Proceedings of the 2004 Winter Simulation Conference
R. Ingalls, M. Rossetti, J. Smith, B. Peters, eds.

ADAPTIVE CONTROL VARIATES

Sujin Kim
Shane G. Henderson

School of Operations Research and Industrial Engineering
Cornell University
Ithaca, NY 14853, U.S.A.

ABSTRACT

Adaptive Monte Carlo methods are specialized Monte Carlo simulation techniques where the methods are adaptively tuned as the simulation progresses. The primary focus of such techniques has been in adaptively tuning importance sampling distributions to reduce the variance of an estimator. We instead focus on adaptive control variate schemes, developing asymptotic theory for the performance of two adaptive control variate estimators. The first estimator is based on a stochastic approximation scheme for identifying the optimal choice of control variate. It is easily implemented, but its performance is sensitive to certain tuning parameters, the selection of which is nontrivial. The second estimator uses a sample average approximation approach. It has the advantage that it does not require any tuning parameters, but it can be computationally expensive and requires the availability of nonlinear optimization software.

1 INTRODUCTION

Suppose that we wish to estimate $\mu = EX$, where X is a real-valued random variable. Suppose also that $EY(\theta) = 0$ for any $\theta \in \Theta$, where Θ is a parameter set. Then $X - Y(\theta)$ is an unbiased estimator for μ, where $Y(\theta)$ serves as a control variate, and one is free to select the parameter θ so as to minimize the variance of $X - Y(\theta)$. We propose two adaptive procedures that tune the parameter θ while estimating μ. We study the asymptotic properties of these procedures as the simulation runlengths become large.

Our interest in this problem stems partly from the simulation analysis of multiclass processing networks. When the networks are heavily loaded, simulation estimators can suffer from large variance. Therefore, some form of variance reduction is needed. The simulation estimators developed in Henderson and Meyn (1997), Henderson and Meyn (2003) give large variance reductions, but the asymptotic rates of growth in the variance are the same as for the naive estimator; see Meyn (2003). One approach to improving these estimators is to develop parameterized estimators. Further motivation comes from the problem of estimating the "expected cost to absorption" in a Markov chain. This problem has received a great deal of attention because of its applications in radiation transport problems; see Kollman et al. (1999).

The first of our procedures is based on a stochastic approximation scheme. At iteration k, one has a current parameter choice θ_{k-1}. Several instances of $X - Y(\theta_{k-1})$ are generated, and the sample variance is computed. The gradient of the sample variance is also computed, and this allows one to perform a stochastic approximation step giving θ_k, and the procedure is iterated. This procedure is easily implemented and, when the step sizes of the algorithm are chosen appropriately, gives very good numerical results. It has the disadvantage that the finite-time performance of the algorithm is strongly impacted by the choice of step sizes, which are not always easily selected.

The second procedure does not require tuning parameters (apart from the selection of an initial runlength) and is based on the theory of sample average approximation. Here a fixed sample is generated, and then the parameter θ that minimizes the sample variance for the fixed sample is determined. One then makes a "production run" using the value of θ chosen in the first stage. The initial optimization can be computationally expensive relative to the stochastic approximation procedure, but for very long simulation runs will occupy a vanishingly small fraction of the effort required.

Henderson, Meyn, and Tadić (2003) also studied adaptive control variate schemes using a stochastic approximation procedure for Markov chains in the steady-state setting. They give conditions for the minimization of an approximation of the steady-state variance. Tadić and Meyn (2004) give the mathematical analysis of the stochastic approximation scheme described in Henderson, Meyn, and Tadić (2003).

Henderson and Simon (2004) show that under certain conditions, adaptive control variate estimators can converge at an exponential rate. One of the key assumptions there is the existence of a "perfect" control variate, i.e., a parameter value θ^* with the property that $\text{var}(X - Y(\theta^*)) = 0$. For the applications we have in mind this assumption is unlikely to hold. Maire (2003) expresses the estimation problem as an integration problem over the unit hypercube, and expands the integrand in an orthonormal series. An iterative procedure for estimating the first few terms in the expansion is given that converges exponentially fast. The residual terms are not estimated iteratively, so that in general the convergence rate of the procedure cannot exceed the canonical rate. In contrast, our parameterization $Y(\theta)$ is much more general, and we do not require an orthonormal series of controls.

In this paper we focus attention on the case where the optimal variance is still positive. Consequently, the rates of convergence for our proposed estimators are typically the canonical $n^{-1/2}$ as evidenced by central limit theorems. This precludes the exponential rates of convergence that are obtained in Henderson and Simon (2004). However, we do briefly consider the case of a perfect control variate in the linearly-parameterized case in Section 3. This section sheds further light on the analysis in Henderson and Simon (2004), taking a somewhat different approach to constructing an estimator.

This paper is organized as follows. In Section 2 we give a motivating example from Markov chain theory. We then explore the linearly parameterized case in Section 3, which is precisely that of standard control variate theory. We then turn to the more complicated nonlinear-parameterization case. First, in Section 4 we outline the general problem and discuss gradient estimation. Second, in Section 5 we explore an approach based on stochastic approximation. Third, in Section 6 we explore the sample average approximation approach. In Section 7 we describe the results of some limited experiments with the example of Section 2. Section 8 contains some concluding remarks.

Space reasons prevent the inclusion of most proofs, which may be found in Kim and Henderson (2004). Unless otherwise stated, all vectors are column vectors and all norms are Euclidean.

2 A MOTIVATING EXAMPLE

Let $Z = (Z_n : n \geq 0)$ be a discrete time Markov chain on the finite state space S. Suppose that Z reaches the absorbing state 0 almost surely starting from any $Z_0 > 0$, and let $T = \inf\{n \geq 0 : Z_n = 0\}$ be the time till absorption. Let $f : S \to \mathbb{R}$ be a given cost function. Define

$$\mu(x) = E(\sum_{k=0}^{T-1} f(Z_k)|Z_0 = x)$$

for all $x \in S - \{0\}$ and set $\mu(0) = 0$, so that μ is the expected cost accrued until absorption. If we view f and μ as column vectors, then μ satisfies $\mu = f + P\mu$, where P is the transition matrix of Z. Suppose that μ is unknown and that we wish to estimate it.

Let $u : S \to \mathbb{R}$ be a real-valued function on the state space S with $u(0) = 0$, and for $n \geq 0$ let

$$M_n(u) = u(Z_n) - u(Z_0) - \sum_{j=0}^{n-1}(P - I)u(Z_j),$$

where I is the identity matrix. Then $(M_n(u) : n \geq 0)$ is the well-known Dynkin martingale; see, e.g., Karlin and Taylor (1981), p. 308). The optional sampling theorem ensures that $E_x M_T(u) = 0$ for any u, where E_x denotes expectation under the initial condition $Z_0 = x$. Therefore, one can estimate $\mu(x)$ via iid replications of

$$\left[\sum_{k=0}^{T-1} f(Z_k)\right] - M_T(u)$$

under initial state $Z_0 = x$ and $M_T(u)$ serves as a parameterized control variate. In our general notational scheme, X is the accrued cost till absorption and $Y(\theta)$ is $M_T(u)$, where u depends on a parameter θ as described below. Since $(P - I)\mu = -f$,

$$\sum_{k=0}^{T-1} f(Z_k) - M_T(\mu) = \mu(x),$$

so if $u = \mu$, then we have a zero-variance estimator.

So it is desirable to find a good choice of the function u. Suppose that $u(x) = u(x;\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}^p$ is a $p-$dimensional vector of parameters. A linear parameterization arises if

$$u(x;\theta) = \sum_{i=1}^{p} \theta(i)u_i(x),$$

where $u_i(\cdot)$ are given basis functions, $i = 1, \ldots, p$. In this case $M_n(u)$ can be shown to be a linear combination of martingales corresponding to the basis functions u_i, $i = 1, \ldots, p$. This observation makes it easy to recompute the value of $X - Y(\theta)$ when the value of θ changes. One simply computes the reweighted linear combination.

The situation is more complicated when $u(x;\theta)$ has a nonlinear parameterization. An example of such a parameterization is given by $u(x;\theta) = \theta_1 x^{\theta_2}$, where $p = 2$. Here it is difficult to recompute the value of $X - Y(\theta)$ when θ changes. Essentially one needs to store the sample path of the chain, explicitly or implicitly, in order to be able to do this.

In the linear case,

$$Y(\theta) = \sum_{i=1}^{p} \theta(i) M_T(u_i)$$

is simply a linear combination of zero-mean random variables. In this sense, the linearly parameterized case leads us back to the theory of linear control variates.

3 THE LINEAR CASE

The theory of linear control variates is very well understood; see, for example, Glynn and Szechtman (2002) or Glasserman (2004) for detailed treatments. The standard theory does not cover the perfect (zero-variance) control variate case, so after a brief review of the key ideas we turn to this case.

Suppose that

$$Y(\theta) = \sum_{i=1}^{p} \theta(i) C(i),$$

where $C(i)$ is a real-valued square-integrable random variable with $EC(i) = 0$ for each $i = 1, \ldots, p$. This is the standard multiple control variates setting. Let θ and C be the corresponding column vectors in $\mathbb{R}^p$, so that $Y(\theta) = \theta^T C$, where x^T denotes the transpose of the matrix x. Assuming that the covariance matrix $\Lambda = \mathrm{cov}\,(C, C)$ is nonsingular, the optimal choice of weights θ^* is $\theta^* = \Lambda^{-1}\beta$, where $\beta = \mathrm{cov}\,(X, C)$ is a column vector whose ith component is $\mathrm{cov}\,(X, C(i))$, $i = 1, \ldots, p$. Since θ^* involves moment quantities that are generally unknown, it can be estimated using the sample analogue $\theta_n = \Lambda_n^{-1}\beta_n$ where

$$\begin{aligned} \beta_n &= \frac{1}{n}\sum_{j=1}^{n} X_j C_j - \bar{X}_n \bar{C}_n \text{ and} \\ \Lambda_n &= \frac{1}{n}\sum_{j=1}^{n} C_i C_i^T - \bar{C}_n \bar{C}_n^T. \end{aligned}$$

Here $(X_j : j \geq 1)$ are i.i.d. replicates of X, $(C_j : j \geq 1)$ are i.i.d. replicates of the vector C, and $\bar{X}_n$ and $\bar{C}_n$ are the usual sample means of the first n observations.

Since Λ is nonsingular and $\Lambda_n \to \Lambda$ as $n \to \infty$ almost surely componentwise, it follows that Λ_n is also nonsingular for sufficiently large n, so that the estimator θ_n is well-defined for sufficiently large n. The corresponding estimator for $\mu = EX$ is $\mu_n = \bar{X}_n - \theta_n^T \bar{C}_n$.

One can show that μ_n satisfies a central limit theorem of the form

$$\sqrt{n}(\mu_n - \mu) \Rightarrow \sigma N(0, 1), \tag{1}$$

where $\Rightarrow$ denotes convergence in distribution, $N(0, 1)$ is a normal random variable with mean 0 and variance 1 and $\sigma^2 = \mathrm{var}\,(X - Y(\theta^*))$. One can develop an alternative estimator for θ_n that exploits the fact that $EC = 0$. This makes no difference to the central limit theorem (1); see Glynn and Szechtman (2002).

Hence, if $\sigma^2 > 0$, the estimator μ_n converges to μ at the canonical rate $n^{-1/2}$ as is well known. In the case where $\sigma^2 = 0$ the central limit theorem (1) shows that the convergence is faster than the canonical rate, but the exact asymptotic behaviour is not as clear. It is worth exploring this case in a bit more detail, partly because it is possible to construct perfect (zero-variance) control variates in certain settings (Henderson and Glynn 2002, Henderson and Simon 2004). Of course, as discussed in the introduction, the perfect-control-variate case is unlikely to arise in the applications we have in mind but, partly to provide another perspective on the results of Henderson and Simon (2004) and partly for completeness, we outline the asymptotic behavior of μ_n in this case.

Let

$$\mathbf{X}_n = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \text{ and } \mathbf{C}_n = \begin{bmatrix} 1 & C_1(1) & \cdots & C_1(p) \\ 1 & C_2(1) & \cdots & C_2(p) \\ \vdots & \vdots & \vdots & \ddots \quad \vdots \\ 1 & C_n(1) & \cdots & C_n(p) \end{bmatrix}$$

be the column vector of observations of X and the matrix with jth row containing a 1 together with C_j^T.

Define $N = \inf\{n \geq 1 : \mathbf{C}_n \text{ has full column rank}\}$. Proposition 1 below shows that N is almost surely finite when Λ is nonsingular and $\mu_N = \bar{X}_N - \theta_N^T \bar{C}_N = \mu$ almost surely. Therefore, if we know that a perfect control exists, then we can continue the simulation until time N and report $\bar{X}_N - \theta_N^T \bar{C}_N$ as an estimate of μ that is almost-surely correct. Therefore, in the case when a perfect control variate exists, *the controlled estimator gives the exact answer in finite time*.

It will typically be the case that $N = p+1$ a.s. However, in certain situations N may be random.

Example 1. Suppose that with probability 0.5, $C(1)$ is uniformly distributed on the interval $(-1, 1)$ and $C(2) = C(1) - 1$, and with probability 0.5, $C(1)$ and $C(2)$ are independent uniform random variables on $(-1, 1)$ and $(0, 2)$ respectively. Suppose further that $X = 2C(1) + C(2) + \mu$. Then with probability 0.5^n, $C_i(2) = C_i(1) - 1$ for $i = 1, \ldots, n$. Hence, $P(N = 3) = 7/8$ and for $n \geq 4$, $P(N = n) = (1/2)^n$. At time N we learn the exact coefficients of the linear function that defines X and not before. This then gives μ. If $X = 2C(1) + C(2) + \mu$ except at, say, $C = (1, 1)$ then the linear relationship still holds with probability 1, but now μ_N does not equal μ on all sample paths, but instead only with probability 1. In this

example N has an exponential tail. This observation is true in general assuming only second moments on X and C.

Proposition 1. *Suppose that $X \in \mathbb{R}$ and $C \in \mathbb{R}^p$ have finite second moments, $EC = 0$, $\Lambda = \text{cov}(C, C)$ is positive definite and $X = C^T\theta^* + \mu$ a.s. Then N as defined above is finite a.s., $\mu_N = \mu$ a.s., and N has an exponentially decaying tail, i.e., $P(N > n) \leq ar^n$ for some $a > 0$ and $r < 1$.*

4 NONLINEAR PRELIMINARIES

We now turn to the case where $Y(\theta)$ is a nonlinear function of a random element Y and a parameter vector θ.

Assumption A1 The random variable X is square integrable, and for all $\theta \in \Theta \subseteq \mathbb{R}^p$, $EY^2(\theta) < \infty$ and $EY(\theta) = Eh(Y, \theta) = 0$.

Define $X(\theta) = X - Y(\theta)$ and set $v(\theta) = \text{var}\, X(\theta) = \text{var}\,(X - Y(\theta))$ to be the variance as a function of θ. As before our overall goal is to estimate $\mu = EX$. Our *intermediate* goal is to identify θ^* which minimizes $v(\theta)$ over $\theta \in \Theta$. In general we cannot expect to find a closed form expression for θ^* as in the linear case, and so we approach this problem from the point of view of stochastic optimization. Regardless of which stochastic optimization method we adopt, we need to impose some structure in order to make progress. We now give conditions under which $v(\cdot)$ is differentiable.

Let H denote the support of the probability distribution of (X, Y), i.e., H is the smallest closed set such that $P((X, Y) \in H) = 1$. Let H_2 be the set of y values that appear in H, i.e., $H_2 = \{y : \exists x \text{ with } (x, y) \in H\}$.

Assumption A2 The set Θ is compact. Also, for all $y \in H_2$, the real-valued function $h(y, \cdot)$ is $\mathcal{C}^1$ (i.e., continuously differentiable) on $\mathcal{U}$, where $\mathcal{U}$ is an open set containing Θ.

Recall that a $\mathcal{C}^1$ function is Lipschitz on a compact set. The following observation is then immediate.

Lemma 2. *For all $y \in H_2$, there exists $K(y) > 0$ such that for all $\theta_1, \theta_2 \in \Theta$,*

$$|h(y, \theta_1) - h(y, \theta_2)| \leqq K(y)\, \|\theta_1 - \theta_2\|,$$

where $\|\cdot\|$ is a metric on $\mathbb{R}^p$. Therefore,

$$\sup_{\theta \in \Theta} \left| \frac{\partial h(y, \theta)}{\partial \theta(i)} \right| \leq K(y)$$

for all $y \in H_2$ and $i = 1, ..., p$.

A2 implies that for each $(x, y) \in H$, $(x - h(y, \cdot) - \mu)^2$ is a $\mathcal{C}^1$ function on $\mathcal{U}$. Therefore we have pathwise differentiability. We also need some integrability conditions.

Assumption A3 $$E\left(K(Y) \left[1 + \sup_{\theta \in \Theta} |X(\theta)| \right] \right) < \infty.$$

As noted below, these conditions are sufficient to ensure that v is $\mathcal{C}^1$. An unbiased gradient estimator of $v(\theta)$ can be obtained by noting that the sample variance of i.i.d. observations is an unbiased estimator of the variance, so that under **A1**, and for any $m \geq 2$,

$$v(\theta) = EV(m, \theta) := E \frac{1}{m-1} \sum_{i=1}^{m} (X_i(\theta) - \bar{X}_m(\theta))^2,$$

where

$$\bar{X}_m(\theta) = \frac{1}{m} \sum_{i=1}^{m} X_i(\theta),$$

for all $\theta \in \Theta$ and $(X_1, Y_1), ..., (X_m, Y_m)$ are i.i.d. replications of (X, Y). Note that we include the terms $h(Y_i, \theta)$ in the sample average $\bar{X}_m(\theta)$ even though we know that they have zero mean. We can construct an unbiased gradient estimator from this expression as

$$\begin{aligned} g_m(\theta_0) &= \nabla V(m, \theta_0) \\ &= \frac{1}{m-1} \sum_{i=1}^{m} \nabla_\theta (X_i(\theta) - \bar{X}_m(\theta))^2 \Big|_{\theta = \theta_0}. \end{aligned}$$

Proposition 3. *If **A1-A3** hold, then v is $\mathcal{C}^1$ on Θ, and for $\theta_0 \in \Theta$,*

$$Eg_m(\theta_0) = g(\theta_0).$$

So under the assumptions **A1 - A3**, the variance function $v(\theta)$ is continuously differentiable in $\theta \in \Theta$, and we have an IPA-based unbiased gradient estimator at our disposal. We are now equipped to attempt to minimize $v(\theta)$ over $\theta \in \Theta$.

5 STOCHASTIC APPROXIMATION

Stochastic approximation is a class of methods used to solve differentiable optimization problems similar to the one we face. The general form of the algorithm is a recursion where an approximation θ_n for the optimal solution is updated to θ_{n+1} using an estimator $g_n(\theta_n)$ of the gradient $g(\theta_n)$ of the objective function evaluated at θ_n. For a minimization problem, the recursion is of the form

$$\theta_{n+1} = \Pi_\Theta(\theta_n - a_n g_n(\theta_n)), \tag{2}$$

Initialization: Choose θ_0.
For $k = 1$ to n
  Generate the i.i.d. sample $(X_{k,i}, Y_{k,i}) \sim (X, Y)$, $i = 1, ..., m$, independent of all else.
  Compute

$$A_k(\theta_{k-1}) = \frac{1}{m}\sum_{i=1}^{m}[X_{k,i} - h(Y_{k,i}, \theta_{k-1})],$$

$$g_{k-1}(\theta_{k-1}) = \frac{1}{m-1}\sum_{i=1}^{m}\nabla_\theta[X_{k,i} - h(Y_{k,i}, \theta) - A_k(\theta)]^2|_{\theta=\theta_{k-1}}, \text{ and}$$

$$\theta_k = \Pi_\Theta(\theta_{k-1} - a_{k-1}g_{k-1}(\theta_{k-1})).$$

Next k
Set $\mu_n = n^{-1}\sum_{k=1}^{n} A_k(\theta_{k-1})$.

Figure 1: Stochastic Approximation Algorithm

where Π_Θ denotes a projection of points outside Θ back into Θ, and $\{a_n\}$ is a sequence of positive real numbers.

Our algorithm for finding θ^* and estimating EX is given in Figure 1, where $\{a_n\}_{n=1}^{\infty}$ is a sequence of positive numbers such that

$$\sum_{n=1}^{\infty} a_n = \infty \text{ and } \sum_{n=1}^{\infty} a_n^2 < \infty, \tag{3}$$

and $m \geq 2$ is a fixed positive integer.

The analysis below requires that θ_n converge to some θ^*. Establishing this result requires some care, so we state our main results assuming that this convergence holds and then give sufficient conditions for the convergence of θ_n. The theory does not require that θ^* be a minimizer of $v(\theta)$ over Θ although we would certainly prefer that this is the case. We first show consistency of the estimator μ_n of μ.

Proposition 4. *Assume **A1-A3** and that θ_n converges to some fixed $\theta^* \in \Theta$ a.s. Then $\hat{\mu}_n \to \mu$ a.s. as $n \to \infty$.*

We now assess the rate of convergence of μ_n to μ through a central limit theorem. First we need another assumption.

Assumption A4 There is a neighbourhood $\mathcal{N}$ of θ^* such that the collection $\{X^2(\theta) : \theta \in \mathcal{N}\}$ is uniformly integrable. In other words, for all $\epsilon > 0$, there exists $K_\epsilon > 0$ such that

$$E[X^2(\theta)I(X^2(\theta) > K_\epsilon)] \leq \epsilon, \text{ for all } \theta \in \mathcal{N}.$$

Remark 1. A set of sufficient conditions for **A4** is **A1, A2** and $EK^2(Y) < \infty$. To see why, note that

$$(X - Y(\theta))^2 \leq 2X^2 + 2Y^2(\theta).$$

For any fixed $\theta_0 \in \Theta$,

$$\begin{aligned} Y^2(\theta) &= h^2(Y, \theta) \\ &= [h(Y, \theta_0) + (h(Y, \theta) - h(Y, \theta_0))]^2 \\ &\leq 2h^2(Y, \theta_0) + 2(h(Y, \theta) - h(Y, \theta_0))^2 \\ &\leq 2h^2(Y, \theta_0) + 2K^2(Y)\|\theta - \theta_0\|^2. \end{aligned}$$

But Θ is compact, and hence $\|\theta - \theta_0\|^2$ is bounded. Therefore $X^2(\theta)$ is uniformly (in θ) bounded by an integrable random variable.

Theorem 5. *Assume **A1-A4** and that $\theta_n \to \theta^*$ as $n \to \infty$. Then*

$$\sqrt{mn}(\mu_n - \mu) \Rightarrow N(0, v(\theta^*))$$

as $n \to \infty$. Moreover, μ_n is an unbiased estimator for μ and

$$mn\text{var}\,\mu_n \to v(\theta^*)$$

as $n \to \infty$.

Hence we see that the stochastic approximation estimator μ_n satisfies a strong law and central limit theorem as $n \to \infty$. Note that it will almost invariably be the case that $v(\theta^*) > 0$ so that the rate of convergence of μ_n is the canonical rate $n^{-1/2}$.

Recall that our motivation for choosing $m > 1$ was to obtain an unbiased gradient estimator with low variance. This additional averaging of m terms in each step of the algorithm does not slow convergence, at least to first order, in the sense that the variance of the estimator and the limiting variance that appear in the central limit theorem are each reduced by a factor of m. Hence the choice of $m \geq 2$ is essentially immaterial from the central-limit-theorem point of view. Of course, these are large sample results, so there may be some benefit to carefully choosing m in small samples. We do not explore that possibility here.

In the rather special case where $v(\theta^*) = 0$ the central limit theorem above still holds in the sense that $\sqrt{n}(\mu_n - \mu) \Rightarrow 0$ as $n \to \infty$. The rate of convergence is then faster than $n^{-1/2}$, and the actual rate of convergence depends on the rate at which $\theta_n \to \theta^*$. We do not explore this case further here, because we believe that the case $v(\theta^*) = 0$ is unlikely to arise in the applications we have in mind. See Henderson and Simon (2004) for an exploration of increased convergence rates when $v(\theta^*) = 0$.

We now give conditions under which θ_n converges to some fixed θ^* as $n \to \infty$, using Kushner and Yin (2003), Theorem 2.1, p. 127). We first need some definitions from that text and one more assumption.

A *box* $B \subset \mathbb{R}^p$ is a set of the form

$$B = \{x \in \mathbb{R}^p : a(i) \leq x(i) \leq b(i), i = 1, \ldots, p\}.$$

For $x \in B$ define the set $\mathcal{C}(x)$ as follows. For x in the interior of B, $\mathcal{C}(x) = \{0\}$. For x on the boundary of B, $\mathcal{C}(x)$ is the convex cone generated by the outward normals of the faces on which x lies. A *first-order critical point* x of a $\mathcal{C}^1$ function $f : B \to \mathbf{R}$ satisfies

$$-\nabla f(x) = z \text{ for some } z \in C(x).$$

A first-order critical point is either a point where the gradient $\nabla f(x)$ is zero, or a point on the boundary of B where the gradient "points towards the interior of B." Let $S(f, B)$ be the set of first-order critical points of f in B. We define the distance from a point x to a set S to be

$$\rho(x, S) = \inf_{y \in S} \|x - y\|.$$

The projection $y = \Pi_B x$ is a pointwise projection defined by

$$y_i = \begin{cases} a(i) & \text{if } x(i) \le a(i), \\ x(i) & \text{if } a(i) < x(i) < b(i), \text{ and} \\ b(i) & \text{if } b(i) \le x(i). \end{cases}$$

for each $i = 1, \ldots, p$.

Assumption A5 The random variables X, $K(Y)$ and $Y(\theta_0)$ for some fixed $\theta_0 \in \Theta$ all have finite 4th moments.

When **A1, A2** and **A5** hold, $EY^4(\theta)$ is bounded in $\theta \in \Theta$, as can be shown using a similar argument to that of Remark 1.

Corollary 6. *Let Θ be a box in $\mathbb{R}^p$ and suppose **A1 - A3, A5** hold. Then $\rho(\theta_n, S(v, \Theta)) \to 0$ as $n \to \infty$ a.s.*

Corollary 6 does not ensure that θ_n converges to a fixed θ^* as $n \to \infty$. For that we need to impose further conditions. One simple condition is that the set of first-order critical points $S(v, \Theta)$ consists of a single element θ^*. This condition is unlikely to be easily verified in practice.

In experiments we have found that the stochastic approximation procedure works well so long as the parameters of the procedure are chosen appropriately. However, as with any stochastic approximation procedure, it can be difficult to select good values for these parameters. For this reason we also consider a second estimator based on quite a different approach.

6 SAMPLE AVERAGE APPROXIMATION

In the stochastic approximation method the estimation of θ^* occurs simultaneously with the estimation of μ. An alternative is to first compute an estimate $\hat{\theta}$ of θ^*, where θ^* solves the optimization problem

$$\mathcal{P} : \qquad \min_{\theta \in \Theta} v(\theta).$$

We can then use $\hat{\theta}$ in a second phase where μ is estimated using

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} [X_i - h(Y_i, \hat{\theta})]. \tag{4}$$

If $\hat{\theta}$ is a deterministic approximation for θ^*, then we have the following immediate consequence of the ordinary strong law and central limit theorem.

Theorem 7. *Suppose that $\hat{\theta}$ is deterministic and $E|X_1 - h(Y_1, \hat{\theta})| < \infty$. Then $\hat{\mu}_n \to \mu$ as $n \to \infty$ a.s. If, in addition, $E[X_1 - h(Y_1, \hat{\theta})]^2 < \infty$ then*

$$\sqrt{n}(\hat{\mu}_n - \mu) \Rightarrow N(0, v(\hat{\theta}))$$

as $n \to \infty$.

It will typically be the case, however, that $\hat{\theta}$ is a random variable depending on some initial sample. This is exactly what happens in the sample average approximation method; see Shapiro (2003) for an introduction to this approach. Let m be a positive integer and suppose that we generate, and then fix, the random sample $(\tilde{X}_1, \tilde{Y}_1), (\tilde{X}_2, \tilde{Y}_2), ..., (\tilde{X}_m, \tilde{Y}_m)$. Let $\tilde{X}_i(\theta) = \tilde{X}_i - h(\tilde{Y}_i, \theta)$. Then for a fixed θ, the sample variance of $(\tilde{X}_i(\theta) : 1 \le i \le m)$ is

$$V(m, \theta) = \frac{1}{m-1} \sum_{i=1}^{m} (\tilde{X}_i(\theta) - \bar{\tilde{X}}_m(\theta))^2$$

where

$$\bar{\tilde{X}}_m(\theta) = \frac{1}{m} \sum_{i=1}^{m} \tilde{X}_i(\theta).$$

Then an approximation to the problem $(\mathcal{P})$ is

$$\mathcal{P}_m : \qquad \min_{\theta \in \Theta} V(m, \theta)$$

We refer to $\mathcal{P}_m$ as the *sample average approximation* (SAA) problem corresponding to the original problem $\mathcal{P}$. Once the sample is fixed, the SAA problem can be solved using any convenient optimization software. The software can exploit the IPA gradients derived earlier, which are exact gradients of $V(m, \theta)$. In our implementation we used a quasi-Newton procedure that exploits the IPA gradients.

Strictly speaking, the term "sample average approximation" refers to an approximation of a function $f(\cdot)$ by a sample average $m^{-1} \sum_{i=1}^{m} f(\cdot, \xi_i)$ of random functions.

The function $V(m, \cdot)$ is not of this form. It is, instead, essentially a nonlinear function of sample averages, because we can write

$$V(m, \theta) = \frac{m}{m-1}\left(\frac{1}{m}\sum_{i=1}^{m}\tilde{X}_i^2(\theta) - \bar{X}_m^2(\theta)\right). \quad (5)$$

The standard theory for sample average approximation is readily extended to this setting. We give extensions that we require below.

Let $\hat{\theta}_m$ be a first-order critical point for the problem $\mathcal{P}_m$. We can then estimate μ via (4), using $\hat{\theta}_m$ in place of $\hat{\theta}$. Note that now $\hat{\theta}_m$ is a random variable, and it is no longer clear a priori that versions of the strong law and central limit theorem of Theorem 7 hold. Nevertheless, versions of these results *do* hold, and can be shown using a uniform version of the strong law and some straightforward arguments.

We now state a version of Theorem 7 for the case where $\hat{\theta}$ is random. In this result there is no need for $\hat{\theta}$ to be a solution of $\mathcal{P}_m$; it can be any random variable taking values in Θ. To emphasize the dependence of $\hat{\mu}_n$ on θ we write $\hat{\mu}_n(\theta)$.

Theorem 8. *Suppose that **A1** and **A2** hold, that $EK(Y) < \infty$, and that the samples used in constructing $\hat{\theta}$ are independent of those used in computing $\hat{\mu}_n$. Then $\hat{\mu}_n(\hat{\theta}) \to \mu$ as $n \to \infty$ a.s., and*

$$\sqrt{n}(\hat{\mu}_n(\hat{\theta}) - \mu) \Rightarrow v^{1/2}(\hat{\theta})N(0, 1)$$

as $n \to \infty$, where $N(0, 1)$ is independent of $\hat{\theta}$.

Hence the strong law and central limit theorem continue to hold in the case where $\hat{\theta}$ is random. In particular, if we first solve, or approximately solve, $\mathcal{P}_m$ to get $\hat{\theta}_m$, and then compute $\mu_n(\hat{\theta}_m)$, then the resulting estimator is "well behaved" as the number of samples n gets large.

As the computational budget gets large, one would naturally want to eventually zero in on a fixed θ^* that solves $\mathcal{P}$ using some vanishing fraction of the budget, and use the remainder of the budget to estimate μ. We can model this by assuming that $m = m(n)$ is a function of n such that $m(n) \to \infty$ as $n \to \infty$. In this case, $\hat{\mu}_n(\hat{\theta}_{m(n)})$ behaves the same as $\hat{\mu}_n(\theta^*)$ as $n \to \infty$, at least to first order.

Theorem 9. *Suppose that $\hat{\theta}_{m(n)} \to \theta^*$ as $n \to \infty$ a.s., for some fixed $\theta^* \in \Theta$. Suppose further that **A1 - A3** hold and the samples used in computing $\hat{\theta}_{m(n)}$ are independent of those used to compute $\hat{\mu}_n$ for every n. Then $E\hat{\mu}_n(\hat{\theta}_{m(n)}) = \mu$ for every n, $\hat{\mu}_n(\hat{\theta}_{m(n)}) \to \mu$ as $n \to \infty$ a.s., and $n\mathrm{var}\,\hat{\mu}_n(\hat{\theta}_{m(n)}) \to v(\theta^*)$ as $n \to \infty$. If, in addition, $EK^2(Y) < \infty$, then*

$$\sqrt{n}(\hat{\mu}_n(\hat{\theta}_{m(n)}) - \mu) \Rightarrow N(0, v(\theta^*))$$

as $n \to \infty$.

It remains to give conditions under which $\hat{\theta}_m \to \theta^*$ as $m \to \infty$ a.s. If we could guarantee that θ_m solved problem $\mathcal{P}_m$ exactly then, as in Shapiro (2003), this would follow using standard arguments and an extension of a uniform law of large numbers to nonlinear functions of means. (Recall from (5) that $V(m, \theta)$ is essentially a nonlinear function of sample means, rather than a sample mean itself.) However, the best that we can hope for from a computational point of view is that $\hat{\theta}_m$ is a first-order critical point for the problem $\mathcal{P}_m$. So to obtain convergence to a fixed θ^* we first prove convergence of first-order critical points to those of the true problem $\mathcal{P}$. Our next result extends Theorem 3.1 in Bastin, Cirillo, and Toint (2004) for sample averages to nonlinear functions of sample averages.

Let $f(\theta, \xi)$ be a $\mathbf{R}^q$-valued function of $\theta \in \Theta \subset \mathbf{R}^p$ and a random vector ξ and let $f(\theta) = Ef(\theta, \xi)$. Let

$$\bar{f}_m(\cdot) = \frac{1}{m}\sum_{i=1}^{m} f(\cdot, \xi_i)$$

denote a sample average of m i.i.d. realizations of the function $f(\cdot, \xi)$. We seek conditions under which first-order critical points of $g \circ \bar{f}_m = g(\bar{f}_m(\cdot))$ on Θ converge to those of $g \circ f$.

Theorem 10. *Consider the functions defined immediately above. Let H denote the support of the probability distribution of ξ. Suppose that Θ is convex and compact, the samples $\xi_1, ..., \xi_m$ are i.i.d. and*

- *(i) for all $\xi \in H$, $f(\cdot, \xi) = (f_1(\cdot, \xi), \ldots, f_q(\cdot, \xi))$ is $\mathcal{C}^1$ on Θ,*
- *(ii) the component functions $f_j(\theta, \xi)$ $(j = 1, \ldots, q)$ are dominated by an integrable function, and*
- *(iii) the gradient components $\partial f_j(\theta, \xi)/\partial\theta(i)$ are dominated by an integrable function ($i = 1, ...p$, $j = 1, \ldots, q$).*

Suppose that $g : \boldsymbol{R}^q \to \boldsymbol{R}$ is $\mathcal{C}^1$ on an open set D, where D contains the range of f and $\bar{f}_m$ for all m (and all realizations). Let $\hat{\theta}_m \in S(g \circ \bar{f}_m, \Theta)$, the set of first-order critical points of $g \circ \bar{f}_m$ on Θ. Then $\rho(\hat{\theta}_m, S(g \circ f, \Theta)) \to 0$ as $m \to \infty$ a.s.

Corollary 11. *Suppose that **A1-A3** hold and $EK^2(Y) < \infty$. Then $\rho(\hat{\theta}_m, S(v, \Theta)) \to 0$ as $m \to \infty$ a.s.*

Corollary 11 shows that $\hat{\theta}_m$ converges to the set of first-order critical points of f as $m \to \infty$. This does not guarantee that the sequence $\{\hat{\theta}_m\}$ is convergent of course. A simple sufficient condition that ensures this is that there is a unique first-order critical point, but this condition is clearly difficult to verify in practice.

7 NUMERICAL RESULTS

In this section, we return to the example presented in Section 2 in the context of nonlinear parameterizations.

Let $u(\cdot;\theta)$ be given, where $u(0,\theta) = 0$ for all $\theta \in \Theta$. Let $M_T(u(\cdot;\theta)) = -u(x;\theta) - \sum_{j=0}^{T-1}(P-I)u(Z_j;\theta)$ under some fixed initial state $Z_0 = x$. Then $X(\theta) = X - M_T(u(\cdot;\theta))$ is an estimator of $\mu(x)$. Kim and Henderson (2004) show that **A1-A5** are satisfied in the following examples, so that all of our previous results apply.

For the simulation experiment, we consider a chain on the states $\{0,\ldots,d\}$, where 0 is an absorbing state. The nonzero transition probabilities are $P(x,x+1) = p(x) = 1 - q(x) = P(x,x-1)$, valid for $1 \le x < d$, and $P(d,d-1) = 1$. We take $u(y;\theta) = \theta(1)y^{\theta(2)}$, so that $\theta = (\theta(1),\theta(2)) \in \Theta$, $\Theta = \{x \in \mathbb{R}^2 : a(j) \le x(j) \le b(j),\ j = 1,2\}$ and $a(j) \ge 0$, $j = 1,2$. We take $f(x) = 1$, so that the random variable $X = T$ is the time till absorption in state 0. We set $d = 30$.

The terms naive, SA and SAA represent the estimators obtained through naive Monte Carlo estimation, the stochastic approximation method and the sample average approximation method respectively. In Algorithm 1, we take $m = 100$ and

$$a_k = \frac{e}{C + k^\alpha},$$

where e, $C > 0$ and $\alpha \in [1/2, 1]$ are tunable constants. This form of the gain sequence is suggested in Spall (2003). We used the sample variance of $A_1(\theta_0), ..., A_n(\theta_n - 1)$ as an estimator of $v(\theta^*)/m$. This estimator is shown to be strongly consistent in Kim and Henderson (2004). For the SAA estimator, we first replicated $m = 100$ samples. We solved $\mathcal{P}_m$ using a quasi-Newton method with a linesearch (supplied as part of the MATLAB^{TM} package) using IPA gradients. As an estimator of the variance $v(\hat{\theta})$, we used the sample variance of $X(\hat{\theta})$ over n replicates, where $\hat{\theta}$ is viewed as fixed, in the sense of Theorem 8. We used the same CPU time for all three estimators for a given initial state x to allow a fair comparison.

Example 2. In this example, we let $p(y) = .0001 + .4998/y$ and $\theta_0 = (1,1)$. In Table 1, we show the squared standard errors of the three estimators for varying initial states x. We see that the SAA estimators outperform the SA estimators, and the SA estimators outperform the naive estimator. A problem with the SA estimator is that it is very sensitive to the step size parameters a_k and the initial point θ_0. We performed preliminary simulations with this method, tuning the parameters heuristically until reasonable performance was observed.

Table 1: Estimated Squared Standard Errors in Example 2

x	CPU time (sec)	Naive	SA	SAA
5	16.8	4.4E-4	2.3E-5	1.7E-14
10	20.2	0.0012	5.7E-5	4.1E-14
15	21.8	0.0024	7.5E-5	2.8E-14
20	25.8	0.0035	1.5E-4	5.5E-15
25	28.6	0.0047	9.4E-4	1.3E-6
30	29.8	0.0058	0.003	6.4E-5

Example 3. In this example, $p(x) = .25$ and $\theta_0 = (2,1)$. The results are given in Table 2 and are similar to those of Example 2. The SAA estimator again outperforms the other estimators, but not by as large a margin. A contour plot of the variance surface as a function of θ for initial state $x = 15$ appears in Figure 2. We see that the function is not convex, but appears to have a unique first-order critical point, so that we can expect convergence of the parameter estimates to θ^*, which from the plot appears to be the point $(2,1)$.

Table 2: Estimated Squared Standard Errors in Example 3

x	CPU time (sec)	Naive	SA	SAA
5	15.5	3.7E-4	5.8E-5	1.1E-6
10	17.0	5.2E-4	5.5E-5	6.1E-6
15	17.6	6.8E-4	4.8E-5	1.2E-5
20	19.5	7.4E-4	3.5E-4	1.7E-5
25	21.2	8.0E-4	1.1E-4	2.2E-5
30	21.8	9.1E-4	3.5E-4	2.5E-5

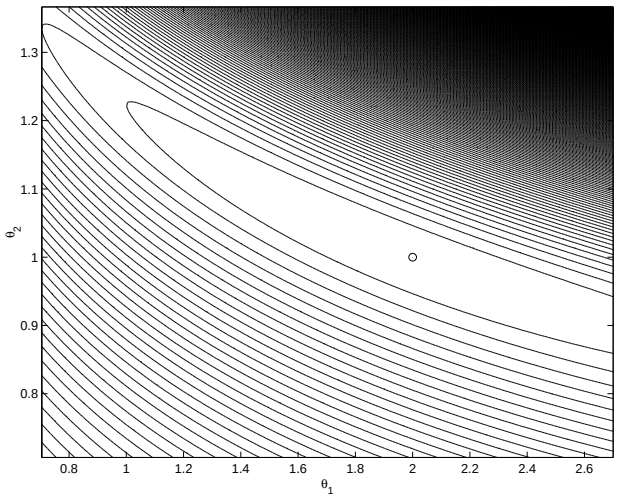

Figure 2: Contour Plot of $v(\cdot)$ for Example 3 With Initial State $x = 15$ And Runlength 1000

8 CONCLUSIONS

The two adaptive estimation procedures developed in this paper have somewhat complementary characteristics. The stochastic approximation scheme has a low computational effort per replication, but typically requires some tuning of the gain sequence to achieve satisfactory performance. The sample average approximation method is more robust, but can be computationally expensive in the initial optimization phase.

The examples in the previous section should be viewed as a simple demonstration of the methods rather than a comprehensive comparison. They serve to demonstrate the feasibility of the two approaches. Both adaptive methods outperform a naive approach.

We are currently exploring the asymptotic theory of variance estimators and more complicated examples with higher-dimensional parameter vectors.

ACKNOWLEDGMENTS

We would like to thank Soren Asmussen for the proof of the exponential tail property of the random variable N in Proposition 1. This research was supported by National Science Foundation grants DMI 0230528, DMI 0224884 and DMI 0400287.

REFERENCES

Bastin, F., C. Cirillo, and P. L. Toint. 2004. Convergence theory for nonconvex stochastic programming with an application to mixed logit. *Mathematical Programming*. To appear.

Glasserman, P. 2004. *Monte Carlo Methods in Financial Engineering*. New York: Springer-Verlag.

Glynn, P. W., and R. Szechtman. 2002. Some new perspectives on the method of control variates. In *Monte Carlo and Quasi-Monte Carlo Methods 2000*, ed. K. T. Fang, F.J.Hickernell, and H. Niederreiter, 27–49. Berlin: Springer-Verlag.

Henderson, S. G., and P. W. Glynn. 2002. Approximating martingales for variance reduction in Markov process simulation. *Mathematics of Operations Research* 27:253–271.

Henderson, S. G., and S. P. Meyn. 1997. Efficient simulation of multiclass queueing networks. In *Proceedings of the 1997 Winter Simulation Conference*, ed. S. Andradóottir, K. J. Healy, D. H. Withers, and B. L. Nelson, 216–223. Piscataway NJ: IEEE.

Henderson, S. G., and S. P. Meyn. 2003. Variance reduction for simulation in multiclass queueing networks. *IIE Transactions*. Submitted.

Henderson, S. G., S. P. Meyn, and V. Tadić. 2003. Performance evaluation and policy selection in multiclass networks. *Discrete Event Dynamic Systems* 13:149–189. Special issue on learning and optimization methods.

Henderson, S. G., and B. Simon. 2004. Adaptive simulation using perfect control variates. *Journal of Applied Probability* 41 (3). To appear.

Karlin, S., and H. M. Taylor. 1981. *A Second Course in Stochastic Processes*. Boston: Academic Press.

Kim, S., and S. G. Henderson. 2004. Adaptive control variates. Manuscript.

Kollman, C., K. Baggerly, D. Cox, and R. Picard. 1999. Adaptive importance sampling on discrete Markov chains. *Ann. Appl. Probab.* 9 (2): 391–412.

Kushner, H. J., and G. G. Yin. 2003. *Stochastic Approximation and Recursive Algorithms and Applications*. 2nd ed. New York: Springer-Verlag.

Maire, S. 2003. Reducing varinace using iterated control variates. *Journal of Statistical Computation and Simulation* 73 (1): 1–29.

Meyn, S. P. 2003. Value functions, optimization and performance evaluation in stochastic network models. *IEEE Transactions on Automatic Control*. Submitted.

Shapiro, A. 2003. Monte Carlo sampling methods. In *Stochastic Programming*, ed. A. Ruszczynski and A. Shapiro, Handbooks in Operations Research and Management Science. Elsevier. To appear.

Spall, J. C. 2003. *Introduction to Stochastic Search and Optimization: Estimation, Simulation and Control*. Hoboken, New Jersey: Wiley.

Tadić, V. B., and S. P. Meyn. 2004. Adaptive Monte Carlo algorithms using control variates. *Manuscript*.

AUTHOR BIOGRAPHIES

SUJIN KIM is a Ph.D. student in Operations Research at Cornell University. Her research interests are in stochastic optimization and efficiency improvement techniques. Her email address is <skim@orie.cornell.edu>.

SHANE G. HENDERSON is an associate professor in the School of Operations Research and Industrial Engineering at Cornell University. He has previously held positions at the University of Michigan and the University of Auckland. He is an associate editor for the *ACM Transactions on Modeling and Computer Simulation*, *Operations Research Letters*, and *Mathematics of Operations Research*, and the secretary of the INFORMS College on Simulation. His research interests include discrete-event simulation, queueing theory and scheduling problems. His e-mail address is <sgh9@cornell.edu>, and his web page URL is <www.orie.cornell.edu/~shane>.