Version: February 1, 1997

BUDGET-DEPENDENT CONVERGENCE RATE OF STOCHASTIC APPROXIMATION

PIERRE L'ECUYER* AND GEORGE YIN†

Abstract. Convergence rate results are derived for a stochastic optimization problem where a performance measure is minimized with respect to a vector parameter θ. Assuming that a gradient estimator is available and that both the bias and the variance of the estimator are (known) functions of the budget devoted to its computation, the gradient estimator is employed in conjunction with a stochastic approximation (SA) algorithm. Our interest is to figure out how to allocate the total available computational budget to the successive SA iterations. The effort is devoted to solving the asymptotic version of this problem by finding the convergence rate of SA towards the optimizer, first as a function of the number of iterations, and then as a function of the total computational effort. As a result the optimal rate of increase of the computational budget per iteration can be found. Explicit expressions for the case where the computational budget devoted to an iteration is a polynomial in the iteration number, and where the bias and variance of the gradient estimator are polynomials of the computational budget, are derived. Applications include the optimization of steady-state simulation models with likelihood ratio, perturbation analysis, or finite-difference gradient estimators, optimization of infinite-horizon models with discounting, optimization of functions of several expectations, and so on. Several examples are discussed. Our results readily generalize to general root-finding problems through stochastic approximation.

Key words. stochastic optimization, discrete-event systems, stochastic approximation, gradient estimate, rate of convergence, limit theorems

AMS subject classifications. 60F05, 60F17, 62L20, 93E23, 93E25, 93E30

1. Introduction, Motivation, and Examples.

Consider a stochastic model parameterized by a vector $\theta \in G \subseteq \mathbb{R}^r$, where G is either the entire space $\mathbb{R}^r$ or a closed and convex set. The objective is to minimize a cost functional $\alpha(\cdot)$ over G, assuming that α has a unique minimizer θ^* in the interior of G. We are mainly interested in the situation where the exact value of $\alpha(\theta)$ (or its gradient) is too hard to compute, and where only a gradient estimator is available, which can be computed, say, by simulation. The quality of the estimator (e.g., bias, variance, ...) might depend on the parameter value θ at which it is evaluated, and also on the computing budget that we are ready to spend to perform the simulation. To simplify the notation, we assume henceforth that $G = \mathbb{R}^r$. Algorithms with projections, which we use when $G \neq \mathbb{R}^r$, are discussed at the end of the paper. Their asymptotic properties are similar, provided that θ^* is in the interior of G. We do not consider the more complicated situation where the optimum lies on the boundary of G; It requires different techniques for obtaining the rates of convergence. Although we concentrate on the stochastic optimization problems, for which we are finding a root of the gradient of α, the results obtained are also applicable to more general root-finding problems under noisy measurements or observations.

* Département d'I.R.O., Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, H3C 3J7, Canada. The research of this author was supported by NSERC-Canada grant no. OGP0110050 and FCAR-Québec grant no. EQ2831.

† Department of Mathematics, Wayne State University, Detroit, MI 48202, USA. The research of this author was supported in part by the National Science Foundation under grants DMS-9529738 and DMS-9600245, and in part by Wayne State University.

We consider the stochastic approximation (SA) algorithm

$$\theta_{n+1} = \theta_n - a_n \psi_n \tag{1.1}$$

for $n \geq 1$, where $\theta_1 \in G$ is the initial parameter estimate, ψ_n is the value of a gradient estimator (at parameter value θ_n) obtained at iteration n, and $\{a_n,\ n \geq 1\}$ is a decreasing sequence such that $a_n > 0$, $\sum_n a_n = \infty$ and $\sum_n a_n^2 < \infty$. Let T_n be the size of the computing effort made to compute ψ_n. Here, ψ_n and T_n are (generally correlated) random variables whose distributions depend in general on the parameter value θ_n, on the initial conditions (or state) of the simulation at step n (say, s_n), on the simulation time-horizon and/or number of replications or number of regenerative cycles at step n, and perhaps on other factors, such as the size of the finite differences used at step n in the case where a finite-difference (FD) gradient estimator is used. We assume that the probability distribution of (ψ_n, T_n, s_{n+1}) depends on (n, s_n, θ_n) and that conditional on (n, s_n, θ_n), that distribution is independent of the past. In other words, $\{(n, \psi_n, T_n, s_{n+1}, \theta_{n+1}), n \geq 1\}$ evolves as a Markov chain. In some situations, T_n could be a deterministic computing budget devoted to iteration n (its distribution is then degenerate), while in other cases the number of replications is fixed and T_n is random (although we have some indirect control over it). This formulation is quite general. Denote by E_n the expectation conditional on (n, s_n, θ_n).

Let $\alpha_\theta(\cdot)$ denote the gradient of $\alpha(\cdot)$. We define the conditional bias and variance of the gradient estimator at step n as

$$B_n = E_n[\psi_n] - \alpha_\theta(\theta_n) \tag{1.2}$$

and

$$V_n = E[\|\psi_n - E_n \psi_n\|^2], \tag{1.3}$$

respectively, where $\|\cdot\|$ denotes the Euclidean norm. B_n is a random variable, but not V_n. Note that V_n is the trace of the "covariance" matrix $E[(\psi_n - E_n\psi_n)(\psi_n - E_n\psi_n)']$ (here and throughout the paper, the prime denotes matrix transpose). Note that in V_n, E_n is used in lieu of E as in the standard definition of variance, but for convenience, we still call it variance. We shall use the following notation: if f and g are positive-valued functions defined over the natural numbers, we say that "$f(n)$ is of the *order* $g(n)$ as $n \to \infty$," denoted by $f(n) = O(g(n))$, if there exist a positive integer n_0 and a constant $K > 0$ such that $f(n) \leq K g(n)$ for all $n \geq n_0$. If $f(n) = O(g(n))$ and $g(n) = O(f(n))$, we say that "$f(n)$ is of the *exact order* $g(n)$," denoted by $f(n) = \Theta(g(n))$. If $f(n)/g(n) \to 1$, we say that "$f(n)$ is *similar* to $g(n)$", denoted by $f(n) \sim g(n)$.

SA algorithms have been studied extensively and employed in a wide range of applications; see for example Wasan [40], Nevel'son and Khasminskii [32], Kushner [19], Kushner and Clark [20], L'Ecuyer and Glynn [29], among others. For an extensive survey on general stochastic approximation for parameterized models of either continuous or discrete event systems, see Kushner and Vázquez-Abad [22] and the references therein. Those analyses generally assume that the computational effort T_n is (roughly) the same at all iterations, and the convergence rates are obtained in terms of the number of iterations only.

In this paper, we consider a more general case where the computing budget could vary between iterations, and analyze the convergence rate of the mean square error

(MSE), defined as $E\|\theta_n - \theta^*\|^2$, in terms of the total computational effort

$$C_n = \sum_{i=1}^{n} T_i. \tag{1.4}$$

We seek to achieve

$$E\left[\|\theta_n - \theta^*\|^2\right] = O((E[C_n])^{-\eta}) \tag{1.5}$$

for the largest possible constant η. This is motivated by the fact that in certain situations (see the examples below), B_n will converge to zero (and θ_n will converge to the optimizer) only if the computing time T_n increases towards infinity as a function of n, thus the T_n's are not the same at all the iterations (and are also typically random). In that context, the convergence rate expressed as a function of the total computational effort is more meaningful than that expressed as a function of n. Typically, both B_n and V_n are (roughly) polynomial functions of $E[T_n]$ and the rate of change of the latter can be chosen directly or indirectly. An interesting question is then: What is the optimal rate of increase of $E[T_n]$? A first step toward answering this question is to figure out how the mean square error depends on B_n and V_n.

Glynn and Whitt [16] have developed a framework for studying the asymptotic efficiency of simulation estimators as a function of the available computational budget. Their goal was to capture the interplay between the variability of an estimator and the computational effort required. They obtained limit theorems for several examples including two SA settings: classical Robbins-Monro and Kiefer-Wolfowitz based on central finite differences. Their analysis for these two examples is based on convergence rate results (in terms of n) by Ruppert [36]. Our results generalize those studies. To further motivate our development, we now introduce some examples where our framework is appropriate.

EXAMPLE 1.1. Consider a $GI/GI/1$ queue with mean arrival rate $\lambda = 1$ and mean service time $\theta \geq 0$, such that θ is a scale parameter of the service time distribution. More specifically, each service time can be written as $S = \theta Z$ where Z has mean 1 and a distribution independent of θ. For $0 \leq \theta < 1$, the system is stable. Let $w(\theta)$ be the steady-state average wait (in the queue) per customer. Suppose that we want to minimize the function $\alpha(\theta) = w(\theta) + C(\theta)$, where C is a smooth convex function with known derivative, over the interval $G = [0, \bar{\theta}]$, for some $0 < \bar{\theta} < 1$. We design the algorithm so that the iterate is projected back to $G = [0, \bar{\theta}]$ whenever it goes outside, and we take $\bar{\theta} < 1$ to make sure that the system is stable for each θ considered (in fact, this is necessary if one simulates a given number of busy cycles at each step of the SA algorithm, but not if one simulates a given number of customers at each step). Here, to estimate the derivative of w, one can start the queue empty and simulate the system until the tth customer starts its service, or until the end of the tth busy cycle, and compute a (generally biased) derivative estimator from that. Then, one adds $dC(\theta)/d\theta$ to obtain a derivative estimator for α. Several specific estimators for the derivative of $w(\theta)$, based on a finite-horizon simulation, are discussed and experimented in L'Ecuyer and Glynn [29] and L'Ecuyer, Giroux, and Glynn [28]. Generally, those estimators have a bias of $O(t^{-1})$ due to the truncation of the horizon, and sometimes additional bias such as that due to the use of finite differences to estimate the gradient. Their variances have different orders of magnitudes. For example, to mention a few, three estimators called infinitesimal perturbation analysis (IPA), likelihood ratio (LR), and LR with control variates (CLR), all based on the simulation of the first t customers, have variances of $O(t^{-1})$, $O(t)$, and $O(1)$, respectively. Here, the computing cost T_n may be assumed roughly proportional to the number of customers simulated (the value of t) at iteration n of SA. A similar situation occurs for other objective functions, more general queueing networks, and many other Markov chains or discrete-event models. For other specific examples, see Yin, Yan, and Lou [43], or Haurie, L'Ecuyer, and van Delft [17], where IPA gradient estimators for the performance of an unreliable manufacturing system, with respect to threshold (hedging point) values, are used for optimization; these estimators have bias and variance both of $O(t^{-1})$. In those references, weak and w.p.1 convergence were proved, but no convergence rate was obtained. □

EXAMPLE 1.2. Suppose that $\alpha(\theta)$ is expressed as a (differentiable) nonlinear function of one or more mathematical expectations, say, $\alpha(\theta) = g(\mu_1(\theta), \ldots, \mu_d(\theta))$, where $g : \mathbb{R}^d \to \mathbb{R}$ and $\mu_j(\theta) = E[X_j]$ for some random variable X_j whose distribution depends on θ. Suppose that an unbiased estimator Y_j is available for $\mu_{j,\theta}(\theta)$, the gradient of μ_j with respect to θ, which is assumed to exist. Let $X_{j,1}, \ldots, X_{j,N}$ and $Y_{j,1}, \ldots, Y_{j,N}$ be N i.i.d. copies of X_j and Y_j, respectively, and let $\bar{X}_j(N) = (1/N)\sum_{i=1}^N X_{j,i}$ and $\bar{Y}_j(N) = (1/N)\sum_{i=1}^N Y_{j,i}$ be their sample averages. A consistent estimator for the gradient of α is then

$$\psi = \sum_{j=1}^{d} g_j(\bar{X}_1(N), \ldots, \bar{X}_d(N)) \cdot \bar{Y}_j(N),$$

where g_j denotes the derivative of g with respect to its jth parameter. This estimator is generally biased because of the non-linearity of g. Under appropriate conditions on g and on the estimators X_j and Y_j, it can be proved that both the bias and variance of ψ are of $\Theta(N^{-1})$ (see Glynn and Heidelberger [13]). Let $T(N)$ be the computing time for performing the required simulations and computing ψ. If $E[T(N)]/N$ converges to some positive constant as $N \to \infty$, uniformly in θ, then $E[T_n]E[B_n]$ and $E[T_n]V_n$ are $\Theta(1)$.

One special case of this example is when g is (or involves) the ratio of two expectations: $g(\mu_1, \mu_2) = \mu_1/\mu_2$. This case arises for example when the regenerative method is applied to construct steady-state estimators, or when estimating conditional probabilities is desired, as well as in other practical situations. The gradient estimator is then

$$\psi = \frac{\bar{Y}_1(N)\bar{X}_2(N) - \bar{X}_1(N)\bar{Y}_2(N)}{(\bar{X}_2(N))^2}.$$

Gradient estimators for ratios of expectations are studied in Glynn, L'Ecuyer, and Adès [15]. □

EXAMPLE 1.3. Consider a continuous-time stochastic process $\{Z(t),\ t \geq 0\}$, whose probability law depends on θ, and which represents a time-varying cost-rate. Let $\rho > 0$ be the discount rate and suppose that we want to minimize the total expected discounted cost

$$\alpha(\theta) = E\left[\int_0^\infty e^{-\rho t} Z(t)dt\right],$$

with respect to θ, where the expectation depends on θ. An unbiased estimator of α_θ might not be available, but a biased estimator can generally be obtained by simulating the process over a truncated (finite) horizon t, using either finite differences, or perturbation analysis, or perhaps other methods. Fox and Glynn [8] analyze different estimators of $\alpha(\theta)$ in terms of their convergence rates as a function of the computational budget. They show in particular that if the cost rate is non-negative, then the truncated-horizon estimator of $\alpha(\theta)$ has bias $O(e^{-\rho t})$ and bounded variance. When estimating the gradient, the bias and variance would also be functions of the horizon t, depending on the gradient-estimation method. The optimal choice of t as a function of n should then depend on that. □

In the preceding examples, conditional on n and θ_n, the gradient estimator ψ_n is independent of s_n. In Example 1.1, the condition is satisfied because we restarted the system empty at every iteration. However, one can instead restart the queue at iteration n from the same state as it was at the end of iteration $n-1$. In that case, ψ_n depends also on s_n and this is covered by our results.

The rest of the paper is arranged as follows. In Section 2, we give the precise formulation of the problem, state our assumptions, and present a preliminary result saying that convergence w.p.1 to θ^* occurs under our assumptions. Section 3 is devoted to obtaining a bound on the order of magnitude of the MSE. We first obtain such a bound as a function of n. Roughly, we show that if $\|B_n\| = O(n^{-\beta})$ w.p.1, $V_n = O(n^{-\delta})$ and $a_n = n^{-\gamma}$, for some real-valued constants β, δ, and $1/2 < \gamma \leq 1$, and if some other technical conditions on the shape of α are satisfied, then

$$E\|\theta_n - \theta^*\|^2 = O(n^{-\kappa}), \tag{1.6}$$

where

$$\kappa = (2\beta) \wedge (\gamma + \delta) \tag{1.7}$$

and $x \wedge y$ denotes $\min\{x, y\}$. Later on, we express that bound as a function of $E[C_n]$ for the case where $E[T_n]$ grows polynomially fast and where B_n and V_n grow at specific rates as functions of $E[T_n]$.

Section 4 deals with the asymptotic distribution of $n^{\kappa/2}(\theta_n - \theta^*)$, where κ is a constant that gives the appropriate scaling for that distribution to be nontrivial. We show that the asymptotic distribution depends on β and δ, i.e., different decreasing rates for the bias or variance result in different scalings in the asymptotic distribution. Additional conditions to those of Section 2 are required for that analysis. These conditions require in particular that $n^\beta\|B_n\|$ and $n^\delta V_n$ converge to a constant scalar and vector, respectively, and our results imply that $n^{\kappa/2}(\theta_n - \theta^*)$ either converges to a constant or to a normal random variable. In Section 5, we derive the (asymptotically) optimal growth rates of T_n for several practical situations, including variants of the previous examples. As an illustration, if $(E[T_n])^b\|B_n\| = \Theta(1)$, $(E[T_n])^d V_n = \Theta(1)$, $n^{-p}E[T_n] = \Theta(1)$ for some constants $b \geq 0$, d, $p \geq 0$, and under some additional assumptions, one has

$$E\|\theta_n - \theta^*\|^2 = \Theta((E[C_n])^{-((2\beta)\wedge(\gamma+\delta))/(p+1)}) = \Theta((E[C_n])^{-\eta}) \tag{1.8}$$

with $\beta = bp$ and $\delta = dp$. From that, one can then find the optimal values of p and γ, i.e., those that maximize η in (1.8), for any given values of b and d. Note that if there is no bias ($\beta = b = \infty$) and $1/2 < \gamma \leq 1$, then it makes sense to have $E[T_n]$ bounded ($p = 0$), which yields $\delta = 0$ (the variance V_n is bounded) and $\eta = \gamma$. We then recover the basic results of classical SA. On the other hand, if $b < \infty$, then we must take $p > 0$, otherwise we get $-\eta \geq 0$. Of course, the asymptotically optimal growth rate for $E[T_n]$ is not always of the form $\Theta(n^p)$; for instance, for Example 1.3, $\Theta(\log n)$

is better (see Section 5). Some technical details related to those different rates are relegated to Appendix 2. Finally, further discussions and concluding remarks are made in Section 6.

2. Model Assumptions and Almost-Sure Convergence. For notational simplicity, we assume $\theta^* = 0$ henceforth. This can be done without loss of generality, since we can always shift (translate) the origin to make $\theta^* = 0$. We make the following assumptions.

(A0) One has $a_n = n^{-\gamma}$, where $1/2 < \gamma \le 1$. Also, $E\|\theta_1\|^2 < \infty$.

(A1) There are constants $\beta > 0$ and $K_\beta \ge 0$ such that $\gamma + \beta > 1$ and for all n,

$$\|B_n\| \le K_\beta n^{-\beta} \text{ w.p.1 and } E\|B_n\|^2 \le K_\beta^2 n^{-2\beta}.$$

(A2) There are constants $\delta \in \mathbb{R}$ and $K_\delta \ge 0$ such that $2\gamma + \delta > 1$ and for all n,

$$V_n = E\|\psi_n - E_n\psi_n\|^2 \le K_\delta n^{-\delta}.$$

Condition (A0) gives an explicit form for a_n. If in lieu of $n^{-\gamma}$, we use $\Gamma_0 n^{-\gamma}$, where Γ_0 is a constant or a positive definite matrix, then all the results of this paper carry over with minor modifications. However, for notational simplicity, we choose the current setting. Introducing Γ_0 is in fact equivalent to rescaling $\alpha_\theta(\cdot)$ and ψ_n by left-multiplying them by Γ_0 and then using $a_n = n^{-\gamma}$. Therefore, there is no loss of generality in assuming $\Gamma_0 = I$ and the algorithm is thus much simpler. On the other hand, doing such a rescaling generally has an important impact on the convergence speed of the SA algorithm (1.1) and finding the "optimal" rescaling (or even a good one) is generally hard. SA algorithms with averaging (see Section 6) address that problem.

In most applications, especially in Monte Carlo optimization via simulation, the initial value θ_1 is a usually a deterministic constant and the second part of (A0) holds trivially. It is also typically verified when θ_1 is random.

Condition (A1) requires that the bias decreases as $O(n^{-\beta})$. The bound holds uniformly in ω (the sample point), i.e., K_β does not depend on ω (or is non-random). To check that condition, we should normally look for a bound on the conditional bias as a function of the computational effort (e.g., number of replications) uniform over (s_n, θ_n). Condition (A2) would be verified in a similar way. See [29, 30] for examples of how that can be achieved. In (A2), δ can be either positive, zero or negative, which corresponds to a variance that (asymptotically) decreases, remains bounded, or increases with n, respectively. The condition that $2\gamma + \delta > 1$ forces the variance not to increase too fast. $K_\beta = 0$ would mean that there is no bias; $K_\delta = 0$ stands for zero variance.

To carry out the convergence analysis, we also need the following additional assumptions on the function $\alpha_\theta(\cdot)$:

(A3) The function (gradient) $\alpha_\theta(\cdot)$ is continuous.

(A4) The ODE (ordinary differential equation)

$$\dot{\theta} = -\alpha_\theta(\theta) \tag{2.1}$$

has a unique solution for each initial condition and $\alpha_\theta(\theta) = 0$ has a unique root ($= \theta^*$).

(A5) There is a finite constant $K_\alpha > 0$ such that $\|\alpha_\theta(\theta)\| \le K_\alpha(1 + \|\theta\|)$ and there is an $\varepsilon > 0$ and a matrix H such that $\alpha_\theta(\theta) = H\theta + O(\|\theta\|^2)$ for $\|\theta\| \le \varepsilon$. All eigenvalues of H have positive real parts. Let $\lambda_{\min} > 0$ be the minimum of those real parts.

REMARK 2.1. Assumption (A4) requires the corresponding ODE to be "well behaved". Assumption (A5) requires the function α to be quadratic near the optimizer. In the one-dimensional case, this means that the second derivative must be bounded away from zero. It also requires α_θ to grow at most linearly. To derive our convergence rate results in Theorem 3.1 and Section 4, when $\gamma = 1$, we will need λ_{min} to be "large enough". If it is not, then the problem should be rescaled by some matrix Γ_0, as discussed previously. In principle, one can always make λ_{min} as large as needed by rescaling with an appropriate Γ_0. On the other hand, finding the appropriate Γ_0 is not necessarily easy. If the growth rate of the objective function is not linearly bounded, then one possible solution is to project θ_n over a compact set G at each step of (1.1), as in Example 1.1 (see also Section 6, or Komlos and Révész [18], Kushner and Clark [20], Azadivar and Talavage [2], and the references therein). That compact set could vary adaptively between iterations, yielding a SA algorithm with varying bounds, as suggested in, for example, Chen and Zhu [4], Yin and Zhu [45], and Andradóttir [1].

Before studying convergence rates, we first state the convergence of the algorithm in the sense of w.p.1. By virtue of the ODE approach developed in Kushner and Clark [20] (see in particular the argument of their Theorem 2.3.1), the following proposition holds.

PROPOSITION 2.1. *Under assumptions* (A0)–(A5), $\theta_n \to \theta^*$ *w.p.1 as* $n \to \infty$.

3. An Asymptotic Bound on the MSE $E\|\theta_n - \theta^*\|^2$. We now give an upper bound on $E\|\theta_n - \theta^*\|^2$, again assuming that $\theta^* = 0$. As a first step, the next theorem gives such a bound as a function of n. Then, in Corollary 3.2, the bound is related to the expected computational expenditures $E[T_n]$ and $E[C_n]$, for the case of a polynomially-growing (in n) $E[C_n]$. Note that the condition (3.2) is not really restrictive; see the remarks that follow (A5). The proof of Theorem 3.1 is given in Appendix 1.

THEOREM 3.1. *Under the conditions* (A0)–(A5), *with* $\gamma < 1$, *there is an* n_0 *such that for all* $n \geq n_0$,

$$
\begin{aligned}
E\|\theta_{n+1}\|^2 &\leq \frac{K_\beta^2}{\lambda_{\mathsf{min}}^2} n^{-2\beta} + \frac{3K_\delta}{2\lambda_{\mathsf{min}}} n^{-\gamma-\delta} + O(n^{-2\beta-1} + n^{-\gamma-\delta-1}) \\
&= O(n^{-((2\beta)\wedge(\gamma+\delta))}) \;=\; O(n^{-\kappa}).
\end{aligned}
\tag{3.1}
$$

The estimate above also holds for $\gamma = 1$, *under the additional condition that*

$$
\lambda_{\mathsf{min}} > \max\{\beta,\ (1+\delta)/2\}. \tag{3.2}
$$

REMARK 3.1. We can see from the previous theorem how the convergence rate is tied up with the values of β and δ. In absence of bias, the order is purely determined by how the "noise" behaves. In particular, for $\delta = 0$ (bounded variance), our result agrees with the classical result (cf., Kushner [19]) which says that $E\|\theta_n\|^2 = O(n^{-\gamma})$. If $\gamma + \delta > 2\beta$, then the error bound becomes $E\|\theta_n\|^2 = O(n^{-2\beta})$, i.e., the convergence speed depends on how fast the bias diminishes. If $\delta < 0$, then $2\beta \wedge (\gamma + \delta) < \gamma$, so the convergence rate will be slower than that of the classical algorithm, no matter how fast the bias tends to 0. If $\delta = 0$, the order cannot be better than $O(n^{-\gamma})$. Finally, for the totally degenerate case where $\beta = \delta = \infty$, that is, no bias and no noise, the theorem implies that $E\|\theta_n\|^2$ converges faster (asymptotically) than $n^{-\kappa}$ for any κ, which agrees with the classical results of nonlinear programming.

We now express the MSE bound in terms of the computer budget, for the important special case where $E[T_n]$ grows polynomially fast: $n^{-p}E[T_n] \to \kappa_1$ for some $p \geq 0$.

The total computing budget $C_n = \sum_{i=1}^{n} T_i$ then satisfies $n^{-p-1}E[C_n] \to \kappa_1/(p+1)$ as $n \to \infty$. We further assume that the bias and variance are bounded uniformly in terms of the computer budget T_n, as follows: $(E[T_n])^b\|B_n\| \le K_b$ w.p.1 and $(E[T_n])^d V_n \le K_d$ for some positive (fixed) constants K_b and K_d. The convergence rate in Corollary 3.2 depends on p in a significant way.

COROLLARY 3.2. *Under the above assumptions and the conditions of Theorem* 3.1, *with $n^{-\beta}$ and $n^{-\delta}$ in* (A1) *and* (A2) *replaced by $(E[T_n])^{-b}$ and $(E[T_n])^{-d}$, respectively, we have that*

$$E\|\theta_{n+1}\|^2 \quad = \quad O((E[C_n])^{-(2bp)\wedge(\gamma+dp)/(p+1)}). \tag{3.3}$$

Proof. This is a direct consequence of Theorem 3.1. Replace β, δ, K_β, and K_δ in Theorem 3.1 by bp, dp, $K_b\kappa_1^{-b}$, and $K_d\kappa_1^{-d}$, respectively, and observe that $n^{-1}(E[C_n])^{1/(p+1)} = \Theta(1)$. □

4. Asymptotic distribution of $n^{\kappa/2}\theta_n$. In this section, we devote our attention to studying the asymptotic distribution of a suitably scaled sequence of the estimation errors by means of the weak convergence methods. Again, we assume that $\theta^* = 0$ for notational simplicity. In the literature, the rate of convergence is often studied through the asymptotic distribution of a normalized error sequence. The most relevant question in the present context is to derive a nontrivial limit result for $n^{\kappa/2}\theta_n$ and calculate explicitly the asymptotic covariance matrix. The exponent $\kappa/2$ represents the rate of convergence (see, e.g., Kushner and Clark [20], p.233). The covariance matrix of the limiting normal distribution is another important characterization of the convergence speed. Such results are more precise than the bounds obtained in the previous section and may provide further opportunities to improve the performance of the algorithms. Limit theorems of that sort already appeared in Fabian [7] and earlier papers cited there, under assumptions different than ours. The derivation of limit theorems, in the analysis that follows, exploits the roles played by the bias and variance under different sets of assumptions and indicates the dominating factor and main influence to the asymptotic distribution and rate of convergence in each case. In the next section, those results will be applied to several examples, including those introduced in Section 1.

Again, we first derive the results in terms of n and later discuss their interpretation in terms of C_n. In view of (3.3), the relative values of 2β and $\gamma + \delta$ are expected to play an important role in the result. For simplicity, the discussion is focused on the case of $\gamma = 1$, which is the most common case in practice. Similar results hold for the case $1/2 < \gamma < 1$, although one then needs to work with an interpolated sequence as in Kushner and Huang [21] to obtain the limit theorems. In order not to disrupt the flow of discussion, the proofs of the technical results are placed in Appendix 1. Our study will be divided into the following three cases: (1) $1+\delta < 2\beta$; (2) $1+\delta > 2\beta$; (3) $1+\delta = 2\beta$.

4.1. Case 1: $1+\delta < 2\beta$. In this case, from Theorem 3.1, one has

$$E\|\theta_n\|^2 = O(n^{-(1+\delta)}).$$

To obtain the desired asymptotic distribution, we first need to determine an appropriate scaling. We wish to select a real number $\kappa > 0$ such that $n^{\kappa/2}\theta_n$ converges weakly to a nontrivial distribution. It turns out that in this case, the correct order is given by $\kappa = 1 + \delta$, which is motivated by the classical central limit theorem. We derive a functional invariance principle, or functional central limit theorem, which shows how the scaled sequence evolves as a stochastic process. This gives a stronger result than convergence in distribution alone.

To begin with, define

$$W_n(t) = \frac{\lfloor nt \rfloor^{1+\delta/2}}{\sqrt{n}} \theta_{\lfloor nt \rfloor + 1}$$

for $t \in [0,1]$, where $\lfloor z \rfloor$ denotes the integral part of z, for $z \in \mathbb{R}$. Assume that the following conditions hold.

(A6) There is a positive definite matrix R such that

$$n^{\delta} E_n [E_n \psi_n - \psi_n][E_n \psi_n - \psi_n]' \xrightarrow{\text{p}} R \tag{4.1}$$

as $n \to \infty$, where $\xrightarrow{\text{p}}$ denotes convergence in probability.

THEOREM 4.1. *Suppose that the conditions* (A0)–(A6) *and* (3.2) *are in force. Then* $W_n(\cdot)$ *converges weakly to a process* $W(\cdot)$ *which has independent Gaussian increments and covariance matrix* Σt*, where*

$$\Sigma = \int_0^{\infty} \exp(-\tilde{H} u) R \exp(-\tilde{H}' u) du. \tag{4.2}$$

Setting $t = 1$ in Theorem 4.1, the following corollary is immediate, which gives us the usual asymptotic normality result although the scaling factor $n^{(1+\delta)/2}$ is different from the classical theorems due to the formulation of our problem. For $\delta = 0$, we recover the canonical convergence rate of the classical results. In the case where the parameter θ is one-dimensional, (4.2) simplifies to $\Sigma = R/(2\tilde{H})$, where $\tilde{H} = \alpha_{\theta\theta}(0) - (1+\delta)I/2$.

COROLLARY 4.2. *Under the conditions of Theorem* 4.1, $n^{(1+\delta)/2}\theta_n$ *converges in distribution to the normal distribution* $N(0, \Sigma)$.

Theorem 4.1 and Corollary 4.2 indicate that in case $1+\delta < 2\beta$, the bias diminishes rather fast and is asymptotically negligible compared to the variance. The next case is the opposite situation.

4.2. Case 2: $2\beta < 1 + \delta$**.** In this case, the bias term becomes the dominating factor. Similar to Case 1, one may wish to have a limiting distribution result for the scaled sequence $n^{\beta}\theta_n$. We show in what follows that the limiting distribution is degenerate. To proceed, suppose that the following assumptions hold.

(A7) There is a $\bar{B} \in \mathbb{R}^r$ such that as $n \to \infty$,

$$n^{\beta} B_n \xrightarrow{\text{p}} -\bar{B},$$

where B_n is the bias defined in (1.2).

THEOREM 4.3. *Suppose that the conditions of Theorem* 4.1 *are satisfied, with* (A6) *replaced by* (A7)*. Then, as* $n \to \infty$,

$$n^{\beta}\theta_n \xrightarrow{\text{p}} H_b^{-1}\bar{B}.$$

REMARK 4.1. Owing to (3.2), H_b is nonsingular and stable. Theorem 4.3 indicates that the limiting distribution of $n^{\beta}\theta_n$ is degenerate. Hence one cannot expect any asymptotic normality. As n gets large, the noise effect becomes negligible and $\theta_n \sim n^{-\beta} H_b^{-1} \bar{B}$ in the sense of in probability. There is also a functional form of that limiting result. However, due to the degenerate nature, the functional limit does not provide much more than the current statement of the theorem.

4.3. Case 3. $2\beta = 1+\delta$**.** In accordance with the results of the last two subsections, we find that in the current situation, both the noise and the bias contribute to the asymptotic distribution. The next theorem shows that asymptotic normality holds in this case. However, the limiting normal distribution is generally not centered; its mean is given by $H_b^{-1}\bar{B}$ $(= \tilde{H}^{-1}\bar{B}$ in this case), which is also the constant $n^\beta\theta_n$ converges to in case 2 of the previous subsection.

THEOREM 4.4. *Assume that the conditions of Theorem* 4.1 *and Theorem* 4.3 *are satisfied with* $2\beta = 1+\delta$. *Then*

$$n^{(1+\delta)/2}\theta_n - H_b^{-1}\bar{B} \xrightarrow{\mathrm{d}} N(0,\Sigma)$$

as $n\to\infty$, *where* Σ *is given by* (4.2) *and* $\xrightarrow{\mathrm{d}}$ *denotes convergence in distribution.*

Theorem 4.4 could be viewed in a sense as a generalization of the results of the two previous cases. Indeed, if (A6) and (A7) hold with $1+\delta < 2\beta$, then (A7) will hold with $\bar{B}=0$ and β replaced by $(1+\delta)/2 < \beta$. Then, applying Theorem 4.4 with β replaced by $(1+\delta)/2$, we recover Corollary 4.2. Similarly, if $1+\delta > 2\beta$, then, with δ replaced by $2\beta - 1 < \delta$, (A6) holds with $R \equiv 0$ and we also recover Theorem 4.3 from Theorem 4.4.

4.4. Asymptotic distribution in terms of the budget C_n**.** The results obtained in this section can easily be transferred to statements with $n^{\kappa/2}$ replaced by $C_n^{\eta/2}$ for some real number η. To be more specific, suppose that

(A8) $n^{-p-1}C_n \to \kappa_b$ in probability for some $\kappa_b > 0$.

Under the conditions of the previous theorems, we obtain the following:

COROLLARY 4.5. *Assume that conditions* (A0)–(A8) *and* (3.2) *are satisfied. If* $1+\delta < 2\beta$, *then*

$$(C_n/\kappa_b)^{(1+\delta)/(2(p+1))}\theta_n \xrightarrow{\mathrm{d}} N(0,\Sigma) \quad \text{as } n\to\infty.$$

If $1+\delta > 2\beta$, *then*

$$(C_n/\kappa_b)^{\beta/(p+1)}\theta_n \xrightarrow{\mathrm{p}} H_b^{-1}\bar{B} \quad \text{as } n\to\infty.$$

If $1+\delta = 2\beta$, *then*

$$(C_n/\kappa_b)^{\beta/(p+1)}\theta_n - H_b^{-1}\bar{B} \xrightarrow{\mathrm{d}} N(0,\Sigma) \quad \text{as } n\to\infty.$$

REMARK 4.2. If $n^{-p}T_n \to \kappa_1$ in probability as $n\to\infty$, then (A8) holds with $\kappa_b = \kappa_1/(p+1)$. In accordance with Theorem 2 of Glynn and Whitt [16], if $n^{-p}E_n[T_n] \to \kappa_1$ w.p.1 and $n^{-2p-1+\varepsilon}E[T_n - E_n[T_n]]^2 \to \kappa_2$ w.p.1 for some positive constants κ_1, κ_2, and ε, then $n^{-p-1}C_n \to \kappa_b = \kappa_1/(p+1)$ w.p.1. Hence (A8) holds with $\kappa_b = \kappa_1/(p+1)$. Note, however, that only convergence in probability is needed in our result.

5. Specific Setups and Examples. We now study the implications of our results in different situations related to the examples given in the introduction. We shall assume in this section that $\gamma = 1$ and (sometimes implicitly) that the functions considered satisfy (A3)–(A8) and (3.2). We analyze a series of rather general cases arising in the simulation of stochastic discrete-event systems, most of which are summarized in Tables 5.1–5.3. In view of our previous results, we say that the *optimal asymptotic rate* of convergence is reached if, with the largest possible η, either (1.5) holds or $C_n^{\eta/2}\theta_n$ converges in distribution (with a possibly degenerate limit). We spell out the conditions needed, but relegate some technical details to Appendix 2. It turns out that for several situations of interest, the optimal asymptotic rate can be reached for any (non-negative) polynomial rate of increase of T_n; that is, either by spending increasingly more time to get better estimates or by going quickly with cheap estimates.

From Corollary 4.5, one can obtain not only the convergence rate, but also the asymptotic mean square error of $C_n^{\eta/2}\theta_n$ as a function of the other asymptotic constants in the problem. We give examples of that in Appendix 2. In principle, one could then minimize the asymptotic mean square error as well. This may be hard to implement, because the asymptotic constants such as $\bar{B}$, H_b, etc., are typically unknown, but is nevertheless interesting to study from the theoretical point of view.

5.1. Finite-horizon models.

EXAMPLE 5.1. Let $\alpha(\theta) = E[X]$ where X is a random variable whose distribution $F(\theta, \cdot)$ depends on θ, and for which i.i.d. samples can be obtained via simulation (or by any other method). We suppose that the function α satisfies Assumptions (A3)–(A5) and we want to minimize this function with respect to θ, via SA. At each iteration of SA, we need a gradient estimator ψ_n at $\theta = \theta_n$. A classical way of estimating the gradient $\alpha_\theta(\theta)$ is through the use of finite differences (FD) (see, e.g., Kushner and Clark [20], L'Ecuyer and Perron [27], or Zazanis and Suri [46] for more details on FD methods). To simplify the notation here, let θ be a scalar. A generalization to the multidimensional case is straightforward and the rates (e.g., in Table 5.1) are the same (see also Fabian [7] for multidimensional results).

One FD estimator of $\alpha_\theta(\theta)$ is given by

$$\psi = \frac{X^+ - X^-}{2c}, \tag{5.1}$$

where X^+ and X^- are independent r.v.'s with distributions $F(\theta^+, \cdot)$ and $F(\theta^-, \cdot)$, respectively, $\theta^+ = \theta + c$, $\theta^- = \theta - c$, and $c \geq 0$ is the half-size of the FD interval. The latter estimator is called a *central* (or two-sided) FD estimator with independent random numbers and we shall refer to it by the acronym FDc. A forward (or one-sided) estimator, which we call FDf, is defined by

$$\psi = \frac{X^+ - X}{c}, \tag{5.2}$$

where X now has distribution $F(\theta, \cdot)$ and is also independent of X^+. Both estimators are biased. Their bias decreases to 0 as $c \to 0$, but then their variance increases towards infinity.

To reduce the variance, we can take the average of several copies of ψ at any given iteration of the SA algorithm. Suppose that at step n of SA, we generate N_n i.i.d. replicates of ψ as defined by either (5.1) or (5.2), with $\theta = \theta_n$ and $c = c_n$, and take the average as a derivative estimator ψ_n. Let $\psi_{n,i}$ denote the ith copy of ψ thus generated at step n, and $\tau_{n,i}$ be the (random) computing time required to generate it. The pairs $(\psi_{n,i}, \tau_{n,i})$, $1 \leq i \leq N_n$, are then i.i.d., the gradient estimator at step n is

$$\psi_n = \frac{1}{N_n}\sum_{i=1}^{N_n} \psi_{n,i},$$

and the computational expenditure at step n is $T_n = \sum_{i=1}^{N_n} \tau_{n,i}$ (we neglect the time for computing the average and for updating θ_n from ψ_n in (1.1)). Here, N_n is deterministic, but T_n and $C_n = \sum_{i=1}^{n} C_i$ (the cumulated computing time up to step n) are random.

Suppose that the sequences $\{N_n\}$ and $\{c_n\}$ are chosen such that (as $n \to \infty$) $n^{-p}N_n \to \kappa_N$ and $n^{\nu}c_n \to \kappa_c$ for some constants $p \geq 0$, $\nu > 0$, $\kappa_N > 0$, and $\kappa_c > 0$. Thus, ν reflects the rate of decreasing of the finite difference interval (as a function of the iteration number), while p is the rate of increase of the number of replicates of ψ (or computing budget) per SA iteration. We make the (reasonable) assumption that

$$\sum_{i=1}^{n} T_i / \sum_{i=1}^{n} N_i \stackrel{\text{w.p.1}}{\longrightarrow} \kappa_T \tag{5.3}$$

as $n \to \infty$, where κ_T represents the average cost for computing ψ. The left side in (5.3) is the average computing time per replicate of ψ computed during the first n iterations of SA. If θ was fixed, then (5.3) would follow directly from the strong law of large numbers, so this assumption is quite reasonable when θ converges to the optimizer.

Under these conditions, we have that as $n \to \infty$, w.p.1,

$$C_n = \sum_{i=1}^{n} T_i \;\sim\; \kappa_T \sum_{i=1}^{n} N_i \;\sim\; \kappa_T\kappa_N \sum_{i=1}^{n} n^p \;\sim\; \kappa_T\kappa_N n^{p+1}/(p+1).$$

In other words, the cumulative computing cost C_n increases with n as follows:

$$n^{-p-1}C_n \stackrel{\text{w.p.1}}{\longrightarrow} \kappa_T\kappa_N/(p+1) \stackrel{\text{def}}{=} \kappa_b.$$

(A9) Assume that (3.2) holds, that Var $[X]$ is bounded over G and converges to a positive constant as $\theta \to 0$ (recall that X has distribution $F(\theta, \cdot)$). Suppose also that α is twice continuously differentiable in G if we use FDf and three times continuously differentiable in G if we use FDc.

In Appendix 2, we obtain the optimal values of p and ν, and the corresponding values of κ and η, under Assumption (A9). The optimal values are those that maximize η; they are given in the first two lines of Table 5.1. The optimal η can be obtained by taking any $p \geq 0$, provided that ν (the convergence rate of c_n) is chosen as specified in the table. Recall that η tells us the convergence rate of θ_n in terms of the computing budget C_n, whereas κ gives the convergence rate as a function of the number of iterations. The values of β and δ in the table are the largest values that satisfy Assumptions (A1) and (A2) (or the values that satisfy (3.2)). They give the convergence rate of the "bias" $\|B_n\|$ and "variance" V_n of the gradient estimator ψ_n, as functions of n. A negative value of δ means that V_n increases with n.

The case $p = 0$ means that the number N_n of replicates of ψ is the same at all iterations n. In this case, the optimal values are $(\nu, \kappa, \eta) = (1/4, 1/2, 1/2)$ for FDf and $(\nu, \kappa, \eta) = (1/6, 2/3, 2/3)$ for FDc. These values are well-known (see, e.g., Fabian [7], Kushner and Clark [20], and Kushner and Huang [21]) and the SA algorithm is then called the *Kiefer-Wolfowitz* algorithm.

For $p > 0$, N_n increases with n, so we spend more and more time per iteration as n goes up. This increases the cost per iteration, but reduces the variance of ψ_n. We can then decrease c_n at a faster rate, to make the bias decrease faster, so that the variance and squared bias decrease at the same rate. This gives the optimal ν. When

TABLE 5.1
Convergence rates of SA for a finite-horizon model ($\gamma = 1$)

model	p	ν	β	δ	κ	η
FDf	≥ 0	$(p+1)/4$	ν	$p-2\nu$	$(p+1)/2$	$1/2$
FDc	≥ 0	$(p+1)/6$	2ν	$p-2\nu$	$2(p+1)/3$	$2/3$
FDf (CRN1)	≥ 0	$(p+1)/3$	ν	$p-\nu$	$2(p+1)/3$	$2/3$
FDc (CRN1)	≥ 0	$(p+1)/5$	2ν	$p-\nu$	$4(p+1)/5$	$4/5$
FDf (CRN2)	≥ 0	$\geq (p+1)/2$	ν	p	$p+1$	1
FDc (CRN2)	≥ 0	$\geq (p+1)/4$	2ν	p	$p+1$	1
IPA, LR	≥ 0		∞	p	$p+1$	1

$p > 0$, the convergence rate in terms of the number of iterations is "faster" than that in terms of C_n. For large enough p, we even obtain a supercanonical rate ($\kappa > 1$) in terms of n (but not in terms of C_n). The optimal η is the same for all $p \geq 0$; i.e., for an increase of N_n at any polynomial rate, we get the same asymptotic efficiency, provided that c_n decreases at rate $n^{-\nu}$ with the ν specified in the table. This holds for all the "models" considered in Table 5.1. In Appendix 2, we also examine the asymptotic mean and variance of $C_n^{\eta/2}\theta_n$ in terms of the asymptotic constants κ_c, κ_N, and so on.

Finite difference estimators can be improved by using common random numbers (CRN), as explained by, e.g., Glynn [11], Glasserman and Yao [10], and L'Ecuyer and Perron [27]. The basic idea is to view X as a function of θ and ω, say $X(\theta, \omega)$, where ω represents an underlying sample point whose distribution does not depend on θ. For example, in a simulation model, ω may represent a sequence of i.i.d. $U(0,1)$ random variables used to drive the simulation. CRN means using the same ω in the same way for both X^- and X^+ in (5.1), or for both X and X^+ in (5.2). The aim is to induce a strong positive correlation between X^- and X^+ (or X and X^+), without changing their expectations. For a fixed FD interval, the bias of the FD estimator ψ is unchanged by using CRNs, but the variance can be reduced dramatically. L'Ecuyer and Perron [27] show that under the following assumption (A10), the variance of ψ with CRN is $O(1)$ (i.e., bounded) as $c_n \to 0$. Under (A11), the bound is uniform over G, so (A2) holds with $\delta = p$. Recall that with independent random numbers, the variance of ψ is $O(c_n^{-2})$.

(A10) Assume that in some neighborhood of 0, w.p.1, $X(\theta) \equiv X(\theta, \omega)$ is a continuous function of θ, differentiable except perhaps at a denumerable number of points. Also, the derivative $X_\theta(\theta)$, where it exists, is uniformly bounded by a square-integrable random variable (independent of θ).

(A11) Assumption (A10) holds with the neighborhood of 0 replaced by G.

We shall denote the CRN approach by "CRN2" when (A11) is satisfied. The optimal asymptotic rate of $\eta = 1$ is obtained under the conditions specified in Table 5.1: Take any $p \geq 0$, $\nu \geq (p+1)/2$ for FDf, and $\nu \geq (p+1)/4$ for FDc. So, it suffices that $c_n \to 0$ fast enough. In fact, one can as well take "$c_n = 0$" for all n, i.e., take the limit as $c \to 0$ in (5.1) or (5.2), which yields the sample derivative $X_\theta(\theta)$. Under (A11), this sample derivative turns out to be an unbiased derivative estimator, with variance uniformly bounded over G. It is called the *infinitesimal perturbation analysis* (IPA) estimator (Glasserman [9]). For CRN2 and IPA, θ_n converges at rate

$O(C_n^{-1/2})$, which is the canonical rate in terms of C_n. The convergence in terms of n is supercanonical ($\kappa > 1$) when $p > 0$.

If (A11) fails to hold, the variance may still increase much slower, as $c \to 0$, with CRN than with independent random numbers. Under the following conditions, for instance, the variance of ψ is improved to $O(c^{-1})$ (not quite as good as $O(1)$, but better than $O(c^{-2})$).

(A12) There exist finite constants K_1, K_2, and $\bar{c}$, such that

$$\sup_{\theta_0 \in G} \sup_{\|\theta - \theta_0\| \le \bar{c}} |X(\theta, \omega) - X(\theta_0, \omega)| \le K_1 \qquad \text{w.p.1}$$

and for all $c \le \bar{c}$,

$$P[X(\cdot, \omega) \text{ is continuous in } (\theta_0 - c, \theta_0 + c)] \ge 1 - K_2 c.$$

Using an argument similar to that in the proof of Proposition 3 of L'Ecuyer and Perron [27], one can show that Var $[\psi] = O(c^{-1})$ (uniformly in θ) under (A12). We may further assume (as in Glynn [11]) that there is a constant $\sigma_D > 0$ such that cVar $[\psi] \to \sigma_D^2$ as $c \to 0$ and $\theta \to 0$, and call this the "CRN1" setup. The optimal asymptotic rates for CRN1 are $\eta = 2/3$ for FDf and $\eta = 4/5$ for FDc, which is a significant improvement over the case of independent random numbers. These rates are obtained under the conditions given in Table 5.1: $\nu = (p+1)/3$ for FDf and $\nu = (p+1)/5$ for FDc.

Besides IPA, there is another approach for obtaining an unbiased derivative estimator, called the likelihood ratio (LR) method (Glynn [12], L'Ecuyer [25], Rubinstein and Shapiro [35]). Its variance decreases at the same rate as that of IPA as a function of the computing budget, but it is typically larger by a constant factor. On the other hand, there are several classes of systems for which the LR method applies whereas IPA (in its direct form) does not. □

EXAMPLE 5.2. Let us return to Example 1.2. Suppose that g is twice continuously differentiable, that all X_j and Y_j have finite moments of all orders, and that (A3)–(A5) hold. The gradient can be written as

$$\begin{aligned} \alpha_\theta(\theta) &= \sum_{j=1}^{d} g_j(\mu_1(\theta), \ldots, \mu_d(\theta))\mu_{j,\theta}(\theta) \\ &\stackrel{\text{def}}{=} \tilde{g}(\mu_1(\theta), \ldots, \mu_d(\theta), \mu_{1,\theta}(\theta), \ldots, \mu_{d,\theta}(\theta)). \end{aligned}$$

Suppose that the gradient estimator ψ_n at step n is ψ, as given in Example 1.2, based on $N = N_n$ replications of X_j and Y_j, whose bias and variance are assumed to be bounded uniformly in θ, and let p, κ_N, T_n, κ_T, and κ_b be as in Example 5.1. For $1 \le i, k \le 2d$, let Γ_{ik} denote the (i,k)-th entry of the covariance matrix $\Gamma = $ Cov $(X_1, \ldots, X_d, Y_1, \ldots, Y_d)$, let $\tilde{g}_i$ be the ith component of the gradient of $\tilde{g}$, and let $\tilde{g}_{ik}$ be the (i,k)-th entry of the Hessian of $\tilde{g}$. One has

$$\tilde{g}_i = \begin{cases} \sum_{j=1}^{d} g_{ij}(\mu)\mu_{j,\theta}(\theta) & \text{for } i = 1, \ldots, d; \\ g_i(\mu) & \text{for } i = d+1, \ldots, 2d, \end{cases}$$

and

$$\tilde{g}_{ik} = \begin{cases} \sum_{j=1}^{d} g_{ijk}(\mu)\mu_{j,\theta}(\theta) & \text{for } 1 \le i, k \le d; \\ g_{ik}(\mu) & \text{for } 1 \le i \le d < k \le 2d; \\ 0 & \text{for } d < i, k \le 2d, \end{cases}$$

where $\mu = (\mu_1(\theta), \ldots, \mu_d(\theta))$, while g_{ik} and g_{ijk} denote the second and the third derivatives of g with respect to its parameters (i,k) and (i,j,k), respectively. By applying the results of Glynn and Heidelberger [13] and Glynn and Whitt [16, Example 2] to $\tilde{g}$, we find that (A6) and (A7) hold with $\beta = \delta = p$, $\bar{B} = -(1/2)\sum_{i=1}^{d}\sum_{k=1}^{d}\tilde{g}_{ik}\Gamma_{ik}$, and $R = \sum_{i=1}^{d}\sum_{k=1}^{d}\tilde{g}_i\tilde{g}_k\Gamma_{ik}$.

It then follows from Corollary 4.5 that

$$C_n^{1/2}\theta_n \Rightarrow \begin{cases} \kappa_b^{p/(p+1)} H_b^{-1}\bar{B} & \text{if } 0 \le p < 1; \\ \kappa_b^{1/2} N(H_b^{-1}\bar{B}, \Sigma) & \text{if } p = 1; \\ \kappa_b^{1/2} N(0, \Sigma) & \text{if } p > 1. \end{cases} \tag{5.4}$$

This gives $\eta = 1$ for $p \ge 1$. To get the best asymptotic properties, one should choose $p > 1$, because the asymptotic bias vanishes while the asymptotic covariance matrix remains the same as for $p = 1$.

As a special case, suppose that we are interested in minimizing a ratio of expectations, so $g(\mu_1(\theta), \mu_2(\theta)) = \mu_1(\theta)/\mu_2(\theta))$ and

$$\tilde{g}(\mu_1(\theta), \mu_2(\theta), \mu_{1,\theta}(\theta), \mu_{2,\theta}(\theta)) = \frac{\mu_2(\theta)\mu_{1,\theta}(\theta) - \mu_1(\theta)\mu_{2,\theta}(\theta)}{\mu_2^2(\theta)}.$$

The components of the gradient of $\tilde{g}$ are then

$$\begin{aligned} \tilde{g}_1 &= -\mu_{2,\theta}(\theta)/\mu_2^2(\theta), \\ \tilde{g}_2 &= \mu_{1,\theta}(\theta)/\mu_2^2(\theta), \\ \tilde{g}_3 &= 1/\mu_2^2(\theta), \\ \tilde{g}_4 &= -\mu_1(\theta)/\mu_2^2(\theta), \end{aligned}$$

and the second order derivatives $\tilde{g}_{ik}$ are easily obtained by differentiating these expressions. The matrix Γ is highly problem-dependent. □

5.2. Steady-state models.

EXAMPLE 5.3. Suppose we want to minimize the infinite-horizon time-average of a real-valued stochastic process Z, whose evolution depends on θ:

$$\alpha(\theta) \stackrel{\text{def}}{=} \lim_{t\to\infty} \frac{1}{t} E \int_0^t Z(\theta, s)ds.$$

At step n of the SA algorithm, we intent to estimate the gradient $\alpha_\theta(\theta_n)$ by its truncated version $\alpha_\theta(\theta_n, t_n)$, defined as the gradient of

$$\alpha(\theta_n, t_n) \stackrel{\text{def}}{=} \frac{1}{t_n} E \int_0^{t_n} Z(\theta, s)ds.$$

For (fixed) finite t_n, the latter can be estimated by any of the methods discussed in Example 5.1. However, the bias and variance will also depend on the truncation point t_n. Let $\{t_n, n \ge 1\}$ be a deterministic sequence such that $n^{-q}t_n \to \kappa_t > 0$ as $n \to \infty$, for some constants q and κ_t. We shall assume the following.

(A13) As $t \to \infty$, the bias component due to the truncated horizon satisfies $t(\alpha_\theta(\theta, t) - \alpha_\theta(\theta)) \to \kappa_1(\theta)$, where $\sup_\theta |\kappa_1(\theta)| < \infty$, and the variance of the estimator of $\alpha(\theta, t)$ (for the FD methods) or of $\alpha_\theta(\theta, t)$ (for IPA) decreases as $O(t^{-1})$, uniformly in θ.

TABLE 5.2
Convergence rates for the truncated-horizon steady-state model ($\gamma = 1$, $p \geq 0$)

model	q	ν	β	δ	κ	η
FDf	$\geq (p+1)/3$	$(p+q+1)/4$	ν	$2\nu-1$	2ν	$1/2$
FDc	$\geq (p+1)/2$	$(p+q+1)/6$	2ν	$4\nu-1$	4ν	$2/3$
FDf (CRN1)	$\geq (p+1)/2$	$(p+q+1)/3$	ν	$2\nu-1$	2ν	$2/3$
FDc (CRN1)	$\geq 2(p+1)/3$	$(p+q+1)/5$	2ν	$4\nu-1$	4ν	$4/5$
FDf (CRN2)	$\geq p+1$	$\geq (p+q+1)/2$	$\nu \wedge q$	$p+q$	$p+q+1$	1
FDc (CRN2)	$\geq p+1$	$\geq (p+q+1)/4$	$2\nu \wedge q$	$p+q$	$p+q+1$	1
IPA	$\geq p+1$		q	$p+q$	$p+q+1$	1
LR	$(p+1)/3$		q	$p-q$	$2q$	$1/2$
CLR	$(p+1)/2$		q	p	$2q$	$2/3$

Assumption (A13) typically holds and can be proved rigorously for a large class of regenerative models (L'Ecuyer and Glynn [30]). L'Ecuyer [26] studies the convergence rates of several gradient estimators in that context; here, we are rather interested in the convergence rate of SA. Notice that even when the model is regenerative, it is often much more convenient to use truncated-horizon estimators instead of estimators directly based on the regenerative cycles, because for the former, there is no need to recognize the regeneration epochs during the simulation. These epochs may be rare, or their identification may require non-negligible effort, especially if the system is modeled as a Harris-recurrent Markov chain (see, e.g., Sigman and Wolff [39] or Meyn and Tweedie [31]).

Table 5.2 tells us how the values of p, q, and ν can be chosen in order to maximize η, for the different types of gradient estimators. See Appendix 2 for the technical details. A typical choice would be to take a single replication at each SA iteration ($p = 0$), use an horizon length $t_n = \lfloor \kappa_t n^q \rfloor$ at iteration n, where q is defined by its lower bound in the table. For example, for FDc under CRN1, take $q = 2/3$ and $\nu = 1/3$, so $t_n = \Theta(n^{2/3})$, $c_n = \Theta(n^{-1/3})$, $\|B_n\| = O(n^{-2/3})$, $V_n = O(n^{1/3})$, and $E[\|\theta_n\|^2] = O(n^{-4/3}) = O(C_n^{-4/5})$. The constraints given in Table 5.2 only make sure that η is maximized; a more refined analysis would be required to also minimize the asymptotic MSE constant K_{MSE} defined by (A.19) in Appendix 2. Again, we see that $E\|\theta_n\|^2$ may converge faster than $O(n^{-1})$ as a function of n (for large values of $p+q$), but not faster than $\Theta(C_n^{-1})$ as a function of the total budget. The optimal values of p and ν here are not the same as for the finite horizon case (Table 5.1); however the same η is reached. For example, for FDc with independent random numbers, $p = 0$ and $\nu = 1/6$ gives the optimal rate in the finite-horizon case, but not in the infinite-horizon case.

In the infinite-horizon case, the likelihood ratio (LR) estimator typically behaves much differently than IPA, since its variance typically increases linearly w.r.t. the horizon length t_n (Glynn [12], Rubinstein and Shapiro [35]). One must then select a small value of q to control the variance, and as a result, the overall convergence turns out to be quite slow ($\eta = 1/2$). There is also a *control-variate* variant of LR (L'Ecuyer [26]) for which (under some conditions) the variance V_n is $\Theta(1)$. It is denoted by CLR in the table. □

EXAMPLE 5.4. Let us return to the $GI/GI/1$ queue of Example 1.1. For that problem, L'Ecuyer and Glynn [14] have proven the convergence w.p.1 of SA (using projection over a finite interval) combined with each of the gradient estimation methods discussed so far in this section, under a given set of assumptions. They did not study the convergence rates. By combining their methods of proofs with the results of L'Ecuyer and Glynn [30], it is easy to show that under their assumptions (A–C), together with additional assumptions on $C(\theta)$ to ensure (A3)–(A5), the exact orders found in the previous example for the truncated-horizon case apply to this model (for the common random numbers, the variant that applies is CRN2). In particular, Table 5.2 gives the appropriate convergence rates for several of the variants which were experimented by L'Ecuyer, Giroux, and Glynn [28] for the $M/M/1$ queue, and therefore explains much of the numerical results obtained by these authors.

Since the single queue considered in this example is a regenerative system, the function $\alpha(\theta)$ can also be written as a ratio of two expectations, and the methodology of Example 5.2 can be applied with g defined as $g(\mu_1, \mu_2) = \mu_1/\mu_1$. This gives $\eta = 1$ provided that $p \geq 1$. □

EXAMPLE 5.5. Haurie, L'Ecuyer, and Van Delft [17] proved the convergence w.p.1 of SA combined with IPA for a class of piecewise-deterministic control systems encountered in manufacturing. They did not find the convergence rates. However, under their assumptions 2.1 and 2.2, and using again the results of L'Ecuyer and Glynn [30], one can show that $\|B_n\|$ and V_n are both $\Theta(t_n^{-1})$ for the IPA estimator, so the results of Table 5.2 for IPA apply to this model, provided that (A3)–(A5) also hold for the performance measure considered. □

5.3. An infinite-horizon model with discounting.

EXAMPLE 5.6. We now return to the infinite-horizon discounting model of Example 1.3, assuming that $Z(\cdot)$ is strictly positive and bounded. Consider the truncated-horizon cost estimator

$$X(\theta, t) = \int_0^t e^{-\rho s} Z(s) ds$$

and let $\alpha(\theta, t) = E[X(\theta, t)]$.

(A14) Suppose that Var $[X(\theta, t)]$ and $e^{\rho t}(\alpha(\theta) - \alpha(\theta, t))$ are both bounded away from 0 and from infinity, uniformly in θ. Let Var $[X(\theta, t)] \to \sigma^2$ and $e^{\rho t}(\alpha(\theta) - \alpha(\theta, t)) \to \kappa_\beta$ when $\theta \to 0$ and $t \to \infty$, where σ and κ_β are positive constants.

This assumption is reasonable; see Fox and Glynn [8] for justifications. Suppose also that the time required for computing $X(\theta, t)$ is approximately $\kappa_T t$, for some constant κ_T. Fox and Glynn [8] have shown that to optimize the convergence rate of the MSE when i.i.d. replicates of $X(\theta, t)$ are used to estimate $\alpha(\theta)$ under the constraint of a limited budget of size C, the horizon length t should increase in such a way that $t/\ln C \to 1/(2\rho)$ as $C \to \infty$. The budget C is then split into approximately $2\rho C/(\kappa_T \ln C)$ simulation runs of length $t = \kappa_T \ln C/(2\rho)$.

We show in Appendix 2 that in the SA context, if an horizon length t_n is used at iteration n, then $t_n/\ln n$ must converge to a positive constant in order to optimize the convergence rate of the MSE of θ_n, and the convergence rate actually depends on the product of that constant with the discount rate ρ. We consider IPA and the FD variants introduced in Example 5.1. In the case of IPA, we assume that

$$X_\theta(\theta, t) = \int_0^t e^{-\rho s} \tilde{Z}(s) ds$$

TABLE 5.3
Asymptotic behavior of SA: The infinite-horizon model with discounting

method	$\rho\kappa_t$	ν	β	δ	κ	$\tilde{\eta}$
FDf	$\geq (p+1)/2$	$(p+1)/4$	ν	$p-\nu$	$(p+1)/2$	$1/2$
FDc	$\geq (p+1)/3$	$(p+1)/6$	2ν	$p-2\nu$	$2(p+1)/3$	$2/3$
FDf (CRN1)	$\geq 2(p+1)/3$	$(p+1)/3$	ν	$p-\nu$	$2(p+1)/3$	$2/3$
FDc (CRN1)	$\geq 2(p+1)/5$	$(p+1)/5$	2ν	$p-2\nu$	$4(p+1)/5$	$4/5$
FDf (CRN2)	$\geq p+1$	$\geq (p+1)/2$	$\rho\kappa_t \wedge \nu$	p	$p+1$	1
FDc (CRN2)	$\geq (p+1)/2$	$\geq (p+1)/4$	$\rho\kappa_t \wedge 2\nu$	p	$p+1$	1
IPA	$\geq (p+1)/2$		$\rho\kappa_t$	p	$p+1$	1

is an unbiased estimator of $\alpha_\theta(\theta, t)$, and that the assumptions (A14) hold with X and α replaced by X_θ and α_θ. For the six variants of FD, $X(\theta, t)$ replaces $X(\theta)$ in the setups of Example 5.1. At step n of SA, we compute N_n i.i.d. replications of the gradient estimator ψ over an horizon-length t_n (e.g., $\psi = X_\theta(\theta_n, t_n)$ in the case of IPA, ψ given by (5.1) for FDc, etc.). The average of those N_n replications is the gradient estimator ψ_n. We suppose that $n^p N_n \to \kappa_N$ and $t_n/\ln n \to \kappa_t$ as $n \to \infty$, where $p \geq 0$, $\kappa_N > 0$, and $\kappa_t \in [0, \infty]$ are fixed constants (Here, $\kappa_t = \infty$ is a loose notation to indicate the case where t_n increases faster than $\Theta(\ln n)$.)

Table 5.3 gives the values of κ and $\tilde{\eta}$ such that $E\|\theta_n\|^2 = \Theta(n^{-\kappa}) = \Theta(((E[C_n])^{-1} \ln E[C_n])^{\tilde{\eta}})$ for the largest possible $\tilde{\eta}$, for the different cases. Note that $\tilde{\eta}$ has a slightly different meaning than η in the previous examples. Here, $(E[C_n])^{-1} \ln E[C_n] = \Theta(n^{-p-1})$, so $\tilde{\eta} = \kappa/(p+1)$. The optimal rates are attained when the constant κ_t is finite and large enough, as indicated in the second column of Table 5.3. So, the horizon length t_n must increase at a logarithmic rate, and the factor of increase must be large enough. The optimal strategy is to take $\rho\kappa_t$ equal to (or slightly larger than) the lower bound given in the table (e.g., $\kappa_t \geq (p+1)/(2\rho)$ for IPA). The values of κ and $\tilde{\eta}$ here are the same as κ and η in Example 5.1; However the convergence in terms of C_n is at a slightly slower rate, due to the additional logarithmic factor. □

6. Further remarks and conclusion. In this section, we briefly discuss other possible variants of our setup, such as SA with averaging and projection methods, then give a conclusion.

6.1. Algorithms with averaging. Very recently, some new methods were proposed and suggested for stochastic approximation, by Polyak [33], Ruppert [37], and Bather [3]. See also Yin [42], Polyak and Juditsky [34] and Kushner and Yang [23]. It has been a long time effort (dated back to Chung [5]) to improve the performance of SA algorithms. Among the choices of step size $a_n = 1/n^\gamma$ for $1/2 < \gamma < 1$, $a_n = 1/n$ gives the highest order of convergence. However, if the step size is too small in the initial stage, the iterates wonder around and take a long time to settle down. From this point of view, one may wish to choose a large step size instead. But large step size gives rise to lower rate of convergence. The averaging procedures provide a good alternative in taking care of these trade-offs. The procedures are multistep iterative schemes. With a large initial step size, they produce a "squeezing effect" forcing the iterates to reach a vicinity of $\theta^* = 0$ faster. Meanwhile, by means of averaging, they keep the convergence at the optimal rate with the smallest possible asymptotic covariance.

Inspired by the averaging approach suggested by Polyak [33] and Ruppert [37], for our problem we may consider the algorithm

$$\theta_{n+1} = \theta_n - n^{-\gamma}\psi_n; \qquad \bar{\theta}_n = \frac{1}{n}\sum_{j=1}^{n}\theta_j. \tag{6.1}$$

Notice that the averaging here creates no additional burden since the average can be recursively updated as $\bar{\theta}_{n+1} = \bar{\theta}_n + (\theta_{n+1} - \bar{\theta}_n)/(n+1)$.

Motivated by the work of Bather [3] (see also Schwabe [38] and Yin and Yin [44]), we may consider another algorithm, which uses averaging in both trajectories and observations (measurements). In addition to the advantages mentioned above, that algorithm appears to be more stable in the initial period (see [44]). Consider

$$\theta_{n+1} = \bar{\theta}_n - n^{-\gamma}\sum_{i=1}^{n}\psi_i; \qquad \bar{\theta}_n = \frac{1}{n}\sum_{j=1}^{n}\theta_j, \tag{6.2}$$

where $1/2 < \gamma < 1$. The study of asymptotic properties of the averaging algorithms in conjunction with the setup of this paper could be carried out. The idea would be to combine the approach in Yin [42] and Yin and Yin [44] with the results of this work. This is a topic for further investigation.

6.2. Projection algorithms. The discussion in this paper is based on the basic recursive SA algorithm (1.1). Many variants of the algorithm can also be considered. For instance, our results can easily be incorporated into algorithms with projection and truncations. Such algorithms have the advantage that growth conditions on the underlying functions are no longer required.

Suppose G is a closed and convex set (for example, a closed ball or a closed hyper-rectangle in $\mathbb{R}^r$), and suppose that all the desirable equilibrium points of $\alpha(\theta)$ are interior to G. If $x \notin G$, define $\pi_G(x)$ to be the nearest point to x in G. Then the algorithm (1.1) can be replaced by

$$\theta_{n+1} = \pi_G(\theta_n - n^{-\gamma}\psi_n). \tag{6.3}$$

The limiting ordinary differential equation then reads:

$$\dot{\theta} = \pi_G(\alpha_\theta(\theta)).$$

The results of this paper still hold, and the discussions are the same as before. A variant of such a projection procedure is an algorithm with random truncation bounds or random projection regions, along the lines of, for example, Chen and Zhu [4].

Let $q_i(\cdot)$, $i \le \nu$, be continuously differentiable functions and

$$G = \{\theta : \ q_i(\theta) \le 0 \ \text{ for } i = 1, 2, \ldots, \nu\}.$$

Suppose that G is bounded, convex, and is the closure of its interior. If we use the projection algorithm in this case, the limit of $\{\theta_n\}$ will be related to the set of Fritz-John points or Kuhn-Tucker points (see p. 191 of Kushner and Clark [20] for more details; see also and Kushner and Yin [24] for a comprehensive and updated development of constrained and unconstrained algorithms). Again, our convergence rate results apply if θ^* lies in the interior of G, but not if it is on the boundary. For more general constrained optimization problems, other methods, such as large deviations techniques (see the discussion in Chapter 10 of [24] and the references therein) may be needed to obtain the rate of convergence.

6.3. Conclusion. We have studied budget- and moment-dependent stochastic optimization algorithms, and ascertained their rates of convergence for different situations arising in discrete-event simulation. The corresponding asymptotic distributions were also derived together with some far reaching functional invariance theorems. Several examples showed how those results can be used to find the "optimal" parameters of the algorithm in various SA settings. Such knowledge is important for understanding how SA algorithms behave from the theoretical point of view. As for most optimization algorithms, it is true that the practitioner rarely has all the information available (e.g., the Hessian matrix) to actually implement the algorithm with the optimal parameters and scaling. But some "rough" information could be used when available and adaptive schemes, which estimate the optimal parameters as the algorithm proceeds, may also work in some situations. Recent results suggest that SA methods with averaging are less sensitive to the choice of scaling and algorithm parameters. Studying those averaging methods within the setup of this paper is a topic of ongoing research.

Appendix 1: Proofs of the theorems. This appendix provides the proofs of Theorems 3.1, 4.1, 4.3, and 4.4.

Proof of Theorem 3.1. Define $U(\theta) = (1/2)\|\theta\|^2$. We first treat the case $\gamma < 1$. The proof is divided into two steps. In Step 1, we show that $\sup_n EU(\theta_n) < \infty$. The w.p.1 convergence and the estimate above allow us to infer $EU(\theta_n) \to 0$. Then in Step 2, we refine the order of magnitude estimate.

Step 1. Since θ_n is measurable with respect to the σ-algebra that measures (n, s_n, θ_n), one has

$$E_n[n^{-\gamma}\theta_n'(\psi_n - E_n\psi_n)] = n^{-\gamma}\theta_n' E_n[\psi_n - E_n\psi_n] = 0. \tag{A.4}$$

Take an arbitrary $0 < \varepsilon_0 \le \varepsilon$ and an n_0 such that $2\beta n_0^{\gamma-1} \le \varepsilon_0$. The reason of this choice will become clear in the sequel. For the remainder of this proof, we suppose that $n \ge n_0$. Observe that

$$\|\theta'(H\theta - \alpha_\theta(\theta))\| \;\le\; \|\theta\|(\|H\| \cdot \|\theta\| + K_\alpha(1 + ||\theta||)) \le K_1(1 + U(\theta)),$$

for all θ and for some $K_1 > 0$. From a Taylor series expansion, using

$$\theta_{n+1} = \theta_n - n^{-\gamma}\alpha_\theta(\theta_n) + n^{-\gamma}(\alpha_\theta(\theta_n) - E_n\psi_n) + n^{-\gamma}(E_n\psi_n - \psi_n), \tag{A.5}$$

Eq. (A.4), the fact that $U_\theta(\theta_n) = \theta_n$, and (A5), we obtain that

$$\begin{aligned}
&EU(\theta_{n+1}) - EU(\theta_n) \\
=\;& E(E_nU(\theta_{n+1}) - U(\theta_n)) \\
=\;& EE_nU_\theta'(\theta_n)(\theta_{n+1} - \theta_n) + (1/2)E\|\theta_{n+1} - \theta_n\|^2 \\
=\;& n^{-\gamma}E\theta_n'[-H\theta_n + (H\theta_n - \alpha_\theta(\theta_n)) + \; (\alpha_\theta(\theta_n) - E_n\psi_n)] \\
&+ (1/2)E\|\theta_{n+1} - \theta_n\|^2 \\
\le\;& -2n^{-\gamma}\lambda_{\min}EU(\theta_n) + K_1n^{-\gamma}(1 + EU(\theta_n)) + n^{-\gamma}E\theta_n'(\alpha_\theta(\theta_n) - E_n\psi_n) \\
&\quad + (3/2)n^{-2\gamma}E\left(\|\alpha_\theta(\theta_n)\|^2 + \|\alpha_\theta(\theta_n) - E_n\psi_n\|^2 + \|\psi_n - E_n\psi_n\|^2\right).
\end{aligned} \tag{A.6}$$

Using (A5), we find

$$E\|\alpha_\theta(\theta_n)\|^2 \le 2K_\alpha^2(1 + 2EU(\theta_n)). \tag{A.7}$$

We also have

$$E\|\theta_n'(\alpha_\theta(\theta_n) - E_n\psi_n)\| \;\le\; E^{1/2}\|\theta_n\|^2 E^{1/2}\|\alpha_\theta(\theta_n) - E_n\psi_n\|^2.$$

Using the inequality $ab \le (a^2+b^2)/2$ with

$$a^2 = (K_\beta/\lambda_{\mathsf{min}})n^{-\beta} \quad \text{and} \quad b^2 = E\|\theta_n\|^2/a^2, \tag{A.8}$$

we arrive at

$$\begin{aligned} E^{1/2}\|\theta_n\|^2 &\le (K_\beta/2\lambda_{\mathsf{min}})n^{-\beta} + (\lambda_{\mathsf{min}}/2K_\beta)n^{\beta}E\|\theta_n\|^2 \\ &= (K_\beta/2\lambda_{\mathsf{min}})n^{-\beta} + (\lambda_{\mathsf{min}}/K_\beta)n^{\beta}EU(\theta_n). \end{aligned}$$

From the inequalities above and (A1), we have

$$E\|\theta_n'(\alpha_\theta(\theta_n) - E_n\psi_n)\| \le (K_\beta^2/2\lambda_{\mathsf{min}})n^{-2\beta} + \lambda_{\mathsf{min}}EU(\theta_n). \tag{A.9}$$

By virtue of (A.6), (A.7), and (A.9), we obtain

$$\begin{aligned} &EU(\theta_{n+1}) - EU(\theta_n) \\ \le\; & -\lambda_{\mathsf{min}}n^{-\gamma}EU(\theta_n) + K_1n^{-\gamma}EU(\theta_n) + K_1n^{-\gamma} + (K_\beta^2/2\lambda_{\mathsf{min}})n^{-\gamma-2\beta} \\ & +3K_\alpha^2n^{-2\gamma}(1+2EU(\theta_n)) + (3/2)K_\beta^2n^{-2\gamma-2\beta} \\ & +(3/2)n^{-2\gamma}E\|\psi_n - E_n\psi_n\|^2, \end{aligned} \tag{A.10}$$

where $E\|\psi_n - E_n\psi_n\|^2 \le K_\delta n^{-\delta}$ in accordance with (A2). Choose λ_0 such that $0 < \lambda_0 < \lambda_{\mathsf{min}}$. We may assume that n_0 has been chosen large enough such that

$$0 < \lambda_0 \le \lambda_{\mathsf{min}} - 6K_\alpha^2 n_0^{-\gamma} \quad \text{and} \quad 3n_0^{-\gamma} \le \varepsilon_0/\lambda_{\mathsf{min}}.$$

Notice that there is a $K_2 > 0$ such that for $n \ge n_0$, $K_1n^{-\gamma} + 3K_\alpha^2n^{-2\gamma} \le K_2n^{-\gamma}$. Using the inequalities above, and iterating on (A.10) yields that for $n \ge n_0$,

$$\begin{aligned} & EU(\theta_{n+1}) \\ \le\; & (1-\lambda_0n^{-\gamma})EU(\theta_n) + K_1n^{-\gamma}EU(\theta_n) \\ & +\frac{K_\beta^2}{\lambda_{\mathsf{min}}}n^{-\gamma-2\beta} + \frac{3K_\delta}{2}n^{-2\gamma-\delta} + K_2n^{-\gamma} \\ =\; & A_{n,n_0-1}EU(\theta_{n_0}) + K_1\sum_{i=n_0}^{n} i^{-\gamma}A_{ni}EU(\theta_i) + \frac{K_\beta^2}{\lambda_{\mathsf{min}}}\sum_{i=n_0}^{n} i^{-\gamma-2\beta}A_{ni} \\ & +\frac{3K_\delta}{2}\sum_{i=n_0}^{n} i^{-2\gamma-\delta}A_{ni} + K_2\sum_{i=n_0}^{n} i^{-\gamma}A_{ni}, \end{aligned} \tag{A.11}$$

where

$$A_{nj} \stackrel{\text{def}}{=} \begin{cases} \prod_{k=j+1}^{n}(1-\lambda_0k^{-\gamma}), & j < n; \\ 1, & j = n. \end{cases}$$

Observe that for each fixed $j \ne n$,

$$\begin{aligned} |A_{nj}| &\le \exp\left(-\lambda_0\sum_{k=j+1}^{n} k^{-\gamma}\right) \\ &\le \exp\left(\lambda_0j^{-\gamma} - \lambda_0\int_j^n x^{-\gamma}dx\right) \\ &= \exp\left(\lambda_0j^{-\gamma} - \frac{\lambda_0(n^{1-\gamma} - j^{1-\gamma})}{1-\gamma}\right) \end{aligned} \tag{A.12}$$

As a consequence, the first term on the r.h.s. of (A.11) is $O(\exp(-\lambda_0 n^{1-\gamma}/(1-\gamma)))$. Consider now the third term in the last part of (A.11). By virtue of a partial summation, one has

$$\sum_{i=n_0}^{n} i^{-\gamma-2\beta} A_{ni} = n^{-2\beta} \sum_{i=n_0}^{n} i^{-\gamma} A_{ni} + \sum_{i=n_0}^{n-1} \left(\frac{1}{i^{2\beta}} - \frac{1}{(i+1)^{2\beta}} \right) \sum_{j=n_0}^{i} j^{-\gamma} A_{nj}. \tag{A.13}$$

Since

$$A_{nj} - A_{n,j-1} = A_{nj}(1 - (1 - \lambda_0 j^{-\gamma})) = \lambda_0 j^{-\gamma} A_{nj},$$

we have

$$\sum_{j=n_0}^{i} j^{-\gamma} A_{nj} = \lambda_0{}^{-1} \sum_{j=n_0}^{i} (A_{nj} - A_{n,j-1}) = \lambda_0{}^{-1}(A_{ni} - A_{n,n_0-1}).$$

Owing to (A.12),

$$\sum_{i=n_0}^{n-1} \left(\frac{1}{i^{2\beta}} - \frac{1}{(i+1)^{2\beta}} \right) A_{n,n_0-1} = O(A_{n,n_0-1}) = O(\exp(-n^{1-\gamma})).$$

Noticing the choice of n_0 and that

$$\frac{1}{n^{2\beta}} - \frac{1}{(n+1)^{2\beta}} = n^{-2\beta} \left(2\beta n^{-1} + O(n^{-2}) \right) = 2\beta n^{-2\beta-1} + O(n^{-2\beta-2}),$$

and using (A.12) again, we have that

$$\begin{aligned} \sum_{i=n_0}^{n-1} \left(\frac{1}{i^{2\beta}} - \frac{1}{(i+1)^{2\beta}} \right) A_{ni} &= \sum_{i=n_0}^{n-1} \left(2\beta i^{-2\beta-1} + O(i^{-2\beta-2}) \right) A_{ni} \\ &\le \sum_{i=n_0}^{n} \varepsilon_0 i^{-2\beta-\gamma} A_{ni} + O(n^{-2\beta-1}). \end{aligned}$$

Combining this with (A.13) and observing that $\exp(-n^{1-\gamma}) = O(n^{-2\beta-1})$, we obtain that

$$\sum_{i=n_0}^{n} i^{-\gamma-2\beta} A_{ni} \le \lambda_0^{-1}(A_{nn} - A_{n,n_0-1}) n^{-2\beta} + \varepsilon_0 \lambda_0^{-1} \sum_{i=n_0}^{n} i^{-\gamma-2\beta} A_{ni} + O(n^{-2\beta-1}),$$

and hence

$$(1 - \varepsilon_0/\lambda_0) \sum_{i=n_0}^{n} i^{-\gamma-2\beta} A_{ni} = \lambda_0^{-1} n^{-2\beta} + O(n^{-2\beta-1}).$$

By virtue of (A2), one has $\gamma + \delta > 1 - \gamma > 0$. Repeating the argument above, we also obtain

$$(1 - \varepsilon_0/\lambda_0) \sum_{i=n_0}^{n} i^{-2\gamma-\delta} A_{ni} = \lambda_0^{-1} n^{-\gamma-\delta} + O(n^{-\gamma-\delta-1}). \tag{A.14}$$

Owing to (A.11)–(A.14), there is a $K_3 > 0$ such that

$$\text{(A.15)} \quad EU(\theta_{n+1}) \le \frac{K_\beta^2}{\lambda_{\min}\lambda_0(1-\varepsilon_0/\lambda_0)} n^{-2\beta} + \frac{3K_\delta}{2\lambda_0(1-\varepsilon_0/\lambda_0)} n^{-\gamma-\delta} + O(n^{-2\beta-1} + n^{-\gamma-\delta-1}) + K_1 \sum_{i=n_0}^{n} i^{-\gamma} A_{ni} EU(\theta_i) + K_3.$$

By virtue of the Gronwall's inequality, there is a $K_4 > 0$ such that

$$EU(\theta_{n+1}) \le K_4 \exp\left(K_1 \sum_{i=n_0}^{n} i^{-\gamma} A_{ni}\right),$$

and hence $\sup_n EU(\theta_n) \le K_5$ for some $K_5 > 0$. This together with the w.p.1 convergence of $\theta_n \to 0$ implies that $EU(\theta_n) \to 0$.

Step 2. In view of (A5) and Step 1, $E[\theta_n' \alpha_\theta(\theta_n)] = E[\theta_n' H \theta_n] + o(EU(\theta_n))$ and $E\|\alpha_\theta(\theta_n)\|^2 \le E[\theta_n' H' H \theta_n] + o(E[U(\theta_n)])$. As in the analysis in Step 1, however, we may assume that n_0 is chosen large enough so that $\|o(EU(\theta_n))\| \le (1/2)\lambda_{\min} EU(\theta_n)$. In (A.8), choose $a^2 = (K_\beta/2\lambda_{\min})n^{-\beta}$. Substitute this into (A.6), and proceed exactly the same as before. However, this time, neither the term $K_1 n^{-\gamma} EU(\theta_n)$ nor $K_1 n^{-\gamma}$ appears. In lieu of (A.15), we obtain

$$EU(\theta_{n+1}) \le \frac{K_\beta^2}{\lambda_{\min}\lambda_0(1-\varepsilon_0/\lambda_0)} n^{-2\beta} + \frac{3K_\delta}{2\lambda_0(1-\varepsilon_0/\lambda_0)} n^{-\gamma-\delta} + O(n^{-2\beta-1} + n^{-\gamma-\delta-1}).$$

Now, observe that it is possible to choose ε_0 and λ_0 such that $1/(\lambda_0 - \varepsilon_0) \le (1+\varepsilon)/\lambda_{\min}$; e.g., take $\lambda_0 = \lambda_{\min}/(1+\varepsilon/2)$ and then $0 < \varepsilon_0 < \lambda_0 - \lambda_{\min}/(1+\varepsilon)$. Since $\|\theta_n\|^2 = 2U(\theta_n)$, and since ε can be taken arbitrarily small in (A5), the first part of the theorem (for $1/2 < \gamma < 1$) follows.

The proof for the case $\gamma = 1$ is similar to that of $1/2 < \gamma < 1$. We only point out the difference below. Notice that (A.11) and (A.13) still hold with $\gamma = 1$. Equation (A.12) can be replaced by

$$\begin{aligned} |A_{nj}| &\le \exp\left(\lambda_0/j - \lambda_0 \int_j^n x^{-1} dx\right) \\ &= \exp\left(\lambda_0/j - \lambda_0 \ln(n/j)\right) \\ &= \exp\left(\lambda_0/j\right)(j/n)^{\lambda_0}. \end{aligned}$$

We thus have

$$\begin{aligned} \sum_{i=n_0}^{n-1} i^{-2\beta-1}|A_{ni}| &\le \exp(\lambda_0/n_0) \sum_{i=n_0}^{n-1} i^{-2\beta-1}(i/n)^{\lambda_0} \\ &\le \exp(\lambda_0/n_0) n^{-\lambda_0} \int_{n_0}^{n} x^{\lambda_0-2\beta-1} dx \\ &\le \exp(\lambda_0/n_0)(\lambda_0 - 2\beta)^{-1} n^{-2\beta}. \end{aligned}$$

Similarly,

$$\begin{aligned} \sum_{i=n_0}^{n-1} i^{-2-\delta}|A_{ni}| &\le \exp(\lambda_0/n_0) \sum_{i=n_0}^{n-1} i^{-2-\delta}(i/n)^{\lambda_0} \\ &\le \exp(\lambda_0/n_0)(\lambda_0 - \delta - 1)^{-1} n^{-\delta-1}. \end{aligned}$$

Since n_0 can be chosen sufficiently large, the proof of the second part then follows. □

Note that in the above, we have assumed that without of loss generality, $\lambda_0 \neq 2\beta$ and $\lambda_0 \neq 1+\delta$. If, for example, $\lambda_0 = 2\beta$, then for some $\Delta > 0$, $\tilde{\lambda}_0 \stackrel{\text{def}}{=} \lambda_0 - \tilde{\Delta} < \lambda_0$ such that $(i/n)^{\lambda_0} \leq (i/n)^{\tilde{\lambda}_0}$. We can then use the same calculation above with λ_0 replaced by $\tilde{\lambda}_0$.

Proof of Theorem 4.1. The proof is divided into three steps. In the first step, we establish an asymptotic equivalence, while in the second and third steps, we obtain the weak convergence by virtue of a result of Ethier and Kurtz [6], and compute the asymptotic covariance matrix. We first prove a lemma which gives an asymptotic equivalence.

LEMMA A.1. *Under the conditions of Theorem 4.1,*

$$W_n(t) = M_n(t) + o(1), \quad \textit{for all } t \in [0,1]$$

where

$$M_n(t) \quad = \quad \frac{\lfloor nt \rfloor^{1+\delta/2}}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} i^{-1} D_{\lfloor nt \rfloor i}(E_i\psi_i - \psi_i),$$

D_{ni} is defined by

$$D_{ni} = \begin{cases} \prod_{j=i+1}^{n}(I - H/j), & n > i; \\ I, & n = i, \end{cases} \tag{A.16}$$

and $o(1) \xrightarrow{\text{p}} 0$ as $n \to \infty$, uniformly in t.

Proof of Lemma A.1. Direct computation leads to

$$\begin{aligned} \theta_{\lfloor nt \rfloor+1} &= \left(I - H\lfloor nt \rfloor^{-1}\right)\theta_{\lfloor nt \rfloor} + \lfloor nt \rfloor^{-1} h(\theta_{\lfloor nt \rfloor}) \\ &\quad + \lfloor nt \rfloor^{-1}(\alpha_\theta(\theta_{\lfloor nt \rfloor}) - E_{\lfloor} nt \rfloor \psi_{\lfloor} nt \rfloor) + \lfloor nt \rfloor^{-1}(E_{\lfloor} nt \rfloor \psi_{\lfloor} nt \rfloor - \psi_{\lfloor} nt \rfloor) \\ &= D_{\lfloor nt \rfloor, n_0-1}\theta_{n_0} + \sum_{i=n_0}^{\lfloor nt \rfloor} i^{-1} D_{\lfloor nt \rfloor, i} h(\theta_i) \\ &\quad + \sum_{i=n_0}^{\lfloor nt \rfloor} i^{-1} D_{\lfloor nt \rfloor, i}(\alpha_\theta(\theta_i) - E_i\psi_i) + \sum_{i=n_0}^{\lfloor nt \rfloor} i^{-1} D_{\lfloor nt \rfloor i}\left(E_i\psi_i - \psi_i\right), \end{aligned}$$

where n_0 is the same as in the proof of Theorem 3.1 and $h(\theta) = H\theta - \alpha_\theta(\theta)$. It then yields

$$\begin{aligned} \frac{\lfloor nt \rfloor^{1+\delta/2}}{\sqrt{n}}\theta_{\lfloor nt \rfloor+1} &= \frac{\lfloor nt \rfloor^{1+\delta/2}}{\sqrt{n}} D_{\lfloor nt \rfloor, n_0-1}\theta_{n_0} + \frac{\lfloor nt \rfloor^{1+\delta/2}}{\sqrt{n}} \sum_{i=n_0}^{\lfloor nt \rfloor} i^{-1} D_{\lfloor nt \rfloor, i} h(\theta_i) \\ &\quad + \frac{\lfloor nt \rfloor^{1+\delta/2}}{\sqrt{n}} \sum_{i=n_0}^{\lfloor nt \rfloor} i^{-1} D_{\lfloor nt \rfloor, i}(\alpha_\theta(\theta_i) - E_i\psi_i) \\ &\quad + \frac{\lfloor nt \rfloor^{1+\delta/2}}{\sqrt{n}} \sum_{i=n_0}^{\lfloor nt \rfloor} i^{-1} D_{\lfloor nt \rfloor i}\left(E_i\psi_i - \psi_i\right). \end{aligned} \tag{A.17}$$

By virtue of the structure of $D_{k,j}$, it is easily seen that

$$\frac{\lfloor nt\rfloor^{1+\delta/2}}{\sqrt{n}} D_{\lfloor nt\rfloor, n_0-1}\theta_{n_0} \stackrel{\text{w.p.1}}{\longrightarrow} 0 \quad \text{as } n\to\infty.$$

Next, we obtain

$$\begin{aligned} E\left\|\frac{\lfloor nt\rfloor^{1+\delta/2}}{\sqrt{n}} \sum_{i=n_0}^{\lfloor nt\rfloor} i^{-1} D_{\lfloor nt\rfloor,i} h(\theta_i)\right\| &\le K\frac{\lfloor nt\rfloor^{1+\delta/2}}{\sqrt{n}} \sum_{i=n_0}^{\lfloor nt\rfloor} i^{-1}\|D_{\lfloor nt\rfloor,i}\|E\|\theta_i\|^2 \\ &\le K\frac{1}{\sqrt{n}}\frac{1}{\lfloor nt\rfloor^{1+\delta/2}}\sum_{i=n_0}^{\lfloor nt\rfloor}\left(\frac{i}{\lfloor nt\rfloor}\right)^{\lambda_{\min}-2-\delta} \\ &\le Kn^{-\frac{1+\delta}{2}}\int_0^1 u^{\lambda_{\min}-2-\delta}du \quad \stackrel{n\to\infty}{\longrightarrow} \quad 0 \end{aligned}$$

uniformly in t. Similarly,

$$E\left\|\frac{\lfloor nt\rfloor^{1+\delta/2}}{\sqrt{n}} \sum_{i=n_0}^{\lfloor nt\rfloor} i^{-1} D_{\lfloor nt\rfloor,i} (\alpha_\theta(\theta_i) - E_i\psi_i)\right\| \stackrel{n\to\infty}{\longrightarrow} 0 \text{ uniformly in } t.$$

In addition,

$$\frac{\lfloor nt\rfloor^{1+\delta/2}}{\sqrt{n}} \sum_{i=1}^{n_0-1} i^{-1} D_{\lfloor nt\rfloor i}(E_i\psi_i - \psi_i) \stackrel{\text{p}}{\longrightarrow} 0 \quad \text{as } n\to\infty, \text{ uniformly in } t.$$

Thus the lemma follows. □

First, following from Theorem 3.1, $M_n(\cdot)$ is a square integrable martingale. Similar to Yin [41], it can be shown that

$$E\left(\sup_{t\in[0,1]}\|M_n(t) - M_n(t^-)\|^2\right) \stackrel{n\to\infty}{\longrightarrow} 0,$$
$$E\left(\sup_{t\in[0,1]}\|\tilde{A}_n^{ij}(t) - \tilde{A}_n^{ij}(t^-)\|\right) \stackrel{n\to\infty}{\longrightarrow} 0,$$

where $\tilde{A}_n(\cdot)$ (with $\tilde{A}_n^{ij}(\cdot)$ denoting its ijth entry) is defined by

$$\tilde{A}_n^{ij}(t) = \frac{\lfloor nt\rfloor^{2+\delta}}{n}\sum_{k=1}^{\lfloor nt\rfloor} k^{-2}E_k\left(D_{\lfloor nt\rfloor,k}(E_k\psi_k - \psi_k)\right)^i \left(D_{\lfloor nt\rfloor,k}(E_k\psi_k - \psi_k)\right)^j,$$

with the superscripts i and j (in the summand above) denoting the ith and jth components of the corresponding vectors, respectively. Furthermore, for each $i,j = 1,2,\cdots,r$, $M_n^i(t)M_n^j(t) - \tilde{A}_n^{ij}(t)$ is a martingale, where $M_n^i(\cdot)$ denotes the ith component of $M_n(\cdot)$. To complete the proof of Theorem 4.1, we can thus apply Theorem 7.1.4 of Ethier and Kurtz [6].

Notice that $\tilde{A}_n(\cdot)$ is non-negative definite. By virtue of Condition (A6) and the square integrability of $M_n(\cdot)$,

$$\tilde{A}_n(t) = A_n(t) + o(1),$$

where

$$A_n(t) = \frac{\lfloor nt \rfloor^{2+\delta}}{n} \sum_{i=1}^{\lfloor nt \rfloor} i^{-2-\delta} D_{\lfloor nt \rfloor, i} R D'_{\lfloor nt \rfloor, i}$$

and $o(1) \xrightarrow{\text{P}} 0$ as $n \to \infty$, uniformly in t. In addition, $A_n(\cdot)$ is also non-negative definite.

It is now clear that to derive the limit covariance matrix, we need only look at $A_n^{ij}(\cdot)$. For a fixed $t \in [0,1]$,

$$\begin{aligned} A_n(t) &= \frac{\lfloor nt \rfloor^{2+\delta}}{n} \sum_{i=1}^{\lfloor nt \rfloor} i^{-2-\delta} \exp\left(H \ln(i/\lfloor nt \rfloor)\right) R \exp\left(H' \ln(i/\lfloor nt \rfloor)\right) \\ &\quad + \frac{\lfloor nt \rfloor^{2+\delta}}{n} \sum_{i=1}^{\lfloor nt \rfloor} i^{-2-\delta} \Big(D_{\lfloor nt \rfloor, i} R D'_{\lfloor nt \rfloor, i} \\ &\qquad - \exp\left(H \ln(i/\lfloor nt \rfloor)\right) R \exp\left(H' \ln(i/\lfloor nt \rfloor)\right) \Big). \end{aligned}$$

The last term on the right-hand side above tends to 0 in probability as $n \to \infty$ (uniformly in t). As for the first term, we have

$$\begin{aligned} &\frac{\lfloor nt \rfloor^{2+\delta}}{n} \sum_{i=1}^{\lfloor nt \rfloor} i^{-2-\delta} \exp\left(H \ln(i/\lfloor nt \rfloor)\right) R \exp\left(H' \ln(i/\lfloor nt \rfloor)\right) \\ \xrightarrow{n\to\infty}\ & t \int_0^1 u^{-2-\delta} \exp(H \ln u) R \exp(H' \ln u) du \\ =\ & t \int_0^\infty \exp((1+\delta)v) \exp(-Hv) R \exp(-H'v) dv \qquad \text{(with } v = -\ln u\text{)} \\ =\ & t \int_0^\infty \exp(-\tilde{H}v) R \exp(-\tilde{H}'v) dv. \end{aligned}$$

Up to now, all the conditions in Theorem 7.1.4 of Ethier and Kurtz [6] are satisfied. The desired result then follows from that theorem. □

Proof of Theorem 4.3. Similar to (A.17),

$$\begin{aligned} n^\beta \theta_{n+1} &= n^\beta D_{n,n_0-1} \theta_{n_0} + n^\beta \sum_{i=n_0}^n i^{-1} D_{ni} h(\theta_i) \\ &\quad + n^\beta \sum_{i=n_0}^n i^{-1} D_{ni} (\alpha_\theta(\theta_i) - E_i \psi_i) \\ &\quad + n^\beta \sum_{i=n_0}^n i^{-1} D_{ni} (E_i \psi_i - \psi_i). \end{aligned}$$

As in the derivation of Lemma A.1, upon using the condition $2\beta < 1 + \delta$ and the estimate $E\|\theta_n\|^2 = O(n^{2\beta})$, we have

$$n^\beta \theta_{n+1} = n^\beta \sum_{i=1}^n i^{-1} D_{ni} (\alpha_\theta(\theta_i) - E_i \psi_i) + o(1) \tag{A.18}$$

where $o(1) \xrightarrow{\text{p}} 0$ as $n \to \infty$.

Concentrating on the first term on the right-hand side above, we obtain

$$\begin{aligned}
& n^{\beta} \sum_{i=1}^{n} i^{-1} D_{ni}(\alpha_\theta(\theta_i) - E_i\psi_i) \\
= \quad & -n^{\beta} \sum_{i=1}^{n} i^{-1} D_{ni} B_i \\
= \quad & n^{-1} \sum_{i=1}^{n} (i/n)^{-1-\beta} D_{ni} \bar{B} - n^{-1} \sum_{i=1}^{n} (i/n)^{-1-\beta} D_{ni}[i^{\beta} B_i + \bar{B}] \\
\xrightarrow{\text{p}} \quad & \int_0^1 u^{-1-\beta} \exp(H \ln u) \bar{B} du \quad \text{as } n \to \infty \\
= \quad & \int_0^{\infty} \exp(-H_b v) \bar{B} dv \\
= \quad & H_b^{-1} \bar{B}.
\end{aligned}$$

This yields the desired result. □

Proof of Theorem 4.4. Owing to the fact that $\delta = 2\beta - 1$, we have

$$\begin{aligned}
n^{(1+\delta)/2}\theta_{n+1} \quad = \quad & n^{(1+\delta)/2} \sum_{i=1}^{n} i^{-1} D_{ni}(E_i\psi_i - \psi_i) \\
& + n^{(1+\delta)/2} \sum_{i=1}^{n} i^{-1} D_{ni}(\alpha_\theta(\theta_i) - E_i\psi_i) + o(1),
\end{aligned}$$

where $o(1) \xrightarrow{\text{p}} 0$ as $n \to \infty$. Subtracting $H_b^{-1}\bar{B}$ from both sides of the equation above, noticing that $\beta = (1+\delta)/2$ and applying Theorem 4.3, we establish that

$$n^{(1+\delta)/2}\theta_{n+1} - H_b^{-1}\bar{B} \quad = \quad n^{(1+\delta)/2} \sum_{i=1}^{n} i^{-1} D_{ni}(E_i\psi_i - \psi_i) + o(1),$$

where $o(1) \xrightarrow{\text{p}} 0$ as $n \to \infty$. This, together with Theorem 4.1, completes the proof. □

Appendix 2: Details on convergence for specific setups. This appendix provides some technical details on the convergence of SA for the specific setups studied in Section 5.

Finite-horizon models (Example 5.1). Let the assumptions made in Example 5.1 hold and assume that Var $[X] \to \sigma^2$ as $\theta \to 0$. In the case of FD with independent random numbers, we have the following (see L'Ecuyer and Perron [27] or Zazanis and Suri [46]): For FDf, if α is twice continuously differentiable at $\theta^* = 0$ with second derivative $H = \lambda_{\min} > 0$, then $n^{\nu} B_n \to H\kappa_c/2$ and $n^{p-2\nu} V_n \to 2\kappa_N^{-1}\kappa_c^{-2}\sigma^2$. For FDc, if α is three times continuously differentiable at 0 and H_3 denotes the third derivative, then $n^{2\nu} B_n \to H_3\kappa_c^2/6$ and $n^{p-2\nu} V_n \to \kappa_N^{-1}\kappa_c^{-2}\sigma^2/2$. Assumptions (A6) and (A7) then hold for $\delta = p - 2\nu$, $\beta = \nu$ for FDf, and $\beta = 2\nu$ for FDc. We assume that (3.2) holds and that θ is restricted (e.g., by projection if necessary) to a convex region of G where the second and the third derivatives of α and the variance of X are bounded uniformly in θ. Then, (A1) and (A2) also hold, and we have all the ingredients to apply Theorem 3.1 as well as the results of Section 4.

According to Corollary 4.5, if H is "large enough" to satisfy (3.2), the optimal values of ν and p must satisfy $1+\delta = 2\beta$; i.e., $\nu = (1+p)/4$ for FDf and $\nu = (1+p)/6$ for FDc. This yields

$$C_n^{\eta/2}\theta_n \xrightarrow{\text{d}} \kappa_b^{\eta/2} N(-\mu, \Sigma),$$

where $\eta = 1/2$, $\mu = H\kappa_c/(2(H-\nu))$, and $\Sigma = \sigma^2/(\kappa_N\kappa_c^2(H-\nu))$ for FDf, while $\eta = 2/3$, $\mu = H_3\kappa_c^2/(6(H-2\nu))$, and $\Sigma = \sigma^2/(4\kappa_N\kappa_c^2(H-2\nu))$ for FDc. The asymptotic rate η does not depend on p, but the asymptotic MSE constant

$$K_{\text{MSE}} = \kappa_b^\eta(\mu^2+\Sigma) \overset{\text{w.p.1}}{=} \lim_{n\to\infty} C_n^\eta E[\|\theta_n\|^2] \tag{A.19}$$

does. Note that $p=0$ is not necessarily the optimal choice. If we restrict ourselves to $p=0$, the constant K_{MSE} is minimized for FDf by choosing κ_N and κ_c such that $\kappa_N\kappa_c^4 = 4(H-1/4)\sigma^2 H^{-2}$, which yields $K_{\text{MSE}} = \sigma\kappa_T^{1/2}(H-1/4)^{-3/2}$. For FDc, again for $p=0$, the constant K_{MSE} is minimized by choosing $\kappa_N\kappa_c^6 = (9/2)(H-1/3)\sigma^2 H_3^{-2}$, which yields $K_{\text{MSE}} = (\sigma^4 H_3^2(H-1/3)^{-4}\kappa_T^2/9)^{1/3}(2^{1/3}+2^{-2/3})/4$.

In the case where common random numbers are used and the CRN2 conditions hold, let $\sigma_P^2 = \text{Var}\,[X_\theta(0)]$. Then, for both (5.1) and (5.2), Var $[\psi]$ converges to σ_P^2 as $c\to 0$ and $\theta\to 0$. This gives $n^p V_n \to \kappa_N^{-1}\sigma_P^2$ and $\Sigma = \sigma_P^2/(2\kappa_N(H-\nu))$. The bias satisfies $n^\nu B_n \to H\kappa_c/2$ for FDf, and $n^{2\nu}B_n \to H_3\kappa_c^2/6$ for FDc. Under the conditions given in Table 5.1, we obtain $\eta = 1$. More precisely, under those conditions,

$$C_n^{1/2}\theta_n \xrightarrow{\text{d}} \kappa_b^{1/2} N(-\mu, \Sigma),$$

where $\kappa_b = \kappa_T\kappa_N/(p+1)$, $\Sigma = \sigma_P^2/(2\kappa_N(H-(1+p)/2))$, $\mu = I[2\nu = p+1]\cdot H\kappa_c/(2(H-\nu))$ for FDf, $\mu = I[4\nu = p+1]\cdot H_3\kappa_c^2/(6(H-2\nu))$ for FDc, and $I[A]$ denotes the indicator function which takes the value 1 when the underlying variable belongs to A and is 0 otherwise. It is best to take ν such that $2\beta > p+1$, because it cancels out the bias term μ. Then, $K_{\text{MSE}} = \kappa_b\Sigma$ can be minimized by taking $p = H-1$, which yields $K_{\text{MSE}} = \sigma_P^2\kappa_T H^{-2}$. Recall that the value of H can also be changed by rescaling the problem.

In the CRN1 case, we have $\mu = H\kappa_c/(2(H-\nu))$ and $\Sigma = \sigma_D^2/(2\kappa_N(H-\nu))$ for FDf, while $\mu = H_3\kappa_c^2/(6(H-2\nu))$ and $\Sigma = \sigma_D^2/(4\kappa_N(H-\nu))$ for FDc. Again, the value of p that minimizes the asymptotic MSE constant is not necessarily zero.

Steady-state average cost (Example 5.3). The expressions for V_n and B_n here are related to those of Example 5.1, modified as follows: Multiply V_n by $\Theta(t_n^{-1}) = \Theta(n^{-q})$ and add $\Theta(t_n^{-1})$ to $\|B_n\|$. It is reasonable to suppose that (5.3) holds with N_n replaced by $N_n t_n$, which implies that $n^{-p-q-1}C_n \to \kappa_b$ w.p.1 for some positive constant κ_b.

Under those assumptions (and those made in Example 5.1 for the finite horizon), we obtain the following. For the FD estimators with independent random numbers, we have $V_n = \Theta(t_n^{-1}N_n^{-1}c_n^{-2}) = \Theta(n^{2\nu-p-q})$, for both FDf and FDc, $\|B_n\| = \Theta(t_n^{-1} + c_n) = \Theta(n^{-(q\wedge\nu)})$ for FDf, and $\|B_n\| = \Theta(t_n^{-1} + c_n^2) = \Theta(n^{-(q\wedge 2\nu)})$ for FDc. With common random numbers, B_n is the same, while $V_n = \Theta((t_n^{-1}N_n^{-1}c_n^{-1}) = \Theta(n^{\nu-p-q})$ for CRN1 and $V_n = \Theta((t_n^{-1}N_n^{-1}) = \Theta(n^{-p-q})$ for CRN2. For IPA, both $\|B_n\|$ and V_n are $\Theta(t_n^{-1}N_n) = \Theta(n^{-p-q})$.

For the LR estimator, we assume here that $V_n = \Theta(t_n) = \Theta(n^q)$ and $\|B_n\| = \Theta(t_n^{-1}) = \Theta(n^{-q})$. For CLR, we have $V_n = \Theta(1)$ and $\|B_n\| = \Theta(t_n^{-1}) = \Theta(n^{-q})$. From that, the optimal values given in Table 5.2 are easily obtained.

Infinite-horizon discounted model (Example 5.6). Let $t_n/\ln n \to \kappa_t$ as $n \to \infty$, where $\kappa_t \in [0,\infty]$. We want to show that κ_t must be finite and large enough (as specified in Table 5.3) to maximize $\tilde{\eta}$, and that the convergence rate actually depends on the product $\rho\kappa_t$. Recall that the derivative estimator ψ_n is the average of $N_n = \kappa_N n^p$ i.i.d. replications of either $X_\theta(\theta_n, t_n)$ (for IPA) or of a finite difference (for FD).

For IPA, we have $n^p V_n \to \sigma^2 \equiv R$ as $\theta_n \to 0$, $B_n \sim \kappa_\beta \exp(-\rho t_n) \sim \kappa_\beta n^{-\rho\kappa_t}$, and the results of Sections 3–4 apply with $\beta = \rho\kappa_t$ and $\delta = p$ (assuming that $\max(1+2\rho\kappa_t, 2+p) < \lambda_{\mathsf{min}}$). This yields $E[\|\theta_n\|^2] = \Theta(n^{-\kappa})$ with $\kappa = 2\rho\kappa_t \wedge (p+1)$. We also have (w.p.1) $T_n \sim \kappa_T\kappa_N\kappa_t n^p \ln n$, $E[C_n] \sim \kappa_T\kappa_N\kappa_t n^{(p+1)} \ln n/(p+1)$, and $(\ln E[C_n])/E[C_n] \sim (p+1)^2(\kappa_T\kappa_N\kappa_t)^{-1} n^{-(p+1)}$. For $\kappa_t \geq (p+1)/(2\rho)$, this gives $E[\|\theta_n\|^2] = \Theta(n^{-p-1}) = \Theta((\ln E[C_n])/E[C_n])$. Upon a closer examination, we also see that the latter MSE increases linearly with κ_t and that the bias term disappears ($2\beta > 1+\delta$) when $\kappa_t > (p+1)/(2\rho)$. If $\kappa_t = \infty$ (i.e., t_n increases faster than $\Theta(\ln n)$), then the MSE must increase faster than $\Theta((\ln E[C_n])/E[C_n])$. For $\kappa_t < (p+1)/(2\rho)$ (including $\kappa_t = 0$), we obtain $E[\|\theta_n\|^2] = \Theta(n^{-2\rho\kappa_t})$, which converges at a slower rate than $\Theta(n^{-p-1})$. Therefore, the optimal strategy is to take κ_t equal to (or slightly larger than) $(p+1)/(2\rho)$. We just showed that the optimal rate in terms of $E[C_n]$ is obtained if and only if $t_n/\ln n \to \kappa_t$ where κ_t is a large enough but finite constant.

Similar analysis can be carried out when $\alpha_\theta(\theta, t)$ is estimated via finite differences, using $X(\theta, t)$ instead of $X(\theta)$ in each of the setups of Example 5.1. For FDc with independent random numbers, for example, one has $\|B_n\| = \Theta(n^{-2\nu} + e^{-\rho t_n}) = \Theta(n^{-(2\nu\wedge\rho\kappa_t)})$ and $V_n = \Theta(n^{2\nu-p})$. Take $\nu = (p+1)/6$ and $\kappa_t \geq (p+1)/(3\rho)$. Then, $\|B_n\| = \Theta(n^{-(p+1)/3})$, $V_n = \Theta(n^{(p+1)/3})$, and $E\|\theta_n\|^2 = \Theta(n^{-2(p+1)/3}) = \Theta(((E[C_n])^{-1}\ln E[C_n])^{2/3})$. It is not hard to see that one cannot improve upon that convergence rate, and that the rate is lower if $\kappa_t < (p+1)/(3\rho)$ or if $t_n/\ln n \to \infty$. The results of the other lines of Table 5.3 can be obtained in a similar fashion, we leave out the details.

Acknowledgment. We thank the Associate Editor John Birge, the referees, and Qian-Yu Tang for their constructive comments.

REFERENCES

[1] S. Andradóttir, *A stochastic approximation algorithm with varying bounds*, Operations Research, 43 (1995), pp. 1037–1048.

[2] F. Azadivar and J. Talavage, *Optimization of stochastic simulation models*, Mathematics and Computers in Simulation, XXII (1980), pp. 231–241.

[3] J. A. Bather, *Stochastic approximation: A generalization of the Robbins-Monro procedure*, in Proceedings of the Fourth Prague Symposium Asymptotic Statist., P. Mandl and M. Hušková, eds., 1989, pp. 13–27.

[4] H. F. Chen and Y. M. Zhu, *Stochastic approximation procedure with randomly varying truncations*, Scientic Sinica, 29 (1986), pp. 914–926.

[5] K. L. Chung, *On a stochastic approximation method*, Annals of Mathematical Statistics, 25 (1954), pp. 463–483.

[6] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*, J. Wiley, New York, 1986.

[7] V. Fabian, *On asymptotic normality in stochastic approximation*, Annals of Mathematical Statistics, 39 (1968), pp. 1327–1332.

[8] B. L. Fox and P. W. Glynn, *Simulating discounted costs*, Management Science, 35 (1989), pp. 1297–1315.

[9] P. Glasserman, *Gradient Estimation Via Perturbation Analysis*, Kluwer Academic, Norwell, MA, 1991.

[10] P. Glasserman and D. D. Yao, *Some guidelines and guarantees for common random numbers*, Management Science, 38 (1992), pp. 884–908.

[11] P. W. Glynn, *A GSMP formalism for discrete event systems*, Proceedings of the IEEE, 77 (1989), pp. 14–23.

[12] ———, *Likelihood ratio gradient estimation for stochastic systems*, Communications of the ACM, 33 (1990), pp. 75–84.

[13] P. W. Glynn and P. Heidelberger, *Bias properties of budget constrained simulations*, Operations Research, 38 (1990), pp. 801–814.

[14] P. W. Glynn and P. L'Ecuyer, *On the existence and estimation of performance measure derivatives for stochastic recursions*, in Lectures Notes in Control and Information Sciences, vol. 199, June 1994, pp. 429–435. Proceedings of the 11th International Conference on Analysis and Optimization of Systems.

[15] P. W. Glynn, P. L'Ecuyer, and M. Adès, *Gradient estimation for ratios*, in Proceedings of the 1991 Winter Simulation Conference, Dec 1991, pp. 986–993.

[16] P. W. Glynn and W. Whitt, *The asymptotic efficiency of simulation estimators*, Operations Research, 40 (1992), pp. 505–520.

[17] A. Haurie, P. L'Ecuyer, and Ch. Van Delft, *Monte-Carlo optimization of parametrized policies in a class of piecewise deterministic control systems arising in manufacturing flow control*, Discrete Event Dynamic Systems: Theory and Applications, 4 (1994), pp. 87–111.

[18] J. Komlòs and P. Révéz, *A modification of the Robbins–Monro process*, Studia Sci. Math. Hungar, 8 (1973), pp. 329–340.

[19] H. J. Kushner, *Approximation and Weak Convergence Methods for Random Processes*, The MIT Press, Cambridge, Mass., 1984.

[20] H. J. Kushner and D. S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, vol. 26 of Applied Mathematical Sciences, Springer-Verlag, 1978.

[21] H. J. Kushner and H. Huang, *Rates of convergence for stochastic approximation type algorighms*, SIAM Journal of Control and Optimization, 17 (1979), pp. 607–617.

[22] H. J. Kushner and F. J. Vázquez-Abad, *Stochastic approximation algorithms for systems over an infinite horizon*, SIAM Journal on Control and Optimization, 34 (1996), pp. 712–756.

[23] H. J. Kushner and J. Yang, *Stochastic approximation with averaging of the iterates: Optimal asymptotic rate of convergence for general processes*, SIAM Journal of Control and Optimization, 31 (1993), pp. 1045–1062.

[24] H. J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer Verlag, New York, 1997.

[25] P. L'Ecuyer, *An overview of derivative estimation*, in Proceedings of the 1991 Winter Simulation Conference, IEEE Press, Dec 1991, pp. 207–217.

[26] ———, *Convergence rates for steady-state derivative estimators*, Annals of Operations Research, 39 (1992), pp. 121–136.

[27] ———, *On the interchange of derivative and expectation for likelihood ratio derivative estimators*, Management Science, 41 (1995), pp. 738–748.

[28] P. L'Ecuyer, N. Giroux, and P. W. Glynn, *Stochastic optimization by simulation: Numerical experiments with the $M/M/1$ queue in steady-state*, Management Science, 40 (1994), pp. 1245–1261.

[29] P. L'Ecuyer and P. W. Glynn, *Stochastic optimization by simulation: Convergence proofs for the $GI/G/1$ queue in steady-state*, Management Science, 40 (1994), pp. 1562–1578.

[30] ———, *Uniform convergence of truncated–horizon estimators for parametrized regenerative systems*. In preparation, 1996.

[31] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer Verlag, 1993.

[32] M. B. Nevelson and R. Z. Khasminskii, *Stochastic Approximation and Recursive Estimation*, vol. 47 of Translation of Math. Monographs, American Mathematical Society, 1976.

[33] B. T. Polyak, *New method of stochastic approximation type*, Avtomatika i telemekhanika, 7 (1990), pp. 937–946. Translation in Automation and Remote Control (1991), 937–946.

[34] B. T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM Journal on Control and Optimization, 30 (1992), pp. 838–855.

[35] R. Y. Rubinstein and A. Shapiro, *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*, Wiley, New York, 1993.

[36] R. Ruppert, *Almost-sure approximation to the Robbins-Monro and Kiefer-Wolfowitz processes with dependent noise*, Annals of Probability, 10 (1982), pp. 178–187.

[37] ———, *Handbook in sequential analysis*, in Stochastic Approximation, B. K. Ghosh and P. K. Sen, eds., Marcel Dekker, New York, 1991, pp. 503–529.

[38] R. Schwabe, *Stability results for smoothed stochastic approximation procedures*, Tech. Rep. Preprint Nr. A-92-14, Fachbereich Mathematik, Freie Universitat Berlin, 1992.

[39] K. Sigman and R. W. Wolff, *A review of regenerative processes*, SIAM Review, 35 (1993), pp. 269–288.
[40] M. T. Wasan, *Stochastic Approximation*, Cambridge Press, 1969.
[41] G. Yin, *A stopping rule for the Robbins–Monro method*, Journal of Optimization Theory and Applications, 67 (1990), pp. 151–173.
[42] ———, *On extensions of Polyak's averaging approach to stochastic approximation*, Stochastics, 36 (1992), pp. 245–264.
[43] G. Yin, H. Yan, and X. Lou, *On a class of stochastic optimization algorithms with applications to manufacturing models*, in Proceedings of the Third Model-Oriented Data Analysis, H. P. W. W. G. Müller and A. A. Zhigljavsky, eds., Heidelberg, 1993, Physica-Verlag, pp. 213–226.
[44] G. Yin and K. Yin, *Asymptotically optimal rate of convergence of smoothed stochastic recursive algorithms*, Stochastics, 47 (1994), pp. 21–46.
[45] G. Yin and Y. M. Zhu, *Almost sure convergence of stochastic approximation algorithms with non–additive noise*, International J. Control, 47 (1989), pp. 1361–1376.
[46] M. A. Zazanis and R. Suri, *Convergence rates of finite-difference sensitivity estimates stochastic systems*, Operations Research, 41 (1993), pp. 694–703.