]\big] = \mathbf{E}\left[\frac{Y}{2}\right] = \frac{\mathbf{E}[Y]}{2} = \frac{\ell}{4}.$$

Example 4.17. Averaging Quiz Scores by Section. A class has n students and the quiz score of student i is x_i. The average quiz score is

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The class consists of S sections, with n_s students in section s. The average score in section s is

$$m_s = \frac{1}{n_s}\sum_{\text{stdnts. } i \text{ in sec. } s} x_i.$$

The average score over the whole class can be computed by taking the average score m_s of each section, and then forming a *weighted average*; the weight given to section s is proportional to the number of students in that section, and is n_s/n. We verify that this gives the correct result:

$$\begin{aligned}
\sum_{s=1}^{S} \frac{n_s}{n} m_s &= \sum_{s=1}^{S} \frac{n_s}{n} \cdot \frac{1}{n_s} \sum_{\text{stdnts. } i \text{ in sec. } s} x_i \\
&= \frac{1}{n} \sum_{s=1}^{S} \sum_{\text{stdnts. } i \text{ in sec. } s} x_i \\
&= \frac{1}{n} \sum_{i=1}^{n} x_i. \\
&= m.
\end{aligned}$$

How is this related to conditional expectations? Consider an experiment in which a student is selected at random, each student having probability $1/n$ of being selected. Consider the following two random variables:

$$X = \text{quiz score of a student,}$$
$$Y = \text{section of a student, } (Y \in \{1, \ldots, S\}).$$

We then have

$$\mathbf{E}[X] = m.$$

Conditioning on $Y = s$ is the same as assuming that the selected student is in section s. Conditional on that event, every student in that section has the same probability $1/n_s$ of being chosen. Therefore,

$$\mathbf{E}[X \mid Y = s] = \frac{1}{n_s} \sum_{\text{stdnts. } i \text{ in sec. } s} x_i = m_s.$$

A randomly selected student belongs to section s with probability n_s/n, i.e., $\mathbf{P}(Y = s) = n_s/n$. Hence,

$$\mathbf{E}\big[\mathbf{E}[X \mid Y]\big] = \sum_{s=1}^{S} \mathbf{E}[X \mid Y = s]\mathbf{P}(Y = s) = \sum_{s=1}^{S} \frac{n_s}{n} m_s.$$

As shown earlier, this is the same as m. Thus, averaging by section can be viewed as a special case of the law of iterated expectations.

Example 4.18. Forecast Revisions. Let Y be the sales of a company in the first semester of the coming year, and let X be the sales over the entire year. The company has constructed a statistical model of sales, and so the joint distribution of X and Y is assumed to be known. In the beginning of the year, the expected value $\mathbf{E}[X]$ serves as a forecast of the actual sales X. In the middle of the year, the first semester sales have been realized and the experimental value of the random value Y is now known. This places us in a new "universe," where everything is conditioned on the realized value of Y. We then consider the mid-year revised forecast of yearly sales, which is $\mathbf{E}[X \mid Y]$.

We view $\mathbf{E}[X \mid Y] - \mathbf{E}[X]$ as the forecast revision, in light of the mid-year information. The law of iterated expectations implies that

$$\mathbf{E}\big[\mathbf{E}[X \mid Y] - \mathbf{E}[X]\big] = 0.$$

This means that, in the beginning of the year, we do not expect our forecast to be revised in any specific direction. Of course, the actual revision will usually be positive or negative, but the probabilities are such that it is zero on the average. This is quite intuitive. For example, if a positive revision was expected, the original forecast should have been higher in the first place.

The Conditional Variance

The conditional distribution of X given $Y = y$ has a mean, which is $\mathbf{E}[X \mid Y = y]$, and by the same token, it also has a variance. This is defined by the same formula as the unconditional variance, except that everything is conditioned on $Y = y$:

$$\operatorname{var}(X \mid Y = y) = \mathbf{E}\Big[\big(X - \mathbf{E}[X \mid Y = y]\big)^2 \mid Y = y\Big].$$

Note that the conditional variance is a function of the experimental value y of the random variable Y. Hence, it is a function of a random variable, and is itself a random variable that will be denoted by $\operatorname{var}(X \mid Y)$.

Arguing by analogy to the law of iterated expectations, we may conjecture that the expectation of the conditional variance $\operatorname{var}(X \mid Y)$ is related to the unconditional variance $\operatorname{var}(X)$. This is indeed the case, but the relation is more complex.

Law of Conditional Variances:

$$\operatorname{var}(X) = \mathbf{E}\big[\operatorname{var}(X \mid Y)\big] + \operatorname{var}\big(\mathbf{E}[X \mid Y]\big)$$

To verify the law of conditional variances, we start with the identity

$$X - \mathbf{E}[X] = \big(X - \mathbf{E}[X \mid Y]\big) + \big(\mathbf{E}[X \mid Y] - \mathbf{E}[X]\big).$$

We square both sides and then take expectations to obtain

$$\begin{aligned}\operatorname{var}(X) &= \mathbf{E}\Big[\big(X - \mathbf{E}[X]\big)^2\Big] \\ &= \mathbf{E}\Big[\big(X - \mathbf{E}[X \mid Y]\big)^2\Big] + \mathbf{E}\Big[\big(\mathbf{E}[X \mid Y] - \mathbf{E}[X]\big)^2\Big] \\ &\quad + 2\mathbf{E}\Big[\big(X - \mathbf{E}[X \mid Y]\big)\big(\mathbf{E}[X \mid Y] - \mathbf{E}[X]\big)\Big].\end{aligned}$$

Using the law of iterated expectations, the first term in the right-hand side of the above equation can be written as

$$\mathbf{E}\Big[\mathbf{E}\Big[\big(X - \mathbf{E}[X \mid Y]\big)^2 \mid Y\Big]\Big],$$

which is the same as $\mathbf{E}\big[\operatorname{var}(X \mid Y)\big]$. The second term is equal to $\operatorname{var}\big(\mathbf{E}[X \mid Y]\big)$, since $\mathbf{E}[X]$ is the mean of $\mathbf{E}[X \mid Y]$. Finally, the third term is zero, as we now show. Indeed, if we define $h(Y) = 2\big(\mathbf{E}[X \mid Y] - \mathbf{E}[X]\big)$, the third term is

$$\begin{aligned}\mathbf{E}\Big[\big(X - \mathbf{E}[X \mid Y]\big)h(Y)\Big] &= \mathbf{E}\big[Xh(Y)\big] - \mathbf{E}\big[\mathbf{E}[X \mid Y]h(Y)\big] \\ &= \mathbf{E}\big[Xh(Y)\big] - \mathbf{E}\Big[\mathbf{E}\big[Xh(Y) \mid Y\big]\Big] \\ &= \mathbf{E}\big[Xh(Y)\big] - \mathbf{E}\big[Xh(Y)\big] \\ &= 0.\end{aligned}$$

Example 4.16. (continued) Consider again the problem where we break twice a stick of length ℓ, at randomly chosen points, with Y being the length of the stick after the first break and X being the length after the second break. We calculated the mean of X as $\ell/4$, and now let us use the law of conditional variances to calculate $\mathrm{var}(X)$. We have $\mathbf{E}[X \,|\, Y] = Y/2$, so since Y is uniformly distributed between 0 and ℓ,

$$\mathrm{var}\big(\mathbf{E}[X \,|\, Y]\big) = \mathrm{var}(Y/2) = \frac{1}{4}\mathrm{var}(Y) = \frac{1}{4} \cdot \frac{\ell^2}{12} = \frac{\ell^2}{48}.$$

Also, since X is uniformly distributed between 0 and Y, we have

$$\mathrm{var}(X \,|\, Y) = \frac{Y^2}{12}.$$

Thus, since Y is uniformly distributed between 0 and ℓ,

$$\mathbf{E}\big[\mathrm{var}(X \,|\, Y)\big] = \frac{1}{12}\int_0^\ell \frac{1}{\ell} y^2 dy = \frac{1}{12}\frac{1}{3\ell} y^3 \Big|_0^\ell = \frac{\ell^2}{36}.$$

Using now the law of conditional variances, we obtain

$$\mathrm{var}(X) = \mathbf{E}\big[\mathrm{var}(X \,|\, Y)\big] + \mathrm{var}\big(\mathbf{E}[X \,|\, Y]\big) = \frac{\ell^2}{48} + \frac{\ell^2}{36} = \frac{7\ell^2}{144}.$$

Example 4.19. Averaging Quiz Scores by Section – Variance. The setting is the same as in Example 4.17 and we consider the random variables

$$X = \text{quiz score of a student},$$
$$Y = \text{section of a student}, \quad (Y \in \{1, \ldots, S\}).$$

Let n_s be the number of students in section s, and let n be the total number of students. We interpret the different quantities in the formula

$$\mathrm{var}(X) = \mathbf{E}\big[\mathrm{var}(X \,|\, Y)\big] + \mathrm{var}\big(\mathbf{E}[X \,|\, Y]\big).$$

In this context, $\mathrm{var}(X \,|\, Y = s)$ is the variance of the quiz scores within section s. Then, $\mathbf{E}\big[\mathrm{var}(X \,|\, Y)\big]$ is the average of the section variances. This latter expectation is an average over the probability distribution of Y, i.e.,

$$\mathbf{E}\big[\mathrm{var}(X \,|\, Y)\big] = \sum_{s=1}^{S} \frac{n_s}{n} \mathrm{var}(X \,|\, Y = s).$$

Recall that $\mathbf{E}[X \,|\, Y = s]$ is the average score in section s. Then, $\mathrm{var}\big(\mathbf{E}[X \,|\, Y]\big)$ is a measure of the variability of the averages of the different sections. The law of conditional variances states that the total quiz score variance can be broken into two parts:

(a) The average score variability $\mathbf{E}\big[\mathrm{var}(X \mid Y)\big]$ *within* individual sections.

(b) The variability $\mathrm{var}\big(\mathbf{E}[X \mid Y]\big)$ *between* sections.

We have seen earlier that the law of iterated expectations (in the form of the total expectation theorem) can be used to break down complicated expectation calculations, by considering different cases. A similar method applies to variance calculations.

Example 4.20. Computing Variances by Conditioning. Consider a continuous random variable X with the PDF given in Fig. 4.7. We define an auxiliary random variable Y as follows:

$$Y = \begin{cases} 1, & \text{if } x < 1, \\ 2, & \text{of } x \geq 1. \end{cases}$$

Here, $\mathbf{E}[X \mid Y]$ takes the values $1/2$ and $3/2$, with probabilities $1/3$ and $2/3$, respectively. Thus, the mean of $\mathbf{E}[X \mid Y]$ is $7/6$. Therefore,

$$\mathrm{var}\big(\mathbf{E}[X \mid Y]\big) = \frac{1}{3}\left(\frac{1}{2} - \frac{7}{6}\right)^2 + \frac{2}{3}\left(\frac{3}{2} - \frac{7}{6}\right)^2 = \frac{2}{9}.$$

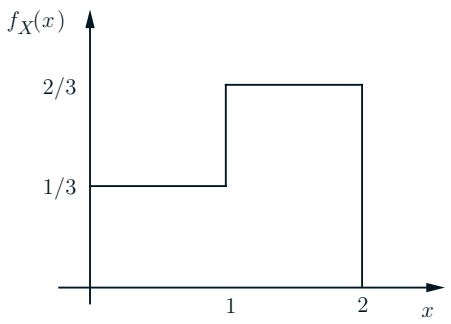

Figure 4.7: The PDF in Example 4.20.

Conditioned on either value of Y, X is uniformly distributed on a unit length interval. Therefore, $\mathrm{var}(X \mid Y = y) = 1/12$ for each of the two possible values of y, and $\mathbf{E}\big[\mathrm{var}(X \mid Y)\big] = 1/12$. Putting everything together, we obtain

$$\mathrm{var}(X) = \mathbf{E}\big[\mathrm{var}(X \mid Y)\big] + \mathrm{var}\big(\mathbf{E}[X \mid Y]\big) = \frac{1}{12} + \frac{2}{9} = \frac{11}{36}.$$

We summarize the main points in this section.

The Mean and Variance of a Conditional Expectation

- $\mathbf{E}[X \mid Y = y]$ is a number, whose value depends on y.
- $\mathbf{E}[X \mid Y]$ is a function of the random variable Y, hence a random variable. Its experimental value is $\mathbf{E}[X \mid Y = y]$ whenever the experimental value of Y is y.
- $\mathbf{E}\big[\mathbf{E}[X \mid Y]\big] = \mathbf{E}[X]$ (law of iterated expectations).
- $\mathrm{var}(X \mid Y)$ is a random variable whose experimental value is $\mathrm{var}(X \mid Y = y)$, whenever the experimental value of Y is y.
- $\mathrm{var}(X) = \mathbf{E}\big[\mathrm{var}(X \mid Y)\big] + \mathrm{var}\big(\mathbf{E}[X \mid Y]\big)$.

4.4 SUM OF A RANDOM NUMBER OF INDEPENDENT RANDOM VARIABLES

In our discussion so far of sums of random variables, we have always assumed that the number of variables in the sum is known and fixed, i.e., it is nonrandom. In this section we will consider the case where the number of random variables being added is itself random. In particular, we consider the sum

$$Y = X_1 + \cdots + X_N,$$

where N is a random variable that takes nonnegative integer values, and $X_1, X_2, \ldots$ are identically distributed random variables. We assume that $N, X_1, X_2, \ldots$ are independent, meaning that any finite subcollection of these random variables are independent.

We first note that the randomness of N can affect significantly the character of the random sum $Y = X_1 + \cdots + X_N$. In particular, the PMF/PDF of $Y = \sum_{i=1}^{N} Y_i$ is much different from the PMF/PDF of the sum $\overline{Y} = \sum_{i=1}^{\mathbf{E}[N]} Y_i$ where N has been replaced by its expected value (assuming that $\mathbf{E}[N]$ is integer). For example, let X_i be uniformly distributed in the interval $[0, 1]$, and let N be equal to 1 or 3 with probability 1/2 each. Then the PDF of the random sum Y takes values in the interval $[0, 3]$, whereas if we replace N by its expected value $\mathbf{E}[N] = 2$, the sum $\overline{Y} = X_1 + X_2$ takes values in the interval $[0, 2]$. Furthermore, using the total probability theorem, we see that the PDF of Y is a mixture of the uniform PDF and the PDF of $X_1 + X_2 + X_3$, and has considerably different character than the triangular PDF of $\overline{Y} = X_1 + X_2$ which is given in Fig. 4.4.

Let us denote by μ and σ^2 the common mean and the variance of the X_i. We wish to derive formulas for the mean, variance, and the transform of Y . The

method that we follow is to first condition on the event $N = n$, under which we have the sum of a *fixed* number of random of random variables, a case that we already know how to handle.

Fix some number n. The random variable $X_1 + \cdots + X_n$ is independent of N and, therefore, independent of the event $\{N = n\}$. Hence,

$$\begin{aligned}\mathbf{E}[Y \mid N = n] &= \mathbf{E}[X_1 + \cdots + X_N \mid N = n] \\ &= \mathbf{E}[X_1 + \cdots + X_n \mid N = n] \\ &= \mathbf{E}[X_1 + \cdots + X_n] \\ &= n\mu.\end{aligned}$$

This is true for every nonnegative integer n and, therefore,

$$\mathbf{E}[Y \mid N] = N\mu.$$

Using the law of iterated expectations, we obtain

$$\mathbf{E}[Y] = \mathbf{E}\big[\mathbf{E}[Y \mid N]\big] = \mathbf{E}[\mu N] = \mu\mathbf{E}[N].$$

Similarly,

$$\begin{aligned}\mathrm{var}(Y \mid N = n) &= \mathrm{var}(X_1 + \cdots + X_N \mid N = n) \\ &= \mathrm{var}(X_1 + \cdots + X_n) \\ &= n\sigma^2.\end{aligned}$$

Since this is true for every nonnegative integer n, the random variable $\mathrm{var}(Y \mid N)$ is equal to $N\sigma^2$. We now use the law of conditional variances to obtain

$$\begin{aligned}\mathrm{var}(Y) &= \mathbf{E}\big[\mathrm{var}(Y \mid N)\big] + \mathrm{var}\big(\mathbf{E}[Y \mid N]\big) \\ &= \mathbf{E}[N]\sigma^2 + \mathrm{var}(N\mu) \\ &= \mathbf{E}[N]\sigma^2 + \mu^2\mathrm{var}(N).\end{aligned}$$

The calculation of the transform proceeds along similar lines. The transform associated with Y, conditional on $N = n$, is $\mathbf{E}[e^{sY} \mid N = n]$. However, conditioned on $N = n$, Y is the sum of the independent random variables $X_1, \ldots, X_n$, and

$$\begin{aligned}\mathbf{E}[e^{sY} \mid N = n] &= \mathbf{E}\big[e^{sX_1} \cdots e^{sX_N} \mid N = n\big] = \mathbf{E}\big[e^{sX_1} \cdots e^{sX_n}\big] \\ &= \mathbf{E}[e^{sX_1}] \cdots \mathbf{E}[e^{sX_n}] = \big(M_X(s)\big)^n.\end{aligned}$$

Using the law of iterated expectations, the (unconditional) transform associated with Y is

$$\mathbf{E}[e^{sY}] = \mathbf{E}\big[\mathbf{E}[e^{sY} \mid N]\big] = \mathbf{E}\big[\big(M_X(s)\big)^N\big] = \sum_{n=0}^{\infty}(M_X(s))^n p_N(n).$$

This is similar to the transform $M_N(s) = \mathbf{E}[e^{sN}]$ associated with N, except that e^s is replaced by $M_X(s)$.

Example 4.21. A remote village has three gas stations, and each one of them is open on any given day with probability 1/2, independently of the others. The amount of gas available in each gas station is unknown and is uniformly distributed between 0 and 1000 gallons. We wish to characterize the distribution of the total amount of gas available at the gas stations that are open.

The number N of open gas stations is a binomial random variable with $p = 1/2$ and the corresponding transform is

$$M_N(s) = (1 - p + pe^s)^3 = \frac{1}{8}(1 + e^s)^3.$$

The transform $M_X(s)$ associated with the amount of gas available in an open gas station is

$$M_X(s) = \frac{e^{1000s} - 1}{1000s}.$$

The transform associated with the total amount Y available is the same as $M_N(s)$, except that each occurrence of e^s is replaced with $M_X(s)$, i.e.,

$$M_Y(s) = \frac{1}{8}\left(1 + \left(\frac{e^{1000s} - 1}{1000s}\right)\right)^3.$$

Example 4.22. Sum of a Geometric Number of Independent Exponential Random Variables. Jane visits a number of bookstores, looking for *Great Expectations*. Any given bookstore carries the book with probability p, independently of the others. In a typical bookstore visited, Jane spends a random amount of time, exponentially distributed with parameter λ, until she either finds the book or she decides that the bookstore does not carry it. Assuming that Jane will keep visiting bookstores until she buys the book and that the time spent in each is independent of everything else, we wish to determine the mean, variance, and PDF of the total time spent in bookstores.

The total number N of bookstores visited is geometrically distributed with parameter p. Hence, the total time Y spent in bookstores is the sum of a geometrically distributed number N of independent exponential random variables $X_1, X_2, \ldots$. We have

$$\mathbf{E}[Y] = \mathbf{E}[N]\mathbf{E}[X] = \frac{1}{p} \cdot \frac{1}{\lambda}.$$

Using the formulas for the variance of geometric and exponential random variables, we also obtain

$$\mathrm{var}(Y) = \mathbf{E}[N]\mathrm{var}(X) + (\mathbf{E}[X])^2\mathrm{var}(N) = \frac{1}{p} \cdot \frac{1}{\lambda^2} + \frac{1}{\lambda^2} \cdot \frac{1-p}{p^2} = \frac{1}{\lambda^2 p^2}.$$

In order to find the transform $M_Y(s)$, let us recall that

$$M_X(s) = \frac{\lambda}{\lambda - s}, \qquad M_N(s) = \frac{pe^s}{1-(1-p)e^s}.$$

Then, $M_Y(s)$ is found by starting with $M_N(s)$ and replacing each occurrence of e^s with $M_X(s)$. This yields

$$M_Y(s) = \frac{pM_X(s)}{1-(1-p)M_X(s)} = \frac{\dfrac{p\lambda}{\lambda - s}}{1-(1-p)\dfrac{\lambda}{\lambda - s}},$$

which simplifies to

$$M_Y(s) = \frac{p\lambda}{p\lambda - s}.$$

We recognize this as the transform of an exponentially distributed random variable with parameter $p\lambda$, and therefore,

$$f_Y(y) = p\lambda e^{-p\lambda y}, \qquad y \geq 0.$$

This result can be surprising because the sum of a *fixed* number n of independent exponential random variables is not exponentially distributed. For example, if $n = 2$, the transform associated with the sum is $\big(\lambda/(\lambda - s)\big)^2$, which does not correspond to the exponential distribution.

Example 4.23. Sum of a Geometric Number of Independent Geometric Random Variables. This example is a discrete counterpart of the preceding one. We let N be geometrically distributed with parameter p. We also let each random variable X_i be geometrically distributed with parameter q. We assume that all of these random variables are independent. Let $Y = X_1 + \cdots + X_N$. We have

$$M_N(s) = \frac{pe^s}{1-(1-p)e^s}, \qquad M_X(s) = \frac{qe^s}{1-(1-q)e^s}.$$

To determine $M_Y(s)$, we start with the formula for $M_N(s)$ and replace each occurrence of e^s with $M_X(s)$. This yields

$$M_Y(s) = \frac{pM_X(s)}{1-(1-p)M_X(s)},$$

and, after some algebra,

$$M_Y(s) = \frac{pqe^s}{1-(1-pq)e^s}.$$

We conclude that Y is geometrically distributed, with parameter pq.

Properties of Sums of a Random Number of Independent Random Variables

Let $X_1, X_2, \ldots$ be random variables with common mean μ and common variance σ^2. Let N be a random variable that takes nonnegative integer values. We assume that all of these random variables are independent, and consider

$$Y = X_1 + \cdots + X_N.$$

Then,

- $\mathbf{E}[Y] = \mu\mathbf{E}[N]$.
- $\text{var}(Y) = \sigma^2\mathbf{E}[N] + \mu^2\text{var}(N)$.
- The transform $M_Y(s)$ is found by starting with the transform $M_N(s)$ and replacing each occurrence of e^s with $M_X(s)$.

4.5 COVARIANCE AND CORRELATION

The **covariance** of two random variables X and Y is denoted by $\text{cov}(X, Y)$, and is defined by

$$\text{cov}(X, Y) = \mathbf{E}\big[\big(X - \mathbf{E}[X]\big)\big(Y - \mathbf{E}[Y]\big)\big].$$

When $\text{cov}(X, Y) = 0$, we say that X and Y are **uncorrelated**.

Roughly speaking, a positive or negative covariance indicates that the values of $X - \mathbf{E}[X]$ and $Y - \mathbf{E}[Y]$ obtained in a single experiment "tend" to have the same or the opposite sign, respectively (see Fig. 4.8). Thus the sign of the covariance provides an important qualitative indicator of the relation between X and Y.

If X and Y are independent, then

$$\text{cov}(X, Y) = \mathbf{E}\big[\big(X - \mathbf{E}[X]\big)\big(Y - \mathbf{E}[Y]\big)\big] = \mathbf{E}\big[X - \mathbf{E}[X]\big]\mathbf{E}\big[Y - \mathbf{E}[Y]\big] = 0.$$

Thus if X and Y are independent, they are also uncorrelated. However, the reverse is not true, as illustrated by the following example.

Example 4.24. The pair of random variables (X, Y) takes the values $(1, 0)$, $(0, 1)$, $(-1, 0)$, and $(0, -1)$, each with probability $1/4$ (see Fig. 4.9). Thus, the marginal PMFs of X and Y are symmetric around 0, and $\mathbf{E}[X] = \mathbf{E}[Y] = 0$. Furthermore, for all possible value pairs (x, y), either x or y is equal to 0, which implies that $XY = 0$ and $\mathbf{E}[XY] = 0$. Therefore,

$$\text{cov}(X, Y) = \mathbf{E}\big[\big(X - \mathbf{E}[X]\big)\big(Y - \mathbf{E}[Y]\big)\big] = \mathbf{E}[XY] = 0,$$

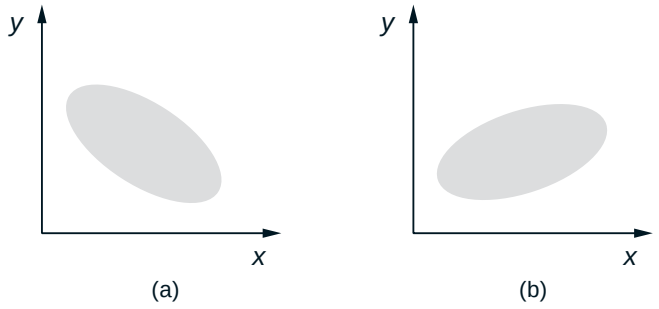

Figure 4.8: Examples of positively and negatively correlated random variables. Here X and Y are uniformly distributed over the ellipses shown. In case (a) the covariance $\text{cov}(X, Y)$ is negative, while in case (b) it is positive.

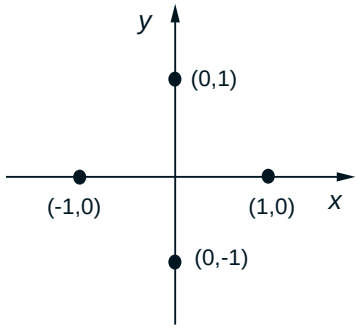

Figure 4.9: Joint PMF of X and Y for Example 4.21. Each of the four points shown has probability 1/4. Here X and Y are uncorrelated but not independent.

and X and Y are uncorrelated. However, X and Y are not independent since, for example, a nonzero value of X fixes the value of Y to zero.

The **correlation coefficient** ρ of two random variables X and Y that have nonzero variances is defined as

$$\rho = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}.$$

It may be viewed as a normalized version of the covariance $\text{cov}(X, Y)$, and in fact it can be shown that ρ ranges from -1 to 1 (see the end-of-chapter problems).

If $\rho > 0$ (or $\rho < 0$), then the values of $x - \mathbf{E}[X]$ and $y - \mathbf{E}[Y]$ "tend" to have the same (or opposite, respectively) sign, and the size of $|\rho|$ provides a normalized measure of the extent to which this is true. In fact, always assuming that X and Y have positive variances, it can be shown that $\rho = 1$ (or $\rho = -1$) if and only if there exists a positive (or negative, respectively) constant c such that

$$y - \mathbf{E}[Y] = c\big(x - \mathbf{E}[X]\big), \qquad \text{for all possible numerical values } (x, y)$$

(see the end-of-chapter problems). The following example illustrates in part this property.

Example 4.25. Consider n independent tosses of a biased coin with probability of a head equal to p. Let X and Y be the numbers of heads and of tails, respectively, and let us look at the correlation of X and Y. Here, for all possible pairs of values (x, y), we have $x + y = n$, and we also have $\mathbf{E}[X] + \mathbf{E}[Y] = n$. Thus,

$$x - \mathbf{E}[X] = -\big(y - \mathbf{E}[Y]\big), \qquad \text{for all possible } (x, y).$$

We will calculate the correlation coefficient of X and Y, and verify that it is indeed equal to -1.

We have

$$\begin{aligned}\operatorname{cov}(X, Y) &= \mathbf{E}\big[\big(X - \mathbf{E}[X]\big)\big(Y - \mathbf{E}[Y]\big)\big] \\ &= -\mathbf{E}\left[(X - \mathbf{E}[X])^2\right] \\ &= -\operatorname{var}(X).\end{aligned}$$

Hence, the correlation coefficient is

$$\rho(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}} = \frac{-\operatorname{var}(X)}{\sqrt{\operatorname{var}(X)\operatorname{var}(X)}} = -1.$$

The covariance can be used to obtain a formula for the variance of the sum of several (not necessarily independent) random variables. In particular, if $X_1, X_2, \ldots, X_n$ are random variables with finite variance, we have

$$\operatorname{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{var}(X_i) + 2 \sum_{\substack{i,j=1 \\ i<j}}^{n} \operatorname{cov}(X_i, X_j).$$

This can be seen from the following calculation, where for brevity, we denote $\tilde{X}_i = X_i - \mathbf{E}[X_i]$:

$$\begin{aligned}\operatorname{var}\left(\sum_{i=1}^{n} X_i\right) &= \mathbf{E}\left[\left(\sum_{i=1}^{n} \tilde{X}_i\right)^2\right] \\ &= \mathbf{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{X}_i \tilde{X}_j\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{E}[\tilde{X}_i \tilde{X}_j] \\ &= \sum_{i=1} \mathbf{E}[\tilde{X}_i{}^2] + 2 \sum_{\substack{i,j=1 \\ i<j}}^{n} \mathbf{E}[\tilde{X}_i \tilde{X}_j] \\ &= \sum_{i=1}^{n} \operatorname{var}(X_i) + 2 \sum_{\substack{i,j=1 \\ i<j}}^{n} \operatorname{cov}(X_i, X_j).\end{aligned}$$

The following example illustrates the use of this formula.

Example 4.26. Consider the hat problem discussed in Section 2.5, where n people throw their hats in a box and then pick a hat at random. Let us find the variance of X, the number of people that pick their own hat. We have

$$X = X_1 + \cdots + X_n,$$

where X_i is the random variable that takes the value 1 if the ith person selects his/her own hat, and takes the value 0 otherwise. Noting that X_i is Bernoulli with parameter $p = \mathbf{P}(X_i = 1) = 1/n$, we obtain

$$\text{var}(X_i) = \frac{1}{n}\left(1 - \frac{1}{n}\right).$$

For $i \neq j$, we have

$$\begin{aligned}
\text{cov}(X_i, X_j) &= \mathbf{E}\big[\big(X_i - \mathbf{E}[X_i]\big)\big(X_j - \mathbf{E}[X_j]\big)\big] \\
&= \mathbf{E}[X_i X_j] - \mathbf{E}[X_i]\mathbf{E}[X_j] \\
&= \mathbf{P}(X_i = 1 \text{ and } X_j = 1) - \mathbf{P}(X_i = 1)\mathbf{P}(X_j = 1) \\
&= \mathbf{P}(X_i = 1)\mathbf{P}(X_j = 1 \,|\, X_i = 1) - \mathbf{P}(X_i = 1)\mathbf{P}(X_j = 1) \\
&= \frac{1}{n}\frac{1}{n-1} - \frac{1}{n^2} \\
&= \frac{1}{n^2(n-1)}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\text{var}(X) &= \text{var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \sum_{i=1}^{n} \text{var}(X_i) + 2 \sum_{\substack{i,j=1 \\ i<j}}^{n} \text{cov}(X_i, X_j) \\
&= n\frac{1}{n}\left(1 - \frac{1}{n}\right) + 2\frac{n(n-1)}{2}\frac{1}{n^2(n-1)} \\
&= 1.
\end{aligned}$$

4.6 LEAST SQUARES ESTIMATION

In many practical contexts, we want to form an estimate of the value of a random variable X given the value of a related random variable Y, which may be viewed

as some form of "measurement" of X. For example, X may be the range of an aircraft and Y may be a noise-corrupted measurement of that range. In this section we discuss a popular formulation of the estimation problem, which is based on finding the estimate c that minimizes the expected value of the squared error $(X-c)^2$ (hence the name "least squares").

If the value of Y is not available, we may consider finding an estimate (or prediction) c of X. The estimation error $X-c$ is random (because X is random), but the mean squared error $\mathbf{E}\big[(X-c)^2\big]$ is a number that depends on c and can be minimized over c. With respect to this criterion, it turns out that the best possible estimate is $c=\mathbf{E}[X]$, as we proceed to verify.

Let $m=\mathbf{E}[X]$. For any estimate c, we have

$$\begin{aligned}\mathbf{E}\big[(X-c)^2\big] &= \mathbf{E}\big[(X-m+m-c)^2\big]\\ &= \mathbf{E}\big[(X-m)^2\big]+2\mathbf{E}\big[(X-m)(m-c)\big]+\mathbf{E}\big[(m-c)^2\big]\\ &= \mathbf{E}\big[(X-m)^2\big]+2\mathbf{E}[X-m](m-c)+(m-c)^2\\ &= \mathbf{E}\big[(X-m)^2\big]+(m-c)^2,\end{aligned}$$

where we used the fact $\mathbf{E}[X-m]=0$. The first term in the right-hand side is the variance of X and is unaffected by our choice of c. Therefore, we should choose c in a way that minimizes the second term, which leads to $c=m=\mathbf{E}[X]$ (see Fig. 4.10).

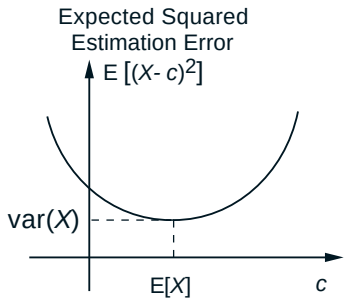

Figure 4.10: The mean squared error $\mathbf{E}\big[(X-c)^2\big]$, as a function of the estimate c, is a quadratic in c and is minimized when $c=\mathbf{E}[X]$. The minimum value of the mean squared error is $\mathrm{var}(X)$.

Suppose now that we observe the experimental value y of some related random variable Y, before forming an estimate of X. How can we exploit this additional information? Once we are told that Y takes a particular value y, the situation is identical to the one considered earlier, except that we are now in a new "universe," where everything is conditioned on $Y=y$. We can therefore adapt our earlier conclusion and assert that $c=\mathbf{E}[X\,|\,Y=y]$ minimizes the

conditional mean squared error $\mathbf{E}\big[(c - X)^2 \,|\, Y = y\big]$. Note that the resulting estimate c depends on the experimental value y of Y (as it should). Thus, we call $\mathbf{E}[X \,|\, Y = y]$ the *least-squares estimate* of X given the experimental value y.

Example 4.27. Let X be uniformly distributed in the interval $[4, 10]$ and suppose that we observe X with some random error W, that is, we observe the experimental value of the random variable

$$Y = X + W.$$

We assume that W is uniformly distributed in the interval $[-1, 1]$, and independent of X. What is the least squares estimate of X given the experimental value of Y?

We have $f_X(x) = 1/6$ for $4 \le x \le 10$, and $f_X(x) = 0$, elsewhere. Conditioned on X being equal to some x, Y is the same as $x + W$, and is uniform over the interval $[x - 1, x + 1]$. Thus, the joint PDF is given by

$$f_{X,Y}(x, y) = f_X(x) f_{Y|X}(y \,|\, x) = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12},$$

if $4 \le x \le 10$ and $x - 1 \le y \le x + 1$, and is zero for all other values of (x, y). The slanted rectangle in the right-hand side of Fig. 4.11 is the set of pairs (x, y) for which $f_{X,Y}(x, y)$ is nonzero.

Given an experimental value y of Y, the conditional PDF $f_{X|Y}$ of X is uniform on the corresponding vertical section of the slanted rectangle. The optimal estimate $\mathbf{E}[X \,|\, Y = y]$ is the midpoint of that section. In the special case of the present example, it happens to be a piecewise linear function of y.

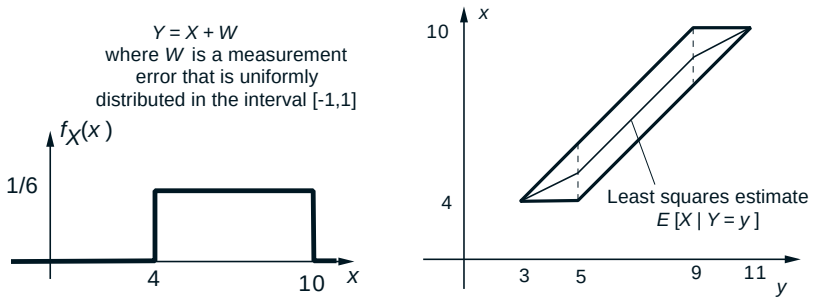

Figure 4.11: The PDFs in Example 4.27. The least squares estimate of X given the experimental value y of the random variable $Y = X + W$ depends on y and is represented by the piecewise linear function shown in the figure on the right.

As Example 4.27 illustrates, the estimate $\mathbf{E}[X \,|\, Y = y]$ depends on the observed value y and should be viewed as a function of y; see Fig. 4.12. To

amplify this point, we refer to any function of the available information as an **estimator**. Given an experimental outcome y of Y, an estimator $g(\cdot)$ (which is a function) produces an estimate $g(y)$ (which is a number). However, if y is left unspecified, then the estimator results in a random variable $g(Y)$. The expected value of the squared estimation error associated with an estimator $g(Y)$ is

$$\mathbf{E}\Big[\big(X - g(Y)\big)^2\Big].$$

Out of all estimators, it turns out that the mean squared estimation error is minimized when $g(Y) = \mathbf{E}[X \,|\, Y]$. To see this, note that if c is any number, we have

$$\mathbf{E}\Big[\big(X - \mathbf{E}[X \,|\, Y = y]\big)^2 \;\Big|\; Y = y\Big] \leq \mathbf{E}\big[(X - c)^2 \,|\, Y = y\big].$$

Consider now an estimator $g(Y)$. For a given value y of Y, $g(y)$ is a number and, therefore,

$$\mathbf{E}\Big[\big(X - \mathbf{E}[X \,|\, Y = y]\big)^2 \,|\, Y = y\Big] \leq \mathbf{E}\Big[\big(X - g(y)\big)^2 \;\Big|\; Y = y\Big].$$

This inequality is true for *every* possible experimental value y of Y. Thus,

$$\mathbf{E}\Big[\big(X - \mathbf{E}[X \,|\, Y]\big)^2 \;\Big|\; Y\Big] \leq \mathbf{E}\Big[\big(X - g(Y)\big)^2 \;\Big|\; Y\Big],$$

which is now an inequality between random variables (functions of Y). We take expectations of both sides, and use the law of iterated expectations, to conclude that

$$\mathbf{E}\Big[\big(X - \mathbf{E}[X \,|\, Y]\big)^2\Big] \leq \mathbf{E}\Big[\big(X - g(Y)\big)^2\Big]$$

for all functions $g(Y)$.

Figure 4.12: The least squares estimator.

Key Facts about Least Mean Squares Estimation

- $\mathbf{E}\big[(X-c)^2\big]$ is minimized when $c=\mathbf{E}[X]$:

$$\mathbf{E}\Big[\big(X-\mathbf{E}[X]\big)^2\Big] \le \mathbf{E}\big[(X-c)^2\big], \qquad \text{for all } c.$$

- $\mathbf{E}\big[(X-c)^2 \,|\, Y=y\big]$ is minimized when $c=\mathbf{E}[X \,|\, Y=y]$:

$$\mathbf{E}\Big[\big(X-\mathbf{E}[X \,|\, Y=y]\big)^2 \;\big|\; Y=y\Big] \le \mathbf{E}\big[(X-c)^2 \,|\, Y=y\big], \quad \text{for all } c.$$

- Out of all estimators $g(Y)$ of X based on Y, the mean squared estimation error $\mathbf{E}\Big[\big(X-g(Y)\big)^2\Big]$ is minimized when $g(Y)=\mathbf{E}[X \,|\, Y]$:

$$\mathbf{E}\Big[\big(X-\mathbf{E}[X \,|\, Y]\big)^2\Big] \le \mathbf{E}\Big[\big(X-g(Y)\big)^2\Big], \qquad \text{for all functions } g(Y).$$

Some Properties of the Estimation Error

Let us introduce the notation

$$\hat{X}=\mathbf{E}[X \,|\, Y], \qquad\qquad \tilde{X}=X-\hat{X},$$

for the (optimal) estimator and the associated estimation error, respectively. Note that both $\hat{X}$ and $\tilde{X}$ are random variables, and by the law of iterated expectations,

$$\mathbf{E}[\tilde{X}]=\mathbf{E}\big[X-\mathbf{E}[X \,|\, Y]\big]=\mathbf{E}[X]-\mathbf{E}[X]=0.$$

The equation $\mathbf{E}[\tilde{X}]=0$ remains valid even if we condition on Y, because

$$\mathbf{E}[\tilde{X} \,|\, Y]=\mathbf{E}[X-\hat{X} \,|\, Y]=\mathbf{E}[X \,|\, Y]-\mathbf{E}[\hat{X} \,|\, Y]=\hat{X}-\hat{X}=0.$$

We have used here the fact that $\hat{X}$ is completely determined by Y and therefore $\mathbf{E}[\hat{X} \,|\, Y]=\hat{X}$. For similar reasons,

$$\mathbf{E}\big[\big(\hat{X}-\mathbf{E}[X]\big)\tilde{X} \,|\, Y\big]=\big(\hat{X}-\mathbf{E}[X]\big)\mathbf{E}[\tilde{X} \,|\, Y]=0.$$

Taking expectations and using the law of iterated expectations, we obtain

$$\mathbf{E}\big[\big(\hat{X}-\mathbf{E}[X]\big)\tilde{X}\big]=0.$$

Note that $X = \hat{X} + \tilde{X}$, which yields $X - \mathbf{E}[X] = \hat{X} - \mathbf{E}[X] + \tilde{X}$. We square both sides of the latter equality and take expectations to obtain

$$\begin{aligned}
\mathrm{var}(X) &= \mathbf{E}\big[\big(X - \mathbf{E}[X]\big)^2\big] \\
&= \mathbf{E}\Big[\big(\hat{X} - \mathbf{E}[X] + \tilde{X}\big)^2\Big] \\
&= \mathbf{E}\Big[\big(\hat{X} - \mathbf{E}[X]\big)^2\Big] + \mathbf{E}[\tilde{X}^2] + 2\mathbf{E}\big[\big(\hat{X} - \mathbf{E}[X]\big)\tilde{X}\big] \\
&= \mathbf{E}\Big[\big(\hat{X} - \mathbf{E}[X]\big)^2\Big] + \mathbf{E}[\tilde{X}^2] \\
&= \mathrm{var}(\hat{X}) + \mathrm{var}(\tilde{X}).
\end{aligned}$$

(The last equality holds because $\mathbf{E}[\hat{X}] = \mathbf{E}[X]$ and $\mathbf{E}[\tilde{X}] = 0$.) In summary, we have established the following important formula, which is just another version of the law of conditional variances introduced in Section 4.3.

$$\mathrm{var}(X) = \mathrm{var}(\hat{X}) + \mathrm{var}(\tilde{X}).$$

Example 4.28. Let us say that the observed random variable Y is *uninformative* if the mean squared estimation error $\mathbf{E}[\tilde{X}^2] = \mathrm{var}(\tilde{X})$ is the same as the unconditional variance $\mathrm{var}(X)$ of X. When is this the case?

Using the formula

$$\mathrm{var}(X) = \mathrm{var}(\hat{X}) + \mathrm{var}(\tilde{X}),$$

we see that Y is uninformative if and only if $\mathrm{var}\big(\hat{X}\big) = 0$. The variance of a random variable is zero if and only if that random variable is a constant, equal to its mean. We conclude that Y is uninformative if and only if $\hat{X} = \mathbf{E}[X \,|\, Y] = \mathbf{E}[X]$, for every realization of Y.

If X and Y are independent, we have $\mathbf{E}[X \,|\, Y] = \mathbf{E}[X]$ and Y is indeed uninformative, which is quite intuitive. The converse, however, is not true. That is, it is possible for $\mathbf{E}[X \,|\, Y]$ to be always equal to the constant $\mathbf{E}[X]$, without X and Y being independent. (Can you construct an example?)

Estimation Based on Several Measurements

So far, we have discussed the case where we estimate one random variable X on the basis of another random variable Y. In practice, one often has access to the experimental values of several random variables $Y_1, \ldots, Y_n$, that can be used to estimate X. Generalizing our earlier discussion, and using essentially

the same argument, the mean squared estimation error is minimized if we use $\mathbf{E}[X \mid Y_1, \ldots, Y_n]$ as our estimator. That is,

$$\mathbf{E}\Big[\big(X - \mathbf{E}[X \mid Y_1, \ldots, Y_n]\big)^2\Big] \leq \mathbf{E}\Big[\big(X - g(Y_1, \ldots, Y_n)\big)^2\Big],$$

for all functions $g(Y_1, \ldots, Y_n)$.

This provides a complete solution to the general problem of least squares estimation, but is sometimes difficult to implement, because:

(a) In order to compute the conditional expectation $\mathbf{E}[X \mid Y_1, \ldots, Y_n]$, we need a complete probabilistic model, that is, the joint PDF $f_{X,Y_1,\ldots,Y_n}(\cdot)$ of $n+1$ random variables.

(b) Even if this joint PDF is available, $\mathbf{E}[X \mid Y_1, \ldots, Y_n]$ can be a very complicated function of $Y_1, \ldots, Y_n$.

As a consequence, practitioners often resort to approximations of the conditional expectation or focus on estimators that are not optimal but are simple and easy to implement. The most common approach involves *linear estimators*, of the form

$$a_1 Y_1 + \cdots + a_n Y_n + b.$$

Given a particular choice of $a_1, \ldots, a_n, b$, the corresponding mean squared error is

$$\mathbf{E}\big[(X - a_1 Y_1 - \cdots - a_n Y_n - b)^2\big],$$

and it is meaningful to choose the coefficients $a_1, \ldots, a_n, b$ in a way that minimizes the above expression. This problem is relatively easy to solve and only requires knowledge of the means, variances, and covariances of the different random variables. We develop the solution for the case where $n = 1$.

Linear Least Mean Squares Estimation Based on a Single Measurement

We are interested in finding a and b that minimize the mean squared estimation error $\mathbf{E}\big[(X - aY - b)^2\big]$, associated with a linear estimator $aY + b$ of X. Suppose that a has already been chosen. How should we choose b? This is the same as having to choose a constant b to estimate the random variable $aX - Y$ and, by our earlier results, the best choice is to let $b = \mathbf{E}[X - aY] = \mathbf{E}[X] - a\mathbf{E}[Y]$.

It now remains to minimize, with respect to a, the expression

$$\mathbf{E}\Big[\big(X - aY - \mathbf{E}[X] + a\mathbf{E}[Y]\big)^2\Big],$$

which is the same as

$$\begin{aligned}
&\mathbf{E}\Big[\big((X - \mathbf{E}[X]) - a(Y - \mathbf{E}[Y])\big)^2\Big] \\
&\quad = \mathbf{E}\big[(X - \mathbf{E}[X])^2\big] + a^2\mathbf{E}\big[(Y - \mathbf{E}[Y])^2\big] - 2a\mathbf{E}\big[\big(X - \mathbf{E}[X]\big)\big(Y - \mathbf{E}[Y]\big)\big] \\
&\quad = \sigma_X^2 + a^2\sigma_Y^2 - 2a \cdot \mathrm{cov}(X, Y),
\end{aligned}$$

where $\mathrm{cov}(X,Y)$ is the covariance of X and Y:

$$\mathrm{cov}(X,Y) = \mathbf{E}\big[(X-\mathbf{E}[X])(Y-\mathbf{E}[Y])\big].$$

This is a quadratic function of a, which is minimized at the point where its derivative is zero, that is, if

$$a = \frac{\mathrm{cov}(X,Y)}{\sigma_Y^2} = \frac{\rho\sigma_X\sigma_Y}{\sigma_Y^2} = \rho\frac{\sigma_X}{\sigma_Y},$$

where

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}$$

is the correlation coefficient. With this choice of a, the mean squared estimation error is given by

$$\begin{aligned}\sigma_X^2 + a^2\sigma_Y^2 - 2a\cdot\mathrm{cov}(X,Y) =& \sigma_X^2 + \rho^2\frac{\sigma_X^2}{\sigma_Y^2}\sigma_Y^2 - 2\rho\frac{\sigma_X}{\sigma_y}\rho\sigma_X\sigma_Y \\ =& (1-\rho^2)\sigma_X^2.\end{aligned}$$

Linear Least Mean Squares Estimation Formulas

The least mean squares linear estimator of X based on Y is

$$\mathbf{E}[X] + \frac{\mathrm{cov}(X,Y)}{\sigma_Y^2}\big(Y-\mathbf{E}[Y]\big).$$

The resulting mean squared estimation error is equal to

$$(1-\rho^2)\mathrm{var}(X).$$

4.7 THE BIVARIATE NORMAL DISTRIBUTION

We say that two random variables X and Y have a *bivariate normal* distribution if there are two independent normal random variables U and V and some scalars a, b, c, d, such that

$$X = aU + bV, \qquad Y = cU + dV.$$

To keep the discussion simple, we restrict ourselves to the case where U, V (and therefore, X and Y as well) have zero mean.

A most important property of the bivariate normal distribution is the following:

> If two random variables X and Y have a bivariate normal distribution and are uncorrelated, then they are independent.

This property can be verified using multivariate transforms. We assume that X and Y have a bivariate normal distribution and are uncorrelated. Recall that if z is a zero-mean normal random variable with variance σ_Z^2, then $\mathbf{E}[e^Z] = M_Z(1) = \sigma_Z^2/2$. Fix some scalars s_1, s_2 and let $Z = s_1X + s_2Y$. Then, Z is the sum of the independent normal random variables $(as_1 + cs_2)U$ and $(bs_1 + ds_2)V$, and is therefore normal. Since X and Y are uncorrelated, the variance of Z is $s_1^2\sigma_X^2 + s_2^2\sigma_Y^2$. Then,

$$\begin{aligned} M_{X,Y}(s_1, s_2) &= \mathbf{E}\left[e^{s_1X+s_2Y}\right] \\ &= \mathbf{E}[e^Z] \\ &= e^{(s_1^2\sigma_X^2+s_2^2\sigma_Y^2)/2.} \end{aligned}$$

Let $\overline{X}$ and $\overline{Y}$ be *independent* zero-mean normal random variables with the same variances σ_X^2 and σ_Y^2 as X and Y. Since they are independent, they are uncorrelated, and the same argument as above yields

$$M_{\overline{X},\overline{Y}}(s_1, s_2) = e^{(s_1^2\sigma_X^2+s_2^2\sigma_Y^2)/2.}.$$

Thus, the two pairs of random variables (X, Y) and $(\overline{X}, \overline{Y})$ are associated with the same multivariate transform. Since the multivariate transform completely determines the joint PDF, it follows that the pair (X, Y) has the same joint PDF as the pair $(\overline{X}, \overline{Y})$. Since $\overline{X}$ and $\overline{Y}$ are independent, X and Y must also be independent.

Let us define

$$\hat{X} = \frac{\mathbf{E}[XY]}{\mathbf{E}[Y^2]}Y, \qquad \tilde{X} = X - \hat{X}.$$

Thus, $\hat{X}$ is the best *linear* estimator of X given Y, and $\tilde{X}$ is the estimation error. Since X and Y are linear combinations of independent normal random variables U and V, it follows that Y and $\tilde{X}$ are also linear combinations of U and V. In particular, Y and $\tilde{X}$ have a bivariate normal distribution. Furthermore,

$$\operatorname{cov}(Y, \tilde{X}) = \mathbf{E}[Y\tilde{X}] = \mathbf{E}[YX] - \mathbf{E}[Y\hat{X}] = \mathbf{E}[YX] - \frac{\mathbf{E}[XY]}{\mathbf{E}[Y^2]}\mathbf{E}[Y^2] = 0.$$

Thus, Y and $\tilde{X}$ are uncorrelated and, therefore, independent. Since $\hat{X}$ is a scalar multiple of Y, we also see that $\hat{X}$ and $\tilde{X}$ are independent.

We now start from the identity

$$X = \hat{X} + \tilde{X},$$

which implies that

$$\mathbf{E}[X \,|\, Y] = \mathbf{E}[\hat{X} \,|\, Y] + \mathbf{E}[\tilde{X} \,|\, Y].$$

But $\mathbf{E}[\hat{X} \,|\, Y] = \hat{X}$ because $\hat{X}$ is completely determined by Y. Also, $\tilde{X}$ is independent of Y and

$$\mathbf{E}[\tilde{X} \,|\, Y] = \mathbf{E}[\tilde{X}] = \mathbf{E}[X - \hat{X}] = 0.$$

(The last equality was obtained because X and Y are assumed to have zero mean and $\hat{X}$ is a constant multiple of Y.) Putting everything together, we come to the important conclusion that the best linear estimator $\hat{X}$ is of the form

$$\hat{X} = \mathbf{E}[X \,|\, Y].$$

Differently said, the optimal estimator $\mathbf{E}[X \,|\, Y]$ turns out to be linear.

Let us now determine the conditional density of X, conditioned on Y. We have $X = \hat{X} + \tilde{X}$. After conditioning on Y, the value of the random variable $\hat{X}$ is completely determined. On the other hand, $\tilde{X}$ is independent of Y and its distribution is not affected by conditioning. Therefore, the conditional distribution of X given Y is the same as the distribution of $\tilde{X}$, shifted by $\hat{X}$. Since $\tilde{X}$ is normal with mean zero and some variance $\sigma^2_{\tilde{X}}$, we conclude that the conditional distribution of X is also normal with mean $\hat{X}$ and variance $\sigma^2_{\tilde{X}}$.

We summarize our conclusions below. Although our discussion used the zero-mean assumption, these conclusions also hold for the non-zero mean case and we state them with this added generality.

Properties of the Bivariate Normal Distribution

Let X and Y have a bivariate normal distribution. Then:

- X and Y are independent if and only if they are uncorrelated.
- The conditional expectation is given by

$$\mathbf{E}[X \,|\, Y] = \mathbf{E}[X] + \frac{\mathrm{cov}(X, Y)}{\sigma^2_Y}(Y - \mathbf{E}[Y]).$$

  It is a linear function of Y and has a normal distribution.

- The conditional distribution of X given Y is normal with mean $\mathbf{E}[X \mid Y]$ and variance
$$\sigma_{\hat{X}}^2 = (1 - \rho^2)\sigma_X^2.$$

Finally, let us note that while if X and Y have a bivariate normal distribution, then X and Y are (individually) normal random variables, the reverse is not true even if X and Y are uncorrelated. This is illustrated in the following example.

Example 4.29. Let X have a normal distribution with zero mean and unit variance. Let z be independent of X, with $\mathbf{P}(Z = 1) = \mathbf{P}(Z = -1) = 1/2$. Let $Y = ZX$, which is also normal with zero mean (why?). Furthermore,

$$\mathbf{E}[XY] = \mathbf{E}[ZX^2] = \mathbf{E}[Z]\mathbf{E}[X^2] = 0 \times 1 = 0,$$

so X and Y are uncorrelated. On the other hand X and Y are clearly dependent. (For example, if $X = 1$, then Y must be either -1 or 1.) This may seem to contradict our earlier conclusion that zero correlation implies independence? However, in this example, the joint PDF of X and Y is *not* multivariable normal, even though both marginal distributions are normal.