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].

We restrict ourselves to simulations on a discrete time grid $0 = t_0 < t_1 < \cdots < t_m = T$. We assume the only source of randomness in the simulated model is a d-dimensional Brownian motion. The increment of the Brownian motion from t_{i-1} to t_i is simulated as $\sqrt{t_i - t_{i-1}}Z_i$, where $Z_1, Z_2, \ldots, Z_m$ are independent d-dimensional standard normal random vectors. Denote by Z the concatenation of the Z_i into a single vector of length $n \equiv md$. Each outcome of Z determines a path of underlying assets or state variables, and each such path determines the discounted payoff of an option. If we let G denote the composition of these mappings, then $G(Z)$ is the discounted payoff derived from Z. Our task is to estimate $\mathsf{E}[G(Z)]$, the expectation taken with Z having the n-dimensional standard normal distribution.

An example will help fix ideas. Consider a single underlying asset modeled as geometric Brownian motion $\mathrm{GBM}(r, \sigma^2)$ and simulated using

$$S(t_i) = S(t_{i-1}) \exp\left([r - \tfrac{1}{2}\sigma^2](t_i - t_{i-1}) + \sigma\sqrt{t_i - t_{i-1}}Z_i \right), \quad i = 1, \ldots, m. \tag{4.86}$$

Consider an Asian call option on the arithmetic average $\bar{S}$ of the $S(t_i)$. We may view the payoff of the option as a function of the Z_i and thus write

$$G(Z) = G(Z_1, \ldots, Z_m) = e^{-rT}[\bar{S} - K]^+.$$

Pricing the option means evaluating $\mathsf{E}[G(Z)]$, the expectation taken with $Z \sim N(0, I)$.

Change of Drift: Linearization

Through importance sampling we can change the distribution of Z and still obtain an unbiased estimator of $\mathsf{E}[G(Z)]$, provided we weight each outcome by the appropriate likelihood ratio. We restrict ourselves to changes of distribution that change the mean of Z from 0 to some other vector μ. Let P_μ and E_μ denote probability and expectation when $Z \sim N(\mu, I)$. From the form of the likelihood ratio given in Example 4.6.1 for normal random vectors, we find that

$$\mathsf{E}[G(Z)] = \mathsf{E}_\mu \left[G(Z) e^{-\mu^\top Z + \frac{1}{2}\mu^\top \mu} \right]$$

for any $\mu \in \Re^n$. We may thus simulate as follows:

for replications $i = 1, \ldots, N$
  generate $Z^{(i)} \sim N(\mu, I)$
  $Y^{(i)} \leftarrow G(Z^{(i)}) \exp\left(-\mu^\top Z^{(i)} + \frac{1}{2}\mu^\top \mu\right)$
return $(Y^{(1)} + \cdots + Y^{(N)})/N$.

This estimator is unbiased for any choice of μ; we would like to choose a μ that produces a low-variance estimator.

If G takes only nonnegative values (as is typical of discounted option payoffs), we may write $G(z) = \exp(F(z))$, with the convention that $F(z) = -\infty$

if $G(z) = 0$. Also, note that taking an expectation over Z under P_μ is equivalent to replacing Z with $\mu + Z$ and taking the expectation under the original measure. (In the algorithm above, this simply means that we can sample from $N(\mu, I)$ by sampling from $N(0, I)$ and adding μ.) Thus,

$$\begin{aligned}\mathsf{E}[G(Z)] = \mathsf{E}\left[e^{F(Z)}\right] &= \mathsf{E}_\mu\left[e^{F(Z)}e^{-\mu^\top Z+\frac{1}{2}\mu^\top\mu}\right] \\ &= \mathsf{E}\left[e^{F(\mu+Z)}e^{-\mu^\top(\mu+Z)+\frac{1}{2}\mu^\top\mu}\right] \\ &= \mathsf{E}\left[e^{F(\mu+Z)}e^{-\mu^\top Z-\frac{1}{2}\mu^\top\mu}\right]. \end{aligned} \tag{4.87}$$

For any μ, the expression inside the expectation in (4.87) is an unbiased estimator with Z having distribution $N(0, I)$. To motivate a particular choice of μ, we now expand F to first order to approximate the estimator as

$$e^{F(\mu+Z)}e^{-\mu^\top Z-\frac{1}{2}\mu^\top\mu} \approx e^{F(\mu)+\nabla F(\mu)Z}e^{-\mu^\top Z-\frac{1}{2}\mu^\top\mu}, \tag{4.88}$$

with $\nabla F(\mu)$ the gradient of F at μ. If we can choose μ to satisfy the fixed-point condition

$$\nabla F(\mu) = \mu^\top, \tag{4.89}$$

then the expression on the right side of (4.88) collapses to a constant with no dependence on Z. Thus, applying importance sampling with μ satisfying (4.89) would produce a zero-variance estimator if (4.88) held exactly, and it should produce a low-variance estimator if (4.88) holds only approximately.

Change of Drift: Normal Approximation and Optimal Path

We now present an alternative argument leading to an essentially equivalent choice of μ. Recall from the discussion surrounding (4.76) that the optimal importance sampling density is the normalized product of the integrand and the original density. For the problem at hand, this means that the optimal density is proportional to

$$e^{F(z)-\frac{1}{2}z^\top z},$$

because $\exp(F(z))$ is the integrand and $\exp(-z^\top z/2)$ is proportional to the standard normal density. Normalizing this function by its integral produces a probability density but not, in general, a normal density. Because we have restricted ourselves to changes of mean, we may try to select μ so that $N(\mu, I)$ approximates the optimal distribution. One way to do this is to choose μ to be the mode of the optimal density; i.e., choose μ to solve

$$\max_z F(z) - \tfrac{1}{2}z^\top z. \tag{4.90}$$

The first-order condition for the optimum is $\nabla F(z) = z^\top$, which coincides with (4.89). If, for example, the objective in (4.90) is strictly concave, and if the first-order condition has a solution, this solution is the unique optimum.

We may interpret the solution z_* to (4.90) as an optimal *path*. Each $z \in \Re^n$ may be interpreted as a path because each determines a discrete Brownian path and thus a path of underlying assets. The solution to (4.90) is the most "important" path if we measure importance by the product of payoff $\exp(F(z))$ and probability density $\exp(-z^\top z/2)/(2\pi)^{n/2}$. In choosing $\mu = z_*$, we are therefore choosing the new drift to push the process along the optimal path.

GHS [139] give conditions under which this approach to importance sampling has an asymptotic optimality property. This property is based on introducing a parameter ϵ and analyzing the second moment of the estimator as ϵ approaches zero. From a practical perspective, a small ϵ should be interpreted as a nearly linear F.

Asian Option

We illustrate the selection and application of the optimal change of drift in the case of the Asian call defined above, following the discussion in GHS [139]. Solving (4.90) is equivalent to maximizing $G(z)\exp(-z^\top z/2)$ with G the discounted payoff of the Asian option. The discount factor e^{-rT} is a constant in this example, so for the purpose of optimization we may ignore it and redefine $G(z)$ to be $[\bar{S} - K]^+$. Also, in maximizing it clearly suffices to consider points z at which $\bar{S} > K$ and thus at which G is differentiable.

For the first-order conditions, we differentiate

$$[\bar{S} - K]e^{-z^\top z/2}$$

to get

$$\frac{\partial \bar{S}}{\partial z_j} - [\bar{S} - K]z_j = 0.$$

Using (4.86), we find that

$$\frac{\partial \bar{S}}{\partial z_j} = \frac{1}{m}\sum_{i=j}^{m} \frac{\partial S(t_i)}{\partial z_j} = \frac{1}{m}\sum_{i=j}^{m} \sigma\sqrt{t_i - t_{i-1}}S(t_i).$$

The first-order conditions thus become

$$z_j = \frac{\sum_{i=j}^{m} \sigma\sqrt{t_i - t_{i-1}}S(t_i)}{mG(z)}.$$

Now we specialize to the case of an equally spaced time grid with $t_i - t_{i-1} \equiv h$. This yields

$$z_1 = \frac{\sigma\sqrt{h}(G(z)+K)}{G(z)}, \quad z_{j+1} = z_j - \frac{\sigma\sqrt{h}S(t_j)}{mG(z)}, \quad j = 1,\ldots,m-1. \quad (4.91)$$

Given the value of $G(z)$, (4.91) and (4.86) determine z. Indeed, if $y \equiv G(z)$, we could apply (4.91) to calculate z_1 from y, then (4.86) to calculate $S(t_1)$, then (4.91) to calculate z_2, and so on. Through this iteration, each value of y determines a $z(y)$ and path $S(t_i, y)$, $i = 1,\ldots,m$. Solving the first-order conditions reduces to finding the y for which the payoff at $S(t_1, y),\ldots,S(t_m, y)$ is indeed y; that is, it reduces to finding the root of the equation

$$\frac{1}{m}\sum_{j=1}^{m} S(t_j, y) - K - y = 0.$$

GHS [139] report that numerical examples suggest that this equation has a unique root. This root can be found very quickly through a one-dimensional search. Once the root y_* is found, the optimization problem is solved by $z_* = z(y_*)$. To simulate, we then set $\mu = z_*$ and apply importance sampling with mean μ.

Combined Importance Sampling and Stratification

In GHS [139], further (and in some cases enormous) variance reduction is achieved by combining importance sampling with stratification of a linear projection of Z. Recall from Section 4.3.2 that sampling Z so that $v^\top Z$ is stratified for some $v \in \Re^n$ is easy to implement. The change of mean does not affect this. Indeed, we may sample from $N(\mu, I)$ by sampling from $N(0, I)$ and then adding μ; we can apply stratified sampling to $N(0, I)$ before adding μ.

Two strategies for selecting the stratification direction v are considered in GHS [139]. One simply sets $v = \mu$ on the grounds that μ is an important path and thus a potentially important direction for stratification. The other strategy expands (4.88) to get

$$e^{F(\mu+Z)}e^{-\mu^\top Z - \frac{1}{2}\mu^\top\mu} \approx e^{F(\mu)+\nabla F(\mu)Z+\frac{1}{2}Z^\top H(\mu)Z}e^{-\mu^\top Z-\frac{1}{2}\mu^\top\mu},$$

with $H(\mu)$ the Hessian matrix of F at μ. Importance sampling with $\mu^\top = \nabla F(\mu)$ eliminates the linear term in the exponent, and this suggests that the stratification should be tailored to the quadratic term.

In Section 4.3.2, we noted that the optimal stratification direction for estimating an expression of the form $\mathsf{E}[\exp(\frac{1}{2}Z^\top AZ)]$ with A symmetric is an eigenvector of A. The optimal eigenvector is determined by the eigenvalues of A through the criterion in (4.50). This suggests that we should stratify along the optimal eigenvector of the Hessian of F at μ. This entails numerical calculation of the Hessian and its eigenvectors.

Table 4.5 shows results from GHS [139]. The table shows variance ratios (i.e., variance reduction factors) using importance sampling and two combinations of importance sampling with stratified sampling, using the two strategies just described for selecting a stratification direction. All results use $S(0) = 50$, $r = 0.05$, and $T = 1$ and are estimated from one million paths for each case. The results show that importance sampling by itself can produce noteworthy variance reduction (especially for out-of-the-money options) and that the combined impact with stratification can be astounding. The combination reduces variance by factors in the thousands.

n	σ	K	Price	Importance Sampling	IS & Strat. (μ)	IS & Strat. (v_{j^*})
16	0.10	45	6.05	11	1,097	1,246
		50	1.92	7	4,559	5,710
		55	0.20	21	15,520	17,026
16	0.30	45	7.15	8	1,011	1,664
		50	4.17	9	1,304	1,899
		55	2.21	12	1,746	2,296
64	0.10	45	6.00	11	967	1,022
		50	1.85	7	4,637	5,665
		55	0.17	23	16,051	17,841
64	0.30	45	7.02	8	1,016	1,694
		50	4.02	9	1,319	1,971
		55	2.08	12	1,767	2,402

Table 4.5. Estimated variance reduction ratios for Asian options using importance sampling and combinations of importance sampling with stratified sampling, stratifying along the optimal μ or the optimal eigenvector v_{j^*}. Stratified results use 100 strata.

The results in Table 4.5 may seem to suggest that stratification has a greater impact than importance sampling, and one may question the value of importance sampling in this example. But the effectiveness of stratification is indeed enhanced by the change in mean, which results in more paths producing positive payoffs. The positive-part operator $[\cdot]^+$ applied to $\bar{S} - K$ diminishes the effectiveness of stratified sampling, because it tends to produce many strata with a constant (zero) payoff — stratifying a region of constant payoff is useless. By shifting the mean of Z, we implicitly move more of the distribution of the stratification variable $v^\top Z$ (and thus more strata) into the region where the payoff varies. In this particular example, the region defined by $\bar{S} > K$ is reasonably easy to characterize and could be incorporated into the selection of strata; however, this is not the case in more complex examples.

A further notable feature of Table 4.5 is that the variance ratios are quite similar whether we stratify along μ or along the optimal eigenvector v_*. In fact, GHS [139] find that the vectors μ and v_* nearly coincide, once normalized

to have the same length. They find similar patterns in other examples. This phenomenon can occur when G is well approximated by a nonlinear function of a linear combination of $z_1, \ldots, z_m$. For suppose $G(z) \approx g(v^\top z)$; then, the gradient of G is nearly proportional to $v^\top$ and so μ will be nearly proportional to v. Moreover, the Hessian of G will be nearly proportional to the rank-1 matrix $vv^\top$, whose only nonzero eigenvectors are multiples of v. Thus, in this setting, the optimal mean is proportional to the optimal eigenvector.

Application in the Heath-Jarrow-Morton Framework

GHS [140] apply the combination of importance sampling and stratified sampling in the Heath-Jarrow-Morton framework. The complexity of this setting necessitates some approximations in the calculation of the optimal path and eigenvector to make the method computationally feasible. We comment on these briefly.

We consider a three-factor model (so $d = 3$) discretized in time and maturity as detailed in Section 3.6.2. The discretization interval is a quarter of a year and we consider maturities up to 20 years, so $m = 80$ and the vector Z of random inputs has dimension $n = md = 240$. The factor loadings for each of the three factors are as displayed in Figure 4.13, where they are plotted against time to maturity.

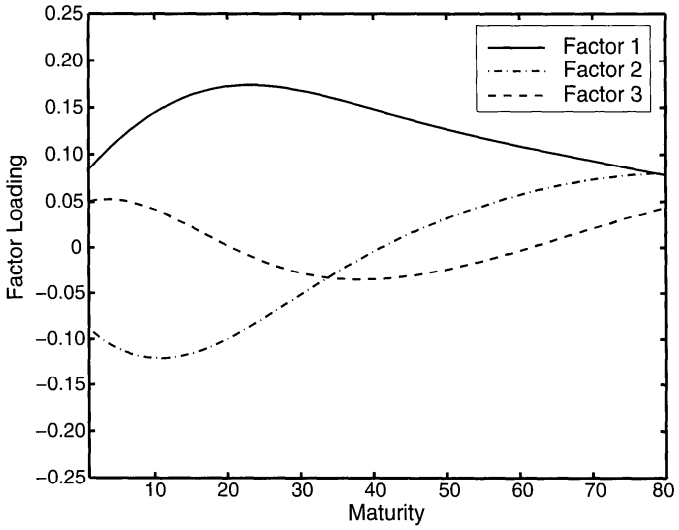

Fig. 4.13. Factor loadings in three-factor HJM model used to illustrate importance sampling.

As in Section 3.6.2, we use $\hat{f}(t_i, t_j)$ to denote the forward rate at time t_i for the interval $[t_j, t_{j+1}]$. We use an equally spaced time grid so $t_i = ih$ with h a quarter of a year. The initial forward curve is

$$\hat{f}(0, t_j) = \log(150 + 12j)/100, \quad j = 0, 1, \ldots, 80,$$