Proceedings of the 1998 Winter Simulation Conference
D.J. Medeiros, E.F. Watson, J.S. Carson and M.S. Manivannan, eds.

GAUSSIAN IMPORTANCE SAMPLING & STRATIFICATION: COMPUTATIONAL ISSUES

Paul Glasserman

Graduate School of Business
Columbia University
New York, NY 10027, U.S.A.

Philip Heidelberger

IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, U.S.A.

Perwez Shahabuddin

IE and OR Department
Columbia University
New York, NY 10027, U.S.A.

ABSTRACT

This paper deals with efficient algorithms for simulating performance measures of Gaussian random vectors. Recently, we developed a simulation algorithm which consists of doing importance sampling by shifting the mean of the Gaussian random vector. Further variance reduction is obtained by stratification along a key direction. A central ingredient of this method is to compute the optimal shift of the mean for the importance sampling. The optimal shift is also a convenient, and in many cases, an effective direction for the stratification. In this paper, after giving a brief overview of the basic simulation algorithms, we focus on issues regarding the computation of the optimal change of measure. A primary application of this methodology occurs in computational finance for pricing path dependent options.

1 INTRODUCTION

We consider Monte Carlo methods driven by Gaussian random variables, a primary application of which is pricing path dependent options. In this finance application, the Gaussian random variables represent the increments of Brownian motion. Only very simple options, e.g., a European call, can be priced analytically in closed form. For the more complicated ones, either numerical methods or Monte Carlo techniques are used. Monte Carlo methods are usually used for higher dimensional problems, or problems with stochastic parameters (like interest rates, volatilities etc.), for which finite difference methods are very time consuming. A recent review of Monte Carlo methods for security pricing may be found in Boyle, Broadie and Glasserman (1997).

Recently, in Glasserman, Heidelberger, Shahabuddin (1998) (we will denote this by GHS98), we presented an efficient Monte Carlo algorithm for estimating $\alpha = E[G(Z)\mathbf{1}_D(Z)]$ where Z is a vector of m independent standard normal random variables, G is some nonnegative function and $\mathbf{1}_D(Z)$ is the indicator that $Z \in D$ for some set D. If we let $N(a, A)$ denote a multivariate random vector with mean (drift) vector a and covariance matrix A, then $Z \sim N(0, I_m)$ where I_m is the $m \times m$ identity matrix. (Since any m dimensional multivariate normal distribution can easily be generated from $N(0, I_m)$, no loss of generality is suffered in this formulation.) The method consists first of doing an importance sampling change of measure, which is chosen to be the best (in an appropriate asymptotic setting) from among all independent multivariate distributions, i.e., distributions of the form $N(a, I_m)$. Let μ denote the optimal drift vector. As will be discussed in Section 2, μ is found by solving a nonlinear optimization problem. The related problem of finding the optimal drift for estimating the probability, $E[\mathbf{1}_D(Z)]$ ($G(Z) \equiv 1$ in our formulation), where D is a rare set, was addressed in Chen, Lu, Sadowsky, and Yao (1993). Further variance reduction is obtained by stratifying along some direction a, i.e., by stratifying upon a linear combination $a'Z$. The selection of a good stratification direction was analyzed in GHS98, but a particularly convenient and often effective direction is to simply let $a = \mu$, the optimal drift vector. See, e.g., Hammersley and Handscomb (1964) for general discussions of both importance sampling and stratification.

A central ingredient in this method is thus to compute the optimal change of measure for the importance sampling. In GHS98, a bisection procedure was used for the specific case of the Asian option (see Section 2) and non-linear optimization techniques were used for the other more general cases. In this paper, we use the special structure of certain instances of this problem to derive a closed form approximation for the optimal change of measure. We also prove that this approximation is close to the true optimum in an appropriate asymptotic setting. This approximation can be interpreted as the first iteration of a refined fixed point iterative method developed in GHS98. In particular, the approximation is obtained by assuming that G is linear and explicitly solving the optimization. We then examine the computational overhead incurred in the optimization part of the overall importance sampling and stratification procedure, where the optimization is done by using the refined fixed point iterative method mentioned above, the bisection method (for the case of Asian options), and a non-linear optimization package.

2 BACKGROUND AND MAIN ALGORITHMS

To motivate this problem, consider the case of using Monte Carlo to price an arithmetic Asian option on a single asset, under standard Black-Scholes assumptions. The price of the underlying asset under the equivalent martingale measure is described by the stochastic differential equation $dS_t = S_t r dt + \sigma S_t dW_t$, where r is the interest rate and σ is the volatility, both of which are assumed to be constants, and W_t is the standard Brownian motion. Let T be time horizon, and let there be n equally spaced time intervals between $[0, T]$ each of length $\Delta = T/n$. The solution of the above equation can be simulated without discretization error on a discrete grid of points $(\Delta, 2\Delta, \ldots, n\Delta)$ by setting S_i, the stock price at the ith grid point, as $S_i = S_0 \exp((r - \sigma^2/2)\Delta i + \sigma\sqrt{\Delta}\sum_{j=1}^{i} Z_j)$, where Z_j's are independent standard normals, i.e., $N(0, 1)$'s. Let $Z = (Z_1, Z_2, \ldots Z_n)$. The discounted payoff for the arithmetic Asian option is given by $G(Z)\mathbf{1}_D(Z)$ where $G(Z) = e^{-rT}(\sum_{i=1}^{n} S_i/n - K)$ and D is the region $\{G(Z) \geq 0\}$. The objective is then to estimate the expected discounted payoff $\alpha = E[G(Z)\mathbf{1}_D(Z)]$, which falls into our general framework ($m = n$ in this case).

We now outline the method presented in GHS98. Let $g(z)$ be the m dimensional multivariate normal density with mean 0 and covariance matrix I_m As is well known from the theory of importance sampling, a zero-variance estimate is obtained by choosing the importance sampling density to be

$$h(z) = G(z)g(z)\mathbf{1}_D(z)/\alpha. \tag{1}$$

However, it is not possible to use this change of measure because the desired quantity α must be known from the outset and, even if it were known, it may be difficult to sample from h. 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.

In GHS98, for tractability, the only $h(z)$ that is considered is $h_a(z)$, which is defined to be the original multivariate normal measure $g(z)$ (that had mean zero) shifted so that the mean vector is now a. One way of achieving a good approximation to (1) is to align the mode of the integrand (assuming it exists and is unique) $G(z)g(z)\mathbf{1}_D(z)$ with the mode of the shifted measure, i.e., choose a to be a vector μ that solves

$$\max_{z \in D} \quad G(z)\ e^{-z'z/2}. \tag{2}$$

Assuming $G(z)$ is appropriately smooth, this tends to assign high probability to regions of D where $G(z)g(z)$ is large. It was shown in GHS98 that such a change of measure is "asymptotically optimal" in an appropriate setting.

The problem then is to compute the optimal drift vector μ. Three main methods were mentioned/used in GHS98 to compute the optimal drift. Assuming that $G(z)$ is positive in the interior of the set D, one can use $F(z) = \ln G(z)$. Hence the problem becomes to find the maximum of $F(z) - z'z/2$ over the set D. Assuming that the maximum occurs in the interior of the set D, the optimal drift μ satisfies the fixed point equation $\nabla F(\mu) = \mu$. The first method is to use the usual fixed point iterative method $\mu_{i+1} = F(\mu_i)$. However, this method did not always converge and was thus discarded. A more refined fixed point iterative method which appears to converge more generally (and faster) was also developed as follows. First rewrite the condition $\nabla F(\mu) = \mu$ as $\nabla G(\mu)/G(\mu) = \mu$. After i iterations one can approximate $G(\mu)$ by $G(\mu_i) + \nabla G(\mu_i)(\mu - \mu_i)$ and $\nabla G(\mu)$ by $\nabla G(\mu_i)$ and thus set μ_{i+1} to be the solution of

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

This set of equations has two roots, the relevant one being given by

$$\mu_{i+1} = \frac{-B(\mu_i) + \sqrt{B(\mu_i)^2 + 4\|\nabla G(\mu_i)\|^2}}{2\|\nabla G(\mu_i)\|^2}\nabla G(\mu_i) \tag{3}$$

where $B(\mu_i) = G(\mu_i) - \nabla G(\mu_i)\mu_i$. The third method was to use general purpose optimization code. An iterative method involving bisection, that was more specific to the Asian option with non-random volatility, was also developed in GHS98.

Given a drift vector μ, the likelihood ratio $g(z)/h_\mu(z) = \exp(-\mu'z + \frac{1}{2}\mu'\mu)$. Thus applying importance sampling and using the fact that $Z + \mu$ (where $Z \sim N(0, I_m)$) has density h_μ we obtain

$$\alpha = E[G(Z + \mu)\mathbf{1}_D(Z + \mu)e^{-\mu' Z - \frac{\mu'\mu}{2}}]. \tag{4}$$

Equation (4) suggests the importance sampling estimator that we use. The form of this estimator motivates the use of stratifying upon $\mu' Z$, which is equivalent to stratifying upon the likelihood ratio.

3 APPROXIMATIONS FOR THE OPTIMAL CHANGE OF MEASURE

We will now approximate solutions to the unconstrained version of (2), i.e., where the constraint $z \in D$ is removed.

Under an appropriate asymptotic setting described below, we show that these approximations are close to the exact optimal. We end with a discussion of what happens when we re-introduce the constraint $z \in D$.

To motivate this asymptotics, consider the problem of pricing Asian options as mentioned in Section 2. The approximation we propose makes use of the fact that certain parameters in the equation for $G(Z)$ are small, especially the prefactor σ in front of the Z_i's. A typical value of σ is 0.2. Hence, we let $\sigma = \delta$ where δ is a small parameter. Also, a typical value of r is 0.05 and so, in this case, $r = \Theta(\delta^2)$ (a function $f(\delta)$ is said to be $\Theta(\delta^c)$ iff there exist positive constants K_1 and K_2 such that $K_1\delta^c \leq f(\delta) \leq K_2\delta^c$ for all δ small enough). When $G(0) > 0$, S_0 is usually relatively close to the strike price K. For example $S_0 = 50$ and $K = 55$. Hence $K = S_0(1 + O(\delta))$.

In general we now consider $G(z)$ of the form $\tilde{G}(\delta z, \delta)$ and study the behavior of the solution of the unconstrained version of

$$\max_{z \in D} \quad \tilde{G}(\delta z, \delta)\ e^{-z'z/2} \tag{5}$$

as $\delta \rightarrow 0$. To capture the basic flavor of the discussion below, consider the simplified version $G(z) = \tilde{G}(\delta z)$ of the asymptotics above. Using the substitution $v = \delta z$, we can transform the corresponding simplified version of (5) to

$$\max_{v \in \delta D} \quad \tilde{G}(v)\ e^{-v'v/(2\delta^2)}. \tag{6}$$

If we assume modest limitations on the growth rate of $\tilde{G}(v)$ with increasing $\|v\|$, then for all sufficiently small δ, only the region in a small neighborhood of 0 matters in computing the maximum. This is illustrated in Figure 1. In this small neighborhood, $\tilde{G}(v)$ may be approximated by $\tilde{G}(0) + \nabla\tilde{G}(0)v$, and so the solution of the unconstrained version of (6) tends to be close to that of

$$\max_{v} \quad (\tilde{G}(0) + \nabla\tilde{G}(0)v)\ e^{-v'v/(2\delta^2)}.$$

The solution of the latter is easily obtained in closed form as

$$\left(\frac{-\tilde{G}(0) + \sqrt{\tilde{G}(0)^2 + 4\delta^2\|\nabla\tilde{G}(0)\|^2}}{2\|\nabla\tilde{G}(0)\|^2}\right)\nabla\tilde{G}(0).$$

In order to state and prove this idea rigorously for the more difficult case of $G(z) = \tilde{G}(\delta z, \delta)$, we need the following assumptions:

Assumption 1 **(a)** *$\tilde{G}(x, \delta)$, $\nabla\tilde{G}(x, \delta)$ and $\nabla_\delta\tilde{G}(x, \delta)$ are continuous at $(0, 0)$. ($\nabla\tilde{G}(x, \delta)$ is the vector of derivatives with respect to the arguments given by x; ∇_δ is the derivative with respect to the last argument.)*

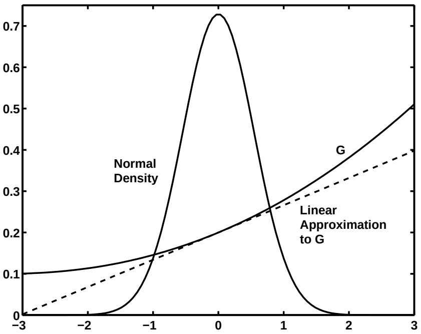

Figure 1: Illustrating the linear approximation to G in a neighborhood of 0. As $\delta \rightarrow 0$, the normal density becomes more peaked around 0, and thus optimizing the density times the linear approximation is almost the same as optimizing the density times G.

(b) *$\tilde{G}(0, \delta) > 0$ (equivalent to $G(0) > 0$) and $\Theta(\delta^c)$ for some $c \geq 0$.*

(c) *$\|\nabla\tilde{G}(0, \delta)\|$ is $\Theta(1)$ (i.e., at least one of the elements of $\nabla\tilde{G}(0, 0)$ is non-zero).*

(d) *The solution to the unconstrained version of (5) is unique for each δ. Call it $\mu \equiv \mu_\delta$.*

(e) *$\|\mu_\delta\|$ is $O(1)$.*

Assumption 1(d) is not necessary, but we have it here in order to simplify the presentation. Sample conditions under which Assumption 1(e) is true are provided in Proposition 1 below.

Theorem 1 *Suppose Assumption 1 holds. Let*

$$\tilde{\mu} \equiv \tilde{\mu}_\delta = \left(\frac{-G(0) + \sqrt{G^2(0) + 4\|\nabla G(0)\|^2}}{2\|\nabla G(0)\|^2}\right)\nabla G(0). \tag{7}$$

Then

$$\frac{\|\tilde{\mu}_\delta - \mu_\delta\|}{\|\mu_\delta\|} \rightarrow 0$$

as $\delta \rightarrow 0$.

Remark 1: The $\tilde{\mu}$ may also be expressed in terms of $\tilde{G}(\cdot, \cdot)$ by using the fact that $G(0) = \tilde{G}(0, \delta)$ and $\nabla G(0) = \delta\nabla\tilde{G}(0, \delta)$. But the representation given by (7) is more practical.

Remark 2: The $\tilde{\mu}$ is also the first iteration of the refined fixed point iteration method given by (3), with the starting point μ_0 being 0.

The proof of Theorem 1 is deferred to Appendix A. The approach we adapt in the proof is as follows. Under the transformation $G(z) = \tilde{G}(\delta z, \delta)$, the $\mu \equiv \mu_\delta$ should satisfy the fixed point equation

$$\nabla \tilde{G}(\delta \nu, \delta) = \nu \tilde{G}(\delta \nu, \delta)/\delta. \tag{8}$$

Let A_δ be the set of solutions to this fixed point equation for a given δ. We "approximate" *all* $\nu \in A_\delta$ that lie in a ball of a sufficiently large but constant (i.e., independent of δ) radius around 0, and then choose the relevant one.

Now we formally specify one of the circumstances under which $\|\mu_\delta\|$ is $O(1)$.

Assumption 2 $[\tilde{G}(x, \delta)]^+ e^{-\|x\|^2/2} \to 0$, *uniformly in* δ, *for all* δ *small enough, as* $\|x\| \to \infty$.

Assumption 2 is true in most problems where $E[G(Z)] = E[\tilde{G}(\delta Z, \delta)]$ is finite.

Proposition 1 *Suppose Assumption 1(a) - 1(d) and Assumption 2 hold. Consider the case where* $\tilde{G}(0, \delta)$ *is* $\Theta(1)$. *Then* $\|\mu_\delta\|$ *is* $O(1)$.

The proof of this proposition is given in Appendix B. Similar results may also be shown for other cases, but they are much more tedious.

Now let us see what happens when we re-introduce the constraint $z \in D$ in the maximization problem. Note that D may also depend on δ (e.g., if D is of the form $\{z : G(z) \geq 0\}$), so we denote D by D_δ. Obviously, the above approximations would go through if the set D_δ included a ball of sufficiently large (constant) radius around 0, for all (sufficiently small) $\delta > 0$. In practice, we found this is procedure to be quite accurate even when only a weaker condition is satisfied: 0 lies in the interior of the set D_δ for all (sufficiently small) $\delta > 0$. If we found out before hand that $\inf_{z \in D_\delta} \|z\| \to \infty$ as $\delta \to 0$, then due to Assumption 1(e), μ_δ will not lie in the feasible region (for all sufficiently small δ). In fact, in many cases where $0 \notin D_\delta$, it turns out that this is the case. Then the approximation cannot be expected to be close, although one still use the iterative procedure of (3) starting at 0.

Let us take a closer look at (6) corresponding to the simpler asymptotics $G(z) = \tilde{G}(\delta z)$, when $0 \in \delta D_\delta$ is not satisfied. For simplicity, assume that $\delta D_\delta = \bar{D}$, i.e., it is independent of δ. Then one can expect that the optimal solution to (6) (that can be expressed as $\delta\mu$ where μ now denotes the optimal solution to the version of (5) with the simpler asymptotics) will occur close to $v_{min} \equiv \min_{v \in \bar{D}} \|v\|$ (see Figure 2) and $\tilde{G}(v)$ is approximately linear in the small region around v_{min}. Hence if one were to start the iteration procedure given by (3) at v_{min}/δ then one may again expect asymptotic convergence after one iteration. The problem is that in most cases determining v_{min} is as difficult as determining the solution to (6) itself.

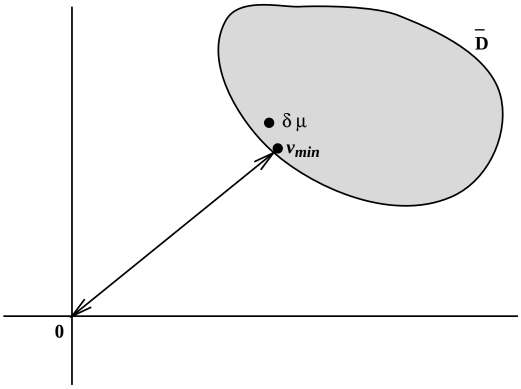

Figure 2: Illustrating the situation when $0 \notin D$ and μ is in the interior of D.

4 COMPUTATIONAL ISSUES AND NUMERICAL RESULTS

In this section we

1. test the accuracy of the approximation scheme described in the previous section and
2. investigate, numerically, the overhead involved in solving the optimization problem to compute the optimal drift vector μ.

Regarding the second point, a reasonable measure of this overhead is to compare the number of function evaluations, M, required to achieve optimality (to a given level of precision) to the number of replications, N, that a standard simulation would require to achieve a desired level of accuracy. As each such function evaluation or replication both require evaluating G, their computational costs are comparable. If M/N is small, then IS+stratification is cost-effective for even a modest reduction in variance (or more precisely, for a modest reduction in the variance times the work per sample, see GHS98). On the other hand, if M/N is large, then the method must produce large variance reductions to be cost-effective. For the purpose of this paper, our definition of the required sample size will be the number of samples required for a 95% confidence interval to have a relative half width of $\pm 1\%$, i.e., if the estimated per sample standard deviation using standard simulation is S and the estimated price of the option is $\hat{P}$ we require $1.96S/\sqrt{N} = 0.01\hat{P}$, or $N = (1.96S/0.01\hat{P})^2$.

We consider four models, which are described in full detail in GHS98: the arithmetic Asian option with constant volatility (denoted Asian), the arithmetic Asian option with the Hull-White stochastic volatility model (denoted HW), the Cox, Ingersoll, Ross interest rate model for pricing a bond (denoted CIR), and the Cox, Ingersoll, Ross interest rate cap model (denoted CIR-Cap). In the interest of space, we do not describe these models or their parameters here, but refer the reader to GHS98. For n time steps, the Asian model has $m = n$ and the HW, CIR, and CIR-Cap models have $m = 2n$.

For the Asian option, it was shown in GHS98 that the optimal μ could be found by reducing the set of n optimality equations to a single nonlinear equation that can be solved by bisection. A high degree of accuracy was typically found in only about a dozen function evaluations, which is negligible compared to the cost of the simulation. For example, in the notation of Section 2, consider the parameter settings $n = 16, \sigma = 0.3, K = 50, S_0 = 50, r = 0.05$ and $T = 1$. With these parameters the required sample size using standard simulation is about 88,000 replications. IS and stratification reduces the variance by about a factor of 1,300. However, each such sample requires about 10% more CPU time than standard simulation (due to the increased cost of sampling from the stratified distribution). Thus the method improves the the efficiency (work times variance) by about a factor of 1,180 ($\approx 1,300/1.1$).

We will also use this example to illustrate the accuracy of the approximation $\tilde{\mu}$ described in the previous section. Define the Relative Error (RE) to be $\|\tilde{\mu} - \mu\|/\ \|\mu\|$. For the same parameter setting as in the above example, the RE is 0.0596. Note that the payoff at zero is positive in this case, as required for the approximation. It also worthwhile to see how the RE behaves as a function of the volatility (σ), as in this case, δ may be interpreted directly as σ. As σ decreases from 0.3 to 0.03, the RE decreases to 0.00754.

For the other three models, nearly closed-form optimal solutions (like the bisection algorithm mentioned above) to the optimization problem are not available. We therefore compared two general purpose approaches that do not take advantage of any problem specific structure. The first approach is to apply a nonlinear optimization package. The specific package we chose was GRG2, which was developed by Leon Lasdon of the University of Texas and is marketed by Optimal Methods, Inc. The second approach is to apply (3) with a suitable starting point. Both these approaches require derivative information, which were estimated using finite differences. The conditions under which these methods converge to a global optimum are different, and difficult to verify in practice. However, we did not find convergence to be a problem, provided a "reasonable" initial point z_0 was selected (see discussion below). In addition, a direct comparison of function evaluation counts is somewhat misleading since the termination criteria are different. In the case of GRG2, the termination criteria are complex as they include checking that the objective function does not change by more than a factor of ϵ for a certain number of iterations (line searches). In addition, GRG2 also computes the gradient at the final point. We set a similar termination criterion for the fixed point iteration; the iteration terminated when the objective function changed by less than a factor of ϵ. Throughout, we used $\epsilon = 10^{-4}$.

Table 1: Number of function evaluations to achieve convergence to optimality in the Hull-White stochastic volatility model. All results use $S_0 = 50$, $V_0 = 0.09$, $\nu = 0$, $r = 0.05$ and $T = 1.0$.

K	ξ	n	GRG2	Fixed Point
50	0.5	32	402	260
50	1.0	32	402	325
50	2.0	32	403	390
55	0.5	32	469	260
55	1.0	32	403	325
55	2.0	32	471	455
50	0.5	64	912	516
50	1.0	64	784	645
50	2.0	64	1044	774
55	0.5	64	915	516
55	1.0	64	915	645
55	2.0	64	1043	903

For the HW model the (undiscounted) payoff function takes the form $G(z) = (A(z) - K)^+$ where K is the strike price and $A(z)$ is the arithmetic average of the underlying stock prices. Note that if μ optimizes $g(z)G(z)$, it also optimizes $H(z) = g(z)(A(z) - K)$, which has the advantage of not losing information about the shape of $A(z)$ when $A(z) \leq K$. Thus for the HW model, we optimized $H(z)$. Recall that $z = 0$ corresponds to not doing IS, which as described earlier is a natural starting point. However, it turns out that if $z = 0$, then the partial derivatives of $H(z)$ are all 0 for $z(n+1), \ldots, z(2n)$. Therefore, we set $z_0(i) = 0.01$ for all i and all parameter settings.

Table 1 shows the total function evaluation counts (including those used for the finite difference approximations to the derivatives) for the two methods. In all cases, the optimization problem is solved in between 250 and 1,100 function evaluations. As expected, the cost to solve the optimization increases as n increases. Recall that for $n = 32$ this is a 64 dimensional problem while for $n = 64$ this is a 128 dimensional problem. Thus most of the function evaluations could be eliminated if partial derivatives were computed analytically (although it is by no means easy to compute them). While the fixed point iteration appears to converge more quickly, part of the difference is due to the different termination criteria.

As it may not be necessary to actually solve the optimization to such a high degree of accuracy in order to obtain good variance reduction, we next investigate

Table 2: IS + Stratification performance for the Hull-White stochastic volatility model. All results use $n = 32$, $S_0 = 50$, $V_0 = 0.09$, $\nu = 0$, $\xi = 1.0$, $r = 0.05$ and $T = 1.0$.

K	IS Vector	Function Calls	Relative Error	Variance Ratio
50	$\tilde{\mu}$	65	0.188	28.8
50	$\hat{\mu}_1$	135	0.144	31.8
50	$\hat{\mu}_2$	202	0.023	32.3
50	μ	402	0.0	30.6
55	$\tilde{\mu}$	65	0.206	27.9
55	$\hat{\mu}_1$	135	0.162	34.9
55	$\hat{\mu}_2$	203	0.015	43.5
55	μ	403	0.0	43.0

Table 3: Number of function evaluations to achieve convergence to optimality in the Cox, Ingersoll, Ross interest rate model. All results use $d = 2$, $\kappa = 0.05$, $\sigma = 0.08$ and $T = 1.0$. The starting point is $z_0(i) = 0.0$ for all i.

r_0	n	GRG2	Fixed Point	Relative Error
0.044	16	142	66	0.006
0.064	16	141	66	0.005
0.084	16	141	66	0.004
0.044	64	526	258	0.013
0.064	64	526	258	0.010
0.084	64	526	258	0.009

Table 4: Number of function evaluations to achieve convergence to optimality in the Cox, Ingersoll, Ross interest rate cap model. All results use $d = 2$, $\kappa = 0.05$, $\sigma = 0.08$, $r_0 = 0.064$ and $T = 1.0$.

K	n	GRG2	Fixed Point	Relative Error
0.064	16	209	66	0.004
0.074	16	314	132	0.168
0.084	16	526	132	0.407
0.064	64	1173	387	0.057
0.074	64	1436	387	0.025
0.084	64	1566	516	0.026

how effective the IS+stratification procedure is for "nearly optimal" changes of measure. To study this, let $\hat{\mu}_j$ denote the solution when GRG2 is prematurely terminated after j line searches, and let μ denote the GRG2 solution when solved to optimality, i.e., within the accuracy criteria as described above. For the parameters listed in Table 2, μ is obtained in 4 line searches. Table 2 lists the number of function calls required to compute an IS vector, μ', where $\mu' = \tilde{\mu}, \hat{\mu}_1, \hat{\mu}_2$, or $\mu (= \hat{\mu}_4)$. This represents the cost to solve the problem to partial optimality. In addition, it lists the relative error, $\|\mu' - \mu\|/\|\mu\|$ as well as the variance ratio (estimated variance of standard simulation to that of the IS+stratification procedure). To obtain accurate variance estimates, we used 1,000,000 replications and all stratification used 100 strata. Table 2 illustrates that the procedure can be highly effective even when the optimization problem is not solved exactly. In this example, it becomes more important to obtain a good solution when the strike price K increases, in which case the problem takes on the flavor of a rare event simulation.

Results for the CIR model are reported in Table 3. The problem is solved in between 50 and 550 function evaluations. Again, fixed point iteration appears to converge more rapidly, however GRG2 again produces near-optimal results earlier. The relative error column lists $\|\tilde{\mu} - \mu\|/\|\mu\|$. In this problem, $\tilde{\mu}$ is extremely close to μ, as all cases of this problem satisfy the conditions for the approximations to be asymptotically close, especially the condition $G(0) > 0$.

For the CIR-Cap model, starting in a neighborhood of $z_0 = 0$ produces a payoff of 0 for many of the parameter settings (leading to a gradient estimate of 0). We therefore chose $z_0(i) = 0.2$ for all i, which produces a positive payoff for all parameter settings. As reported in Table 4, this problem is solved in between 50 and 1,600 function evaluations, with fixed point iteration converging more rapidly. As for the $\tilde{\mu}$ (this is not exactly $\tilde{\mu}$, as by definition, $\tilde{\mu}$ is always computed with $z_0(i) = 0$), in some cases they seem to be very close, whereas in other cases there is a wide difference. Note that in this case the payoff is of the form $\sum_{i=1}^{l} G_i(Z)\mathbf{1}_{D_i}(Z)$, and thus strictly speaking it does not fit the framework described earlier. However, one can easily show that if for each i, $G_i(0) > 0$, then one can again expect a good approximation. In none of the above mentioned cases is this condition satisfied. With $K = 0.054$ this condition is satisfied and with $z_0(i) = 0$ the relative error was 0.033 for $n = 16$ and 0.034 for $n = 64$.

We also experimented with using $\tilde{\mu}$ as the starting point for GRG2, which sometimes, but not always, produced savings compared to GRG2 initialized with the values of z_0 described above.

Table 5 reports the variance ratio (estimated variance of standard simulation divided by estimated variance of IS + stratification, obtained from GHS98), the required sample sizes (in thousands) for $\pm 1\%$ relative accuracy (derived from data reported in GHS98) and the percentage optimization overhead. The overhead is defined to be the corresponding number of function evaluations required by GRG2 as shown in Tables 1-3 divided by

Table 5: Parameters are as in Tables 1, 3 and 4 with $n = 32$ for the HW model, and $n = 16$ for the CIR and CIR-Cap models.

Model	Variance Ratio	Sample Size ($\times$1000)	Opt. Overhead
HW, $K = 50, \xi = 1.0$	30.6	117	0.34%
HW, $K = 50, \xi = 2.0$	13.5	190	0.21%
HW, $K = 55, \xi = 1.0$	43.0	251	0.16%
HW, $K = 55, \xi = 2.0$	21.5	466	0.10%
CIR, $r_0 = 0.044$	105	1.54	9.2%
CIR, $r_0 = 0.064$	152	1.06	13.3 %
CIR, $r_0 = 0.084$	200	0.79	17.8 %
CIR-Cap, $K = 0.064$	39.1	65	0.22%
CIR-Cap, $K = 0.074$	36.8	167	0.08%
CIR-Cap, $K = 0.084$	48.2	413	0.03%

the required sample sizes (expressed as a percentage). This is a conservative estimate of the overhead, since the optimization may not need to be solved to such a high degree of accuracy. For the HW and CIR-Cap models, the required sample sizes are in the tens to hundreds of thousands, the optimization overheads are less than 1%, and the variance reductions range from 13.5 to 200. Interpreting, for example, the first row of Table 5, we see that IS + stratification would obtain the same $\pm 1\%$ relative accuracy in about 3,800 ($\approx 117,000/30.6$) replications. However, each such sample takes about 17% more CPU time than standard simulation. Including the cost of the optimization (400), IS + stratification achieves comparable accuracy in effectively $4,850 (\approx 400 + 1.17 \times 3,800)$ replications.

For the CIR model, which prices a bond paying 100 at maturity, the accuracy was defined relative to $(100 - \hat{P})$. Standard simulation achieves this level of accuracy in about 1,000 replications. Thus the relative cost of the optimization is high, and in fact this type of elaborate variance reduction technique seems unnecessary. However, if one wants to achieve "penny accuracy", i.e., ± 0.01 absolute accuracy, then the required sample sizes increase to between 30,000 and 50,000, the percentage optimization overhead decreases and the procedure becomes computationally attractive.

ACKNOWLEDGMENT

This work is partially supported by NSF National Young Investigator Award Grant DMI-94-57189, NSF University Industry Cooperative Research Program Grant DMS-95-08709, and NSF CAREER Award Grant DMI-96-25297.

APPENDIX A

Proof of Theorem 1: We will first give some definitions and prove a lemma. For any set $A \subset \mathbf{R}^m$, define $\|A\|$ as $\sup_{z \in A} \|z\|$. Two vectors ν'_δ and ν_δ in $\mathbf{R}^m$ (where $\|\nu'_\delta\|$ and $\|\nu_\delta\|$ are positive for all sufficiently small $\delta > 0$) are said to be "asymptotically close" iff $\|\tilde{\nu}_\delta - \nu'_\delta\| / \|\nu'_\delta\| \to 0$ as $\delta \to 0$. Let $(\nu_{\delta,1}, \ldots, \nu_{\delta,k})$ be a set of vectors in $\mathbf{R}^m$ with $\|\nu_{\delta,i}\| \neq 0$ for all i and all sufficiently small $\delta > 0$. Assume that none of them are asymptotically close to each other. A set $A_\delta \subset \mathbf{R}^m$ is said to be "asymptotically close" to $\nu_{\delta,1}$ iff $\sup_{\nu \in A_\delta} \frac{\|\nu - \nu_{\delta,1}\|}{\|\nu_{\delta,1}\|} \to 0$ as $\delta \to 0$. The set A_δ is said to be asymptotically close to $(\nu_{\delta,1}, \ldots, \nu_{\delta,k})$ iff

$$\sup_{\nu \in A_\delta} \left(\min_{1 \leq i \leq k} \frac{\|\nu - \nu_{\delta,i}\|}{\|\nu_{\delta,i}\|} \right) \to 0 \qquad (9)$$

as $\delta \to 0$. In a rough sense, (9) means that all the elements of A_δ are asymptotically close to at least one $\nu_{\delta,i}$.

We consider three sub-cases for Assumption 1(a):

Case 1: $\tilde{G}(0, \delta)$ is $\Theta(1)$, i.e., $\tilde{G}(0, 0) > 0$.

Case 2: $\tilde{G}(0, \delta)$ is $\Theta(\delta)$, i.e., $\tilde{G}(0, 0) = 0$ and $|\nabla_\delta \tilde{G}(0, 0)| > 0$.

Case 3: $\tilde{G}(0, \delta)$ is $\Theta(\delta^c)$, with $c > 1$.

Other cases (e.g., $\tilde{G}(0, \delta)$ is $\Theta(\delta^c)$ with $0 < c < 1$) can similarly be handled.

Lemma 1 *Suppose Assumption 1 holds. Let $A_{\delta,R} = A_\delta \cap B_R$, where B_R is a ball of constant radius r around 0. Then*

Case 1: *For all $R > 0$, $A_{\delta,R}$ is asymptotically close to $\nabla \tilde{G}(0, 0)\delta / \tilde{G}(0, 0)$. Also $\nabla \tilde{G}(0, 0)\delta / \tilde{G}(0, 0)$ and $\nabla G(0)/G(0)$ (which is equivalent to $\nabla \tilde{G}(0, \delta)\delta / \tilde{G}$ $(0, \delta)$) are asymptotically close.*

Case 2: *Let (γ_1, γ_2), be given by*

$$\left(\frac{-\nabla_\delta \tilde{G}(0,0) \pm \sqrt{\nabla_\delta \tilde{G}(0,0)^2 + 4\|\nabla \tilde{G}(0,0)\|^2}}{2\|\nabla \tilde{G}(0,0)\|^2} \right)$$

$$\nabla \tilde{G}(0, 0),$$

respectively. Then for all $R > \max(\|\gamma_1\|, \|\gamma_2\|)$, $A_{\delta,R}$ is asymptotically close to (γ_1, γ_2). Also, γ_1 and γ_2 are asymptotically close to

$$\left(\frac{-G(0) \pm \sqrt{G^2(0) + 4\|\nabla G(0)\|^2}}{2\|\nabla G(0)\|^2} \right) \nabla G(0), \qquad (10)$$

respectively.

Case 3: *For all $R > 1$, $A_{\delta,R}$ is asymptotically close to $\pm\nabla\tilde{G}(0,0)/\|\nabla\tilde{G}(0,0)\|$. Also, $\pm\nabla\tilde{G}(0,\delta)/\|\nabla\tilde{G}(0,\delta)\|$ are asymptotically close to $\pm\nabla G(0)/\|\nabla G(0)\|$, respectively.*

Proof of Lemma 1: Consider Case 1. In this case we only make use of the fact that both $\tilde{G}(x,\delta)$ and $\nabla\tilde{G}(x,\delta)$ are continuous at $(0,0)$. In that case, for all $\epsilon > 0$, one can select a δ_0 such that for all $\delta \leq \delta_0$ and $\nu \in B_R$ (i.e., bounded) $|\tilde{G}(\nu\delta,\delta) - \tilde{G}(0,0)| < \epsilon$ and $\|\nabla\tilde{G}(\nu\delta,\delta) - \nabla\tilde{G}(0,0)\| < \epsilon$. Since $\tilde{G}(0,0) > 0$, this implies that for all $\epsilon' > 0$, one can find δ_0' such that for all $\delta \leq \delta_0'$ and $\nu \in B_R$,

$$\|\frac{\nabla\tilde{G}(\nu\delta,\delta)}{\tilde{G}(\nu\delta,\delta)} - \frac{\nabla\tilde{G}(0,0)}{\tilde{G}(0,0)}\| \leq \epsilon'.$$

Then from (8) we get that for $\nu \in A_{\delta,R}$

$$\|\frac{\nu}{\delta} - \frac{\nabla\tilde{G}(0,0)}{\tilde{G}(0,0)}\| \leq \epsilon'.$$

Now consider Case 2. Again, due to the continuity of $\nabla\tilde{G}(x,\delta)$ we have that for all $\epsilon > 0$, there exists δ_0, such that $\|\nabla\tilde{G}(\nu\delta,\delta) - \nabla\tilde{G}(0,0)\| < \epsilon$ for all $\delta \leq \delta_0$. By the mean value theorem and the fact that $\tilde{G}(0,0) = 0$, we have that for all ν and δ

$$\tilde{G}(\nu\delta,\delta)/\delta = \nabla_\delta\tilde{G}(\theta\nu\delta,\theta\delta) + \nabla\tilde{G}(\theta\nu\delta,\theta\delta)\nu$$

where θ is some quantity between 0 and 1. Again by continuity of the derivatives we have that $|\nabla_\delta\tilde{G}(\theta\nu\delta,\theta\delta) - \nabla_\delta\tilde{G}(0,0)| \to 0$, uniformly over all $\nu \in B_R$. Similarly for $\|\nabla\tilde{G}(\theta\nu\delta,\theta\delta) - \nabla\tilde{G}(0,0)\|$. Now since $\nu \in B_R$, this also holds for $\|\nabla\tilde{G}(\theta\nu\delta,\theta\delta)\nu - \nabla\tilde{G}(0,0)\nu\|$. Hence $|\tilde{G}(\nu\delta,\delta) - (\nabla_\delta\tilde{G}(0,0) + \nabla\tilde{G}(0,0)\nu)| \to 0$ uniformly over $\nu \in B_R$ as $\delta \to 0$.

Note that $|\nabla_\delta\tilde{G}(0,0) + \nabla\tilde{G}(0,0)\nu|$ is always positive for $\nu \in A_{\delta,R}$. (Because if it were zero, then the norm of the righthand side of (8) will be converging to zero, uniformly over $\nu \in A_{\delta,R}$, as $\delta \to 0$, whereas the norm of the left hand side will be converging uniformly to $\|\nabla\tilde{G}(0,0)\| > 0$.) In that case one can show that

$$\|\frac{\nabla\tilde{G}(\nu\delta,\delta)\delta}{\tilde{G}(\nu\delta,\delta)} - \frac{\nabla\tilde{G}(0,0)}{\nabla_\delta\tilde{G}(0,0) + \nabla\tilde{G}(0,0).\nu}\| \to 0$$

uniformly over $\nu \in A_{\delta,R}$, as $\delta \to 0$. Equivalently,

$$\|h(\nu)\| \to 0 \qquad (11)$$

uniformly over $\nu \in A_{\delta,R}$ as $\delta \to 0$, where

$$h(\nu) \equiv \nu - \frac{\nabla\tilde{G}(0,0)}{\nabla_\delta\tilde{G}(0,0) + \nabla\tilde{G}(0,0).\nu}.$$

Now the solutions of $h(\nu) = 0$ is given by γ_1 and γ_2. For any $\epsilon > 0$, let $B_{\epsilon,1}$ and $B_{\epsilon,2}$ be two balls around γ_1 and γ_2, respectively, such that $\|h(z)\| \geq \epsilon$ for $z \in B_R - (B_{\epsilon,1} \cup B_{\epsilon,2})$. Let $r_i(\epsilon)$ be the radius of $B_{\epsilon,i}$. Note that due to the continuity of $h(z)$, for any $\epsilon' > 0$, there exists $\epsilon > 0$, such $B_{\epsilon,1}, B_{\epsilon,2} \subset B_R$, $B_{\epsilon,1} \cap B_{\epsilon,2} = \emptyset$ and $\max(r_1(\epsilon), r_2(\epsilon)) \leq \epsilon'$. Using (11), for any $\epsilon > 0$, there exists δ_0, such that for all $\delta \leq \delta_0$, the set $A_{\delta,R}$ will be a subset of $B_{\epsilon,1} \cup B_{\epsilon,2}$. Hence $\max_{\nu\in A_{\delta,R}} \min(\|\nu - \gamma_1\|, \|\nu - \gamma_2\|) \leq \epsilon'$.

The proof of the third case is very similar. □

Proof of Theorem 1: For Case 1, using the fact that $\tilde{G}(0,\delta)$ is $\Theta(1)$, it can easily be shown that $\tilde{\mu}_\delta$ and $\nabla\tilde{G}(0,\delta)\delta/\tilde{G}(0,\delta)$ are asymptotically close. For Case 2 and Case 3 we need only the positive root in Lemma 1, because for all $\nu \in A_{\delta,R}$ that is close to the negative root, $\tilde{G}(\nu\delta,\delta) < 0$ (for all sufficiently small δ). In Case 2, the positive root in (10) is exactly $\tilde{\mu}_\delta$. For Case 3, using the fact that $\tilde{G}(0,\delta)$ is $\Theta(\delta^c)$, $c > 1$, one can easily show that $\tilde{\mu}_\delta$ and $\nabla\tilde{G}(0,\delta)/\|\nabla\tilde{G}(0,\delta)\|$ are asymptotically close. □.

APPENDIX B

Proof of Proposition 1: $\tilde{G}(0,\delta)$ being $\Theta(1)$ and Assumption 1(a)-(b) imply that $\tilde{G}(0,0) > 0$. Using $v = \delta z$ we can transform the unconstrained version of (5) to $\max_v \tilde{G}(v,\delta)\exp(-v'v/\delta^2)$. Using Assumption 2, there exists a $d > 0$ and $\delta_0 > 0$, such that for all v and δ, such that $\|v\| > d$, and $\delta \leq \delta_0$, $\tilde{G}(v,\delta)e^{-v'v/(2\delta^2)} \leq \tilde{G}(v,\delta)e^{-v'v/2} \leq \tilde{G}(0,0)/2$. Hence, due to the continuity of $\tilde{G}(x,\delta)$ and the fact that $\tilde{G}(0,0)$ is positive, we get that for all sufficiently small δ, the maximum of $\tilde{G}(v,\delta)e^{-v'v/(2\delta^2)}$ cannot occur in the region $\|v\| \geq d$.

So now let us consider the region $\|v\| \leq d$. Let M be the maximum of $\tilde{G}(v,\delta)$ over $\|v\| \leq d$ and $\delta \leq \delta_0$. The compactness of the feasible region and the conditions on $\tilde{G}(v,\delta)$ ensure that M exists and is positive. Then for all v such that $2\delta\sqrt{\ln(\tilde{G}(0,0)/2M)} \leq \|v\| \leq d$ we see that

$$\tilde{G}(v,\delta)e^{-v'v/(2\delta^2)} \leq Me^{-v'v/(2\delta^2)} \leq \tilde{G}(0,0)/2.$$

So the corresponding optimal of the unconstrained version of (5) cannot occur in the region $z \geq 2\sqrt{\ln(\tilde{G}(0,0)/2M)}$. □

REFERENCES

Boyle, P. , Broadie, M. and Glasserman, P. 1997. Simulation Methods for Security Pricing. *Journal of Economic Dynamics and Control* **21**, 1267-1321.

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 in Communications*, **11**, 289–299.

Glasserman, P., Heidelberger, P. and Shahabuddin, P. 1998. Asymptotically Optimal Importance Sampling and Stratification for Pricing Path Dependent Options. Working paper, Columbia University.

Hammersley, J., and D. Handscomb. 1964. *Monte Carlo Methods*. Methuen & Co. Ltd., London.

AUTHOR BIOGRAPHIES

PAUL GLASSERMAN is a Professor in the Management Science division of the Columbia University Graduate School of Business. Prior to joining the Columbia faculty he was a Member of Technical Staff in the Operations Research department of AT&T Bell Laboratories in Holmdel, NJ. He holds a Ph.D. from Harvard University and an A.B. from Princeton University.

PHILIP HEIDELBERGER received a B.A. in mathematics from Oberlin College in 1974 and a Ph.D. in Operations Research from Stanford University in 1978. He has been a Research Staff Member at the IBM T.J. Watson Research Center since 1978.

PERWEZ SHAHABUDDIN is an Associate Professor at the Industrial Eng. and Operations Research Dept. at Columbia University, New York, NY. From 1990 to 1995, he was a Research Staff Member at the IBM T.J. Watson Research Center. He received his B.Tech in Mechanical Eng. from I.I.T/Delhi, in 1984, worked at Engineers India Limited, India, from 1984 to 1985, and then obtained a M.S. in Statistics and a Ph.D in Operations Research from Stanford University in 1987 and 1990, respectively.