Example 4.5.6 *Least-squares weights.* The simplest objective to consider in (4.70) is the quadratic

$$H(w_1, \ldots, w_n) = \tfrac{1}{2}w^\top w,$$

with w the vector of weights $(w_1, \ldots, w_n)^\top$. We will show that the estimator $w^\top Y = \sum w_i Y_i$ produced by these weights is identical to the control variate estimator $\bar{Y}(\hat{b}_n)$ defined by (4.16).

Define an $n \times (d+1)$ matrix A whose ith row is $(1, X_i^\top - \mu_X^\top)$. (Here we assume that $X_1, \ldots, X_n$ do not contain an entry identically equal to 1.) Constraints (4.67)–(4.69) can be expressed as $w^\top A = (1, \mathbf{0})$, where $\mathbf{0}$ is a row vector of d zeros. The Lagrangian becomes

$$\tfrac{1}{2}w^\top w + w^\top A\lambda$$

with $\lambda \in \Re^d$. The first-order conditions are $w = -A\lambda$. From the constraint we get

$$\begin{aligned}(1, \mathbf{0}) = w^\top A = -\lambda^\top A^\top A &\Rightarrow -\lambda^\top = (1, \mathbf{0})(A^\top A)^{-1} \\ &\Rightarrow w^\top = (1, \mathbf{0})(A^\top A)^{-1}A^\top,\end{aligned}$$

assuming the matrix A has full rank. The weighted Monte Carlo estimator of the expectation of the Y_i is thus

$$w^\top Y = (1, \mathbf{0})(A^\top A)^{-1}A^\top Y, \quad Y = (Y_1, \ldots, Y_n)^\top. \tag{4.72}$$

The control variate estimator is the first entry of the vector $\beta \in \Re^{d+1}$ that solves

$$\min_\beta \tfrac{1}{2}(Y - A\beta)^\top(Y - A\beta);$$

i.e., it is the value fitted at $(1, \mu_X)$ in a regression of the Y_i against the rows of A. From the first-order conditions $(Y - A\beta)^\top A = 0$, we find that the optimal β is $(A^\top A)^{-1}A^\top Y$. The control variate estimator is therefore

$$(1, \mathbf{0})\beta = (1, \mathbf{0})(A^\top A)^{-1}A^\top Y,$$

which coincides with (4.72). When written out explicitly, the weights in (4.72) take precisely the form displayed in (4.20), where we first noted the interpretation of a control variate estimator as a weighted Monte Carlo estimator. □

This link between the general strategy in (4.68) and(4.70) for constructing "moment-matched" estimators and the more familiar method of control variates suggests that (4.68) provides at best a small refinement of the control variate estimator. As the sample size n increases, the refinement typically vanishes and using knowledge of μ_X as a constraint in (4.67) becomes equivalent to using it in a control variate estimator. A precise result to this effect is proved in Glasserman and Yu [147]. This argues in favor of using control variate estimators rather than (4.68), because they are easier to implement and because more is known about their sampling properties.

4.6 Importance Sampling

4.6.1 Principles and First Examples

Importance sampling attempts to reduce variance by changing the probability measure from which paths are generated. Changing measures is a standard tool in financial mathematics; we encountered it in our discussion of pricing principles in Section 1.2.2 and several places in Chapter 3 in the guise of changing numeraire. Appendix B.4 reviews some of the underlying mathematical theory. When we switch from, say, the objective probability measure to the risk-neutral measure, our goal is usually to obtain a more convenient representation of an expected value. In importance sampling, we change measures to try to give more weight to "important" outcomes thereby increasing sampling efficiency.

To make this idea concrete, consider the problem of estimating

$$\alpha = \mathsf{E}[h(X)] = \int h(x)f(x)\,dx$$

where X is a random element of $\Re^d$ with probability density f, and h is a function from $\Re^d$ to $\Re$. The ordinary Monte Carlo estimator is

$$\hat{\alpha} = \hat{\alpha}(n) = \frac{1}{n}\sum_{i=1}^n h(X_i)$$

with $X_1, \ldots, X_n$ independent draws from f. Let g be any other probability density on $\Re^d$ satisfying

$$f(x) > 0 \Rightarrow g(x) > 0 \tag{4.73}$$

for all $x \in \Re^d$. Then we can alternatively represent α as

$$\alpha = \int h(x)\frac{f(x)}{g(x)}g(x)\,dx.$$

This integral can be interpreted as an expectation with respect to the density g; we may therefore write

$$\alpha = \tilde{\mathsf{E}}\left[h(X)\frac{f(X)}{g(X)}\right], \tag{4.74}$$

$\tilde{\mathsf{E}}$ here indicating that the expectation is taken with X distributed according to g. If $X_1, \ldots, X_n$ are now independent draws from g, the importance sampling estimator associated with g is

$$\hat{\alpha}_g = \hat{\alpha}_g(n) = \frac{1}{n}\sum_{i=1}^n h(X_i)\frac{f(X_i)}{g(X_i)}. \tag{4.75}$$

The weight $f(X_i)/g(X_i)$ is the *likelihood ratio* or *Radon-Nikodym derivative* evaluated at X_i.

It follows from (4.74) that $\tilde{\mathsf{E}}[\hat{\alpha}_g] = \alpha$ and thus that $\hat{\alpha}_g$ is an unbiased estimator of α. To compare variances with and without importance sampling it therefore suffices to compare second moments. With importance sampling, we have

$$\tilde{\mathsf{E}}\left[\left(h(X)\frac{f(X)}{g(X)}\right)^2\right] = \mathsf{E}\left[h(X)^2\frac{f(X)}{g(X)}\right].$$

This could be larger or smaller than the second moment $\mathsf{E}[h(X)^2]$ without importance sampling; indeed, depending on the choice of g it might even be *infinitely* larger or smaller. Successful importance sampling lies in the art of selecting an effective importance sampling density g.

Consider the special case in which h is nonnegative. The product $h(x)f(x)$ is then also nonnegative and may be normalized to a probability density. Suppose g is this density. Then

$$g(x) \propto h(x)f(x), \tag{4.76}$$

and $h(X_i)f(X_i)/g(X_i)$ equals the constant of proportionality in (4.76) regardless of the value of X_i; thus, the importance sampling estimator $\hat{\alpha}_g$ in (4.75) provides a *zero-variance* estimator in this case. Of course, this is useless in practice: to normalize $h \cdot f$ we need to divide it by its integral, which is α; the zero-variance estimator is just α itself.

Nevertheless, this optimal choice of g does provide some useful guidance: in designing an effective importance sampling strategy, we should try to sample in proportion to the product of h and f. In option pricing applications, h is typically a discounted payoff and f is the risk-neutral density of a discrete path of underlying assets. In this case, the "importance" of a path is measured by the product of its discounted payoff and its probability density.

If h is the indicator function of a set, then the optimal importance sampling density is the original density conditioned on the set. In more detail, suppose $h(x) = \mathbf{1}\{x \in A\}$ for some $A \subset \Re^d$. Then $\alpha = P(X \in A)$ and the zero-variance importance sampling density $h(x)f(x)/\alpha$ is precisely the conditional density of X given $X \in A$ (assuming $\alpha > 0$). Thus, in applying importance sampling to estimate a probability, we should look for an importance sampling density that approximates the conditional density. This means choosing g to make the event $\{X \in A\}$ more likely, especially if A is a rare set under f.

Likelihood Ratios

In our discussion thus far we have assumed, for simplicity, that X is $\Re^d$-valued, but the ideas extend to X taking values in more general sets. Also, we have assumed that X has a density f, but the same observations apply if f is a probability mass function (or, more generally, a density with respect to some reference measure on $\Re^d$, possibly different from Lebesgue measure).

For option pricing applications, it is natural to think of X as a discrete path of underlying assets. The density of a path (if one exists) is ordinarily not specified directly, but rather built from more primitive elements. Consider, for example, a discrete path $S(t_i)$, $i = 0, 1, \ldots, m$, of underlying assets or state variables, and suppose that this process is Markov. Suppose the conditional distribution of $S(t_i)$ given $S(t_{i-1}) = x$ has density $f_i(x, \cdot)$. Consider a change of measure under which the transition densities f_i are replaced with transition densities g_i. The likelihood ratio for this change of measure is

$$\prod_{i=1}^{m} \frac{f_i(S(t_{i-1}), S(t_i))}{g_i(S(t_{i-1}), S(t_i))}.$$

More precisely, if E denotes expectation under the original measure and $\tilde{\mathsf{E}}$ denotes expectation under the new measure, then

$$\mathsf{E}[h(S(t_1), \ldots, S(t_m))] = \tilde{\mathsf{E}}\left[h(S(t_1), \ldots, S(t_m)) \prod_{i=1}^{m} \frac{f_i(S(t_{i-1}), S(t_i))}{g_i(S(t_{i-1}), S(t_i))}\right], \tag{4.77}$$

for all functions h for which the expectation on the left exists and is finite.

Here we have implicitly assumed that $S(t_0)$ is a constant. More generally, we could allow it to have density f_0 under the original measure and density g_0 under the new measure. This would result in an additional factor of $f_0(S(t_0))/g_0(S(t_0))$ in the likelihood ratio.

We often simulate a path $S(t_0), \ldots, S(t_m)$ through a recursion of the form

$$S(t_{i+1}) = G(S(t_i), X_{i+1}), \tag{4.78}$$

driven by i.i.d. random vectors $X_1, X_2, \ldots, X_m$. Many of the examples considered in Chapter 3 can be put in this form. The X_i will often be normally distributed, but for now let us simply assume they have common density f. If we apply a change of measure that preserves the independence of the X_i but changes their common density to g, then the corresponding likelihood ratio is

$$\prod_{i=1}^{m} \frac{f(X_i)}{g(X_i)}.$$

This means that

$$\mathsf{E}[h(S(t_1), \ldots, S(t_m))] = \tilde{\mathsf{E}}\left[h(S(t_1), \ldots, S(t_m)) \prod_{i=1}^{m} \frac{f(X_i)}{g(X_i)}\right], \tag{4.79}$$

where, again, E and $\tilde{\mathsf{E}}$ denote expectation under the original and new measures, respectively, and the expectation on the left is assumed finite. Equation (4.79) relies on the fact that $S(t_1), \ldots, S(t_m)$ are functions of $X_1, \ldots, X_m$.

Random Horizon

Identities (4.77) and (4.79) extend from a fixed number of steps m to a random number of steps, provided the random horizon is a stopping time. We demonstrate this in the case of i.i.d. inputs, as in (4.79). For each $n = 1, 2, \ldots$ let h_n be a function of n arguments and suppose we want to estimate

$$\mathsf{E}[h_N(S(t_1), \ldots, S(t_N))], \tag{4.80}$$

with N a random variable taking values in $\{1, 2, \ldots\}$. For example, in the case of a barrier option with barrier b, we might define N to be the index of the smallest t_i for which $S(t_i) > b$, taking $N = m$ if all $S(t_0), \ldots, S(t_m)$ lie below the barrier. We could then express the discounted payoff of an up-and-out put as $h_N(S(t_1), \ldots, S(t_N))$ with

$$h_n(S(t_1), \ldots, S(t_n)) = \begin{cases} e^{-rt_m}(K - S(t_m))^+, & n = m; \\ 0, & n = 0, 1, \ldots, m-1. \end{cases}$$

The option price then has the form (4.80).

Suppose that (4.78) holds and, as before, E denotes expectation when the X_i are i.i.d. with density f and $\tilde{\mathsf{E}}$ denotes expectation when they are i.i.d. with density g. For concreteness, suppose that $S(t_0)$ is fixed under both measures. Let N be a stopping time for the sequence $X_1, X_2, \ldots$; for example, N could be a stopping time for $S(t_1), S(t_2), \ldots$ as in the barrier option example. Then

$$\begin{aligned} &\mathsf{E}[h_N(S(t_1), \ldots, S(t_N))\mathbf{1}\{N < \infty\}] \\ &= \tilde{\mathsf{E}}\left[h_N(S(t_1), \ldots, S(t_N)) \prod_{i=1}^{N} \frac{f(X_i)}{g(X_i)} \mathbf{1}\{N < \infty\}\right], \end{aligned}$$

provided the expectation on the left is finite. This identity (sometimes called Wald's identity or the fundamental identity of sequential analysis — see, e.g., Asmussen [20]) is established as follows:

$$\begin{aligned} &\mathsf{E}[h_N(S(t_1), \ldots, S(t_N))\mathbf{1}\{N < \infty\}] \\ &= \sum_{n=1}^{\infty} \mathsf{E}[h_n(S(t_1), \ldots, S(t_n))\mathbf{1}\{N = n\}] \\ &= \sum_{n=1}^{\infty} \tilde{\mathsf{E}}\left[h_n(S(t_1), \ldots, S(t_n))\mathbf{1}\{N = n\} \prod_{i=1}^{n} \frac{f(X_i)}{g(X_i)}\right] \\ &= \tilde{\mathsf{E}}\left[h_N(S(t_1), \ldots, S(t_N)) \prod_{i=1}^{N} \frac{f(X_i)}{g(X_i)} \mathbf{1}\{N < \infty\}\right]. \end{aligned} \tag{4.81}$$

The second equality uses the stopping time property: because N is a stopping time the event $\{N = n\}$ is determined by $X_1, \ldots, X_n$ and this allows us to apply (4.79) to each term in the infinite sum. It is entirely possible for the event $\{N < \infty\}$ to have probability 1 under one of the measures but not the other; we will see an example of this in Example 4.6.3.

Long Horizon

We continue to consider two probability measures under which random vectors $X_1, X_2, \ldots$ are i.i.d., P giving the X_i density f, $\tilde{P}$ giving them density g.

It should be noted that even if f and g are mutually absolutely continuous, the probability measures P and $\tilde{P}$ will not be. Rather, absolute continuity holds for the restrictions of these measures to events defined by a finite initial segment of the infinite sequence. For $A \subseteq \Re^d$, the event

$$\left\{\lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\{X_i \in A\} = \int_A f(x)\, dx\right\}$$

has probability 1 under P; but some such event must have probability 0 under $\tilde{P}$ unless f and g are equal almost everywhere. In short, the strong law of large numbers forces P and $\tilde{P}$ to disagree about which events have probability 0.

This collapse of absolute continuity in the limit is reflected in the somewhat pathological behavior of the likelihood ratio as the number of terms grows, through an argument from Glynn and Iglehart [157]. Suppose that

$$\tilde{\mathsf{E}}\left[|\log(f(X_1)/g(X_1))|\right] < \infty;$$

then the strong law of large numbers implies that

$$\frac{1}{m} \sum_{i=1}^{m} \log(f(X_i)/g(X_i)) \to \tilde{\mathsf{E}}\left[\log(f(X_1)/g(X_1))\right] \equiv c \tag{4.82}$$

with probability 1 under $\tilde{P}$. By Jensen's inequality,

$$c \le \log \tilde{\mathsf{E}}[f(X_1)/g(X_1)] = \log \int \frac{f(x)}{g(x)} g(x)\, dx = 0,$$

with strict inequality unless $\tilde{P}(f(X_1) = g(X_1)) = 1$ because log is strictly concave. But if $c < 0$, (4.82) implies

$$\sum_{i=1}^{m} \log(f(X_i)/g(X_i)) \to -\infty;$$

exponentiating, we find that

$$\prod_{i=1}^{m} \frac{f(X_i)}{g(X_i)} \to 0$$

with $\tilde{P}$-probability 1. Thus, the likelihood ratio converges to 0 though its expectation equals 1 for all m. This indicates that the likelihood ratio becomes highly skewed, taking increasingly large values with small but non-negligible probability. This in turn can result in a large increase in variance if the change of measure is not chosen carefully.

Output Analysis

An importance sampling estimator does not introduce dependence between replications and is just an average of i.i.d. replications. We can therefore supplement an importance sampling estimator with a large-sample confidence interval in the usual way by calculating the sample standard deviation across replications and using it in (A.6). Because likelihood ratios are often highly skewed, the sample standard deviation will often underestimate the true standard deviation, and a very large sample size may be required for confidence intervals based on the central limit theorem to provide reasonable coverage. These features should be kept in mind in comparing importance sampling estimators based on estimates of their standard errors.

Examples

Example 4.6.1 *Normal distribution: change of mean.* Let f be the univariate standard normal density and g the univariate normal density with mean μ and variance 1. Then simple algebra shows that

$$\prod_{i=1}^{m} \frac{f(Z_i)}{g(Z_i)} = \exp\left(-\mu \sum_{i=1}^{m} Z_i + \frac{m}{2}\mu^2\right).$$

A bit more generally, if we let g_i have mean μ_i, then

$$\prod_{i=1}^{m} \frac{f(Z_i)}{g_i(Z_i)} = \exp\left(-\sum_{i=1}^{m} \mu_i Z_i + \tfrac{1}{2}\sum_{i=1}^{m} \mu_i^2\right). \tag{4.83}$$

If we simulate Brownian motion on a grid $0 = t_0 < t_1 < \cdots < t_m$ by setting

$$W(t_n) = \sum_{i=1}^{n} \sqrt{t_i - t_{i-1}} Z_i,$$

then (4.83) is the likelihood ratio for a change of measure that adds mean $\mu_i\sqrt{t_i - t_{i-1}}$ to the Brownian increment over $[t_{i-1}, t_i]$. □

Example 4.6.2 *Exponential change of measure.* The previous example is a special case of a more general class of convenient measure transformations. For a cumulative distribution function F on $\Re$, define

$$\psi(\theta) = \log \int_{-\infty}^{\infty} e^{\theta x}\, dF(x).$$

This is the *cumulant generating function* of F, the logarithm of the moment generating function of F. Let $\Theta = \{\theta : \psi(\theta) < \infty\}$ and suppose that Θ is nonempty. For each $\theta \in \Theta$, set

$$F_\theta(x) = \int_{-\infty}^{x} e^{\theta u - \psi(\theta)}\, dF(u);$$

each F_θ is a probability distribution, and $\{F_\theta, \theta \in \Theta\}$ form an *exponential family* of distributions. The transformation from F to F_θ is called exponential tilting, exponential twisting, or simply an exponential change of measure. If F has a density f, then F_θ has density

$$f_\theta(x) = e^{\theta x - \psi(\theta)} f(x).$$

Suppose that $X_1, \ldots, X_n$ are initially i.i.d. with distribution $F = F_0$ and that we apply a change of measure under which they become i.i.d. with distribution F_θ. The likelihood ratio for this transformation is

$$\prod_{i=1}^{n} \frac{dF_0(X_i)}{dF_\theta(X_i)} = \exp\left(-\theta \sum_{i=1}^{n} X_i + n\psi(\theta)\right). \tag{4.84}$$

The standard normal distribution has $\psi(\theta) = \theta^2/2$, from which we see that this indeed generalizes Example 4.6.1. A key feature of exponential twisting is that the likelihood ratio — which is in principle a function of all $X_1, \ldots, X_n$ — reduces to a function of the sum of the X_i. In statistical terminology, the sum of the X_i is a *sufficient statistic* for θ.

The cumulant generating function ψ records important information about the distributions F_θ. For example, $\psi'(\theta)$ is the mean of F_θ. To see this, let E_θ denote expectation with respect to F_θ and note that $\psi(\theta) = \log \mathsf{E}_0[\exp(\theta X)]$. Differentiation yields

$$\psi'(\theta) = \frac{\mathsf{E}_0[Xe^{\theta X}]}{\mathsf{E}_0[e^{\theta X}]} = \mathsf{E}_0[Xe^{\theta X - \psi(\theta)}] = \mathsf{E}_\theta[X].$$

A similar calculation shows that $\psi''(\theta)$ is the variance of F_θ. The function ψ passes through the origin; Hölder's inequality shows that it is convex, so that $\psi''(\theta)$ is indeed positive. For further theoretical background on exponential families see, e.g., Barndorff-Nielson [35].

We conclude with some examples of exponential families. The normal distributions $N(\theta, \theta\sigma^2)$ form an exponential family in θ for all $\sigma > 0$. The gamma densities

$$\frac{1}{\Gamma(a)\theta^a} x^{a-1} e^{-x/\theta}, \quad x \geq 0,$$

form an exponential family in θ for each value of the shape parameter $a > 0$. With $a = 1$, this is the family of exponential distributions with mean θ. The Poisson distributions

$$e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, \ldots,$$

form an exponential family in $\theta = \log \lambda$. The binomial distributions

$$\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}, \quad k = 0, 1, \ldots, n,$$

form an exponential family in $\theta = \log(p/(1-p))$. □

Example 4.6.3 *Ruin probabilities.* A classic application of importance sampling arises in estimating ruin probabilities in the theory of insurance risk. Consider an insurance firm earning premiums at a constant rate p per unit of time and paying claims that arrive at the jumps of a Poisson process with rate λ. Letting $N(t)$ denote the number of claims arriving in $[0, t]$ and Y_i the size of the ith claim, $i = 1, 2, \ldots$, the net payout of the firm over $[0, t]$ is given by

$$\sum_{i=1}^{N(t)} Y_i - pt.$$

Suppose the firm has a reserve of x; then ruin occurs if the net payout ever exceeds x. We assume the claims are i.i.d. and independent of the Poisson process. We further assume that $\lambda\mathsf{E}[Y_i] < p$, meaning that premiums flow in at a faster rate than claims are paid out; this ensures that the probability of eventual ruin is less than 1.

If ruin ever occurs, it must occur at the arrival of a claim. It therefore suffices to consider the discrete-time process embedded at the jumps of the Poisson process. Let $\xi_1, \xi_2, \ldots$ be the interarrival times of the Poisson process; these are independent and exponentially distributed with mean $1/\lambda$. The net payout between the $(n-1)$th and nth claims (including the latter but not the former) is $X_n = Y_n - p\xi_n$. The net payout up to the nth claim is given by the random walk $S_n = X_1 + \cdots + X_n$. Ruin occurs at

$$\tau_x = \inf\{n \geq 0 : S_n > x\},$$

with the understanding that $\tau_x = \infty$ if S_n never exceeds x. The probability of eventual ruin is $P(\tau_x < \infty)$. Figure 4.10 illustrates the notation for this example.

The particular form of the increments X_n is not essential to the problem so we generalize the setting. We assume that $X_1, X_2, \ldots$ are i.i.d. with $0 < P(X_i > 0) < 1$ and $\mathsf{E}[X_i] < 0$, but we drop the specific form $Y_n - p\xi_n$. We add the assumption that the cumulant generating function ψ_X of the X_i (cf. Example 4.6.2) is finite in a neighborhood of the origin. This holds in the original model if the cumulant generating function ψ_Y of the claim sizes Y_i is finite in a neighborhood of the origin.

For any point θ in the domain of ψ_X, consider the exponential change of measure with parameter θ and let E_θ denote expectation under this measure. Because τ_x is a stopping time, we may apply (4.81) to write the ruin probability $P(\tau_x < \infty)$ as an E_θ-expectation. Because we have applied an exponential change of measure, the likelihood ratio simplifies as in (4.84); thus, the ruin probability becomes

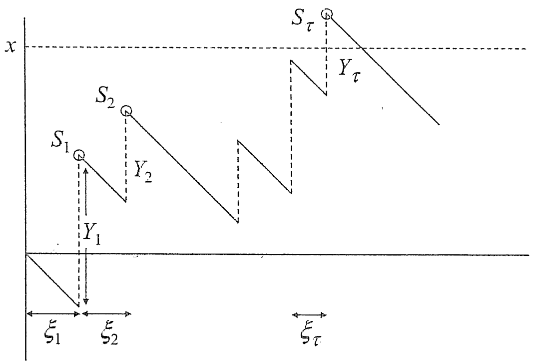

Fig. 4.10. Illustration of claim sizes Y_i, interarrival times ξ_i, and the random walk S_n. Ruin occurs at the arrival of the τth claim.

$$P(\tau_x < \infty) = \mathsf{E}_\theta\left[e^{-\theta S_{\tau_x} + \psi_X(\theta)\tau_x}\mathbf{1}\{\tau_x < \infty\}\right]. \tag{4.85}$$

If $0 < \psi_X'(\theta) < \infty$ (which entails $\theta > 0$ because $\psi(0) = 0$ and $\psi'(0) = \mathsf{E}[X_n] < 0$), then the random walk has positive drift $\mathsf{E}_\theta[X_n] = \psi_X'(\theta)$ under the twisted measure, and this implies $P_\theta(\tau_x < \infty) = 1$. We may therefore omit the indicator inside the expectation on the right. It also follows that we may obtain an unbiased estimator of the ruin probability by simulating the random walk under P_θ until τ_x and returning the estimator $\exp(-\theta S_{\tau_x} + \psi_X(\theta)\tau_x)$. This would not be feasible under the original measure because of the positive probability that $\tau_x = \infty$.

Among all θ for which $\psi_X'(\theta) > 0$, one is particularly effective for simulation and indeed optimal in an asymptotic sense. Suppose there is a $\theta > 0$ at which $\psi_X(\theta) > 0$. There must then be a $\theta_* > 0$ at which $\psi_X(\theta_*) = 0$; convexity of ψ_X implies uniqueness of θ^* and positivity of $\psi_X'(\theta_*)$, as is evident from Figure 4.11. In the insurance risk model with $X_n = Y_n - p\xi_n$, θ_* is the unique positive solution to

$$\psi_Y(\theta) + \log\left(\frac{\lambda}{\lambda + p\theta}\right) = 0,$$

where ψ_Y is the cumulant generating function for the claim-size distribution.

With the parameter θ_*, (4.85) becomes

$$P(\tau_x < \infty) = \mathsf{E}_{\theta_*}\left[e^{-\theta_* S_{\tau_x}}\right] = e^{-\theta_* x}\mathsf{E}_{\theta_*}\left[e^{-\theta_*(S_{\tau_x} - x)}\right].$$

Because the overshoot $S_{\tau_x} - x$ is nonnegative, this implies the simple bound $P(\tau_x < \infty) \leq e^{-\theta_* x}$ on the ruin probability. Under modest additional regularity conditions (for example, if the X_n have a density), $\mathsf{E}_{\theta_*}\left[e^{-\theta_*(S_{\tau_x} - x)}\right]$ converges to a constant c as $x \to \infty$, providing the classical approximation

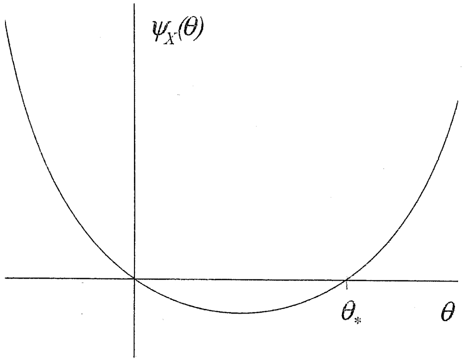

Fig. 4.11. Graph of a cumulant generating function ψ_X. The curve passes through the origin and has negative slope there because $\psi'_X(0) = \mathsf{E}[X] < 0$. At the positive root θ_*, the slope is positive.

$$P(\tau_x < \infty) \sim ce^{-\theta_* x},$$

meaning that the ratio of the two sides converges to 1 as $x \to \infty$. Further details and some of the history of this approximation are discussed in Asmussen [20] and in references given there.

From the perspective of simulation, the significance of θ_* lies in the variance reduction achieved by the associated importance sampling estimator. The unbiased estimator $\exp(-\theta_* S_{\tau_x})$, sampled under P_{θ_*}, has second moment

$$\mathsf{E}_{\theta_*}\left[e^{-2\theta_* S_{\tau_x}}\right] \le e^{-2\theta_* x}.$$

By Jensen's inequality, the second moment of any unbiased estimator must be at least as large as the square of the ruin probability, and we have seen that this probability is $O(e^{-\theta_* x})$. In this sense, the second moment of the importance sampling estimator based on θ_* is asymptotically optimal as $x \to \infty$.

This strategy for developing effective and even asymptotically optimal importance sampling estimators originated in Siegmund's [331] application in sequential analysis. It has been substantially generalized, particularly for queueing and reliability applications, as surveyed in Heidelberger [175]. □

Example 4.6.4 *A knock-in option.* As a further illustration of importance sampling through an exponential change of measure, we apply the method to a down-and-in barrier option. This example is from Boyle et al. [53]. The option is a digital knock-in option with payoff

$$\mathbf{1}\{S(T) > K\} \cdot \mathbf{1}\{\min_{1\le k\le m} S(t_k) < H\},$$

with $0 < t_1 < \cdots < t_m = T$, S the underlying asset, K the strike, and H the barrier. If H is much smaller than $S(0)$, most paths of an ordinary simulation will result in a payoff of zero; importance sampling can potentially make knock-ins less rare.

Suppose the underlying asset is modeled through a process of the form

$$S(t_n) = S(0)\exp(L_n), \quad L_n = \sum_{i=1}^n X_i,$$

with $X_1, X_2, \ldots$ i.i.d. and $L_0 = 0$. This includes geometric Brownian motion but many other models as well; see Section 3.5. The option payoff is then

$$\mathbf{1}\{L_m > c, \tau < m\}$$

where $c = \log(K/S(0))$, τ is the first time the random walk L_n drops below $-b$, and $-b = \log(H/S(0))$. If b or c is large, the probability of a payoff is small. To increase the probability of a payoff, we need to drive L_n down toward $-b$ and then up toward c.

Suppose the X_i have cumulant generating function ψ and consider importance sampling estimators of the following form: exponentially twist the distribution of the X_i by some θ_- (with drift $\psi'(\theta_-) < 0$) until the barrier is crossed, then twist the remaining $X_{\tau+1}, \ldots, X_m$ by some θ_+ (with drift $\psi'(\theta_+) > 0$) to drive the process up toward the strike. On the event $\{\tau < m\}$, the likelihood ratio for this change of measure is (using (4.81) and (4.84))

$$\begin{aligned}&\exp\left(-\theta_- L_\tau + \psi(\theta_-)\tau\right) \cdot \exp\left(-\theta_+[L_m - L_\tau] + \psi(\theta_+)[m-\tau]\right)\\ &\quad = \exp\left((\theta_+ - \theta_-)L_\tau - \theta_+ L_m + (\psi(\theta_-) - \psi(\theta_+))\tau + m\psi(\theta_+)\right).\end{aligned}$$

The importance sampling estimator is the product of this likelihood ratio and the discounted payoff.

We now apply a heuristic argument to select the parameters θ_-, θ_+. We expect most of the variability in the estimator to result from the barrier crossing time τ, because for large b and c we expect $L_\tau \approx -b$ and $L_m \approx c$ on the event $\{\tau < m, L_m > c\}$. (In other words, the undershoot below $-b$ and the overshoot above c should be small.) If we choose θ_-, θ_+ to satisfy $\psi(\theta_-) = \psi(\theta_+)$, the likelihood ratio simplifies to

$$\exp\left((\theta_+ - \theta_-)L_\tau - \theta_+ L_m + m\psi(\theta_+)\right),$$

and we thus eliminate explicit dependence on τ.

To complete the selection of the parameters $\theta_\pm$, we impose the condition that traveling in a straight-line path from 0 to $-b$ at rate $|\psi'(\theta_-)|$ and then from $-b$ to c at rate $\psi'(\theta_+)$, the process should reach c at time m; i.e.,

$$\frac{-b}{\psi'(\theta_-)} + \frac{c+b}{\psi'(\theta_+)} = m.$$

These conditions uniquely determine $\theta_\pm$, at least if the domain of ψ is sufficiently large. This is illustrated in Figure 4.12.

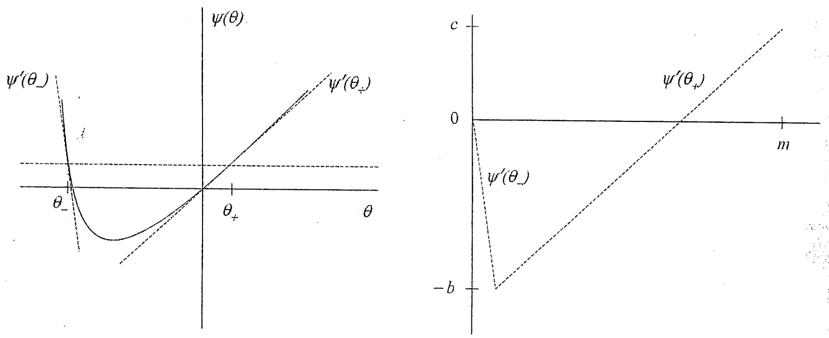

Fig. 4.12. Illustration of importance sampling strategy for a knock-in option. Twisting parameters $\theta_\pm$ are chosen so that (a) $\psi(\theta_-) = \psi(\theta_+)$ and (b) straight-line path with slopes $\psi'(\theta_-)$ and $\psi'(\theta_+)$ reaches $-b$ and then c in m steps.

In the case of geometric Brownian motion GBM(μ, σ^2) with equally spaced time points $t_n = nh$, we have

$$X_n \sim N((\mu - \tfrac{1}{2}\sigma^2)h, \sigma^2 h),$$

and the cumulant generating function is

$$\psi(\theta) = (\mu - \tfrac{1}{2}\sigma^2)h\theta + \tfrac{1}{2}\sigma^2 h\theta^2.$$

Because this function is quadratic in θ, it is symmetric about its minimum and the condition $\psi(\theta_-) = \psi(\theta_+)$ implies that $\psi'(\theta_-) = -\psi'(\theta_+)$. Thus, under our proposed change of measure, the random walk moves at a constant speed of $|\psi'(\theta_\pm)|$. To traverse the path down to the barrier and up to the strike in m steps, we must have

$$|\psi'(\theta_\pm)| = \frac{2b + c}{m}.$$

We can now solve for the twisting parameters to get

$$\theta_\pm = \left(\frac{1}{2} - \frac{\mu}{\sigma^2}\right) \pm \frac{2b + c}{m\sigma^2 h}.$$

The term in parentheses on the right is the point at which the quadratic ψ is minimized. The twisting parameters $\theta_\pm$ are symmetric about this point.

Table 4.4 reports variance ratios based on this method. The underlying asset $S(t)$ is GBM(r,σ^2) with $r = 5\%$, $\sigma = 0.15$, and initial value $S(0) = 95$. We consider an option paying 10,000 if not knocked out, hence having price $10,000 \cdot e^{-rT} P(\tau < m, S(T) > K)$. As above, m is the number of steps and $T \equiv t_m$ is the option maturity. The last column of the table gives the estimated ratio of the variance per replication using ordinary Monte Carlo to the variance using importance sampling. It is thus a measure of the speed-up produced by importance sampling. The estimates in the table are based on 100,000 replications for each case. The results suggest that the variance ratio depends primarily on the rarity of the payoff, and not otherwise on the maturity. The variance reduction can be dramatic for extremely rare payoffs.

An entirely different application of importance sampling to barrier options is developed in Glasserman and Staum [146]. In that method, at each step along a simulated path, the value of an underlying asset is sampled conditional on not crossing a knock-out barrier so that all paths survive to maturity. The one-step conditional distributions define the change of measure in this approach. □

	H	K	Price	Variance Ratio
$T = 0.25$, $m = 50$	94	96	3017.6	2
	90	96	426.6	10
	85	96	5.6	477
	90	106	13.2	177
$T = 1$, $m = 50$	90	106	664.8	6
	85	96	452.0	9
$T = 0.25$, $m = 100$	85	96	6.6	405
	90	106	15.8	180

Table 4.4. Variance reduction using importance sampling in pricing a knock-in barrier option with barrier H and strike K.

4.6.2 Path-Dependent Options

We turn now to a more ambitious application of importance sampling with the aim of reducing variance in pricing path-dependent options. We consider models of underlying assets driven by Brownian motion (or simply by normal random vectors after discretization) and change the drift of the Brownian motion to drive the underlying assets into "important" regions, with "importance" determined by the payoff of the option. We identify a specific change of drift through an optimization problem.

The method described in this section is from Glasserman, Heidelberger, and Shahabuddin (henceforth abbreviated GHS) [139], and that reference contains a more extensive theoretical development than we provide here. This method restricts itself to deterministic changes of drift over discrete time steps. It is theoretically possible to eliminate all variance through a stochastic change of drift in continuous time, essentially by taking the option being priced as the numeraire asset and applying the change of measure associated with this change of numeraire. This however requires knowing the price of the option in advance and is not literally feasible, though it potentially provides a basis for approximations. Related ideas are developed in Chapter 16 of Kloeden and Platen [211], Newton [278, 279], and Schoenmakers and Heemink [319].