Adaptive Control Variates

Sujin Kim and Shane G. Henderson
School of Operations Research and Industrial Engineering
Cornell University

November 5, 2004

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, 2003] give large variance reductions, but the asymptotic rates of growth in the variance are the same as for the naïve 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. 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 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 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 et al. [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 et al. [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 such that the error in the estimates of the first few terms converges to 0 at an exponential rate. 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 our attention on the case where the optimal variance is still positive. Consequently, the rates of convergence for our proposed estimators are typically the canonical $n^{-1/2}$ rate, where n is proportional to the computational effort, as evidenced by central limit theorems. This precludes the exponential rates of convergence that are demonstrated 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 perfect control variate case treated 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 final remarks.

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 $S = \{0, 1, ..., d\}$. 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

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

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, and we take $f(0) = 0$. 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 i.i.d. replications of

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

where $Z_0 = x$ 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. By (1),

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

and hence, 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,

$$\begin{aligned}
M_n(u) &= u(Z_n;\theta) - u(Z_0;\theta) - \sum_{j=0}^{n-1}[(P - I)u](Z_j;\theta) \\
&= \sum_{i=1}^{p} \theta(i) u_i(Z_n) - \sum_{i=1}^{p} \theta(i) u_i(Z_0) - \sum_{j=0}^{n-1}[(P - I)\sum_{i=1}^{p} \theta(i) u_i](Z_j) \\
&= \sum_{i=1}^{p} \theta(i) u_i(Z_n) - \sum_{i=1}^{p} \theta(i) u_i(Z_0) - \sum_{i=1}^{p} \theta(i) \sum_{j=0}^{n-1}[(P - I)u_i](Z_j) \\
&= \sum_{i=1}^{p} \theta(i) \left[u_i(Z_n) - u_i(Z_0) - \sum_{j=0}^{n-1}[(P - I)u_i](Z_j) \right] \\
&= \sum_{i=1}^{p} \theta(i) M_n(u_i) \qquad (2)
\end{aligned}$$

so that $M_n(u)$ is simply 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 $u(x;\theta)$ arises from 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.

From (2) we see that 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 discuss this case in some detail.

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 $Y(\theta) = \theta^\top C$, where $x^\top$ 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

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

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 $n \to \infty$ element-wise, 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

$$\mu_n = \bar{X}_n - {\theta_n}^\top \bar{C}_n.$$

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

$$\sqrt{n}(\mu_n - \mu) \Rightarrow \sigma N(0,1), \tag{3}$$

where $\Rightarrow$ denotes convergence in distribution, $N(a,b)$ is a normal random variable with mean a and variance b 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 (3); 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 (3) 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

$$\mathbf{X}_n = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \text{ and } \mathbf{C}_n = \begin{bmatrix} 1 & C_1(1) & C_1(2) & \cdots & C_1(p) \\ 1 & C_2(1) & C_2(2) & \cdots & C_2(p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & C_n(1) & C_n(2) & \cdots & C_n(p) \end{bmatrix}$$

be the column vector of observations of X and the matrix with jth row containing a 1 together with ${C_j}^\top$.

Define $N = \inf\{n \geq 1 : \mathbf{C}_n \text{ has full column rank}\}$. Proposition 2 below shows that N is almost surely finite when Λ is nonsingular and

$$\mu_N = \bar{X}_N - {\theta_N}^\top \bar{C}_N = \mu$$

almost surely. Hence, if we know that a perfect control exists, then we can continue the simulation until time N and report $\bar{X}_N - {\theta_N}^\top \bar{C}_N$ 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. Before stating this result precisely we need a lemma.

Lemma 1 *The matrix $\mathbf{C}_n$ has full column rank if and only if Λ_n is nonsingular.*

Proof. It is well-known (e.g., Rice [1988, p. 477]) that $\mathbf{C}_n$ has full column rank if and only if ${\mathbf{C}_n}^\top \mathbf{C}_n$ is nonsingular. Define

$$\Sigma_n = \frac{1}{n}\sum_{i=1}^{n} C_i {C_i}^\top.$$

Then

$$\begin{aligned} {\mathbf{C}_n}^\top \mathbf{C}_n &= \begin{bmatrix} 1 & 1 & \dots & 1 \\ C_1 & C_2 & \dots & C_n \end{bmatrix} \begin{bmatrix} 1 & {C_1}^\top \\ 1 & {C_2}^\top \\ \vdots & \vdots \\ 1 & {C_n}^\top \end{bmatrix} \\ &= n \begin{bmatrix} 1 & {\bar{C}_n}^\top \\ \bar{C}_n & \Sigma_n \end{bmatrix}. \end{aligned} \tag{4}$$

Premultiplying ${\mathbf{C}_n}^\top \mathbf{C}_n$ by the nonsingular elementary matrix

$$B = \begin{bmatrix} 1 & 0 \\ -\bar{C}_n & I \end{bmatrix}$$

where I is the $p \times p$ identity matrix, we obtain

$$B{\mathbf{C}_n}^\top \mathbf{C}_n = n \begin{bmatrix} 1 & {\bar{C}_n}^\top \\ 0 & \Lambda_n \end{bmatrix},$$

which is nonsingular if and only if Λ_n is nonsingular. ■

We can now state the main result of this section.

Proposition 2 *Suppose that $X \in \mathbb{R}$ and $C \in \mathbb{R}^p$ have finite second moments, $EC = 0$, $\Lambda = \operatorname{cov}(C, C)$ is nonsingular and $X = C^\top \theta^* + \mu$ 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., $P(N > n) \le ar^n$ for some $a > 0$ and $r < 1$.*

Proof. From Lemma 1, N can be alternatively defined as

$$\inf\{n \ge 1 : \Lambda_n \text{ is nonsingular}\}. \tag{5}$$

Since Λ_n converges elementwise to Λ under the second moment assumption almost surely, it follows that N is finite almost surely.

Next, $X = C^\top \theta^* + \mu$ a.s., and so

$$\mathbf{X}_n = \mathbf{C}_n \begin{bmatrix} \mu \\ \theta^* \end{bmatrix} \tag{6}$$

almost surely, for any $n \geq 1$. The relation (6) also holds at time N, since

$$\begin{aligned} P\left(\mathbf{X}_N \neq \mathbf{C}_N \begin{bmatrix}\mu \\ \theta^*\end{bmatrix}\right) &= \sum_{n=1}^{\infty} P\left(\mathbf{X}_n \neq \mathbf{C}_n \begin{bmatrix}\mu \\ \theta^*\end{bmatrix}, N = n\right) \\ &\leq \sum_{n=1}^{\infty} P\left(\mathbf{X}_n \neq \mathbf{C}_n \begin{bmatrix}\mu \\ \theta^*\end{bmatrix}\right) \\ &= 0. \end{aligned}$$

Taking (6) at time N and premultiplying by ${\mathbf{C}_N}^\top$, we then get

$${\mathbf{C}_N}^\top \mathbf{X}_N = {\mathbf{C}_N}^\top \mathbf{C}_N \begin{bmatrix}\mu \\ \theta^*\end{bmatrix} \text{ a.s.}$$

If we use the representation (4) to expand out this relation we find that

$$\bar{X}_N = \mu + {\bar{C}_N}^\top \theta^* \text{ and} \tag{7}$$

$$\frac{1}{N}\sum_{i=1}^{N} C_i X_i = \bar{C}_N \mu + \Sigma_N \theta^* \tag{8}$$

almost surely. From (7), ${\bar{C}_N}^\top \theta^* = \bar{X}_N - \mu$ a.s., so that

$$\bar{C}_N {\bar{C}_N}^\top \theta^* = \bar{C}_N \bar{X}_N - \bar{C}_N \mu \text{ a.s.} \tag{9}$$

Adding (8) and (9) and rearranging we then see that

$$\Lambda_N \theta^* = \beta_N \text{ a.s.},$$

so that

$$\theta^* = \Lambda_N^{-1} \beta_N = \theta_N \text{ a.s.}$$

It follows from this relation and (7) that

$$\mu_N = \bar{X}_N - {\bar{C}_N}^\top \theta_N = \mu \text{ a.s.}$$

as claimed.

To prove the exponentially decaying tail property, note that $\mathbf{C}_n$ has full column rank if and only if at least $p+1$ of the vectors $C_1, \ldots, C_n$ are affinely independent [Bazaraa et al., 1993, p. 36]. Since Λ is nonsingular, it follows that there exist $p+1$ affinely-independent points $c_1, \ldots, c_{p+1}$ contained in the support of C_1. Now let $\epsilon > 0$ be such that the open balls $B(c_i, \epsilon)$ centered at c_i with radius ϵ are disjoint, and moreover if $x_i \in B(c_i, \epsilon)$ for all $i = 1, \ldots, p+1$, then $\{x_1, \ldots, x_{p+1}\}$ are affinely independent. Let $\tau_i = \inf\{k : C_k \in B(c_i, \epsilon)\}$, and let $N' = \max_i \tau_i$. Then at least $p+1$ of $C_1, \ldots, C_{N'}$ are affinely independent, and so $\mathbf{C}_{\mathbf{N}'}$ has full column rank. It follows that $N \leq N'$. Furthermore, $P(C_1 \in B(c_i, \epsilon)) > 0$ since c_i is contained in the support of C_1. Hence, each τ_i is a geometric random variable and therefore N' has a geometric tail. Since $N \leq N'$ this gives the result. ■

4 The Nonlinear Case: Preliminaries

We now turn to the case where $Y(\theta)$ is a nonlinear function of a random element Y and a parameter vector $\theta \in \Theta \subset \mathbb{R}^p$. 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

$$\{y : \exists x \text{ with } (x, y) \in H\},$$

i.e., the set of y values that appear in H.

Assumption A1 The parameter set Θ is compact. 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 a bounded open set containing Θ.

Assumption A2 The random variable X is square integrable. Also, for all $\theta \in \mathcal{U}$, $EY^2(\theta) < \infty$ and $EY(\theta) = Eh(Y, \theta) = 0$.

For convenience we define $X(\theta) = X - Y(\theta)$. Define

$$v(\theta) = \operatorname{var} X(\theta) = \operatorname{var} (X - Y(\theta))$$

to be the variance of the estimator 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 develop some machinery that will allow us to conclude that $v(\cdot)$ is differentiable.

Assumption A3 For all $y \in H_2$, $h(y, \cdot)$ is Lipschitz on $\mathcal{U}$, i.e. there exists $K(y) > 0$ such that for all $\theta_1, \theta_2 \in \mathcal{U}$,

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

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

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

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

Remark 1 *Recall that a $\mathcal{C}^1$ function is Lipschitz on a compact set. If $h(y, \cdot)$ is $\mathcal{C}^1$ on $\mathbb{R}^p$ (or on an open set containing the closure of $\mathcal{U}$), then **A3** is immediate.*

To establish the required differentiability we use the following result on Infinitesimal Perturbation Analysis (IPA) from L'Ecuyer [1995]. Let $f(\theta) = Ef(\theta, \xi)$ for some random variable ξ whose distribution does not depend on θ. The basic idea in IPA is to take $\nabla_\theta f(\theta, \xi)$, the gradient of $f(\theta, \xi)$ for fixed ξ, as an estimate of $\nabla_\theta f(\theta)$. This yields an unbiased estimator if the gradient and expectation can be exchanged. The following theorem gives sufficient conditions for the interchange to be valid. Since each component of the gradient can be dealt with separately, there is no loss of generality if we assume for the purposes of this theorem that $p = 1$.

Theorem 3 *[L'Ecuyer, 1995] Let $\theta_0 \in \Upsilon$, where Υ is an open interval, and let H be a measurable set such that $P(\xi \in H) = 1$. Suppose that for every $z \in H$, there is a $D(z)$, where $D(z)$ is at most countable, such that*

(i) $\forall z \in H$, $f(\cdot, z)$ *is continuous everywhere in* Υ,

(ii) $\forall z \in H$, $f(\cdot, z)$ *is differentiable everywhere in* $\Upsilon \backslash D(z)$,

(iii) there exists a function $\phi : H \to [0, \infty)$ *such that*

$$\sup_{\theta \in \Upsilon \backslash D(z)} |f'(\theta, z)| \leq \phi(z)$$

$\forall z \in H$ *with* $E\phi(\xi) < \infty$, *and*

(iv) $f(\theta, \xi)$ *is almost surely differentiable at* $\theta = \theta_0$, *i.e.,*

$$P\left(\xi \in \left\{ z : f'(\theta_0, z) = \lim_{\delta \to 0} \frac{f(\theta_0 + \delta, z) - f(\theta_0, z)}{\delta} \right\}\right) = 1.$$

Then $f(\cdot)$ *is differentiable at* $\theta = \theta_0$, *and*

$$f'(\theta_0) = Ef'(\theta_0, \xi).$$

Assumption **A1** implies that for each $(x, y) \in H$, $(x - h(y, \cdot) - \mu)^2$ is a $\mathcal{C}^1$ function on $\mathcal{U}$. Differentiation then gives that for $\theta \in \mathcal{U}$,

$$\nabla_\theta [(x - h(y, \theta) - \mu)^2] = -2(x - h(y, \theta) - \mu) \nabla_\theta h(y, \theta).$$

Therefore we have pathwise differentiability. We also need some integrability conditions.

Assumption A4 $E\left(K(Y)\left[1 + \sup_{\theta \in\, \mathcal{U}} |X(\theta)|\right]\right) < \infty$.

Proposition 4 *If* ***A1 - A4*** *hold then* $v(\cdot)$ *is* $\mathcal{C}^1$ *on* $\mathcal{U}$ *and*

$$\begin{aligned} g(\theta_0) &:= \nabla_\theta v(\theta)|_{\theta=\theta_0} \\ &= E\, \nabla_\theta (X(\theta) - \mu)^2 \big|_{\theta=\theta_0} \\ &= -2E\left[(X - h(Y, \theta_0) - \mu)(\nabla_\theta h(Y, \theta)|_{\theta=\theta_0})\right]. \end{aligned} \tag{10}$$

Proof. We apply Theorem 3 to each component separately, with $\xi = (X, Y)$ and $f(\theta, \xi) = (X(\theta) - \mu)^2$. The only condition of Theorem 3 that needs explicit verification is Condition (iii). To this end, fix attention on the jth component. By **A3**,

$$\begin{aligned} \left|\frac{\partial}{\partial \theta_j}(X(\theta) - \mu)^2\right| &= \left|2(X(\theta) - \mu)\frac{\partial}{\partial \theta(j)} X(\theta)\right| \\ &= 2\left|(X(\theta) - \mu)\frac{\partial h(Y, \theta)}{\partial \theta(j)}\right| \\ &\leq 2|X(\theta)|K(Y) + 2|\mu|K(Y) \\ &\leq 2K(Y) \sup_{\theta \in \mathcal{U}} |X(\theta)| + 2|\mu|K(Y). \end{aligned}$$

But this final expression is integrable, by **A4**. Thus $v(\theta)$ is differentiable at any $\theta_0 \in \mathcal{U}$, and we have the expression for the gradient $g(\theta)$ given in the statement of the result. It remains to establish that the

gradient is continuous in $\mathcal{U}$. But this follows almost immediately from **A3** and **A4** and the dominated convergence theorem. ■

Based on (10) we can estimate $g(\theta_0)$ via

$$\frac{-2}{n}\sum_{i=1}^{n}(X_i - h(Y_i,\theta_0) - \bar{X}_m)(\nabla_\theta h(Y_i,\theta)|_{\theta=\theta_0}),$$

where $(X_1,Y_1),...,(X_m,Y_m)$ are i.i.d. replications of (X,Y) and

$$\bar{X}_m = \frac{1}{m}\sum_{i=1}^{m} X_i.$$

In numerical experiments we found that this estimator has a large variance. It is also biased. An unbiased gradient estimator with greatly reduced variance can be obtained by noting that the sample variance of i.i.d. observations is an unbiased estimator of the variance, so that under **A2**, 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 $\theta \in \mathcal{U}$. 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 5 *If **A1-A4** hold, then for $\theta_0 \in \mathcal{U}$,*

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

Proof. We again apply Theorem 3 to the sample variance $V(m,\theta)$ component by component. Consider the jth component, for some $j = 1,\ldots,p$. The only condition that requires explicit verification is Condition (iii), where we require that $\partial V(m,\theta)/\partial\theta(j)$ be dominated by an integrable function of $(\mathbf{X},\mathbf{Y}) = ((X_i,Y_i) : 1 \leq i \leq m)$. We exploit the alternative formula

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

to find that

$$\frac{\partial V(m,\theta)}{\partial\theta(j)} = \frac{m}{m-1}\left(\frac{-1}{m}\sum_{i=1}^{m} 2X_i(\theta)\frac{\partial h(Y_i,\theta)}{\partial\theta(j)} + 2\bar{X}_m(\theta)\frac{1}{m}\sum_{i=1}^{m}\frac{\partial h(Y_i,\theta)}{\partial\theta(j)}\right) \tag{11}$$

The first term in the parentheses in (11) is integrable by **A4**. As for the second term, we apply **A3** and split the sums to obtain

$$\begin{aligned}
&\left| \bar{X}_m(\theta) \frac{1}{m} \sum_{i=1}^{m} \frac{\partial h(Y_i, \theta)}{\partial \theta(j)} \right| \\
&\leq \frac{1}{m^2} \sum_{i=1}^{m} \sup_{\theta \in \mathcal{U}} |X_i(\theta)| K(Y_i) + \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j \neq i} \sup_{\theta \in \mathcal{U}} |X_i(\theta)| K(Y_j).
\end{aligned}$$

If $E \sup_{\theta \in \mathcal{U}} |X_i(\theta)|$ is finite then **A4** implies integrability of this bound and the proof will be complete. Fix $\theta_0 \in \mathcal{U}$. By **A3**,

$$\begin{aligned}
|X_1(\theta)| &\leq |X_1| + |h(Y_1, \theta)| \\
&\leq |X_1| + |h(Y_1, \theta_0)| + |h(Y_1, \theta) - h(Y_1, \theta_0)| \\
&\leq |X_1| + |h(Y_1, \theta_0)| + K(Y_1) \|\theta - \theta_0\|.
\end{aligned}$$

But $\|\theta - \theta_0\|$ is bounded on the bounded set $\mathcal{U}$, and so $\sup_{\theta \in \mathcal{U}} |X_1(\theta)|$ is integrable. ■

So under the assumptions **A1 - A4**, the variance function $v(\theta)$ is continuously differentiable in $\theta \in \mathcal{U}$, 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. In the presence of nonconvexity the algorithm may only converge to a local minimum. 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

$$\theta_{n+1} = \Pi_\Theta(\theta_n - a_n g_n(\theta_n)), \tag{12}$$

where Π_Θ denotes a projection of points outside Θ back into Θ, and $\{a_n\}$ is a sequence of positive real numbers. The sequence $\{a_n\}$ is usually chosen to be of the form $a_n = a/n$ for all n, where a ia a positive scalar, although other forms have their merits. We use IPA to obtain $g_n(\theta_n)$, as discussed in the previous section.

Our algorithm for finding θ^* and estimating EX is as follows. Let $\{a_n\}_{n=1}^{\infty}$ be a sequence of positive numbers such that

$$\sum_{n=1}^{\infty} a_n = \infty \text{ and } \sum_{n=1}^{\infty} a_n^2 < \infty, \tag{13}$$

and let $m \geq 2$ be a fixed positive integer.

Stochastic Approximation

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

The analysis below is based on martingale theory. We first show consistency of the estimator μ_n of μ. We apply the following martingale strong law of large numbers which can be found in Liptser and Shiryayev [1989, p. 144]. Let $(\mathcal{F}_n : n \geq 0)$ be a filtration, i.e. an increasing sequence of σ-fields. A martingale sequence $(M_n, \mathcal{F}_n : n \geq 1)$ is a collection of integrable random variables M_n such that M_n is measurable with respect to $\mathcal{F}_n$ and $E(M_n|\mathcal{F}_{n-1}) = M_{n-1}$ for all $n \geq 1$.

Theorem 6 (Liptser and Shiryayev 1989) *Let $(M_n, \mathcal{F}_n : n \geq 0)$ be a martingale with $M_0 = 0$ and $EM_n^2 < \infty$ for all n. Let $(L_n : n \geq 0)$ be a sequence of random variables such that $L_0 = 0$, L_n is measurable with respect to $\mathcal{F}_n$ for all n and for any fixed sample path, L_n is a nondecreasing function of n. Define*

$$V_n = \sum_{k=1}^{n} E((M_k - M_{k-1})^2|\mathcal{F}_{k-1})$$

and assume that

$$\sum_{n=1}^{\infty}\frac{V_{n+1} - V_n}{(1 + L_n)^2} < \infty \text{ a.s. and}$$

$$P(L_\infty = \infty) = 1,$$

where $L_\infty = \lim_{n\to\infty} L_n$. Then

$$\frac{M_n}{L_n} \to 0 \text{ a.s.}$$

Let $\mathcal{F}_n = \sigma\{(X_{k,i}, Y_{k,i}) : 1 \leq k \leq n, 1 \leq i \leq m\}$ be the sigma field containing the information from the first n steps of the stochastic algorithm. Let $\mathcal{F}_0$ be the trivial sigma field and θ_0 be any deterministic guess for θ^*. (If θ_0 is not deterministic then we can extend $\mathcal{F}_0$ appropriately, so there is no loss of generality in this convention.)

Proposition 7 *Assume **A1-A4**. Then $\mu_n \to \mu$ a.s. as $n \to \infty$.*

Proof. For $k \geq 1$ and $n \geq 1$, define

$$\begin{aligned}\zeta_k(\theta_{k-1}) &= A_k(\theta_{k-1}) - \mu \text{ and}\\ M_n &= \sum_{k=1}^{n}\zeta_k(\theta_{k-1}).\end{aligned}$$

Then

$$\mu_n = \mu + \frac{M_n}{n},$$

and hence it suffices to show that $M_n/n \to 0$ a.s. as $n \to \infty$. We apply Theorem 6.

Define $M_0 = 0$. Since $E(\zeta_k(\theta_{k-1})|\mathcal{F}_{k-1}) = 0$ for all $k \geq 1$, $(M_n, \mathcal{F}_n : n \geq 0)$ is a martingale. Moreover, for all $n \geq 1$,

$$\begin{aligned}
E(M_n^2) &= \sum_{k=1}^{n} E(\zeta_k^2(\theta_{k-1})) && (14)\\
&= \sum_{k=1}^{n} \text{var}\,(A_k(\theta_{k-1})) \\
&= \sum_{k=1}^{n} (E\text{var}\,[A_k(\theta_{k-1})|\mathcal{F}_{k-1}] + \text{var}\,E[A_k(\theta_{k-1})|\mathcal{F}_{k-1}]) \\
&= \sum_{k=1}^{n} \frac{1}{m}\left(E[(X_{k,1} + h(Y_{k,1}, \theta_{k-1}) - \mu)^2|\mathcal{F}_{k-1}] + 0\right) \\
&= \sum_{k=1}^{n} \frac{1}{m} E(v(\theta_{k-1})) < \infty && (15)
\end{aligned}$$

where (14) follows from the fact that for $j \neq l$,

$$E(\zeta_j(\theta_{j-1})\zeta_l(\theta_{l-1})) = 0$$

and (15) follows from the fact that $v(\cdot)$ is continuous on the compact set Θ and therefore $E(v(\theta_{k-1}))$ is bounded. Define $L_n = n$ for all $n \geq 0$ and

$$V_n = \sum_{k=1}^{n} E((M_k - M_{k-1})^2|\mathcal{F}_{k-1}) = \sum_{k=1}^{n} E(\zeta_k^2(\theta_{k-1})|\mathcal{F}_{k-1}) = \frac{1}{m}\sum_{k=1}^{n} v(\theta_{k-1}).$$

Then $P(L_\infty = \infty) = 1$ and

$$\sum_{n=1}^{\infty} \frac{V_{n+1} - V_n}{(1+L_n)^2} = \frac{1}{m}\sum_{n=1}^{\infty} \frac{v(\theta_n)}{(1+n)^2} \leq \frac{\sup_{\theta\in\Theta} v(\theta)}{m} \sum_{n=1}^{\infty} \frac{1}{(1+n)^2} < \infty \text{ a.s.}$$

Therefore, by Theorem 6, $M_n/n \to 0$ a.s. as $n \to \infty$. ■

Remark 2 *The proof of Proposition 7 is based on the square integrability of $X_1(\cdot)$ and the continuity of $v(\cdot)$ on Θ. The square-integrability condition may seem too strong. But if $\theta_k \to \theta^*$ for some random variable θ^* and $\{\theta^*(\omega)|\omega \in \Omega\}$ is a countable set, then under the Lipschitz continuity of $h(y,\cdot)$ and finite* first *moment conditions, μ_n is still strongly consistent.*

We now assess the *rate* of convergence of μ_n through a central limit theorem. We need the following martingale central limit theorem which can be found in Liptser and Shiryayev [1989, p. 444]. A martingale difference sequence $(\xi_{k,n}, \mathcal{F}_{k,n} : n \geq 1, 1 \leq k \leq n)$ is a collection of mean-zero random variables $\xi_{k,n}$ and filtrations $(\mathcal{F}_{k,n} : k = 1, \ldots, n)$ such that $\xi_{k,n}$ is measurable with respect to $\mathcal{F}_{k,n}$ for all $n \geq 1$ and $1 \leq k \leq n$, and $E(\xi_{k,n}|\mathcal{F}_{k-1,n}) = 0$ for all $n \geq 1$ and $k = 1, \ldots, n$. Here we have adopted the convention that $\mathcal{F}_{0,n}$ is the trivial sigma field for all $n \geq 1$, so that θ_0 is a deterministic approximation for θ^*.

Theorem 8 (Liptser and Shiryayev 1989) *Assume that* $(\mathcal{F}_{k,n} : 1 \leq k \leq n, n \geq 1)$ *is nested i.e.* $\mathcal{F}_{k,n} \subseteq \mathcal{F}_{k,n+1}$, *for all* $k \leq n$, $n \geq 1$. *Let* η^2 *be a* $\mathcal{G}$*-measurable random variable where*

$$\mathcal{G} \subseteq \sigma(\bigcup_{n\geq 1} \mathcal{F}_{n,n}).$$

Let Z *be a random variable with characteristic function*

$$E(e^{itZ}) = E\left[\exp\left(-\frac{t^2}{2}\eta^2\right)\right], t \in \mathbb{R},$$

so that Z *is a mixture of mean-zero normal random variables. Let* $(\xi_{k,n}, \mathcal{F}_{k,n} : n \geq 1, 1 \leq k \leq n)$ *be a martingale difference sequence with* $E(\xi_{k,n}^2) < \infty$, *for all* $n \geq 1, 1 \leq k \leq n$. *Assume that*

(i) $\sum_{k=1}^n E(\xi_{k,n}^2 I(|\xi_{k,n}| > \delta)|\mathcal{F}_{k-1,n}) \to 0$ *in probability, for all* $\delta \in (0, 1]$,

(ii) $\sum_{k=1}^n E(\xi_{k,n}^2|\mathcal{F}_{k-1,n}) \to \eta^2$ *in probability, and*

(iii) $\sum_{k=1}^{\lfloor nc_n \rfloor} E(\xi_{k,n}^2|\mathcal{F}_{k-1,n}) \to 0$ *in probability*

for a certain sequence $(c_n)_{n\geq 1}$ *with* $c_n \downarrow 0$, $nc_n \to \infty$ *as* $n \to \infty$. *Then*

$$S_n = \sum_{k=1}^n \xi_{k,n} \Rightarrow Z$$

as $n \to \infty$, *where* $\Rightarrow$ *denotes convergence in distribution.*

The central limit theorem below assumes that θ_n converges to some random variable θ^* a.s. 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 this to be the case. Before stating the central limit theorem we need another assumption. Let

$$E = \{\omega : \theta_k(\omega) \to \theta^*(\omega) \text{ as } k \to \infty\}$$

so that $P(E) = 1$ and

$$\Gamma = \{\theta^*(\omega) = \lim_{k\to\infty} \theta_k(\omega) : \omega \in E\} \subseteq \Theta.$$

Assumption A5 For any $\gamma \in \Gamma$, there is a neighbourhood $\mathcal{N}(\gamma)$ of γ such that the collection $\{X^2(\theta) : \theta \in \mathcal{N}(\gamma)\}$ is uniformly integrable. In other words, for all $\epsilon > 0$, there exists $K_\epsilon > 0$ such that

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

Remark 3 A set of sufficient conditions for **A5** is **A1-A3** and $EK^2(Y) < \infty$. To see why, note that

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

For any fixed $\theta_0 \in \mathcal{U}$,

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

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

Theorem 9 *Assume **A1-A5** and that $\theta_n \to \theta^*$ for some random variable θ^* a.s. as $n \to \infty$. Let Z be a random variable with characteristic function*

$$E(e^{itZ}) = E\left[\exp\left(-\frac{t^2}{2}v(\theta^*)\right)\right], t \in \mathbb{R},$$

i.e., $Z = v^{1/2}(\theta^*)N(0,1)$ is a mixture of mean-zero normal random variables. Then

$$\sqrt{mn}(\mu_n - \mu) \Rightarrow Z$$

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

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

as $n \to \infty$.

Proof. To show the central limit theorem we apply Theorem 8. Let

$$\xi_{k,n} = \frac{\sqrt{m}(A_k(\theta_{k-1}) - \mu)}{\sqrt{n}}$$

so that

$$\sqrt{mn}(\mu_n - \mu) = \sum_{k=1}^{n} \xi_{k,n}.$$

Notice that

$$E[\xi_{k,n}|\mathcal{F}_{k-1}] = \sqrt{\frac{m}{n}}E[A_k(\theta_{k-1}) - \mu|\mathcal{F}_{k-1}) = 0. \tag{16}$$

Moreover, as in (15)

$$E\xi_{k,n}^2 = \frac{m}{n}\text{var}\,A_k(\theta_{k-1}) = \frac{Ev(\theta_{k-1})}{n} < \infty, \tag{17}$$

so that $(\xi_{k,n}, \mathcal{F}_{k,n} : n \geq 1, 1 \leq k \leq n)$ is a martingale difference sequence with $E\xi_{k,n}^2 < \infty$, where $\mathcal{F}_{k,n} = \mathcal{F}_k$ for all n. Then $(\mathcal{F}_{k,n})$ is nested and θ^* is $\mathcal{F}_\infty$-measurable, where $\mathcal{F}_\infty = \sigma(\bigcup_{k\geq 0}\mathcal{F}_k)$. Fix $\delta > 0$ and let

$$W_n = \sum_{k=1}^{n} E(\xi_{k,n}^2 I(|\xi_{k,n}| > \delta)|\mathcal{F}_{k-1,n}).$$

If $\zeta_k(\theta) = A_k(\theta) - \mu$, then

$$\begin{aligned} W_n &= \frac{m}{n}\sum_{k=1}^{n} E[\zeta_k^2(\theta_{k-1})I(\zeta_k^2(\theta_{k-1}) > n\delta^2/m)|\mathcal{F}_{k-1,n}] \\ &= \frac{m}{n}\sum_{k=1}^{n} E[\zeta_k^2(\theta_{k-1})I(\zeta_k^2(\theta_{k-1}) > n\delta^2/m)|\theta_{k-1}] \\ &= \frac{m}{n}\sum_{k=1}^{n} f(\theta_{k-1}, n\delta^2/m), \end{aligned}$$

where

$$f(\theta, b) = E[\zeta_1^2(\theta) I(\zeta_1^2(\theta) > b)].$$

Assumption **A5** implies that for any $\omega \in E$, the collection $(\zeta_1^2(\theta) : \theta \in \mathcal{N}(\theta^*(\omega)))$ is also uniformly integrable and so for all $\epsilon > 0$, there exists $K_\epsilon > 0$ such that $f(\theta, K_\epsilon) \leq \epsilon$ for all $\theta \in \mathcal{N}(\theta^*(\omega))$. Fix $\omega \in E$ and $\epsilon > 0$. Let $n_1 \geq 1$ be such that $\theta_n \in \mathcal{N}(\theta^*)$ for all $n \geq n_1$ and let $n_2 \geq 1$ be such that $n\delta^2/m \geq K_\epsilon$ for all $n \geq n_2$. Let $n^* = \max\{n_1, n_2\} + 1$. Then

$$\begin{aligned} W_n &= \frac{m}{n} \sum_{k=1}^{n} f(\theta_{k-1}, n\delta^2/m) \\ &= \frac{m}{n} \sum_{k=1}^{n^*} f(\theta_{k-1}, n\delta^2/m) + \frac{m}{n} \sum_{k=n^*+1}^{n} f(\theta_{k-1}, n\delta^2/m) \\ &\leq \frac{m}{n} \sum_{k=1}^{n^*} f(\theta_{k-1}, 0) + \frac{m}{n} \sum_{k=n^*+1}^{n} f(\theta_{k-1}, K_\epsilon). \end{aligned}$$

Hence

$$0 \leq \limsup_{n\to\infty} W_n \leq 0 + \limsup_{n\to\infty} \frac{m}{n} \sum_{k=n^*+1}^{n} \epsilon = m\epsilon.$$

Since $\omega \in E$ and ϵ were arbitrary, we conclude that $W_n \to 0$ as $n \to \infty$ a.s.

The second and third conditions of Theorem 8 are easily dealt with. We see that

$$\sum_{k=1}^{n} E(\xi_{k,n}^2 | \mathcal{F}_{k-1}) = \sum_{k=1}^{n} \frac{m}{n} E((A_k(\theta_{k-1}) - \mu)^2 | \mathcal{F}_{k-1}) = \frac{1}{n} \sum_{k=1}^{n} v(\theta_{k-1}).$$

But $\theta_{k-1}(\omega) \to \theta^*(\omega)$ and $v(\cdot)$ is continuous at $\theta^*(\omega)$ for all $\omega \in E$, and so

$$\frac{1}{n} \sum_{k=1}^{n} v(\theta_{k-1}) \to v(\theta^*)$$

as $n \to \infty$ a.s. For the third condition, let $c_n = n^{-1/2}$. Then

$$\sum_{k=1}^{\lfloor nc_n \rfloor} E(\xi_{k,n}^2 | \mathcal{F}_{k-1}) = \frac{1}{n} \sum_{k=1}^{\lfloor n^{1/2} \rfloor} v(\theta_{k-1}) \leq \frac{n^{1/2} \sup_{\theta\in\Theta} v(\theta)}{n} \to 0$$

as $n \to \infty$. The central limit theorem is therefore a consequence of Theorem 8.

The fact that $E\mu_n = \mu$ is an immediate consequence of (16). It remains to establish the variance result. From (17) and the fact that the $\xi_{k,n}$'s are martingale differences we see that

$$\operatorname{var} \mu_n = \frac{1}{mn^2} \sum_{k=1}^{n} Ev(\theta_{k-1}).$$

But $v(\theta_n) \to v(\theta^*)$ as $n \to \infty$ a.s., and the sequence $(v(\theta_n) : n \geq 1)$ is bounded and therefore uniformly integrable. Thus $Ev(\theta_n) \to Ev(\theta^*)$ and

$$mn \operatorname{var} \mu_n \to E(v(\theta^*))$$

as $n \to \infty$. ■

Hence we see that the stochastic approximation estimator μ_n satisfies a strong law and central limit theorem as $n \to \infty$. It will almost invariably be the case that $v(\theta^*) > 0$ a.s. so that the rate of convergence of μ_n is the canonical rate $n^{-1/2}$. This is the best that can be hoped for with the Monte Carlo nature of the estimation procedure we used.

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. Therefore 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 $v(\theta^*) = 0$ a.s. 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 $\theta_n \to \theta^*$ a.s. We do not explore this case further here, because we believe that the case $v(\theta^*) = 0$ a.s. is unlikely to arise in the applications we have in mind. See Henderson and Simon [2004] for an exploration of increased convergence rates when θ^* is constant and $v(\theta^*) = 0$.

The central limit theorem suggests a confidence interval procedure, provided that the variance can be estimated. Suppose that $\theta_k \to \theta^*$ a.s. for some fixed $\theta^* \in \Theta$, so that the variance appearing in the central limit theorem is deterministic and equal to $v(\theta^*)$. To estimate $v(\theta^*)$ we can use any one of the three estimators

$$\begin{aligned} S_n^2 &= \frac{1}{mn-1} \sum_{k=1}^{n} \sum_{i=1}^{m} \left(X_{k,i}(\theta_{k-1}) - \mu_n \right)^2, \\ \hat{S}_n^2 &= \frac{1}{n} \sum_{k=1}^{n} \left(\frac{1}{m-1} \sum_{i=1}^{m} (X_{k,i}(\theta_{k-1}) - A_k(\theta_{k-1}))^2 \right), \text{ and} \\ \tilde{S}_n^2 &= \frac{m}{n-1} \sum_{k=1}^{n} (A_k(\theta_{k-1}) - \mu_n)^2. \end{aligned}$$

The estimator S_n^2 is the sample variance using all mn samples, $\hat{S}_n^2$ is the average of the sample variances of m terms in each iteration, and $\tilde{S}_n^2$ is m times the sample variance of the averages computed at each iteration. The following proposition shows that all three estimators are strongly consistent, so they can be used to construct asymptotically valid confidence intervals. The proof appears in the appendix.

Proposition 10 *Assume **A1-A4** and that θ_n converges to some fixed $\theta^* \in \Theta$ a.s. Then*

(i) $S_n^2, \hat{S}_n^2, \tilde{S}_n^2 \to v(\theta^*)$ as $n \to \infty$ a.s.

*(ii) Assume also **A5**. Then*

$$\frac{\sqrt{nm}(\mu_n - \mu)}{\eta_n} \Rightarrow N(0,1)$$

as $n \to \infty$, where η_n can be $S_n, \hat{S}_n$ or $\tilde{S}_n$.

Under the conditions of Proposition 10(ii), an asymptotic $100(1-\alpha)\%$ confidence interval for μ is

$$\left[\mu_n - z\frac{\eta_n}{\sqrt{nm}}, \mu_n + z\frac{\eta_n}{\sqrt{nm}} \right],$$

where η_n can be $S_n, \hat{S}_n$ or $\tilde{S}_n$ and z is chosen such that $P(-z \leq N(0,1) \leq z) = 1 - \alpha$.

But which variance estimator should we use? Some insight into this question can be obtained by assuming that $\theta_k = \theta^*$ for all k, and then considering the second-order behavior of the variance estimators as given by central limit theorems. This case is easier to analyze than the general case because the $X_{k,i}(\theta^*)$s are i.i.d. random variables. The proof of the following result is given in the appendix.

Proposition 11 *Suppose that $\theta_k = \theta^*$ for all $k \geq 0$. Suppose that $EX^4(\theta^*) < \infty$. Then*

$$\begin{aligned}
\sqrt{mn}(S_n^2 - v(\theta^*)) &\Rightarrow \sigma N(0,1), \\
\sqrt{mn}(\hat{S}_n^2 - v(\theta^*)) &\Rightarrow \hat{\sigma} N(0,1), \text{ and} \\
\sqrt{mn}(\tilde{S}_n^2 - v(\theta^*)) &\Rightarrow \tilde{\sigma} N(0,1)
\end{aligned}$$

as $n \to \infty$, where

$$\begin{aligned}
\sigma^2 &= E[X_1(\theta^*) - \mu]^4 - v^2(\theta^*), \\
\hat{\sigma}^2 &= E[X_1(\theta^*) - \mu]^4 - \frac{m-3}{m-1} v^2(\theta^*), \text{ and} \\
\tilde{\sigma}^2 &= E[X_1(\theta^*) - \mu]^4 + (2m-3) v^2(\theta^*).
\end{aligned}$$

Notice that $\tilde{\sigma}^2 > \sigma^2, \hat{\sigma}^2$ for $m \geq 2$, so on that basis we prefer either S_n^2 or $\hat{S}_n^2$ to $\tilde{S}_n^2$. The difference between σ^2 and $\hat{\sigma}^2$ is much smaller and vanishes as m grows. So the choice between these estimators essentially comes down to computational convenience, so long as m is large enough. We used $\hat{S}_n^2$ in our experiments.

We now give conditions under which θ_n converges to some random variable θ^* a.s. as $n \to \infty$. Theorem 12 below is an immediate specialization of Kushner and Yin [2003, Theorem 2.1, p. 127]. We first need some definitions.

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 $f : B \to \mathbb{R}$ satisfies

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

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

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

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

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

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

Let $(\mathcal{G}_n : n \geq 0)$ be a filtration, where the initial guess θ_0 is measurable with respect to $\mathcal{G}_0$ and G_n (an estimate for the gradient of f at θ_n) is measurable with respect to $\mathcal{G}_{n+1}$ for all $n \geq 0$.

Theorem 12 *Let B be a box in $\mathbb{R}^p$ and $f : \mathbb{R}^p \to \mathbb{R}$ be $\mathcal{C}^1$. Suppose that for $n \geq 0$, $\theta_{n+1} = \Pi_B(\theta_n - a_n G_n)$ with the following additional conditions.*

(i) The conditions (13) hold.

(ii) $\sup_n E\|G_n\|^2 < \infty$.

(iii) $E[G_n|\mathcal{G}_n] = \nabla f(\theta_n)$ *for all* $n \geq 0$.

Then,

$$d(\theta_n, S(f, B)) \to 0$$

as $n \to \infty$ a.s. Moreover, suppose that $S(f, B)$ is a discrete set. Then, on almost all sample paths, θ_n converges to a unique point in $S(f, B)$ as $n \to \infty$.

The limiting points in $S(f, B)$ can be random. We can apply Theorem 12 in our context, but first we need one more assumption.

Assumption A6 The random variables X, $K(Y)$ and $Y(\theta_0)$, for some fixed $\theta_0 \in \Theta$, all have finite 4th moments.

When **A1-A3** and **A6** hold, $EY^4(\theta)$ is bounded in $\theta \in \Theta$, since the Lipschitz condition gives

$$\begin{aligned} |Y(\theta)| &\leq |h(Y, \theta_0)| + |h(Y, \theta) - h(Y, \theta_0)| \\ &\leq |h(Y, \theta_0)| + K(Y)\|\theta - \theta_0\| \end{aligned}$$

so that

$$Y^4(\theta) \leq ch^4(Y, \theta_0) + cK^4(Y)\|\theta - \theta_0\|^4$$

for some constant c. (Compactness of Θ ensures that $\|\theta - \theta_0\|^4$ is bounded.)

Corollary 13 *Let Θ be a box in $\mathbb{R}^p$ and suppose **A1 - A4, A6** hold. Then $d(\theta_n, S(v, \Theta)) \to 0$ as $n \to \infty$ a.s. Moreover, suppose that $S(v, \Theta)$ is a discrete set. Then, on almost all sample paths, θ_n converges to a unique point in $S(v, \Theta)$ as $n \to \infty$.*

Proof. The only condition of Theorem 12 that needs verification is the condition $\sup_n E\|G_n\|^2 < \infty$. In our case, $G_n = g_n(\theta_n)$, and

$$\|g_n(\theta_n)\|^2 \leq \sup_{\theta \in \Theta} \|g_n(\theta)\|^2.$$

But the distribution of $g_n(\theta)$ does not depend on n, so the result follows if

$$\sup_{\theta \in \Theta} E\|g_1(\theta)\|^2 < \infty.$$

The argument is similar to that used in Propositon 5 and is omitted. It is this argument that requires the stronger moment assumption **A6**. ■

Corollary 13 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.

Corollary 14 *In addition to the conditions of Corollary 13 suppose that $S(v, \Theta)$ consists of a single element θ^*. Then $\theta_n \to \theta^*$ as $n \to \infty$ a.s.*

We will see in Section 7 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

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} [X_i - h(Y_i, \hat{\theta})]. \tag{18}$$

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 15 *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 [2004] 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 $(\tilde{X}_i(\theta) : 1 \leq i \leq m)$ is

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

where

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

Then an approximation to 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

$$V(m, \theta) = \frac{m}{m-1} \left(\frac{1}{m} \sum_{i=1}^{m} \tilde{X}_i^2(\theta) - \bar{X}_m^2(\theta) \right). \tag{19}$$

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

Let $\hat{\theta}_m$ be a first-order critical point for problem $\mathcal{P}_m$. We can then estimate μ via (18), using $\hat{\theta}_m$ in place of $\hat{\theta}$. 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 15 hold. Nevertheless, versions of these results *do* hold, and can be shown using a uniform version of the strong law and some straightforward arguments.

Proposition 16 is a uniform version of the strong law and appears as Proposition 7 in Shapiro [2004]. We say that $f(y,\theta)$ is *dominated* by an integrable function $f(\cdot)$ if $Ef(Y) < \infty$ and for every $\theta \in \Theta$, $|f(Y,\theta)| \leq f(Y)$ a.s.

Proposition 16 (Shapiro 2003) *Suppose that for every $y \in H_2$, the function $f(y,\cdot)$ is continuous on (the compact set) Θ, and $f(y,\theta)$ is dominated by an integrable function. Then $Ef(Y,\theta)$ is continuous as a function of $\theta \in \Theta$ and*

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} f(Y_i,\theta) - Ef(Y,\theta)\right| \to 0$$

as $n \to \infty$ a.s.

We can now state a version of Theorem 15 for the case where $\hat{\theta}$ is random. 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 17 *Suppose that **A1-A3** 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 $\hat{\theta}$.

Proof. For the strong law note that

$$\begin{aligned} |\hat{\mu}_n(\hat{\theta}) - \mu| &\leq \left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\right| + \left|\frac{1}{n}\sum_{i=1}^{n} h(Y_i,\hat{\theta})\right| \\ &\leq \left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)\right| + \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} h(Y_i,\theta)\right|. \end{aligned} \tag{20}$$

The first term in (20) converges to 0 as $n \to \infty$ by the strong law of large numbers. The second term converges to 0 by an application of Theorem 16. To see why $h(Y,\theta)$ is dominated by an integrable function fix $\theta_0 \in \Theta$ and then

$$h(Y,\theta) = h(Y,\theta_0) + (h(Y,\theta) - h(Y,\theta_0)),$$

so that

$$\begin{aligned} |h(Y,\theta)| &\leq |h(Y,\theta_0)| + |h(Y,\theta) - h(Y,\theta_0)| \\ &\leq |h(Y,\theta_0)| + K(Y)\|\theta - \theta_0\|. \end{aligned} \tag{21}$$

But Θ is compact and so $\|\theta - \theta_0\|$ is bounded. Hence, (21) is bounded by an integrable random variable uniformly in $\theta \in \Theta$, which is the required domination.

For the central limit theorem, first note that conditional on $\hat{\theta}$, μ_n is an average of i.i.d. random variables with finite variance. Hence the ordinary central limit theorem ensures that for each fixed $x \in \mathbb{R}$,

$$P\left(\sqrt{n}(\hat{\mu}_n(\hat{\theta}) - \mu) \leq x \,|\, \hat{\theta}\right) \to \Phi\left(\frac{x}{v^{1/2}(\hat{\theta})}\right) I(v(\hat{\theta}) > 0) + I(x \geq 0) I(v(\hat{\theta}) = 0) \tag{22}$$

as $n \to \infty$, where Φ is the distribution function of a normal random variable with mean 0 and variance 1 and $I(\cdot)$ is an indicator function. The dominated convergence theorem ensures that we can take expectations through (22), and so

$$\begin{aligned} & P(\sqrt{n}(\hat{\mu}_n(\hat{\theta}) - \mu) \leq x) \\ \to \;& E\left[\Phi\left(\frac{x}{v^{1/2}(\hat{\theta})}\right) I(v(\hat{\theta}) > 0) + I(x \geq 0) I(v(\hat{\theta}) = 0)\right] \\ = \;& P(v^{1/2}(\hat{\theta}) N(0,1) \leq x) \end{aligned}$$

for all $x \in \mathbb{R}$, which is the desired central limit theorem. ■

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.

Now, 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. This can be modelled 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 18 *Suppose that $\hat{\theta}_{m(n)} \to \theta^*$ as $n \to \infty$ a.s., for some random variable θ^*. 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 $n\mathrm{var}\,\hat{\mu}_n(\hat{\theta}_{m(n)}) \to E(v(\theta^*))$ as $n \to \infty$. If, in addition, $EK^2(Y) < \infty$, then*

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

as $n \to \infty$.

Proof. Proofs of the strong law, unbiasedness, and asymptotic expression for the variance are very similar to the analogous results in the previous section and therefore omitted.

To prove the central limit theorem, note that

$$\begin{aligned} \sqrt{n}(\hat{\mu}_n(\hat{\theta}_{m(n)}) - \mu) &= \sqrt{n}(\hat{\mu}_n(\theta^*) - \mu) + \sqrt{n}(\hat{\mu}_n(\hat{\theta}_{m(n)}) - \hat{\mu}(\theta^*)) \\ &= D_{1,n} - D_{2,n}, \text{ say.} \end{aligned}$$

Notice that θ^* is independent of the samples to compute $\hat{\mu}_n$ for every n. By Theorem 17, $D_{1,n} \Rightarrow v^{1/2}(\theta^*) N(0,1)$ as $n \to \infty$. Thus, it suffices to show that

$$D_{2,n} = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} [h(Y_j, \hat{\theta}_{m(n)}) - h(Y_j, \theta^*)] \Rightarrow 0$$

as $n \to \infty$.

Chebyshev's inequality ensures that for any fixed $\epsilon > 0$

$$\begin{aligned} P(|D_{2,n}| > \epsilon) &\leq \epsilon^{-2} E D_{2,n}^2 \\ &= \frac{1}{n\epsilon^2} E\left[\sum_{j=1}^{n} [h(Y_j, \hat{\theta}_{m(n)}) - h(Y_j, \theta^*)]\right]^2 \\ &= \frac{1}{n\epsilon^2} \sum_{j=1}^{n} E[h(Y_j, \hat{\theta}_{m(n)}) - h(Y_j, \theta^*)]^2 \qquad (23) \\ &= \frac{1}{\epsilon^2} E[h(Y_1, \hat{\theta}_{m(n)}) - h(Y_1, \theta^*)]^2, \qquad (24) \end{aligned}$$

where (23) follows from the fact that for $i \neq j$,

$$E[h(Y_i, \hat{\theta}_{m(n)}) - h(Y_i, \theta^*)][h(Y_j, \hat{\theta}_{m(n)}) - h(Y_j, \theta^*)] = 0.$$

Now,

$$[h(Y_1, \hat{\theta}_{m(n)}) - h(Y_1, \theta^*)]^2 \to 0$$

as $n \to \infty$ a.s. Moreover,

$$[h(Y_1, \hat{\theta}_{m(n)}) - h(Y_1, \theta^*)]^2 \leq K^2(Y_1) \, \|\hat{\theta}_{m(n)} - \theta^*\|^2. \qquad (25)$$

But $\hat{\theta}_{m(n)}$ and θ^* are contained in Θ which is compact. Hence the normed term in (25) is bounded. The dominated convergence theorem then implies that (24) converges to 0 as $n \to \infty$ and the central limit theorem is established. ■

It remains to give conditions under which $\hat{\theta}_m \to \theta^*$ as $m \to \infty$ a.s. If we could guarantee that $\hat{\theta}_m$ solved problem $\mathcal{P}_m$ exactly then, as in Shapiro [2004], this would follow using standard arguments and an extension of a uniform law of large numbers to nonlinear functions of means. (Recall from (19) 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 et al. [2004] for sample averages to nonlinear functions of sample averages.

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

Theorem 19 *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 $\xi \in H$, $f(\cdot, \xi) = (f_1(\cdot, \xi), \ldots, f_d(\cdot, \xi))$ is $\mathcal{C}^1$ on an open set containing Θ,

(ii) the component functions $f_j(\theta, \xi)$ $(j = 1, \ldots, d)$ 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 ($i=1,\ldots,p,\ j=1,\ldots,d$).

Suppose that $g(x)$ is a real-valued $\mathcal{C}^1$ fuction of $x\in D\subset\mathbb{R}^d$, where D is an open set containing the range of $\bar{f}$ and $\bar{f}_m$ for all m. Let $\hat{\theta}_m\in S(g\circ\bar{f}_m,\Theta)$ be the set of first-order critical points of $g\circ\bar{f}_m$ on Θ. Then $d(\hat{\theta}_m,S(g\circ\bar{f},\Theta))\to 0$ as $m\to\infty$ a.s.

Proof. If $d(\hat{\theta}_m,S(g\circ\bar{f},\Theta))\not\to 0$, then by passing to a subsequence if necessary, we can assume that for some $\epsilon>0$, $d(\hat{\theta}_m,S(g\circ\bar{f},\Theta))\geq\epsilon$ for all $m\geq 1$. Since Θ is compact, by passing to a further subsequence if necessary, we can assume that $\hat{\theta}_m$ converges to a point $\theta^*\in\Theta$. It follows that $\theta^*\notin S(g\circ\bar{f},\Theta)$. On the other hand, by Propositon 16, $\bar{f}_m(\hat{\theta}_m)\to\bar{f}(\theta^*)$ and $\nabla_\theta\bar{f}_m(\hat{\theta}_m)\to\nabla_\theta\bar{f}(\theta^*)$ as $m\to\infty$ a.s.

Since Θ is convex, each $\hat{\theta}_m$ satisfies the first order condition

$$\langle g'(\bar{f}_m(\hat{\theta}_m))\nabla_\theta\bar{f}_m(\hat{\theta}_m),u-\hat{\theta}_m\rangle\geq 0,\ \text{for all } u\in\Theta,\ \text{a.e.}$$

Taking the limit as $m\to\infty$, we obtain that

$$\langle g'(\bar{f}(\theta^*))\nabla_\theta\bar{f}(\theta^*),u-\theta^*\rangle\geq 0,\ \text{for all } u\in\Theta,\ \text{a.e.}$$

Therefore, $\theta^*\in S(g\circ\bar{f},\Theta)$ and we obtain a contradiction. ■

We now obtain the following corollary.

Corollary 20 *Suppose that **A1-A4** hold, Θ is convex and $EK^2(Y)<\infty$. Then $d(\hat{\theta}_m,S(v,\Theta))\to 0$ as $m\to\infty$ a.s.*

Proof. If $g(x,y)=x-y^2$, then

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

Notice that

$$S(V(m,\cdot),\Theta)=S(g\left(\frac{1}{m}\sum_{i=1}^m X_i^2(\cdot),\frac{1}{m}\sum_{i=1}^m X_i(\cdot)\right),\Theta),$$

i.e., the sets of first-order critical points of these two functions coincide.

By the proof of Proposition 5 and Remark 3,

$$X(\theta),X^2(\theta),\frac{\partial h(Y,\theta)}{\partial\theta(i)}\ \text{and}\ 2X(\theta)\frac{\partial h(Y,\theta)}{\partial\theta(i)}$$

are all dominated by an integrable function ($i=1,\ldots,p$). By Theorem 19, it follows that

$$d(\hat{\theta}_m,S(g(EX^2(\cdot),EX(\cdot)),\Theta))=d(\hat{\theta}_m,S(v,\Theta))\to 0$$

as $m\to\infty$. ■

Corollary 20 shows that $\hat{\theta}_m$ converges to the set of first-order critical points of v as $m\to\infty$. This does not guarantee that the sequence $\{\hat{\theta}_m\}$ converges almost surely, as was the case for stochastic approximation. In general we cannot guarantee this because when there are multiple critical points, the particular critical point chosen depends, among other things, on the optimization algorithm that is used. Of course, a simple sufficient condition that ensures this is the existence of a unique first-order critical point. This condition is clearly difficult to verify in practice.

Corollary 21 *In addition to the conditions of Corollary 20 suppose that $S(v,\Theta)$ consists of a single element θ^*. Then $\hat{\theta}_m\to\theta^*$ as $m\to\infty$ a.s.*

7 Numerical Results

In this section, we return to the discrete time finite state space Markov chain 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 $M_T(u(\theta)) = -u(x;\theta) - \sum_{j=0}^{T-1}(P-I)u(Z_j;\theta)$ under some fixed initial state $Z_0 = x$. Then $X(\theta) = X - M_T(u(\theta))$ is an estimator of $\mu(x)$. Let $V = (0, V(1), ..., V(d))^\top$, where $V(j) = \sum_{k=0}^{T-1} I(Z_k = j)$ is the number of visits to state j before absorption. Then

$$\begin{aligned} X(\theta) &= \sum_{j=0}^{T-1} f(Z_j) + u(x;\theta) + \sum_{j=0}^{T-1}[(P-I)u(\theta)](Z_j) \\ &= u(x;\theta) + \sum_{j=0}^{T-1}[(P-I)(u(\theta)-\mu)](Z_j) \\ &= u(x;\theta) + \sum_{k=0}^{d} V(k)[(P-I)(u(\theta)-\mu)](k) \\ &= u(x;\theta) + V^\top(P-I)(u(\theta)-\mu). \end{aligned}$$

To verify that **A1-A6** are satisfied we proceed as follows. First suppose that Θ is convex and compact, that there exists a bounded open set $\mathcal{U}$ such that $\Theta \subset \mathcal{U}$, and that $u(y;\cdot) : \mathcal{U} \to \mathbb{R}$ is $\mathcal{C}^1$ and Lipschitz for all $y \in S$ (these assumptions are all satisfied in our particular example below). Since S is finite and $\mathcal{U}$ is bounded, there exists a $K > 0$ such that for all $\theta_1, \theta_2 \in \mathcal{U}$ and $y \in S$,

$$|u(y;\theta_1) - u(y;\theta_2)| \le K\|\theta_1 - \theta_2\|,$$

and $\{u(y;\theta), \frac{\partial}{\partial\theta(i)}u(y;\theta) : \theta \in \mathcal{U}, y \in S, i = 1, ..., p\}$ are uniformly bounded, i.e.

$$C = \sup_{\theta\in\mathcal{U}, y\in S, i=1,...,p} \left\{ |u(y;\theta)|, \left| \frac{\partial u(y;\theta)}{\partial\theta(i)} \right| \right\} < \infty.$$

Moreover, for any $\theta_1, \theta_2 \in \mathcal{U}$,

$$\begin{aligned} |M_T(u(\theta_1)) - M_T(u(\theta_2))| &\le |u(x;\theta_1) - u(x;\theta_2)| + \sum_{j=0}^{T-1} |[(P-I)(u(\theta_1) - u(\theta_2))](Z_j)| \\ &\le K\|\theta_1 - \theta_2\| + T\|P - I\| \, \|u(\theta_1) - u(\theta_2)\| \\ &\le K\|\theta_1 - \theta_2\| + T\|P - I\| \cdot dK\|\theta_1 - \theta_2\|. \end{aligned}$$

For any $\theta \in \mathcal{U}$,

$$\begin{aligned} |X(\theta)| &\le |u(x;\theta)| + |V^\top(P-I)(u(\theta)-\mu)| \\ &\le |u(x;\theta)| + \|V^\top(P-I)\| \, \|(u(\theta)-\mu)\| \\ &\le C + dT\|(P-I)\|(dC + \|\mu\|), \end{aligned}$$

and similarly,

$$\begin{aligned} |\frac{\partial}{\partial\theta(i)}X(\theta)| &\le |\frac{\partial}{\partial\theta(i)}u(x;\theta)| + |V^\top(P-I)(\frac{\partial}{\partial\theta(i)}u(\theta) - \mu)| \\ &\le C + dT\|(P-I)\|(dC + \|\mu\|). \end{aligned}$$

Since all of these bounds depend only on the random variable T, which has a finite moment generating function in a neighborhood of 0, we can easily verify that assumptions **A1-A6** are satisfied.

For the simulation experiment, we use the "random walk" transition matrix P given by

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ q(1) & 0 & p(1) & 0 & \dots & 0 & 0 & 0 \\ 0 & q(2) & 0 & p(2) & \dots & 0 & 0 & 0 \\ \vdots & \ddots & & \ddots & & & & \vdots \\ 0 & 0 & 0 & 0 & \dots & q(d-1) & 0 & p(d-1) \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & 0 \end{pmatrix},$$

where $q(i) > 0$ for all $i = 1, ..., d-1$. We take

$$u(y; \theta) = \theta_1 y^{\theta_2},$$

where $\theta = (\theta_1, \theta_2) \in \Theta$, $\Theta = \{x \in \mathbb{R}^2 : a(j) \leq x(j) \leq b(j), j = 1, 2\}$ and $a(j) \geq 0, j = 1, 2$. Then $u(y; \cdot)$ is $\mathcal{C}^1$ for all $y \in S$ and the moment generating function of T is defined in a neighborhood of 0. We took $d = 30$ and $f(x) = 1$, so that the random variable $X = T$ is the time till absorption in state 0.

We use the terms naïve, SA and SAA to represent the estimators obtained through naïve Monte Carlo estimation, the stochastic approximation method and the sample average approximation method, respectively. In the stochastic approximation algorithm, we took $m = 100$ 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 advocated in Spall [2003]. We used the average of the sample variances of m terms in each step as an estimator of $v(\theta^*)$. For the SAA estimator, we first replicated $m = 100$ samples. We obtained $\hat{\theta}_m$ by applying a quasi-Newton method with a linesearch (supplied as part of the MATLAB$^{\text{TM}}$ package) using IPA gradients to solve the sample average approximation problem $\mathcal{P}_m$. 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 17. 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 $p(x) = .25$ and $\theta_0 = (1, 1)$. In Table 1, we show the squared standard errors of the three estimators. We see that the SAA estimators outperform the SA estimators, and the SA estimators outperform the naïve 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. A contour plot of the variance surface as a function of θ for initial state $x = 15$ appears in Figure 1. 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 $(2, 1)$.

Remark 4 If the simulation run length n is long enough, then from Theorems 9 and 18 we would expect the SA and SAA estimators to be fairly similar in performance.

Example 3 In this example, $p(x) = .0001 + .4998/x$ and $\theta_0 = (2, 1)$. The results are given in Table 2 and are similar to those of Example 2. The SAA estimator outperforms the other estimators, but not by as large a margin.

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

Table 1: Estimated squared standard errors in Example 2

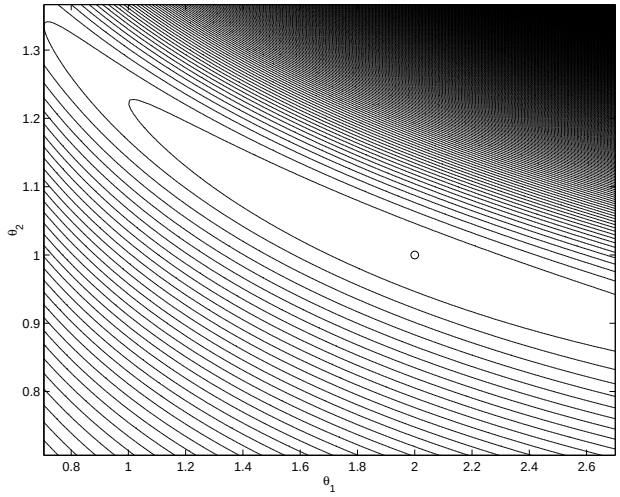

Figure 1: Contour Plot of $v(\cdot)$ for Example 2 with initial state $x = 15$ and runlength 1000

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

Table 2: Estimated squared standard errors in Example 3

8 Final Remarks

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 naïve approach.

We are currently exploring the asymptotic theory of the 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 2. This research was supported by National Science Foundation grants DMI 0230528, DMI 0224884 and DMI 0400287. This paper is a considerable outgrowth of Kim and Henderson [2004].

Appendix: Additional Proofs

Proof of Proposition 10

For part (i), write

$$\begin{aligned} S_n^2 &= \frac{1}{nm-1}\sum_{k=1}^{n}\sum_{i=1}^{m} X_{k,i}^2(\theta_{k-1}) - \frac{nm}{nm-1}\mu_n^2 \\ &= \frac{1}{nm-1}\sum_{k=1}^{n}\sum_{i=1}^{m} X_{k,i}^2(\theta^*) - \frac{nm}{nm-1}\mu_n^2 + \frac{1}{nm-1}\sum_{k=1}^{n}\sum_{i=1}^{m}(X_{k,i}^2(\theta_{k-1}) - X_{k,i}^2(\theta^*)) \end{aligned} \tag{26}$$

By the SLLN and Proposition 7,

$$\frac{1}{nm-1}\sum_{k=1}^{n}\sum_{i=1}^{m} X_{k,i}^2(\theta^*) - \frac{nm}{nm-1}\mu_n^2 \to E(X_1^2(\theta^*)) - \mu^2 = v(\theta^*)$$

as $n \to \infty$ a.s. Therefore it suffices to show that the last term in (26) converges to 0 a.s. as $n \to \infty$.

Since $\theta_k \to \theta^*$ as $k \to \infty$ a.s., for any given $\epsilon > 0$, there exists a random N such that for all $k \geq N$,

$\|\theta^* - \theta_k\| < \epsilon$ a.s. Then

$$\begin{aligned}
&\frac{1}{nm-1}\sum_{k=1}^{n}\sum_{i=1}^{m}(X_{k,i}^2(\theta_{k-1}) - X_{k,i}^2(\theta^*)) \\
&\leq \frac{1}{nm-1}\sum_{k=1}^{n}\sum_{i=1}^{m}|X_{k,i}(\theta_{k-1}) - X_{k,i}(\theta^*)|\ |X_{k,i}(\theta_{k-1}) + X_{k,i}(\theta^*)| \\
&\leq \frac{2}{nm-1}\sum_{k=1}^{n}\sum_{i=1}^{m}K(Y_{k,i})\sup_{\theta\in\mathcal{U}}|X_{k,i}(\theta)|\|\theta_{k-1} - \theta^*\| \\
&\leq \frac{2}{nm-1}\sum_{k=1}^{N}\sum_{i=1}^{m}K(Y_{k,i})\sup_{\theta\in\mathcal{U}}|X_{k,i}(\theta)|\|\theta_{k-1} - \theta^*\| \qquad (27) \\
&\quad + \frac{2}{nm-1}\sum_{k=N+1}^{n}\sum_{i=1}^{m}K(Y_{k,i})\sup_{\theta\in\mathcal{U}}|X_{k,i}(\theta)|\epsilon \qquad (28)
\end{aligned}$$

Now, (27) converges to 0 a.s. as $n \to \infty$ since N is finite. **A4** implies that $K(Y_1)\sup_{\theta\in\mathcal{U}}|X_1(\theta)|$ is integrable and hence the SLLN ensures that

$$\frac{2}{nm-1}\sum_{k=N+1}^{n}\sum_{i=1}^{m}K(Y_{k,i})\sup_{\theta\in\mathcal{U}}|X_{k,i}(\theta)|\epsilon \to 2\epsilon E\left(K(Y_1)\sup_{\theta\in\mathcal{U}}|X_1(\theta)|\right)$$

as $n \to \infty$ a.s. Since ϵ is arbitrary, (28) converges to 0 a.s. as $n \to \infty$.

Essentially the same argument can be applied to $\hat{S}_n$ and $\tilde{S}_n$. We omit the details.

Part (ii) is an immediate consequence of Part (i) and the converging together lemma (e.g., Chung [1974, p. 93]). ■

Proof of Proposition 11

First consider S_n^2. Notice that the $X_{k,i}(\theta^*)$s are i.i.d. Therefore

$$\begin{aligned}
&\sqrt{nm}\left(S_n^2(\theta^*) - v(\theta^*)\right) \\
&= \sqrt{nm}\left(\frac{1}{nm-1}\sum_{k=1}^{n}\sum_{i=1}^{m}X_{k,i}^2(\theta^*) - \frac{nm}{nm-1}\mu_n^2 - v(\theta^*)\right) \\
&= \sqrt{nm}\left(\frac{1}{nm}\sum_{k=1}^{n}\sum_{i=1}^{m}X_{k,i}^2(\theta^*) - \mu_n^2 - v(\theta^*) + o_p((nm)^{-1/2})\right).
\end{aligned}$$

Let $g(x,y) = x - y^2$. Then

$$\frac{1}{nm}\sum_{k=1}^{n}\sum_{i=1}^{m}X_{k,i}^2(\theta^*) - \mu_n^2 - v(\theta^*) = g(\frac{1}{nm}\sum_{k=1}^{n}\sum_{i=1}^{m}X_{k,i}^2(\theta^*), \mu_n) - g(E(X_1^2(\theta^*)), \mu).$$

By the delta method,

$$\sqrt{nm}\left(g(\frac{1}{nm}\sum_{k=1}^{n}\sum_{i=1}^{m}X_{k,i}^2(\theta^*), \mu_n) - g(E(X_1^2(\theta^*)), \mu)\right) \Rightarrow \sigma N(0,1),$$

where

$$\begin{aligned}\sigma^2 &= \nabla g\left(E[X_1(\theta^*)]^2, \mu\right)^T \operatorname{cov}\left(X_1^2(\theta^*), X_1(\theta^*)\right)\nabla g\left(E(X_1^2(\theta^*)), \mu\right)\\ &= E(X_1^4(\theta^*)) - 4\mu E(X_1^3(\theta^*)) + 8\mu^2 E(X_1^2(\theta^*)) - [E(X_1^2(\theta^*))]^2 - 4\mu^4\\ &= E(X_1(\theta^*) - \mu)^4 - v^2(\theta^*).\end{aligned}$$

The central limit theorem for $\hat{S}_n^2$ follows from the ordinary central limit theorem. We get

$$\sqrt{nm}(\hat{S}_n^2(\theta^*) - v(\theta^*)) \Rightarrow \hat{\sigma}N(0,1),$$

where

$$\begin{aligned}\hat{\sigma}^2 &= m\text{var}\left(\frac{1}{m-1}\sum_{i=1}^{m}(X_{1,i}(\theta^*) - A_1(\theta^*))^2\right)\\ &= m\frac{1}{m}\left(E(X_1(\theta^*) - \mu)^4 - \frac{m-3}{m-1}E(X_1(\theta^*) - \mu)^2\right)\\ &= E(X_1(\theta^*) - \mu)^4 - \frac{m-3}{m-1}v^2(\theta^*).\end{aligned}$$

(The second equality above requires some algebra.)

The proof of the central limit theorem for $\tilde{S}_n^2$ follows essentially the same argument that we used for S_n^2 and is omitted. ■

References

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

M. S. Bazaraa, H. D. Sherali, and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms*. Wiley, New York, 2nd edition, 1993.

K. L. Chung. *A Course in Probability Theory*, volume 21 of *Probability and Mathematical Statistics*. Academic Press, San Diego, 2nd edition, 1974.

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

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

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

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

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

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

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

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

S. Kim and S. G. Henderson. Adaptive control variates. In R. Ingalls, M. Rossetti, J. Smith, and B. Peters, editors, *Proceedings of the 2004 Winter Simulation Conference*, Piscataway NJ, 2004. IEEE.

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

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

P. L'Ecuyer. On the interchange of derivative and expectation for likelihood ratio derivative estimators. *Management Science*, 41:738–748, 1995.

R. S. Liptser and A. N. Shiryayev. *Theory of Martingales*. Kluwer Academic, Boston, 1989.

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

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

J. A. Rice. *Mathematical Statistics and Data Analysis*. Wadsworth & Brooks/Cole, Pacific Grove, California, 1988.

A. Shapiro. Monte Carlo sampling methods. In A. Ruszczynski and A. Shapiro, editors, *Stochastic Programming*, Handbooks in Operations Research and Management Science. Elsevier, 2004.

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

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