Mathematical Finance, Vol. 9, No. 2 (April 1999), 117–152

ASYMPTOTICALLY OPTIMAL IMPORTANCE SAMPLING AND STRATIFICATION FOR PRICING PATH-DEPENDENT OPTIONS

PAUL GLASSERMAN

Graduate School of Business, Columbia University

PHILIP HEIDELBERGER

IBM Research Division, T.J. Watson Research Center

PERWEZ SHAHABUDDIN

IEOR Department, Columbia University

This paper develops a variance reduction technique for Monte Carlo simulations of path-dependent options driven by high-dimensional Gaussian vectors. The method combines *importance sampling* based on a change of drift with *stratified sampling* along a small number of key dimensions. The change of drift is selected through a large deviations analysis and is shown to be optimal in an asymptotic sense. The drift selected has an interpretation as the path of the underlying state variables which maximizes the product of probability and payoff—the most important path. The directions used for stratified sampling are optimal for a quadratic approximation to the integrand or payoff function. Indeed, under differentiability assumptions our importance sampling method eliminates variability due to the linear part of the payoff function, and stratification eliminates much of the variability due to the quadratic part of the payoff. The two parts of the method are linked because the asymptotically optimal drift vector frequently provides a particularly effective direction for stratification. We illustrate the use of the method with path-dependent options, a stochastic volatility model, and interest rate derivatives. The method reveals novel features of the structure of their payoffs.

KEY WORDS: Monte Carlo methods, variance reduction, large deviations, Laplace principle

1. INTRODUCTION

This paper develops a variance reduction technique for Monte Carlo simulations driven by high-dimensional Gaussian vectors, with particular emphasis on the pricing of path-dependent options. The method combines *importance sampling* based on a change of drift with *stratified sampling* along a small number of key dimensions. The change of drift is selected through a large deviations analysis and is shown to be optimal in an asymptotic sense. The directions used for stratified sampling are optimal for a quadratic approximation to the integrand or payoff function. Indeed, under differentiability assumptions our importance sampling method eliminates variability due to the linear part of the payoff function, and stratification eliminates much of the variability due to the quadratic part of the payoff. The two parts of the method are linked because the asymptotically optimal drift vector frequently provides a particularly effective direction for stratification.

This research is supported in part by NSF grants DMI-9457189, DMS-9508709, and DMI-9625297. The authors thank Fernando Bernstein for computational assistance.

Manuscript received January 1998; final revision received September 1998.

Address correspondence to P. Glasserman at the Graduate School of Business, Columbia University, 3022 Broadway, Room 403, New York, NY 10027-6902; e-mail: pg20@columbia.edu.

Monte Carlo simulation is frequently the only method available for the pricing of complex path-dependent options, particularly if the number of relevant underlying assets is large or if additional sources of randomness like stochastic volatility, stochastic interest rates, or a stochastic convenience yield are included in the model. The computational demands of simulation have motivated substantial interest in the financial industry in methods for increased efficiency, as evidenced in part by the growth in commercial software offering enhancements to Monte Carlo and a large number of publications on the subject in the practitioner literature. (See Boyle, Broadie, and Glasserman 1997 for an overview of simulation methods for option pricing.) If a particular derivative security is to be valued just once or only rarely, the investment required to value it efficiently may not be justified. If, however, a pricing routine is to be run often—because a firm regularly quotes prices on the instrument, because it rebalances its hedges frequently, or because prices in many different scenarios are required for daily value-at-risk calculations—it may well be worthwhile to carry out some analysis that leads to more efficient pricing. The method proposed here requires this type of initial investment, but it has the potential to provide enormous variance reduction as a result.

We restrict attention to simulations driven by a sequence of independent standard normal random variables. Independent normals can of course be converted into correlated normals through a linear transformation; and if we allow sufficiently complicated transformations, a sample from a normal distribution can at least theoretically be converted into a sample from any other distribution. However, we have in mind simulations of discretized diffusion processes using, for example, an *Euler scheme* or higher-order discretization (see Kloeden and Platen 1992), or an exact solution to a stochastic differential equation if available. We do not address the issue of discretization bias. Rather, we assume that an acceptable discretization has already been determined and thus we focus attention on obtaining precise estimates at that level of discretization. This is a reasonable perspective in practice because unless an extremely coarse discretization is chosen the statistical error associated with simulation is likely to overwhelm the bias associated with discretization.

An example helps to illustrate the setting. Consider the pricing of an arithmetic Asian option on a single underlying asset under standard Black–Scholes assumptions. The price of the underlying asset under the equivalent martingale measure is described by the stochastic differential equation

$$\frac{dS_t}{S_t} = r\,dt + \sigma\,dW_t, \tag{1.1}$$

with r the risk-free, continuously compounded interest rate, σ the asset's volatility, W_t a standard Wiener process, and S_0 fixed. The solution to this equation can be simulated (without discretization error) on a discrete grid of points $0 = t_0 < t_1 < \cdots < t_n = T$ by setting

$$S_{t_i} = S_{t_{i-1}} \exp([r - \tfrac{1}{2}\sigma^2](t_i - t_{i-1}) + \sigma\sqrt{t_i - t_{i-1}}Z_i), \qquad i = 1, \ldots, n, \tag{1.2}$$

where $Z_1, \ldots, Z_n$ are independent standard normals. The discounted payoff on an arithmetic Asian option with strike K is given by

$$G(Z_1, \ldots, Z_n) = e^{-rT}(\bar{S} - K)^+, \tag{1.3}$$

with $\bar{S} = \frac{1}{n}\sum_{i=1}^{n} S_{t_i}$. More generally, $\bar{S}$ could be specified as the average over a subset of the S_{t_i} or a weighted average of these values. The objective is to evaluate $E[G(Z_1, \ldots, Z_n)]$. A crude Monte Carlo simulation would simply average replications of $G(Z)$ over independent draws of the vector $Z' = (Z_1, \ldots, Z_n)$; the first step in our approach is to add a drift vector to Z.

The integrand G implicitly defined by equation (1.3) is typical of option payoffs in that it is nonnegative. We may therefore write

$$G(z) = e^{F(z)}\mathbf{1}_D = e^{F(z)}\mathbf{1}_D(z), \quad z \in \mathbf{R}^n,$$

where $D = \{z \in \mathbf{R}^n : G(z) > 0\}$, $\mathbf{1}_D$ is the indicator of the set D, and $F(z) \stackrel{\triangle}{=} \log G(z)$ on D. (Alternatively, we could set $F(z) = -\infty$ off D and omit the indicator.) To identify an appropriate drift vector, we examine what happens if we replace the driving Wiener process W in equation (1.1) with a smaller-noise process $\sqrt{\epsilon}W$ and simultaneously scale the payoff by raising it to the power of $1/\epsilon$. Through a generalization of the classical Laplace method for integrals known as Varadhan's Integral Lemma (Dembo and Zeitouni 1993) or a Laplace Principle (Dupuis and Ellis 1997), we will see in Section 2 that we frequently have

$$E[e^{F(\sqrt{\epsilon}Z)/\epsilon}\mathbf{1}_D] \approx e^{\psi(z_*)/\epsilon}, \qquad \epsilon \downarrow 0,$$

in a sense to be made precise, where z_* maximizes the function $\psi(z) = F(z) - \frac{1}{2}z'z$ over D. Maximizing ψ is equivalent to maximizing the product of the payoff and the probability density. We use z_* as the new drift vector and analyze the variance of the resulting importance sampling estimator. To obtain further variance reduction, we then stratify (in effect, numerically integrate) along one or possibly more directions in $\mathbf{R}^n$. The vector z_* frequently provides a good choice of direction.

There is a large body of work on the use of large deviations asymptotics to identify an effective change of measure for estimating *rare event probabilities* by simulation; see, for example, Glasserman and Wang (1997), Heidelberger (1995), and Shahabuddin (1995) for numerous references and discussions of the literature, and see Chen et al. (1993) for the Gaussian case in particular. In the option pricing setting, Reider (1993) uses importance sampling to price deep out-of-the-money calls, and Boyle et al. (1997) give an application to pricing knock-in options far from the barrier. In the rare event setting, the integrand is usually the indicator of a rare set and importance sampling is used to make the set less rare. In our setting, the set on which a positive payoff occurs need not be rare at all, and yet large deviations asymptotics are useful in identifying a change of measure. Instead of simply increasing the probability of an event, the change of measure provided by this analysis balances the magnitude and probability of payoffs and puts the mean on the trajectory that effectively maximizes the product of the two.

Importance sampling for diffusion processes is treated in Kloeden and Platen (1992) and Newton (1994) in a general setting; further developments in the application of the method to option pricing appear in Andersen (1995), Fournié, Lasry, and Touzi (1997), Newton (1997), and Schoenmakers and Heemink (1997). For problems that can be formulated as path-*independent* options in continuous time (possibly through augmentation of the number of state variables), a zero-variance estimator can be obtained through a *stochastic* change of drift. Determining the optimal drift, however, requires knowing the option price in advance.

Most of these references are therefore based on using approximations to the option price to find approximations to the optimal drift. Andersen (1995) applies the optimal drift from a simple model to a less tractable one; Newton (1994, 1997) and Schoenmakers and Heemink (1997) use approximations calculated numerically, and Fournié et al. (1997) use asymptotics around the Black–Scholes model. Some of these approaches rely on PDE methods, which may make them difficult to use with high-dimensional problems. Because we restrict ourselves to deterministic changes of drift, the viability of our method should be less sensitive to problem dimension. Additionally, we explicitly analyze the variance of our estimators and use this analysis (rather than approximations *per se*) to guide the method.

The rest of this paper is organized as follows. Section 2 reviews background on importance sampling and then identifies and analyzes an asymptotically optimal change of measure. Section 3 addresses the computation of the drift vector for this change of measure. Section 4 contains a general discussion of stratification, the optimal direction for quadratic F, and the connection with the optimal drift. Section 5 gives numerical results and practical considerations for implementation. Section 6 concludes the paper, and an Appendix contains some proofs.

2. IMPORTANCE SAMPLING

2.1. Preliminaries

Consider the general problem of estimating $c \stackrel{\triangle}{=} E[G(Z)\mathbf{1}_D]$ for some $G : \mathbf{R}^n \to [0, \infty)$, with Z an n-dimensional random vector having multivariate density g. Shortly, we limit ourselves to the multivariate normal case, but for the moment we keep the setting general. Let h be another density with the property that $g(z) > 0 \Rightarrow h(z) > 0$ for all $z \in D$. Then

$$\begin{aligned} E[G(Z)\mathbf{1}_D] \equiv E_g[G(Z)\mathbf{1}_D] &= \int_D G(z)g(z)\,dz \\ &= \int_D G(z)\frac{g(z)}{h(z)}h(z)\,dz \equiv E_h\left[G(Z)\left(\frac{g(Z)}{h(Z)}\right)\mathbf{1}_D\right], \end{aligned} \tag{2.1}$$

where the subscript on the expectation indicates the density with respect to which the integral is computed. The factor $g(Z)/h(Z)$ is called the *likelihood ratio* or Radon–Nikodym derivative. It follows that, with Z drawn from h, the estimate $G(Z)[g(Z)/h(Z)]\mathbf{1}_D$ is unbiased for c; this is an *importance sampling* estimate. Its variance is

$$\int_D \left(G(z)\frac{g(z)}{h(z)} - c\right)^2 h(z)\,dz.$$

If $G \geq 0$, a zero-variance estimate is thus obtained by choosing

$$h(z) \propto G(z)g(z)\mathbf{1}_D; \tag{2.2}$$

but the proportionality constant required to make the right side a density is $1/c$, precluding this choice of density unless the desired quantity c is known from the outset. Nevertheless, this observation provides a useful insight: An effective importance sampling density should weight points according to the product of their probability and their payoff.

Now suppose the original probability density g is the multivariate normal density over $\mathbf{R}^n$ with mean vector 0 and covariance matrix I_n, the $n \times n$ identity. For any $\mu \in \mathbf{R}^n$, let h_μ be the normal density with mean vector μ and the same covariance matrix. Write E_μ for expectation with respect to h_μ. Simple algebra shows that the importance sampling identity (2.1) becomes

$$E[G(Z)\mathbf{1}_D] = E_\mu\left[G(Z)e^{-\mu' Z+(1/2)\mu'\mu}\mathbf{1}_D\right].$$

Thus, drawing Z from h_μ and multiplying $G(Z)\mathbf{1}_D$ by the likelihood ratio $\exp(-\mu' Z + \frac{1}{2}\mu'\mu)$ yields an unbiased estimator. Of course, the distribution of Z under h_μ is the same as that of $Z + \mu$ under the original density $g = h_0$; making this substitution yields

$$E_\mu\left[G(Z)e^{-\mu' Z+(1/2)\mu'\mu}\mathbf{1}_D(z)\right] = E\left[G(Z+\mu)e^{-\mu' Z-(1/2)\mu'\mu}\mathbf{1}_D(Z+\mu)\right] \tag{2.3}$$

and the unbiased estimator

$$G(Z+\mu)e^{-\mu' Z-(1/2)\mu'\mu}\mathbf{1}_D(Z+\mu)$$

under the original measure. The second moment of this estimator is

$$E_\mu\left[G(Z)^2 e^{-2\mu' Z+\mu'\mu}\mathbf{1}_D\right] = E\left[G(Z)^2 e^{-\mu' Z+(1/2)\mu'\mu}\mathbf{1}_D\right]. \tag{2.4}$$

Among all μ, the optimal one solves the problem

$$\min_\mu E\left[G(Z)^2 e^{-\mu' Z+(1/2)\mu'\mu}\mathbf{1}_D\right]. \tag{2.5}$$

Even if the optimal μ can be found, it will not in general provide a zero-variance estimator. Its effectiveness will depend on the extent to which h_μ approximates the optimal density identified in equation (2.2).

In practice, finding the optimal μ exactly is infeasible and some approximation is required. As in Section 1, let us write $F(z)$ for $\log G(z)$, $z \in D$, so that equation (2.5) becomes

$$\min_\mu E\left[e^{2F(Z)-\mu' Z+(1/2)\mu'\mu}\mathbf{1}_D\right]. \tag{2.6}$$

The classical Laplace method for integrals (e.g., Bleistein and Handelsman 1975, Chap. 8) suggests that, for any fixed μ,

$$\begin{aligned} E\left[e^{2F(Z)-\mu' Z+(1/2)\mu'\mu}\mathbf{1}_D\right] &\equiv (2\pi)^{n/2}\int_D e^{2F(z)-\mu' z+(1/2)\mu'\mu}e^{-(1/2)z'z}\,dz \\ &\approx \text{constant} \times \exp\left(\max_{z\in D}\{2F(z) - \mu' z + \tfrac{1}{2}\mu'\mu - \tfrac{1}{2}z'z\}\right). \end{aligned}$$

Substituting this approximation into equation (2.6) suggests that we may approximate the optimal μ by solving

$$\min_{\mu} \max_{z \in D} \left\{ 2F(z) - \mu' z + \tfrac{1}{2}\mu'\mu - \tfrac{1}{2} z' z \right\}. \tag{2.7}$$

Alternatively, we may interpret (2.2) as stating that an effective importance sampling density assigns high probability to regions of D on which $e^{F(z)}g(z)$ is large. In the multivariate normal case, this suggests choosing the drift vector to be the point μ that solves

$$\max_{z \in D} \{F(z) - \tfrac{1}{2} z' z\}. \tag{2.8}$$

The next subsection develops a setting in which these approximations become equivalent and, in fact, optimal in an asymptotic sense.

2.2. Large Deviations and Laplace Principles

We begin with brief generalities on large deviations, then specialize to our context. For additional background see Dembo and Zeitouni (1993), Deuschel and Stroock (1989), and Dupuis and Ellis (1997).

A sequence of probability measures ν_ϵ on a topological space $(\mathbf{X}, \mathcal{B})$ ($\mathcal{B}$ the completion of the Borel sets on $\mathbf{X}$) satisfies a *large deviations principle with good rate function* I if the function $I : \mathbf{X} \to [0, \infty]$ is lower semicontinuous with compact level sets and if the following conditions hold:

(i) for all open sets $B \subseteq \mathbf{X}$

$$\liminf_{\epsilon \to 0} \epsilon \log \nu_\epsilon(B) \geq - \inf_{x \in B} I(x),$$

(ii) for all closed sets $C \subseteq \mathbf{X}$

$$\limsup_{\epsilon \to 0} \epsilon \log \nu_\epsilon(C) \leq - \inf_{x \in C} I(x).$$

Part of the content of this definition is that if both limits hold for some set D, then we may write

$$\nu_\epsilon(D) = e^{[-\inf_{x \in D} I(x) + o(1)]/\epsilon}.$$

We will use the following version of a result called Varadhan's Integral Lemma in Dembo and Zeitouni (1993). This type of result is called a *Laplace principle* in Dupuis and Ellis (1997). To accommodate our intended application, we explicitly consider functions that may take the value $-\infty$. For a set $D \in \mathcal{B}$, D^o denotes the interior of D and $\bar{D}$ the closure of D.

LEMMA 2.1. *Suppose that ν_ϵ satisfies a large deviations principle with good rate function I. Suppose that $f : \mathbf{X} \to [-\infty, \infty)$ is continuous and that for some $q > 1$*

$$\limsup_{\epsilon \to 0} \epsilon \log \int e^{q f(x)/\epsilon} d\nu_\epsilon < \infty. \tag{2.9}$$

Then

(i) *for all open sets B*

$$\liminf_{\epsilon \to 0} \epsilon \log \int_B e^{f(x)/\epsilon} d\nu_\epsilon \le \sup_{x \in B} \{ f(x) - I(x) \},$$

(ii) *for all closed sets C*

$$\limsup_{\epsilon \to 0} \epsilon \log \int_C e^{f(x)/\epsilon} d\nu_\epsilon \ge \sup_{x \in C} \{ f(x) - I(x) \}.$$

Thus, if $\nu_\epsilon(D^o) = \nu_\epsilon(\bar{D})$ for all ϵ and if I is continuous on $\bar{D}$, then

$$\lim_{\epsilon \to 0} \epsilon \log \int_D e^{f(x)/\epsilon} d\nu_\epsilon = \sup_{x \in D} \{ f(x) - I(x) \}. \tag{2.10}$$

REMARK 2.1. Writing equation (2.10) as

$$\int_D e^{f(x)/\epsilon} d\nu_\epsilon = e^{[\sup_{x \in D} \{ f(x) - I(x) \} + o(1)]/\epsilon}$$

makes its interpretation as an approximation more transparent.

Proof. The proof of (i) is the same as that given for Lemma 4.3.4 of Dembo and Zeitouni (1993); the possibility that f takes the value $-\infty$ and the restriction to the integral over the open set B do not affect the validity of the argument given there. For (ii), first observe that if $f \equiv -\infty$ the result holds trivially, so suppose this is not the case. For $M \in \mathbf{R}$ define $C_M = C \cap \{x : f(x) \ge -M\}$; C_M is closed because f is continuous. For any $\delta > 0$ let $x_0 \in C$ satisfy

$$f(x_0) - I(x_0) > \sup_{x \in C} \{ f(x) - I(x) \} - \delta$$

and observe that $f(x_0) > -\infty$. Choose $M \ge -f(x_0)$. Then

$$\begin{aligned} \limsup_{\epsilon \to 0} \epsilon \log \int_C e^{f(x)/\epsilon} \, d\nu_\epsilon &\ge \limsup_{\epsilon \to 0} \epsilon \log \int_{C_M} e^{f(x)/\epsilon} \, d\nu_\epsilon \\ &\ge \sup_{x \in C_M} \{ f(x) - I(x) \} \end{aligned}$$

$$\begin{aligned} &\geq f(x_0) - I(x_0) \\ &> \sup_{x \in C}\{f(x) - I(x)\} - \delta. \end{aligned}$$

The second inequality follows from equation (2.9) and Theorem 4.3.1 of Dembo and Zeitouni (1993). (In that result, the integral is taken over the entire space. The restriction to a closed subset appears in Dembo and Zeitouni as Exercise 4.3.11 and in Deuschel and Stroock (1989) as 2.1.24.) Since $\delta > 0$ is arbitrary, the result follows. □

We apply Lemma 2.1 to the measures $\nu_\epsilon(\cdot) = P(\sqrt{\epsilon}Z \in \cdot)$ where Z is a vector of n independent standard normal random variables. These measures satisfy a large deviations principle with good rate function $I(z) = \frac{1}{2}z'z$. This function is evidently continuous, and under each ν_ϵ open sets have the same measure as their closures and closed sets have the same measure as their interiors. The last case in Lemma 2.1 therefore applies. We record this important special case in the following lemma.

LEMMA 2.2. *Suppose that $f : \mathbf{R}^n \to [-\infty, \infty)$ is continuous and satisfies $f(z) \leq c_1 + c_2 z'z$ for some $c_2 < 1/2$ and all $z \in D$, where D is either open or closed. Then*

$$\lim_{\epsilon \to 0} \epsilon \log E[e^{f(\sqrt{\epsilon}Z)/\epsilon}\mathbf{1}_D] = \sup_{z \in D}\{f(z) - \tfrac{1}{2}z'z\}.$$

Proof. We may assume $c_2 > 0$ and choose any $1 < q < 1/(2c_2)$. Then

$$E[e^{qf(\sqrt{\epsilon}Z)/\epsilon}\mathbf{1}_D] \leq e^{qc_1/\epsilon}E[e^{qc_2Z'Z}\mathbf{1}_D].$$

Since $qc_2 < 1/2$, we have $E[e^{qc_2Z'Z}\mathbf{1}_D] \leq E[e^{qc_2Z'Z}] < \infty$ and thus

$$\lim_{\epsilon \to 0} \epsilon \log E[e^{qf(\sqrt{\epsilon}Z)/\epsilon}\mathbf{1}_D] \leq qc_1 < \infty.$$

This verifies equation (2.9). The result now follows from Lemma 2.1. □

2.3. Asymptotic Optimality

We return now to the problem posed in equation (2.6) of finding the optimal drift vector μ for estimating $E[\exp(F(Z))\mathbf{1}_D]$. To make precise the approximations following (2.6), we generalize the setting and consider the problem of estimating

$$\alpha(\epsilon) = E[e^{F(\sqrt{\epsilon}Z)/\epsilon}\mathbf{1}_D], \qquad \epsilon > 0.$$

Notice that the quantity of interest $c = E[\exp(F(Z))\mathbf{1}_D]$ is $\alpha(\epsilon)$ at $\epsilon = 1$. The Laplace Principle approximations give us information about $\alpha(\epsilon)$ as $\epsilon \to 0$. In order to use this information, we embed the problem of estimating c in the more general problem of estimating $\alpha(\epsilon)$ and analyze the behavior of estimators of $\alpha(\epsilon)$ when ϵ is small. We will see

shortly that letting $\epsilon \to 0$ can be interpreted as making a linear approximation to F in the original problem of estimating c.

Although ultimately we work only with unbiased estimators, to show that our estimators remain attractive within a broader class we introduce the following definition.

DEFINITION 2.1. A family of estimators $\{\hat{\alpha}(\epsilon)\}$ is *asymptotically relatively unbiased* if

$$\frac{E[\hat{\alpha}(\epsilon)] - \alpha(\epsilon)}{\alpha(\epsilon)} \to 0 \quad \text{as } \epsilon \to 0. \tag{2.11}$$

In comparing such estimators (and unbiased estimators in particular) we compare their second moments as in the following definition.

DEFINITION 2.2. A family of asymptotically relatively unbiased estimators $\{\hat{\alpha}_o(\epsilon)\}$ is *asymptotically optimal* if

$$\limsup_{\epsilon \to 0} \epsilon \log E[\hat{\alpha}_o^2(\epsilon)] = \inf_{\{\hat{\alpha}(\epsilon)\}} \limsup_{\epsilon \to 0} \epsilon \log E[\hat{\alpha}^2(\epsilon)], \tag{2.12}$$

the infimum taken over all $\{\hat{\alpha}(\epsilon)\}$ satisfying equation (2.11).

In other words, $\hat{\alpha}_o(\epsilon)$ is asymptotically optimal if its second moment achieves the smallest exponential rate in ϵ. The degenerate estimator $\hat{\alpha}(\epsilon) \equiv \alpha(\epsilon)$ is trivially unbiased and in fact asymptotically optimal. Hence, the infimum in equation (2.12) is

$$\limsup_{\epsilon \to 0} \epsilon \log \alpha^2(\epsilon) = 2 \limsup_{\epsilon \to 0} \epsilon \log \alpha(\epsilon).$$

An asymptotically optimal estimator is thus one whose second moment achieves twice the exponential rate of $\alpha(\epsilon)$ itself. This condition is sometimes termed *asymptotic efficiency*.

For any candidate drift vector μ_ϵ, by the μ_ϵ-IS estimator we mean

$$\exp\left\{\epsilon^{-1} F(\sqrt{\epsilon} Z) - \mu_\epsilon' Z + \tfrac{1}{2} \mu_\epsilon' \mu_\epsilon\right\} \mathbf{1}_D,$$

with $Z \sim N(\mu_\epsilon, I_n)$. (The symbol $N(m, \Sigma)$ denotes the normal distribution with mean vector m and covariance matrix Σ; as before, I_n is the $n \times n$ identity matrix.) Our main result on importance sampling gives conditions under which solving for the optimal μ in equations (2.7) or (2.8) and setting $\mu_\epsilon = \mu/\sqrt{\epsilon}$ results in an asymptotically optimal μ_ϵ-IS estimator.

THEOREM 2.1. *Suppose that $F : \mathbf{R}^n \to [-\infty, \infty)$ is continuous and satisfies $F(z) \le c_1 + c_2 z'z$ for some $c_2 < 1/4$ and all z in a closed set D. There is then a $\mu \in D$ at which*

$$F(\mu) - \tfrac{1}{2} \mu' \mu = \max_{z \in D} \{F(z) - \tfrac{1}{2} z' z\} \tag{2.13}$$

and $\epsilon \log \alpha(\epsilon) \to F(\mu) - \frac{1}{2}\mu'\mu$. *A necessary and sufficient condition for the* μ_ϵ*-IS estimator with* $\mu_\epsilon = \mu/\sqrt{\epsilon}$ *to be asymptotically optimal is that*

$$F(\mu) - \tfrac{1}{2}\mu'\mu \geq F(z) - \tfrac{1}{2}z'z + \tfrac{1}{4}(z-\mu)'(z-\mu), \tag{2.14}$$

for all $z \in D$. *Moreover, in this case the optimal* μ *in equation (2.13) is unique.*

Proof. Under the hypotheses of the theorem, $F(z) - \frac{1}{2}z'z$ is continuous and bounded so it attains a maximum on the closed set D; thus, μ in (2.13) exists. The convergence of $\epsilon \log \alpha(\epsilon)$ to $F(\mu) - \frac{1}{2}\mu'\mu$ follows from Lemma 2.2.

Next, we identify necessary and sufficient conditions for asymptotic optimality of any asymptotically unbiased estimator $\hat{\alpha}(\epsilon)$. Let $b(\epsilon) = E[\hat{\alpha}(\epsilon)] - \alpha(\epsilon)$. Using first Jensen's inequality and then the definition of $b(\epsilon)$ yields

$$\begin{aligned}
\log E[\hat{\alpha}^2(\epsilon)] &\geq 2 \log E[\hat{\alpha}(\epsilon)] \\
&= 2 \log(\alpha(\epsilon) + b(\epsilon)) \\
&= 2 \log(\alpha(\epsilon)) + 2 \log(1 + b(\epsilon)/\alpha(\epsilon)) \\
&= 2 \log(\alpha(\epsilon)) + O(b(\epsilon)/\alpha(\epsilon)),
\end{aligned}$$

with $b(\epsilon)/\alpha(\epsilon) \to 0$ by equation (2.11). Thus, from Lemma 2.2 we find that

$$\begin{aligned}
\liminf_{\epsilon \to 0} \epsilon \log E[\hat{\alpha}^2(\epsilon)] &\geq \liminf_{\epsilon \to 0} 2\epsilon \log(\alpha(\epsilon)) \\
&= 2 \liminf_{\epsilon \to 0} \epsilon \log E[\exp\{F(\sqrt{\epsilon}Z)/\epsilon\}\mathbf{1}_D] \\
&= 2 \sup_{z \in D}\{F(z) - \tfrac{1}{2}z'z\} \\
&= 2\{F(\mu) - \tfrac{1}{2}\mu'\mu\}.
\end{aligned}$$

Moreover, the degenerate estimator $\hat{\alpha}(\epsilon) = \alpha(\epsilon)$ achieves this rate, so an estimator $\hat{\alpha}(\epsilon)$ is asymptotically optimal if and only if

$$\limsup_{\epsilon \to 0} \epsilon \log E[\hat{\alpha}^2(\epsilon)] = 2[F(\mu) - \tfrac{1}{2}\mu'\mu].$$

The second moment of the $\mu/\sqrt{\epsilon}$-IS estimator is

$$m(\epsilon) = E\left[\exp\left\{\epsilon^{-1}\left(2F(\sqrt{\epsilon}Z) - \mu'\sqrt{\epsilon}Z + \tfrac{1}{2}\mu'\mu\right)\right\}\mathbf{1}_D\right].$$

Lemma 2.2 applied to the function $2F(z) - \mu'z + \frac{1}{2}\mu'\mu$ yields

$$\limsup_{\epsilon \to 0} \epsilon \log m(\epsilon) = \sup_{z \in D}\left\{2F(z) - \mu'z + \tfrac{1}{2}\mu'\mu - \tfrac{1}{2}z'z\right\}. \tag{2.15}$$

Thus, the $\mu/\sqrt{\epsilon}$-IS estimator is asymptotically optimal if and only if

$$\sup_{z\in D}\left\{2F(z)-\mu' z+\tfrac{1}{2}\mu'\mu-\tfrac{1}{2}z'z\right\}\le 2[F(\mu)-\tfrac{1}{2}\mu'\mu]; \tag{2.16}$$

that is, if and only if for all $z\in D$

$$2[F(z)-\tfrac{1}{2}z'z]+\tfrac{1}{2}z'z-\mu' z+\tfrac{1}{2}\mu'\mu\le 2[F(\mu)-\tfrac{1}{2}\mu'\mu],$$

which is equivalent to equation (2.14). The uniqueness of μ follows from the fact that the last (quadratic) term on the right-hand side of (2.14) is strictly positive at all $z\neq\mu$. □

REMARKS 2.2. (i) The condition in equation (2.14) is analogous to the *forbidden set* condition of Chen et al. (1993) in the rare event setting—that is, when $F\equiv 0$.

(ii) The first-order condition $\mu'=\nabla F(\mu)$ provides a further indication of the attractiveness of this importance sampling scheme without explicit appeal to asymptotics in ϵ. Under the zero-drift measure, the importance sampling estimator becomes

$$e^{F(\mu+Z)-\mu' Z-(1/2)\mu'\mu}\mathbf{1}_D.$$

A first-order Taylor approximation to the exponent suggests that this is

$$e^{F(\mu)+\nabla F(\mu)Z+O(Z'Z)-\mu' Z-(1/2)\mu'\mu}\mathbf{1}_D=e^{F(\mu)-(1/2)\mu'\mu}e^{O(Z'Z)}\mathbf{1}_D.$$

So, our choice of drift vector may be viewed as eliminating the variance contribution due to the linear part of F. Indeed, when F is exactly linear and $D=\mathbf{R}^n$, our importance sampling estimator has zero variance. In Section 4, we return to this perspective and show how to reduce the variance contribution arising from the quadratic part of F.

(iii) At the end of Section 2.1 we suggested two heuristics for choosing an effective change of drift: one based on minimizing the worst-case contribution to the variance, leading to equation (2.7), and one based on maximizing the product of payoff and probability, leading to equation (2.8). Assuming in each case that the maximum is attained in the interior of D and that F is differentiable there, the first-order conditions become

$$2\nabla F(z)-\mu'-z'=0,\qquad z'-\mu'=0$$

for equation (2.7) (equivalently, for the maximum over μ of equation (2.15)) and

$$\nabla F(\mu)=\mu' \tag{2.17}$$

for equation (2.8) (equivalently, (2.13)). Clearly, any solution to one provides a solution to the other, so the heuristics are consistent. Viewed another way, the optimization problem (2.13) finds the dominant term in the expansion for $\alpha(\epsilon)$, the quantity to be estimated. Thus,

if the optimum is unique, solving the first-order conditions (2.17) yields an approximation to $\alpha(\epsilon)$. Furthermore, because the first-order conditions to (asymptotically) minimize the second moment are equivalent to (2.17), the single set of equations $\nabla F(\mu) = \mu'$ finds both an approximation to $\alpha(\epsilon)$ and an asymptotically optimal importance sampling procedure.

3. COMPUTING THE OPTIMAL DRIFT

In this section, we consider various ways of finding the asymptotically optimal change of drift when one exists. We begin with a simple but convenient consequence of Theorem 2.1.

COROLLARY 3.1. *Let μ solve equation (2.13). Then $\mu/\sqrt{\epsilon}$ is asymptotically optimal if and only if μ also solves*

$$\max_{z \in D} \{F(z) - \tfrac{1}{2} z'z + \tfrac{1}{4}(z-\mu)'(z-\mu)\},$$

in which case the maximal values in the two optimization problems are equal.

Proof. The result follows from the condition identified in equation (2.16) and substitution of μ into the left side of (19). □

Under additional conditions on F and D, this leads to a characterization in which equation (2.14) is automatically satisfied and solving equation (2.13) suffices.

PROPOSITION 3.1. *Suppose $D = \{z : g(z) \leq 0\}$ for some convex function g, with $g(z) < 0$ for at least one z. Suppose that $F(z) - \frac{1}{4} z'z$ is concave on D. If μ solves equation (2.13), then $\mu/\sqrt{\epsilon}$ is asymptotically optimal.*

Proof. Under our assumption on F, the function $f(z) = \frac{1}{2} z'z - F(z)$ is convex. If μ solves equation (2.13), then μ solves the convex program $\min_{z \in D}\{\frac{1}{2} z'z - F(z)\}$. By the Kuhn–Tucker Theorem (Rockafeller 1970, Cor. 28.3.1) there exists a $\lambda \geq 0$ satisfying $\lambda g(\mu) = 0$ and

$$0 \in \partial f(\mu) + \lambda \partial g(\mu),$$

where ∂f and ∂g are the subdifferentials of f and g, respectively (cf. Rockafeller 1970, p. 215). The function

$$h(z) = \tfrac{1}{2} z'z - \tfrac{1}{4}(z-\mu)'(z-\mu) - F(z)$$

is also convex. Observe that

$$\partial h(\mu) + \lambda \partial g(\mu) = \partial f(\mu) + \lambda \partial g(\mu),$$

so this set contains 0. Thus, (μ, λ) satisfies the Kuhn–Tucker conditions for the convex program $\min_{z \in D} h(z)$. It follows that μ maximizes $-h$ over D. In light of Corollary 3.1, $\mu/\sqrt{\epsilon}$ is asymptotically optimal. □

The computation of the optimal drift vector for the case of an Asian option (as in equation (1.3)) is detailed in Section 5.1. For more complicated underlying assets and option payoffs we find candidate drift vectors through numerical optimization. Verifying global optimality is often difficult, but ultimately what matters is the variance reduction achieved, and this can be assessed directly. We give examples in Section 5.

An alternative approach is to view the first-order conditions $\nabla F(\mu) = \mu$ as a fixed point equation and to solve it iteratively. Our next result gives conditions supporting this method. For any matrix A, let $\|A\|$ denote the square-root of the sum of the squared entries of A.

PROPOSITION 3.2. *Suppose that F is twice continuously differentiable on D and that D is convex. Let $H(z)$ denote the Hessian of F at z.*

(i) *If ∇F maps D into D and $\sup_{z\in D}\|H(z)\| < 1$, then there is just one $\mu \in D$ satisfying $\mu = \nabla F(\mu)$, and the iterates $\mu_{n+1} = \nabla F(\mu_n)$ converge to it for any $\mu_0 \in D$.*

(ii) *If D^o contains a fixed point μ and $\|H(\mu)\| < 1$, then the iterates $\mu_{n+1} = \nabla F(\mu_n)$ converge to μ for all μ_0 in a neighborhood of μ.*

Proof. Both parts of the proposition follow from standard results on contraction mappings (see, e.g., Buck 1978, pp. 529–530). □

In several examples we find that fixed-point iteration converges to a near-optimal point in just four or five steps. However, we also find that the requirement that ∇F map D into D cannot be taken for granted, particularly when $0 \notin D$.

A simple refinement of the fixed-point iteration can dramatically accelerate its convergence. First note that we may rewrite the condition $\nabla F(\mu) = \mu'$ as $\nabla G(\mu)/G(\mu) = \mu'$. After i iterations, approximate $G(\mu)$ by $G(\mu_i) + \nabla G(\mu_i)(\mu - \mu_i)$ and $\nabla G(\mu)$ by $\nabla G(\mu_i)$, and let μ_{i+1} be the solution to

$$\mu_{i+1} = \frac{\nabla G(\mu_i)}{G(\mu_i) + \nabla G(\mu_i)(\mu_{i+1} - \mu_i)}. \tag{3.1}$$

The solution can be given explicitly. Let $\beta \equiv \beta(\mu_{i+1}) = 1/(G(\mu_i) + \nabla G(\mu_i)(\mu_{i+1} - \mu_i))$, so that

$$\mu_{i+1} = \beta \nabla G(\mu_i). \tag{3.2}$$

Substituting for μ_{i+1} given by equation (3.2) into the definition of β, we get the following quadratic equation for β:

$$\|\nabla G(\mu_i)\|^2 \beta^2 + (G(\mu_i) - \nabla G(\mu_i)\mu_i)\beta - 1 = 0.$$

We use the positive root (since β approximates $1/G(\mu)$) in equation (3.2) to get the solution vector μ_{i+1}. In our examples we find that when $G(0) > 0$ a single iteration starting from $\mu_0 = 0$ gives an excellent approximation to the optimal μ. Further theoretical support for this approximation and a numerical investigation are reported in Glasserman, Heidelberger, and Shahabuddin (1998).

4. OPTIMAL STRATIFICATION

We saw in Section 2.3 (see the remarks following Theorem 2.1) that asymptotically optimal importance sampling has the effect of eliminating variance due to the linear part of F, at least if the optimal μ satisfies the first-order condition $\mu' = \nabla F(\mu)$. By replacing Z with $Z + (\mu/\sqrt{\epsilon})$ we can write the $\mu/\sqrt{\epsilon}$-IS estimator under the original measure and apply a Taylor approximation to get

$$\begin{aligned} e^{\epsilon^{-1}[F(\sqrt{\epsilon}Z)-\mu'\sqrt{\epsilon}Z+(1/2)\mu'\mu]} \text{ (under } P_{\mu/\sqrt{\epsilon}}) &= e^{\epsilon^{-1}[F(\mu+\sqrt{\epsilon}Z)-\mu'\sqrt{\epsilon}Z-(1/2)\mu'\mu]} \text{ (under } P_0) \\ &\approx e^{\epsilon^{-1}[F(\mu)-\epsilon(1/2)Z'H(\mu)Z+o(\epsilon)-(1/2)\mu'\mu]} \\ &\approx e^{\epsilon^{-1}[F(\mu)-(1/2)\mu'\mu]}e^{(1/2)Z'H(\mu)Z+o(1)}, \end{aligned} \tag{4.1}$$

where $H(\mu)$ is the Hessian of F at μ and we have omitted the indicator $\mathbf{1}_D$ for simplicity. This suggests that to obtain further variance reduction we need to address the quadratic component of F. We do this by *stratifying* certain linear combinations of the Z_i; this is essentially the same as numerically integrating along a small number of key directions. In a simulation driven by arbitrary random vectors, stratifying on a linear combination would typically be impractical because of the difficulty of sampling from the distribution of the vector conditional on the linear combination; but in the Gaussian case the conditional distribution is itself Gaussian and this makes the approach practical.

4.1. Preliminaries on Stratification

We begin with a general description of stratifying $u'Z$ with u a vector in $\mathbf{R}^n$ and $Z \sim N(0, I_n)$. Since only the direction determined by u is relevant, we may take $u'u = 1$. In this case, $u'Z$ has a standard normal distribution on $\mathbf{R}$. Consider drawing a sample of $u'Z$ from this distribution and then sampling the vector Z conditional on the value of $u'Z$. This produces samples with the same distribution as drawing directly from $N(0, I_n)$ and has no advantage over direct sampling. If, however, we draw the samples of $u'Z$ in a more structured way than pure random sampling, the resulting samples of the vector Z will be more regularly distributed, at least along the direction determined by u.

For a stratified sample of size N, we partition $\mathbf{R}$ into N subsets $B_1, \ldots, B_N$. Ultimately, we will take these to be equiprobable intervals, but for the moment we keep the setting general. Let X_i have the distribution of $u'Z$ conditional on $u'Z \in B_i$, $i = 1, \ldots, N$. Assuming we can sample from this distribution, we then sample from the distribution of Z conditional on the value of $u'Z$. Using standard properties of the normal distribution (e.g., Johnson and Wichern 1982, p. 136) we find that

$$(Z|u'Z = a) \sim N(ua, I_n - uu'). \tag{4.2}$$

In particular, the conditional covariance matrix does not depend on a, which simplifies implementation.

Suppose now that B_i is the interval between the $(i-1)/N$th and i/Nth quantile of the

standard normal distribution. To sample from $u'Z$ conditional on $u'Z \in B_i$, first set

$$V_i = \frac{i-1}{N} + \frac{U_0^i}{N} \tag{4.3}$$

with U_0^i uniform on $[0, 1)$; this makes V_i uniform on $[(i-1)/N, i/N)$. Now set $X_i = \Phi^{-1}(V_i)$, with Φ the inverse of the cumulative normal distribution. (Algorithms for quickly and accurately evaluating this inverse are detailed in Marsaglia, Zaman, and Marsaglia (1994) and in Moro (1995).) This gives X_i the distribution of a standard normal variate conditioned to lie in the ith stratum B_i. Finally, set

$$\tilde{Z}^i = uX_i + C_u Y^i, \tag{4.4}$$

where $Y^i \sim N(0, I_n)$ (independent of X_i) and C_u is any $n \times n$ matrix satisfying $C_u C_u' = I_n - uu'$. In particular, we may choose $C_u = I_n - uu'$ because

$$C_u C_u' = (I_n - uu')(I_n - uu') = I_n - uu';$$

the matrix $I_n - uu'$ is symmetric and idempotent. This choice of C_u is especially convenient because equation (4.4) becomes

$$\tilde{Z}^i = uX_i + Y^i - u(u'Y^i).$$

This is a $O(n)$ calculation, whereas $O(n^2)$ operations would typically be required to calculate $C_u Y^i$ for other choices of C_u. We have found that taking $C_u = I_n - uu'$ substantially reduces the overhead required for stratified sampling, particularly when the dimension n is large. We refer to $\{\tilde{Z}^1, \ldots, \tilde{Z}^N\}$ (obtained using independent $Y^1, \ldots, Y^N$ and independent $U_0^1, \ldots, U_0^N$) as a stratified sample, stratified along direction u.

This construction easily generalizes if u is replaced with an $n \times k$ matrix U, $1 \le k \le n$, satisfying $U'U = I_k$. In this case, $U'Z \sim N(0, I_k)$, and $(Z|U'Z = a) \sim N(Ua, I - UU')$. To stratify in k dimensions, suppose $N = m^k$ and partition $[0, 1)^k$ into N bins by dividing each of the k coordinates into m intervals of equal width. Choose a point uniformly from each bin and apply Φ^{-1} to each coordinate to obtain a sample of $U'Z$. Since we want m to be reasonably large, this procedure is infeasible unless k is fairly small. (An alternative for large k is Latin hypercube sampling (McKay et al. 1979). This method generates k points in $[0, 1)^k$ stratified separately along each coordinate.)

By setting the jth component of Y in equation (4.4) equal to $\Phi^{-1}(U_j^i)$, with $U_1^i, \ldots, U_{n-1}^i$ independent and uniform on $[0, 1)$, we make the entire sampling procedure a function on $[0, 1)^n$. Indeed, if in equation (4.3) we were to replace U_0^i with $(2i-1)/2N$, we would be using the midpoint rule for numerical integration along one (or more generally k) of the n coordinates of $[0, 1)^n$ and Monte Carlo for the other coordinates. This observation is useful in assessing the variance reduction for a fixed choice of u or U and then identifying a particularly effective choice.

Consider, then, an arbitrary integrand $f : \mathbf{R}^n \to \mathbf{R}$ and the problem of estimating $c = E[f(Z)]$, assuming $E[f^2(Z)] < \infty$. For example, f could be the result of applying

importance sampling to $e^{F(Z)}$, as it will be in our application. If $Z^1, \ldots, Z^N$ are drawn independently from $N(0, I_n)$, then the standard Monte Carlo estimator has variance

$$\mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^{N} f(Z^i)\right] = \frac{\sigma_1^2}{N}$$

with $\sigma_1^2 = E[(f(Z) - c)^2]$. Let

$$\sigma_N^2 = \mathrm{Var}\left[\frac{1}{N}\sum_{i=1}^{N} f(\tilde{Z}^i)\right]$$

be the corresponding variance using a stratified sample obtained by stratifying along $u'Z$. Observe that stratification induces dependence among the $\tilde{Z}^i$. Define $g : \mathbf{R} \to \mathbf{R}$ by setting

$$g(x) = \mathrm{Var}[f(Z)|u'Z = x] \equiv \mathrm{Var}[f(X)], \qquad X \sim N(ux, I - uu'),$$

and let ϕ denote the standard univariate normal density. In the next result, we let N increase through a constant multiple of powers of two; accordingly, each stratification is a refinement of those that precede it.

LEMMA 4.1. *Suppose $E[f^2(Z)] < \infty$. For any u, $\sigma_N \le \sigma_1/\sqrt{N}$ with strict inequality unless the integral of f over each of the equiprobable strata is the same for all strata. Moreover,*

$$\sigma_N^2 = \frac{1}{N}\left(\int g(x)\phi(x)\,dx\right) + o(1/N) \tag{4.5}$$

as N increases through a constant multiple of powers of two.

Proof. Let $\eta_N(x)$ denote the index of the stratum containing x; that is, $x \in B_{\eta_N(x)}$. By standard results on stratified sampling (e.g., Hammersley and Handscomb 1964, p. 55) we have

$$N\sigma_N^2 = E[\mathrm{Var}[f(Z)|\eta_N(u'Z)]] \tag{4.6}$$

and

$$\sigma_1^2 = E[\mathrm{Var}[f(Z)|\eta_N(u'Z)]] + \mathrm{Var}[E[f(Z)|\eta_N(u'Z)]]. \tag{4.7}$$

This shows that $\sigma_N^2 \le \sigma_1^2/N$ and that strict inequality holds unless $E[f(Z)|\eta_N(u'Z)]$ is a constant.

Let $\xi = E[f(Z)|u'Z]$ and define the sequence $\xi_N = E[\xi|\eta_N(u'Z)]$. As N increases through a constant multiple of powers of two, the σ-algebras generated by $\eta_N(u'Z)$ form an increasing family and the ξ_N a martingale. Moreover, $E[|\xi|] < \infty$ so $\xi_N \to \xi$ (Karlin and Taylor 1975, p. 295). Jensen's inequality gives $\sup_N E[\xi_N^2] \le E[f^2(Z)] < \infty$, so

Theorem 5.2 of Karlin and Taylor (1975, p. 282) ensures that $E[\xi_N^2] \to E[\xi^2]$. Using equation (4.6) we now find that

$$\begin{aligned} N\sigma_N^2 &= E[\mathrm{Var}[\,f(Z)|\eta_N(u'Z)]] \\ &= E[\,f(Z)^2] - E[(E[\,f(Z)|\eta_N(u'Z)])^2] \\ &= E[\,f(Z)^2] - E[\xi_N^2] \\ &\to E[\,f(Z)^2] - E[\xi^2] \\ &= E[\,f(Z)^2] - E[(E[\,f(Z)|u'Z])^2] \\ &= E[\mathrm{Var}[\,f(Z)|u'Z]] = \int g(x)\phi(x)\,dx, \end{aligned}$$

as claimed. □

We use this result in the next section to select particularly effective directions for stratification.

4.2. Optimal Directions

In light of Lemma 4.1, an optimal direction u (at least for large N) can be selected by minimizing

$$\int \mathrm{Var}[\,f(Z)|u'Z = x]\phi(x)\,dx.$$

In practice, this is unlikely to be feasible so we consider approximations. Recall from equation (4.1) that we are particularly interested in reducing variance due to the quadratic component of F. This motivates an examination of the optimal stratification direction when F is exactly quadratic. Of course, if F were truly quadratic and D all of $\mathbf{R}^n$, simulation would be unnecessary. Nevertheless, we will see that the quadratic case provides useful guidance in selecting stratification directions for general F.

Consider, then, an $n \times n$ symmetric nonsingular matrix A with eigenvalues $\lambda_1, \ldots, \lambda_n$ all less than $1/2$. Let $f(z) = \exp(\frac{1}{2}z'Az)$. We want to solve

$$\min_{u:u'u=1} E[\mathrm{Var}[e^{(1/2)Z'AZ}|u'Z]]. \tag{4.8}$$

For each $i = 1, \ldots, n$, let v_i be an eigenvector associated with λ_i, normalized so that $v_i'v_i = 1$.

THEOREM 4.1. *If*

$$\left(\frac{\lambda_{j^*}}{1-\lambda_{j^*}}\right)^2 = \max_{i=1,\ldots,n}\left(\frac{\lambda_i}{1-\lambda_i}\right)^2, \tag{4.9}$$

then $u = v_{j^*}$ achieves the minimum in equation (4.8).

REMARK 4.1. The maximum in equation (4.9) is attained by the largest eigenvalue when all eigenvalues are positive, the smallest (most negative) eigenvalue when all are negative,

and is always either the smallest or the largest. Also, $[\lambda/(1-\lambda)]^2 = \lambda^2 + O(\lambda^3)$, so the eigenvalue of largest magnitude achieves the maximum when all the eigenvalues are close to zero.

Proof. See the Appendix.

Number the eigenvalues and eigenvectors of the matrix A so that

$$\left(\frac{\lambda_1}{1-\lambda_1}\right)^2 \geq \left(\frac{\lambda_2}{1-\lambda_2}\right)^2 \geq \cdots \geq \left(\frac{\lambda_n}{1-\lambda_n}\right)^2;$$

in particular, v_1 is optimal in the sense of Theorem 4.1. A simple corollary to Theorem 4.1 is that the optimal direction orthogonal to v_1 is v_2, the optimal direction orthogonal to both v_1 and v_2 is v_3, and so on. To see this, use the spectral representation of A to write

$$e^{(1/2)Z'AZ} = \prod_{i=1}^{n} e^{(\lambda_i/2)(v_i'Z)^2}$$

and notice that $v_1'Z, v_2'Z, \ldots, v_n'Z$ are independent standard normals. Once we condition on the value of $v_1'Z$, we are left with a problem of exactly the same form as the original problem, but involving only the last $n-1$ factors. Applying Theorem 4.1 to the reduced problem we find that v_2 is the optimal direction once we condition on $v_1'Z$, and so on.

As a benchmark of the impact of combined importance sampling and stratification, the next result gives explicit expressions for the variance in the case $F(z) = b'z + \frac{1}{2}z'Az$ for some $b \in \mathbf{R}^n$ and some symmetric $n \times n$ matrix A with eigenvalues less than $1/2$. Of course, in this setting simulation is unnecessary because

$$E[e^{b'Z+(1/2)Z'AZ}] = \frac{\exp\left(\frac{1}{2}b'(I-A)^{-1}b\right)}{|I-A|^{1/2}}, \tag{4.10}$$

where $|\cdot|$ gives the determinant of a matrix. The comparison is nevertheless informative.

PROPOSITION 4.1. (i) *The standard Monte Carlo estimator has variance*

$$e^{2b'(I-2A)^{-1}b}|I-2A|^{-1/2} - e^{b'(I-A)^{-1}b}|I-A|^{-1}.$$

(ii) *The variance minimizing drift vector is* $\mu = (I-A)^{-1}b$ *and the variance of the resulting importance sampling estimator is*

$$e^{b'(I-A)^{-1}b}\left(|I-2A|^{-1/2} - |I-A|^{-1}\right).$$

(iii) *Stratifying the importance sampling estimator along* $v_1, \ldots, v_k$, $1 \leq k \leq n$,

produces a limiting variance constant (in the sense of equation (4.5)) of

$$e^{b'(I-A)^{-1}b}\left(|I-2A|^{-1/2}-|I-A|^{-1}\prod_{i=1}^{k}\frac{1-\lambda_i}{\sqrt{1-2\lambda_i}}\right).$$

A proof is given in the Appendix. It is worth noting that the optimal drift vector μ in (ii) coincides with the asymptotic optimum obtained by maximizing $F(z)-\frac{1}{2}z'z$; this lends further support to the approach suggested by Theorem 4.1. Also, the expression in (iii) shows explicitly how each direction of stratification slices the limiting variance constant, eliminating it completely when $k=n$.

4.3. Computing Directions

Theorem 4.1 provides a strategy for identifying effective stratification directions, but the task of actually computing these directions still remains. We now consider various approaches.

The most straightforward approach evaluates the Hessian of F at the point μ used for importance sampling and then computes the eigenvalues and eigenvectors of this matrix. In simple examples, such as the Asian option in equation (1.3), second derivatives can be found explicitly and evaluated quickly. In more complicated examples it becomes necessary to evaluate them numerically, typically using finite difference approximations. This requires $O(n^2)$ evaluations of F, an effort comparable to simulating $O(n^2)$ paths. (The only savings in the function evaluations compared with simulation runs is that generating random draws of Z becomes unnecessary.) Whether or not this effort is justified depends on n and the total number of planned replications. If, say, $n=100$ then even very modest variance reduction would justify the effort if the precision required would otherwise entail 100,000 replications.

There are settings, however, in which time constraints severely limit the total computational effort that can be expended on pricing, regardless of the variance reduction achieved. In such settings, approximations become necessary. We consider two.

The first method looks for approximations to the eigenvectors of the Hessian in an m-dimensional subspace, $m<n$. We start by choosing an $n\times m$ matrix M whose columns seem likely to span a good approximation to the optimal eigenvector v_1. For example, if we believe v_1 should be approximately piecewise linear (as a function of the index $i=1,\ldots,n$) we could choose M to build an n-vector from an m-vector by linear interpolation. Let $\bar{z}$ denote an element of $\mathbf{R}^m$ and z an element of $\mathbf{R}^n$. The function $F_M(\bar{z})\triangleq F(M\bar{z})$ has Hessian $H_M\triangleq M'HM$, H the Hessian of F at $M\bar{z}$. Because H_M is $m\times m$, it may be much simpler to evaluate (through finite differences of F_M) than H. The next step is to find the eigenvalues $\gamma_1,\ldots,\gamma_m$ and eigenvectors $\bar{v}_1,\ldots,\bar{v}_m$ of the $m\times m$ matrix $(M'M)^{-1}M'HM$, ranked according to the criterion in equation (4.9). Our candidate stratification directions are then $M\bar{v}_1,\ldots,M\bar{v}_m$. This procedure is exact if F happens to depend on z only through $M'z$. Moreover, any eigenvector of H that lies in the range of M is recovered by this procedure along with its eigenvalue. For if $v_i=Mx$ for some x, then $(M'M)^{-1}M'HMx=\lambda_i x$, and x will be among the $\bar{v}_i$, up to a scalar multiple.

The second approximation we consider in selecting stratification directions is even simpler: stratify along $\tilde{\mu}'Z$, where $\tilde{\mu}=\mu/\sqrt{\mu'\mu}$ and μ is the optimal drift used for importance sampling. This choice is appropriate when F is well-approximated by a scalar function

of a linear combination of the Z_i. For if $F(z) = f(b'z)$ with $b \in \mathbf{R}^n$ and $f : \mathbf{R} \to \mathbf{R}$, then $\nabla F(z) = b' \dot{f}(b'z)$ (the dot indicating differentiation), implying that any solution to $\nabla F(\mu) = \mu$ is proportional to b; and then the Hessian of F at z is $bb' \ddot{f}(b'z)$. The only nontrivial eigenvectors of this matrix are the scalar multiples of b, so the optimal μ for the importance sample is also optimal for stratification, once normalized. Indeed, if $F(z) = f(b'z)$ exactly, then stratification along b eliminates all variance as the number of strata increases to infinity. We will see through examples that using μ can be surprisingly effective, even when the relation $F(z) = f(b'z)$ does not hold exactly.

5. EXAMPLES AND NUMERICAL RESULTS

We now illustrate the general results developed in previous sections through specific examples and numerical results.

5.1. Asian Option

Our first example—the arithmetic Asian option defined by equations (1.2) and (1.3)—is arguably the simplest option pricing problem for which simulation is necessary. (Numerical procedures are available for the problem with continuous averaging of the underlying but not with discrete averaging; see Geman and Yor 1993.) For simplicity, we assume that the dates t_i over which the average price of the underlying is computed are evenly spaced and let Δt denote this spacing. We begin with a discussion of the optimization problem developed in Section 2 as a step in the importance sampling procedure. The structure of this example allows for particularly efficient solution of the optimization problem.

We need to maximize $F(z) - \frac{1}{2}z'z$ with $F(z) = \log G(z)$ and G as defined in equation (1.3). The discount factor e^{-rT} has no effect on the optimal solution so we can just as well take $G(z) = [\bar{S} - K]^+$. It clearly suffices to consider points z at which $G(z) \neq 0$ and thus at which G and F are differentiable. The first-order conditions for optimality become

$$z_j = \frac{\sigma\sqrt{\Delta t}\sum_{i=j}^{n} S_i}{nG(z)}, \qquad j = 1, \ldots, n,$$

where we have written S_i for $S_{i\Delta t}$. This implies that

$$z_1 = \frac{\sigma\sqrt{\Delta t}[G(z) + K]}{G(z)}, \qquad z_{j+1} = z_j - \frac{\sigma\sqrt{\Delta t}S_j}{nG(z)}, \qquad j = 1, \ldots, n-1. \tag{5.1}$$

Given a value of $G(z) \equiv y$, equation (5.1) determines z together with

$$S_j = S_{j-1}e^{(r-(1/2)\sigma^2)\Delta t + \sigma\sqrt{\Delta t}z_j}, \qquad j = 1, \ldots, n. \tag{5.2}$$

Indeed, given $G(z) = y$, the values of $z_1, \ldots, z_n$ can be found very quickly starting by recursively applying equations (5.1) and (5.2). Subject to the first-order conditions, we may therefore view the S_i as functions of the scalar y rather than the vector z. The optimization problem thus reduces to finding the y that indeed produces a payoff of y at $S_1(y), \ldots, S_n(y)$;

that is, finding the root of the equation

$$g(y) \equiv \frac{1}{n}\sum_{j=1}^{n} S_i(y) - K - y = 0.$$

We do not have a proof that this equation has a unique root, but numerically this appears to be the case. Bisection finds the root very quickly, and given this scalar y, equations (5.1) and (5.2) recover z efficiently. We denote this vector by μ. Similar recursions apply if r and σ are time-varying (but deterministic) or if the averaging dates are unequally spaced.

Next, we turn to stratification. Second derivatives of $\log[\bar{S} - K]^+$ at any point at which $\bar{S} > K$ are easily expressed in closed form, so the matrix of second derivatives $H(\mu)$ at the point μ is easily computed. We can then find the eigenvalues and eigenvectors of this matrix, though we will see shortly that this is often unnecessary.

To get an indication of the potential effectiveness of stratification in this problem we examine the eigenvalues. Take $S_0 = K = 50$, $\sigma = 0.30$, $r = 0.05$, $T = 1$ (year), and $n = 64$. At the optimal μ, the eigenvalue of largest magnitude is -0.451 and all other eigenvalues have magnitudes smaller than 0.02, suggesting that this problem is a good candidate for stratification. As a further indication, we can examine

$$(5.3)\quad \left(|I - 2H(\mu)|^{-1/2} - |I - H(\mu)|^{-1} \prod_{i=1}^{k} \frac{1-\lambda_i}{\sqrt{1-2\lambda_i}} \right), \qquad k = 0, 1, \ldots, 64,$$

where $\lambda_1, \lambda_2, \ldots, \lambda_{64}$ are the eigenvalues of $H(\mu)$ ranked in decreasing order of $(\lambda_i/(1-\lambda_i))^2$. In view of Proposition 4.1, this shows how stratifying along progressively more directions would drive the variance to zero if F were exactly quadratic. (The values for $k = 1, \ldots, 8$ as a percentage of the value for $k = 0$ are listed in Table 5.3 in the column labeled "$\xi = 0$"; the other columns refer to the model of Section 5.2.) These suggest a dramatic reduction in variance from stratifying along the first eigenvector and negligible additional reduction from stratifying along a small number of additional directions. We find the same pattern across a wide range of parameter values for this model.

Conveniently, the heuristic suggested at the end of Section 4.3 is remarkably effective in this example. Figure 5.1 shows the optimal drift μ and optimal eigenvector $H(\mu)$ for the parameters above. The inner product between the two is 0.9993 when both are normalized to have unit length. We find the same pattern across a wide range of parameter values. This observation is important because it allows us to stratify along the optimal path μ without ever having to compute second derivatives or eigenvectors. Thus, this example provides an ideal set of circumstances for our approach: The optimization problem is easily solved, the Hessian appears to be well-suited to one-dimensional stratification, and we get an effective direction for stratification with no additional computation.

Numerical results in Table 5.1 confirm the effectiveness of the procedure for this problem. The table shows variance reduction ratios, relative to standard Monte Carlo, using importance sampling and importance sampling combined with stratification along either μ or the optimal eigenvector v_{j^*} identified in Theorem 4.1. Each variance reduction ratio is the variance per replication using standard Monte Carlo divided by the variance per replication using the method indicated at the top of the column. The larger the ratio, the greater the improvement. Throughout this section, the number of runs is large enough to provide accurate

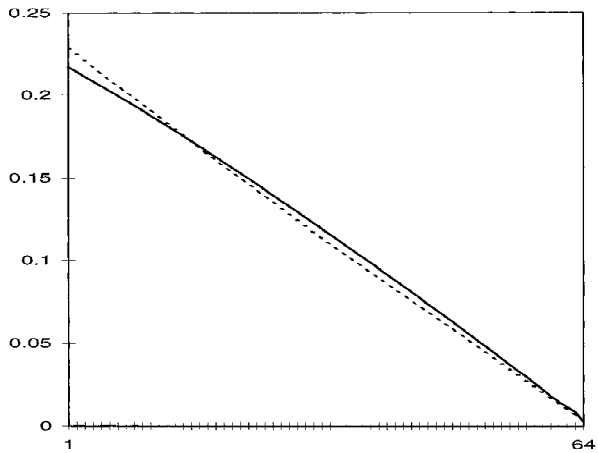

FIGURE 5.1. Optimal drift vector (solid line) and eigenvector (dashed line) for Asian option. The eigenvector has been scaled to have the same norm as the optimal μ. The parameter values are $S_0 = K = 50$, $\sigma = 0.30$, $T = 1$, $r = 0.05$, $n = 64$.

estimates of both the price and its variance; for example, 99 percent confidence intervals for the price are typically within ± 0.01. The results show that, by itself, importance sampling provides moderate variance reduction, but when combined with stratification the effect is astounding. In all our examples the additional computing time per replication required to simulate using importance sampling and stratification is only about 5 percent of the time required to simulate without variance reduction—negligible when compared with the variance reduction achieved. The modest additional effort required to solve the optimization problem is not reflected in the variance reduction ratios, nor is the effort required to find the optimal eigenvector required for the results in the last column. These are fixed costs, so their effect depends on the number of paths generated: the greater the precision required the less significant the time spent on the optimization or the eigenvector computation.

In all our examples with stratified sampling, we use an equal number of replications in each stratum. For equiprobable strata (as is the case here), the optimal allocation of replications to a stratum is proportional to the standard deviation of that stratum. By using estimates of the standard deviations, we can assess the potential of attempting to further improve the procedure by allocating replications to strata in a (near) optimal fashion. Although additional significant potential improvements (e.g., 50 percent) can sometimes be achieved, the typical potential improvement is much less. Furthermore, the best potential improvements usually occurred when the combination of importance sampling and stratification was already highly efficient.

We now modify the example by adding a knock-out provision at expiration: If at expiration the price of the underlying asset is below a barrier B, the option pays $(\bar{S} - K)^+$ (as with an ordinary Asian option), but if the final price is above B the option pays nothing. Since the price of the underlying at expiration is S_n (with n the number of simulation steps to expiration), the payoff becomes $(\bar{S} - K)^+ \mathbf{1}_{\{S_n \le B\}}$. The optimal path is still easily computed if the constraint $S_n < B$ is appended with a Lagrange multiplier. When the constraint is binding, the underlying terminates exactly at the barrier when evaluated at the optimal path;

TABLE 5.1
Estimated Variance Reduction Ratios for the Asian Option[a]

Parameters				Importance sampling	Importance sampling and stratification	
n	σ	K	Price	Variance ratio	Variance ratio (μ)	Variance ratio (v_{j^*})
16	0.10	45	6.05	11	1,097	1,246
		50	1.92	7.0	4,559	5,710
		55	0.20	21	15,520	17,026
16	0.30	45	7.15	8.3	1,011	1,664
		50	4.17	9.2	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	7.3	4,637	5,665
		55	0.17	23	16,051	17,841
64	0.30	45	7.02	8.3	1,016	1,694
		50	4.02	9.2	1,319	1,971
		55	2.08	12	1,767	2,402

[a] All results are based on a total of 1,000,000 runs. Stratified results use 100 strata. All results use $S_0 = 50$, $r = 0.05$, and $T = 1.0$. Of the last two columns, the first shows results based on stratifying along the optimal μ and the second shows results based on stratifying along the optimal eigenvector.

otherwise, the optimal path is the same as in the absence of the barrier. One would expect the presence of the barrier to reduce the effectiveness of our procedure because the linear and quadratic approximations that motivate our approach seem less likely to apply in the presence of a discontinuity in the payoff. If B is large, reaching B is a rare event and one might expect that importance sampling is particularly effective in pricing the corresponding knock-*in* option with payoff $(\bar{S} - K)^+ \mathbf{1}_{\{S_n > B\}}$.

These expectations are borne out by the numerical results in Table 5.2. Importance sampling by itself is less effective in the knock-out examples than in the absence of a barrier, but it can dramatically reduce variance in pricing the knock-ins when the barrier is far from the underlying. (The knock-in price with $K = 50$, $B = 80$, and $\sigma = 0.30$ is only 0.00016 and so is perhaps too small to be of much practical interest, but these examples illustrate the potential of the method.) The stratified results stratify along the optimal μ. The impact of stratification varies; it is more effective for knock-ins than knock-outs and in both cases its effectiveness increases as the barrier becomes more remote.

The numerical results in Table 5.2 suggest a possible indirect method for pricing knock-outs: subtract the price of a knock-in from the price of the corresponding option without a barrier, estimating the two parts in separate simulations using a change of drift and stratification direction tailored for each. This is sometimes advantageous. Furthermore, because the underlying asset in this example is modeled as geometric Brownian motion, it is possible to simulate paths conditional on $S_n \leq B$ or conditional on $S_n > B$ and thus to eliminate the payoff discontinuity caused by the barrier. Because such an approach is applicable only in the simplest models, we do not discuss it further.

TABLE 5.2
Estimated Variance Reduction Ratios for the Asian Option with a Knock-out or a Knock-in Barrier[a]

Parameters			Importance sampling		Importance sampling and stratification	
K	σ	B	Knock-out var. ratio	Knock-in var. ratio	Knock-out var. ratio	Knock-in var. ratio
50	0.10	60	2.4	8.4	6.1	25
		70	6.4	198	240	992
		80	7.0	29,327	3,864	195,055
50	0.30	60	2.0	6.2	2.4	14
		70	2.0	7.0	4.1	16
		80	3.1	12	8.9	34
		100	6.0	46	46	167
55	0.10	60	6.5	19	9.1	43
		70	17	232	351	787
		80	21	30,701	12,988	154,406
55	0.30	60	3.9	10	4.4	41
		70	2.8	9.1	4.5	18
		80	3.8	14	9.2	34
		100	7.5	49	51	157

[a]Results are based on a total of 1,000,000 runs with $S_0 = 50$, $r = 0.05$, $T = 1.0$, and $n = 16$. Stratified results use 100 strata and stratify in direction μ.

5.2. Stochastic Volatility

For pricing path-dependent options in a stochastic volatility model, simulation is typically the only method available. The particular specification we consider is the Hull–White (1987) model,

$$\begin{aligned} dS_t &= rS_t\,dt + \sqrt{V}_t S_t\,dW_t^{(1)} \\ dV_t &= \nu V_t\,dt + \xi V_t dW_t^{(2)}, \end{aligned}$$

with $E[dW^{(1)}dW^{(2)}] = \rho$. We work with the following discrete-time counterpart:

$$\begin{aligned} S_{i+1} &= S_i(1 + r\Delta t + \sqrt{V_i \Delta t}\, Z_i) \\ V_{i+1} &= \min\{m,\, V_i e^{(\nu-(1/2)\xi^2)\Delta t + \xi\sqrt{\Delta t}(\rho Z_i + \sqrt{1-\rho^2} Z_{n+i})}\}, \end{aligned}$$

with m a constant. With $m = \infty$ this would be an exact discretization of the variance process V_t but only one of many possible approximations to the solution of the equation for S_t. In the continuous-time model, S_t has finite mean but infinite variance. Under the linear discretization above, the variance is finite but increases very quickly with the number of steps. Truncating the V_i at m helps reduce this effect. (In our simulation results we take $m = 2$.) To be consistent with our general framework, we have taken the stochastic input

TABLE 5.3
Remaining Variance (as a percentage of total) after Stratification in k Dimensions, as Measured by Equation (5.3)[a]

k	$\xi = 0$	$\xi = 0.5$	$\xi = 1$	$\xi = 2$	$\xi = 3$
1	5.1	10.3	13.8	22.4	30.9
2	4.7	8.8	11.9	19.1	25.5
3	4.3	7.5	10.2	16.2	20.7
4	4.0	6.4	8.7	13.6	16.6
5	3.6	5.4	7.3	11.3	13.1
6	3.3	4.5	6.1	9.3	10.3
7	3.0	3.7	5.1	7.6	8.0
8	2.8	3.0	4.2	6.2	6.2
64	0	0	0	0	0

[a] In order to make all cases 64-dimensional problems, the constant volatility case $\xi = 0$ is based on $\Delta t = 1/64$ whereas the others are based on $\Delta t = 1/32$.

to the model to be a single vector $(Z_1, \ldots, Z_{2n})$ of independent standard normals, though in some respects it might be more natural to think of two separate vectors each of length n.

The payoff we consider is again that of a call option on the arithmetic mean, $(\bar{S} - K)^+$, with the mean $\bar{S}$ computed over n dates spaced $\Delta t = T/n$ time units apart. For this problem, we find the optimal μ using general-purpose optimization code. The quasi-Newton routine e04jbc included in the NAG library (1996) solves the problem extremely quickly; on a Sun Enterprise 4000 workstation running the Solaris 2.51 operating system all our optimization problems took less than one second and many took less than one-tenth of a second. The solver included in the Excel spreadsheet handles this problem well too. For parameter values that result in a positive payoff at $z = 0$, the fixed-point iteration discussed in Section 3 appears to converge quickly, and the approximation in equation (3.2) is very close to the optimum.

Throughout our examples we take $\nu = 0$, $r = 0.05$, $S_0 = 50$, $T = 1$, and truncate at $m = 2$. The truncation has little impact on the mean but makes estimated variances much more stable. Figures 5.2 and 5.3 show, respectively, the optimal μ and the path of S and V evaluated at the optimal μ, all with the parameters $K = 50$, $\sqrt{V_0} = 0.30$, $\rho = 0.5$, $n = 32$, and $\xi = 2$. (Hull and White (1987) suggest that $\xi = 1$ may be more realistic; we use a larger value in the figures to get a sharper contrast with the constant volatility case.) Figure 5.2 displays the optimal μ (the solid line) in two pieces corresponding to the first 32 and second 32 Z_i in the recursions above. The figures indicate that the product of probability and payoff is maximized by a trajectory in which volatility first increases and then decreases, driving the price of the underlying to climb first sharply and then gradually to achieve a large value of $\bar{S}$.

For the same parameter values, Figure 5.2 shows the optimal eigenvector (the dashed line) for the Hessian evaluated at the optimal μ. The eigenvector has been scaled to have the same total length (the same sum of squares) as μ. The first 32 values are similar to those of the optimal μ itself and the second 32 values are nearly the mirror image of the corresponding entries of μ. (Of course, the sign of the eigenvector is indeterminate so the two halves can be interchanged in this statement. If we change the sign of the last 32 values of the eigenvector we must also change the sign of the first 32 values so the two halves in the figure cannot be simultaneously aligned.)

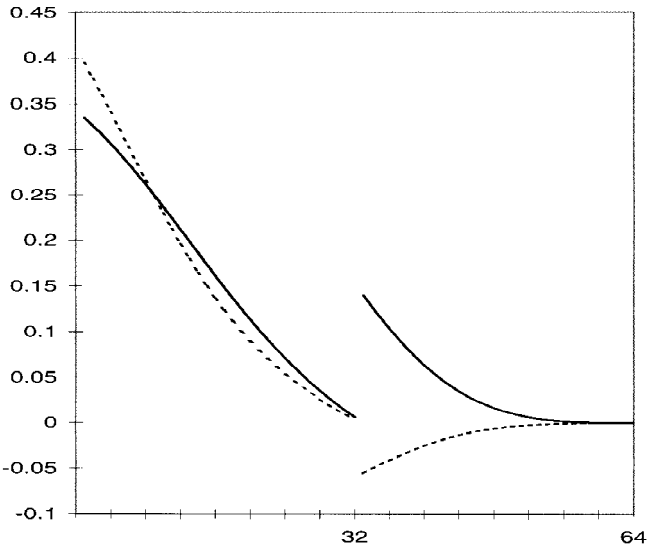

FIGURE 5.2. Optimal drift vector (solid line) and eigenvector (dashed line) for the Asian option in the Hull–White stochastic volatility model. The parameter values are $S_0 = K = 50$, $T = 1$, $r = 0.05$, $\nu = 0$, $\xi = 2$, $\sqrt{V_0} = 0.30$, $n = 32$. In each of the vectors in the figure, both the ith and the $(32 + i)$th coordinates are associated with the ith increment of the process.

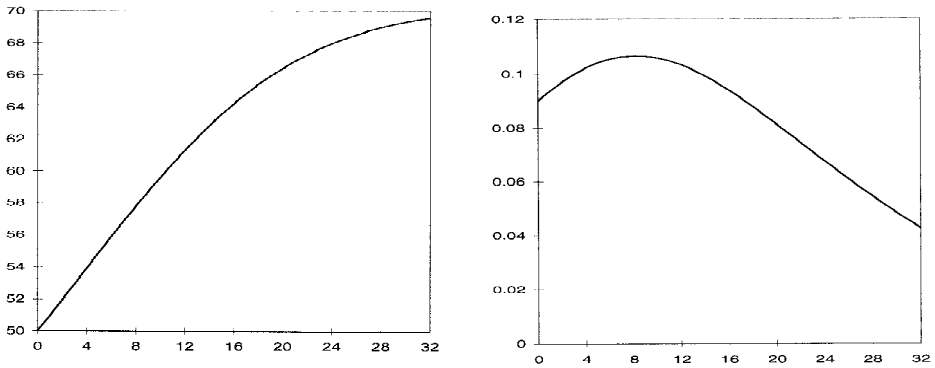

FIGURE 5.3. Path of underlying price S (left panel) and conditional variance V (right panel) evaluated at $z = \mu$ (the optimal drift vector) for the Asian option in the stochastic volatility model. The parameter values are $S_0 = K = 50$, $T = 1$, $r = 0.05$, $\nu = 0$, $\xi = 2$, $\sqrt{V_0} = 0.30$, $n = 32$.

Table 5.3 displays values of equation (5.3) for this example for various values of ξ. Whereas the values for constant volatility ($\xi = 0$) drop very quickly, with $\xi = 3$ the decay is gradual; indeed, the table suggests that it would take stratification in more than eight dimensions when $\xi = 3$ to achieve the same variance reduction as stratification in a

TABLE 5.4
Estimated Variance Reduction Ratios for the Hull–White Stochastic Volatility Model[a]

Parameters				Importance sampling	Importance sampling and stratification	
K	T	ξ	Price	Variance ratio	Variance ratio (μ)	Variance ratio (v_{j^*})
50	0.25	0.0	1.92	8.3	3,864	4,955
		0.5	1.91	8.3	307	431
		1.0	1.91	8.5	103	147
		2.0	1.88	8.8	34	46
50	1.00	0.0	4.07	9.3	1,355	1,947
		0.5	4.05	9.2	84	123
		1.0	4.00	9.7	31	43
		2.0	3.81	9.9	14	18
55	0.25	0.0	0.39	19	8,304	9,447
		0.5	0.41	19	588	673
		1.0	0.44	19	204	235
		2.0	0.47	20	73	82
55	1.00	0.0	2.12	12	1,827	2,363
		0.5	2.17	12	113	145
		1.0	2.19	13	43	55
		2.0	2.12	15	22	25

[a] All results are based on a total of 1,000,000 runs. Stratified results use 100 strata. All results use $S_0 = 50$, $V_0 = 0.09$, $r = 0.05$, $\nu = 0$, $\rho = 0.5$, and $n = 32$. For $\xi > 0$, V_i is capped at 2.0. Results for $\xi = 0$ correspond to constant volatility.

single dimension when $\xi = 0$. This comparison is merely suggestive, however, because equation (5.3) is an exact measure only in the quadratic case.

Numerical results illustrating the actual variance reduction appear in Table 5.4. By itself, importance sampling reduces variance by a factor of 8–20; in combination with stratification (along either μ or the optimal eigenvector v_{j^*}) it remains very effective, particularly for small values of ξ. The diminishing effectiveness with increasing ξ is consistent with what would be predicted by looking at the eigenvalues of the Hessian and what might even be predicted from the specification of (S_t, V_t): this model is far less linear than the constant volatility model.

5.3. CIR Model

Our final example is the interest rate model of Cox, Ingersoll, and Ross (1985):

$$dr_t = \kappa(a - r_t)\,dt + \sigma\sqrt{r_t}\,dW_t. \tag{5.4}$$

As noted by Rogers (1995), when $d \equiv 4a\kappa/\sigma^2$ is an integer, $\{r_t, t \geq 0\}$ has the same law as $\{\|X_t\|^2, t \geq 0\}$, where X_t is the d-dimensional process defined by

$$dX_t = -\frac{\kappa}{2} X_t \, dt + \frac{\sigma}{2} \, dB_t,$$

where B_t is a standard d-dimensional Wiener process, and the components of X_0 are all equal to $\sqrt{r_0/d}$. This equation has solution

$$X_t = e^{-(1/2)\kappa t} X_0 + \frac{\sigma}{2} \int_0^t e^{-(1/2)\kappa(t-s)} \, dB_s.$$

On a discrete grid of points $j\Delta t$, $j = 0, 1, \ldots$, the ith coordinate of this process can be simulated without discretization error by setting

$$X^{(i)}_{(j+1)\Delta t} = e^{-(1/2)\kappa \Delta t} X^{(i)}_{j\Delta t} + \frac{\sigma}{2} \sqrt{\frac{1}{\kappa}\left(1 - e^{-\kappa \Delta t}\right)} Z_{(i-1)n+j}, \qquad j = 0, 1, \ldots, n-1,$$

where $Z_1, \ldots, Z_{dn}$ are independent standard normals. An alternative (applicable even if d is not an integer) discretizes equation (5.4) directly. Our method applies as well to a discretization of (5.4); we have chosen to use the exact procedure available when d is an integer solely to separate the examination of variance reduction from discretization bias unrelated to our method.

The price of a discount bond with maturity T is given by

$$E\left[\exp\left(-\int_0^T r_u \, du\right)\right].$$

This expectation is available in closed form in the CIR model but it is still useful as a numerical illustration. We set $\Delta t = T/(n+1)$ and approximate the integral inside the expectation by a sum

$$\exp\left(-\Delta t \sum_{i=0}^{n} r_{i\Delta t}\right),$$

with the $r_{i\Delta t}$ simulated as described above. With d restricted to integer values, the dimension of this problem is nd.

The parameter values we consider are $a = 0.064$, $\kappa = 0.05$, $\sigma = 0.08$ (hence, $d = 2$), and various values of r_0 and T. In all cases the optimal μ is nearly (if not exactly) linear, increasing through negative values and thereby driving $r_{i\Delta}$ to decrease. The optimal μ is the same for $X^{(1)}$ and $X^{(2)}$. The eigenvalues of the Hessian at μ consistently appear with multiplicity two. The optimal μ is nearly contained in the eigenspace associated with the eigenvalue maximizing $(\lambda/(1-\lambda))^2$; more precisely, the norm of its projection onto the eigenspace is 99.6 percent of the norm of μ itself.

Numerical results for this example are given in Table 5.5. Here we find that importance sampling by itself can produce more than a hundredfold reduction in variance, and that

TABLE 5.5
Estimated Variance Reduction Ratios for Bond Prices in the CIR Model[a]

Parameters			Importance sampling	Importance sampling and stratification	
T	r_0	Price	Variance ratio	Variance ratio(μ)	Variance ratio(ν_{j^*})
0.25	0.044	98.90	227	397	411
	0.064	98.41	330	576	597
	0.084	97.92	433	755	771
1.0	0.044	95.66	57.2	104	108
	0.064	93.81	82.7	150	155
	0.084	91.99	108	197	203

[a] Prices are for face value of 100. All results are based on a total of 1,000,000 runs, with $d = 2$, $\kappa = 0.05$, $\sigma = 0.08$, and $n = 16$. Stratified results use 100 strata.

stratification further reduces variance by roughly a factor of two. There is virtually no difference in performance between stratifying along the optimal eigenvector and stratifying along μ. The effectiveness of importance sampling in this case is likely due to the extent of linearity in the model.

To make the model less linear, we introduce some optionality. We consider an interest rate cap struck at K paying $(r_{i\Delta t} - K)^+$ at time $(i+1)\Delta t$, $i = 0, 1, \ldots, n$. The total discounted payoff is

$$\sum_{i=1}^{n} e^{-\Delta t \sum_{j=0}^{i} r_{j\Delta t}} (r_{i\Delta t} - K)^+.$$

This formulation is slightly nonstandard in that it blurs the distinction between discrete and continuous compounding, but it is nevertheless illustrative.

Like our previous examples, this one shows a great deal of similarity between the optimal drift for importance sampling and the optimal eigenvector for stratification. An example appears in the left panel of Figure 5.4; when μ is normalized to have unit length (like the eigenvector), the two are nearly indistinguishable. So we use the optimal μ for both importance sampling and stratification. Numerical results appear in Table 5.6. By itself, importance sampling is less effective for the cap than for the bond (as expected) and stratification generally has greater impact than before. The combination reduces variance by factors in the range of 30–200. The effectiveness of the two stratification directions is nearly identical.

6. CONCLUDING REMARKS

The framework we have developed in this paper—using importance sampling to eliminate the variance from the linear part of the log-payoff, and using stratification to reduce variance from the quadratic part—lends itself to many more variations than we have investigated here. For example, further approximations to the optimal drift and stratification directions

TABLE 5.6
Estimated Variance Reduction Ratios for Cap Prices in the CIR Model[a]

Parameters			Importance sampling	Importance sampling and stratification	
T	K	Price	Variance ratio	Variance ratio (μ)	Variance ratio (ν_{j^*})
0.25	0.064	4.29	11	52	53
	0.074	0.55	22	63	62
	0.084	0.05	81	199	212
1.0	0.064	8.17	11	38	40
	0.074	3.13	14	38	38
	0.084	1.11	20	49	48

[a]Prices are for face value of 100. All results are based on a total of 1,000,000 runs. Stratified results use 100 strata. All results use $d = 2$, $\kappa = 0.05$, $\sigma = 0.08$, $r_0 = 0.064$, and $n = 16$.

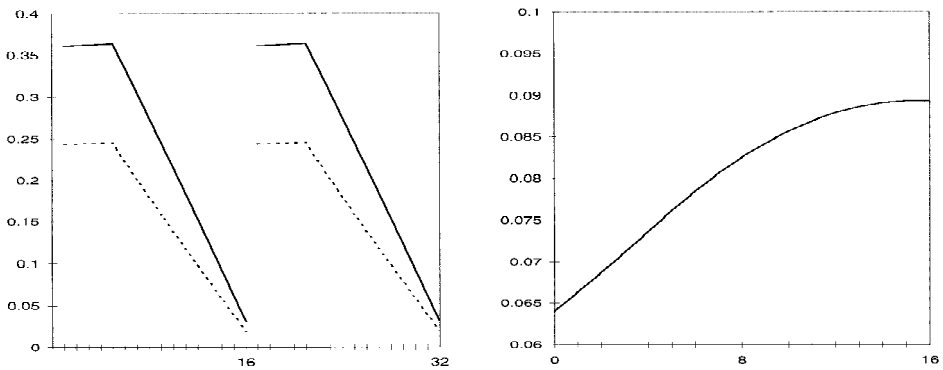

FIGURE 5.4. The left panel shows the optimal drift vector (solid line) and eigenvector (dashed line) for cap in the CIR interest rate model. The right panel shows the path of the interest rate evaluated at the optimal drift. The parameter values are $r_0 = a = 0.064$, $\kappa = 0.08$, $\sigma = 0.05$, $K = 0.074$, $T = 1$, and $n = 16$. In each of the vectors in the left panel, both the ith and the $(16+i)$th coordinates are associated with the ith increment of the interest rate path in the right panel. The optimal drift increases until the cap is in-the-money and then decreases.

could be explored. The optimization problem could be re-solved (exactly or approximately) at various points along a path. Stratification along multiple directions could naturally be combined with quasi-Monte Carlo techniques. Indeed, the ranking provided by Theorem 4.1 seems likely to be useful in other applications of quasi-Monte Carlo and perhaps even with other numerical techniques for pricing path-dependent options.

APPENDIX

This appendix contains proofs of Theorem 4.1 and Proposition 4.1.

Proof of Theorem 4.1. By a standard decomposition of variance, we can rewrite equation (4.8) as

$$\begin{aligned}\mathrm{Var}[e^{(1/2)Z'AZ}] - \mathrm{Var}[E[e^{(1/2)Z'AZ}|u'Z]] \;&=\; \mathrm{Var}[e^{(1/2)Z'AZ}] - E[(E[e^{(1/2)Z'AZ}|u'Z])^2] \\ &\quad + (E[e^{(1/2)Z'AZ}])^2;\end{aligned}$$

hence, (4.8) is equivalent to

$$\max_{u:u'u=1} E[(E[e^{(1/2)Z'AZ}|u'Z])^2]. \tag{A.1}$$

We proceed by recording three useful identities. For any positive definite $m \times m$ matrix M and any m-vector w

$$\frac{1}{(2\pi)^{m/2}} \int_{R^m} e^{-(1/2)x'Mx+w'x}\, dx = |M|^{-1/2} e^{(1/2)w'M^{-1}w}, \tag{A.2}$$

where $|M|$ denotes the determinant of M. For any M partitioned as

$$M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \tag{A.3}$$

with M_{22} nonsingular,

$$|M| = |M_{22}||M_{11} - M_{12}M_{22}^{-1}M_{21}|. \tag{A.4}$$

If $\xi \sim N(0, 1)$, then

$$E[e^{t\xi^2}] = \frac{1}{\sqrt{1-2t}}, \qquad t < 1/2. \tag{A.5}$$

Recognize equation (A.2) as a property of the moment generating function of the normal distribution with covariance matrix M^{-1}; equation (A.4) is standard (see, e.g., Rao and Toutenburg 1995, p. 289); equation (A.5) is the moment generating function of a chi-square random variable with 1 degree of freedom.

Conditional on $u'Z = a$, the vector Z has distribution $N(au, I - uu')$. Consider first the special case $u = e_1 = (1, 0, \ldots, 0)'$, and let X have distribution $N(ae_1, I - e_1e_1')$. Partition A the way M is partitioned in equation (A.3) with A_{11} scalar. The conditional expectation in equation (A.1) is given by

$$\begin{aligned}E[e^{(1/2)X'AX}] \;&=\; E[e^{(1/2)(ae_1+(X-ae_1))'A((X-ae_1)+ae_1)}] \\ &=\; e^{(1/2)a^2A_{11}} E[e^{(1/2)[2ae_1'A(X-ae_1)+(X-ae_1)'A(X-ae_1)]}]\end{aligned}$$

$$= e^{(1/2)a^2A_{11}} \frac{1}{(2\pi)^{\frac{n-1}{2}}} \int_{R^{n-1}} e^{-(1/2)x'x} \cdot e^{(1/2)(2aA_{12}x + x'A_{22}x)}\, dx$$
$$= e^{(1/2)a^2A_{11}} \frac{1}{(2\pi)^{\frac{n-1}{2}}} \int e^{-(1/2)x'(I-A_{22})x} \cdot e^{aA_{12}x}\, dx.$$

To evaluate this expression we may invoke equation (A.2), provided $I - A_{22}$ is invertible. For any unit length $(n-1)$-vector x,

$$x'A_{22}x = (0 \;\; x')A(0 \;\; x')' \leq \max_{i=1,\ldots,n} \lambda_i < 1/2,$$

so all eigenvalues of A_{22} are less than $1/2$, ensuring that $I - A_{22}$ is invertible. Applying (A.2) and noting that the symmetry of A implies $A_{21} = A'_{12}$, we get

$$E[e^{(1/2)X'AX}] = \frac{e^{(1/2)a^2A_{11}}}{|I - A_{22}|^{1/2}} e^{\frac{a^2}{2}A_{12}(I-A_{22})^{-1}A_{21}}.$$

Setting $M = I - A$ in equation (A.4) yields

$$|I - A| = |I - A_{22}| \cdot (1 - A_{11} - A_{12}(I - A_{22})^{-1}A_{21}),$$

and therefore

$$A_{11} + A_{12}(I - A_{22})^{-1}A_{21} = 1 - \frac{|I - A|}{|I - A_{22}|}.$$

We conclude that, with $u = e_1$,

$$E[e^{(1/2)X'AX}] = |I - A_{22}|^{-1/2} \exp\left[\frac{a^2}{2}\left(1 - \frac{|I - A|}{|I - A_{22}|}\right)\right].$$

For general u, let U be any orthogonal matrix with $Ue_1 = u$ and set $G = U'AU$. Observe that if $X \sim N(au, I - uu')$, then

$$X'AX = X'(UU')A(UU')X = (U'X)'G(U'X), \quad U'X \sim N(ae_1, I - e_1e_1').$$

The calculation above for $u = e_1$ thus applies to general u if we replace A with G:

$$E[e^{(1/2)X'AX}] = |I - G_{22}|^{-1/2} \exp\left[\frac{a^2}{2}\left(1 - \frac{|I - G|}{|I - G_{22}|}\right)\right].$$

Recalling now that $a = u'Z$, we get

$$E[(E[e^{(1/2)X'AX}])^2] = |I - G_{22}|^{-1/2} E\left[\exp\left\{(u'Z)^2\left(1 - \frac{|I - G|}{|I - G_{22}|}\right)\right\}\right]$$

$$= |I - G_{22}|^{-1} \left(2\frac{|I-G|}{|I-G_{22}|} - 1 \right)^{-1/2},$$

using equation (A.5) and the fact that $u'Z \sim N(0,1)$. We need to maximize this expression over u. Because $G = U'AU$, G and A have the same eigenvalues, so $|I-G| = \prod_{i=1}^{n}(1-\lambda_i)$. This is independent of u so we may equivalently solve

$$\max_{u:u'u=1} \left(\frac{|I-G_{22}|}{|I-G|} - 1 \right)^2. \tag{A.6}$$

Let $\beta_2, \ldots, \beta_n$ be the eigenvalues of G_{22}, so that $|I - G_{22}| = \prod_{i=2}^{n}(1-\beta_i)$. Now we make the following claim.

Claim: Every $(\beta_2, \ldots, \beta_n)$ arising from some choice of u is contained within the convex hull of the points $(\lambda_{p(1)}, \ldots, \lambda_{p(n-1)})$ generated by all permutations p of the indices $1, \ldots, n$.

We defer the proof of the claim until the end. Replace equation (A.6) with

$$\max_{\beta_2,\ldots,\beta_n} \left(\frac{\prod_{i=2}^{n}(1-\beta_i)}{\prod_{i=1}^{n}(1-\lambda_i)} - 1 \right)^2$$

under the ostensibly weaker constraint that $(\beta_2, \ldots, \beta_n)$ belong to the convex hull described in the claim. Since this is a problem of maximizing a (symmetric) convex function over a convex set, the maximum is attained at an extreme point of the form of $(\lambda_1, \ldots, \lambda_{j-1}, \lambda_{j+1}, \ldots, \lambda_n)$, for some $j = 1, \ldots, n$. In fact, every such point corresponds to a feasible solution for the original problem (A.6) with $u = v_j$: let V be an orthogonal matrix of eigenvectors of A with v_j as first column; then the eigenvalues of $G_{22} = (V'AV)_{22}$ are precisely the λ_i, $i \neq j$. The solution to (A.6) is therefore given by

$$\max_{u=v_j,\, j=1,\ldots,n} \left(\frac{|I-G_{22}|}{|I-G|} - 1 \right)^2 = \max_{j=1,\ldots,n} \left(\frac{1}{1-\lambda_j} - 1 \right)^2 = \max_{j=1,\ldots,n} \left(\frac{\lambda_j}{1-\lambda_j} \right)^2.$$

It remains to prove the claim. First we argue that $\beta_2, \ldots, \beta_n$ are the nonzero eigenvalues of $(I - uu')A(I - uu')$. By definition, they are the eigenvalues of $G_{22} = (U'AU)_{22}$ and therefore also the nonzero eigenvalues of

$$\begin{pmatrix} 0 & 0 \\ 0 & (U'AU)_{22} \end{pmatrix} = (I - e_1e_1')U'AU(I - e_1e_1').$$

But

$$(I-e_1e_1')U'AU(I-e_1e_1') = U'U(I-e_1e_1')U'AU(I-e_1e_1')U'U = U'(I-uu')A(I-uu')U$$

has the same eigenvalues as $(I - uu')A(I - uu')$.

Now let P be an orthogonal matrix of eigenvectors of $(I - uu')A(I - uu')$ and B a diagonal matrix of eigenvalues with $(I - uu')A(I - uu') = PBP'$. We can choose these so that the first column of P is u (the eigenvector associated with eigenvalue 0) and the diagonal entries of B are $0, \beta_2, \ldots, \beta_n$. Expanding, we find that

$$B = P'AP - P'Aue_1' - e_1u'AP + e_1e_1'(u'Au) \equiv P'AP - D,$$

with $D_{jj} = 0$ for $j = 2, \ldots, n$. This shows that $(D_{11}, \beta_2, \ldots, \beta_n)$ is the main diagonal of $P'AP$, and therefore by Schur's Theorem (Marshall and Olkin 1979, p. 218) it is contained in the convex hull of the permutations of $(\lambda_1, \ldots, \lambda_n)$. *A fortiori*, $(\beta_2, \ldots, \beta_n)$ is contained in the convex hull of points generated by permutations of all subsets of size $n - 1$ of $\{\lambda_1, \ldots, \lambda_n\}$. □

Proof of Proposition 4.1. The standard Monte Carlo estimator averages independent replications of $\exp(b'Z + \frac{1}{2}Z'AZ)$. Its variance per replication is the difference between the second moment $E[\exp(2b'Z + Z'AZ)]$ and the square of its mean $E[\exp(b'Z + \frac{1}{2}Z'AZ)]$. Each of these expectations is evaluated using equation (4.10) (which follows from equation (A.2)). Subtracting then yields the expression in (i). For any choice of μ, using equation (2.4) and then equation (4.10) we find that the second moment of the importance sampling estimator based on μ is

$$\text{(A.7)} \qquad E\left[e^{2b'Z+Z'AZ}e^{-\mu'Z+(1/2)\mu'\mu}\right] = \exp\left(\tfrac{1}{2}(2b-\mu)'(I-2A)^{-1}(2b-\mu) + \tfrac{1}{2}\mu'\mu\right)|I-2A|^{-1/2}.$$

Minimizing this expression over μ yields the optimal value $\mu = (I - A)^{-1}b$. Substituting this choice in equation (A.7) yields the expression in (ii).

Some algebra shows that the importance sampling estimator based on the optimal μ can be expressed under the original zero-mean measure (cf. equation (2.3)) as

$$\hat{\alpha} = \exp\left(\tfrac{1}{2}b'(I-A)^{-1}b + \tfrac{1}{2}Z'AZ\right).$$

We need to show that $E[\mathrm{Var}[\hat{\alpha}|v_1'Z, \ldots, v_k'Z]]$ is given by the expression in the proposition. This expectation is equal to

$$E[\hat{\alpha}^2] - E\left[E[\hat{\alpha}|v_1'Z, \ldots, v_k'Z]^2\right].$$

The first of these terms matches the first term in (iii) (again using equation (4.10)). To complete the proof we need to show that $E\left[E[\exp(\frac{1}{2}Z'AZ)|v_1'Z, \ldots, v_k'Z]^2\right]$ equals the second term inside parentheses in the expression in (iii). We have

$$\begin{aligned} E\left[E[\exp(\tfrac{1}{2}Z'AZ)|v_1'Z, \ldots, v_k'Z]^2\right] &= E\left[E[\prod_{i=1}^{n} e^{(1/2)\lambda_i(v_i'Z)^2}|v_1'Z, \ldots, v_k'Z]^2\right] \\ &= E\left[\left(\prod_{i=1}^{k} e^{(1/2)\lambda_i(v_i'Z)^2}\prod_{i=k+1}^{n}\frac{1}{\sqrt{1-\lambda_i}}\right)^2\right] \end{aligned}$$

$$= E\left[\prod_{i=1}^{k} e^{\lambda_i (v_i' Z)^2} \prod_{i=k+1}^{n} \frac{1}{1-\lambda_i}\right]$$
$$= \prod_{i=1}^{k} \frac{1}{\sqrt{1-2\lambda_i}} \prod_{i=k+1}^{n} \frac{1}{1-\lambda_i},$$

the second and fourth equalities following from equation (A.5). This equals the desired term in (iii). □

REFERENCES

ANDERSEN, L. (1995): Efficient Techniques for Simulation of Interest Rate Models Involving Non-Linear Stochastic Differential Equations, working paper, General Re Financial Products, New York, NY.

BLEISTEIN, N., and R. A. HANDELSMAN (1975): *Asymptotic Expansions of Integrals*. New York: Holt, Rinehart, and Winston.

BOYLE, P., M. BROADIE, and P. GLASSERMAN (1997): Simulation Methods for Security Pricing, *J. Econ. Dynam. Control* 21, 1267–1321.

BUCK, R. C. (1978): *Advanced Calculus*, 3rd ed. New York: McGraw-Hill.

CHEN, J.-C., D. LU, J. S. SADOWSKY, and K. YAO (1993): On Importance Sampling in Digital Communications—Part I: Fundamentals, *IEEE J. Selected Areas Commun.* 11, 289–299.

COX, J., J. E. INGERSOLL, and S. A. ROSS (1985): A Theory of the Term Structure of Interest Rates, *Econometrica* 53, 385–407.

DEMBO, A., and O. ZEITOUNI (1993): *Large Deviations Techniques and Applications*. Boston: Jones and Bartlett.

DEUSCHEL, J. D., and D. STROOCK (1989): *Large Deviations*. San Diego: Academic Press.

DUPUIS, P., and R. ELLIS (1997): *A Weak Convergence Approach to the Theory of Large Deviations*. New York: Wiley.

FOURNIÉ, E., J. M. LASRY, and N. TOUZI (1999): Monte Carlo Methods for Stochastic Volatility Models; in *Numerical Methods in Finance*, L. C. G. Rogers and D. Talay, eds. New York: Cambridge University Press, 146–164.

GEMAN, H., and M. YOR (1993): Bessel Processes, Asian Options, and Perpetuities, *Math. Finance* 3, 349–375.

GLASSERMAN, P., P. HEIDELBERGER, and P. SHAHABUDDIN (1998): Gaussian Importance Sampling and Stratification: Computational Issues, *Proceedings of the 1998 Winter Simulation Conference*. New York: IEEE Press, 685–693.

GLASSERMAN, P., and Y. WANG (1997): Counterexamples in Importance Sampling for Rare Event Probabilities, *Annals Appl. Probab.* 7, 731–746.

HAMMERSLEY, J., and D. HANDSCOMB (1964): *Monte Carlo Methods*. London: Methuen & Co. Ltd.

HEIDELBERGER, P. (1995): Fast Simulation of Rare Events in Queueing and Reliability Models, *ACM Trans. Modeling and Computer Simulation* 5(1), 43–85.

HULL, J., and A. WHITE (1987): The Pricing of Options on Assets with Stochastic Volatilities, *J. Finance* 42, 281–300.

JOHNSON, R. A., and D. W. WICHERN (1982): *Applied Multivariate Statistical Analysis*. Englewood Cliffs, NJ: Prentice-Hall.

KARLIN, S., and H. TAYLOR (1975): *A First Course in Stochastic Processes*, 2nd ed. New York: Academic Press.

KLOEDEN, P., and E. PLATEN (1992): *Numerical Solution of Stochastic Differential Equations*, Berlin: Springer-Verlag.

MARSAGLIA, G., A. ZAMAN, and J. MARSAGLIA (1994): Rapid Evaluation of the Inverse of the Normal Distribution Function, *Statistics and Probability Letters* 19, 259–266.

MARSHALL, A. W., and I. OLKIN (1979): *Inequalities: Theory of Majorization and its Applications*. New York: Academic Press.

MCKAY, M. D., W. J. CONOVER, and R. J. BECKMAN (1979): A Comparison of Three Methods for Selecting Input Variables in the Analysis of Output from a Computer Code, *Technometrics* 21, 239–245.

MORO, B. (1995): The Full Monte, *RISK* 8 (February), 57–58.

The Numerical Algorithms Group (1996): *NAG C Library Manual, Mark 4*. Oxford, UK: NAG, Ltd.

NEWTON, N. J. (1994): Variance Reduction for Simulated Diffusions, *SIAM J. Appl. Math.* 54, 1780–1805.

NEWTON, N. J. (1997) Continuous-Time Monte Carlo Methods and Variance Reduction; in *Numerical Methods in Finance*, L. C. G. Rogers and D. Talay, eds. New York: Cambridge University Press, 22–42.

RAO, C. R., and H. TOUTENBURG (1995): *Linear Models: Least Squares and Alternatives*. New York: Springer-Verlag.

REIDER, R. (1993): An Efficient Monte Carlo Technique for Pricing Options, working paper, Wharton School, University of Pennsylvania.

ROCKAFELLER, R. T. (1970): *Convex Analysis*. Princeton, NJ: Princeton University Press.

ROGERS, L. C. G. (1995): Which Model for the Term Structure of Interest Rates Should One Use?; in *Mathematical Finance*, IMA vol. 65, M. H. A. Davis, D. Duffie, and I. Karatzas, eds. New York: Springer, 93–116.

SCHOENMAKERS, J. G., and A. W. HEEMINK (1997): Fast Valuation of Financial Derivatives, *J. Computat. Finance* 1, 47–62.

SHAHABUDDIN, P. (1995): Rare Event Simulation in Stochastic Models; in *Proceedings of the 1995 Winter Simulation Conference*, IEEE Computer Society Press, 178–185.