Proceedings of the 2010 Winter Simulation Conference
B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, eds.

IMPORTANCE SAMPLING FOR PARAMETRIC ESTIMATION

Xiaojin Tang

Division of Systems Engineering
Boston University
Boston, MA 02225

Pirooz Vakili

Division of Systems Engineering &
Mechanical Engineering Department
Boston University
Boston, MA 02225

ABSTRACT

We consider a class of parametric estimation problems where the goal is efficient estimation of a quantity of interest for many instances that differ in some model or decision parameters. We have proposed an approach, called DataBase Monte Carlo (DBMC), that uses variance reduction techniques in a "constructive" way in this setting: Information is gathered through sampling at a set of parameter values and is used to construct effective variance reducing algorithms when estimating at other parameters. We have used DBMC along with the variance reduction techniques of stratification and control variates. In this paper we present results for the application of DBMC in conjunction with importance sampling. We use the optimal sampling measure at a nominal parameter as a sampling measure at neighboring parameters and analyze the variance of the resulting importance sampling estimator. Experimental results for this implementation are provided.

1 INTRODUCTION AND OVERVIEW

A basic step in the analysis, optimization, and control of stochastic systems is evaluating or approximating quantities of interest based on stochastic models of such systems. These quantities, depending on context, may be performance measures such as cost, price, measures of congestion, energy, $\cdots$, or may be sensitivities of these quantities with respect to some model or decision parameters. Whether performance indices or their sensitivities, they most often can be represented as expected values of appropriately defined random variables. Analytic or deterministic approximation methods, when available, are the most efficient way to evaluate such expected values. The reach of deterministic methods, however, is limited and for most stochastic models statistical estimation via sampling, the so-called Monte Carlo (MC) method, is the most general and flexible tool.

1.1 Variance Reduction Techniques

MC simulation, while flexible and widely applicable, has a relatively slow rate of convergence and since its inception significant effort has been devoted to improving its efficiency. The statistical techniques used for efficiency improvement are often referred to as *Variance Reduction Techniques* (VRT), or *Efficiency Improvement Techniques (EIT)* (Glynn 1994).

Assume the quantity of interest, say J, can be represented as the expected value of a random variable Y, with respect to a probability measure P, i.e.,

$$J = \mathsf{E}_P[Y] \tag{1}$$

For some problems sampling from P directly is not feasible or is excessively costly. Consider the following cases: (i) in many continuous-time models in finance or physics generating continuous-time samples according to P is not possible and discrete-time approximations are simulated; (ii) P represents the stationary distribution of a stochastic process and direct sampling from P may not be feasible; (iii) P has a complex joint density on $\mathbf{R}^d$ for some d for which the common i.i.d. sampling methods are highly inefficient; (iv) P is a probability measure on a finite set S and efficient sampling directly from P is not feasible. This would be case, for example, when P has a Gibbs distribution over a

very large set (as a result, the so-called partition function is not easy to obtain). In such cases, sampling according to P', an appropriate approximation to P, is used.

Assume $Y_1, \cdots, Y_n$ are n samples. The MC estimator is defined as the sample average

$$\hat{J}_{MC} = \frac{1}{n}(Y_1 + \cdots + Y_n).$$

$Y_1, \cdots, Y_n$ may be i.i.d. samples according to probability measure P (or P'), or they may be correlated samples with stationary measure P (or P'), the latter corresponding to the so-called Markov Chain Monte Carlo (MCMC) when Y_i's are evaluated on an appropriately defined Markov chain.

VRTs involve finding a random variable Z and a sampling measure Q such that

$$\mathsf{E}_Q[Z] = \mathsf{E}_P[Y] \quad \text{and} \quad \mathsf{Var}_Q(Z) < \mathsf{Var}_P(Y).$$

The method is effective if $\mathsf{Var}_Q(Z)$ is substantially less than $\mathsf{Var}_P(Y)$. More generally, one may allow some estimation bias if the resulting mean square error is less than $\mathsf{Var}_P(Y)$, i.e.,

$$\mathsf{E}_Q[(Z-J)^2] < \mathsf{Var}_P(Y).$$

The most commonly used such techniques are *Control Variate*, *Stratification*, and *Importance Sampling* (see, e.g., (Asmussen and Glynn 2007), (Glasserman 2004)).

It is worth noting that in many instances of practical importance we are interested in simultaneously evaluating a number of performance indices, Y is a random vector, and J is a vector in $\mathbf{R}^k$ for some k. Some VRTs are effective for simultaneous estimation of a number of performance indices; some adapt too closely to single performance indices to be simultaneously effective for a vector of performance indices. Control variate technique belongs to the former group and stratification and importance sampling generally belong to the latter. In this paper, we assume J is a scalar.

1.2 Parametric estimation & DataBase Monte Carlo (DBMC)

We consider a parametric version of the estimation problem (1); namely, we assume that J depends on some model or decision variable θ. Specifically, let $\{P_\theta; \theta \in \Theta\}$ be a family of probability measures on a measurable space $(\Omega, \mathscr{B})$. Let Y be a real-values function on Ω, i.e., $Y : \Omega \to \mathbf{R}$. We are interested in estimating

$$J(\theta) = \mathsf{E}_{P_\theta}[Y] \quad \text{for a number of } \theta \in \Theta.$$

We assume that the goal is efficient estimation of $J(\theta)$ for many θ, or efficient estimation of $J(\theta)$ for some θ under time and budget constraints. More precisely, we consider the following settings

1. Estimate $J(\theta)$ for many $\theta \in \Theta$.
2. Estimate $J(\theta)$ for a θ that is initially not specified. Once it is, there is a time and/or budget constraints to estimate it.
3. A combination of the two settings above.

We have proposed an approach, called DataBase Monte Carlo (DBMC), that uses variance reduction techniques in a "constructive" way in the above settings: Information is gathered through sampling at a set of parameter values and is used to construct effective variance reducing algorithms when estimating at other parameters. In other words, the overall strategy for efficient estimation is to *learn* from computationally solving instances of the problem, say, at $\theta_1, \cdots, \theta_l$, and then use that information to improve the efficiency of MC when solving other instances. The premise of the approach is that in some parametric estimation problems the setup cost of gathering information at $\theta_1, \cdots, \theta_l$ is justifiable by the resulting improved efficiency when solving other instances; in some cases this is so even if the setup cost is substantial. We have considered this approach when using variance reduction techniques of stratification and control variates. (See, e.g., (Zhao and Vakili 2008), (Borogovac and Vakili 2008), and (Borogovac and Vakili 2009).) In this paper we present some preliminary results for the case where DBMC is used in conjunction with importance sampling.

1.3 Relevant literature

The literature on VRTs generally and the Importance Sampling technique specifically is vast and we will not give a general review. See (Asmussen and Glynn 2007) for a general description and (Liu 2001), (Glasserman 2004), (Chen, Shao, and Ibrahim 2000), for specific focus on application domains of, respectively, scientific computing,

computational finance, and Bayesian computation. (Glynn and Iglehart 1989) covers the theoretical basis of importance sampling and its various implementations.

One of the most effective uses of importance sampling has been the estimation of the probability of rare events, called rare event simulation. See (Juneja and Shahabuddin 2006) for a recent review. Assume the objective is to estimate

$$\alpha = \mathsf{E}_P[I\{A\}] = P(A) \quad \text{for some } A \in \mathscr{B}$$

where $I\{A\}$ is the indicator of the event A. Let $X_1, \cdots, X_n$ be an i.i.d. sample with distribution P and let $Y_i = I\{X_i \in A\}$. Then, a crude MC estimator of α is

$$\hat{\alpha}(n) = \overline{Y}(n) = \frac{1}{n}(Y_1 + \cdots + Y_n).$$

The central limit theorem always holds in this case and we have

$$\sqrt{n}(\hat{\alpha}(n) - \alpha) \Rightarrow N(0, \alpha(1-\alpha)).$$

For α close to zero the absolute error of $\hat{\alpha}(n)$ is of the order of $\sqrt{\alpha \cdot (1-\alpha)}n^{-1/2}$ and is small. However, its relative error, which is more relevant given the small value of α, is of the order

$$\sqrt{\frac{1-\alpha}{\alpha}}\, n^{-1/2} \approx \alpha^{-1/2} \cdot n^{-1/2}.$$

For rare events, i.e., for α very small, the number of samples needed to obtain crude MC estimators with acceptable relative error can be prohibitively high (Glynn 1994). Importance sampling has been the main tool employed in this context leading to significant gains in efficiency. *Large deviation theory* has provided one of the guidelines for obtaining effective importance sampling measures in specific problems (see, e.g., (Bucklew 2004)). Another approach is to consider a parametric family of candidate importance sampling measures and then solve a parametric optimization problem to select the optimum sampling measure from the parametric family (see, e.g., (Glasserman, Heidelberger, and Shahabuddin 1999)).

A recent approach to rare event simulation has been the so-called Cross Entropy (CE) method (see, e.g., (de Mello and Rubinstein 2002), (Rubinstein 2005)). CE, similar to the approach we consider in this paper, uses stochastic sampling to search for a good importance sampling measure. The non-parametric approach to estimating the importance sampling density used in the Generalized Cross Entropy (GCE) method (see, (Botev, Kroese, and Taimre 2007)) is relevant to our research.

Now, let us consider the setting we have in mind in this paper, i.e., the case where the estimation parameter $Y(\theta)$ depends on some model or decision parameter θ. We assume that the parameter θ is a parameter of the probability measure in the following sense. Let $\{P_\theta; \theta \in \Theta\}$ be a family of probability measures on a measurable space $(\Omega, \mathscr{B})$. Let Y be a real-values function on Ω, i.e., $Y : \Omega \to \mathbf{R}$. Let E_θ denote expectation with respect to probability measure P_θ. It is well-known that

$$J(\theta) = \mathsf{E}_\theta[Y] = \mathsf{E}_{\theta_0}[Y \cdot \frac{dP_\theta}{dP_{\theta_0}}(Y)] = \mathsf{E}_{\theta_0}[Y \cdot L(\theta, \theta_0, Y))] \tag{2}$$

where $dP_\theta / dP_{\theta_0}(Y) = L(\theta, \theta_0, Y)$ is the likelihood ratio (assume it is well-defined). Therefore, in principle, from sampling at θ_0 we can obtain an estimate of the entire response surface, namely, $J(\theta)$ at all θ. The variance of such an estimator can grow extremely large as θ moves away from θ_0 (see, e.g., (Glynn and Iglehart 1989), Section 8.)

The approach is used in statistical physics for studying phase transitions (see, e.g., (Binder and Heermann 2002)). The parameter of interest is a temperature related parameter and the goal is to obtain information about quantities of interest at multiple temperatures. Markov Chain Monte Carlo (MCMC) is used at one temperature and the samples are re-weighting using likelihood ratio weights. This method is called *Histogram Re-weighting method* (see, e.g., (Ferrenberg and Swendsen 1988), (Ferrenberg, Landau, and Swendsen 1995), and (Binder and Heermann 2002)) and apparently was initially introduced in statistics (see, (Madras and Piccioni 1999), (Barbu and Zhu 2005)). This approach, which is closely related to ours, is effective in a sufficiently small neighborhood of the nominal parameter.

Compared to i.i.d. sampling, MCMC sampling in general moves in a more restricted way in the sample space. In other words, due to correlations introduced by the Markov chain, consecutive samples are often in some sense closer to each other when compared to independent samples. As a result the time until the Markov chain reaches stationarity, the so-called mixing time, is an important consideration for MCMC. Some recent techniques in the context of parametric MCMC, called *Simulated Tempering* and *Replica Exchange*, consider coupling parametric Markov chains at different parameters to improve the mixing times (see, (Madras and Piccioni 1999)). These are relevant to the setting we have in mind where importance sampling is used in conjunction with MCMC.

In this paper we use information gathered at a single parameter value to improve efficiency of the importance sampling algorithm at neighboring parameters. We analyze the variance of the resulting importance sampling estimator in a neighborhood of the nominal parameter and present a DataBase Monte Carlo (DBMC) implementation.

The rest of the paper is organized as follows. We give a brief review of importance sampling in Section 2 and present our approach and results in Section 3. Some experimental results are given in Section 4. We conclude in Section 5.

2 IMPORTANCE SAMPLING

In this section, we give a brief review of the importance sampling technique. Importance sampling is a general method to reduce variance by having samples generated from a different sampling probability measure that implied by the stochastic model. The sample values are corrected by the *likelihood ratio*. In importance sampling, we change the probability measure in order to give more weight to "important" outcomes thereby increasing estimation efficiency.

Let P be a probability measure on $(\Omega,\mathscr{B})$, $X\sim P$ and $Y=h(X)$ a random variable. Let

$$J=\mathsf{E}_P[Y]=\mathsf{E}_P[h(X)].$$

Assume generating samples of X is feasible. Let $X_1,\cdots,X_n$ be i.i.d. P distributed samples. Then the crude MC estimator is

$$\hat{J}_{MC}(n)=\frac{1}{n}(Y_1+\cdots+Y_n)$$

where $Y_i=h(X_i)$. If $\mathsf{Var}(Y)=\sigma_Y^2$ is finite the central limit theorem holds

$$\sqrt{n}(\hat{J}_{MC}(n)-J)\Rightarrow\sigma_Y N(0,1)$$

where $\Rightarrow$ denotes weak convergence and $N(0,1)$ is the standard normal random variable.

Let Q be another probability measure on $(\Omega,\mathscr{B})$ such that P is absolutely continuous with respect to Q ($P\ll Q$), i.e., $P(A)>0$ implies $Q(A)>0$ for all $A\in\mathscr{B}$. Then J can be alternatively represented as

$$J=\mathsf{E}_P[h(X)]=\mathsf{E}_Q[h(X)\frac{dP}{dQ}(X)].$$

Q is the called the importance sampling probability measure.

Therefore, the *Importance Sampling Algorithm* (the i.i.d version) is as follows.

- Generate $X_1,X_2,...,X_n$, an i.i.d. sample, where $X_i\sim Q$.
- Set $Z_i=h(X_i)\frac{dP}{dQ}(X_i)$.
- Evaluate

$$\hat{J}_{IS}(n)=\frac{1}{n}(Z_1+\cdots+Z_n).$$

$\hat{J}_{IS}$ is an unbiased estimator of J. The weight $\frac{dP}{dQ}$ is the *Radon-Nikodym derivative* (or the likelihood ratio) of P with respect to Q.

If

$$\mathsf{Var}_Q(Z)=\mathsf{Var}_Q(h(X)\frac{dP}{dQ}(X))=\sigma_Z^2<\infty$$

again the central limit theorem holds and we have

$$\sqrt{n}(\hat{J}_{IS}(n)-J)\Rightarrow\sigma_Z N(0,1).$$

Importance Sampling is effective if

$$\mathsf{Var}_Q(Z)<\mathsf{Var}_P(Y).$$

One can alternatively consider a more general measure of effectiveness by taking the computational costs of the two approaches of crude MC and importance sampling into account as in (Glynn and Iglehart 1989), Section 5.

(Glynn and Iglehart 1989) shows that the absolute continuity condition can be relaxed on Q as follows. The requirement of $P(A) > 0$ implies $Q(A) > 0$ is replaced by the new requirement

$$\mathsf{E}_p[h(X)I\{A\}] > 0 \quad \text{implies} \quad Q(A) > 0, \qquad A \in \mathscr{B}.$$

where $I\{A\}$ is the indicator function of $A \in \mathscr{B}$. In this case L does not have a likelihood ratio interpretation. In what follows we consider this more general case. It is well known that there exists an optimal sampling measure, in the sense that it minimizes the $\mathsf{Var}_Q(Z)$. (See, e.g., (Asmussen and Glynn 2007), Chapter 5.)

Proposition 1. *Assume the objective is to estimate $J = \mathsf{E}_P[h(X)]$. The following variance minimization problem*

$$\min_Q \{\mathsf{Var}_Q(Z); Z = h(X)\frac{dP}{dQ}(X); \mathsf{E}_p[h(X)I\{A\}] > 0 \text{ implies } Q(A) > 0,\ A \in \mathscr{B}\}$$

has a solution Q_{opt},

$$dQ_{opt}(x) = \frac{|h(x)|dP(x)}{K}$$

where $K = \mathsf{E}_P[|h(X)|]$ and the corresponding minimum variance is $K^2 - J^2$.

An immediate corollary of the above proposition is the following well-known result.

Corollary 1. *If $h(\cdot)$ above is a nonnegative (or nonpositive) function, then Q_{opt} provides a zero-variance estimator.*

At first glance, it seems that we have identified a perfect sampling measure. However, to obtain the optimal sampling measure, we first need to know $\mathsf{E}_P[h(X)]$ which is exactly what we would like to estimate. This corollary has been used as a guideline for selecting near optimal sampling measures. The corollary shows that a good sampling measure should allocate samples approximately proportional to

$$h(x) \cdot dP(x).$$

In Section 3 we show that this apparently circular result can be constructively used in the context of parametric estimation considered in this paper.

2.1 Importance Sampling and "distance" between probability measures

Another approach for obtaining a good importance sampling measure is to select the best from a restricted set of sampling measures. This latter variance minimization problem can alternatively be formulated as looking for a sampling measure from a restricted set that is the "closest" to the optimum sampling measure, for an appropriately defined distance between probability measures.

(Ali and Silvey 1966) defines a general class of measures of divergence between probability measures. Members of this class are often referred to as Ali-Silvey distances. It is worth noting that these measures of divergence do not satisfy all the properties of a distance in a metric space. An Ali-Silvey distance between two probability measures, defined on the same measurable space, is defined as

$$d(P_1, P_2) = \phi(\int_{\mathbf{R}^M} C[\frac{dP_2}{dP_1}(x)]dP_1(x))$$

where $C(\cdot)$ is a continuous convex real-valued function and $\phi(\cdot)$ is an increasing real-valued function of a real variable. Some of the well-known distances between probability measures, such the Kullback-Leibler distance, belong to this class.

The Ali-Silvey distance satisfies the following properties

- $d(P_1, P_2)$ takes its minimum value when $P_1 = P_2$ and its maximum value when $P_1 \perp P_2$.
- In general, $d(P_1, P_2) \neq d(P_2, P_1)$, i.e., $d(\cdot, \cdot)$ is not symmetric.
- In general, $d(P_1, P_2) + d(P_2, P_3)$ is not greater than or equal to $d(P_1, P_3)$, i.e., the distance measure does not satisfy the triangle inequality.

Assume we use Q as the importance sampling probability measure. Then, the variance of the importance sampling estimator is

$$\mathsf{Var}(h(X)\frac{dP}{dQ}(X)) = \int [h(x)\frac{dP}{dQ}(x)]^2 dQ(x) - J^2$$

$$= K^2 \int [\frac{|h(x)|}{K}\frac{dP}{dQ}(X)]^2 dQ(x) - J^2$$

Substituting in dQ_{opt} from Proposition 1, we have

$$\mathsf{Var}(h(X)\frac{dP}{dQ}(X)) = K^2 \int [\frac{dQ_{opt}}{dQ}(x)]^2 dQ(x) - J^2$$

$$= K^2 d_{IS}(Q, Q_{opt}) - J^2$$

where $d_{IS}(\cdot,\cdot)$ denotes the Ali-Silvey distance with $\phi(x) = x$ and $C(x) = x^2$.

Assume that the choice of importance sampling measure is restricted to a subset of probability measures denoted by $\mathscr{G}$. Then, the above implies that the variance minimization problem

$$\min_{Q\in\mathscr{G}} \mathsf{Var}(Z_{IS}) = \min_{Q\in\mathscr{G}} \mathsf{Var}(h(Y)\frac{dP}{dQ}(Y))$$

is equivalent to the following distance minimization problem.

$$\min_{Q\in\mathscr{G}} d_{IS}(Q, Q_{opt}).$$

For an application of this approach, see, e.g., (Orsak and Aazhang 1991).

3 APPROACH AND PRELIMINARY RESULTS

As stated earlier, our strategy for efficient estimation is to *learn* from computationally solving a number of instances of the estimation problem, say, at $\theta_1, \cdots, \theta_l$, and then use what we learnt to find better importance sampling measures for estimation at other parameters. In this section we assume that sampling at a single parameter, denoted by θ_0 is used to obtain information, i.e., to learn.

Let $\{P_\theta; \theta \in \Theta\}$ be a family of probability measures on a measurable space $(\Omega, \mathscr{B})$. Let

$$J(\theta) = E_\theta[h(X)] \quad \text{where} \quad X \sim P_\theta.$$

Assume $P_\theta \ll P_{\theta_0}$, i.e., $P_\theta(A) > 0$ implies $P_{\theta_0}(A) > 0$ for all $\theta \in \Theta$ and all $A \in \mathscr{B}$. Then, we have

$$J(\theta) = \mathsf{E}_{\theta_0}[h(X)\frac{dP_\theta}{dP_{\theta_0}}(X)].$$

This is the well-known result that by re-weighing samples obtained according to the sampling measure at θ_0 one can estimate the quantity of interest $J(\theta)$ for any θ. It is also well-known that while re-weighing leads to an unbiased estimator of $J(\theta)$ the variance of the resulting estimator is not guaranteed to be lower than $\mathsf{Var}_\theta(h(X))$ and in fact it can be much higher.

We consider the implications of using the optimal importance sampling measure for estimating $J(\theta_0)$ as a sampling measure for estimating $J(\theta)$.

As we pointed out in the previous section, if $h(\cdot)$ is either nonnegative or nonpositive, the optimal importance sampling measure produces a zero variance estimator of $J(\theta_0)$. However, knowing the optimal sampling measure requires knowing $J(\theta_0)$ making such a zero variance estimator apparently useless. We ask the question of what can be said if the optimal Importance Sampling measure for estimating $J(\theta_0)$ is used to estimate $J(\theta)$ in a neighborhood of θ_0. The following proposition provides an answer to this question.

Assume θ is a scalar and P_θ is twice continuously differentiable with respect to θ in a neighborhood of θ_0. Then, we have

Proposition 2. *For a fixed parameter θ_0, assume $h(\cdot)$ is a nonnegative (or nonpositive) function and $dQ(x,\theta_0) = h(x)dP(x,\theta_0)/J(\theta_0)$ is the optimal importance sampling probability measure for estimating $J(\theta_0)$. If this sampling*

measure is used for estimating $J(\theta)$ in a neighborhood of θ_0, then the variance of the corresponding importance sampling estimator satisfies the following.

$$\mathsf{Var}_{Q(\theta_0)}[h(X)\frac{dP_\theta}{dQ(\theta_0)}(X)] = O(\theta-\theta_0)^2$$

The proof is given in the appendix.

The above proposition implies that the optimal importance sampling for estimating $J(\theta_0)$ can be a very effective sampling measure for estimating $J(\theta)$ for values of θ close to θ_0. In other words, the effort to obtain $Q(\theta_0)$ pays dividends when estimating $J(\theta)$ in a neighborhood of θ_0.

To operationalize the above result we use the so-called DataBase Monte Carlo approach (DBMC) as follows.

3.1 Setup phase of DBMC

Let $X_1,\cdots,X_N$ be an i.i.d. sample from P_{θ_0} for large N. Let $DB=\{x_1,\cdots,x_N\}$ be the samples generated. We call this set the *database*. Let $\tilde{P}$ denote the empirical measure associated with the sample. Then, $\tilde{P}$ is the uniform measure on DB assuming that if identical samples are generated, they are kept as separate elements of DB. We have

$$\mathsf{E}_{\tilde{P}}[h(X)] = \frac{1}{N}\sum_{i=1}^{N} h(x_i) = \widetilde{J}(\theta_0) \approx J(\theta_0).$$

We consider solving the following approximate problem to our original estimation problem.

$$\text{Estimate} \quad \tilde{J}(\theta) = \mathsf{E}_{\tilde{P}}[h(X)\frac{dP_\theta}{dP_{\theta_0}}(X)].$$

3.2 Estimation phase of DBMC

Define the probability measure $\tilde{Q}$ on DB by

$$\tilde{Q}(x_i) = h(x_i)/\sum_{i=1}^{N} h(x_i).$$

Then the estimation phase of the DBMC is given by Figure 1.

---

1. Sample $Y_1,Y_2,...,Y_M$ from DB according to $\tilde{Q}$.
2. Set
$$Z_i = \frac{dP_\theta}{dP_{\theta_0}}(Y_i)\cdot\widetilde{J}(\theta_0).$$
3. Calculate
$$\hat{J}_{IS}(\theta) = \frac{1}{M}\sum_{i=1}^{M} Z_i$$
.

Figure 1: DBMC and importance sampling algorithm

Z is an unbiased estimator of $\tilde{J}(\theta)$. Note that

$$\begin{aligned}
\mathsf{E}_{\tilde{Q}}[Z] &= \sum_{i=1}^{N} \frac{dP_\theta}{dP_{\theta_0}}(x_i)\tilde{J}(\theta_0)\cdot\tilde{Q}(x_i) \\
&= \sum_{i=1}^{N} \frac{dP_\theta}{dP_{\theta_0}}(x_i)(\frac{1}{N}\sum_{i=1}^{N} h(x_i))\cdot\frac{h(x_i)}{\sum_{i=1}^{N} h(x_i)} \\
&= \sum_{i=1}^{N} \frac{dP_\theta}{dP_{\theta_0}}(x_i)h(x_i)\cdot\frac{1}{N} = \tilde{J}(\theta).
\end{aligned}$$

4 EXPERIMENTAL RESULTS

In this section we provide a number of experimental results for the application of the DBMC algorithm above to give some indications about the performance of the algorithm. We consider estimating prices of three financial options, a simple European call, an Asian call, and a lookback call option. We assume the price of the underlying asset satisfies the Black-Scholes model. The parameters of interest are the volatility parameter σ and the risk-free interest rate r, namely, the drift parameter of the underlying asset under the risk-neutral measure.

In all cases the information is gathered by extensive sampling at a nominal parameter value (N = 1000,000). Using the resulting payoffs, the optimal sampling measure at the nominal parameter was evaluated and used for sampling when perturbing the parameter of interest. $M = 1000$ values were used in resampling.

Figures 2, 3, and 4 show the results for, respectively European, Asian, and lookback options. The x-axis shows different values of the parameter of interest and the y-axis gives the ratio of the variance of the importance sampling estimator to the crude MC estimator. Note that we are talking about the approximate estimation problem.

All results are consistent with Proposition 2 given in Section 3. The variance of the importance sampling approaches zero as the parameter approaches the nominal parameter. In each case there is a neighborhood of the nominal parameter where importance sampling leads to variance reduction. The variance reduction increases as the parameter approaches the nominal parameter value.

The details of the experiments are given below.

4.1 European Call Option

Assume the objective is to estimate the price of a European call option via Monte Carlo. Let $S_0 = 100$ denote the initial price of the underlying asset, $K = 100$ the strike price, $r = 5\%$ the risk-free interest rate, and $T = 1/6$, the time horizon. Let σ, the volatility of the underlying asset be the parameter of interest. Then,

$$J(\sigma) = e^{-rT}E[(S_0 e^{(r-\frac{1}{2}\sigma^2)T+\sigma\sqrt{T}Z} - K)^+].$$

where $Z \sim N(0,1)$.

We use the above algorithm and the optimal density at $\sigma_0 = 0.2$ as importance sampling density to estimate $J(\sigma)$ where σ changes from 0.08 to 0.3 by step size of 0.004.

Similiarly, if we fix $\sigma = 0.2$, let interest rate r of the underlying asset be the parameter of interest. Then,

$$J(r) = e^{-rT}E[(S_0 e^{(r-\frac{1}{2}\sigma^2)T+\sigma\sqrt{T}Z} - K)^+].$$

where $Z \sim N(0,1)$.

We use the optimal density at $r = 0.05$ as importance sampling density to estimate $J(r)$ where r changes from 0 to 0.4 by step size of 0.004.

4.2 Asian Call Option

Let $S_0 = 100$ denote the initial price of the underlying asset, $K = 100$ the strike price, $r = 5\%$ the risk-free interest rate, and $T = 1/12$, the time horizon. Let σ, the volatility of the underlying asset be the parameter of interest. Then,

$$J(\sigma) = e^{-rT}E[(\overline{S}(\sigma) - K)^+].$$

$$\overline{S} = \frac{1}{m}\sum_{j=1}^{m} S(t_j)$$

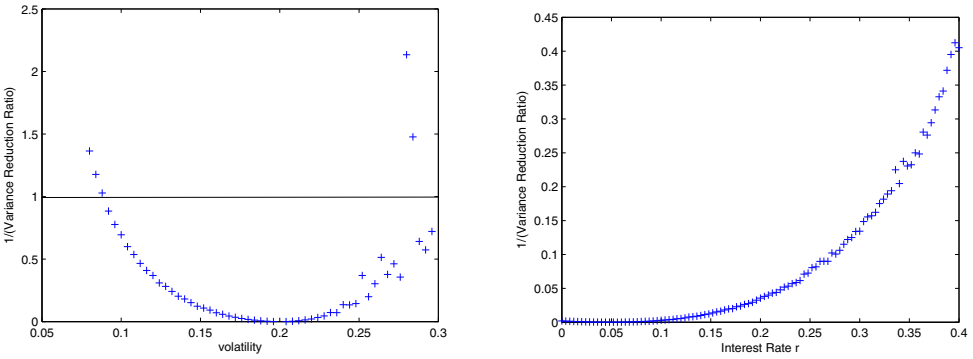

Figure 2: European Call Option - Using optimal IS density at $\sigma = 0.2$ and $r = 0.05$.

$$S(t_{j+1}) = S(t_j)\exp([r - \frac{1}{2}\sigma^2](t_{j+1} - t_j) + \sigma\sqrt{t_{j+1} - t_j}Z_{j+1})$$

where $Z_1, ..., Z_m$ are independent standard normal random variables. Here we assume $m = 10$.

We use the above algorithm and the optimal density at $\sigma_0 = 0.2$ as importance sampling density to estimate $J(\sigma)$ where σ changes from 0.14 to 0.26 by step size of 0.002.

Similarly, if we fix $\sigma = 0.2$, let interest rate r of the underlying asset be the parameter of interest. Then,

$$J(r) = e^{-rT}E[(\overline{S}(r) - K)^+].$$

We use the optimal density at $r = 0.05$ as importance sampling density to estimate $J(r)$ where r changes from 0 to 0.2 by step size of 0.001.

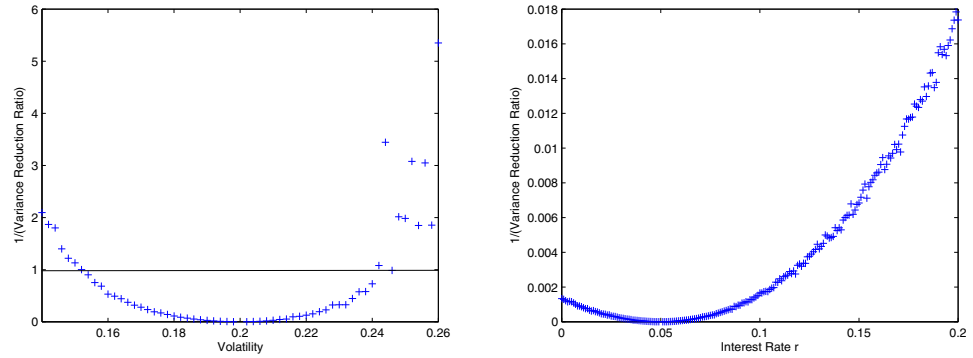

Figure 3: Asian Call Option - Using optimal IS density at $\sigma = 0.2$ and $r = 0.05$.

4.3 Lookback Call Option

Let $S_0 = 100$ denote the initial price of the underlying asset, $K = 100$ the strike price, $r = 5\%$ the risk-free interest rate, and $T = 1/12$, the time horizon. Let σ, the volatility of the underlying asset be the parameter of interest. Then,

$$J(\sigma) = e^{-rT}[S(t_m, \sigma) - \min_{j=1,..,m}(S(t_j, \sigma))].$$

$$S(t_{j+1}, \sigma) = S(t_j, \sigma)\exp([r - \frac{1}{2}\sigma^2](t_{j+1} - t_j) + \sigma\sqrt{t_{j+1} - t_j}Z_{j+1})$$

where $Z_1, ..., Z_m$ are independent standard normal random variables. Here we assume $m = 10$.

We use the above algorithm and the optimal density at $\sigma_0 = 0.2$ as importance sampling density to estimate $J(\sigma)$ where σ changes from 0.14 to 0.26 by step size of 0.002.

Similarly, if we fix $\sigma = 0.2$, let interest rate r of the underlying asset be the parameter of interest. Then,

$$J(r) = e^{-rT}[S(t_m, r) - \min_{j=1,..,m}(S(t_j, r))].$$

We use the optimal density at $r = 0.05$ as importance sampling density to estimate $J(r)$ where r changes from 0 to 0.2 by step size of 0.004.

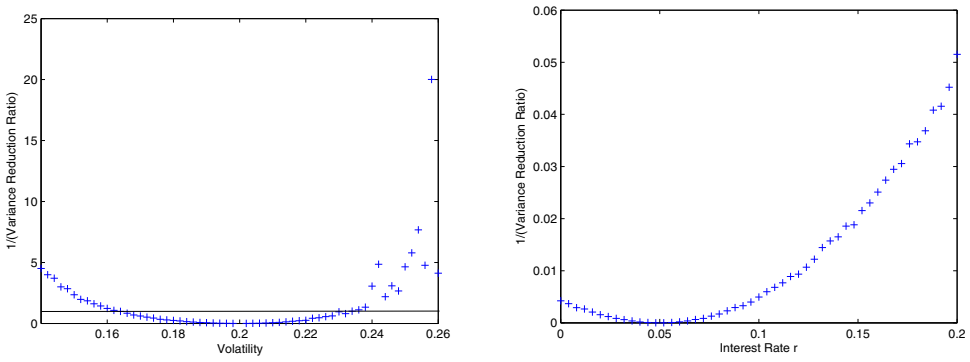

Figure 4: Lookback Call Option - Using optimal IS density at $\sigma = 0.2$ and $r = 0.05$.

5 CONCLUSIONS

We consider the application of DataBase Monte Carlo (DBMC) approach in conjunction with the variance reduction technique of importance sampling in a parametric estimation setting. In DBMC, information, gathered via statistical sampling at a number of parameter values, is used to construct a more effective variance reducing algorithm for estimation at other parameter values. In this paper, information is gathered at a single parameter value. We show that the optimal importance sampling measure can be a very effective sampling measure for estimation at neighboring parameter values and give an analysis of the variance of the resulting importance sampling estimator. Experimental results for a number of examples are provided that show the effectiveness of the approach in a neighborhood of the nominal parameter. A natural extension of this approach to the case where information is gathered at a finite number of parameter values is the subject of our current research.

ACKNOWLEDGMENTS

Research supported in part by the National Science Foundation grant CMMI-0620965.

APPENDIX

Proposition 2. *For a fixed parameter θ_0, assume $h(\cdot)$ is a nonnegative (or nonpositive) function and $dQ(x, \theta_0) = h(x)dP(x, \theta_0)/J(\theta_0)$ is the optimal Importance Sampling probability measure for estimating $J(\theta_0)$. If this sampling measure is used for estimating $J(\theta)$ in a neighborhood of θ_0, then the variance of the corresponding Importance Sampling estimator satisfies the following.*

$$\mathsf{Var}_{Q(\theta_0)}[h(X)\frac{dP_\theta}{dQ(\theta_0)}(X)] = O(\theta - \theta_0)^2$$

Proof: Since $dQ(x,\theta_0) = h(x)dP(x,\theta_0)/J(\theta_0)$, we have

$$\begin{aligned}\mathsf{Var}_{Q(\theta_0)}(h(X)\frac{dP_\theta}{dQ(\theta_0)}(X)) &= \int [h(x)\frac{dP_\theta}{dQ(\theta_0)}(x)]^2 dQ(x,\theta_0) - J(\theta)^2 \\ &= J(\theta_0)^2 \int (\frac{dP_\theta}{dP_{\theta_0}}(x))^2 dQ(x,\theta_0) - J(\theta)^2\end{aligned}$$

Assume P is sufficiently differentiable at θ_0 and consider the Taylor expansions of dP_θ and $J(\theta)$. We have

$$(\frac{dP_\theta}{dP_{\theta_0}})^2 = 1 + 2(\theta-\theta_0)\frac{dP'(\theta_0)}{dP_{\theta_0}} + (\theta-\theta_0)^2[\frac{dP''(\theta_0)}{dP_{\theta_0}} + (\frac{dP'(\theta_0)}{dP_{\theta_0}})^2] + o[(\theta-\theta_0)^2]$$

$$J(\theta)^2 = J(\theta_0)^2 + 2(\theta-\theta_0)J(\theta_0)J'(\theta_0) + (\theta-\theta_0)^2[J(\theta_0)J''(\theta_0) + (J'(\theta_0))^2] + o[(\theta-\theta_0)^2]$$

Therefore,

$$\begin{aligned}\mathsf{Var}_{Q(\theta_0)}[h(X)\frac{dP_\theta}{dQ(\theta_0)}(X)] &= J(\theta_0)^2 \int (\frac{dP(x,\theta)}{dP(x,\theta_0)})^2 dQ(x,\theta_0) - J(\theta)^2 \\ &= J(\theta_0)^2 \\ &+ J(\theta_0)^2 \cdot 2(\theta-\theta_0)\int \frac{dP'(x,\theta_0)}{dP(x,\theta_0)} dQ(x,\theta_0) \\ &+ J(\theta_0)^2(\theta-\theta_0)^2 \int (\frac{dP''(x,\theta_0)}{dP(x,\theta_0)} + [\frac{dP'(x,\theta_0)}{dP(x,\theta_0)}]^2) dQ(x,\theta_0) \\ &- J(\theta_0)^2 \\ &- 2(\theta-\theta_0)J(\theta_0)J'(\theta_0) \\ &- (\theta-\theta_0)^2[J(\theta_0)J''(\theta_0) + (J'(\theta_0))^2]) \\ &+ o[(\theta-\theta_0)^2] \\ &= 2(\theta-\theta_0)J(\theta_0)\int h(x)dP'(x,\theta_0) \\ &+ (\theta-\theta_0)^2 J(\theta_0)(\int h(x)dP''(x,\theta_0) + \int h(x)\frac{dP'(x,\theta_0)^2}{dP(x,\theta_0)}) \\ &- 2(\theta-\theta_0)J(\theta_0)J'(\theta_0) \\ &- (\theta-\theta_0)^2[J(\theta_0)J''(\theta_0) + (J'(\theta_0))^2]) \\ &+ o[(\theta-\theta_0)^2]\end{aligned}$$

Since

$$\begin{aligned}J'(\theta_0) &= \int h(x)dP'(x,\theta_0), \\ J''(\theta_0) &= \int h(x)dP''(x,\theta_0).\end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathsf{Var}_{Q(\theta_0)}[h(X)\frac{dP_\theta}{dQ(\theta_0)}(X)] &= (\theta-\theta_0)^2[J(\theta_0)\int h(x)\frac{dP'(x,\theta_0)^2}{dP(x,\theta_0)} - (J'(\theta_0))^2] + o[(\theta-\theta_0)^2] \\ &= O(\theta-\theta_0)^2.\end{aligned}$$

REFERENCES

Ali, S., and D. Silvey. 1966. A General Class of Coefficients of Divergence of One Distribution from Another. *Journal of the Royal Statistical Society* 28 (1): 131–142.

Asmussen, S., and P. Glynn. 2007. *Stochastic simulation: Algorithms and analysis*. Springer.

Barbu, A., and S. C. Zhu. 2005. Generalizing Swendsen-Wang to Sampling Arbitrary Posterior Probabilities. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27 (8): 1239–1253.

Binder, K. K., and D. Heermann. 2002. *Monte Carlo Simulation in Statistical Physics: An Introduction*. Springer.
Borogovac, T., and P. Vakili. 2008. Control variate technique: A constructive approach. In *Proceedings of the 2008 Winter Simulation Conference*, ed. S. J. Mason, R. R. Hill, L. Moench, and O. Rose, 320–327: IEEE.
Borogovac, T., and P. Vakili. 2009. DataBase Monte Carlo (DBMC) & Generic Control Variates for Parametric Estimation. Technical report, Boston University College of Engineering.
Botev, Z. I., D. P. Kroese, and T. Taimre. 2007. Generalized Cross-entropy Methods with Applications to Rare-event Simulation and Optimization. *Simulation* 83:785–806.
Bucklew, J. A. 2004. *Introduction to rare event simulation*. Springer.
Chen, M., Q. Shao, and J. G. Ibrahim. 2000. *Monte Carlo methods in Bayesian computation*. Springer.
de Mello, T. H., and R. Y. Rubinstein. 2002. Estimation of Rare Event Probabilities using Cross-entropy. In *Proceedings of the 2002 Winter Simulation Conference*, ed. J. L. S. E. Yucesan, C.-H. Chen and J. M. Charnes, 310–319: IEEE.
Ferrenberg, A. M., D. P. Landau, and R. H. Swendsen. 1995. Statistical Errors in Histogram Reweighing. *Physical Review E* 51 (5): 5092–5100.
Ferrenberg, A. M., and R. H. Swendsen. 1988. New Monte Carlo Technique for Studying Phase Transitions. *Physical Review Letters* 61 (23): 2635–2638.
Glasserman, P. 2004. *Monte carlo methods in financial engineering*. Springer.
Glasserman, P., P. Heidelberger, and P. Shahabuddin. 1999. Importance Sampling and Stratification for Value-at-risk. In *Computational Finance 1999 (Proceedings of the Sixth International Conference on Computational Finance)*, ed. A. L. Y.S. Abu-Mostafa, B. LeBaron and A. Weigend, 7–24: MIT Press.
Glynn, P. 1994. Efficiency Improvement Techniques. *Annals of Operations Research* 53:175–197.
Glynn, P. W., and D. L. Iglehart. 1989. Importance Sampling for Stochastic Simulations. *Management Science* 35 (11): 1367–1392.
Juneja, S., and P. Shahabuddin. 2006. Rare-Event Simulation Techniques: An Introduction and Recent Advances. Volume 13 of *Handbook in OR and MS*, Chapter 11, 291–350. Elsevier B.V.
Liu, J. S. 2001. *Monte carlo strategies in scientific computing*. Springer.
Madras, N., and M. Piccioni. 1999. Importance Sampling for Families of Distributions. *The Annals of Applied Probability* 9 (4): 1202–1225.
Orsak, G., and B. Aazhang. 1991. Constrained Solutions in Importance Sampling via Robust Statistics. *IEEE Transactions on Information Theory* 37 (2): 307–316.
Rubinstein, R. Y. 2005. A Stochastic Minimum Cross-Entropy Method for Combinatorial Optimization and Rare-Event Estimation. *Methodology and Computing in Applied Probability* 7 (1): 5–50.
Zhao, G., and P. Vakili. 2008. Monotonicity and stratification. In *Proceedings of the 2008 Winter Simulation Conference*, ed. S. J. Mason, R. R. Hill, L. Moench, and O. Rose, 313–319: IEEE.

AUTHOR BIOGRAPHIES

XIAOJIN TANG is a Ph.D. student of Systems Engineering at Boston University. Her current research interests include efficient Monte Carlo simulation in the areas of finance. Her e-mail address is <xiaojin@bu.edu>

PIROOZ VAKILI is an Associate Professor in the Division of Systems Engineering and the Department of Mechanical Engineering at Boston University. His research interests include Monte Carlo simulation, optimization, computational finance, and bioinformatics. His email address is <vakili@bu.edu>.