Chapter 10

Expectations and Bounds

The concept of expectation, which was originally introduced in the context of discrete random variables, can be generalized to other types of random variables. For instance, the expectation of a continuous random variable is defined in terms of its probability density function (PDF). We know from our previous discussion that expectations provide an effective way to summarize the information contained in the distribution of a random variable. As we will see shortly, expectations are also very valuable in establishing bounds on probabilities.

10.1 Expectations Revisited

The definition of an expectation associated with a continuous random variable is very similar to its discrete counterpart; the weighted sum is simply replaced by a weighted integral. For a continuous random variable X with PDF $f_X(\cdot)$, the *expectation* of $g(X)$ is defined by

$$\mathrm{E}[g(X)] = \int_{\mathbb{R}} g(\xi) f_X(\xi) d\xi.$$

In particular, the *mean* of X is equal to

$$\mathrm{E}[X] = \int_{\mathbb{R}} \xi f_X(\xi) d\xi$$

and its *variance* becomes

$$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = \int_{\mathbb{R}} (\xi - \mathrm{E}[X])^2 f_X(\xi) d\xi.$$

As before, the variance of random variable X can also be computed using $\mathrm{Var}(X) = \mathrm{E}\left[X^2\right] - (E[X])^2$.

Example 84. *We wish to calculate the mean and variance of a Gaussian random variable with parameters m and σ^2. By definition, the PDF of this random variable can be written as*

$$f_X(\xi) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\xi-m)^2}{2\sigma^2}} \quad \xi \in \mathbb{R}.$$

The mean of X can be obtained through direct integration, with a change of variables,

$$\begin{aligned}
\mathrm{E}[X] &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \xi e^{-\frac{(\xi-m)^2}{2\sigma^2}} d\xi \\
&= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(\zeta + \frac{m}{\sigma}\right) e^{-\frac{\zeta^2}{2}} d\zeta \\
&= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \zeta e^{-\frac{\zeta^2}{2}} d\zeta + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{m}{\sigma} e^{-\frac{\zeta^2}{2}} d\zeta = m.
\end{aligned}$$

In finding a solution, we have leveraged the facts that $\zeta e^{-\frac{\zeta^2}{2}}$ is an absolutely integrable, odd function. We also took advantage of the normalization condition which ensures that a Gaussian PDF integrates to one. To derive the variance, we again use the normalization condition. For a Gaussian PDF, this property implies that

$$\int_{-\infty}^{\infty} e^{-\frac{(\xi-m)^2}{2\sigma^2}} d\xi = \sqrt{2\pi}\sigma.$$

Differentiating both sides of this equation with respect to σ, we get

$$\int_{-\infty}^{\infty} \frac{(\xi-m)^2}{\sigma^3} e^{-\frac{(\xi-m)^2}{2\sigma^2}} d\xi = \sqrt{2\pi}.$$

Rearranging the terms yields

$$\int_{-\infty}^{\infty} \frac{(\xi-m)^2}{\sqrt{2\pi}\sigma} e^{-\frac{(\xi-m)^2}{2\sigma^2}} d\xi = \sigma^2.$$

Hence, $\mathrm{Var}(X) = \mathrm{E}\left[(X-m)^2\right] = \sigma^2$. Of course, the variance can also be obtained by more conventional methods.

Example 85. *Suppose that R is a Rayleigh random variable with parameter σ^2. We wish to compute its mean and variance.*

Recall that R is a nonnegative random variable with PDF

$$f_R(r) = \frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}} \quad r \geq 0.$$

Using this distribution, we get

$$\begin{aligned} \mathrm{E}[R] &= \int_0^\infty \xi f_R(\xi) d\xi = \int_0^\infty \frac{\xi^2}{\sigma^2} e^{-\frac{\xi^2}{2\sigma^2}} d\xi \\ &= -\xi e^{-\frac{\xi^2}{2\sigma^2}} \Big|_0^\infty + \int_0^\infty e^{-\frac{\xi^2}{2\sigma^2}} d\xi \\ &= \sqrt{2\pi}\sigma \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\zeta^2}{2\sigma^2}} d\zeta = \frac{\sqrt{2\pi}\sigma}{2}. \end{aligned}$$

Integration by parts is key in solving this expectation. Also, notice the judicious use of the fact that the integral of a standard normal random variable over $[0, \infty)$ must be equal to $1/2$. We compute the second moment of R below,

$$\begin{aligned} \mathrm{E}\left[R^2\right] &= \int_0^\infty \xi^2 f_R(\xi) d\xi = \int_0^\infty \frac{\xi^3}{\sigma^2} e^{-\frac{\xi^2}{2\sigma^2}} d\xi \\ &= -\xi^2 e^{-\frac{\xi^2}{2\sigma^2}} \Big|_0^\infty + \int_0^\infty 2\xi e^{-\frac{\xi^2}{2\sigma^2}} d\xi \\ &= -2\sigma^2 e^{-\frac{\xi^2}{2\sigma^2}} \Big|_0^\infty = 2\sigma^2. \end{aligned}$$

The variance of R is therefore equal to

$$\mathrm{Var}[R] = \frac{(4-\pi)}{2}\sigma^2.$$

Typically, σ^2 is employed to denote the variance of a random variable. It may be confusing at first to have a random variable R described in terms of parameter σ^2 whose variance is equal to $(4-\pi)\sigma^2/2$. This situation is an artifact of the following relation. A Rayleigh random variable R can be generated through the expression $R = \sqrt{X^2 + Y^2}$, where X and Y are independent zero-mean Gaussian variables with variance σ^2. Thus, the parameter σ^2 in $f_R(\cdot)$ is a tribute to this popular construction, not a representation of its actual variance.

For nonnegative random variable X, an alternative way to compute $\mathrm{E}[X]$ is described in Proposition 7.

Proposition 7. *Suppose that X is a nonnegative random variable with finite mean, then*

$$\mathrm{E}[X] = \int_0^\infty \Pr(X > x)dx.$$

Proof. We offer a proof for the special case where X is a continuous random variable, although the result remains true in general,

$$\begin{aligned}\int_0^\infty \Pr(X > x)dx &= \int_0^\infty \int_x^\infty f_X(\xi)d\xi dx \\ &= \int_0^\infty \int_0^\xi f_X(\xi)dxd\xi \\ &= \int_0^\infty \xi f_X(\xi)d\xi = \mathrm{E}[X].\end{aligned}$$

Interchanging the order of integration is justified because X is assumed to have finite mean. □

Example 86. *A player throws darts at a circular target hung on a wall. The dartboard has unit radius, and the position of every dart is distributed uniformly over the target. We wish to compute the expected distance from each dart to the center of the dartboard.*

Let R denote the distance from a dart to the center of the target. For $0 \leq r \leq 1$, the probability that R exceeds r is given by

$$\Pr(R > r) = 1 - \Pr(R \leq r) = 1 - \frac{\pi r^2}{\pi} = 1 - r^2.$$

Then, by Proposition 7, the expected value of R is equal to

$$\mathrm{E}[R] = \int_0^1 \left(1 - r^2\right) dr = \left(r - \frac{r^3}{3}\right)\Big|_0^1 = 1 - \frac{1}{3} = \frac{2}{3}.$$

Notice how we were able to compute the answer without deriving an explicit expression for $f_R(\cdot)$.

10.2 Moment Generating Functions

The *moment generating function* of a random variable X is defined by

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

For continuous random variables, the moment generating function becomes

$$M_X(s) = \int_{-\infty}^{\infty} f_X(\xi) e^{s\xi} d\xi.$$

The experienced reader will quickly recognize the definition of $M_X(s)$ as a variant of the *Laplace Transform*, a widely used linear operator. The moment generating function gets its name from the following property. Suppose that $M_X(s)$ exists within an open interval around $s = 0$, then the nth moment of X is given by

$$\begin{aligned}\frac{d^n}{ds^n} M_X(s)\Big|_{s=0} &= \frac{d^n}{ds^n} \mathrm{E}\left[e^{sX}\right]\Big|_{s=0} = \mathrm{E}\left[\frac{d^n}{ds^n} e^{sX}\right]\Big|_{s=0} \\ &= \mathrm{E}\left[X^n e^{sX}\right]\Big|_{s=0} = \mathrm{E}[X^n].\end{aligned}$$

In words, if we differentiate $M_X(s)$ a total of n times and then evaluate the resulting function at zero, we obtain the nth moment of X. In particular, we have $\frac{dM_X}{ds}(0) = \mathrm{E}[X]$ and $\frac{d^2 M_X}{ds^2}(0) = \mathrm{E}[X^2]$.

Example 87 (Exponential Random Variable)**.** *Let X be an exponential random variable with parameter λ. The moment-generating function of X is given by*

$$M_X(s) = \int_0^{\infty} \lambda e^{-\lambda\xi} e^{s\xi} d\xi = \int_0^{\infty} \lambda e^{-(\lambda - s)\xi} d\xi = \frac{\lambda}{\lambda - s}.$$

The mean of X is

$$\mathrm{E}[X] = \frac{dM_X}{ds}(0) = \left.\frac{\lambda}{(\lambda - s)^2}\right|_{s=0} = \frac{1}{\lambda};$$

more generally, the nth moment of X can be computed as

$$\mathrm{E}[X^n] = \frac{d^n M_X}{ds^n}(0) = \left.\frac{n!\lambda}{(\lambda - s)^{n+1}}\right|_{s=0} = \frac{n!}{\lambda^n}.$$

Incidentally, we can deduce from these results that the variance of X is $1/\lambda^2$.

The definition of the moment generating function applies to discrete random variables as well. In fact, for integer-valued random variables, the moment generating function and the ordinary generating function are related through the equation

$$M_X(s) = \sum_{k \in X(\Omega)} e^{sk} p_X(k) = G_X(e^s).$$

Example 88 (Discrete Uniform Random Variable)**.** *Suppose U is a discrete uniform random variable taking value in $U(\Omega) = \{1, 2, \ldots, n\}$. Then, $p_U(k) = 1/n$ for $1 \leq k \leq n$ and*

$$M_U(s) = \sum_{k=1}^{n} \frac{1}{n} e^{sk} = \frac{1}{n} \sum_{k=1}^{n} e^{sk} = \frac{e^s(e^{ns} - 1)}{n(e^s - 1)}.$$

The moment generating function provides an alternate and somewhat intricate way to compute the mean of U,

$$\begin{aligned} \mathrm{E}[U] = \frac{dM_U}{ds}(0) &= \lim_{s \to 0} \frac{ne^{(n+2)s} - (n+1)e^{(n+1)s} + e^s}{n\left(e^s - 1\right)^2} \\ &= \lim_{s \to 0} \frac{n(n+2)e^{(n+1)s} - (n+1)^2 e^{ns} + 1}{2n\left(e^s - 1\right)} \\ &= \lim_{s \to 0} \frac{n(n+1)(n+2)ne^{(n+1)s} - n(n+1)^2 e^{ns}}{2ne^s} = \frac{n+1}{2}. \end{aligned}$$

Notice the double application of l'Hôpital's rule *to evaluate the derivative of $M_U(s)$ at zero. This may be deemed a more contrived method to derive the expected value of a discrete uniform random variables, but it does not rely on prior knowledge of special sums. Through similar steps, one can derive the second moment of U, which is equal to*

$$\mathrm{E}\left[U^2\right] = \frac{(n+1)(2n+1)}{6}.$$

From these two results, we can show that the variance of U is $(n^2 - 1)/12$.

The simple form of the moment generating function of a standard normal random variable points to its importance in many situations. The exponential function is analytic and possesses many representations.

Example 89 (Gaussian Random Variable)**.** *Let X be a standard normal random variable whose PDF is given by*

$$f_X(\xi) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}}.$$

The moment generating function of X is equal to

$$\begin{aligned} M_X(s) = \mathrm{E}\left[e^{sX}\right] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2}{2}} e^{s\xi} d\xi \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2 + 2s\xi}{2}} d\xi = e^{\frac{s^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{\xi^2 - 2s\xi + s^2}{2}} d\xi \\ &= e^{\frac{s^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(\xi - s)^2}{2}} d\xi = e^{\frac{s^2}{2}}. \end{aligned}$$

The last equality follows from the normalization condition and the fact that the integrand is a Gaussian PDF.

Let $M_X(s)$ be the moment generating function associated with a random variable X, and consider the random variable $Y = aX + b$ where a and b are constant. The moment generating function of Y can be obtained as follows,

$$M_Y(s) = \mathrm{E}\left[e^{sX}\right] = \mathrm{E}\left[e^{s(aX+b)}\right] = e^{sb}\mathrm{E}\left[e^{saX}\right] = e^{sb}M_X(as).$$

Thus, if Y is an affine function of X then $M_Y(s) = e^{sb}M_X(as)$.

Example 90. *We can use this property to identify the moment generating function of a Gaussian random variable with parameters m and σ^2. Recall that an affine function of a Gaussian random variable is also Gaussian. Let $Y = \sigma X + m$, then the moment generating function of Y becomes*

$$M_Y(s) = \mathrm{E}\left[e^{sY}\right] = \mathrm{E}\left[e^{s(\sigma X+m)}\right] = e^{sm}\mathrm{E}\left[e^{s\sigma X}\right] = e^{sm+\frac{s^2\sigma^2}{2}}.$$

From this moment generating function, we get

$$\begin{aligned}
\mathrm{E}[Y] &= \frac{dM_Y}{ds}(0) = \left[\left(m + s\sigma^2\right)e^{sm+\frac{s^2\sigma^2}{2}}\right]\Big|_{s=0} = m \\
\mathrm{E}\left[Y^2\right] &= \frac{d^2M_Y}{ds^2}(0) = \left[\sigma^2 e^{sm+\frac{s^2\sigma^2}{2}} + \left(m + s\sigma^2\right)^2 e^{sm+\frac{s^2\sigma^2}{2}}\right]\Big|_{s=0} \\
&= \sigma^2 + m^2.
\end{aligned}$$

The mean of Y is m and its variance is equal to σ^2, as anticipated.

10.3 Important Inequalities

There are many situations for which computing the exact value of a probability is impossible or impractical. In such cases, it may be acceptable to provide bounds on the value of an elusive probability. The expectation is most important in finding pertinent bounds.

As we will see, many upper bounds rely on the concept of dominating functions. Suppose that $g(x)$ and $h(x)$ are two nonnegative function such that $g(x) \leq h(x)$ for all $x \in \mathbb{R}$. Then, for any continuous random variable X, the

following inequality holds

$$\begin{aligned}\mathrm{E}[g(X)] &= \int_{-\infty}^{\infty} g(x) f_X(x) dx \\ &\leq \int_{-\infty}^{\infty} h(x) f_X(x) dx = \mathrm{E}[h(X)].\end{aligned}$$

This is illustrated in Figure 10.1. In words, the weighted integral of $g(\cdot)$ is dominated by the weighted integral of $h(\cdot)$, where $f_X(\cdot)$ acts as the weighting function. This notion is instrumental in understanding bounding techniques.

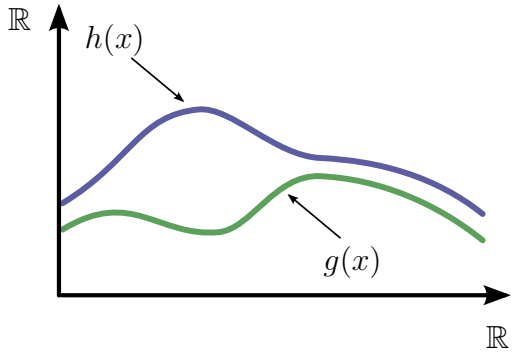

Figure 10.1: If $g(x)$ and $h(x)$ are two nonnegative functions such that $g(x) \leq h(x)$ for all $x \in \mathbb{R}$, then $\mathrm{E}[g(X)]$ is less than or equal to $\mathrm{E}[h(X)]$.

10.3.1 The Markov Inequality

We begin our exposition of classical upper bounds with a result known as the *Markov inequality*. Recall that, for admissible set $S \subset \mathbb{R}$, we have

$$\Pr(X \in S) = \mathrm{E}\left[\mathbf{1}_S(X)\right].$$

Thus, to obtain a bound on $\Pr(X \in S)$, it suffices to find a function that dominates $\mathbf{1}_S(\cdot)$ and for which we can compute the expectation.

Suppose that we wish to bound $\Pr(X \geq a)$ where X is a nonnegative random variable. In this case, we can select $S = [a, \infty)$ and function $h(x) = x/a$. For any $x \geq 0$, we have $h(x) \geq \mathbf{1}_S(x)$, as illustrated in Figure 10.2. It follows that

$$\Pr(X \geq a) = \mathrm{E}\left[\mathbf{1}_S(X)\right] \leq \frac{\mathrm{E}[X]}{a}.$$

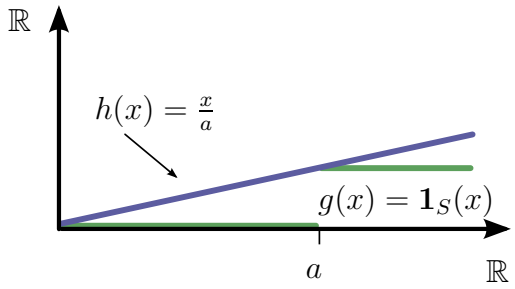

Figure 10.2: Suppose that we wish to find a bound for $\Pr(X \leq a)$. We define set $S = [a, \infty)$ and function $g(x) = \mathbf{1}_S(x)$. Using dominating function $h(x) = x/a$, we conclude that $\Pr(X \geq a) \leq a^{-1}\mathrm{E}[X]$ for any nonnegative random variable X.

10.3.2 The Chebyshev Inequality

The *Chebyshev inequality* provides an extension to this methodology to various dominating functions. This yields a number of bounds that become useful in a myriad of contexts.

Proposition 8 (Chebyshev Inequality)**.** *Suppose $h(\cdot)$ is a nonnegative function and let S be an admissible set. We denote the infimum of $h(\cdot)$ over S by*

$$i_S = \inf_{x \in S} h(x).$$

The Chebyshev inequality asserts that

$$i_S \Pr(X \in S) \leq \mathrm{E}[h(X)] \tag{10.1}$$

where X is an arbitrary random variable.

Proof. This is a remarkably powerful result and it can be shown in a few steps. The definition of i_S and the fact that $h(\cdot)$ is nonnegative imply that

$$i_S \mathbf{1}_S(x) \leq h(x)\mathbf{1}_S(x) \leq h(x)$$

for any $x \in \mathbb{R}$. Moreover, for any such x and distribution $f_X(\cdot)$, we can write $i_S \mathbf{1}_S(x) f_X(x) \leq h(x) f(x)$, which in turn yields

$$\begin{aligned} i_S \Pr(X \in S) = \mathrm{E}\left[i_s \mathbf{1}_S(X)\right] &= \int_{\mathbb{R}} i_S \mathbf{1}_S(\xi) f_X(\xi) d\xi \\ &\leq \int_{\mathbb{R}} h(\xi) f_X(\xi) d\xi = \mathrm{E}[h(X)]. \end{aligned}$$

When $i_S > 0$, this provides the upper bound $\Pr(X \in S) \leq i_S^{-1}\mathrm{E}[h(X)]$. □

Although the proof assumes a continuous random variable, we emphasize that the Chebyshev inequality applies to both discrete and continuous random variables alike. The interested reader can rework the proof using the discrete setting and a generic PMF. We provide special instances of the Chebyshev inequality below.

Example 91. *Consider the nonnegative function $h(x) = x^2$ and let $S = \{x|x^2 \geq b^2\}$ where b is a positive constant. We wish to find a bound on the probability that $|X|$ exceeds b. Using the Chebyshev inequality, we have $i_S = \inf_{x\in S} x^2 = b^2$ and, consequently, we get*

$$b^2 \Pr(X \in S) \leq \mathrm{E}\left[X^2\right].$$

Constant b being a positive real number, we can rewrite this equation as

$$\Pr(|X| \geq b) = \Pr(X \in S) \leq \frac{\mathrm{E}\left[X^2\right]}{b^2}.$$

Example 92 (The Cantelli Inequality)**.** *Suppose that X is a random variable with mean m and variance σ^2. We wish to show that*

$$\Pr(X - m \geq a) \leq \frac{\sigma^2}{a^2 + \sigma^2},$$

where $a \geq 0$.

This equation is slightly more involved and requires a small optimization in addition to the Chebyshev inequality. Define $Y = X - m$ and note that, by construction, we have $\mathrm{E}[Y] = 0$. Consider the probability $\Pr(Y \geq a)$ where $a > 0$, and let $S = \{y|y \geq a\}$. Also, define the nonnegative function $h(y) = (y+b)^2$, where $b > 0$. Following the steps of the Chebyshev inequality, we write the infimum of $h(y)$ over S as

$$i_S = \inf_{y\in S}(y + b)^2 = (a + b)^2.$$

Then, applying the Chebyshev inequality, we obtain

$$\Pr(Y \geq a) \leq \frac{\mathrm{E}\left[(Y + b)^2\right]}{(a + b)^2} = \frac{\sigma^2 + b^2}{(a + b)^2}. \tag{10.2}$$

This inequality holds for any $b > 0$. To produce a better upper bound, we minimize the right-hand side of (10.2) *over all possible values of b. Differentiating this expression and setting the derivative equal to zero yields*

$$\frac{2b}{(a+b)^2} = \frac{2\left(\sigma^2 + b^2\right)}{(a+b)^3}$$

or, equivalently, $b = \sigma^2/a$. A second derivative test reveals that this is indeed a minimum. Collecting these results, we obtain

$$\Pr(Y \geq a) \leq \frac{\sigma^2 + b^2}{(a+b)^2} = \frac{\sigma^2}{a^2 + \sigma^2}.$$

Substituting $Y = X - m$ leads to the desired result.

In some circumstances, a Chebyshev inequality can be tight.

Example 93. *Let a and b be two constants such that $0 < b \leq a$. Consider the function $h(x) = x^2$ along with the set $S = \{x | x^2 \geq a^2\}$. Furthermore, let X be a discrete random variable with PMF*

$$p_X(x) = \begin{cases} 1 - \frac{b^2}{a^2}, & x = 0 \\ \frac{b^2}{a^2}, & x = a \\ 0, & \textit{otherwise.} \end{cases}$$

For this random variable, we have $\Pr(X \in S) = b^2/a^2$. By inspection, we also gather that the second moment of X is equal to $\mathrm{E}\left[X^2\right] = b^2$. Applying the Chebyshev inequality, we get $i_S = \inf_{x \in S} h(x) = a^2$ and therefore

$$\Pr(X \in S) \leq i_S^{-1} \mathrm{E}\left[h(X)\right] = \frac{b^2}{a^2}.$$

Thus, in this particular example, the inequality is met with equality.

10.3.3 The Chernoff Bound

The *Chernoff bound* is yet another upper bound that can be constructed from the Chebyshev inequality. Still, because of its central role in many application domains, it deserves its own section. Suppose that we want to find a bound on the probability $\Pr(X \geq a)$. We can apply the Chebyshev inequality using the

nonnegative function $h(x) = e^{sx}$, where $s > 0$. For this specific construction, $S = [a, \infty)$ and

$$i_S = \inf_{x \in S} e^{sx} = e^{sa}.$$

It follows that

$$\Pr(X \geq a) \leq e^{-sa} \mathrm{E}[e^{sX}] = e^{-sa} M_X(s).$$

Because this inequality holds for any $s > 0$, we can optimize the upper bound over all possible values of s, thereby picking the best one,

$$\Pr(X \geq a) \leq \inf_{s>0} e^{-sa} M_X(s). \tag{10.3}$$

This inequality is called the Chernoff bound. It is sometimes expressed in terms of the log-moment generating function $\Lambda(s) = \log M_X(s)$. In this latter case, (10.3) translates into

$$\log \Pr(X \geq a) \leq - \sup_{s>0} \{sa - \Lambda(s)\}. \tag{10.4}$$

The right-hand side of (10.4) is called the *Legendre transformation* of $\Lambda(s)$. Figure 10.3 plots $e^{s(x-a)}$ for various values of $s > 0$. It should be noted that all these functions dominate $\mathbf{1}_{[a,\infty)}(x)$, and therefore they each provide a different bound on $\Pr(X \geq a)$. It is natural to select the function that provides the best bound. Yet, in general, this optimal $e^{s(x-a)}$ may depend on the distribution of X and the value of a, which explains why (10.3) involves a search over all possible values of s.

10.3.4 Jensen's Inequality

Some inequalities can be derived based on the properties of a single function. The *Jensen inequality* is one such example. Suppose that function $g(\cdot)$ is convex and twice differentiable, with

$$\frac{d^2 g}{dx^2}(x) \geq 0$$

for all $x \in \mathbb{R}$. From the fundamental theorem of calculus, we have

$$g(x) = g(a) + \int_a^x \frac{dg}{dx}(\xi) d\xi.$$

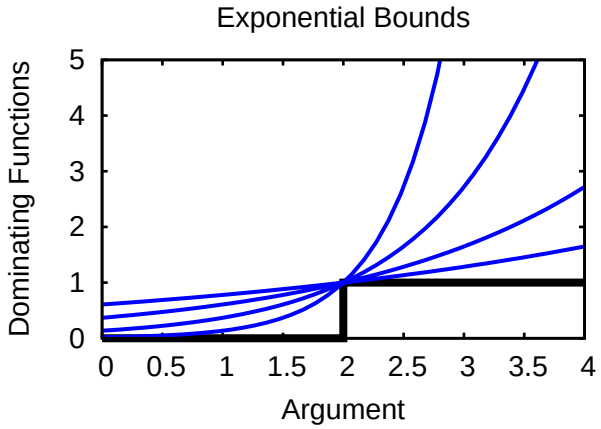

Figure 10.3: This figure illustrates how exponential functions can be employed to provide bounds on $\Pr(X > a)$. Optimizing over all admissible exponential functions, $e^{s(x-a)}$ where $s > 0$, leads to the celebrated Chernoff bound.

Futhermore, because the second derivative of $g(\cdot)$ is a non-negative function, we gather that $\frac{dg}{dx}(\cdot)$ is a monotone increasing function. As such, for any value of a, we have

$$\begin{aligned} g(x) &= g(a) + \int_a^x \frac{dg}{dx}(\xi)d\xi \\ &\geq g(a) + \int_a^x \frac{dg}{dx}(a)d\xi = g(a) + (x-a)\frac{dg}{dx}(a). \end{aligned}$$

For random variable X, we then have

$$g(X) \geq g(a) + (X-a)\frac{dg}{dx}(a).$$

Choosing $a = \mathrm{E}[X]$ and taking expectations on both sides, we obtain

$$\mathrm{E}[g(X)] \geq g(\mathrm{E}[X]) + (\mathrm{E}[X] - \mathrm{E}[X])\frac{dg}{dx}(\mathrm{E}[X]) = g(\mathrm{E}[X]).$$

That is, $\mathrm{E}[g(X)] \geq g(\mathrm{E}[X])$, provided that these two expectations exist. The Jensen inequality actually holds for convex functions that are not twice differentiable, but the proof is much harder in the general setting.

Further Reading

1. Ross, S., *A First Course in Probability*, 7th edition, Pearson Prentice Hall, 2006: Sections 5.2, 7.7, 8.2.

2. Bertsekas, D. P., and Tsitsiklis, J. N., *Introduction to Probability*, Athena Scientific, 2002: Section 3.1, 4.1, 7.1.

3. Gubner, J. A., *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge, 2006: Sections 4.2–4.3.

4. Miller, S. L., and Childers, D. G., *Probability and Random Processes with Applications to Signal Processing and Communications*, 2004: Section 5.2.