Simulation-Based Optimization with Stochastic Approximation Using Common Random Numbers

Nathan L. Kleinman • James C. Spall • Daniel Q. Naiman
Options and Choices, Inc. (OCI), 2232 Dell Range Blvd., Suite 300, Cheyenne, Wyoming 82009
The Johns Hopkins University Applied Physics Laboratory, Johns Hopkins Road, Laurel, Maryland 20723
The Johns Hopkins University Department of Mathematical Sciences, Baltimore, Maryland 21218
kleinman@brutus.mts.jhu.edu • james.spall@jhuapl.edu • dan@jesse.mts.jhu.edu

The method of Common Random Numbers is a technique used to reduce the variance of difference estimates in simulation optimization problems. These differences are commonly used to estimate gradients of objective functions as part of the process of determining optimal values for parameters of a simulated system. Asymptotic results exist which show that using the Common Random Numbers method in the iterative Finite Difference Stochastic Approximation optimization algorithm (FDSA) can increase the optimal rate of convergence of the algorithm from the typical rate of $k^{-1/3}$ to the faster $k^{-1/2}$, where k is the algorithm's iteration number. Simultaneous Perturbation Stochastic Approximation (SPSA) is a newer and often much more efficient optimization algorithm, and we will show that this algorithm, too, converges faster when the Common Random Numbers method is used. We will also provide multivariate asymptotic covariance matrices for both the SPSA and FDSA errors.

(*Common Random Numbers; Simultaneous Perturbation Stochastic Approximation (SPSA); Finite Difference Stochastic Approximation (FDSA); Discrete Event Dynamic Systems*)

1. Introduction

Consider the problem of minimizing (or maximizing) a particular system performance measure by determining values for a set of controllable parameters when the system is modeled by a computer simulation. Often, the form of the performance measure (or loss function) is unknown to the person attempting to optimize, particularly when the simulation is complex or when the person has little or no access to the computer code in which the performance measure is embedded. In this paper, we consider situations where all one can do is obtain (possibly noisy) measurements of the loss function for given parameter values. Here, the gradient of the loss function is also unavailable, thus making gradient-based methods incapable of finding the optimal parameters. Some interest has surfaced in automatic differentiation tools (see Iri 1991, Juedes 1991, and Soulié 1991), but again these require access to the source code of the computer simulation. These also require that the source code be in a certain programming language or a certain format, and some require prohibitively large amounts of computer memory (Soulié 1991).

With these considerations in mind, we turn to non-gradient-based methods. Stochastic approximation (SA) is an iterative technique introduced in the 1950s by Robbins and Monro (1951), Kiefer and Wolfowitz (1952), and Blum (1954), which can often be used to solve optimization problems from both real systems and computer simulations of real systems. The category of SA algorithms introduced by Kiefer

0025-1909/99/4511/1570$05.00
1526-5501 electronic ISSN

and Wolfowitz can be used in situations where loss function extrema are desired and only noisy loss function measurements are available. Although the Robbins-Monro SA algorithms generally converge to the optimum faster than those of the Kiefer-Wolfowitz type, they require direct measurements of the gradient, and thus are not applicable to the above kinds of problems. Section 2 of this paper will discuss how Kiefer-Wolfowitz type stochastic approximation algorithms can be used to optimize these kinds of systems and will review two key examples of SA algorithms, simultaneous perturbation stochastic approximation (SPSA) and finite differences stochastic approximation (FDSA).

Practitioners have often attempted to reduce the variance of the parameter estimates produced from SA algorithms, FDSA in particular, by using the method of Common Random Numbers (CRN) (see Gal et al. 1984, Glasserman and Yao 1992, L'Ecuyer and Perron 1994, and L'Ecuyer and Yin 1998). When models of real systems are represented as computer programs, the noise in the system is modeled by using combinations of computer generated random numbers. To optimize the parameters of such a system, SA algorithms at each iteration use an estimate of the difference in performance measures, $(X - Y)$, for example, to obtain new parameter estimates. Instead of using two independent vectors or streams of independent Uniform (0, 1) random numbers to generate X and Y, the CRN method attempts to increase $\mathrm{Cov}(X, Y)$ and thereby improve the efficiency of the SA algorithm by using the *same* realization of the vector of Uniform (0, 1) random numbers to generate *both* X and Y.

Results exist (for FDSA) that show the use of CRN improves the rate of convergence of the SA estimates to the optimal parameter values under certain conditions (L'Ecuyer and Yin 1998). Without the use of CRN, the optimal FDSA rate is typically $k^{-1/3}$ (see L'Ecuyer and Yin 1998, or Spall 1992), where k is the number of iterations of the algorithm. Using CRN, however, it was shown under certain conditions (see L'Ecuyer and Yin 1998) that the FDSA rate could be $k^{-1/2}$. It is important to note that this is the same rate of convergence attained by the Robbins-Monro SA algorithms, but here it is attained without the use of direct gradient measurements. The results in L'Ecuyer and Yin (1998), however, do not include specific expressions for the multivariate asymptotic mean and covariance matrix of the parameter estimation errors at this faster rate of convergence. In §2 of this paper we will show that the SPSA algorithm also attains this faster rate of convergence when CRN can be used, and furthermore, we will provide multivariate expressions of the asymptotic error means and covariances for both FDSA and SPSA.

Section 3 illustrates the theoretical results of the previous section with a numerical example comparing the use of CRN and independent random numbers (IRN) for both SPSA and FDSA. In the example, we find the parameter estimation error to be smaller when the CRN method is employed than when independent random numbers are used.

Conclusions will be given in §4, and Appendix A provides proofs of the theoretical results in §2.

2. Stochastic Approximation and Common Random Numbers

In this section we focus on simulation-based optimization via stochastic approximation algorithms. First, we give a description of the FDSA and SPSA algorithms along with results that apply when CRN is not used. Then we will discuss how CRN can be implemented in the SA algorithms and give new results.

2.1. Stochastic Approximation Background

If $\theta \in R^p$ is a vector representing the p parameters in the system we can control, and ω is a vector representing (uncontrollable) randomness in the system, let

$$L(\theta) = E[f(\theta, \omega)] \tag{2.1}$$

be the loss function we wish to minimize, where f is a performance measure on the system, and the expectation is with respect to the random vector ω. Let $g(\theta)$ be the gradient of L with respect to θ. Stochastic approximation algorithms attempt to find a local minimizer θ^* by starting at a fixed $\hat{\theta}_0$ and iterating according to the following scheme:

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{g}_k(\hat{\theta}_k), \tag{2.2}$$

where $\hat{g}_k$ is an estimate of the gradient g, and $\{a_k\}$ is a *gain* sequence of positive scalars such that $a_k \to 0$ and $\sum_{k=1}^{\infty} a_k = \infty$. We will focus on non-gradient-based SA algorithms, that is, those for which only loss function measurements are required to obtain $\hat{g}_k$.

The two primary non-gradient-based SA algorithms we will discuss are distinguished by the manner in which the estimate $\hat{g}_k$ is defined. In the classical finite differences stochastic approximation (FDSA) (Kiefer and Wolfowitz 1952, Kushner and Clark 1978), the lth component of the gradient estimate is defined as

$$\hat{g}_{kl}(\hat{\theta}_k) = \frac{y_{kl}^{+} - y_{kl}^{-}}{2c_k}, \tag{2.3}$$

where $\{c_k\}$ is a sequence of positive scalars such that $c_k \to 0$ and $\sum_{k=1}^{\infty} a_k^2 c_k^{-2} < \infty$, and where y_{kl}^{+} and y_{kl}^{-} represent noisy measurements of the loss function and are defined as

$$y_{kl}^{\pm} = f(\hat{\theta}_k \pm c_k e_l, \omega_k^{l\pm}) \tag{2.4}$$

for $l = 1, \ldots, p$ (where the notation ± means the expression is valid using either + throughout or − throughout). Here, e_l is the lth unit vector and $\omega_k^{l\pm}$ represent realizations of the uncontrolled randomness in the system. Note that $2p$ loss function measurements are required to generate one estimate of the gradient. The forward differences version of FDSA (which uses $p + 1$ measurements for each gradient estimate) is not considered here because it has been shown to converge much more slowly than the central differences FDSA algorithm described above (Chin 1997, Kushner and Clark 1978).

A newer SA algorithm, simultaneous perturbation stochastic approximation (SPSA), was introduced and developed by Spall (Spall 1987, 1988, 1992). In this algorithm the gradient estimate is defined as follows. Let $\Delta_k \in R^p$ be a vector of p mutually independent random variables satisfying conditions in Spall (1992). For example, the components could be independent Bernoulli (±1) distributed random variables with each outcome occurring with probability one half (and this distribution, in fact, is asymptotically optimal (Sadegh and Spall 1998) and will be used throughout this paper). Then the lth component of the SPSA gradient estimate is defined as

$$\hat{g}_{kl}(\hat{\theta}_k) = \frac{y_k^{+} - y_k^{-}}{2c_k \Delta_{kl}}, \tag{2.5}$$

where

$$y_k^{\pm} = f(\hat{\theta}_k \pm c_k \Delta_k, \omega_k^{\pm}). \tag{2.6}$$

Observe that the numerator in Equation (2.5) is the same for each l. Thus, only *two* measurements of the loss function are required to obtain one SPSA gradient estimate.[1]

The following assumptions are similar to those in Kiefer and Wolfowitz (1952), Spall (1992), and others. See Spall (1992) for a discussion.

Assumption (A1). *Let α_0, α_1, and α_2 denote positive constants. Consider all $k \geq K$ for some $K < \infty$. Suppose that, for each such k, the $\{\Delta_{ki}\}$ are i.i.d. ($i = 1, 2, \ldots, p$) and symmetrically distributed about 0 with $|\Delta_{ki}| \leq \alpha_0$ a.s. and $E|\Delta_{ki}^{-1}| \leq \alpha_1$. Also, for almost all $\hat{\theta}_k$ (at each $k \geq K$), suppose that for all θ in an open neighborhood of $\hat{\theta}_k$ that is not a function of k or ω, $L^{(3)}(\theta) \equiv \partial^3 L / \partial\theta^T \partial\theta^T \partial\theta^T$ exists continuously with individual elements satisfying $|L_{i_1 i_2 i_3}^{(3)}(\theta)| \leq \alpha_2$.*

Assumption (A2). *For some $\alpha_3, \alpha_4, \alpha_5 > 0$, $\delta \geq 0$ and $\forall k$, assume $E[(y_k^{\pm} - L(\hat{\theta}_k \pm c_k\Delta_k))^{2+\delta}] \leq \alpha_3$, $E[L(\hat{\theta}_k \pm c_k\Delta_k)^{2+\delta}] \leq \alpha_4$, and $E[\Delta_{kl}^{-2-\delta}] \leq \alpha_5$ ($l = 1, 2, \ldots, p$).*

Assumption (A3). *$\|\hat{\theta}_k\| < \infty$ a.s. $\forall k$.*

Assumption (A4). *θ^* is an asymptotically stable solution of the differential equation $dx(t)/dt = -g(x)$.*

Assumption (A5). *Let $D(\theta^*) = \{x_0 : \lim_{t\to\infty} x(t|x_0) = \theta^*\}$, where $x(t|x_0)$ denotes the solution to the differential equation in* (A4) *based on initial condition x_0. There exists a compact $S \subseteq D(\theta^*)$ such that $\hat{\theta}_k \in S$ infinitely often for almost all sample points.*

Assumption (A6). *Let σ^2, ρ^2, and ξ^2 be such that $E[(y_k^{+} - L(\hat{\theta}_k + c_k\Delta_k) - y_k^{-} + L(\hat{\theta}_k - c_k\Delta_k))^2 | \mathcal{F}_k] \to$*

[1] Random directions stochastic approximation (see Kushner and Clark 1978, pp. 58–60) also requires only two measurements per gradient estimate, but its overall performance has not compared favorably with that of SPSA (see Chin 1997).

σ^2 a.s., $E[\Delta_{kl}^{-2}] \to \rho^2$, and $E[\Delta_{kl}^2] \to \xi^2$ as $k \to \infty\ \forall l$, where $\mathcal{F}_k$ is the sigma field generated by $\{\hat{\theta}_0, \ldots, \hat{\theta}_k\}$. Also, $\forall k$ sufficiently large and almost all ω, let the sequence $\{E[(y_k^+ - L(\hat{\theta}_k + c_k\Delta_k) - y_k^- + L(\hat{\theta}_k - c_k\Delta_k))^2 | \mathcal{F}_k, c_k\Delta_k = \tau]\}$ be equicontinuous at $\tau = 0$ and continuous in τ on some compact, connected set containing $c_k\Delta_k$ a.s.

Under (A1)–(A5), Proposition 1 of Spall (1992) proves that $\hat{\theta}_k \to$ *a.s.* θ^* for SPSA (and similarly for FDSA), where θ^* is a local minimizer of L. If we add Assumption (A6) and constrain δ to be strictly positive in (A2), then asymptotic distribution and rate of convergence results for the SPSA parameter estimate errors for the general L and y defined above are given as follows (FDSA results are similar). Again, for ease of notation, we let $\{\Delta_{kl}\}$ be independent Bernoulli (± 1) random variables $\forall k, l$. Define the gain sequences as $a_k = ak^{-\alpha}$ and $c_k = ck^{-\gamma}$ where $a, c, \gamma > 0$, $0 < \alpha \leq 1$, $2\alpha - 2\gamma > 1$, and $\alpha - 2\gamma > 0$. Let $\beta = \alpha - 2\gamma$, and assume $\beta - 4\gamma \leq 0$. Next, let $H(\theta)$ be the Hessian matrix of $L(\theta)$, and let P be an orthogonal matrix satisfying $P^T H(\theta^*) P = a^{-1}\,\mathrm{diag}(\lambda_1, \ldots, \lambda_p)$. If $\alpha = 1$, define $\beta_+ = \beta < 2 \min_i \lambda_i$; otherwise set $\beta_+ = 0$ (note that requiring $\beta < 2 \min_i \lambda_i$ is not restrictive since we can always choose a such that the inequality holds). Then from Fabian (1968) and Proposition 2 of Spall (1992) we have that the normalized estimation error $k^{\beta/2}(\hat{\theta}_k - \theta^*)$ is asymptotically $N_p(\mu, PMP^T)$ distributed in the SPSA case (where N_p denotes the p-variate normal distribution). Here, $\beta/2 \leq 1/3$, and μ is a mean vector that involves the second and third derivatives of $L(\theta)$ at θ^* and for which formulas are given in Spall (1992). We also have

$$M = \frac{a^2\sigma^2}{4c^2}\,\mathrm{diag}\left[\frac{1}{2\lambda_1 - \beta_+}, \ldots, \frac{1}{2\lambda_p - \beta_+}\right], \qquad (2.7)$$

where

$$\sigma^2 = 2\,\mathrm{Var}[f(\theta^*, \omega)]. \qquad (2.8)$$

2.2. Using Common Random Numbers to Reduce Variance

A key ingredient used in Fabian (1968) and Proposition 2 of Spall (1992) in deriving the above asymptotic distributions is the asymptotic variance of $\hat{g}_k$. We will show it is possible to reduce the asymptotic variability of the SPSA and FDSA parameter estimation error in cases where we can reduce the variance of $\hat{g}_k$.

From Equations (2.3) to (2.6) we see each SA gradient estimate is formed from the difference of pairs of observed system performance measures. For example, in the SPSA algorithm

$$\hat{g}_{kl}(\theta_k) = \frac{1}{2c_k\Delta_{kl}}\,[f(\hat{\theta}_k + c_k\Delta_k, \omega_k^+) - f(\hat{\theta}_k - c_k\Delta_k, \omega_k^-)] \qquad (2.9)$$

for $l = 1, \ldots, p$. In a real system, functions of ω_k^+ and ω_k^- in (2.6) (and similarly in (2.4)) represent instances of the system's uncontrolled randomness. On the other hand, if the optimization is based on a computer simulation of a real world system, the researcher may have some control over how the $\omega_k^{\pm}$ are generated. In the code of a computer simulation, it is usually the case that, for each k, ω_k^+ and ω_k^- are implemented as vectors or streams of mutually independent Uniform (0, 1) random variables (and each is generated independently of $\hat{\theta}_0, \ldots, \hat{\theta}_k$). If so, it may be possible (by setting random number seeds, for example) to implement the CRN method by setting $\omega_k^- = \omega_k^+ =: \omega_k\ \forall k$ in the SPSA case, and $\omega_k^{l-} = \omega_k^{l+} =: \omega_k^l\ \forall k, l$ in the FDSA case. Then the SPSA gradient estimate is written as

$$\hat{g}_{kl}(\hat{\theta}_k) = \frac{1}{2c_k\Delta_{kl}}\,[f(\hat{\theta}_k + c_k\Delta_k, \omega_k) - f(\hat{\theta}_k - c_k\Delta_k, \omega_k)] \qquad \forall\, k, l$$

(and similarly for the FDSA gradient estimate). Recall that $f(\theta, \omega)$ is a measurement of the loss function and is the output of a computer simulation that uses θ and ω as inputs. Thus, using the same sequence of Uniform (0, 1) random numbers to run both simulations each iteration results (via the cancellation of terms in the difference of the Taylor expansions of f) in a reduction in the variance of $\hat{g}_k$ compared to the case where ω_k^+ and ω_k^- were independent, and this subsequently leads to smaller asymptotic variances of the SPSA and FDSA estimates. These results are summarized in the following new theorem and corollary. In the SPSA case (Theorem 2.1), for all k, $\hat{g}_k$ is defined as in (2.5), and $y_k^{\pm} = f(\hat{\theta}_k \pm c_k\Delta_k, \omega_k)$. Also, the Δ_{kl} are independent Bernoulli (±1) random variables, with each outcome occurring with probability one half, for all k and l. In the FDSA case (Corollary 2.1), $\hat{g}_k$ is defined as in (2.3), and $y^{\pm}_{kl} = f(\hat{\theta}_k \pm c_k e_l, \omega^l_k)$ for all k and $l = 1, \ldots, p$, where e_l is the lth unit vector. In both results, $L(\theta) = E[f(\theta, \omega)]$.

Theorem 2.1. *Suppose Assumptions* (A1)–(A5) *hold and that* δ *in* (A2) *is strictly positive. Also assume* $|\partial f(\hat{\theta}_k, \omega)/\partial\theta_l| \leq \alpha_6$ *for some* $\alpha_6 > 0$ *and for almost all* $\hat{\theta}_k$ *(*k *sufficiently large), almost all* ω*, and all* $l = 1, \ldots, p$*. Let* $a_k = ak^{-\alpha}$ *and* $c_k = ck^{-\gamma}$ *with* $a, c, \gamma > 0$, $0 < \alpha \leq 1$*, and* $2\alpha - 2\gamma > 1$*. Let* $\beta = \alpha$ *with* $\beta - 4\gamma < 0$ *(note the change from* §2.1 *where* $\beta = \alpha - 2\gamma$*). Define* β_+, P*, and* $\lambda_1, \ldots, \lambda_p$ *as in* §2.1.

Then,

$$k^{\beta/2}(\hat{\theta}_k - \theta^*) \overset{d}{\rightarrow} N_p(0, PMP^T)$$

with the ij-entry of M as follows:

$$M_{ij} = a^2(P^T \sum P)_{ij}\left(\frac{1}{\lambda_i + \lambda_j - \beta_+}\right)$$

where

$$\sum_{ij} = \begin{cases} \sum_{l=1}^{p} E\left[\left(\frac{\partial}{\partial\theta_l} f(\theta, \omega)\Big|_{\theta=\theta^*}\right)^2\right] & \text{if } i = j, \\ 2E\left[\left(\frac{\partial}{\partial\theta_i} f(\theta, \omega)\Big|_{\theta=\theta^*}\right)\left(\frac{\partial}{\partial\theta_j} f(\theta, \omega)\Big|_{\theta=\theta^*}\right)\right] & \text{if } i \neq j. \end{cases} \tag{2.10}$$

Corollary 2.1. *In the FDSA case, assume the conditions of Theorem* 2.1. *Then,*

$$k^{\beta/2}(\hat{\theta}_k - \theta^*) \overset{d}{\rightarrow} N_p(0, P\tilde{M}P^T)$$

with the ij-entry of $\tilde{M}$ *as follows:*

$$\tilde{M}_{ij} = a^2(P^T \tilde{\sum} P)_{ij}\left(\frac{1}{\lambda_i + \lambda_j - \beta_+}\right),$$

where

$$\tilde{\sum}_{ij} = \begin{cases} E\left[\left(\frac{\partial}{\partial\theta_i} f(\theta, \omega)\Big|_{\theta=\theta^*}\right)^2\right] & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{2.11}$$

Note that while the faster rate of convergence ($\beta/2 = 1/2$) for FDSA with CRN has been shown previously (see L'Ecuyer and Yin 1998), this faster rate is a new result for SPSA. Additionally, the multivariate asymptotic covariance results for both SPSA and FDSA shown above have not been reported previously. Note that if we let $p = 1$ above (a one-dimensional optimization problem), then our results coincide with those in 4.1 of L'Ecuyer and Yin (1998), where our β_+ equals $1 + \delta$ in Theorem 4.1 of L'Ecuyer and Yin (1998).[2]

An area for further study would be to compare the asymptotic mean square errors (MSE) of SPSA and FDSA, taking into account that SPSA requires p times fewer loss function measurements than does FDSA. If we consider the special case where the Hessian matrix, H, of L is a diagonal matrix, then P is the identity matrix, and we gain insight by looking at the relative shapes of L and f. Note that in this case, the larger diagonal entries of the matrix $\tilde{\Sigma}$ represent directions from the point θ^* in which f more steeply increases or decreases. Also, if $\lambda_1 > \lambda_2$, say, then at the point θ^*, the slope of L is changing more rapidly in the direction of its first component than in its second. Thus, roughly speaking, if f has a steep slope in the same directions from θ^* as in the directions where the slope of L is rapidly changing, then the larger diagonal entries of $\tilde{\Sigma}$ will be paired with the larger values of $\{\lambda_1, \ldots, \lambda_p\}$. This would then imply that the trace of PMP^T will be larger than p times the trace of $P\tilde{M}P^T$, and FDSA will perform better than SPSA (even though SPSA uses p times fewer measurements each iteration). But, on the other hand, when the larger diagonal entries of $\tilde{\Sigma}$ are paired with the smaller λ values, then $\mathrm{tr}(PMP^T) < \mathrm{tr}(P\tilde{M}P^T) \cdot p$, and SPSA outperforms FDSA. In an even more specific case, when H is some constant times the identity matrix, we have $\mathrm{tr}(PMP^T) = \mathrm{tr}(P\tilde{M}P^T) \cdot p$, and the same number of measurements are required by each algorithm to obtain the same level of accuracy. These specific cases (using

[2] Using CRN, note that when we define $c_k = ck^{-\gamma}$ we have some freedom in how we choose the exponent. We may choose γ large enough (i.e., choose β large enough in L'Ecuyer and Yin (1998)) so that the bias term in the gradient estimate is asymptotically negligible compared to the variance, making Case 1 in L'Ecuyer and Yin (1998) ($4\gamma > \alpha$ in our notation) the applicable case.

CRN) contrast the case where independent random numbers are used. There, SPSA attains greater accuracy than FDSA under fairly general conditions (see Spall 1992, §IV), even when the same number of total measurements are used in each algorithm. Hence, $\mathrm{tr}(PMP^T) < \mathrm{tr}(P\tilde{M}P^T)$, not just $\mathrm{tr}(PMP^T) < \mathrm{tr}(P\tilde{M}P^T) \cdot p$.

3. Numerical Illustration of Results

3.1. Introduction

In this section we present a numerical study that illustrates and compares the performances of the SPSA and FDSA algorithms in different scenarios depending on whether ω_k^+ and ω_k^- use independent or common random numbers. We also examine the situation where an attempt to use CRN is made, but ω_k^+ and ω_k^- do not have all components in common. We call this scenario "Partial Common Random Numbers" (see also the air traffic simulation study in Kleinman et al. 1997). As the theoretical results of the previous section imply, we expect SPSA and FDSA will perform best when using common random numbers and worst when using independent random numbers.

The loss function we wish to minimize is

$$L(\theta) = \theta^T\theta + E\left[\sum_{i=1}^{p} e^{-X_i\theta_i}\right]$$

where θ_i is the ith component of $\theta \in [0, \infty) \times \cdots \times [0, \infty) \subseteq R^p$, and X_i is an exponentially distributed random variable with parameter η_i, $i = 1, \ldots, p$. This loss function is similar to the one Chin (1997) uses in his comparison of several SA algorithms. Using the relatively simple loss function in place of a larger, more complex real simulation allows us to compute a large number of optimization runs and thereby reduce case-dependent variation.

3.2. Numerical Study

In this study we chose $p = 10$ in the above definition of $L(\theta)$. The exponential random variables $\{X_{ki}^{\pm}\}$ were generated independently by the inversion method. That is, for each iteration k and each $i = 1, \ldots, p$, we set $X_{ki}^{\pm} = (-1/\eta_i)\ln(1 - \omega_{ki}^{\pm})$ where $\{\omega_{ki}^{\pm}\}$ are independent Uniform (0, 1) random variables generated using the system C function *drand*48( ). The constant vector η was defined as

$$[1.10254, 1.69449, 1.47894, 1.92617, 0.750471,$$
$$1.32673, 0.842822, 0.724652, 0.769311, 1.3986]^T,$$

each component representing the parameter of the particular exponential distribution. The study was performed on a Sun Sparc10 workstation.

For both SPSA and FDSA, the initial point, $\hat{\theta}_0$, was chosen to be $[1, 1, \ldots, 1]^T$. We found the minimizer,

$$\theta^* = [0.286, 0.229, 0.248, 0.211, 0.325,$$
$$0.263, 0.315, 0.327, 0.323, 0.256]^T,$$

through differentiation and by using the mathematical software package Maple. Both SPSA and FDSA were used to minimize $L(\theta)$ in each of three scenarios. In the first scenario, using independent random numbers (IRN), ω_k^+ and ω_k^- were chosen independently of each other. More specifically, at each iteration we generated $2p = 20$ Uniform (0, 1) random numbers and set $\omega_k^+ = [\omega_{k,1}, \ldots, \omega_{k,10}]^T$ and $\omega_k^- = [\omega_{k,11}, \ldots, \omega_{k,20}]^T$. These random numbers were then used to generate the exponential random numbers.

For the second scenario, Partial Common Random Numbers (PCRN), we desired to simulate a partial CRN case where ω_k^+ and ω_k^- had some components in common and yet did not have full synchronicity. Thus, in this scenario we generated 10 Uniform (0, 1) random variables independently for each k and then set $\omega_k^+ = [\omega_{k,1}, \ldots, \omega_{k,10}]^T$ and $\omega_k^- = [\omega_{k,1}, \ldots, \omega_{k,7}, \omega_{k,10}, \omega_{k,9}, \omega_{k,8}]^T$.

In the third scenario (CRN), full synchronicity was simulated in order to use the method of common random numbers. That is, we set $\omega_{k,i}^+ = \omega_{k,i}^-$ for $i = 1, \ldots, 10$ and for all k.

Each row of Table 1 describes results for a different scenario. Note that for both SPSA and FDSA and all scenarios, we set $a = 0.7$ and $c = 0.5$ in the gain sequences. Also, for IRN and PCRN we used $\gamma = 0.167$, and for CRN we used $\gamma = 0.49$. All three scenarios used $\alpha = 1.0$ (from the conditions on α and γ in §2.1 we have that the values $\alpha = 1$ and $\gamma = 1/6$ provide the fastest rate of convergence in the IRN and PCRN cases, and similarly from the conditions in

Table 1 **Numerical Results for 100 Replications of SPSA and FDSA Runs**

	SPSA		FDSA	
	$L(\hat{\theta}_{10,000})$	$\frac{\lVert\hat{\theta}_{10,000} - \theta^*\rVert}{\lVert\hat{\theta}_0 - \theta^*\rVert}$	$L(\hat{\theta}_{1000})$	$\frac{\lVert\hat{\theta}_{1000} - \theta^*\rVert}{\lVert\hat{\theta}_0 - \theta^*\rVert}$
IRN	8.725	0.0190	8.736	0.0410
PCRN	8.723	0.0071	8.724	0.0110
CRN	8.723	0.0065	8.723	0.0064

Theorem 2.1, we see that $\alpha = 1$ and any $\gamma \in (0.25, 0.5)$ are asymptotically optimal in the CRN case). Since the FDSA algorithm requires $p = 10$ times more loss function evaluations per iteration than does SPSA, all FDSA runs were carried out using 1,000 iterations, and all SPSA runs were carried out using 10,000 iterations in order to equate the total number of function evaluations performed by each algorithm. Furthermore, for both SPSA and FDSA in each of IRN, PCRN, and CRN, 100 independent replications of each minimization run were performed. Thus, the final loss function values obtained in Columns one and three of the table are averages over the 100 replications, as well as the ratios in Columns two and four.

3.3. Interpretation of Results

The results in Table 1 show that for both SPSA and FDSA, this study agrees with the theory of the previous section. The final loss function values are smallest and the errors $\|\hat{\theta}_{10,000} - \theta^*\|_{SPSA}$ and $\|\hat{\theta}_{1000} - \theta^*\|_{FDSA}$ are smallest in the pure CRN scenario. Additionally, the final loss function values are largest and the errors $\|\hat{\theta}_{10,000} - \theta^*\|_{SPSA}$ and $\|\hat{\theta}_{1000} - \theta^*\|_{FDSA}$ are largest in the IRN scenario, with the PCRN scenario results falling in between. Note also that, as the theory implies, $\sqrt{k}\|\hat{\theta}_k - \theta^*\|$ appeared to be bounded as k increased when using CRN.

Furthermore, SPSA outperforms FDSA in the IRN and PCRN scenarios, but performs no better than FDSA in the CRN case. More specifically, the ratio

$$\frac{\|\hat{\theta}_{10,000} - \theta^*\|_{SPSA}}{\|\hat{\theta}_{1000} - \theta^*\|_{FDSA}}$$

is about 0.46 for IRN and 0.65 for PCRN, but 1.02 for CRN.

4. Conclusions

This article describes the use of Common Random Numbers as a means of reducing the variance in stochastic approximation estimates for problems where only loss function measurements are available and simulation-based optimization is performed. We show that when CRN can be used, the estimates of both SPSA and FDSA converge to the optimal values at a faster rate ($k^{-1/2}$) than when CRN is not used ($k^{-1/3}$). Multivariate asymptotic means and variances for the SPSA and FDSA iterates are also given.

This work also contains a numerical study illustrating the theoretical results. We find, in the example, the parameter estimation errors from both SPSA and FDSA take their smallest values in the pure CRN setting and largest in the independent random numbers setting. In the example, when the number of function evaluations performed by SPSA and FDSA is equated, the SPSA estimates give smaller errors than those of FDSA in the independent and partial CRN settings. In the pure CRN setting the two algorithms perform nearly equally well.

Further comparison between the asymptotic performances of SPSA and FDSA in the CRN setting is an area for future research.[3]

[3] This work was partially supported by U.S. Navy Contract N00039-95-C-002 and the JHU/APL IRAD Program.

A. Proofs for the Pure CRN Setting

In this appendix we prove Theorem 2.1 and Corollary 2.1, which give asymptotic distribution and rate of convergence results in the pure CRN setting for the SPSA and FDSA algorithms, respectively.

Proof of Theorem 2.1. As in Spall (1992), define the conditional bias

$$b_k(\hat{\theta}_k) = E[\hat{g}_k(\hat{\theta}_k) - g(\hat{\theta}_k)|\hat{\theta}_k]. \tag{A.1}$$

Note that our loss function, $L(\theta) = E[f(\theta, \omega)]$, and the sequence $\{\hat{\theta}_k\}$, including its use of CRN, satisfy the assumptions (A1)–(A5). Thus, Lemma 1 in Spall (1992) implies $b_k(\hat{\theta}_k) = O(k^{-2\gamma})$ with probability one, and Proposition 1 in Spall (1992) yields $\hat{\theta} \to a.s.\ \theta^*$. The remainder of the proof of Theorem 2.1 will proceed similarly to that of Proposition 2 in Spall (1992); that is, we will show that conditions (2.2.1), (2.2.2), and (2.2.3) in Fabian (1968) hold.

First, Lemma 1 and Proposition 1 in Spall (1992) imply that there exists an open neighborhood of $\hat{\theta}_k$, for all sufficiently large k, wherein H is continuous and that includes θ^*. Then using (A.1),

$$\begin{aligned} E[\hat{g}_k(\hat{\theta}_k)|\mathcal{F}_k] &= E[\hat{g}_k(\hat{\theta}_k)|\hat{\theta}_k] \\ &= g(\hat{\theta}) + b_k(\hat{\theta}) \\ &= H(\bar{\theta}_k)(\hat{\theta}_k - \theta^*) + b_k(\hat{\theta}), \end{aligned} \tag{A.2}$$

where $\bar{\theta}_k$ lies on the line segment between $\hat{\theta}_k$ and θ^* and where $\mathcal{F}_k$ is the sigma field generated by $\{\hat{\theta}_0, \ldots, \hat{\theta}_k\}$. Then as in Fabian (1968) we use (A.2) and write

$$\begin{aligned} \hat{\theta}_{k+1} - \theta^* &= (I - k^{-\alpha}\Gamma_k)(\hat{\theta}_k - \theta^*) \\ &\quad + k^{-(\alpha+\beta)/2}\Phi_k V_k + k^{-\alpha-\beta/2}T_k, \end{aligned} \tag{A.3}$$

where $T_k = -ak^{\beta/2}b_k(\hat{\theta}_k)$, $V_k = k^{(\beta-\alpha)/2}[\hat{g}_k(\hat{\theta}_k) - E(\hat{g}_k(\hat{\theta}_k)|\hat{\theta}_k)]$, $\Phi_k = -aI$, and $\Gamma_k = aH(\bar{\theta}_k)$.

As in Spall (1992), $\Gamma_k \to aH(\theta^*)$ *a.s.* by the continuity of H and the almost sure convergence of $\hat{\theta}_k$. Furthermore, since $b_k(\hat{\theta}_k)$ is $O(k^{-2\gamma})$ with probability one, we have $T_k = O(k^{(\beta/2)-2\gamma})$ *a.s.* Thus, $T_k \to 0$ *a.s.* since $\beta - 4\gamma < 0$, and Fabian's condition (2.2.1) holds.

Next, let $\Delta_k^{-1} = (\Delta_{k1}^{-1}, \ldots, \Delta_{kp}^{-1})^T$. Then we have

$$\begin{aligned} &E[V_k V_k^T|\mathcal{F}_k] \\ &\overset{a.s.}{=} k^{\beta-\alpha}E\left[\Delta_k^{-1}(\Delta_k^{-1})^T\left(\frac{1}{4c_k^2}\right)(f(\hat{\theta}_k + c_k\Delta_k, \omega_k) \right. \\ &\qquad \left. - f(\hat{\theta}_k - c_k\Delta_k, \omega_k))^2 \,\middle|\, \mathcal{F}_k\right] \\ &\qquad - k^{\beta-\alpha}(g(\hat{\theta}_k) + b_k(\hat{\theta}_k))(g(\hat{\theta}_k) + b_k(\hat{\theta}_k))^T \quad \text{by (2.5) and } (A.1), \\ &\overset{a.s.}{=} k^{\beta-\alpha}\left(\frac{1}{4c_k^2}\right)E[\Delta_k^{-1}(\Delta_k^{-1})^T E[(f(\hat{\theta}_k + c_k\Delta_k, \omega_k) \\ &\qquad - f(\hat{\theta}_k - c_k\Delta_k, \omega_k))^2|\mathcal{F}_k, \Delta_k]|\mathcal{F}_k] + o(k^{\beta-\alpha}), \end{aligned}$$

and we may use the $o(k^{\beta-\alpha})$ term since $b_k(\hat{\theta}_k) = a.s.\ O(k^{-2\gamma})$, $\hat{\theta}_k \to a.s.\ \theta^*$, and $g(\theta^*) = 0$. Using Taylor expansions,

$$\begin{aligned} &E[(f(\hat{\theta}_k + c_k\Delta_k, \omega_k) - f(\hat{\theta}_k - c_k\Delta_k, \omega_k))^2|\mathcal{F}_k, \Delta_k] \\ &\overset{a.s.}{=} 4c_k^2 E[(\Delta_k^T \nabla f(\hat{\theta}_k, \omega_k))^2|\mathcal{F}_k, \Delta_k] + O(c_k^4). \end{aligned} \tag{A.4}$$

We point out here that the improvement found in using the CRN method comes from the cancellation of the $O(1)$ terms in the Taylor expansion of the left side of equation (A.4). If the CRN method were not used, ω_k^+ and ω_k^- would replace the two instances of ω_k in the left side above, and the right side would be $O(1)$ instead of $O(c_k^2)$.

Then, using (A.4), the ij-entry of $E[V_k V_k^T|\mathcal{F}_k]$ can, with probability one, be written as

$$k^{\beta-\alpha}\sum_{l=1}^{p}\sum_{m=1}^{p} E\left[\frac{\Delta_{kl}\Delta_{km}}{\Delta_{ki}\Delta_{kj}}\right] E[\nabla_l f(\hat{\theta}_k, \omega_k)\nabla_m f(\hat{\theta}_k, \omega_k)|\mathcal{F}_k] + o(k^{\beta-\alpha})$$

by the independence of Δ_k and $\mathcal{F}_k$, where $\nabla_l f(\hat{\theta}_k, \omega_k)$ is the lth component of the gradient vector ∇f evaluated at $\hat{\theta}_k$ and ω_k. Then, since $\beta = \alpha$ and using the fact that the $\{\Delta_{kl}\}$ are independent Bernoulli (±1) random variables (with parameter 1/2) for all k and $l = 1, \ldots, p$, we have

$$\begin{aligned} &E[V_k V_k^T|\mathcal{F}_k]_{ij} \\ &\overset{a.s.}{=} \begin{cases} \displaystyle\sum_{l=1}^{p} E[(\nabla_l f(\hat{\theta}_k, \omega_k))^2|\mathcal{F}_k] + o(1) & \text{if } i = j, \\ 2E[(\nabla_i f(\hat{\theta}_k, \omega_k))(\nabla_j f(\hat{\theta}_k, \omega_k))|\mathcal{F}_k] + o(1) & \text{if } i \neq j. \end{cases} \end{aligned}$$

By independence of ω_k and $\mathcal{F}_k$,

$$\begin{aligned} E[(\nabla_l f(\hat{\theta}_k, \omega_k))^2|\mathcal{F}_k] &\overset{a.s.}{=} E[(\nabla_l f(\hat{\theta}_k, \omega_k))^2|\hat{\theta}_k] \\ &\overset{a.s.}{=} \int (\nabla_l f(\hat{\theta}_k, \omega_k))^2 dP_{\omega_k} \end{aligned} \tag{A.5}$$

where P_{ω_k} is the probability measure associated with ω_k. Since $\{\omega_k\}$ is an i.i.d. sequence,

$$\int (\nabla_l f(\hat{\theta}_k, \omega_k))^2 dP_{\omega_k} \overset{a.s.}{=} \int (\nabla_l f(\hat{\theta}_k, \omega))^2 dP_{\omega} \tag{A.6}$$

where $\omega = \omega_1$. Then by the boundedness assumption on $\nabla_l f$ and the dominated convergence theorem, Equations (A.5) and (A.6) imply

$$E[(\nabla_l f(\hat{\theta}_k, \omega_k))^2|\mathcal{F}_k] \overset{a.s.}{\longrightarrow} E[(\nabla_l f(\theta^*, \omega))^2] \quad \text{as } k \to \infty$$

and similarly in the off diagonal-elements. Therefore, (2.2.2) of Fabian (1968) is satisfied, and (2.10) holds.

Now we wish to show that (2.2.3) of Fabian (1968) holds. Following the argument in Spall (1992) we have, for $0 < \delta' < \delta/2$,

$$\begin{aligned} &\lim_{k\to\infty} E[1_{\{\|V_k\|^2 \geq rk^\alpha\}}\|V_k\|^2] \\ &\leq \limsup_{k\to\infty} \left(\frac{E\|V_k\|^2}{rk^\alpha}\right)^{\delta'/(1+\delta')} (E\|V_k\|^{2(1+\delta')})^{1/(1+\delta')}. \end{aligned}$$

In our case, we also have

$$\|V_k\|^{2(1+\delta')} \leq 2^{2(1+\delta')}[\|\hat{g}_k(\hat{\theta}_k)\|^{2(1+\delta')} + \|g(\hat{\theta}_k\|^{2(1+\delta')} + \|b_k(\hat{\theta}_k)\|^{2(1+\delta')}].$$

From Assumption (A1) and arguments in the proof of Lemma 1 of Spall (1992), we see that $g(\hat{\theta}_k)$ and $b_k(\hat{\theta}_k)$ are uniformly bounded for large k. Thus, the expected values of the second and third terms above are $O(1)$. Next, similarly to the argument in Spall (1992), assumption (A2), the fact that $0 < \delta' < \delta/2$, and Hölder's inequality imply $E[\|\hat{g}_k(\hat{\theta}_k)\|^{2(1+\delta')}]$ is $O(1)$. Thus $E\|V_k\|^{2(1+\delta')}$ is $O(1)$. This shows that the above limit goes to zero as $k \to \infty$, $\forall r > 0$, and that (2.2.3) of Fabian (1968) is satisfied. □

Proof of Corollary 2.1. First, define $\tilde{b}_k(\hat{\theta}_k)$ to be the conditional bias in $\hat{g}_k(\hat{\theta}_k)$ in a manner analogous to (A.1). Again from

Lemma 1 in Spall (1992), $\tilde{b}_k(\hat{\theta}_k)$ is $O(k^{-2\gamma})$ *a.s.* Then with the notation of Fabian (1968), the convergence of Γ_k, Φ_k, and T_k in the FDSA case are similar to the SPSA case. Furthermore, if $V_k = k^{(\beta-\alpha)/2}[\hat{g}_k(\hat{\theta}_k) - E(\hat{g}_k(\hat{\theta}_k)|\hat{\theta}_k)]$ and $\mathcal{F}_k = \sigma\{\hat{\theta}_1, \ldots, \hat{\theta}_k\}$, then with $\beta = \alpha$,

$$
\begin{aligned}
&E[V_k V_k^T|\mathcal{F}_k]_{ij}\\
&\overset{a.s.}{=} \frac{1}{4c_k^2} E[(f(\hat{\theta}_k + c_k e_i, \omega_k^i) - f(\hat{\theta}_k - c_k e_i, \omega_k^i))\\
&\quad \times (f(\hat{\theta}_k + c_k e_j, \omega_k^j) - f(\hat{\theta}_k - c_k e_j, \omega_k^j))|\mathcal{F}_k] + o(1)\\
&\overset{a.s.}{=} \begin{cases} E[(\nabla_i f(\hat{\theta}_k, \omega_k^i))^2|\mathcal{F}_k] + o(1) & \text{if } i = j\\ \dfrac{1}{4c_k^2}(L(\hat{\theta}_k + c_k e_i) - L(\hat{\theta}_k - c_k e_i))(L(\hat{\theta}_k + c_k e_j) \\ \qquad - L(\hat{\theta}_k - c_k e_j)) + o(1) & \text{if } i \neq j. \end{cases}
\end{aligned}
$$

Therefore, since $(L(\hat{\theta}_k + c_k e_i) - L(\hat{\theta}_k - c_k e_i)) = a.s.\ 2c_k g_i(\hat{\theta}_k) + O(c_k^3)$ and $g(\hat{\theta}_k) \rightarrow a.s.\ 0$, we have, by a dominated convergence theorem argument similar to the one in the SPSA case, that (2.2.2) of Fabian (1968) is satisfied, and (2.11) holds. Proving (2.2.3) of Fabian (1968) holds in the FDSA case is similar to the SPSA case, and the proof of Corollary 2.1 is complete. □

References

Blum, J. R. 1954. Approximation methods which converge with probability one. *Ann. Math. Statist.* **25** 382–386.

Chin, Daniel C. 1997. Comparative study of stochastic algorithms for system optimization based on gradient approximations. *IEEE Trans. Systems, Man, and Cybernetics, Part B* **27** 244–249.

Fabian, Václav. 1968. On asymptotic normality in stochastic approximation. *Ann. Math. Statist.* **39**(4) 1327–1332.

Gal, S., R. Y. Rubinstein, A. Ziv. 1984. On the optimality and efficiency of common random numbers. *Math. Comput. Simulation* **26** 502–512.

Glasserman, Paul, David D. Yao. 1992. Some guidelines and guarantees for common random numbers. *Management Sci.* **38**(6) 884–908.

Iri, Masao. 1991. History of automatic differentiation and rounding error estimation. In Andreas Griewank and George F. Corliss eds. *Automatic Differentiation of Algorithms: Theory, Implementation, and Application.* SIAM, Philadelphia, PA. 3–16.

Juedes, David W. 1991. A taxonomy of automatic differentiation tools. Andreas Griewank and George F. Corliss, eds. *Automatic Differentiation of Algorithms: Theory, Implementation, and Application.* SIAM, Philadelphia, PA. 315–329.

Kiefer, J., J. Wolfowitz. 1952. Stochastic estimation of a regression function. *Ann. Math. Statist.* **23** 462–466.

Kleinman, Nathan L., Stacy D. Hill, Victor A. Ilenda. 1997. SPSA/SIMMOD optimization of air traffic delay cost. *Proc. Amer. Control Conf.* (June) 1121–1125, Albuquerque, New Mexico.

Kushner, Harold J., Dean S. Clark. 1978. *Stochastic Approximation Methods for Constrained and Unconstrained Systems.* Springer-Verlag, New York.

L'Ecuyer, Pierre, Gaëtan Perron. 1994. On the convergence rates of IPA and FDC derivative estimators for finite-horizon stochastic simulations. *Oper. Res.* **42**(4) 643–656.

——, George Yin. 1998. Budget-dependent convergence rate of stochastic approximation. *SIAM J. Optim.* **8** (February) 217–247.

Robbins, H., S. Monro. 1951. A stochastic approximation method. *Ann. Math. Statist.* **22** 400–407.

Sadegh, Payman, James C. Spall. 1998. Optimal random perturbations for stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Automatic Control* **43** 1480–1484. Correction in **44** 231, 1999.

Soulié, Edgar J. 1991. User's experience with FORTRAN compilers in least squares problems. In Andreas Griewank and George F. Corliss, eds. *Automatic Differentiation of Algorithms: Theory, Implementation, and Application.* SIAM, Philadelphia, PA. 297–306.

Spall, James C. 1987. A stochastic approximation technique for generating maximum likelihood parameter estimates. *Proc. Amer. Control Conf.* 1161–1167.

——. 1988. A stochastic approximation algorithm for large-dimensional systems in the Kiefer-Wolfowitz setting. *Proc. IEEE Conf. Decision and Control* 1544–1548.

——. 1992. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Automatic Control* **37**(3) 332–341.

Accepted by Pierre L'Ecuyer, former Departmental Editor; received July 22, 1996. This paper has been with the authors 14 months for 3 revisions.