Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation

James C. Spall, *Senior Member, IEEE*

Abstract—**Consider the problem of finding a root of the multivariate gradient equation that arises in function minimization. When only noisy measurements of the function are available, a stochastic approximation (SA) algorithm of the general Kiefer-Wolfowitz type is appropriate for estimating the root. This paper presents an SA algorithm that is based on a "simultaneous perturbation" gradient approximation instead of the standard finite difference approximation of Kiefer-Wolfowitz type procedures. Theory and numerical experience indicate that the algorithm presented here can be significantly more efficient than the standard finite difference-based algorithms in large-dimensional problems.**

I. Introduction

STOCHASTIC approximation (SA) is a well-known recursive procedure for finding roots of equations in the presence of noisy measurements. Perhaps the most important application of SA is in finding extrema of functions as first described in Kiefer and Wolfowitz [18] for the scalar case and Blum [2] for the multivariate case. This type of SA has potential applications in a number of areas relevant to statistical modeling and control, e.g., sequential parameter estimation, adaptive control, experimental design, stochastic optimization, and neural network weight estimation. This paper describes an SA procedure that has the potential to be significantly more efficient than the usual p-dimensional algorithms (of Kiefer-Wolfowitz/Blum type) that are based on standard finite-difference gradient approximations. It is shown that approximately the same level of estimation accuracy can typically be achieved with only $1/p$th the amount of data needed in the standard approach. The procedure is based on a "simultaneous perturbation" gradient approximation.

Let us now describe the setting. Consider the problem of finding a root θ^* of the gradient equation

$$g(\theta) \equiv \frac{\partial L(\theta)}{\partial \theta} = 0$$

for some differentiable loss function $L: R^p \to R^1$. When L and g are observed directly, there are, of course, many methods for finding θ^* (e.g., steepest descent, Newton-Raphson, scoring). In the case where L is observed in the presence of noise, an SA algorithm of the generic Kiefer-Wolfowitz/Blum type is appropriate (see [25] for a general discussion and related references).

In contrast to SA algorithms based on finite difference methods, which require $2p$ (noisy) measurements of L at each iteration, the "simultaneous perturbation" algorithm here requires only $2q$, $q \geq 1$, measurements of L at each iteration, where for large p we typically have $q \ll p$. Thus there exists the potential for a significant improvement in efficiency provided that the number of iterations does not increase to negate the reduced amount of data per iteration. As we will see this potential can be realized in realistic problems.

The remainder of this paper is organized as follows. Section II presents the simultaneous perturbation gradient approximation and the associated SA algorithm. Section III provides theoretical justification for the algorithm, including results on the strong convergence and asymptotic distribution of the iterate. Section IV discusses the efficiency of the algorithm relative to the multivariate form of the Kiefer-Wolfowitz finite difference algorithm. Section V presents a numerical evaluation of the simultaneous perturbation and finite difference algorithms on a fairly large-dimensional problem and Section VI offers some concluding remarks, including a mention of some areas for future research.

II. The SA Algorithm and Associated Gradient Approximation

This section includes a brief discussion of the SA algorithm that is of interest here and discusses the "simultaneous perturbation" estimate for $g(\theta)$, $\hat{g}(\theta)$, that will be used in the SA algorithm.

Letting $\hat{\theta}_k$ denote the estimate for θ at the kth iteration, the SA algorithm of interest here has the standard form

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{g}_k(\hat{\theta}_k) \tag{2.1}$$

where the gain sequence $\{a_k\}$ satisfies certain well-known conditions (see Section III). Note the close relationship of (2.1) to the method of steepest descent, the difference being that in steepest descent $g(\cdot)$ replaces $\hat{g}_k(\cdot)$.

There are a number of techniques for accelerating the convergence of SA algorithms under special conditions (e.g., second-order algorithms such as that in [26] or adaptive

Manuscript received January 23, 1990; revised June 10, 1990, November 15, 1990, and June 20, 1991. Paper recommended by Associate Editor, B. Pasik-Duncan. This work was supported in part by a JHU/APL Independent Research and Development Grant and by the U.S. Navy under Contract N00039-89-C-5301.

The author is with The Johns Hopkins University, Applied Physics Laboratory, Laurel, MD 20723-6099.

IEEE Log Number 9105739.

0018-9286/92$03.00 © 1992 IEEE

algorithms as in [17] or [19]), but they will not be considered in this paper. Rather, we will focus on the performance of (2.1) with $\hat{g}_k(\cdot)$ as defined in the following, and contrast (in Sections IV and V) this performance with that of the Kiefer–Wolfowitz algorithm where a finite difference gradient approximation replaces $\hat{g}_k(\cdot)$ in (2.1). We believe, however, that this baseline study provides insight into the potential usefulness of $\hat{g}_k(\cdot)$ as it might apply in an accelerated algorithm. The gradient approximation here may also have applications in the field of perturbation analysis since it provides insight into the shape of the performance measure without requiring exact derivatives or a large number of function evaluations (see, e.g., [3], [15], or [14]).

We now define the "simultaneous perturbation" estimate for $g(\cdot)$. Let $\Delta_k \in R^p$ be a vector of p mutually independent mean-zero random variables $\{\Delta_{k1}, \Delta_{k2}, \cdots, \Delta_{kp}\}$ satisfying conditions given in Section III (note: perhaps the main condition is that $E|\Delta_{ki}^{-1}|$, or some higher inverse moment of Δ_{ki}, be bounded, which precludes Δ_{ki} being uniformly or normally distributed). Furthermore, let $\{\Delta_k\}$ be a mutually independent sequence with Δ_k independent of $\hat{\theta}_0, \hat{\theta}_1, \cdots, \hat{\theta}_k$. We need make no assumptions regarding the specific type of distribution for Δ_{ki}, although the numerical studies in Section V take the Δ_{ki} as symmetrically Bernoulli distributed. Consistent with the usual SA framework, we have available noisy measurements of $L(\cdot)$. In particular, at design levels $\hat{\theta}_k \pm c_k \Delta_k$, with c_k a positive scalar, let

$$y_k^{(+)} = L(\hat{\theta}_k + c_k \Delta_k) + \epsilon_k^{(+)}$$

$$y_k^{(-)} = L(\hat{\theta}_k - c_k \Delta_k) + \epsilon_k^{(-)}$$

where $\epsilon_k^{(+)}, \epsilon_k^{(-)}$ represent measurement noise terms that satisfy

$$E(\epsilon_k^{(+)} - \epsilon_k^{(-)} | \mathscr{F}_k, \Delta_k) = 0 \text{ a.s. } \forall k,$$

$$\mathscr{F}_k \equiv \{\hat{\theta}_0, \hat{\theta}_1, \cdots, \hat{\theta}_k\};$$

this condition closely resembles the common martingale difference noise assumption that appears in the literature for the setting where Δ_k is deterministic, differing only in the additional conditioning on Δ_k. The fact that $\{\epsilon_k^{(+)}, \epsilon_k^{(-)}\}$ need not be an independent sequence is of practical concern, e.g., in adaptive control problems (e.g., [32] or [28, pp. 375–376]) as well as in general parameter estimation problems involving the minimization of an integral-based loss function (such as mean-square error) by taking observed values of the integrand (e.g., [7]).

One form for the estimate of $g(\cdot)$ at the kth iteration is then

$$\hat{g}_k(\hat{\theta}_k) = \begin{bmatrix} \dfrac{y_k^{(+)} - y_k^{(-)}}{2c_k \Delta_{k1}} \\ \vdots \\ \dfrac{y_k^{(+)} - y_k^{(-)}}{2c_k \Delta_{kp}} \end{bmatrix}. \tag{2.2}$$

Note that this estimate differs from the usual finite difference approximation in that only two measurements (instead of $2p$) are used. (The name "simultaneous perturbation" as applied to (2.2) arises from the fact that all elements of the $\hat{\theta}_k$ vector are being varied simultaneously.) Aside from evaluating (2.1) with $\hat{g}_k(\cdot)$ as in (2.2), we will also consider using (2.1) with several (conditional on $\mathscr{F}_k$) independent simultaneous perturbation approximations averaged at each iteration. In particular, $\hat{g}_k(\cdot)$ in (2.2) is replaced by

$$\hat{g}_k(\hat{\theta}_k) = q^{-1} \sum_{j=1}^{q} \hat{g}_k^{(j)}(\hat{\theta}_k) \tag{2.3}$$

where each $\hat{g}_k^{(j)}(\cdot)$ is generated as in (2.2) based on a new pair of measurements that are conditionally (on $\mathscr{F}_k$) independent of the other measurement pairs. It will be demonstrated in Sections III, IV, and V that averaging as in (2.3) can sometimes enhance the performance of the SA algorithm (relative to using $\hat{g}_k(\cdot)$ as in (2.2)). Obviously, other averaging methods may also be applicable. It does not appear, however, that the averaging method of Fabian [8] would directly apply since the elements of $\hat{g}_k^{(j)}$ are (conditional on $\mathscr{F}_k$) dependent, violating a key assumption of the Fabian technique.

Gradient approximations similar to (2.2) or (2.3) in the sense that only two measurements are used in the analog to (2.2) have been considered in Kushner and Clark [20, pp. 58–60, 254–256] and Ermoliev [6], [7]. Motivated partly by the problem of weight estimation (learning) in neural networks, Styblinski and Tang [33] consider an algorithm similar to those of Kushner and Clark and Ermoliev for the noise $\equiv 0$ setting and compare this algorithm to one based on simulated annealing. The gradient approximations of these authors have somewhat different forms and regularity conditions than (2.2) and (2.3). Kushner and Clark state that SA with their random directions approximation is *not* superior to SA with a finite difference approximation (i.e., the number of iterations increases enough to nullify the reduced number of measurements per iteration). Some numerical experience described in Section V seems to corroborate this statement; also, theoretical results in [4] comparing asymptotic mean square errors (analogous to results in Section IV here) indicate that random directions SA will not generally be superior to finite difference SA. Note that neither of the gradient approximations, random directions or simultaneous perturbation, is a special case of the other ([4] discusses this further).[1] Ermoliev focuses mainly on how this gradient approximation applies in general stochastic optimization problems; so it is not clear how effective it is in SA algorithms of the type (2.1). As mentioned earlier, Sections IV and V consider the use of (2.2) and (2.3), and finds that they can offer significant savings in data (> order of magnitude) for a moderately large-scale problem with even greater savings possible for larger dimensional problems.

[1] For example, the random directions procedure is based on sampling directions uniformly in a p-dimensional sphere (see [20, pp. 58–60] or [12, pp. 29–31]). In the simultaneous perturbation technique, this type of sampling is forbidden by the regularity conditions.

III. Strong Convergence and Asymptotic Normality

This section presents several results that form the theoretical basis for the simultaneous perturbation SA (SPSA) algorithm. The following sections consider the bias in $\hat{g}_k(\cdot)$ and establish conditions for the strong convergence and asymptotic normality of $\hat{\theta}_k$. The asymptotic normality result allows us to theoretically compare the relative efficiency of SPSA and the multivariate Kiefer-Wolfowitz finite difference SA (FDSA) algorithm, as discussed in Section IV.

Before presenting the main results, note that $E(\hat{g}_k(\hat{\theta}_k) \mid \hat{\theta}_k) = E(\hat{g}_k(\hat{\theta}_k) \mid \mathscr{F}_k)$ a.s. (this follows easily using the fact that Δ_k is independent of $\mathscr{F}_k$ and fact that $E|\Delta_{ki}^{-1}|$ exists, which implies that $P(\Delta_{ki} = 0) = 0$). For convenience in this section and in Section V, we will frequently write $\overline{\Delta}_k$ for $c_k \Delta_k$, with corresponding elements $\overline{\Delta}_{ki}$. All norms $\|\cdot\|$ are taken as the Euclidean norm. We also focus on the $q = 1$ case in the proofs; trivial modifications accommodate the $q > 1$ case.

A. The Bias in $\hat{g}(\cdot)$

Lemma 1 below gives conditions under which the bias in $\hat{g}_k(\cdot)$ as an estimator of $g(\cdot)$ goes to 0 as $k \to \infty$. An explicit bound for the bias is given in expression (3.2); α_0, α_1, and α_2 will denote positive constants, and $\Omega = \{\omega\}$ will denote the sample space generating the sequence $\hat{\theta}_1, \hat{\theta}_2, \cdots$.

Lemma 1: Consider all $k \geq K$ for some $K < \infty$. Suppose that for each such k the $\{\Delta_{ki}\}$ are i.i.d. ($i = 1, 2, \cdots, p$) and symmetrically distributed about 0 with $|\Delta_{ki}| \leq \alpha_0$ a.s. and $E|\Delta_{ki}^{-1}| \leq \alpha_1$. For almost all $\hat{\theta}_k$ (at each $k \geq K$) suppose that $\forall\ \theta$ 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 $|L^{(3)}_{i_1 i_2 i_3}(\theta)| \leq \alpha_2$. Then for almost all $\omega \in \Omega$

$$\begin{aligned} b_k(\hat{\theta}_k) &\equiv E(\hat{g}_k(\hat{\theta}_k) - g(\hat{\theta}_k) \mid \hat{\theta}_k) \\ &\quad (= E(\hat{g}_k(\hat{\theta}_k) - g(\hat{\theta}_k) \mid \mathscr{F}_k)) \\ &= O(c_k^2)\ (c_k \to 0). \end{aligned}$$

Proof: Consider any $l \in \{1, 2, \cdots, p\}$. First, note that $E[\epsilon_k^{(+)} - \epsilon_k^{(-)})/2\overline{\Delta}_{kl} \mid \hat{\theta}_k] = 0$ a.s. Then by the continuity of $L^{(3)}$ near $\hat{\theta}_k$ and uniform boundedness of $|\Delta_{ki}|$ for all k sufficiently large, we have by Taylor's theorem for all such k

$$\begin{aligned} b_{kl}(\hat{\theta}_k) = \frac{1}{12} E\{\overline{\Delta}_{kl}^{-1}[& L^{(3)}(\overline{\theta}_k^{+}) \\ & + L^{(3)}(\overline{\theta}_k^{-})] \overline{\Delta}_k \otimes \overline{\Delta}_k \otimes \overline{\Delta}_k \mid \hat{\theta}_k\} \end{aligned} \tag{3.1}$$

where $\overline{\theta}_k^{+}, \overline{\theta}_k^{-}$ are on the line segment between $\hat{\theta}_k$ and $\hat{\theta}_k \pm \overline{\Delta}_k$, respectively, and b_{kl} denotes the lth term of the bias b_k. By the mean value theorem, the term on the r.h.s. of (3.1), is bounded in magnitude by

$$\begin{aligned} \frac{\alpha_2 c_k^2}{6} \sum_{i_1} \sum_{i_2} \sum_{i_3} E\left|\frac{\Delta_{ki_1} \Delta_{ki_2} \Delta_{ki_3}}{\Delta_{kl}}\right| &\leq \frac{\alpha_2 c_k^2}{6} \\ \cdot \{[p^3 - (p-1)^3]\alpha_0^2 + (p-1)^3 \alpha_1 \alpha_0^3\} \end{aligned} \tag{3.2}$$

where the upper bound follows from the fact that $(p-1)^3$ summands on the l.h.s. will have no Δ_{kl} term in the numerator. Combining (3.1) and (3.2) completes the proof. Q.E.D.

Note: The randomness in $\{\Delta_{ki}\}$ plays a critical role in ensuring that the bias in $\hat{g}_k(\cdot)$ is $O(c_k^2)\ \forall\ q = 1, 2, \cdots$. (This follows since terms of the form Δ_{ki}/Δ_{kl} that arise in the expansion $b_{kl}(\theta) = E(\sum_{i \neq l} g_i(\theta) \Delta_{ki}/\Delta_{kl}) + O(c_k^2)$ have expectation 0, thereby removing the $O(1)$ contribution to the bias that would otherwise result.) It is, however, possible to construct gradient estimators similar in spirit to SP that have deterministic perturbations and have $O(c_k^2)$ bias if q is large enough (usually $\gg 1$). One such estimator is based on picking $\Delta_k^{(j)}$ for use in (2.2) and (2.3) as $\Delta_k^{(j)} = (\pm\delta, \pm\delta, \cdots, \pm\delta, \delta)^T$, δ a positive constant, where the $\pm$ indicates that we are taking all possible combinations of δ and $-\delta$ in the first $p - 1$ elements of the perturbation vector as j ranges from 1 to q. Thus we need $q = 2^{p-1}$, which is unacceptable for large p. An alternate implementation of this deterministic perturbation idea (suggested by a reviewer of this paper) is to use $q = 1$ and cycle through all 2^{p-1} directions over 2^{p-1} successive iterations. Although $\hat{g}_k(\cdot)$ is $O(1)$ biased in this approach, it might be hoped that the biases would tend to cancel each other over the block of 2^{p-1} iterations. This cancellation is easily seen to be highly unlikely when p is large since it depends on $\hat{\theta}_k$ and a_k being approximately equal over the block. Without bias cancellation, certain elements of $\hat{\theta}_k$ could get repeatedly pushed in a (wrong) direction to a point from which they may not be able to recover.[2]

B. Strong Convergence of $\hat{\theta}_k$

We now present Proposition 1, which establishes conditions under which $\hat{\theta}_k$ converges almost surely to θ^*. Defining the error term

$$e_k(\hat{\theta}_k) = \hat{g}_k(\hat{\theta}_k) - E(\hat{g}_k(\hat{\theta}_k) \mid \hat{\theta}_k)$$

we can rewrite (2.1) as

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k[g(\hat{\theta}_k) + b_k(\hat{\theta}_k) + e_k(\hat{\theta}_k)]$$

which is in the form of a generalized Robbins-Monro algorithm considered, e.g., in [20, pp. 38-39], [23], or [24].

Let us introduce the following assumptions, which are very similar to those of a number of other authors, as discussed below.

A1: $a_k, c_k > 0\ \forall\ k$; $a_k \to 0, c_k \to 0$ as $k \to \infty$;

$$\sum_{k=0}^{\infty} a_k = \infty, \quad \sum_{k=0}^{\infty} \left(\frac{a_k}{c_k}\right)^2 < \infty.$$

[2] The randomness in the perturbations (especially, independence and symmetry of $\{\Delta_{ki}\}$) is also critical in the asymptotic distribution theory of Section III-C in providing an invertible asymptotic covariance matrix for $\hat{\theta}_k$. Under the bounded moment conditions given there, cov $\hat{g}_k(\theta)$ is invertible even when $q = 1$, which via (3.5) and (3.6) ensures that the asymptotic covariance matrix for $\hat{\theta}_k$ is also invertible. For any deterministic perturbations of the SP type (irrespective of bias considerations), a *minimum* of $q = p$ is needed to achieve this invertibility.

A2: For some α_0, α_1, $\alpha_2 > 0$ and $\forall\, k$, $E\epsilon_k^{(\pm)^2} \leq \alpha_0$, $EL(\hat{\theta}_k \pm \bar{\Delta}_k)^2 \leq \alpha_1$, and $E\,\Delta_{kl}^{-2} \leq \alpha_2$ $(l = 1, 2, \cdots, p)$.

A3: $\|\hat{\theta}_k\| < \infty$ a.s. $\forall\, k$.

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

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

Remarks on regularity conditions: A1 and A2 are typical SA conditions. A3 is perhaps the most difficult to verify in practice. Kushner and Clark [20, pp. 40–41] discuss why A3 is, in fact, not a restrictive condition and could be expected to hold in most applications.[3] A4 and A5 are motivated by considering a limiting form of the deterministic version of (2.1), i.e., $\hat{\theta}_{k+1} = \hat{\theta}_k - a_k g(\hat{\theta}_k)$ as $k \to \infty$. Lai [22] presents a brief tutorial on the role of A4 and A5 in SA algorithms. See also the discussion in [23, Section 6]. Slightly weaker assumptions than A4 and A5 (leading to convergence to different roots θ^* along different sample paths) are discussed in [24] and [1].

Proposition 1: Let A1–A5 and the conditions of Lemma 1 hold. Then as $k \to \infty$

$$\hat{\theta}_k \to \theta^* \qquad \text{for almost all } \omega \in \Omega. \tag{3.3}$$

Proof: Given A1 and A3–A5, we know from [20, Lemma 2.2.1 and Theorem 2.3.1] (see also [24] or [22]), that (3.3) holds if

i)

$$\|b_k(\hat{\theta}_k)\| < \infty\ \forall\, k \text{ and } b_k(\hat{\theta}_k) \to 0 \text{ a.s.},$$

ii)

$$\lim_{k\to\infty} P\left(\sup_{m\geq k} \Big\|\sum_{i=k}^{m} a_i e_i(\hat{\theta}_i)\Big\| \geq \eta\right) = 0 \qquad \text{for any } \eta > 0.$$

Now, i) follows immediately by Lemma 1, (3.2), and A1.

Consider ii). Since $\{\sum_{i=k}^{m} a_i e_i\}_{m\geq k}$ is a martingale sequence, we have from an inequality in [5, p. 315] (see also [20, p. 27])

$$P\left(\sup_{m\geq k} \Big\|\sum_{i=k}^{m} a_i e_i\Big\| \geq \eta\right) \leq \eta^{-2} E\Big\|\sum_{i=k}^{\infty} a_i e_i\Big\|^2$$

$$= \eta^{-2} \sum_{i=k}^{\infty} a_i^2 E\|e_i\|^2 \tag{3.4}$$

where the equality follows by the fact that $E(e_i^T e_j) = E(e_i^T E(e_j \mid \hat{\theta}_j)) = 0\ \forall\ i < j$. Now for any $l \in \{1, 2, \cdots, p\}$, we have

$$\begin{aligned} E\hat{g}_{kl}(\hat{\theta}_k)^2 &\leq \tfrac{1}{4} E\big[L(\hat{\theta}_k + \bar{\Delta}_k) - L(\hat{\theta}_k - \bar{\Delta}_k) \\ &\qquad + \epsilon_k^{(+)} - \epsilon_k^{(-)}\big]^2 E\bar{\Delta}_{kl}^{-2} \\ &\leq 2(\alpha_1 + \alpha_0)\alpha_2 c_k^{-2} \qquad \text{(using A2)} \end{aligned}$$

and so $E\|e_k\|^2 \leq 2p(\alpha_1 + \alpha_0)\alpha_2 c_k^{-2}$. Then from (3.4) and A1, ii) has been shown, which completes the proof. Q.E.D.

C. Asymptotic Normality of $\hat{\theta}_k$

Using a result of Fabian [9], Proposition 2 below establishes asymptotic normality for scaled $\hat{\theta}_k$. Section IV comments on how this result can be used to draw conclusions about the relative efficiency of SPSA and FDSA.

Proposition 2 considers gains of the standard form $a_k = a/k^\alpha$ and $c_k = c/k^\gamma$ where $a, c, \alpha, \gamma > 0$. The proposition is stated for the case where $\hat{g}_k(\cdot)$ is not formed from an average of several $\hat{g}_k^{(j)}(\cdot)$'s (i.e., $q = 1$ in (2.3)); the note after the proof considers the $q \geq 2$ case. The proposition also relies on the following strengthened version of condition A2 of Proposition 1.

A2′: For some $\delta, \alpha_0, \alpha_1, \alpha_2 > 0$ and $\forall\ k$, $E|\epsilon_k^{(\pm)}|^{2+\delta} \leq \alpha_0$, $E|L(\hat{\theta}_k \pm \bar{\Delta}_k)|^{2+\delta} \leq \alpha_1$, and $E|\Delta_{kl}|^{-2-\delta} \leq \alpha_2$ $(l = 1, 2, \cdots, p)$.

Finally, let $H(\theta)$ denote the Hessian matrix for $L(\theta)$; notation of [9] will also be used when appropriate.

Proposition 2: Assume that the conditions of Lemma 1 and Proposition 1 hold but with A2 strengthened to A2′. Let σ^2, ρ^2, and ξ^2 be such that $E[(\epsilon_k^{(+)} - \epsilon_k^{(-)})^2 \mid \mathscr{F}_k| \to \sigma^2$ a.s., $E\Delta_{kl}^{-2} \to \rho^2$, and $E\Delta_{kl}^2 \to \xi^2$ as $k \to \infty\ \forall\ l$. Also, $\forall\ k$ sufficiently large and almost all ω let the sequence $\{E[(\epsilon_k^{(+)} - \epsilon_k^{(-)})^2 \mid \mathscr{F}_k, \bar{\Delta}_k = \eta]\}$ be equicontinuous at $\eta = 0$ and continuous in η on some compact, connected set containing $\bar{\Delta}_k$ a.s.[4] Furthermore, let $\beta = \alpha - 2\gamma > 0$, $3\gamma - \alpha/2 \geq 0$, and P be orthogonal with $PH(\theta^*)P^T = a^{-1}$ diag $(\lambda_1, \cdots, \lambda_p)$. Then

$$k^{\beta/2}(\hat{\theta}_k - \theta^*) \overset{\text{dist}}{\to} N(\mu, PMP^T), \qquad k \to \infty$$

where $M = \frac{1}{4}a^2 c^{-2}\sigma^2\rho^2$ diag $[(2\lambda_1 - \beta_+)^{-1}, \cdots, (2\lambda_p - \beta_+)^{-1}]$ with $\beta_+ = \beta < 2\min_i \lambda_i$ if $\alpha = 1$ and $\beta_+ = 0$ if $\alpha < 1$, and

$$\mu = \begin{cases} 0 & \text{if } 3\gamma - \alpha/2 > 0 \\ \big(aH(\theta^*) - \tfrac{1}{2}\beta_+ I\big)^{-1} T & \text{if } 3\gamma - \alpha/2 = 0 \end{cases}$$

where the lth element of T is

$$-\tfrac{1}{6}ac^2\xi^2\left[L_{lll}^{(3)}(\theta^*) + 3\sum_{\substack{i=1 \\ i\neq l}}^{p} L_{iil}^{(3)}(\theta^*)\right].$$

[3] Ljung [23] replaces A3 with a weaker condition, $\liminf_{k\to\infty}\|\hat{\theta}_k\| < \infty$ a.s., and gives sufficient conditions for this weaker condition to hold. This author found, however, that certain other conditions were not as easily verified as those of Kushner and Clark [20], Lai [22], or Metivier and Priouret [24], which will be used in the proof of Proposition 1.

[4] We emphasize that in the equicontinuity and continuity assumptions, $\mathscr{F}_k$ depends on the actual values of $\bar{\Delta}_i$, not on $\bar{\Delta}_i = \eta$, $\forall\ i < k$.

Proof: The result will be shown if conditions (2.2.1), (2.2.2), and (2.2.3) of Fabian [9] hold; all statements conditioned on $\hat{\theta}_k$ or $\mathscr{F}_k$ are assumed to hold a.s. (on Ω). By Lemma 1, we know that ∃ an open neighborhood of $\hat{\theta}_k$ ($\forall$ k sufficiently large) such that $H(\cdot)$ is continuous in the neighborhood, and by Proposition 1 we know for large enough k that θ^* lies in this neighborhood. Thus

$$E[\hat{g}_k(\hat{\theta}_k) \mid \hat{\theta}_k] = H(\bar{\theta}_k)(\hat{\theta}_k - \theta^*) + b_k(\hat{\theta}_k) \; (= E[\hat{g}_k(\hat{\theta}_k) \mid \mathscr{F}_k])$$

where $\bar{\theta}_k$ is on the line segment between $\hat{\theta}_k$ and θ^*. In the notation of [9] we can now write

$$\hat{\theta}_{k+1} - \theta^* = (I - k^{-\alpha}\Gamma_k)(\hat{\theta}_k - \theta^*) + k^{-(\alpha+\beta)/2}\Phi_k V_k + k^{-\alpha-\beta/2} T_k$$

where $\Gamma_k = aH(\bar{\theta}_k)$, $V_k = k^{-\gamma}[\hat{g}_k(\hat{\theta}_k) - E(\hat{g}_k(\hat{\theta}_k) \mid \hat{\theta}_k)]$, $\Phi_k = -aI$, and $T_k = -ak^{\beta/2} b_k(\hat{\theta}_k)$. We now show that Γ_k, T_k, and $E(V_k V_k^T \mid \mathscr{F}_k)$ are a.s. convergent, i.e., conditions (2.2.1) and (2.2.2) of [9] hold.

We see that $\Gamma_k \to aH(\theta^*)$ a.s. by the continuity of $H(\cdot)$ and a.s. convergence of $\hat{\theta}_k$. Now consider the convergence of T_k. If $3\gamma - \alpha/2 > 0$, we see that $T_k \to 0$ a.s. by the fact that $b_k(\hat{\theta}_k) = O(k^{-2\gamma})$ a.s. If $3\gamma - \alpha/2 = 0$, (3.1) and the facts that $\hat{\theta}_k \to \theta^*$ a.s. and $L^{(3)}$ is uniformly bounded near θ^*, imply by the dominated convergence theorem

$$k^{2\gamma} b_{kl}(\hat{\theta}_k) - \tfrac{1}{6}c^2 L^{(3)}(\theta^*) E[\Delta_{kl}^{-1}(\Delta_k \otimes \Delta_k \otimes \Delta_k)] \to 0 \quad \text{a.s.}$$

Then the fact that $\{\Delta_{ki}\}$ are symmetrically i.i.d. for each k implies that the lth element of T_k satisfies

$$T_{kl} \to -\tfrac{1}{6}ac^2\xi^2 \left\{ L_{lll}^{(3)}(\theta^*) + \sum_{\substack{i=1 \\ i \neq l}}^{p} \left[L_{lii}^{(3)}(\theta^*) + L_{ili}^{(3)}(\theta^*) + L_{iil}^{(3)}(\theta^*) \right] \right\} \quad \text{a.s.}$$

(The value for $\mu \neq 0$ in the proposition statement uses the fact that $L_{lii}^{(3)} = L_{ili}^{(3)} = L_{iil}^{(3)}$ at θ^* by continuity, which is guaranteed by the uniform conditions in Lemma 1 and a.s. convergence of $\hat{\theta}_k$.) We have thus shown that T_k converges for $3\gamma - \alpha/2 \geq 0$. Now define $\Delta_k^{-1} = (\Delta_{k1}^{-1}, \cdots, \Delta_{kp}^{-1})^T$. We have

$$\begin{aligned} E(V_k V_k^T \mid \mathscr{F}_k) = {} & k^{-2\gamma} E\left\{ \Delta_k^{-1}(\Delta_k^{-1})^T \cdot \left[\frac{L(\hat{\theta}_k + \bar{\Delta}_k) - L(\hat{\theta}_k - \bar{\Delta}_k)}{2ck^{-\gamma}} \right]^2 \middle| \hat{\theta}_k \right\} \\ & + k^{-2\gamma} E\left\{ \Delta_k^{-1}(\Delta_k^{-1})^T \left[\frac{\epsilon_k^{(+)} - \epsilon_k^{(-)}}{2ck^{-\gamma}} \right] \cdot \left[\frac{L(\hat{\theta}_k + \bar{\Delta}_k) - L(\hat{\theta}_k - \bar{\Delta}_k)}{2ck^{-\gamma}} \right] \middle| \mathscr{F}_k \right\} \\ & + k^{-2\gamma} E\left\{ \Delta_k^{-1}(\Delta_k^{-1})^T \left[\frac{\epsilon_k^{(+)} - \epsilon_k^{(-)}}{2ck^{-\gamma}} \right]^2 \middle| \mathscr{F}_k \right\} \\ & - k^{-2\gamma}[g(\hat{\theta}_k) + b_k(\hat{\theta}_k)][g(\hat{\theta}_k) + b_k(\hat{\theta}_k)]^T. \end{aligned} \tag{3.5}$$

Since the conditions of Lemma 1 imply that for all k sufficiently large and almost all $\omega \in \Omega$ $L(\hat{\theta}_k \pm \bar{\Delta}_k)$ is uniformly bounded in Δ_k, A2′, Holder's inequality, and the dominated convergence theorem imply that the first and second terms on the RHS of (3.5) approach 0 a.s. Also, (3.2) implies the same for the fourth term. Now, let us consider the third term in (3.5). By the independence of Δ_k and $\mathscr{F}_k$

$$E\left[\Delta_k^{-1}(\Delta_k^{-1})^T(\epsilon_k^{(+)} - \epsilon_k^{(-)})^2 \mid \mathscr{F}_k\right] = \int_{\Omega_\Delta} \Delta_k^{-1}(\Delta_k^{-1})^T E\left[(\epsilon_k^{(+)} - \epsilon_k^{(-)})^2 \mid \mathscr{F}_k, \bar{\Delta}_k\right] dP_\Delta \tag{3.6}$$

where Ω_Δ is the sample space generating the Δ_k's and P_Δ is the corresponding probability measure. By the fact that $E[(\epsilon_k^{(+)} - \epsilon_k^{(-)})^2 \mid \mathscr{F}_k] \to \sigma^2$ a.s., we know by the mean value theorem and by the equicontinuity in η at 0 that $E[(\epsilon_k^{(+)} - \epsilon_k^{(-)})^2 \mid \mathscr{F}_k, \bar{\Delta}_k = \eta_k] \to \sigma^2$ a.s. $\forall$ sequences $\{\eta_k\}$ in the assumed compact, connected sets containing $\{\bar{\Delta}_k\}$ a.s. such that $\eta_k \to 0$. Then applying the mean value theorem to each of the diagonal and off-diagonal elements in (3.6) yields

$$E(V_k V_k^T \mid \mathscr{F}_k) \to \tfrac{1}{4}c^{-2}\sigma^2\rho^2 I \qquad \text{a.s.}$$

This completes the proof of Fabian's conditions (2.2.1) and (2.2.2).

We now show that condition (2.2.3) of Fabian [9] holds, which is

$$\lim_{k\to\infty} E(\mathscr{I}_{\{\|V_k\|^2 \geq rk^\alpha\}} \|V_k\|^2) = 0 \qquad \forall\, r > 0$$

where $\mathscr{I}_{\{\cdot\}}$ denotes the indicator for $\{\cdot\}$. By Holder's inequality and for any $0 < \delta' < \delta/2$, the above limit is bounded above by

$$\begin{aligned} & \limsup_{k\to\infty} P(\|V_k\|^2 \geq rk^\alpha)^{\delta'/(1+\delta')} (E\|V_k\|^{2(1+\delta')})^{1/(1+\delta')} \\ & \leq \limsup_{k\to\infty} \left(\frac{E\|V_k\|^2}{rk^\alpha} \right)^{\delta'/(1+\delta')} (E\|V_k\|^{2(1+\delta')})^{1/(1+\delta')}. \end{aligned} \tag{3.7}$$

Note that

$$\|V_k\|^{2(1+\delta')} \leq 2^{2(1+\delta')} k^{-2(1+\delta')\gamma} [\|\hat{g}(\hat{\theta}_k)\|^{2(1+\delta')} + \|g(\hat{\theta}_k)\|^{2(1+\delta')} + \|b_k(\hat{\theta}_k)\|^{2(1+\delta')}]. \tag{3.8}$$

From Lemma 1 (including (3.2)), $g(\theta_k)$ and $b_k(\theta_k)$ are uniformly bounded for almost all $\omega \in \Omega$ and $\forall k \geq K$; $L(\hat{\theta}_k \pm \bar{\Delta}_k)$ is uniformly bounded for almost all $(\omega, \omega_\Delta) \in \Omega \times \Omega_\Delta$

and $\forall\ k \geq K_0$ where $K_0 \geq K$. Thus, the expectation of the second and third terms within $[\cdot]$ in (3.8) approaches 0 as $k \to \infty$. Further, invoking A2$'$ and the fact that $\delta' < \delta/2$, Holder's inequality implies that the expected value of the first term in $[\cdot]$ is $O(k^{2(1+\delta')\gamma})$. Thus $E\|V_k\|^{2(1+\delta')} = O(1)$, which shows by (3.7) that (2.2.3) of [9] has been proved. Q.E.D.

Note: Proposition 2 is stated for the case where $q = 1$ in (2.3). If $q \geq 2$ and a, c are held constant asymptotic normality also holds, but with σ^2 (in M) replaced by σ^2/q. With $q \geq 2$ the asymptotic mean-squared error (with the same number of measurements y) can generally be reduced. Let us assume that the first and second moments of the asymptotic distribution correspond to the first two asymptotic moments of $k^{\beta/2}(\hat{\theta}_k - \theta^*)$, which is true if $\|k^{\beta/2}(\hat{\theta}_k - \theta^*)\|^2$ is uniformly integrable (see, e.g., [21, pp. 138–140]). Then, letting n denote the number of measurements and recalling that $k = n/2q$, the mean square error (MSE) matrix for $\hat{\theta}_k - \theta^*$ is asymptotic to

$$k^{-\beta}\left(\frac{PMP^T}{q} + \mu\mu^T\right) = \left(\frac{2}{n}\right)^{\beta}\left(q^{\beta-1}PMP^T + q^{\beta}\mu\mu^T\right). \tag{3.9}$$

Straightforward calculations yield an asymptotically optimal q, that is, one that minimizes $E\|\hat{\theta}_k - \theta^*\|^2$ based on (3.9), as the integer either immediately above or below $(1 - \beta) \cdot \text{tr}\, PMP^T/\beta\mu^T\mu$ when $\mu \neq 0$.[5]

IV. Relative Efficiency of Simultaneous Perturbation SA and Finite Difference SA

To gain some insight into the relative performance of the SPSA algorithm and the FDSA algorithm this section compares the result of Proposition 2 with the analogous result for FDSA. In particular, we focus on relative asymptotic mean square error (MSE). Under fairly general conditions, it is shown that SPSA will achieve lower MSE than FDSA for the same amount of data, which is equivalent to SPSA using less data to achieve the same level of MSE as FDSA.

In the FDSA algorithm, the finite difference gradient estimate, say $\tilde{g}_k(\cdot)$, will replace $\hat{g}_k(\cdot)$ in (2.1). In particular the lth component of $\tilde{g}_k(\cdot)$, $l = 1, 2, \cdots, p$, is given by

$$\tilde{g}_{kl}(\tilde{\theta}_k) = \frac{L(\tilde{\theta}_k + c_k u_l) + \epsilon_k^{(l+)} - L(\tilde{\theta}_k - c_k u_l) - \epsilon_k^{(l-)}}{2c_k}$$

where $\tilde{\theta}_k$ denotes the FDSA estimate at the kth iteration, u_l denotes a unit vector in the direction of the lth coordinate in R^p and $\epsilon_k^{(l+)}, \epsilon_k^{(l-)}$ denote measurement noise terms with $\epsilon_k^{(l+)} - \epsilon_k^{(l-)}$ satisfying the usual martingale difference condition (so $\tilde{g}_{kl}(\tilde{\theta}_k)$ is based on measurements at design levels $\tilde{\theta}_k \pm c_k u_l$).

First, note that, as with Lemma 1 and Proposition 1, $\tilde{g}_k(\cdot)$ has $O(c_k^2)$ bias and $\tilde{\theta}_k \to \theta^*$ a.s. (see, e.g., [20, pp. 51–52]). Using these results it can be shown (in the manner of Proposition 2) that

$$k^{\beta/2}(\tilde{\theta}_k - \theta^*) \overset{\text{dist}}{\to} N(\tilde{\mu}, P\tilde{M}P^T)$$

where P is as in Proposition 2, $\tilde{M} = \frac{1}{4}a^2c^{-2}\tilde{\sigma}^2\, \text{diag}\,[(2\lambda_1 - \beta_+)^{-1}, \cdots, (2\lambda_p - \beta_+)^{-1}]$, $E[(\epsilon_k^{(l+)} - \epsilon_k^{(l-)})^2 \mid \mathscr{F}_k] \to \tilde{\sigma}^2$ a.s. $\forall\ l$, and

$$\tilde{\mu} = \begin{cases} 0 & \text{if } 3\gamma - \alpha/2 > 0 \\ \left(aH(\theta^*) - \frac{1}{2}\beta_+ I\right)^{-1}\tilde{T} & \text{if } 3\gamma - \alpha/2 = 0 \end{cases}$$

where the lth element of $\tilde{T}$ is $-\frac{1}{6}ac^2L_{lll}^{(3)}(\theta^*)$.

Let us now discuss the relative MSE for SPSA and FDSA for an identical number of measurements n (not iterations). Taking $\sigma^2 = \tilde{\sigma}^2$, and invoking the caveat in the note after Proposition 2 regarding the moments of the asymptotic distribution corresponding to the asymptotic moments of scaled $\hat{\theta}_k$ and $\tilde{\theta}_k$, we have as $n \to \infty$

$$\frac{E\|\hat{\theta}_k - \theta^*\|^2}{E\|\tilde{\theta}_{k'} - \theta^*\|^2} \to \left(\frac{q}{p}\right)^{\beta}\frac{\text{tr}\, PMP^T/q + \mu^T\mu}{\text{tr}\, P\tilde{M}P^T + \tilde{\mu}^T\tilde{\mu}} \tag{4.1}$$

where $k = n/2q$ and $k' = n/2p$.

Broadly speaking, we can say that if $\|\mu\| \approx \|\tilde{\mu}\|$ as p gets large, then the ratio in (4.1) is $O(1/p^{\beta})$, indicating that SPSA is the superior algorithm for large enough p. However, if $\|\tilde{\mu}\|/\|\mu\|$ gets small as p gets large (as would happen if $3\gamma - \alpha/2 = 0$ and the $\{L_{iil}^{(3)}(\theta^*)\}$ tended to be of the same nonzero magnitude and sign $\forall\ i, l$), then the ratio in (4.1) might tend to favor FDSA. The first case (i.e., that favoring SPSA) seems to arise more frequently in practice, especially in light of the common setting where $\mu = \tilde{\mu} = 0$. It does not seem possible to make more general statements about the relative superiority of SPSA or FDSA since: i) ρ^2 and ξ^2, which enter M and μ, are functions of the distribution chosen for the $\{\Delta_{ki}\}$ and ii) the mixed third partials of L may or may not reduce the $\mu^T\mu$ contribution to MSE relative to the $\tilde{\mu}^T\tilde{\mu}$ contribution (where no mixed partials are present); this obviously depends on the function L being considered.

Nevertheless, we can illustrate an application of (4.1) by assuming that $\rho^2 = \xi^2 = 1$ (as in the Bernoulli ± 1 examples of Section V) and $\mu \approx \tilde{\mu}$ (e.g., as when the magnitudes of the mixed partial $L_{iil}^{(3)}(\theta^*)$ terms, $i \neq l$, are small relative to the $L_{lll}^{(3)}(\theta^*)$ terms $\forall\ l$ or when $3\gamma - \alpha/2 > 0$ [so $\mu = \tilde{\mu} = 0$]). Then the ratio in (4.1) is approximately (or exactly if $\mu = \tilde{\mu}$) equal to $1/p^{\beta}$ when $q = 1$ and both algorithms are run with the same values of a, c (even those chosen optimally for FDSA, such as in [10]).[6] Thus, when $\beta = 2/3$ (as in Kush-

[5] In contrast to the aforementioned, where a and c are held fixed, it can be shown (as in [20, pp. 253–254] or [10]) that if $\beta = 2/3$ and c is chosen to minimize the MSE (as a function of q), then averaging does *not* reduce the MSE. This follows from a calculation effect between the increased quality of estimate for $g(\cdot)$ and the decreased number of iterations.

[6] Choosing $\rho^2 = 1$ forces the *effective* measurement noise in SPSA and FDSA (i.e., measurement noise$/2c_k\Delta_{kl}$ and measurement noise$/2c_k$) to have equal variance in the important special case where Δ_k is independent of the noise. This provides a fair basis for comparing the algorithms since the effective measurement noise plays a major role in the stability/convergence of the algorithm. An alternate approach for the Bernoulli setting (which the discussion above and in the previous sections does not require) is to choose the Δ_{kl} magnitude such that the length of the SPSA and FDSA difference intervals for the gradient approximations, $2c_k\|\Delta_k\|$ and $2c_k$, are equal; here we choose $|\Delta_{kl}| = 1/\sqrt{p}$ (now, however, the effective measurement noise variance in SPSA is p times larger than that for FDSA). From (4.1) it follows that when $\mu \neq 0$ the MSE ratio is $O(1/p^{2+\beta})$ $(p \to \infty)$ but that when $\mu = 0$ the ratio equals $p^{1-\beta} > 1$. This is unlike the $|\Delta_{kl}| = 1$ case, where SPSA generally has lower MSE than FDSA in both the zero and nonzero μ settings.

ner and Clark [20, pp. 252–253] and Fabian [10]) and $p = 20$, as in the examples of Section 5, FDSA has asymptotic MSE approximately 7.4 times larger than SPSA when $q = 1$. For other values of q the relative efficiency of SPSA to FDSA is likely to increase depending on the relative magnitudes of tr PMP^T (= tr $P\tilde{M}P^T$) and $\mu^T\mu(\approx \tilde{\mu}^T\tilde{\mu})$. We will see in Section V some other examples where the ratio in (4.1) can be readily computed. In particular, when $3\gamma - \alpha/2 > 0$ (so $\mu = \tilde{\mu} = 0$) and $\rho^2 = \xi^2 = 1$ (so $M = \tilde{M}$) the ratio simplifies to $q^{\beta-1}/p^{\beta}$ for any common values of a and c.

An equivalent, but enlightening, way of looking at the relative MSE is to compare the number of measurements needed in FDSA to achieve the same level of MSE as SPSA. Under the earlier-mentioned assumptions that $\mu = \tilde{\mu}$ (at least approximately) and $M = \tilde{M}$, we have by setting the right-hand side of (4.1) to unity

$$\frac{\#\text{ measurements in SPSA}}{\#\text{ measurements in FDSA}} \to \frac{q}{p}\left(\frac{\operatorname{tr} PMP^T/q + \mu^T\mu}{\operatorname{tr} PMP^T + \mu^T\mu}\right)^{1/\beta} \tag{4.2}$$

as the number of measurements for both procedures gets large. Since β and the term inside $(\cdot)$ are both ≤ 1 we know that (asymptotically) FDSA will require at least p/q times more measurements than SPSA. More specifically if $\mu = 0$, at least p times more measurements are needed in FDSA.

V. Numerical Evaluation of the Algorithm

A. Introduction

This section presents the results of a study that illustrate how the SA algorithm with the simultaneous perturbation gradient approximation of (2.2) or (2.3) compares to the finite difference procedure. The study was performed on an IBM 3091 mainframe with MVS operating system; pseudorandom numbers were generated using the IMSL DRNNOR normal random number generator and a standard uniform pseudorandom algorithm based on the remainder of the division of two large numbers.

The function we seek to minimize in this study is

$$L(\theta) = \sum_{i=1}^{N}\left[\log\det\left(\Sigma(\theta) + Q_i\right) + x_i^T\left(\Sigma(\theta) + Q_i\right)^{-1}x_i\right]$$

where $\Sigma(\theta) = \operatorname{diag}(\theta_1, \theta_2, \cdots, \theta_p)$, Q_i is positive semidefinite $\forall\ i$, and $x_i \in R^p$ are constants.[7] Recall that we have noisy observations of the function $L(\cdot)$ at various design levels. It is of interest to note that the function $L(\theta)$ arises in a signal-plus-noise MLE problem (where the x_i represent data distributed $N(0, \Sigma + Q_i)$), and that it has applications, e.g., in Kalman filter model estimation (the author's primary interest) and dose response curve estimation (see, e.g., [29] and [16]).[8] More relevant for our purposes here is the fact that L is a function $R^p \to R^1$ for which third-order (actually any order) derivatives exists continuously on $(0, \infty) \times (0, \infty) \times \cdots \times (0, \infty)$, and for which $L^{(3)}(\theta)$ is uniformly bounded in magnitude on any compact subset of this domain (in particular, for θ_i's uniformly bounded away from 0). This implies (subject to conditions on Δ_k, of course) that Lemma 1 is satisfied for $\hat{\theta}_k$ in this compact subset, i.e., $\hat{g}_k(\hat{\theta}_k)$ is an unbiased estimator of $g(\hat{\theta}_k)$ to within an $O(c_k^2)$ bias term. The $\{\Delta_{ki}\}$ will be generated as independent Bernoulli random variables with outcomes ±1, and so the conditions of Propositions 1 and 2 on Δ_k are satisfied. The measurement noise for both SPSA and FDSA will be generated as i.i.d. normal random variables, which clearly satisfies the relevant conditions. Also note that in the nonidentical Q_i case being considered here, no closed form solution to $g(\theta) = 0$ exists; so an iterative algorithm would be needed even if $L(\cdot)$ were observed without measurement noise.

There are two remaining sections. Section V-B, which presents the main results, considers the case where $p = 20$ and evaluates the relative performance of SPSA and FDSA for several different SA gain sequences and measurement noise distributions. Section V-C discusses some ancillary studies that provide additional insight into the relative performance of the algorithms, including a comparison of SPSA to the random directions method of Kushner and Clark [20] that was mentioned in Section II.

B. Main Study

1) Design of Study and Results: This section compares the SPSA to the FDSA algorithms for the case where $p = 20$. The constants x_i in $L(\theta)$ were generated randomly according to an $N(0, \Sigma + Q_i)$ distribution where: $\Sigma = 225I_{20}$, $N = 60$, $Q_i = A_iA_i^T$ with $A_i \in R^{20\times 30}$, and where each element of the A_i matrices was generated uniformly (and independently) on $(-1, 1)$. This, of course, leads to the nonidentical $\{Q_i\}$ situation mentioned in Section V-A. The value θ^* (the root of $g(\theta) = 0$) was found by application of a scoring algorithm to $L(\theta)$ with no measurement noise; this value for θ^* was corroborated by application of a Newton–Raphson algorithm.

The table below compares the performance of SPSA and FDSA for two different distributions of the measurement noise. (The notation ϵ_k is used to generically represent $\epsilon_k^{(+)}, \epsilon_k^{(-)}, \epsilon_k^{(l+)}, \epsilon_k^{(l-)}$ as appropriate.) Two different SA gain sequences were used ($a_k = 300/k^{0.7501}$ and a "standard" $O(1/k)$ gain, $a_k = 300/k$) while c_k was the same for all studies ($c_k = 100/k^{0.25}$). Note that the combinations of a_k and c_k satisfy the relevant conditions in Propositions 1 and 2. The notation SPSA-q denotes SPSA with an average of q individual $\hat{g}_k(\cdot)$'s at each iteration of the algorithm as discussed in Section II (2.3). To effect a fair comparison of the SPSA to the FDSA algorithm, the runs were initialized with $\hat{\theta}_0 = \tilde{\theta}_0$ ($\|\hat{\theta}_0 - \theta^*\| = 770.8$ and $\|\theta^*\| = 1064.6$) and same random number seed (for generating the ϵ_k's). Several additional runs with different x_i's, Q_i's, and $\hat{\theta}_0$ were also made, which confirmed that the relative behavior of SPSA and FDSA reported in Table I is representative (Spall [31] discusses some of these additional studies).

[7] Chin [4] considers a different function $L(\cdot)$ and performs a study similar to that here. His numerical results are qualitatively the same.

[8] In an MLE context observing the likelihood $L(\cdot)$ in the presence of noise and using SA to find the root might be useful if $L(\cdot)$ is approximated in some computationally efficient random manner. Such stochastic optimization techniques have been considered in [34], where it is assumed that the function to be minimized is approximated in an efficient random way.

TABLE I

Values of $\|\hat{\theta}_k - \theta^*\|/\|\hat{\theta}_0 - \theta^*\|$ and $\|\tilde{\theta}_k - \theta^*\|/\|\tilde{\theta}_0 - \theta^*\|$ for Slowly Decaying and Standard Gain Sequences

	$O(1/k^{0.7501})$ Gain Sequence				$O(1/k)$ Gain Sequence			
	$\epsilon_k \sim N(0, 400)$		$\epsilon_k \sim N(0, 1600)$		$\epsilon_k \sim N(0,400)$		$\epsilon_k \sim N(0, 1600)$	
	$n = 80$	$n = 3000$	$n = 80$	$n = 3000$	$n = 80$	$n = 3000$	$n = 80$	$n = 3000$
SPSA-1	0.61	0.19	0.96	0.41	0.73	0.40	0.91	0.57
SPSA-2	0.65	0.15	0.80	0.32	0.77	0.42	0.85	0.51
SPSA-4	0.73	0.14	0.79	0.27	0.81	0.46	0.85	0.50
FDSA	0.86	0.49	0.90	0.70	0.87	0.59	0.88	0.73

The noises in Table I were chosen to be large enough to have a significant effect in degrading $\hat{g}_k(\cdot)$ (and the FDSA approximation $\tilde{g}_k(\cdot)$), which is appropriate for testing the efficacy of the algorithms under realistic conditions. The $N(0, 1600)$ noise contributes, on average, over 50% to the magnitude of the $\hat{g}_k(\cdot)$ terms in SPSA-1 for the first few iterations of the algorithm and a greater percentage in later iterations. That is, for low values of k the magnitude of $L(\hat{\theta}_k + \overline{\Delta}_k) - L(\hat{\theta}_k - \overline{\Delta}_k)$ was generally near 40 while the mean absolute noise contribution (i.e., $E|\epsilon_k^{(+)} - \epsilon_k^{(-)}|$) was slightly over 45. For larger values of k, the difference $L(\hat{\theta}_k + \overline{\Delta}_k) - L(\hat{\theta}_k - \overline{\Delta}_k)$ approached 0 (due to the greater flatness of $L(\cdot)$ near θ^* and fact that the magnitude of $\overline{\Delta}_k$ approached 0), and so the relative noise contribution to the calculation $\hat{g}_k(\cdot)$ approached 100% (which in turn was compensated for by the fact that $a_k \to 0$). The two columns under each noise distribution contain the values of normalized $\|\hat{\theta}_k - \theta^*\|$ and $\|\tilde{\theta}_k - \theta^*\|$ after $n = 80$ or $n = 3000$ measurements y have been processed. Since $p = 20$, each iteration of FDSA requires 40 measurements while each iteration of SPSA-q requires $2q$ measurements. Thus, for example, in the columns labeled $n = 3000$, SPSA-2 has gone through 750 iterations while FDSA has gone through 75 iterations. The two values of n were chosen to illustrate small- and moderate-sample behavior of the algorithms.

2) Interpretation of Results: For the small sample $n = 80$ case, the results were mixed, although overall SPSA tended to achieve a smaller normalized $\|\cdot\|$ value (especially for the $O(1/k^{0.7501})$ gain). For the larger sample $n = 3000$ setting, the evidence for SPSA is much stronger. In every case, the norm value for SPSA was significantly lower than the value for FDSA. Furthermore, based on SPSA norm values not included in the table, it was found that SPSA reached the terminal FDSA ($n = 3000$) value with approximately 1/20 to 1/10 the number of measurements for the $O(1/k^{0.7501})$ gain and 1/8 to 1/4 the number of measurements for the $O(1/k)$ gain.

Let us now discuss the values in the table in light of the discussion in Section IV on asymptotic relative mean square error. Since $M = \tilde{M}$ and $\mu = \tilde{\mu} = 0$, we have from (4.1)

$$\frac{\text{RMS(SPSA-}q)}{\text{RMS(FDSA)}} \to \left(\frac{q^{\beta-1}}{p^{\beta}}\right)^{1/2} \qquad (5.1)$$

as $n \to \infty$ where RMS($\cdot$) represents root mean square error. For the $O(1/k^{0.7501})$ gain the ratio in (5.1) varies from 0.41 to 0.69, while the observed $n = 3000$ ratios vary from 0.29 to 0.59. As predicted by (5.1), the SPSA values in the table decrease as q increases (of course, q cannot be increased indefinitely since the number of iterations must also be large enough to justify the asymptotic theory). For the $O(1/k)$ gain, the ratio in (5.1) varies from 0.33 to 0.47 while the observed $n = 3000$ ratios vary from 0.68 to 0.78; however, unlike the $O(1/k^{0.7501})$ gain, the SPSA values do not necessarily decrease as q increases.

Although there is some consistency between the asymptotically based (5.1) and the observed norm values in the table (especially for $O(1/k^{0.7501})$), there is also some evidence that $n = 3000$ is not large enough for asymptotic theory to be fully valid. Three observations provide this evidence: 1) the norm values in the table for the $O(1/k)$ gain are larger than those for the $O(1/k^{0.7501})$ gain despite the fact that β for the former is greater than β for the latter and $\hat{\theta}_k - \theta^*$ is $O_p(k^{-\beta/2})$ (this observed superiority for slowly decaying gains in a *finite-sample* setting is consistent with the author's other experiences with SA [e.g., [30], [31]] although asymptotic theory (e.g., [10], [11] and [13]) indicates that $O(1/k)$ gains are optimal), 2) the norm values are still dropping relatively fast near the terminal iterations for the $O(1/k)$ gain (although the values are dropping only slightly for the $O(1/k^{0.7501})$ gain), and 3) the previously mentioned observed ratio of number of measurements for equivalent accuracy (1/20 to 1/4) tends to be larger than the asymptotically valid ratios of 1/20 to less than 1/1000 as given by (4.2), although predicted by (4.2) the observed ratios tend to be smaller for the $O(1/k^{0.7501})$ gain. The above confirms that while asymptotic theory provides some basis for understanding the relative and absolute behavior of SPSA, it must be used with caution in practical problems.

C. Some Ancillary Studies

To acquire further insight into the SPSA algorithm we performed several additional studies involving SPSA and FDSA; we also considered the random directions (RDSA) method of Kushner and Clark [20, pp. 58–60], which was mentioned in Section II (see [4] for a more detailed theoretical and numerical analysis of RDSA relative to SPSA). Spall [31] includes details on a study involving a larger measurement noise variance ($\epsilon_k \sim N(0, 10)^4)$) and a study where $p = 5$. As expected, these studies show that gradient averaging ($q > 1$) has a beneficial effect in stabilizing the algorithm and avoiding divergence when the noise contribution is large and that the advantage of SPSA over FDSA is not as great as when $p = 5$ (relative to $p = 20$). Spall and Cristion [32] consider a setting where $p = 302$ in the context of weight estimation for a neural network; as might be expected in such a large-dimensional case, SPSA is shown to have a dramatic advantage over FDSA. We now discuss a study based on "optimal" gains as well as the above mentioned study involving RDSA.

The results of Section V-B illustrate the relative performance of SPSA and FDSA when both algorithms have the same gain sequence, which is representative of what would happen in a practical application where the "optimal" gain is unknown or uncomputable. We also considered the relative performance for $p = 20$ when each algorithm was run with its optimal gain (determined numerically). It was found that SPSA continued to significantly outperform FDSA. In particular for the eight runs made at the two noise levels of the table (four different random number seeds at each noise level), SPSA-4 (the only SPSA algorithm considered) reached the terminal FDSA ($n = 1200$) level of accuracy with approximately $1/14$ to $1/4$ the number of measurements. This is comparable to the observed $1/20$ to $1/4$ range for SPSA-4 that is associated with the identical gains of the table in Section V-B.

RDSA is similar to SPSA in that it too requires only $2q$ measurements per iteration; it differs, however, in the form of the gradient estimate and assumptions on the random variables characterizing the search direction (analogous to Δ_k here). With each algorithm run with its respective optimal gain (as described above) it was found that when $p = 20$, SPSA-4 (the only SPSA algorithm considered) was significantly more efficient than RDSA-4 (and RDSA-4 was more effective than RDSA-1 or RDSA-2). In particular for the eight runs made at the two noise levels of the table, SPSA used from $1/15$ to $1/3$ the number of measurements used (1200) by RDSA to reach a given level of accuracy. This range is close to that mentioned above for FDSA versus SPSA with optimal gains, which is consistent with the claim of Kushner and Clark that RDSA and FDSA perform about equally well. As with FDSA, the relative advantage of SPSA appears to be higher in larger-dimensional problems. In particular, for the smaller $p = 5$ case, neither algorithm appeared to have a significant advantage.

VI. Concluding Remarks

For problems in the multivariate Kiefer–Wolfowitz setting, this paper has presented an SA algorithm based on a "simultaneous perturbation" gradient approximation. The bias in the gradient estimate was characterized and the algorithm was shown to have the almost sure convergence property of standard (Kiefer-Wolfowitz) finite difference SA algorithms. Conditions were also presented under which the SPSA procedure is asymptotically normally distributed. This result allowed us to show that, in a wide variety of problems, the asymptotic MSE for SPSA will be smaller than that for FDSA and that the relative MSE for SPSA (to FDSA) gets smaller as the problem dimension (p) gets larger.

The algorithm was compared in a numerical study to FDSA and the random directions SA procedure of Kushner and Clark in finding the minimum of a fairly complicated function. The comparison was performed with the algorithms run under a common SA gain sequence and also with each algorithm run under its own optimal gain. It was found that SPSA was significantly superior to either FDSA or RDSA for a variety of scenarios, and appeared to be relatively more effective as the amount of data increases.

There remain several open questions in the theory and application of SPSA. One is to determine a "best" (or at least suboptimal) distribution for the perturbation random variables (the Δ_{ki}'s); we employed a Bernoulli distribution in the numerical studies but it is possible that other (perhaps continuous) distributions may be superior, at least in certain problems. The asymptotic normality result provides the basis for determining optimal SA gain coefficients (a, c); carrying out these calculations would be of some interest. It would also be of interest to examine whether an averaging scheme such as that in [8], [27], or [33], which might enhance the rate of convergence, could be developed for SPSA. Similarly, it may be possible to develop an SPSA-type technique for accelerated SA algorithms (e.g., those with adaptive gains or second-order effects, as mentioned in Section II). With respect to applications, the problem of learning in neural networks is one area of particular promise since the dimension of the weight vector to be estimated is inherently high (easily on the order of 10^2 or 10^3); a preliminary look at network learning via SPSA is given in [32] for a problem in adaptive control. Another area, not directly related to SA, is the potential application of the simultaneous perturbation gradient approximation to the field of perturbation analysis, as discussed briefly in Section II. Solving any of these problems is likely to make SPSA an even more efficient algorithm or a more widely applicable technique.

References

[1] N. Berman, A. Feuer, and E. Wahnon, "Convergence analysis of smoothed stochastic gradient-type algorithm," *Int. J. Syst. Sci.*, vol. 18, pp. 1061-1078, 1987.

[2] J. R. Blum, "Multidimensional stochastic approximation methods," *Ann. Math. Stat.*, vol. 25, pp. 737-744, 1954.

[3] X. R. Cao, "Convergence of parameter sensitivity estimates in a stochastic experiment," *IEEE Trans. Automat. Contr.*, vol. AC-30, pp. 845-853.

[4] D. C. Chin, "Comparative study of several multivariate stochastic approximation algorithms," in *Proc. Amer. Stat. Assoc., Stat. Comp. Sect.*, 1990, pp. 223-228.

[5] J. L. Doob, *Stochastic Processes*. New York: Wiley, 1953.

[6] Y. Ermoliev, "On the method of generalized stochastic gradients and quasi-Fejer sequences," *Cybernetics*, vol. 5, pp. 208-220, 1969.

[7] Y. Ermoliev, "Stochastic quasigradient methods and their application to system optimization," *Stochastics*, vol. 9, pp. 1-36, 1983.

[8] V. Fabian, "Stochastic approximation of minima with improved asymptotic speed," *Ann. Math. Stat.*, vol. 38, pp. 191-200, 1967.

[9] —, "On asymptotic normality in stochastic approximation," *Ann. Math. Stat.*, vol. 39, pp. 1327-1332, 1968.

[10] —, "Stochastic approximation," in *Optimizing Methods in Statistics*, J. J. Rustagi, Ed. New York: Academic, 1971, pp. 439-470.

[11] —, "Asymptotically efficient stochastic approximation: The RM case," *Ann. Stat.*, vol. 1, pp. 486-495, 1973.

[12] W. Feller, *An Introduction to Probability Theory and its Applications, Vol. 2*. New York: Wiley, 1971.

[13] L. Goldstein, "On the choice of step size in the Robbins-Monro procedure," *Stat. Prob. Lett.*, vol. 6, pp. 299-303, 1988.

[14] Y. C. Ho, "Perturbation analysis explained," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 761-763, 1989.

[15] J. M. Holtzman, "On using perturbation analysis to do sensitivity analysis: Derivatives versus differences," in *Proc. IEEE Conf. Decision Control*, 1989, pp. 2018-2023.

[16] S. L. Hui and J. O. Berger, "Empirical Bayes estimation of rates in

longitudinal studies," *J. Amer. Stat. Assoc.*, vol. 78, pp. 753-760, 1983.

[17] M. Kesten, "Accelerated stochastic approximation," *Ann. Math. Stat.*, vol. 29, pp. 41-59, 1958.

[18] J. Kiefer and J. Wolfowitz, "Stochastic estimation of a regression function," *Ann. Math. Stat.*, vol. 23, pp. 462-466, 1952.

[19] H. J. Kushner and T. Gavin, "Extensions of Kesten's adaptive stochastic approximation method," *Ann. Stat.*, vol. 1, pp. 851-861, 1973.

[20] H. J. Kushner and D. S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems.* New York: Springer-Verlag, 1978.

[21] R. G. Laha and V. K. Rohatgi, *Probability Theory.* New York: Wiley, 1979.

[22] T. L. Lai, "Stochastic approximation and sequential search for optimum," in *Proc. Berkeley Conf. in Honor of Jerzy Newman and Jack Kiefer, Vol. II*, L. M. Le Cam and R. A. Olshen, Eds. Belmont, CA: Wadsworth, 1985, pp. 557-577.

[23] L. Ljung, "Strong convergence of a stochastic approximation algorithm," *Ann. Stat.*, vol. 6, pp. 680-696, 1978.

[24] M. Metivier and P. Priouret, "Applications of a Kushner and Clark lemma to general classes of stochastic algorithms," *IEEE Trans. Informat. Theory*, vol. 30, pp. 140-151, 1984.

[25] D. Ruppert, "Kiefer-Wolfowitz procedure," in *Encyclopedia of Statistical Sciences, Vol. 4*, S. Kotz and N. L. Johnson, Eds. New York: Wiley, 1983, pp. 379-381.

[26] ——, "A Newton-Raphson version of multivariate Robbins-Monro procedure," *Ann. Stat.*, vol. 13, pp. 236-245, 1985.

[27] A. Ruszczynski and W. Syski, "Stochastic approximation method with gradient averaging for unconstrained problems," *IEEE Trans. Automat. Contr.*, vol. AC-28, pp. 1097-1105.

[28] G. N. Saridis, *Self-Organizing Control of Stochastic Systems.* New York: Marcel-Dekker.

[29] R. H. Shumway, D. E. Olsen, and L. J. Levy, "Estimation and tests of hypotheses for the initial mean and covariance in the Kalman filter model," *Commun. Stat. Theory Meth.*, vol. A-10, pp. 1625-1641, 1981.

[30] J. C. Spall, "Bayesian error isolation for models of large-scale systems," *IEEE Trans. Automat. Contr.*, vol. 33, pp. 341-347, 1988.

[31] ——, "A stochastic approximation algorithm for large-dimensional systems in the Kiefer-Wolfowitz setting," in *Proc. IEEE Conf. Decision Contr.*, 1988, pp. 1544-1548.

[32] J. C. Spall and J. A. Cristion, "Neural networks for control of uncertain systems," in *Proc. Test Technology Symp. IV* (sponsored by U.S. Army Test and Evaluation Command), 1991, pp. 575-588.

[33] M. A. Styblinski and T.-S. Tang, "Experiments in nonconvex optimization: Stochastic approximation with function smoothing and simulated annealing," *Neural Networks*, vol. 3, pp. 467-483, 1990.

[34] Y. Wardi, "A stochastic steepest-descent algorithm," *J. Optimiz. Theory Appl.*, vol. 59, pp. 307-323, 1988.

James C. Spall (S'82-M'83-SM'90) received the S. M. degree in technology and policy from the Massachusetts Institute of Technology, Cambridge, in 1981, and the Ph.D. degree in systems engineering from the University of Virginia, Charlottesville, in 1983.

Since 1983, he has been with The Johns Hopkins University, Applied Physics Laboratory, where he is currently a Project Leader for several research efforts focusing on problems in statistical modeling. In 1991, he was appointed to the Principal Professional Staff of the Laboratory.

Dr. Spall has published numerous research articles in the areas of statistics and control, including Kalman filtering, non-Gaussian modeling, stochastic approximation, system identification, neural networks, and general Bayesian analysis, and served as editor and coauthor of the book, *Bayesian Analysis of Time Series and Dynamic Models* (1988). He is a Member of The American Statistical Association and Sigma Xi, and a Fellow of Tau Beta Pi.