The Annals of Applied Probability
1999, Vol. 9, No. 2, 391–412

ADAPTIVE IMPORTANCE SAMPLING ON DISCRETE MARKOV CHAINS[1]

BY CRAIG KOLLMAN, KEITH BAGGERLY, DENNIS COX AND RICK PICARD

National Marrow Donor Program, Rice University, Rice University and Los Alamos National Laboratory

In modeling particle transport through a medium, the path of a particle behaves as a transient Markov chain. We are interested in characteristics of the particle's movement conditional on its starting state, which take the form of a "score" accumulated with each transition. Importance sampling is an essential variance reduction technique in this setting, and we provide an adaptive (iteratively updated) importance sampling algorithm that converges exponentially to the solution. Examples illustrating this phenomenon are provided.

1. Introduction and motivation. The motivation for this work involves modeling the behavior of particles (primarily neutrons and photons) as they move within a medium. Applications include such wide ranging areas as well logging for oil exploration, heat generation calculations for nuclear reactors, radiation dosage calculations for exposures to medical X-rays, development of radiation detection devices, design of radiation shielding, criticality safety calculations for storage of nuclear materials and assessment of nuclear weapons performance. Because it is vastly more expensive to perform the related physical experiments than to pursue computer simulation for these applications, it is common to use simulation software such as the Monte Carlo particle transport code MCNP [Briesmeister (1993)].

We give, as background, a brief and necessarily superficial description of particle transport physics [further information on the subject and discussion on the computer simulation thereof can be found in the texts of Carter and Cashwell (1975), Kalos and Whitlock (1986) and Lux and Koblinger (1991)]. Once a particle is emitted from a source, its subsequent behavior is inherently stochastic. Particles move in random directions, collide with atoms in their paths and interact in various ways upon collision. The movement of a particle through a medium can be simulated through a series of (1) path lengths, that is, the distance until the next collision, and (2) types of interactions upon collision with atoms in the material; for example, the particle may be absorbed or may continue on with possibly altered energy and direction. Such particle motions are often well modeled by a Markov chain with states which include location of a collision (or source for the initial state), and either absorption or

Received November 1996; revised May 1998.
[1]Research supported by Los Alamos National Laboratory.
AMS 1991 *subject classification*. 65C05.
Key words and phrases. Adaptive procedures, exponential convergence, Monte Carlo, particle transport, zero-variance solution.

escape, or direction and energy if the particle continues. The Markov chain is transient in that either the particle is eventually absorbed during a collision or leaves the region of interest and is presumed never to return (backscattering is ignored).

From a particle's simulated history, a "score" is obtained. In practice, the score denotes some physical quantity of interest and the objective of the simulation is to estimate the expected score as a function of the initial state. Simple examples related to the above applications include estimating: (1) the proportion of particles emitted from a particular source that are prevented ("shielded") from passing through a specified volume, (2) the energy released within a specified region (i.e., the summed differences between the before-and-after energies of all collisions within the region), and (3) the proportion of particles entering a specified location which are in a certain energy range.

Obtaining results from so-called "analog" simulations that use nature's transition probabilities can be extremely time-consuming. For many problems involving complex material geometries, simulating a single particle history is nontrivial, and only a small fraction of those histories may contribute nonzero scores. Such situations are ideally suited for importance sampling (sometimes called "biasing the random walk"). Several other variance reduction techniques have been proposed in the literature and successfully implemented in current transport code; see Hammersley and Handscomb (1964), Lux and Koblinger (1991) and Briesmeister (1993) for details. These methods, while vastly more efficient than analog Monte Carlo, still only achieve a "constant" speedup in that the variance of such procedures still decreases as $O(n^{-1})$, where n is the number of histories simulated.

This rate is built into the procedures through the use of independent samples; the outcome of one realization does not affect another. To improve on this rate, then, we must use intelligently chosen dependent samples. The "intelligent choice" investigated here corresponds to allowing the process to learn and adapt at various stages, a notion which has been termed sequential or adaptive Monte Carlo. Early work in the field (on a two-stage procedure) includes that of Marshall (1956), which was soon followed by the more general and extensive efforts of Halton (1962). More recently, work by Booth (1985, 1986, 1988, 1989) on guiding various Monte Carlo methods towards zero-variance solutions highlighted adaptive methods once again. Booth found cases where, using adaptive Monte Carlo methods, empirical convergence to the solution was not merely $O(n^{-1})$, but rather $O(e^{-\theta n})$ for some positive constant θ. We refer to this as "exponential convergence." Kollman (1993) later proved that such convergence was possible, albeit in a setting that requred histories starting from every state in the state space.

The purpose of this work is to prove that under certain conditions [which extend those in Kollman (1993)] adaptive importance sampling for discrete Markov chains with scoring converges exponentially. Examples presented here show that this exponential convergence can occur with a reasonably small number of simulation runs. These assumptions include (1) the state space is finite, (2) the vector $\boldsymbol{\mu}$ of expected scores conforms to a linear model $\mathbf{X}\boldsymbol{\beta}$ and

(3) there are sufficiently many replications of the initial states used in the simulation.

In Section 2, we set out the basic mathematical model. We also discuss importance sampling for Markov chains, establish the existence of a zero-variance chain and introduce the adaptive importance sampling algorithm. In Section 3, we prove that this algorithm achieves exponential convergence. In Section 4, we examine some examples showing the empirical performance of the algorithm in simulations and illustrating various features of the theory developed here. Finally, in Section 5 we discuss several practical issues.

2. Description of the method. Consider a Markov chain $\langle X_n \rangle_{n=0}^{\infty}$ with state space $\{1, \ldots, d, \Delta\}$ where Δ is the cemetery or death state. Denote the transition probability matrix for the nonabsorbing states $\{1, \ldots, d\}$ by $\mathbf{P}$, with the (i, j) entry denoted p_{ij}. We assume that eventual death is certain, so

$$\lim_{n \to \infty} \mathbf{P}^n = \mathbf{0}. \tag{1}$$

Of course $p_{\Delta\Delta} = 1$.

When the particle moves from state i to state j, a "score" $s_{ij} \geq 0$ is incurred. The concept of scoring arises naturally in particle physics, and there are also applications to queueing theory [Glasserman (1993a, b)]. We assume $s_{\Delta\Delta} = 0$, so no score is accumulated after death. Denoting the transition at which death occurs by τ, the total score for a particle history is

$$Y_{\mathbf{P}} = \sum_{n=1}^{\tau} s_{X_{n-1}, X_n}.$$

We are interested in estimating

$$\mu_i \equiv E\big[Y_{\mathbf{P}} \big| X_0 = i\big],$$

the expected total score for a particle starting in state i. To indicate a practical application, if s_{ij} is the energy loss of a particle in going from state i to j resulting from a collision, then μ_i is the average energy released per particle for particles starting in state i. Combined with a distribution of sources and a particle density, this can be used to compute the total energy released in a region. For another example, let $s_{ij} = 0$ for all $j \neq \Delta$, and $s_{i\Delta} = 1$ for i in a certain region. Then μ_i is the probability of absorption or escape in that region. Scores of this type are frequently used in reactor safety calculations.

We assume that $\boldsymbol{\mu}^T = (\mu_1, \ldots, \mu_d)$ is given by a linear model

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \tag{2}$$

where $\mathbf{X}$ is a known $d \times p$ matrix of column rank p and $\boldsymbol{\beta}$ is an unknown p-vector. This of course includes the case where $\boldsymbol{\mu}$ ranges over all of $\mathbb{R}^d$: take $\boldsymbol{\beta} = \boldsymbol{\mu}$ and $\mathbf{X}$ the $d \times d$ identity matrix. This case, which is simultaneously the least presumptive and the most computationally intensive (as it requires running the algorithm for more starting values; see Section 2.2 below), was addressed by Kollman (1993).

2.1. *Importance sampling on Markov chains.* Instead of simulating particle movement under nature's transition probabilities $\{p_{ij}\}$, we can often benefit by simulating with different (well chosen) transition probabilities $\{q_{ij}\}$. To keep the expected scores unbiased, each score is weighted by the likelihood ratio between $\mathbf{P}$ and $\mathbf{Q}$. For all times up to death $\{n \le \tau\}$, this ratio is given by

$$L_n = \prod_{i=1}^{n} \frac{p_{X_{i-1}, X_i}}{q_{X_{i-1}, X_i}}.$$

For L_n to be well defined, we require that $\mathbf{Q}$ dominates $\mathbf{P}$, denoted $\mathbf{Q} \gg \mathbf{P}$, which means that $p_{ij} > 0$ implies $q_{ij} > 0$, even if $j = \Delta$. Note that $\mathbf{Q} \gg \mathbf{P}$ guarantees that the chain is transient under $\mathbf{Q}$ (i.e., $\mathbf{Q}^n \to \mathbf{0}$) because of the finite state space. When sampling from $\mathbf{Q}$, the total (weighted) score is

$$Y_{\mathbf{Q}} = \sum_{n=1}^{\tau} s_{X_{n-1}, X_n} L_n.$$

It is easily shown that the expectation of $Y_{\mathbf{Q}}$ given $X_0 = i$ is μ_i. The importance sampling scheme introduced here is an example of *filtered Monte Carlo* introduced in Glasserman (1993b). Under conditions given in Section 3 of Glasserman (1993a), the variance of $Y_{\mathbf{Q}}$ is less than that of the classical importance sampling estimator $\sum_{n=1}^{\tau} s_{X_{n-1}, X_n} L_\tau$.

Now, let $v_i = \mathrm{Var}(Y_{\mathbf{Q}} | X_0 = i)$. By the Markov property,

$$E(Y_{\mathbf{Q}} | X_0 = i,\ X_1 = j) = \frac{p_{ij}}{q_{ij}}(s_{ij} + \mu_j)$$

and

$$\mathrm{Var}(Y_{\mathbf{Q}} | X_0 = i,\ X_1 = j) = \left(\frac{p_{ij}}{q_{ij}}\right)^2 v_j.$$

From the variance decomposition

$$\mathrm{Var}(Y_{\mathbf{Q}} | X_0) = \mathrm{Var}[E(Y_{\mathbf{Q}} | X_0, X_1) | X_0] + E[\mathrm{Var}(Y_{\mathbf{Q}} | X_0, X_1) | X_0],$$

we obtain

$$v_i = q_{i\Delta}\left(\frac{p_{i\Delta} s_{i\Delta}}{q_{i\Delta}}\right)^2 + \sum_{j=1}^{d}\left[q_{ij}\left(\frac{p_{ij}}{q_{ij}}\right)^2 (s_{ij} + \mu_j)^2\right] - \mu_i^2 + \sum_{j=1}^{d}\left[q_{ij}\left(\frac{p_{ij}}{q_{ij}}\right)^2 v_j\right].$$

Noting that $E[\mathrm{Var}(Y_{\mathbf{Q}} | X_0 = \Delta)] = 0$,

$$v_i = \left(\frac{p_{i\Delta}^2 s_{i\Delta}^2}{q_{i\Delta}} + \sum_j\left[\frac{p_{ij}^2 (s_{ij} + \mu_j)^2}{q_{ij}}\right] - \mu_i^2\right) + \sum_j \frac{p_{ij}^2}{q_{ij}} v_j,$$

(with $0/0 = 0$ is defined). If we let

$$f_i = \frac{p_{i\Delta}^2 s_{i\Delta}^2}{q_{i\Delta}} + \sum_j\left[\frac{p_{ij}^2 (s_{ij} + \mu_j)^2}{q_{ij}}\right] - \mu_i^2, \tag{3}$$

and define $\mathbf{R}$ to be the matrix whose (i, j)th element is given by

$$r_{ij} = \begin{cases} p_{ij}^2/q_{ij}, & \text{if } q_{ij} > 0, \\ 0, & \text{if } q_{ij} = 0, \end{cases}$$

then in matrix form the variance equation becomes

$$\mathbf{v} = \mathbf{f} + \mathbf{R}\mathbf{v}. \tag{4}$$

Note that when

$$q_{ij}(\boldsymbol{\mu}) = \frac{p_{ij}(s_{ij} + \mu_j)}{p_{i\Delta}s_{i\Delta} + \sum_{k=1}^{d} p_{ik}(s_{ik} + \mu_k)}$$

and

$$q_{i\Delta}(\boldsymbol{\mu}) = \frac{p_{i\Delta}s_{i\Delta}}{p_{i\Delta}s_{i\Delta} + \sum_{k=1}^{d} p_{ik}(s_{ik} + \mu_k)},$$

we get

$$f_i = \left(\frac{p_{i\Delta}^2 s_{i\Delta}^2}{p_{i\Delta}s_{i\Delta}} + \sum_{j=1}^{d} \frac{p_{ij}^2(s_{ij} + \mu_j)^2}{p_{ij}(s_{ij} + \mu_j)}\right)\left(p_{i\Delta}s_{i\Delta} + \sum_{k=1}^{d} p_{ik}(s_{ik} + \mu_k)\right) - \mu_i^2,$$

which reduces to

$$f_i = \left(p_{i\Delta}s_{i\Delta} + \sum_{j=1}^{d} p_{ij}(s_{ij} + \mu_j)\right)\left(p_{i\Delta}s_{i\Delta} + \sum_{k=1}^{d} p_{ik}(s_{ik} + \mu_k)\right) - \mu_i^2 = 0.$$

This choice of $\mathbf{Q}$ gives $\mathbf{v} = \mathbf{0}$ by (4) and, as can also be shown by an induction on the value of τ, a zero-variance importance scheme. That is, $Y_{\mathbf{Q}} = \mu_i$ if $X_0 = i$. This choice of $\mathbf{Q}$ depends on the unknown solution $\boldsymbol{\mu}$. However, it suggests an adaptive procedure, described next.

2.2. *The adaptive algorithm.* For $\tilde{\boldsymbol{\mu}} \in \mathbb{R}^d$ with $\tilde{\mu}_j > 0$ for $1 \le j \le d$, define

$$q_{ij}(\tilde{\boldsymbol{\mu}}) = \frac{p_{ij}(s_{ij} + \tilde{\mu}_j)}{p_{i\Delta}s_{i\Delta} + \sum_{k=1}^{d} p_{ik}(s_{ik} + \tilde{\mu}_k)},$$

where of course $\tilde{\mu}_\Delta = 0$ and $q_{i\Delta}(\tilde{\boldsymbol{\mu}}) = 1 - \sum_{j=1}^{d} q_{ij}(\tilde{\boldsymbol{\mu}})$. This gives a transition probability matrix $\mathbf{Q}(\tilde{\boldsymbol{\mu}})$ with $\mathbf{Q}(\tilde{\boldsymbol{\mu}}) \gg \mathbf{P}$. Similarly, define $\mathbf{f}(\tilde{\boldsymbol{\mu}})$ to be the $\mathbf{f}$ of (3) for $\mathbf{Q}(\tilde{\boldsymbol{\mu}})$; that is, the ith component of $\mathbf{f}(\tilde{\boldsymbol{\mu}})$ is given by

$$f_i(\tilde{\boldsymbol{\mu}}) = \left(p_{i\Delta}s_{i\Delta} + \sum_{j=1}^{d} \frac{p_{ij}^2(s_{ij} + \mu_j)^2}{p_{ij}(s_{ij} + \tilde{\mu}_j)}\right)\left(p_{i\Delta}s_{i\Delta} + \sum_{k=1}^{d} p_{ik}(s_{ik} + \tilde{\mu}_k)\right) - \mu_i^2.$$

In a similar way, define $\mathbf{v}(\tilde{\boldsymbol{\mu}})$ and $\mathbf{R}(\tilde{\boldsymbol{\mu}})$. The idea behind the algorithm is to simulate under $\mathbf{Q}(\hat{\boldsymbol{\mu}})$ where the vector $\hat{\boldsymbol{\mu}}$ is an estimate of $\boldsymbol{\mu}$ from previous simulations.

We must be careful to ensure that $\mathbf{Q}(\hat{\boldsymbol{\mu}}) \gg \mathbf{P}$. Note that if there exists a $\delta > 0$ such that $s_{i\Delta} \geq \delta$ for all i then $\mu_i \geq \delta$ for all i; that is, the particle is assured of scoring at least δ since it eventually dies. Because $\mu_i \geq \delta$, we can take the maximum of δ and the estimate without increasing the error. This ensures that $\hat{\mu}_i \geq \delta$ and hence $\mathbf{Q}(\hat{\boldsymbol{\mu}}) \gg \mathbf{P}$. If such a value for δ is unknown, we can easily alter the problem by adding a known $\delta > 0$ to each $s_{i\Delta}$. Since every particle dies exactly once, this just adds δ to each μ_i. We can subtract δ from each $\hat{\mu}_i$ at the end of the simulation to return to the original problem. Thus, without loss of generality, there exists a known $\delta > 0$ such that $\mu_i \geq \delta$ for all $i \neq \Delta$.

We describe how outcomes of simulation runs started in various states are used to estimate $\boldsymbol{\mu}$. Let D denote a fixed *base design* with n_D initial states, that is, $D = (i_1, \ldots, i_{n_D})$ and each $i_j \neq \Delta$. Given $\mathbf{Q}$, we obtain data $\mathbf{y}^T = (Y_{1\mathbf{Q}}, \ldots, Y_{n_D\mathbf{Q}})$, where $Y_{j\mathbf{Q}}$ is the total weighted score from simulation of the chain having initial state $X_0 = i_j$ and using transition kernel $\mathbf{Q}$. Of course, the components of $\mathbf{y}$ are mutually independent conditional on the current estimate of $\boldsymbol{\mu}$. We use the ordinary least squares estimator of $\boldsymbol{\beta}$ given by

$$\hat{\boldsymbol{\beta}}_D = (\mathbf{X}_D^T \mathbf{X}_D)^{-1} \mathbf{X}_D^T \mathbf{y},$$

where $\mathbf{X}_D$ is the $n_D \times p$ matrix of column rank p whose jth row is the i_jth row of $\mathbf{X}$. The corresponding estimator of $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}_D$. The case where $\mathbf{X}$ is the $d \times d$ identity matrix [as considered in Kollman (1993)] requires, obviously, the largest (minimum) base design D in order to estimate $\boldsymbol{\mu}$.

We now describe the design that is used in each iteration of the algorithm. Assume the base design D, which is fixed throughout the algorithm, is such that $\mathbf{X}_D$ has full rank p, and let D_r be the design obtained by replicating D r times. Exponential convergence is assured if the number of replications exceeds a (nonconstructive) lower bound given in the proof of the main theorem.

The algorithm is as follows.

I. Obtain an initial estimate $\hat{\boldsymbol{\mu}}^{(0)}$, for example, one based on previous experience with similar problems or based on a simulation using the analog transition probabilities p_{ij}.

II. Given that m iterations of the algorithm have produced the estimate $\hat{\boldsymbol{\mu}}^{(m)}$, iterate the following steps to convergence:
   1. Using the design D_r, run independent replications of the chain with $\mathbf{Q}(\hat{\boldsymbol{\mu}}^{(m)})$ as the transition matrix.
   2. For a given replication in 1, let τ be the transition at which death occurs and $Y = \sum_{n=1}^{\tau} s_{X_{n-1}, X_n} L_n$, where the $\{X_n\}$ and $\{L_n\}$ are obtained from the given simulation run.
   3. Using the linear model $\boldsymbol{\mu}_D = \mathbf{X}_D\boldsymbol{\beta}$, use the (i, Y) pairs (i is the starting value) to estimate $\hat{\boldsymbol{\beta}}^{(m+1)} = (\mathbf{X}_D^T\mathbf{X}_D)^{-1}\mathbf{X}_D^T\bar{\mathbf{y}}^{(m+1)}$ by ordinary least squares.
   4. Define $\hat{\boldsymbol{\mu}}^{(m+1)}$ by $\hat{\mu}_i^{(m+1)} = \max([\mathbf{X}\hat{\boldsymbol{\beta}}^{(m+1)}]_i, \delta)$, $i = 1, \ldots, d$.

Note that this last step of the algorithm ensures that $\hat{\mu}_i^{(m+1)} \geq \delta$, $1 \leq i \leq d$. Moreover, the computational burden can be reduced for very large problems (where only a fraction of the states are visited in each iteration) by computing $q_{ij}(\hat{\boldsymbol{\mu}}^{(m)})$ only when needed.

We mention two modifications of the algorithm. The first involves utilizing what we call "path information." When a state i is visited, we may think of it as an initial state and begin accumulating a total weighted score. The computations needed to do this are basically the same as those associated with the actual (first) initial state of the history of that path, plus a little bookkeeping. However, there is now dependence between the total weighted scores from any paths that share a common history. Path information was considered for the case where $\mathbf{X}$ is the identity matrix in Kollman (1993), where only the first visit to i is allowed in the estimation (so that for a given initial state, all histories are independent), and it was proved that it improves convergence in theory for the case where $\mathbf{X}$ is the identity.

The second modification is the use of previous $\boldsymbol{\mu}$ estimates, which we refer to as "incorporating previous information." It will also be necessary for technical reasons to assume a known upper bound H on μ_i. We modify step 4 of the algorithm to

$$\hat{\mu}_i^{(m+1)} = w \min\bigl\{\max\{[\mathbf{X}\hat{\boldsymbol{\beta}}^{(m+1)}]_i, \delta\}, H\bigr\} + (1-w)\hat{\mu}_i^{(m)}, \qquad i = 1, \ldots, d, \tag{5}$$

where the weight $w \in (0, 1]$ can be chosen. We show in Proposition 1 below that one obtains exponential convergence with this algorithm as well. By choosing the weight w properly, convergence can be improved over the choice $w = 1$ corresponding to the original algorithm.

3. Proof of exponential convergence. We have made three assumptions about the problem.

ASSUMPTION 1. $\lim_{n\to\infty} \mathbf{P}^n = \mathbf{0}$.

ASSUMPTION 2. There exists a known $\delta > 0$ such that $\mu_i \geq \delta$ for every i.

ASSUMPTION 3. The mean total score $\boldsymbol{\mu}$ is given by the linear model $\mathbf{X}\boldsymbol{\beta}$ in (2).

Denote the usual norm on Euclidean d-space by $\|\mathbf{z}\| = (\mathbf{z}^T\mathbf{z})^{1/2}$. Let D be a given base design as in the previous section and note that for $M = \operatorname{tr}(\mathbf{X}_D^T\mathbf{X}_D)^{-1}(\mathbf{X}^T\mathbf{X})$,

$$E\bigl[\|\mathbf{X}\hat{\boldsymbol{\beta}}_D - \boldsymbol{\mu}\|^2\bigr] \leq M \max_i \operatorname{Var}(Y_{\mathbf{Q}} | X_0 = i) \tag{6}$$

for all importance sampling transition matrices $\mathbf{Q}$, where $\hat{\boldsymbol{\beta}}_D$ is the least squares estimator of $\boldsymbol{\beta}$ based on data with design D. For the r-fold replicated design D_r, $Y_{\mathbf{Q}}$ in the previous display is replaced by the average of the

r independent replicates of $Y_{\mathbf{Q}}$ at the same design point, and

$$E\big[\|\mathbf{X}\hat{\boldsymbol{\beta}}_{D_r} - \boldsymbol{\mu}\|^2\big] \le \frac{M}{r} \max_i \operatorname{Var}(Y_{\mathbf{Q}} | X_0 = i). \tag{7}$$

It should be understood when we write $\hat{\boldsymbol{\beta}}^{(m+1)}$ that this estimate is derived from the $Y_{\mathbf{Q}(\hat{\boldsymbol{\mu}}^{(m)})}$ values utilizing the design D_r; that is, it is $\hat{\boldsymbol{\beta}}_{D_r}$ for the current iteration.

THEOREM. *Under the assumptions stated above, given a design D there exist deterministic constants $\theta > 0$ and R such that if the adaptive algorithm is run with design D_r where $r \ge R$, then with probability* 1, $e^{\theta m}\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\| \to 0$ *as* $m \to \infty$.

Before presenting the proof we need to derive some preliminary results. We use $I[\cdots]$ to denote the indicator of an event specified by $[\cdots]$.

LEMMA 1. *There exists a matrix* $\mathbf{A}$ *and an* $\varepsilon > 0$ *such that*

$$\mathbf{1}^T\mathbf{v}(\tilde{\boldsymbol{\mu}}) \le (\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu})^T\mathbf{A}(\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}),$$

whenever $\|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\| < \varepsilon$, *where* $\mathbf{1}$ *denotes a vector of ones*.

PROOF. We establish that $\mathbf{v}(\tilde{\boldsymbol{\mu}})$ is twice continuously differentiable in a neighborhood of $\boldsymbol{\mu}$. Since $\mathbf{v}(\boldsymbol{\mu}) = 0$ is a global minimum of $\mathbf{v}$, the result then follows by application of Taylor's formula.

When all the components of $\tilde{\boldsymbol{\mu}}$ are positive, $\mathbf{Q}(\tilde{\boldsymbol{\mu}}) \gg \mathbf{P}$ and both $r_{ij}(\tilde{\boldsymbol{\mu}})$ and $f_i(\tilde{\boldsymbol{\mu}})$ [see discussion following (4)] are infinitely differentiable in $\tilde{\boldsymbol{\mu}}$. Thus, in view of (4) we need only show that $(\mathbf{I} - \mathbf{R}(\tilde{\boldsymbol{\mu}}))^{-1}$ exists and is infinitely differentiable in a neighborhood of $\boldsymbol{\mu}$ to complete the proof.

By Assumption 2 we know that $\boldsymbol{\mu} \ge \delta\mathbf{1}$, which implies that $\mathbf{Q}(\boldsymbol{\mu}) \gg \mathbf{P}$. If $p_{ij} > 0$, then $q_{ij}(\boldsymbol{\mu}) > 0$ and the elements of $\mathbf{R}(\boldsymbol{\mu})$ satisfy

$$r_{ij}(\boldsymbol{\mu}) = \frac{p_{ij}^2}{q_{ij}(\boldsymbol{\mu})} = \frac{p_{ij}^2[p_{i\Delta}s_{i\Delta} + \sum_{l=1}^d p_{il}(s_{il} + \mu_l)]}{p_{ij}(s_{ij} + \mu_j)} = \frac{p_{ij}\mu_i}{s_{ij} + \mu_j} \le \frac{p_{ij}\mu_i}{\mu_j}.$$

Let $r_{ij}^{(n)}(\boldsymbol{\mu})$ and $p_{ij}^{(n)}$ denote the (i, j)th elements of $\mathbf{R}^n(\boldsymbol{\mu})$ and $\mathbf{P}^n$, respectively. An induction argument shows that $r_{ij}^{(n)}(\boldsymbol{\mu}) \le p_{ij}^{(n)}\mu_i/\mu_j$ for all n. Since the state space is finite, Assumption 1 implies that $\sum_{n=0}^{\infty} \mathbf{P}^n < \infty$ and thus $\sum_{n=0}^{\infty} \mathbf{R}^n(\boldsymbol{\mu}) < \infty$. Consequently,

$$(\mathbf{I} - \mathbf{R}(\boldsymbol{\mu}))^{-1} = \sum_{n=0}^{\infty} \mathbf{R}^n(\boldsymbol{\mu}),$$

and the requisite differentiability of $(\mathbf{I} - \mathbf{R}(\tilde{\boldsymbol{\mu}}))^{-1}$ follows. □

LEMMA 2. *There exist a constant* $c \in (0,1)$, *an* $R_1 > 0$, *an* $\varepsilon > 0$ *and a* $\nu > 0$ *such that if the learning algorithm is run with* $\|\hat{\boldsymbol{\mu}}^{(0)} - \boldsymbol{\mu}\| < \varepsilon$ *and a design consisting of* $r \geq R_1$ *replications of* D, *then*

$$\begin{aligned}\Pr\bigl\{\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\| \leq c^m \|\hat{\boldsymbol{\mu}}^{(0)} - \boldsymbol{\mu}\|,\ \forall\ m \mid \hat{\boldsymbol{\mu}}^{(0)}\bigr\} I[\|\hat{\boldsymbol{\mu}}^{(0)} - \boldsymbol{\mu}\| < \varepsilon] \\ \geq \nu I\bigl[\|\hat{\boldsymbol{\mu}}^{(0)} - \boldsymbol{\mu}\| < \varepsilon\bigr].\end{aligned}$$

PROOF. By Assumption 2 and steps 3 and 4 of the algorithm, we have

$$\begin{aligned}E\bigl[\|\hat{\mu}_i^{(m+1)} - \mu_i\|^2 \big| \hat{\boldsymbol{\mu}}^{(m)}\bigr] &\leq E\bigl[\|\mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{(m+1)} - \mu_i\|^2 \big| \hat{\boldsymbol{\mu}}^{(m)}\bigr] \\ &\leq \frac{M}{r} \max_i v_i\bigl(\hat{\boldsymbol{\mu}}^{(m)}\bigr),\end{aligned}$$

by (7), where $\mathbf{x}_i^T$ is the ith row of $\mathbf{X}$ and M is given in (6). Let ε and $\mathbf{A}$ be as in Lemma 1. Then

$$\begin{aligned}&E\bigl[\|\hat{\boldsymbol{\mu}}^{(m+1)} - \boldsymbol{\mu}\|^2 I\bigl[\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\| < \varepsilon\bigr] \mid \hat{\boldsymbol{\mu}}^{(m)}\bigr] \\ &\quad = \sum_{i=1}^{d} E\bigl[\|\hat{\mu}_i^{(m+1)} - \mu_i\|^2 \big| \hat{\boldsymbol{\mu}}^{(m)}\bigr] I\bigl[\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\| < \varepsilon\bigr] \qquad (8) \\ &\quad \leq b\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\|^2,\end{aligned}$$

where, using the fact that the maximum of the $v_i(\hat{\boldsymbol{\mu}}^{(m)})$ is less than or equal to $\mathbf{1}^T \mathbf{v}(\hat{\boldsymbol{\mu}}^{(m)})$,

$$b \equiv dMr^{-1} \sup_{\|\phi\|=1} \|\mathbf{A}\phi\|.$$

Choose $R_1 > 0$ so that $r \geq R_1$ implies $b < 1$.

Let $T = \inf\{m \geq 0\colon \|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\| \geq \varepsilon\}$ and define $\boldsymbol{\lambda}^{(m)}$ as

$$\boldsymbol{\lambda}^{(m)} \equiv \begin{cases} \hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}, & m \leq T, \\ \mathbf{0}, & m > T. \end{cases}$$

Note that if $\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu} = 0$ for some m, then $\hat{\boldsymbol{\mu}}^{(m^\star)} - \boldsymbol{\mu} = 0$ for all $m^\star > m$ by the zero-variance property. One can then check that

$$\boldsymbol{\lambda}^{(m+1)} = (\hat{\boldsymbol{\mu}}^{(m+1)} - \boldsymbol{\mu}) I[0 < \|\boldsymbol{\lambda}^{(m)}\| < \varepsilon].$$

Thus

$$\begin{aligned}&E\bigl[\|\boldsymbol{\lambda}^{(m+1)}\|^2 \mid \langle \hat{\boldsymbol{\mu}}^{(n)} \rangle_{n=0}^{m}\bigr] \\ &\quad = E\bigl[\|\hat{\boldsymbol{\mu}}^{(m+1)} - \boldsymbol{\mu}\|^2 I[0 < \|\boldsymbol{\lambda}^{(m)}\| < \varepsilon] \mid \langle \hat{\boldsymbol{\mu}}^{(n)} \rangle_{n=0}^{m}\bigr] \\ &\quad = E\bigl[\|\hat{\boldsymbol{\mu}}^{(m+1)} - \boldsymbol{\mu}\|^2 I[\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\| < \varepsilon] \big| \langle \hat{\boldsymbol{\mu}}^{(n)} \rangle_{n=0}^{m}\bigr] I[0 < \|\boldsymbol{\lambda}^{(m)}\| < \varepsilon] \qquad (9) \\ &\quad \leq b\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\|^2 I[0 < \|\boldsymbol{\lambda}^{(m)}\| < \varepsilon] \qquad (10) \\ &\quad \leq b\|\boldsymbol{\lambda}^{(m)}\|^2. \qquad (11)\end{aligned}$$

In the above, (9) follows because $\{0 < \|\boldsymbol{\lambda}^{(m)}\| < \varepsilon\} \subset \{\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\| < \varepsilon\}$, (10) by (8) and (11) by definition of $\boldsymbol{\lambda}^{(m)}$. By induction, this yields

$$E\big[\|\boldsymbol{\lambda}^{(m)}\|^2 \big| \langle \hat{\boldsymbol{\mu}}^{(n)} \rangle_{n=0}^{m-1}\big] \le b^m E\big[\|\boldsymbol{\lambda}^{(0)}\|^2 \big| \hat{\boldsymbol{\mu}}^{(0)}\big] \qquad \forall\ m \ge 1. \tag{12}$$

Now, choose a value c such that $b < c^2 < 1$, and define events

$$F_m = \begin{cases} \{\|\boldsymbol{\lambda}^{(0)}\| < \varepsilon\}, & \text{if } m = 0, \\ \{\|\boldsymbol{\lambda}^{(m)}\| \le c^m \|\boldsymbol{\lambda}^{(0)}\|\}, & \text{if } m > 0. \end{cases}$$

Using Markov's inequality and (12), for $m \ge 1$,

$$\begin{aligned} &P\Big(F_m^c \cap \bigcap_{j=0}^{m-1} F_j \Big| \hat{\boldsymbol{\mu}}^{(0)}\Big) \\ &\quad = P\Big(\big\{\|\boldsymbol{\lambda}^{(m)}\|^2 > c^{2m}\|\boldsymbol{\lambda}^{(0)}\|^2\big\} \cap \bigcap_{j=0}^{m-1} F_j \Big| \hat{\boldsymbol{\mu}}^{(0)}\Big) \\ &\quad = E\Big(P\big[\|\boldsymbol{\lambda}^{(m)}\|^2 > c^{2m}\|\boldsymbol{\lambda}^{(0)}\|^2 \big| \langle \hat{\boldsymbol{\mu}}^{(j)} \rangle_{j=0}^{m-1}\big] I\Big[\bigcap_{j=0}^{m-1} F_j\Big] \Big| \hat{\boldsymbol{\mu}}^{(0)}\Big) \\ &\quad \le E\Big(\frac{E(\|\boldsymbol{\lambda}^{(m)}\|^2 | \langle \hat{\boldsymbol{\mu}}^{(j)} \rangle_{j=0}^{m-1})}{c^{2m}\|\boldsymbol{\lambda}^{(0)}\|^2} I\Big[\bigcap_{j=0}^{m-1} F_j\Big] \Big| \hat{\boldsymbol{\mu}}^{(0)}\Big) \\ &\quad \le E\Big(\frac{b^m\|\boldsymbol{\lambda}^{(0)}\|^2}{c^{2m}\|\boldsymbol{\lambda}^{(0)}\|^2} I\Big[\bigcap_{j=0}^{m-1} F_j\Big] \Big| \hat{\boldsymbol{\mu}}^{(0)}\Big) \\ &\quad = \Big(\frac{b}{c^2}\Big)^m P\Big(\bigcap_{j=0}^{m-1} F_j \Big| \hat{\boldsymbol{\mu}}^{(0)}\Big) \end{aligned}$$

and hence

$$P\Big(\bigcap_{j=0}^{m} F_j \Big| \hat{\boldsymbol{\mu}}^{(0)}\Big) \ge \Big[1 - \Big(\frac{b}{c^2}\Big)^m\Big] P\Big(\bigcap_{j=0}^{m-1} F_j \Big| \hat{\boldsymbol{\mu}}^{(0)}\Big). \tag{13}$$

So

$$P\Big(\bigcap_{j=0}^{\infty} F_j \Big| \hat{\boldsymbol{\mu}}^{(0)}\Big) = \lim_{m\to\infty} P\Big(\bigcap_{j=0}^{m} F_j \Big| \hat{\boldsymbol{\mu}}^{(0)}\Big) \ge \nu I[F_0],$$

where $\nu > 0$ is given by

$$\nu \equiv \prod_{m=1}^{\infty}\Big[1 - \Big(\frac{b}{c^2}\Big)^m\Big].$$

Note that $\nu > 0$ since $\sum_{m=1}^{\infty}(b/c^2)^m < \infty$ [see Theorem 12-52 of Apostol (1957)].

Now

$$\bigcap_{m=0}^{\infty} F_m \subseteq \{T = \infty\} = \{\boldsymbol{\lambda}^{(m)} = \hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\ \forall\ m\}.$$

That is, the $\boldsymbol{\lambda}$ and $\hat{\boldsymbol{\mu}}$ processes never decouple on the event $\bigcap_{m=0}^{\infty} F_m$. So,

$$P\{\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\| \le c^m \|\hat{\boldsymbol{\mu}}^{(0)} - \boldsymbol{\mu}\| \ \forall\ m \mid \hat{\boldsymbol{\mu}}^{(0)}\} I[\|\hat{\boldsymbol{\mu}}^{(0)} - \boldsymbol{\mu}\| < \varepsilon]$$
$$\ge P\left\{\bigcap_{m=0}^{\infty} F_m\right\} \ge \nu I[\|\hat{\boldsymbol{\mu}}^{(0)} - \boldsymbol{\mu}\| < \varepsilon].$$

This completes the proof. □

We must start the initial estimate, $\hat{\boldsymbol{\mu}}^{(0)}$, close enough to $\boldsymbol{\mu}$ for the probability bound of Lemma 2 to hold. However, even if we start with $\|\hat{\boldsymbol{\mu}}^{(0)} - \boldsymbol{\mu}\| \ge \varepsilon$ we can wait to see if $\|\hat{\boldsymbol{\mu}}^{(m^*)} - \boldsymbol{\mu}\| < \varepsilon$ for some m^*. If this happens, the strong Markov property implies

$$\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\| < c^{m-m^*} \|\hat{\boldsymbol{\mu}}^{(m^*)} - \boldsymbol{\mu}\| \qquad \forall\ m \ge m^*$$

holds with probability at least ν. That is, every time the process $\{\hat{\boldsymbol{\mu}}^{(m)}\}_{m=0}^{\infty}$ enters the ε-neighborhood of $\boldsymbol{\mu}$, there is probability at least ν of exponential convergence. If such entry occurred infinitely often, then exponential convergence would be certain. We are now ready to complete the proof.

PROOF OF THEOREM. Let $\bar{\mathbf{y}}^{(m)} = (\bar{Y}_1^{(m)}, \ldots, \bar{Y}_{n_D}^{(m)})$ denote the n_D-dimensional vector of the averages of the r values of $Y_{j,\,\mathbf{Q}(\hat{\boldsymbol{\mu}}^{(m)})}$ at each design point i_j in D at the mth iteration (i.e., $\bar{Y}_j^{(m)}$ is the average of the r replicates of $Y_{j,\,\mathbf{Q}(\hat{\boldsymbol{\mu}}^{(m)})}$). The average total scores $\bar{Y}_j^{(m)}$ are nonnegative and have finite mean μ_{i_j}, so by Markov's inequality,

$$\Pr\{\bar{Y}_j^{(m+1)} > k\mu_{i_j} \mid \hat{\boldsymbol{\mu}}^{(m)}\} \le \frac{1}{k},$$

for $j = i_1, i_2, \ldots, i_{n_D}$. Conditional on $\hat{\boldsymbol{\mu}}^{(m)}$, the rn_D chains of iteration $m+1$ are independent, and hence

$$\Pr\{\bar{Y}_j^{(m+1)} \le k\mu_{i_j} \ \forall\ j \mid \hat{\boldsymbol{\mu}}^{(m)}\} \ge \left(1 - \frac{1}{k}\right)^{n_D}.$$

Choose k such that $(1 - 1/k)^{n_D} > 1/2$.

Denote by $\boldsymbol{\mu}_D^T = (\mu_{i_1}, \ldots, \mu_{i_{n_D}})$ the vector of component $\boldsymbol{\mu}$ values corresponding to the states in D. Let $[\mathbf{0}, k\boldsymbol{\mu}_D]$ be the rectangle set in R^{n_D} where $0 \le y_j \le k\mu_{i_j}$, $1 \le j \le n_D$ and let

$$H = \sup_{\mathbf{y} \in [\mathbf{0},\, k\boldsymbol{\mu}_D]} \max_{1 \le i \le d} |\mathbf{x}_i^T (\mathbf{X}_D^T \mathbf{X}_D)^{-1} \mathbf{X}_D^T \mathbf{y}|. \tag{14}$$

Then

$$\Pr\{\hat{\boldsymbol{\mu}}^{(m+1)} \le H\mathbf{1} \mid \hat{\boldsymbol{\mu}}^{(m)}\} \ge \tfrac{1}{2}. \tag{15}$$

Let $\mathscr{U}$ denote the set of vectors in R^d having all positive components. For a vector $\tilde{\boldsymbol{\mu}}$ in $\mathscr{U}$ let $\mathscr{L}_i(\cdot|\tilde{\boldsymbol{\mu}})$ denote the probability law (distribution) of $Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}})}$ when simulating under $\mathbf{Q}(\tilde{\boldsymbol{\mu}})$ and starting at $X_0 = i$. That is, for sets $A \subseteq R$,

$$\mathscr{L}_i(A|\tilde{\boldsymbol{\mu}}) = \Pr\{Y_{\mathbf{Q}(\hat{\boldsymbol{\mu}}^{(m)})} \in A \mid X_0 = i,\ \hat{\boldsymbol{\mu}}^{(m)} = \tilde{\boldsymbol{\mu}}\}.$$

Note that the right-hand side of the above equation does not depend on m. The transition probabilities $q_{ij}(\tilde{\boldsymbol{\mu}})$ are continuous functions of $\tilde{\boldsymbol{\mu}}$. For $\tilde{\boldsymbol{\mu}} \in \mathscr{U}$, $\mathbf{Q}(\tilde{\boldsymbol{\mu}}) \gg \mathbf{P}$, so the likelihood ratios L_n and total scores $Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}})}$ are continuous functions of $\tilde{\boldsymbol{\mu}}$. It follows that the map $\tilde{\boldsymbol{\mu}} \mapsto \mathscr{L}_i(\cdot|\tilde{\boldsymbol{\mu}})$ is continuous in the topology of weak convergence; that is,

$$\text{if } \tilde{\boldsymbol{\mu}}_n \to \tilde{\boldsymbol{\mu}}, \text{ then } \mathscr{L}_i(\cdot|\tilde{\boldsymbol{\mu}}_n) \Rightarrow \mathscr{L}_i(\cdot|\tilde{\boldsymbol{\mu}}).$$

Let $E^{\tilde{\boldsymbol{\mu}}}(\cdot)$ denote expectation under $\mathscr{L}_i(\cdot|\tilde{\boldsymbol{\mu}})$. Fix $\alpha > 0$. Suppose that we have a sequence of vectors $\langle \tilde{\boldsymbol{\mu}}_n \rangle$ in $\mathscr{U}$ with $\tilde{\boldsymbol{\mu}}_n \to \tilde{\boldsymbol{\mu}}$. Then the distributions for the random variables $Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}}_n)} I[Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}}_n)} \le \alpha]$ under $\mathbf{Q}(\tilde{\boldsymbol{\mu}}_n)$ converge in distribution to the distribution of $Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}})} I[Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}})} \le \alpha]$. By bounded convergence,

$$E^{\tilde{\boldsymbol{\mu}}_n}(Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}}_n)} I[Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}}_n)} \le \alpha]) \to E^{\tilde{\boldsymbol{\mu}}}(Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}}_n)} I[Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}}_n)} \le \alpha]).$$

That is, $E^{\tilde{\boldsymbol{\mu}}}(Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}}_n)} I[Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}}_n)} \le \alpha])$ is a continuous function of $\tilde{\boldsymbol{\mu}}$ for each fixed α. For any $\tilde{\boldsymbol{\mu}} \in \mathscr{U}$, $E^{\tilde{\boldsymbol{\mu}}}(Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}})}) = \mu_i$, which is a constant and hence a continuous function of $\tilde{\boldsymbol{\mu}}$. So, on $\mathscr{U}$,

$$E^{\tilde{\boldsymbol{\mu}}}(Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}})} I[Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}})} > \alpha]) = E^{\tilde{\boldsymbol{\mu}}}(Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}})}) - E^{\tilde{\boldsymbol{\mu}}}(Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}})} I[Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}})} \le \alpha])$$

is continuous in $\tilde{\boldsymbol{\mu}}$.

Because $Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}})}$ has finite mean,

$$\lim_{\alpha \to \infty} E^{\tilde{\boldsymbol{\mu}}}\big(Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}})} I[Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}})} > \alpha]\big) = 0.$$

If we restrict $\tilde{\boldsymbol{\mu}}$ to $\{\tilde{\boldsymbol{\mu}}\colon \delta\mathbf{1} \le \tilde{\boldsymbol{\mu}} \le H\mathbf{1}\}$ then we can think of $E^{\tilde{\boldsymbol{\mu}}}(Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}}_n)} I[Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}}_n)} > \alpha])$ as a family of continuous functions of $\tilde{\boldsymbol{\mu}}$ on a compact set indexed by α. These functions tend monotonically to zero pointwise as $\alpha \to \infty$, so the convergence is uniform by Dini's theorem. That is,

$$\lim_{\alpha \to \infty} \sup_{\delta\mathbf{1} \le \tilde{\boldsymbol{\mu}} \le H\mathbf{1}} E^{\tilde{\boldsymbol{\mu}}}(Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}}_n)} I[Y_{\mathbf{Q}(\tilde{\boldsymbol{\mu}}_n)} > \alpha]) = 0 \tag{16}$$

so that the family of probability measures $\{\mathscr{L}_i(\cdot|\tilde{\boldsymbol{\mu}})\colon \delta\mathbf{1} \le \tilde{\boldsymbol{\mu}} \le H\mathbf{1}\}$ is uniformly integrable. By Parzen (1954), this implies that the weak law of large numbers holds uniformly over $\{\tilde{\boldsymbol{\mu}}\colon \delta\mathbf{1} \le \tilde{\boldsymbol{\mu}} \le H\mathbf{1}\}$. Let ε be as in Lemma 2. Then

$$\lim_{r \to \infty} \sup_{\delta\mathbf{1} \le \tilde{\boldsymbol{\mu}} \le H\mathbf{1}} \Pr\left\{ |\mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{(m+1)} - \mu_i| > \frac{\varepsilon}{\sqrt{d}} \,\middle|\, \hat{\boldsymbol{\mu}}^{(m)} = \tilde{\boldsymbol{\mu}} \right\} = 0. \tag{17}$$

This is where the assumption $\mu_i \ge \delta \ \forall\ i$ is critical. We must have a compact subset of $\tilde{\boldsymbol{\mu}}$'s in $\mathscr{U}$ in order for (16) to hold and so $\tilde{\boldsymbol{\mu}}$ must be bounded away from $\mathbf{0}$.

By (17), we can choose R_2 large enough so that $r \ge R_2$ implies

$$\sup_{\delta\mathbf{1} \le \tilde{\boldsymbol{\mu}} \le H\mathbf{1}} \Pr\left\{ |\mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{(m+1)} - \mu_i| > \frac{\varepsilon}{\sqrt{d}} \,\middle|\, \hat{\boldsymbol{\mu}}^{(m)} = \tilde{\boldsymbol{\mu}} \right\} < \frac{1}{2d}. \tag{18}$$

If the algorithm is run with $r > R_2$, then a union bound yields

$$\Pr\big\{ \|\hat{\boldsymbol{\mu}}^{(m+1)} - \boldsymbol{\mu}\| < \varepsilon \,\big|\, \hat{\boldsymbol{\mu}}^{(m)} \big\} \ge \tfrac{1}{2} I\big[\delta\mathbf{1} \le \hat{\boldsymbol{\mu}}^{(m)} \le H\mathbf{1}\big]. \tag{19}$$

Let R_1, c, and ν be as in Lemma 2, and suppose that the algorithm is run with number of replications $r \geq R = \max(R_1, R_2)$. The sequence $\langle \hat{\boldsymbol{\mu}}^{(m)} \rangle_{m=0}^{\infty}$ is Markov, so step 4 of the algorithm, (15) and (19) imply that

$$\begin{aligned}
&\Pr\{\|\hat{\boldsymbol{\mu}}^{(m+2)} - \boldsymbol{\mu}\| < \varepsilon \,|\, \hat{\boldsymbol{\mu}}^{(m)}\} \\
&\quad \geq \Pr\{\|\hat{\boldsymbol{\mu}}^{(m+2)} - \boldsymbol{\mu}\| < \varepsilon \,|\, \delta\mathbf{1} \leq \hat{\boldsymbol{\mu}}^{(m+1)} \leq H\mathbf{1}\} P\{\delta\mathbf{1} \leq \hat{\boldsymbol{\mu}}^{(m+1)} \leq H\mathbf{1} | \hat{\boldsymbol{\mu}}^{(m)}\} \\
&\quad \geq \tfrac{1}{4}.
\end{aligned}$$

Thus, regardless of the value $\hat{\boldsymbol{\mu}}^{(m)}$, if the algorithm is run with $r \geq R$ there is at least 1/4 probability of $\hat{\boldsymbol{\mu}}^{(m+2)}$ being within an ε-neighborhood of $\boldsymbol{\mu}$. Letting

$$\eta_l = P\{\|\hat{\boldsymbol{\mu}}^{(l+2)} - \boldsymbol{\mu}\| < \varepsilon \,|\, \hat{\boldsymbol{\mu}}^{(l)}\},$$

then $\sum_l \eta_l$ diverges, and by the conditional Borel–Cantelli lemma [Corollary 5.29 of Brieman (1968)],

$$\Pr\{\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\| < \varepsilon \text{ infinitely often}\} = 1. \tag{20}$$

We have shown that, if $\hat{\boldsymbol{\mu}}$ starts within an ε-neighborhood of $\boldsymbol{\mu}$, there is a positive probability of exponential convergence occurring by Lemma 2, and have now shown that if $\hat{\boldsymbol{\mu}}$ leaves this ε-neighborhood, it returns to it with probability 1. We now combine the two to complete the proof. Define two sequences of stopping times $\{U_n\}$ and $\{W_n\}$ inductively as follows:

$$\begin{aligned}
&W_0 = 0, \\
&U_n = \inf\{m > W_{n-1} \colon \|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\| < \varepsilon\}, \qquad n \geq 1, \\
&W_n = \inf\{m > U_n \colon \|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\| > c^{m-U_n} \|\hat{\boldsymbol{\mu}}^{(U_n)} - \boldsymbol{\mu}\|\}, \qquad n \geq 1.
\end{aligned}$$

The W_n mark the transitions at which exponential convergence fails after each U_n, and the U_{n+1} mark the transitions thereafter that $\{\hat{\boldsymbol{\mu}}^{(m)}\}$ enters the ε-neighborhood of $\boldsymbol{\mu}$. By Lemma 2, the strong Markov property, and (20),

$$\Pr\{W_n = \infty \,|\, U_n < \infty\} \geq \nu, \tag{21}$$

$$\Pr\{U_n < \infty \,|\, W_{n-1} < \infty\} = 1. \tag{22}$$

Let $G_n = \{W_{n-1} < \infty \text{ and } W_n = \infty\}$. The G_n are obviously disjoint. Note that

$$\bigcap_{l=1}^{n-1} G_l^c = \bigcap_{l=1}^{n-1} \{W_l < \infty\}.$$

By the Markov property, (21) and (22),

$$\begin{aligned}
\Pr\Big\{G_n \,\Big|\, \bigcap_{l=1}^{n-1} G_l^c\Big\} &= \Pr\{G_n \,|\, W_{n-1} < \infty\} \\
&\geq \Pr\{U_n < \infty \,|\, W_{n-1} < \infty\} \Pr\{W_n = \infty \,|\, U_n < \infty\} \\
&= 1\nu = \nu.
\end{aligned}$$

It follows by the conditional Borel–Cantelli lemma that

$$\Pr\left\{\bigcup_{n=1}^{\infty} G_n\right\} = 1.$$

Now, $G_n = \{W_{n-1} < \infty \text{ and } U_n = \infty\} \cup \{U_n < \infty \text{ and } W_n = \infty\}$. By (22), the event on the left has probability zero, so

$$P\left\{\bigcup_{n=1}^{\infty} \{U_n < \infty \text{ and } W_n = \infty\}\right\} = 1. \tag{23}$$

Recall from Lemma 2 that $c < 1$ and hence $-\log_e(c) > 0$. Choose $0 < \theta < -\log_e(c)$ and note $\{e^{m\theta}\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\| \to 0\} \supset \{U_n < \infty \text{ and } W_n = \infty\}$ for all n. Thus, by (23),

$$P\{e^{m\theta}\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\| \to 0\} \geq P\left\{\bigcup_{n=1}^{\infty} \{U_n < \infty \text{ and } W_n = \infty\}\right\} = 1$$

and the theorem is proved. □

We close this section with a proof that exponential convergence is preserved under the modified algorithm which uses previous information [i.e., (5) replaces step 4]. To obtain this result, we assume there is a known upper bound on the true $\boldsymbol{\mu}$. This holds in many situations, for example, when $\boldsymbol{\mu}$ is a probability. We also give some discussion indicating that the exponential rate can be improved by good choice of the weight w in (5).

PROPOSITION 1. *Under the same assumptions as the theorem, the same conclusions hold for the modified algorithm with* (5) *in place of step* 4.

PROOF. The steps in the proof of the theorem are equally valid for the modified algorithm, except for (8), (15) and (20). To deal with (8), note that the same argument shows

$$\begin{aligned} &E\big[\|\hat{\boldsymbol{\mu}}^{(m+1)} - \boldsymbol{\mu}\|^2 I[\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\| < \varepsilon] \,\big|\, \hat{\boldsymbol{\mu}}^{(m)}\big] \\ &\quad \leq \big\{w^2 b\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\|^2 + (1-w)^2\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\|^2\big\} I[\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\| < \varepsilon] \\ &\quad = \big[w^2 b + (1-w)^2\big]\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\|^2 I[\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\| < \varepsilon]. \end{aligned}$$

Then (8) holds but with b replaced by

$$b^\star = w^2 b + (1-w)^2. \tag{24}$$

Since $0 < w \leq 1$, it follows that $0 < b^\star < 1$.

A simple way to establish (15) is to bound each $\hat{\boldsymbol{\mu}}^{(m)}$ from above. Redefining H as a credible upper bound for the solution $\boldsymbol{\mu}$, such truncation ensures the compactness used in (16). Truncated in this fashion, the modified algorithm conforms to (15).

To establish a version of (20), first let

$$E_m = \left\{ |\mathbf{x}_i^T \hat{\boldsymbol{\beta}}^{(m)} - \mu_i| \le \frac{\varepsilon}{2\sqrt{d}},\ 1 \le i \le d \right\}.$$

Now (18) holds for r sufficiently large, and we take r large enough that by subadditivity,

$$\sup_{\delta \mathbf{1} \le \tilde{\boldsymbol{\mu}} \le H\mathbf{1}} \Pr\bigl[E_{m+1} \bigm| \hat{\boldsymbol{\mu}}^{(m)} = \tilde{\boldsymbol{\mu}}\bigr] \ge \tfrac{1}{2}. \tag{25}$$

For any $N > 0$,

$$\hat{\mu}_i^{(m+N)} = (1-w)^N \hat{\mu}_i^{(m)} + w \sum_{n=1}^{N} (1-w)^{N-n} \min\bigl\{\max\bigl\{[\mathbf{X}\hat{\boldsymbol{\beta}}^{(m+n)}]_i, \delta\bigr\}, H\bigr\}.$$

As $0 < w \le 1$, we can choose N sufficiently large that $(1-w)^N < \varepsilon/(4H\sqrt{d})$, and then if $\hat{\boldsymbol{\mu}}^{(m)} \le H\mathbf{1}$, we have $0 < (1-w)^N \hat{\mu}_i^{(m)} < \varepsilon/(4\sqrt{d})$. We also assume $(1-w)^N \mu_i < \varepsilon/(4\sqrt{d})$. If in addition,

$$\bigl|\min\bigl\{\max\{[\mathbf{X}\hat{\boldsymbol{\beta}}^{(m+n)}]_i, \delta\}, H\bigr\} - \mu_i\bigr| < \frac{\varepsilon}{2\sqrt{d}} \quad \text{for } n = 1, 2, \ldots, N,$$

then

$$\begin{aligned}
\bigl|\hat{\mu}_i^{(m+N)} - \mu_i\bigr| &\le (1-w)^N \bigl|\hat{\mu}_i^{(m)} - \mu_i\bigr| \\
&\qquad + w \sum_{n=1}^{N} (1-w)^{N-n} \bigl|\min\bigl\{\max\{[\mathbf{X}\hat{\boldsymbol{\beta}}^{(m+n)}]_i, \delta\}, H\bigr\} - \mu_i\bigr| \\
&\le \frac{\varepsilon}{2\sqrt{d}} + [1 - (1-w)^N] \frac{\varepsilon}{2\sqrt{d}} \\
&\le \frac{\varepsilon}{\sqrt{d}},
\end{aligned}$$

and we conclude that

$$\bigl\{\|\hat{\boldsymbol{\mu}}^{(m+N)} - \boldsymbol{\mu}\| < \varepsilon\bigr\} \supset \bigl\{\hat{\boldsymbol{\mu}}^{(m)} \le H\mathbf{1}\bigr\} \cap \bigcap_{n=1}^{N} E_{m+n}.$$

Note that the newly defined $\langle \hat{\boldsymbol{\mu}}^{(m)} \colon m = 1, 2, \ldots \rangle$ is still a Markov chain. By the Markov property, (15) and (25),

$$\Pr\bigl\{\|\hat{\boldsymbol{\mu}}^{(m+N+1)} - \boldsymbol{\mu}\| < \varepsilon \bigm| \hat{\boldsymbol{\mu}}^{(m)}\bigr\} \ge 2^{-(N+1)}.$$

The result (20) now follows by the conditional Borel–Cantelli lemma, which completes the proof of the proposition. □

If one assumes that an exact exponential rate applies to the original algorithm (rather than as an upper bound, as the theorem asserts), then it is possible to obtain a more or less optimal weight w for the modified algorithm.

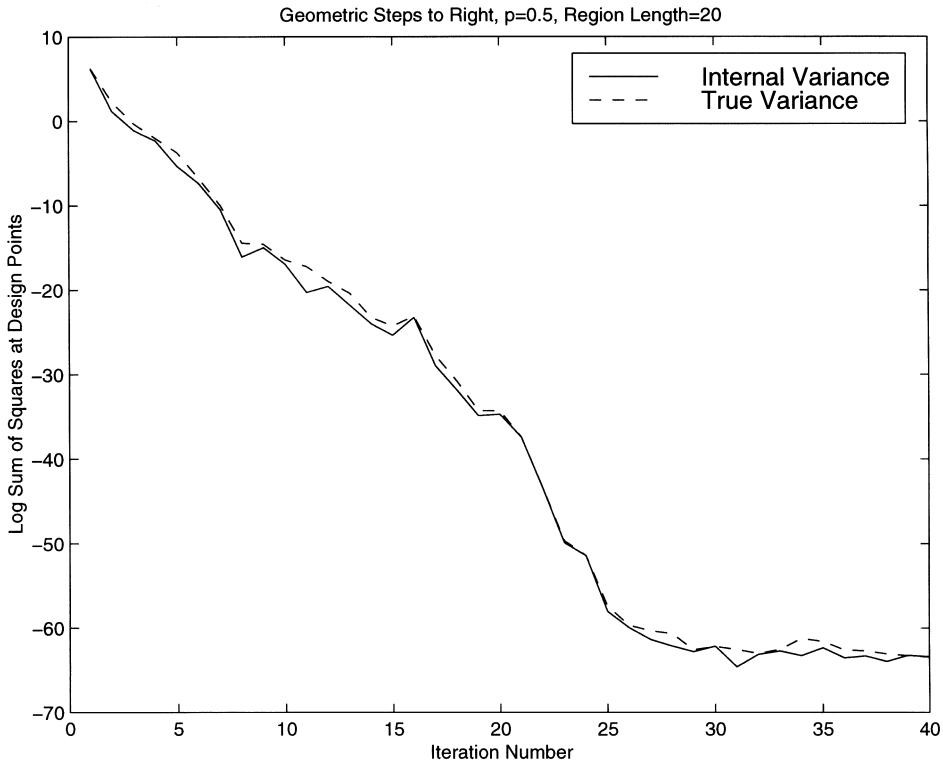

FIG. 1. *Plot of the logarithm of the error versus iteration number for Example* 1.

By an "exact exponential rate," we mean something like: there exists $0 < \kappa < 1$ such that

$$E\left[\|\hat{\boldsymbol{\mu}}^{(m+1)} - \boldsymbol{\mu}\|^2 \middle| \hat{\boldsymbol{\mu}}^{(m)}\right] \approx \kappa \|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\|^2.$$

Our simulation results suggest this may well be the case (see Figure 1). Then, as in (24), the modified algorithm satisfies the same approximation but with κ replaced by $\kappa^\star$. The approximately optimal weight factor is

$$w \approx 1/(\kappa + 1), \tag{26}$$

giving $\kappa^\star \approx \kappa/(\kappa + 1)$. For κ close to 1, this reduces the exponential factor by roughly $1/2$. This heuristic discussion is corroborated by computational results in Section 4.

4. Examples. We now illustrate various features of the algorithm in some simple examples.

EXAMPLE 1. Consider the problem of counting the number of steps to absorption in a simple system. A particle "moves" on the integers $\{1, 2, \ldots\}$ ac-

cording to the transition probability matrix $\mathbf{P}$ having entries

$$p_{ij} = \begin{cases} 0.5^{j-i+1}, & \text{if } i \le j \text{ and } i < 20, \\ 1, & \text{if } i \ge 20 \text{ and } j = \Delta; \\ 0, & \text{otherwise,} \end{cases}$$

which is a random walk with geometric steps except for when the particle moves beyond 19, when it is absorbed. With the score function $s_{ij} = 1$ for all i and j, $i \neq \Delta$, μ_i is the expected number of steps until absorption.

The memoryless property of the geometric distribution implies that $\mu_i = 21 - i$. We used the linear model $\mu_i = \beta_0 + \beta_1 i$ with six replications of the four-point design $(1, 7, 13, 19)$. The results are shown in Figure 1. Convergence is measured in two ways. First, by the quantity $\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\|^2$ as in the main theorem and also by the "internal" variance $A_m = \|\bar{\mathbf{y}}^{(m)} - \hat{\boldsymbol{\mu}}_D^{(m)}\|^2$. The latter estimate could be used in practice when $\boldsymbol{\mu}$ is unknown. Note that the two quantities track each other quite well, which suggests a simple stopping criterion for the algorithm: stop when A_m is deemed sufficiently small. The accuracy in the example levels off at machine precision in about 30 iterations. The roughly linear (in log scale) decrease prior to leveling off exemplifies "exponential convergence."

EXAMPLE 2. To illustrate the use of previous information as in Proposition 1, we consider a situation where convergence is slower. We "overfit" the mean function for the geometric step model of Example 1 with a fourth degree polynomial using a single replication of the seven-point design $(1, 4, 7, 10, 13, 16, 19)$. Using seven simulated scores to estimate five parameters at each iteration slows convergence: a histogram of the iteration-to-iteration slopes in the log-sum-of-squares plot observed over 100 simulated runs is shown in the top panel of Figure 2. The slopes were estimated by fitting a simple linear regression to $\log \|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\|^2$ as a function of m. The slopes are clustered about -1 (the observed mean is -0.99) so $\hat{\kappa} \approx e^{-1}$ and, correspondingly, an estimated optimal weight as in (26) is $\hat{w} = e/(1 + e)$. An additional 100 simulations were then run, this time using $\hat{w}$ to weight contributions from the preceding iterations. A histogram of the slopes is given in the bottom panel of Figure 2, showing a clustering about -1.3. We expect to see $\hat{\kappa}^\star = 1/(1 + e)$ and a corresponding slope of $-\log(1 + e) = -1.31$, matching the observed data fairly well. Note that the extra factor of $e^{-0.3}$ in the exponential convergence rate means that one obtains an order of magnitude reduction in error for each eight iterations with virtually no extra effort.

EXAMPLE 3. To illustrate the importance of the number of replications, we consider another situation where convergence is slow. We again overfit in the geometric step model with a second degree polynomial and use a three-point base design (1,5,9). The results of simulations using 10, 20, 40 and 80 replications of the base design (50 simulations each) are shown in Figure 3, with the

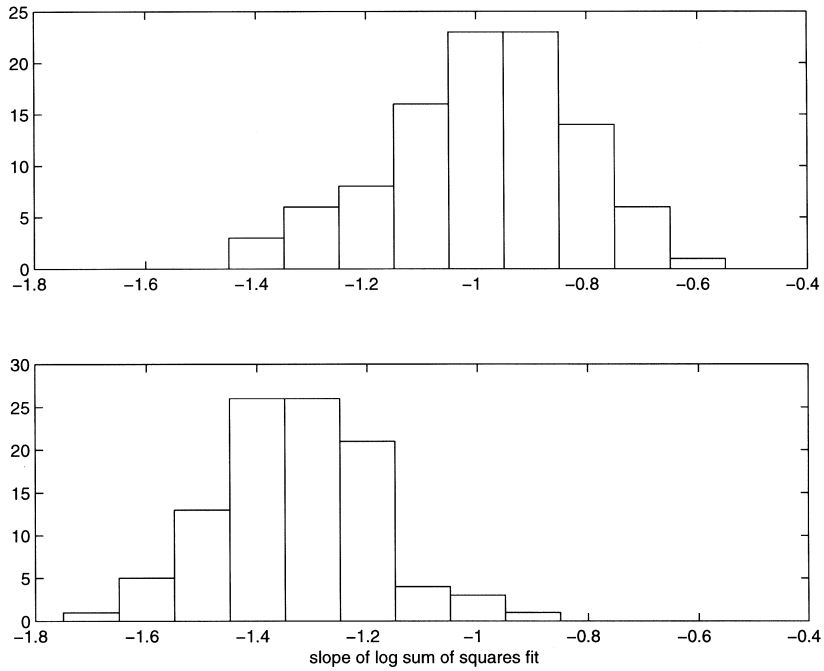

FIG. 2. *Histograms of* 50 *slope estimates for Example* 2. *The upper plot is for the standard algorithm and the bottom plot for the algorithm that uses previous information.*

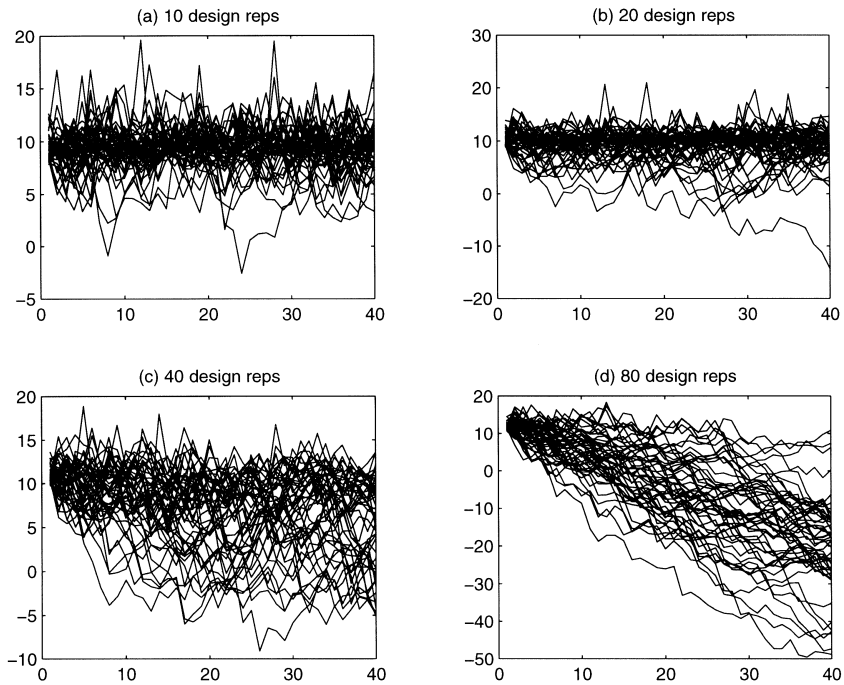

FIG. 3. *Illustration of required number of replications needed to achieve exponential convergence in Example* 3. *Note that the vertical scales are different among the plots.*

true log sum of squares at the design points being plotted against the iteration number. Note the different ranges of the y-axes. For 10 and 20 replications, there is very little if any indication of exponential convergence in Figure 3. The situation with 40 replications is not totally clear, but the results for 80 replications are unequivocal. These results suggest how one might implement a method for obtaining a requisite replication number to obtain exponential convergence: keep doubling the number of replications until evidence of exponential convergence is apparent.

EXAMPLE 4. Next, we consider the gambler's ruin problem and use it to illustrate the behavior of the algorithm when the linear model (2) is not exact. A gambler with i dollars bets repeatedly against the house, whose initial capital is $z - i$ dollars. At stake on each bet is one dollar, and bets are independent. The gambler has probability p of winning each bet, and betting continues until either the gambler is reduced to zero dollars (and is "ruined") or the gambler's fortune reaches z dollars (at which point he has "broken the bank"). Analog transition probabilities are

$$p_{ij} = \begin{cases} p, & \text{if } j - i = 1, \text{ or } i = z - 1 \text{ and } j = \Delta; \\ q, & \text{if } i - j = 1, \text{ or } i = 1 \text{ and } j = \Delta; \\ 0, & \text{otherwise.} \end{cases}$$

We are interested in the probability of the gambler's eventual ruin, which is $\boldsymbol{\mu}$ when the scores s_{ij} are all zero except $s_{1\Delta} = 1$. A well known result, the probability of ruin as a function of the initial fortune i is

$$\mu_i = \frac{(q/p)^z - (q/p)^i}{(q/p)^z - 1},$$

when $q = 1 - p \neq p$, and $\mu_i = 1 - i/z$ when $q = p = 1/2$.

Consider estimating the parameter β in the linear model $\mu_i = \beta[(q/p)^z - (q/p)^i]$. Results for the case of $z = 20$ dollars in the game, $p = 0.6$ and $\delta = 0.05$ added to all terminal transitions, using six replications of the four-point design $(1, 7, 13, 19)$, are shown in Figure 4. Convergence results using only the actual norm $\|\hat{\boldsymbol{\mu}}^{(m)} - \boldsymbol{\mu}\|^2$ are shown (the internal variance estimates track very closely).

Also shown in Figure 4 are the results of three simulations where an error is deliberately introduced into the model used to fit the data. We do this by replacing (q/p) with $(q/p) + \text{err}$ in the formula for μ_i given above. The three values used for err are 10^{-3}, 10^{-5} and 10^{-7}. Exponential convergence to zero variance is not possible for adapting to this class of functions because it does not contain the true $\boldsymbol{\mu}$. Nonetheless, exponential convergence to a limiting accuracy is achieved, after which no further improvement occurs. Of course, n^{-1} convergence can be obtained thereafter by averaging the observed $\hat{\boldsymbol{\mu}}^{(m)}$. We conjecture that the behavior in Figure 4 is typical when adapting to classes of importance functions that do not contain the true solution.

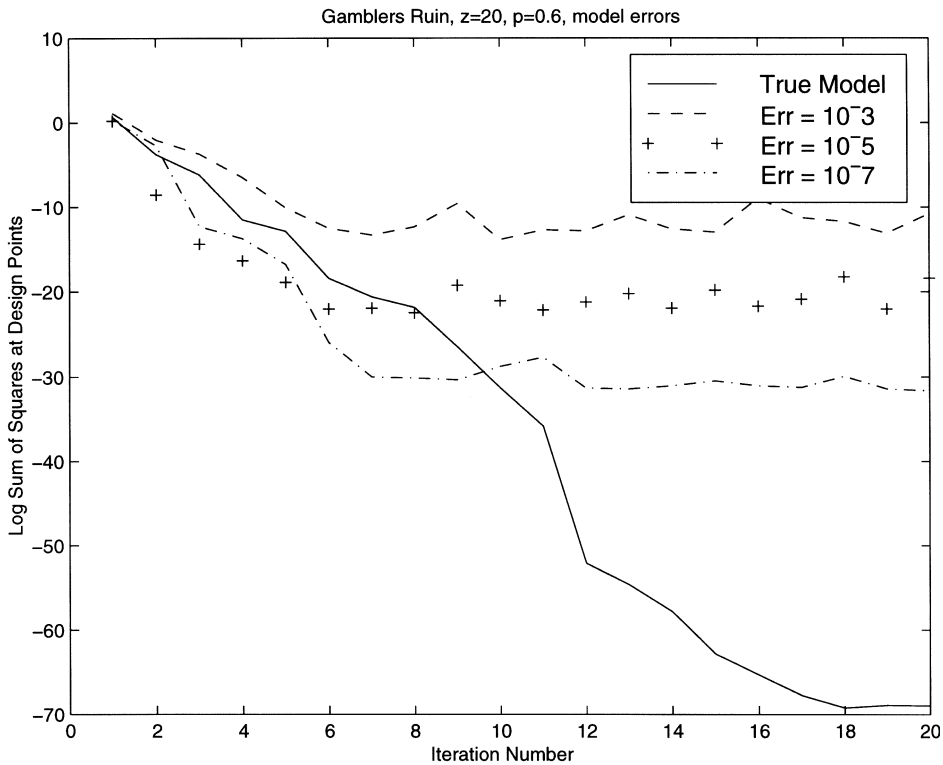

FIG. 4. *Plots of accuracy versus number of iterations for Example* 4.

5. Discussion. The main theorem establishes an exponential upper bound on the rate of convergence under certain conditions (e.g., a linear model for $\boldsymbol{\mu}$ and a minimal number of replications r of the base design). We expect that the upper bound $\exp[-\theta m]$ which can be dug out of the proof is probably exceedingly conservative and useless in practice. As discussed in Example 1, the accuracy measure $A_m = \|\bar{\mathbf{y}}^{(m)} - \hat{\boldsymbol{\mu}}_D^{(m)}\|^2$ can be monitored to evaluate convergence, and the algorithm terminated when this is sufficiently small. The theorem provides no explicit value for a minimal r to obtain exponential convergence. In our experience, insufficient replication is reflected by no improvement in the accuracy measure A_m during early iterations. Such behavior was introduced into Example 3 by using a model with too many terms and a poor base design D. These difficulties were easily overcome by doubling r until exponential convergence was evident.

The requirement of an (exact) linear model for the solution $\boldsymbol{\mu}$ in our main theorem is a limitation. One generalization would be to models nonlinear in their underlying parameters. As the calculation of variances and distribution theory for nonlinear least squares estimators is difficult, this issue remains open. In the majority of applications there is no known parametric form for $\boldsymbol{\mu}$. General qualities of $\boldsymbol{\mu}$ may be known from physical principles (such as conti-

nuity of the solution), but these do not explicitly translate into a parametric model. A linear form for $\boldsymbol{\mu}$ may be used as an approximation. As illustrated in Example 4, the better the approximation, the better the performance of adaptive importance sampling, which of course will not converge without further modification (e.g., averaging across iterations). One area we plan to investigate further is the use of linear models based on series approximations which become more accurate as the order is increased.

Most problems in particle transport involve continuous state spaces. Although the algorithm described herein is easily extended to a general state space, rigorously establishing its convergence properties in that domain is nontrivial. Simulation evidence suggests that exponential convergence continues to hold. There is also the possibility of utilizing adaptive importance sampling in conjunction with other Monte Carlo variance reduction techniques and deterministic algorithms. These techniques may be especially useful in early iterations of the algorithm when the fitted model $\mathbf{X}\hat{\boldsymbol{\beta}}_D$ is sufficiently far from the actual $\boldsymbol{\mu}$, but we do not explore them further here. When the state space has small dimension, deterministic methods are probably superior to adaptive importance sampling. As the dimension of the space grows, comparisons become more problematic—deterministic methods typically scale exponentially in the number of dimensions, while Monte Carlo methods can often be run with a limited amount of computer time and provide a solution with some estimable uncertainty (which may be large). As such, adaptive importance sampling may become an attractive alternative to deterministic methods and classical Monte Carlo in these settings.

The potential of the adaptive method is great. It escapes n^{-1} convergence by intelligently exploiting dependent data (i.e., by learning) and achieves exponential variance reduction. The algorithm described here used to attain such convergence is conceptually straightforward enough to extend to other problems.

REFERENCES

APOSTOL, T. A. (1957). *Mathematical Analysis*. Addison-Wesley, Reading, MA.

BOOTH, T. E. (1985). Exponential convergence for Monte Carlo particle transport? *Trans. Amer. Nuclear Soc.* **50** 267–268.

BOOTH, T. E. (1986). A Monte Carlo learning/biasing experiment with intelligent random numbers. *Nuclear Science and Engineering* **92** 465–481.

BOOTH, T. E. (1988). The intelligent random number technique in MCNP. *Nuclear Science and Engineering* **100** 248–254.

BOOTH, T. E. (1989). Zero-variance solutions for linear Monte Carlo. *Nuclear Science and Engineering* **102** 332–340.

BRIEMAN, L. (1968). *Probability*. Addison-Wesley, Reading, MA.

BRIESMEISTER, J. F., ed. (1993). MCNP—A general Monte Carlo N-particle transport code. Version 4A, Los Alamos National Laboratory Report LA-12625-MS.

CARTER, L. L. and CASHWELL, E. D. (1975). *Particle Transport Simulation and the Monte Carlo Method*. Technical Information Center, Springfield, VA, Energy Research and Development Administration.

GLASSERMAN, P. (1993a). Stochastic monotonicity and conditional Monte Carlo for likelihood ratios. *Adv. in Appl. Probab.* **25** 103–115.

GLASSERMAN, P. (1993b). Filtered Monte Carlo. *Math. Oper. Res.* **18** 610–634.
HALTON, J. H. (1962). Sequential Monte Carlo. *Proc. Cambridge Philos. Soc.* **58** 57–78.
HAMMERSLEY, J. M. and HANDSCOMB, D. C. (1964). *Monte Carlo Methods*. Chapman and Hall, London.
KALOS, M. H. and WHITLOCK, P. A. (1986). *Monte Carlo Methods*. Wiley, New York.
KOLLMAN, C. (1993). Rare event simulation in radiation transport. Ph.D. dissertation, Dept. Statistics, Stanford Univ.
LUX, I. and KOBLINGER, L. (1991). *Monte Carlo Particle Transport Methods: Neutron and Photon Calculations*. CRC Press, Boca Raton, FL.
MARSHALL, A. W. (1956). The use of multi-stage sampling schemes in Monte Carlo computation. In *University of Florida Symposium on Monte Carlo Methods* 123–140. Wiley, New York.
PARZEN, E. (1954). On uniform convergence of families of sequences of random variables. *Univ. California Publications in Statistics* **2** 23–54.

C. KOLLMAN
NATIONAL MARROW DONOR PROGRAM
MINNEAPOLIS, MINNESOTA 55413

K. BAGGERLY
DEPARTMENT OF STATISTICS
RICE UNIVERSITY
HOUSTON, TEXAS 77251-1892

D. COX
DEPARTMENT OF STATISTICS
RICE UNIVERSITY
HOUSTON, TEXAS 77251-1892
E-MAIL: dcox@stat.rice.edu

R. PICARD
STATISTICS GROUP
LOS ALAMOS NATIONAL LABORATORY
LOS ALAMOS, NEW MEXICO 87545