Adaptive Stochastic Approximation by the Simultaneous Perturbation Method

James C. Spall, *Senior Member, IEEE*

***Abstract*—Stochastic approximation (SA) has long been applied for problems of minimizing loss functions or root finding with noisy input information. As with all stochastic search algorithms, there are adjustable algorithm coefficients that must be specified, and that can have a profound effect on algorithm performance. It is known that choosing these coefficients according to an SA analog of the deterministic Newton–Raphson algorithm provides an optimal or near-optimal form of the algorithm. However, directly determining the required Hessian matrix (or Jacobian matrix for root finding) to achieve this algorithm form has often been difficult or impossible in practice. This paper presents a general adaptive SA algorithm that is based on a simple method for estimating the Hessian matrix, while concurrently estimating the primary parameters of interest. The approach applies in both the gradient-free optimization (Kiefer–Wolfowitz) and root-finding/stochastic gradient-based (Robbins–Monro) settings, and is based on the "simultaneous perturbation (SP)" idea introduced previously. The algorithm requires only a small number of loss function or gradient measurements per iteration—independent of the problem dimension—to adaptively estimate the Hessian and parameters of primary interest. Aside from introducing the adaptive SP approach, this paper presents practical implementation guidance, asymptotic theory, and a nontrivial numerical evaluation. Also included is a discussion and numerical analysis comparing the adaptive SP approach with the iterate-averaging approach to accelerated SA.**

***Index Terms*—Adaptive estimation, optimization, parameter estimation, root-finding, simultaneous perturbation stochastic approximation (SPSA), stochastic approximation.**

I. Introduction

STOCHASTIC approximation (SA) represents an important class of stochastic search algorithms. Many well-known techniques are special cases of SA, including neural-network backpropagation, perturbation analysis for discrete-event systems, recursive least squares and least mean squares, and some forms of simulated annealing. Therefore, progress in general SA methodology can have a potential bearing on a wide range of practical implementations. This paper presents an approach for accelerating the convergence of SA algorithms. The results apply in both the gradient-free (Kiefer–Wolfowitz) and stochastic gradient-based (Robbins–Monro root-finding) SA settings.[1] The essential idea is to use the "simultaneous perturbation" concept to efficiently and easily estimate the Hessian matrix of the loss function to be optimized (or, equivalently, the Jacobian matrix for root finding). This Hessian matrix is then used in an SA recursion that is a stochastic analog of the well-known Newton–Raphson algorithm of deterministic optimization to accelerate convergence.

The problem of minimizing a (scalar) differentiable loss function $L(\theta)$, where $\theta \in \boldsymbol{R}^p, p \geq 1$ is considered. A typical example of $L(\theta)$ would be some measure of mean-square error for the output of a process as a function of some design parameters θ. For many cases of practical interest, this is equivalent to finding the unique minimizing θ^* such that

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

For the gradient-free setting, it is assumed that measurements of $L(\theta)$, say $y(\theta)$, are available at various values of θ. These measurements may or may not include random noise. No direct measurements (either with or without noise) of $g(\theta)$ are assumed available in this setting. In the gradient-based case, it is assumed that direct measurements of $g(\theta)$ are available, usually in the presence of added noise. The basic problem is to take the available information (measurements of $L(\theta)$ and/or $g(\theta)$), and attempt to estimate θ^*. This is essentially a local unconstrained optimization problem (although this is also the form when differentiable penalty functions are used for constrained optimization). Although there are extensions of SA to finding the global optimum in the presence of multiple local minima and for optimizing in the presence of constraints (see, e.g., Styblinski and Tang [39], Chin [5], Kushner and Yin [15, pp. 77–79, 100–101, etc.], and Sadegh [29])—and it is expected that the approach here would apply in the context of these extensions—we will not focus on those generalizations in this paper.

The adaptive simultaneous perturbation (ASP) approach here is based on the simple idea of creating two parallel recursions, one for estimating θ and the other for estimating the Hessian matrix $H(\theta)$. The first recursion is a stochastic analog of the Newton–Raphson algorithm, and the second recursion yields the sample mean of per-iteration Hessian estimates. The second recursion provides the Hessian estimate for use in the first recursion. The simultaneous perturbation idea of varying all of the parameters in the problem together (rather than one at a time) is

Manuscript received December 31, 1998; revised September 21, 1999 and November 8, 1999. Recommended by Associate Editor, T. Parisini. This work was supported in part by the JHU/APL IRAD Program, and the U.S. Navy under Contract N00024-98-D-8124.

The author is with the Applied Physics Laboratory, The Johns Hopkins University, Laurel, MD 20723-6099 USA (e-mail: james.spall@jhuapl.edu).

Publisher Item Identifier S 0018-9286(00)09446-0.

[1]Although this paper is written largely in the language of optimization, the ideas would also apply in the stochastic root-finding context of the Robbins–Monro algorithm. In particular, the "gradient" in this paper is equivalent to the function for which a zero is to be found, and the Hessian matrix is equivalent to the Jacobian matrix of this function.

used to form the per-iteration Hessian estimates in the second recursion. This leads to an efficient means for achieving a second-order adaptive algorithm. In particular, in the gradient-free case, only *four* function measurements $y(\cdot)$ are needed at each iteration to estimate both the gradient and Hessian for any dimension p. In the gradient-based case, *three* gradient measurements are needed at each iteration, again for any p. (In practical implementations, one or more additional $y(\cdot)$ values may be useful as a check on algorithm behavior as discussed in Section II-D below.) Although ASP is a *relatively* simple adaptive approach, care is required in implementation just as in any other second-order-type approach (deterministic or stochastic); this includes the choice of initial condition and choice of "step size" coefficients to avoid divergence. (However, simply choosing the step size $a_k = 1/k$ in the notation below provides *asymptotically* optimal or near-optimal performance.) These issues are discussed in the sections to follow.

Although the *concept* of adaptive SA has been known for some time (e.g., Venter [41], Nevel'son and Has'minskii [23, ch. 7]), the *implementation* has been far less successful: no adaptive method appears to have been proposed that is practically implementable in a wide range of general multivariate problems (e.g., "...the optimal choice [of gain sequence] involves the Hessian of the risk [loss] function, which is typically unknown and hard to estimate," from Yakowitz *et al.* [44]). Let us summarize some of the existing approaches to illustrate the difficulties. Fabian [10] forms estimates of the gradient and Hessian for an adaptive SA algorithm by using, respectively, a finite-difference approximation and a set of differences of finite-difference approximations. This leads to $O(p^2)$ measurements $y(\cdot)$ per update of the θ estimate, which is extremely costly when p is large. Kao *et al.* [13] present a heuristic approach based on analogies to quasi-Newton methods of deterministic optimization; at each iteration, this approach uses $O(p)$ function measurements $y(\cdot)$ plus some additional measurements for a separate line search. For the gradient-based case, Ruppert [27] forms a Hessian estimate by taking finite differences of gradient measurements. In a similar spirit, Wei [43] presents a multivariate extension of the Venter [41] and Nevel'son and Has'minskii [23, ch. 7] approaches for adaptive Robbins–Monro algorithms. Both the Ruppert and Wei approaches require $O(p)$ measurements of $g(\cdot)$ for each iteration. These approaches differ from the ASP approach in the potentially large number of function or gradient measurements required per iteration. Related to the above, there are also numerous means for adaptively estimating a Hessian matrix in special SA estimation settings where one has detailed knowledge of the underlying model (see, e.g., Macchi and Eweda [19], Benveniste *et al.* [1, ch. 3–4], Ljung [17], and Yin and Zhu [45]); while these are more practically implementable than the general adaptive approaches mentioned above, they are restricted in their range of application.

The concept of iterate averaging, as reported in Ruppert [28] and Polyak and Juditsky [25] for the gradient-based case and Dippon and Renz [7] for the gradient-free case, also provides a form of second-order (optimal or near-optimal) convergence for SA. This appealingly simple idea is based on using a sample mean of the iterates coming from a "basic" first-order SA recursion as the final estimate of θ. For the gradient-based case, it can be shown that the asymptotic mean-square error for the averaged iterations is identical to that which would be obtained by using the true Hessian in a stochastic Newton–Raphson-like algorithm, i.e., the iterate averaging method achieves the minimum possible mean-square error *without* requiring knowledge—or even an estimate—of the Hessian matrix. For the gradient-free case, the iterate averaging solution is *nearly* asymptotically optimal in a precise sense defined by Dippon and Renz [7]. Some numerical studies provide support for the benefits of iterate averaging (e.g., Yin and Zhu [45], Kushner and Yin [15, ch. 11]). However, finite-sample analysis by this author and others (e.g., Maryak [21], Spall and Cristion [38], and, Section V here) has shown that the asymptotic promise of iterate averaging may be difficult to realize in practice. This is not surprising upon reflection. For iterate averaging to be successful, it is necessary that a large proportion of the individual iterates hover approximately uniformly around θ^* in $\boldsymbol{R}^p$, leading to the average of the iterates producing a mean that is nearer θ^* than the bulk of the individual iterates. However, since a well-designed ("stable") algorithm will not be jumping approximately uniformly around θ^* when the iterates are far from the solution (or else it is likely to diverge), the only way for the bulk of the iterates to be distributed uniformly around the solution is for the individual iterates to be near the solution. In most practical settings with a well-designed algorithm, one observes that the components of θ move in a *roughly* (subject to the inherent stochastic variability) monotonic manner toward the solution, and that the user will terminate the algorithm when either the "budget" of iterations has been exceeded or when the iterates begin to move very slowly near (one hopes!) θ^*. But the latter situation is precisely when iterate averaging starts to work well (in fact, while the algorithm is in its monotonic phase, iterate averaging will tend to *hurt* the accuracy of those components in θ that have not yet settled near θ^*!). This suggests that, despite the simplicity and asymptotic justification, the role of iterate averaging in practical finite-sample problems may not be in achieving true second-order efficiency (one role may be in enhancing algorithm stability by feeding back the averaged solution into the iteration process as in Kushner and Yin [15, ch. 11]).[2]

Section II describes the general ASP approach, and summarizes the essential methodology related to the simultaneous perturbation form of the basic first-order SA algorithm (i.e., the SPSA algorithm). This section also summarizes some of the practical guidelines the user should consider in a real implementation. Section III and the associated Appendixes A and B provide part of the theoretical justification for ASP, establishing conditions for the almost sure (a.s.) convergence of both the θ iterate and the Hessian estimate. Section IV and Appendix A then build on this convergence to establish the asymptotic normality of ASP in both the gradient-based and gradient-free case. Most importantly, Section IV uses the asymptotic normality to analyze the statistical estimation error of the ASP iterates, showing that the errors are either asymptotically optimal or nearly op-

[2] Note the contrast of iterate averaging with a "true" second-order algorithm, where knowledge of the Hessian, even at iterations not near θ^*, may enhance convergence by improving the search direction and scaling for potentially large differences in the magnitudes of the θ elements.

timal. Section V performs a numerical analysis of ASP, and Section VI offers some concluding remarks.

II. The Adaptive Simultaneous Perturbation Approach: Methodology and Implementation Issues

A. Basic Form of Algorithm

The second-order ASP approach is composed of two parallel recursions: one for θ and one for the Hessian of $L(\theta)$. The two core recursions are, respectively,

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \overline{\overline{H}}_k^{-1} G_k(\hat{\theta}_k), \quad \overline{\overline{H}}_k = f_k(\overline{H}_k) \tag{2.1a}$$

$$\overline{H}_k = \frac{k}{k+1}\,\overline{H}_{k-1} + \frac{1}{k+1}\,\hat{H}_k, \quad k = 0, 1, 2, \cdots \tag{2.1b}$$

where a_k is a nonnegative scalar gain coefficient, $G_k(\hat{\theta}_k)$ is the input information related to $g(\hat{\theta}_k)$ (i.e., the gradient approximation from $y(\cdot)$ measurements in the gradient-free case or the direct observation as in the Robbins–Monro gradient-based case), f_k: $\boldsymbol{R}^{p\times p} \rightarrow \{positive\ definite\ p \times p\ matrices\}$ is a mapping designed to cope with possible nonpositive definiteness of $\overline{H}_k$, and $\hat{H}_k$ is a per-iteration estimate of the Hessian discussed below.[3] Equation (2.1a) is a stochastic analog of the well-known Newton–Raphson algorithm of deterministic search and optimization. Equation (2.1b) is simply a recursive calculation of the sample mean of the per-iteration Hessian estimates.[4] Initialization of the two recursions is discussed in Section II-D below. Since $G_k\ (\hat{\theta}_k)$ has a known form, the parallel recursions in (2.1a), (2.1b) can be implemented once $\hat{H}_k$ is specified. The remainder of this paper will focus on two specific implementations of the ASP approach above: 2SPSA (second-order SPSA) for applications in the gradient-free case, and 2SG (second-order stochastic gradient) for applications in the Robbins–Monro gradient-based case.

We now present the per-iteration Hessian estimate $\hat{H}_k$. As with the "basic" first-order SPSA algorithm, let c_k be a positive scalar (decaying to 0 for formal convergence; conditions given below), and let $\Delta_k \in R^p$ be a user-generated mean-zero random vector satisfying certain regularity conditions discussed in Section III below. (Typical conditions are that the individual components Δ_{ki} be mutually independent, bounded, symmetrically distributed, and have finite *inverse* moments of order ≥ 2, e.g., Δ_k being a vector of independent Bernoulli ± 1 random variables satisfies these conditions, but a vector of uniformly or normally distributed random variables does not.) It will prove convenient to work with a "vector-divide" operation where the ijth element of the resulting matrix corresponds to the ratio of the jth element of the numerator row vector to the ith element of the denominator column vector. Applying the vector-divide operator, the formula for estimating the Hessian at each iteration is

$$\hat{H}_k = \frac{1}{2}\left[\frac{\delta G_k^T}{2c_k\Delta_k} + \left(\frac{\delta G_k^T}{2c_k\Delta_k}\right)^T\right] \tag{2.2}$$

where

$$\delta G_k = G_k^{(1)}(\hat{\theta}_k + c_k\Delta_k) - G_k^{(1)}(\hat{\theta}_k - c_k\Delta_k)$$

and $G_k^{(1)}(\cdot)$ may or may not equal $G_k(\cdot)$, depending on the setting.[5] In particular, for 2SPSA, there are advantages to using a *one-sided* gradient approximation in order to reduce the total number of function evaluations [versus the two-sided form usually recommended for $G_k(\cdot)$], while for 2SG, usually $G_k^{(1)}(\cdot) = G_k(\cdot)$. The term "simultaneous perturbation" in ASP comes from the fact that all elements of $\hat{\theta}_k$ are varied simultaneously (and randomly) in forming $\hat{H}_k$, as opposed to the finite-difference forms in, e.g., Fabian [10] and Ruppert [27], where the elements of θ are changed deterministically one at a time.

B. Specific Gradient Forms

While the ASP structure in (2.1a), (2.1b), and (2.2) is general, we will largely restrict ourselves in our choice of $G_k(\cdot)$ (and $G_k^{(1)}(\cdot)$) in the remainder of the discussion in order to present concrete theoretical and numerical results. For 2SPSA, we will consider the simultaneous perturbation approach for generating $G_k(\cdot)$ and $G_k^{(1)}(\cdot)$, while for 2SG, we will suppose that $G_k(\cdot) = G_k^{(1)}(\cdot)$ is an unbiased direct measurement of $g(\cdot)$ (i.e., $G_k(\cdot) = G_k^{(1)}(\cdot) = g(\cdot)$ + *mean-zero noise*). The rationale for basic SPSA in the gradient-free case has been discussed extensively elsewhere (e.g., Spall [33], Chin [6], Dippon and Renz [7], and Gerencsér [12]), and hence will not be discussed in detail here. (In summary, it tends to lead to more efficient optimization than the classical finite-difference Kiefer–Wolfowitz method while being no more difficult to implement; the relative efficiency grows with the problem dimension p.) In the gradient-based case, SG methods include as special cases the well-known approaches mentioned at the beginning of Section I (backpropagation, etc.). SG methods are themselves special cases of the general Robbins–Monro root-finding framework and, in fact, most of the results here can apply in this root-finding setting as well.

For 2SPSA, the core gradient approximation $G_k(\hat{\theta}_k)$ requires two measurements of $L(\cdot)$, $y(\hat{\theta}_k + c_k\Delta_k)$ and $y(\hat{\theta}_k - c_k\Delta_k)$, representing measurements at design levels $\hat{\theta}_k + c_k\Delta_k$ and $\hat{\theta}_k - c_k\Delta_k$, where c_k and Δ_k are as defined above for $\hat{H}_k$ (see Spall [32], [33]). These two measurements will be used to generate $G_k(\hat{\theta}_k)$ in the conventional SPSA manner, in addition to being employed toward generating the one-sided gradient approximations $G_k^{(1)}(\hat{\theta}_k \pm c_k\Delta_k)$ used in forming $\hat{H}_k$. Two additional mea-

[3]In the general Robbins–Monro root-finding case, the mapping f_k would be into the set of nonsingular (but not necessarily symmetric) matrices.

[4]It is also possible to use a weighted average or "sliding window" method (where only the most recent $\hat{H}_k$ values are used in the recursion) to determine $\overline{H}_k$. Formal convergence of $\overline{H}_k$ (à la Theorems 2a, b) may still hold under such weighting provided that the analog to expressions (A10) and (A13) in the proof of Theorem 2a holds.

[5]The symmetrizing operation in (2.2) is convenient in the optimization case being emphasized here to maintain a symmetric Hessian estimate in finite samples. In the general root-finding case, where $H(\theta)$ represents a Jacobian matrix, the symmetrizing operation should not be used since the Jacobian is not necessarily symmetric.

surements $y(\hat{\theta}_k \pm c_k\Delta_k + \tilde{c}_k\tilde{\Delta}_k)$ are used in generating the one-sided approximations as follows:

$$G_k^{(1)}(\hat{\theta}_k \pm c_k\Delta_k) = \frac{y(\hat{\theta}_k \pm c_k\Delta_k + \tilde{c}_k\tilde{\Delta}_k) - y(\hat{\theta}_k \pm c_k\Delta_k)}{\tilde{c}_k} \begin{bmatrix} \tilde{\Delta}_{k1}^{-1} \\ \tilde{\Delta}_{k2}^{-1} \\ \vdots \\ \tilde{\Delta}_{kp}^{-1} \end{bmatrix} \tag{2.3}$$

with $\tilde{\Delta}_k = (\tilde{\Delta}_{k1}, \tilde{\Delta}_{k2}, \cdots, \tilde{\Delta}_{kp})^T$ generated in the same statistical manner as Δ_k, but independently of Δ_k (in particular, choosing $\tilde{\Delta}_{ki}$ as independent Bernoulli ± 1 random variables is a valid—but not necessary—choice), and with $\tilde{c}_k$ satisfying conditions similar to c_k (although the numerical value of $\tilde{c}_k$ may be best chosen larger than c_k; see Section II-D).[6]

C. Motivation for Form of Hessian Recursion

To illuminate the underlying simplicity of ASP, let us now provide some informal motivation for the $\hat{H}_k$ form in (2.2). The arguments below are formalized in the theorems of Sections III and IV. Let $H(\theta)$ represent the true Hessian matrix, and suppose that $g(\theta)$ is three-times continuously differentiable in a neighborhood of $\hat{\theta}_k$. Then, simple Taylor series arguments show that

$$\begin{aligned} E(\delta G_k | \hat{\theta}_k, \Delta_k) &= g(\hat{\theta}_k + c_k\Delta_k) - g(\hat{\theta}_k - c_k\Delta_k) + O(c_k^3) \\ &\equiv \delta g_k + O(c_k^3) \quad (O(c_k^3) = 0 \text{ in the SG case}) \end{aligned}$$

where this result is immediate in the SG case, and follows easily (as in Spall [33, Lemma 1]) by a Taylor series argument in the SPSA case (where the $O(c_k^3)$ term is the difference of the two $O(c_k^2)$ bias terms in the one-sided SP gradient approximations and $\tilde{c}_k = O(c_k)$). Hence, by an expansion of each of $g(\hat{\theta}_k \pm c_k\Delta_k)$, we have for any i, j

$$\begin{aligned} &E\left(\frac{\delta G_{ki}}{2c_k\Delta_{kj}} \middle| \hat{\theta}_k, \Delta_k \right) \\ &\quad = E\left(\frac{\delta g_{ki}}{2c_k\Delta_{kj}} \middle| \hat{\theta}_k, \Delta_k \right) + O(c_k^2) \\ &\quad = H_{ij}(\hat{\theta}_k) + \sum_{\ell \neq j} H_{\ell j}(\hat{\theta}_k) \frac{\Delta_{k\ell}}{\Delta_{kj}} + O(c_k^2) \end{aligned}$$

where the $O(c_k^2)$ term in the second line absorbs higher order terms in the expansion of δg_k. Then, since $E(\Delta_{k\ell}/\Delta_{kj}) = 0 \; \forall \, j \neq \ell$ by the assumptions for Δ_k, we have

$$E\left(\frac{\delta G_{ki}}{2c_k\Delta_{kj}} \middle| \hat{\theta}_k \right) = H_{ij}(\hat{\theta}_k) + O(c_k^2)$$

implying that the Hessian estimate is "nearly unbiased," with the bias disappearing at rate $O(c_k^2)$. The addition operation in (2.2) simply forces the per-iteration estimate to be symmetric.

[6]An alternative SPSA gradient approximation not explored here is the one-measurement form in Spall [34]. Here, only *one* function evaluation is required to get an $O(c_k^2)$ "almost unbiased" approximation of $g(\cdot)$. Although this will increase the variability of $\hat{H}_k$, it may be beneficial in nonstationary settings where the underlying true gradient and Hessian are changing in time since the reduced number of measurements will reduce the potential bias that may otherwise be introduced.

D. Implementation Aspects

The two recursions in (2.1a), (2.1b) are the foundation for the ASP approach. However, as is typical in all stochastic algorithms, the specific implementation details are important. Equations (2.1a), (2.1b) do not fully define these details. The five points below have been found important in making ASP perform well in practice.

1) θ and H Initialization: Typically, (2.1a) is initialized at some $\hat{\theta}_0$ believed to be near θ^*. One may wish to run standard first-order SA (i.e., (2.1a) without $\overline{\overline{H}}_k^{-1}$) or some other "rough" optimization approach for some period to move the initial θ for ASP closer to θ^*. Although, with the indexing shown in (2.1b), no initialization of the $\overline{H}_k$ recursion is required since $\overline{H}_0$ is computed directly from $\hat{H}_0$, the recursion may be trivially modified to allow for an initialization if one has useful prior information. If this is done, then the recursion may be initialized at (say) scale $\cdot \, I_{p\times p}$, scale ≥ 0, or some other positive semidefinite matrix reflecting available prior information (e.g., if one knows that the θ elements will have very different magnitudes, then the initialization may be chosen to approximately scale for the differences). It is also possible to run (2.1b) in parallel with the rough search methods that might be used for initializing θ; the resulting Hessian estimate can be used to initialize (2.1b) when the full ASP method (2SPSA or 2SG) is used. Since $\hat{H}_k$ has (at most) rank 2 (and may not be positive semidefinite), having a positive-definite initialization helps provide for the invertibility of $\overline{H}_k$, especially for small k (if $\overline{H}_k$ is positive definite, $f_k(\cdot)$ in (2.1a) may be taken as the identity transformation).

2) Numerical Issues in Choice of Δ_k and $\overline{\overline{H}}_k$: Generating the elements of Δ_k according to a Bernoulli ± 1 distribution is easy and theoretically valid (and was shown to be asymptotically optimal in Sadegh and Spall [30] for basic SPSA; its potential optimality for the adaptive approach here is an open question). In some applications, however, it may be worth exploring other valid choices of distributions since the generation of Δ_k represents a trivial part of the cost of optimization, and a different choice may yield improved finite-sample performance (this was done, e.g., in Maeda and De Figueiredo [20] in "basic" SPSA). Because $\overline{H}_k$ may not be positive definite, especially for small k (even if $\overline{H}_0$ is initialized based on prior information to be positive definite), it is recommended that $\overline{H}_k$ in (2.1b) not generally be used directly in (2.1a). Hence, as shown in (2.1a), it is recommended that $\overline{H}_k$ be replaced by another matrix $\overline{\overline{H}}_k$ that is closely related to $\overline{H}_k$. One useful form when p is not too large has been to take $\overline{\overline{H}}_k = (\overline{H}_k\overline{H}_k)^{1/2} + \delta_k I$, where the indicated square root is the (unique) positive semidefinite square root and $\delta_k \geq 0$ is some small number. For large p, a more efficient method is to simply set $\overline{\overline{H}}_k = \overline{H}_k + \delta_k I$, but this is likely to require a larger δ_k to ensure positive definiteness of $\overline{\overline{H}}_k$. For very large p, it may be advantageous to have $\overline{\overline{H}}_k$ be only a diagonal matrix based on the diagonal elements of $\overline{H}_k + \delta_k I$. This is a way of capturing large scaling differences in the θ elements (unavailable to first-order algorithms) while eliminating the potentially onerous computations associated with the inverse operation in (2.1a). Note that $\overline{\overline{H}}_k$ should only be used in (2.1a), as (2.1b) should remain in terms of $\overline{H}_k$ to ensure a.s. consistency (see Theorems 2a, b in Section III). By Theorems 2a, b, one can set $\overline{\overline{H}}_k = \overline{H}_k$ for sufficiently large k. Also, for general nondiagonal $\overline{\overline{H}}_k$, it is numerically advantageous to avoid a direct inversion of $\overline{\overline{H}}_k$ in (2.1a), preferring a method such as Gaussian elimination (which, e.g., is directly available as the MATLAB "\" operator).

3) Gradient/Hessian Averaging: At each iteration, it may be desirable to compute and average several $\hat{H}_k$ and $G_k(\hat{\theta}_k)$ values despite the additional cost. This may be especially true in a high-noise environment. Also see item 5) for additional potentially useful averaging.

4) Gain Selection: The principles outlined in Brennan and Rogers [3] and Spall [36] are useful here as well for practical selection of the gain sequences $\{a_k\}$, $\{c_k\}$, and in the 2SPSA case, $\{\tilde{c}_k\}$. For 2SPSA and 2SG, the critical gain a_k can be simply chosen as $1/k, k \geq 1$ to achieve asymptotic near optimality or optimality, respectively (see Section IV-B), although this may not be ideal in practical finite-sample problems. For the remainder, let us focus on the 2SPSA case; similar ideas apply for 2SG case, but the problem is slightly easier since there is no $\{\tilde{c}_k\}$ sequence. We can choose $a_k = a/(k+A)^\alpha, c_k = c/k^\gamma$, and $\tilde{c}_k = \tilde{c}/k^\gamma, a, c, \tilde{c}, \alpha, \gamma > 0, A \geq 0$ for $k \geq 1$. In finite-sample practice, it may be better to choose α and γ lower than their asymptotically optimal values of α = 1 and γ = 1/6 (see Section IV-B), and, in particular, α = 0.602 and γ = 0.101 are practically effective and approximately the lowest theoretically valid values allowed (see Theorems 1a, 2a, and 3a in Sections III and IV). Choosing a_k so that the typical change in $\hat{\theta}_k$ to $\hat{\theta}_{k+1}$ is of "reasonable" magnitude, especially in the critical early iterations, has proven effective. Setting A approximately equal to 5–10% of the total expected number of iterations enhances practical convergence by allowing for a larger a than possible with the more typical $A = 0$. However, in slight contrast to Spall [36] for the first-order algorithm, we recommend that c have a magnitude greater (by roughly a factor of 2–10) than the typical ("one-sigma") noise level in the $y(\cdot)$ measurements. Further, setting $\tilde{c} > c$ has been effective. These recommendations for larger c (and $\tilde{c}$) values than given in Spall [36] are made due to the greater inherent sensitivity of a second-order algorithm to noise effects.

5) Blocking: At each iteration, block "bad" steps if the new estimate for θ fails a certain criterion (i.e., set $\hat{\theta}_{k+1} = \hat{\theta}_k$ in going from k to $k+1$). $\overline{H}_k$ should typically continue to be updated even if $\hat{\theta}_{k+1}$ is blocked. The most obvious blocking applies when θ must satisfy constraints; an updated value may be blocked or modified if a constraint is violated. There are two ways [5a) and 5b)] that one might implement blocking when constraints are not the limiting factor, with 5a) based on $\hat{\theta}_k$ and $\hat{\theta}_{k+1}$ directly, and 5b) based on loss measurements. Both of 5a) and 5b) may be implemented in a given application. In 5a), one simply blocks the step from $\hat{\theta}_k$ to $\hat{\theta}_{k1}$ if $\|\hat{\theta}_{k+1} - \hat{\theta}_k\| >$ tolerance$_1$, where the norm is any convenient distance measure and tolerance$_1 > 0$ is some "reasonable" maximum distance to cover in one step. The rationale behind 5a) is that a well-behaving algorithm should be moving toward the solution in a smooth manner, and very large steps are indicative of potential divergence. The second potential method, 5b), is based on blocking the step if $y(\hat{\theta}_{k+1}) > y(\hat{\theta}_k) -$ tolerance$_2$, where tolerance$_2 \geq 0$ might be set at about one or two times the approximate standard deviation of the noise in the $y(\cdot)$ measurements. In a setting where the noise in the loss measurements tends to be large (say, much larger than the allowable difference between $L(\theta^*)$ and $L(\hat{\theta}_{\text{final}})$), it may be undesirable to use 5b) due to the difficulty in obtaining meaningful information about the relative old and new loss values. For any nonzero noise levels, it may be beneficial to average several $y(\cdot)$ measurements in making the decision about whether to block the step; this may be done even if the averaging mentioned in guideline 3) is not used (then the standard deviation for choosing tolerance$_2$ should be normalized by the amount of averaging). Having tolerance$_2 > 0$ as specified above when there is noise in the $y(\cdot)$'s builds some conservativeness into the algorithm by allowing a new step only if there is relatively strong statistical evidence of an improved loss value.

Let us close this subsection with a few summary comments about the implementation aspects above. Without the second blocking procedure 5b) in use, 2SPSA requires *four* measurements $y(\cdot)$ per iteration, *regardless* of the dimension p (two for the standard $G_k(\cdot)$ estimate and two new values for the one-sided SP gradients $G_k^{(1)}(\cdot)$). For 2SG, *three* gradient measurements $G_k(\cdot)$ are needed, again independent of p. If the second blocking procedure 5b) is used, one or more additional $y(\cdot)$ measurements are needed for both 2SPSA and 2SG. The use of gradient/Hessian averaging 3) would increase the number of loss or gradient evaluations, of course. The standard deviation for the measurement noise (used in items 4) and 5b)) can be estimated by collecting several $y(\cdot)$ values at $\theta = \hat{\theta}_0$; neither 4) nor 5a) requires this estimate to be precise (so relatively few $y(\cdot)$ values are needed). In general, 5a) can be used anytime, while 5b) is more appropriate in a low- or no-noise setting. Note that 5a) helps to prevent divergence, but lacks direct insight into whether the loss function is improving, while 5b) does provide that insight, but requires additional $y(\cdot)$ measurements, the number of which might grow prohibitively in a high-noise setting.

III. STRONG CONVERGENCE

This section presents results related to the strong (a.s.) convergence of $\hat{\theta}_k \to \theta^*$ and $\overline{H}_k \to H(\theta^*)$ (all limits are as $k \to \infty$ unless otherwise noted). This section establishes separate results for 2SPSA and for 2SG. One of the challenges, of course, in establishing convergence is the coupling between the recursions for $\hat{\theta}_k$ and $\overline{H}_k$. We present a martingale approach that seems to provide a relatively simple solution with reasonable regularity conditions. Alternative conditions for convergence might be available using the ordinary differential equation approach of Metivier and Priouret [22] and Benveniste *et al.* [1, ch. II.1], which includes a certain Markov dependence that would, in principle, accommodate the recursion coupling. However, this approach was not pursued here due to the difficulty of checking certain regularity conditions associated with the Markov dependence (e.g., those related to the solution of the "Poisson equation").

The results below are in two parts, with the first part (Theorems 1a, b) establishing conditions for the convergence of $\hat{\theta}_k$, and the second part (Theorems 2a, b) doing the same for $\overline{H}_k$. The proofs of the theorems are in Appendix A. We let $\|\cdot\|$ denote the standard Euclidean vector norm or compatible matrix spectral norm (as appropriate), $(\theta^*)_i$ and $(\theta-\theta^*)_i$ represent the ith components of the indicated vectors (notation chosen to avoid confusion with the iteration subscript k), i.o. represent infinitely often, and $\overline{g}_k(\hat{\theta}_k) \equiv \overline{\overline{H}}_k^{-1} g(\hat{\theta}_k)$. Below are some regularity conditions that will be used in Theorem 1a for 2SPSA and, in part, in the succeeding theorems. Some comments on the practical implications of the conditions are given immediately following their statement. Appendix B provides some additional comments on the relationship of the conditions here to the conditions of other adaptive approaches mentioned in Section I.

Note that some conditions show a dependence on $\hat{\theta}_k$ and $\overline{H}_k$, the very quantities for which we are showing convergence. Although such "circularity" is generally undesirable, it is fairly common is the SA field (e.g., Kushner and Yin [15, Theorem 5.2.1], Benveniste *et al.* [1, p. 238]). Appendix B elaborates on the circularity issue relative to conditions in other adaptive algorithms. The Appendix points out that adaptive algorithms without circularity conditions have *other* conditions that are difficult to check and/or easily violated. The inherent difficulty in establishing theoretical properties of adaptive approaches comes from the need to couple the estimates for the parameters of interest and for the Hessian (Jacobian) matrix. Note that the bulk of the conditions here showing dependence on $\hat{\theta}_k$ and $\overline{H}_k$ are conditions on the measurement noise and smoothness of the loss function (C.0, C.2, and C.3 below; C.0′, C.2′, C.3′, C.8, and C.8′ in later theorems); the explicit dependence on $\hat{\theta}_k$ can be removed by assuming that the relevant condition holds uniformly for all "reasonable" θ. The dependence in C.5 is handled in the lemma below.

- C.0: $E(\varepsilon_k^{(+)} - \varepsilon_k^{(-)} | \hat{\theta}_k; \Delta_k; \overline{H}_k) = 0$ a.s. $\forall k$, where $\varepsilon_k^{(\pm)}$ is the effective SA measurement noise, i.e., $\varepsilon_k^{(\pm)} \equiv y(\hat{\theta}_k \pm c_k\Delta_k) - L(\hat{\theta}_k \pm c_k\Delta_k)$.
- C.1: $a_k, c_k > 0\ \forall k; a_k \to 0, c_k \to 0$ as $k \to \infty$; $\sum_{k=0}^{\infty} a_k = \infty, \sum_{k=0}^{\infty}(a_k/c_k)^2 < \infty$.
- C.2: For some $\delta, \rho > 0$ and $\forall k, \ell, E(|y(\hat{\theta}_k \pm c_k\Delta_k)/\Delta_{k\ell}|^{2+\delta}) \le \rho, |\Delta_{k\ell}| \le \rho, \Delta_{k\ell}$ is symmetrically distributed about 0, and $\{\Delta_{k\ell}\}$ are mutually independent.
- C.3: For some $\rho > 0$ and almost all $\hat{\theta}_k$, the function $g(\cdot)$ is continuously twice differentiable with a uniformly (in k) bounded second derivative for all θ such that $\|\hat{\theta}_k - \theta\| \le \rho$.
- C.4: For each $k \ge 1$ and all θ, there exists a $\rho > 0$ not dependent on k and θ, such that $(\theta-\theta^*)^T \overline{g}_k(\theta) \ge \rho\|\theta - \theta^*\|$.
- C.5: For each $i = 1, 2, \cdots, p$ and any $\rho > 0$, $P(\{\overline{g}_{ki}(\hat{\theta}_k) \ge 0 \text{ i.o.}\} \cap \{\overline{g}_{ki}(\hat{\theta}_k) < 0 \text{ i.o.}\} | \{|\hat{\theta}_{ki} - (\theta^*)_i| \ge \rho\ \forall k\}) = 0$ (see lemma for sufficient conditions).
- C.6: $\overline{\overline{H}}_k^{-1}$ exists a.s. $\forall k, c_k^2 \overline{\overline{H}}_k^{-1} \to 0$ a.s., and for some $\delta, \rho > 0, E(\|\overline{\overline{H}}_k^{-1}\|^{2+\delta}) \le \rho$.
- C.7: For any $\tau > 0$ and nonempty $S \subseteq \{1, 2, \cdots, p\}$, there exists a $\rho'(\tau, S) > \tau$ such that

$$\limsup_{k\to\infty} \left| \frac{\sum_{i \notin S} (\theta - \theta^*)_i \overline{g}_{ki}(\theta)}{\sum_{i \in S} (\theta - \theta^*)_i \overline{g}_{ki}(\theta)} \right| < 1 \quad \text{a.s.}$$

for all $|(\theta-\theta^*)_i| < \tau$ when $i \notin S$ and $|(\theta-\theta^*)_i| \ge \rho'(\tau, S)$ when $i \in S$ (see lemma for sufficient conditions).

Comments on Conditions C.0–C.7: C.0 and C.1 are common martingale-difference noise and gain sequence conditions. C.2 provides a structure to ensure that the gradient approximations $G(\cdot)$ and $G_k^{(1)}(\cdot)$ are well behaved. The conditions on Δ_k are very close to those for "basic" SPSA, and would usually *exclude* Δ_k from being uniformly or normally distributed due to their violation of the implied finite inverse moments condition in $E(|y(\hat{\theta}_k \pm c_k\Delta_k)/\Delta_{k\ell}|^{2+\delta}) \le \rho$ (note that Holder's inequality makes the finite inverse moment condition explicit since the expectation of interest exists if $E(|y(\hat{\theta}_k \pm c_k\Delta_k)|^{2+\delta'})$ and $E(|1/\Delta_{k\ell}|^{2+\delta''})$ are uniformly bounded for $\delta', \delta'' > \delta$). An independent Bernoulli ± 1 distribution is frequently used for the elements of Δ_k as discussed in Section II-D. C.3 and C.4 provide basic assumptions about the smoothness and steepness of $L(\theta)$. C.3 holds, of course, if $g(\theta)$ is twice continuously differentiable with a bounded second derivative on R^p. C.5 is a modest condition that says that $\hat{\theta}_k$ cannot be bouncing around in a manner that causes the signs of the normalized gradient elements to be changing an infinite number of times if $\hat{\theta}_k$ is uniformly bounded away from θ^* (see the sufficient conditions below). C.6 provides some conditions on the surrogate for the Hessian estimate that appears in (2.1). Since the user has full control over the definition of $\overline{\overline{H}}_k$, these conditions should be relatively easy to satisfy. Note that the middle part of C.6 ($\overline{\overline{H}}_k^{-1} = o(c_k^{-2})$ a.s.) allows for $\overline{\overline{H}}_k^{-1}$ to "occasionally" be large provided that the boundedness of moments in the last part of the condition is satisfied. The example for $\overline{\overline{H}}_k$ given in Section II-D [guideline 2)] would satisfy this potential growth condition, for instance, if $\delta_k = c_k^{\rho}, 0 < \rho < 2$. Finally, C.7 ensures that, for k sufficiently large, each element of $\overline{g}_k(\theta)$ tends to make a non-negligible contribution to products of the form $(\theta-\theta^*)^T \overline{g}_k(\theta)$ (see C.4). A sufficient condition for C.7 is that, for each $i, \overline{g}_{ki}(\theta)$ be uniformly (in k) bounded > 0 and $< \infty$ when $|(\theta-\theta^*)_i|$ is bounded away from 0 for all i; C.7 is unnecessary when $\hat{\theta}_k$ is bounded as stated in the lemma below. Note that, although no explicit conditions are shown on $\{\tilde{c}_k\}$, there are implicit conditions in C.4–C.7 given $\tilde{c}_k$'s effect on $\overline{\overline{H}}_k$ (via $\overline{H}_k$). In Theorem 2a on the convergence of $\overline{H}_k$, there are explicit conditions on $\{\tilde{c}_k\}$.

Conditions C.5 and C.7 are relatively unfamiliar. So, before showing the main theorems on convergence for 2SPSA and 2SG, we give sufficient conditions for these two conditions in the lemma below. The main sufficient condition is the well-known boundedness condition on the SA iterate (e.g., Kushner and Yin [15, Theorem 5.2.1], Benveniste *et al.* [1, Theorem II.15]). Although some authors have relaxed this boundedness condition (e.g., Gerencsér [12]), the condition imposes no *practical* limitation. This boundedness condition also formally eliminates the need for the explicit dependence of other conditions (C.2 and C.3 above; C.0$'$, C.2$'$, C.3$'$, C.8, and C.8$'$ below) on $\hat{\theta}_k$ since the conditions can be restated to hold for all θ in the bounded set containing $\hat{\theta}_k$. Note also that the condition $a_k/c_k^2 \to 0$ holds automatically for gains in the standard form discussed in Section IV. One example of when the remaining condition of the lemma, (3.1), is trivially satisfied is when $\overline{\overline{H}}_k$ is chosen as a diagonal matrix, as suggested in guideline 2) of Section II-D.

Lemma—Sufficient Conditions for C.5 and C.7: Assume that C.1–C.4 and C.6 hold, and $\limsup_{k\to\infty} \|\hat{\theta}_k\| < \infty$ a.s. Then condition C.7 is not needed. Further, let $a_k/c_k^2 \to 0$, and suppose that, for any $\rho > 0$,

$$P(\text{sign } \overline{g}_{ki}(\hat{\theta}_k) \neq \text{sign } g_i(\hat{\theta}_k) \text{ i.o.} \,\Big|\, |\hat{\theta}_{ki} - (\theta^*)_i| \geq \rho) = 0 \qquad \forall i. \quad (3.1)$$

Then C.5 is automatically satisfied.

Theorem 1a—2SPSA: Consider the SPSA estimate for $G_k(\cdot)$ with $G_k^{(1)}(\cdot)$ given by (2.4). Let conditions C.0–C.7 hold. Then $\hat{\theta}_k - \theta^* \to 0$ a.s.

Theorem 1b below on the 2SG approach is a straightforward modification of Theorem 1a on 2SPSA. We replace C.0, C.1, and C.2 with the following SG analogs. Equalities hold a.s. where needed.

C.0$'$: $E(e_k|\hat{\theta}_k; \Delta_k; \overline{H}_k) = 0$ where $e_k = G_k(\hat{\theta}_k) - g(\hat{\theta}_k)$.

C.1$'$: $a_k > 0\ \forall k; a_k \to 0; \sum_{k=0}^{\infty} a_k = \infty, \sum_{k=0}^{\infty} a_k^2 < \infty$.

C.2$'$: For some $\delta, \rho > 0, E(\|G_k(\hat{\theta}_k)\|^{2+\delta}) \leq \rho\ \forall k$.

Comments on C.0$'$–C.2$'$: Note (analogous to $\{\tilde{c}_k\}$ in Theorem 1a) that there are no explicit conditions on $\{c_k\}$ here. These conditions are implicit via the conditions on $\overline{\overline{H}}_k$, and will be made explicit when we consider the convergence of $\overline{H}_k$ in Theorem 2b.

Theorem 1b—2SG: Consider the setting where $G_k(\cdot)$ is a direct measurement of the gradient. Suppose that C.0$'$–C.2$'$ and C.3–C.7 hold. Then $\hat{\theta}_k - \theta^* \to 0$ a.s.

Theorem 2a below treats the convergence of $\overline{H}_k$ in the SPSA case. We introduce several new conditions as follows, which are largely self-explanatory:

C.1$''$: The conditions of C.1 hold plus $\sum_{k=0}^{\infty}(k+1)^{-2}(c_k\tilde{c}_k)^{-2} < \infty$ with $\tilde{c}_k = O(c_k)$.

C.3$'$: Change "thrice differentiable" in C.3 to "four-times differentiable" with all else unchanged.

C.8: For some $\rho > 0$ and all k, ℓ, m,

$$E[y(\hat{\theta}_k \pm c_k\Delta_k + \tilde{c}_k\tilde{\Delta}_k)^2/(\Delta_{k\ell}\tilde{\Delta}_{km})^2] \leq \rho$$

and

$$E[y(\hat{\theta}_k \pm c_k\Delta_k)^2/(\Delta_{k\ell}\Delta_{km})^2] \leq \rho;$$
$$E(\tilde{\varepsilon}_k^{(\pm)} - \varepsilon_k^{(\pm)}|\hat{\theta}_k; \tilde{\Delta}_k; \overline{H}_k) = 0$$

and

$$E[(\tilde{\varepsilon}_k^{(\pm)} - \varepsilon_k^{(\pm)})^2/(\Delta_{k\ell}\tilde{\Delta}_{km})^2] \leq \rho$$

where $\tilde{\varepsilon}_k^{(\pm)} = y(\hat{\theta}_k \pm c_k\Delta_k + \tilde{c}_k\tilde{\Delta}_k) - L(\hat{\theta}_k \pm c_k\Delta_k + \tilde{c}_k\tilde{\Delta}_k)$.

C.9: $\tilde{\Delta}_k$ satisfies the assumptions for Δ_k in C.2 (i.e., $\forall k, \ell, |\tilde{\Delta}_{k\ell}| \leq \rho$ and $\tilde{\Delta}_{k\ell}$ is symmetrically distributed about 0; $\{\tilde{\Delta}_{k\ell}\}$ are mutually independent); Δ_k and $\tilde{\Delta}_k$ are independent; $E(\Delta_{k\ell}^{-2}) \leq \rho$, $E(\tilde{\Delta}_{k\ell}^{-2}) \leq \rho\ \forall k, \ell$ and some $\rho > 0$.

Theorem 2a—2SPSA: Let conditions C.0, C.1$''$, C.2, C.3$'$, and C.4–C.9 hold. Then, $\overline{H}_k \to H(\theta^*)$ a.s.

Our final strong convergence result is for the Hessian estimate in 2SG. As above, we introduce some additional modified conditions.

C.1$'''$: The conditions of C.1$'$ hold plus $c_k > 0$, $c_k \to 0$, and $\sum_{k=0}^{\infty}(k+1)^{-2}c_k^{-2} < \infty$.

C.8$'$: For some $\rho > 0$ and all k, ℓ,

$$E(\|g(\hat{\theta}_k \pm c_k\Delta_k)/\Delta_{k\ell}\|^2) \leq \rho$$
$$E(\|(e_k^{(+)} - e_k^{(-)})/\Delta_{k\ell}\|^2) \leq \rho$$

and

$$E((e_k^{(+)} - e_k^{(-)})/\Delta_{k\ell}|\hat{\theta}_k) = 0$$

where $e_k^{(\pm)} = G_k(\hat{\theta}_k \pm c_k\Delta_k) - g(\hat{\theta}_k \pm c_k\Delta_k)$.

C.9$'$: For some $\rho > 0$ and all $k, \ell, |\Delta_{k\ell}| \leq \rho, \Delta_{k\ell}$ is symmetrically distributed about 0, $\{\Delta_{k\ell}\}$ are mutually independent, and $E(\Delta_{k\ell}^{-2}) \leq \rho$.

Comments on C.1$'''$, C.8$'$, C.9$'$: Unlike this theorem's companion result for 2SG (Theorem 1b), explicit conditions are necessary on $\{c_k\}$ to control the convergence of the Hessian iteration. Note that due to the simpler structure of 2SG (versus 2SPSA), the conditions in C.9$'$ are a subset of the conditions in C.9 for Theorem 2a.

Theorem 2b—2SG: Suppose that C.0$'$, C.1$'''$, C.2$'$, C.3$'$, C.4–C.7, C.8$'$, and C.9$'$ hold. Then $\overline{H}_k \to H(\theta^*)$ a.s.

IV. Asymptotic Distributions and Efficiency Analysis

A. Asymptotic Distributions of ASP Iterate

This subsection builds on the convergence results in the previous section, establishing the asymptotic normality of the 2SPSA and 2SG formulations of ASP. The asymptotic normality is then used in Section IV-B to analyze the asymptotic efficiency of the algorithms. Proofs are in Appendix A.

2SPSA Setting: As before, we consider 2SPSA before 2SG. Asymptotic normality or the related issue of convergence of moments in basic first-order SPSA has been established under slightly differing conditions by Spall [33], Chen *et al.* [4], Dippon and Renz [7], Kushner and Yin [15, ch. 10], and Gerencsér [12]. We consider gains of the typical form $a_k = a/(k+A)^\alpha$ and $c_k = c/k^\gamma, a, c, \alpha, \gamma > 0, A \geq 0, k \geq 1$, and take $\beta = \alpha - 2\gamma, \rho^2 = E(\Delta_{ki}^{-2}), \xi^2 = E(\Delta_{ki}^2)\ \forall k, i$. The asymptotic mean below relies on the third derivative of $L(\theta)$; we let $L_{ijk}^{(3)}(\theta^*)$ represent the third derivative of L with respect to elements i, j, k of θ evaluated at θ^*. The following regularity conditions will be used in the asymptotic normality result.

C.10: $E(\varepsilon_k^{(+)} - \varepsilon_k^{(-)})^2|\hat{\theta}_k, \overline{H}_k) \to \sigma^2$ a.s. for some $\sigma^2 > 0$. For almost all $\hat{\theta}_k$, $\{E((\varepsilon_k^{(+)} - \varepsilon_k^{(-)})^2|\hat{\theta}_k, c_k\Delta_k = \eta)\}$ is an equicontinuous sequence at η = 0, and is continuous in η on some compact, connected set containing the actual (observed) value of $c_k\Delta_k$ a.s.

C.11: In addition to implicit conditions an α and γ via C.1$''$, $3\gamma - \alpha/2 \geq 0$ and $\beta > 0$. Further, when $\alpha = 1, a > \beta/2$. Let $f_k(\cdot)$ in (2.1a) be chosen such that $\overline{\overline{H}}_k - \overline{H}_k \to 0$ a.s.

Comments on C.10 and C.11: Although, in some applications, the "$\to$" for the noise second moments in C.10 may be replaced by "=," the limiting operation allows for a more general setting, and is relevant in the example of Section V. Since the user has full control over $f_k(\cdot)$, it is not difficult to guarantee in C.11 that $\overline{\overline{H}}_k - \overline{H}_k \to 0$ a.s.; most examples in Section II-D2) satisfy this condition.

Theorem 3a—2SPSA: Suppose that C.0, C.1$''$, C.2, C.3$'$, and C.4–C.9 hold (implying convergence of $\hat{\theta}_k$ and $\overline{H}_k$). Then, if C.10 and C.11 hold and $H(\theta^*)^{-1}$ exists,

$$k^{\beta/2}(\hat{\theta}_k - \theta^*) \xrightarrow{\text{dist}} N(\mu, \Omega) \tag{4.1}$$

where $\mu = \{0$ if $3\gamma - \alpha/2 > 0$; $H(\theta^*)^{-1}T/(a - \beta_+/2)$ if $3\gamma - \alpha/2 = 0\}$, the jth element of T is

$$-\tfrac{1}{6}ac^2\xi^2 \left[L^{(3)}_{jjj}(\theta^*) + 3 \sum_{\substack{i=1 \\ i\neq j}}^{p} L^{(3)}_{iij}(\theta^*) \right] \tag{4.2}$$

$\Omega = a^2c^{-2}\sigma^2\rho^2 H(\theta^*)^{-2}/(8a - 4\beta_+)$, and $\beta_+ = \beta$ if $\alpha = 1$ and $\beta_+ = 0$ if $\alpha < 1$.

2SG Setting: We now consider the 2SG setting of direct bias-free gradient measurements. There are a number of references on the asymptotic distribution and/or moments of second-order SG algorithms when the Hessian is estimated adaptively in particular ways (e.g., Nevel'son and Has'minskii [23, ch. 7], Fabian [11], Ruppert [27], Wei [43], Benveniste *et al.* [1, pp. 115–116], Ljung [17], and Walk [42]). These references show that the asymptotic properties—such as distribution—of the adaptive algorithms are identical to those that would result from using the true (unknown) Hessian. We will do likewise for 2SG implementation. As above, consider gains of the typical form $a_k = a/(k + A)^\alpha$.

Before introducing the asymptotic normality result, we introduce an additional regularity condition.

C.12: $E(e_k e_k^T | \hat{\theta}_k, \overline{H}_k) \to \Sigma$ a.s. for some positive semidefinite matrix $\Sigma, a > 1/2$ if α = 1, and $f_k(\cdot)$ is chosen such that $\overline{\overline{H}}_k - \overline{H}_k \to 0$ a.s.

Comments on C.12: As with C.10, frequently, "$\to$" can be replaced with "=" in the limiting covariance expression. Likewise, see the comments following C.11 regarding the condition $\overline{\overline{H}}_k - \overline{H}_k \to 0$ a.s.

Theorem 3b—2SG: Suppose that C.0$'$, C.1$'''$, C.2$'$, C.3$'$, C.4–C.7, C.8$'$, and C.9$'$ hold (implying convergence of $\hat{\theta}_k, \overline{H}_k$) and that C.12 holds with $H(\theta^*)^{-1}$ existing. Then,

$$k^{\alpha/2}(\hat{\theta}_k - \theta^*) \xrightarrow{\text{dist}} N(0, \Omega') \tag{4.3}$$

where $\Omega' = a^2 H(\theta^*)^{-1}\Sigma H(\theta^*)^{-1}/(2a - \beta_+)$ with $\beta_+ = 1$ if $\alpha = 1$ and $\beta_+ = 0$ if $\alpha < 1$.

B. *Efficiency Analysis*

Using the distribution results in Section IV-A, we now analyze the asymptotic efficiency of the second-order approaches. For convenience here and in Section V, let 1SPSA and 1SG denote the standard first-order SPSA and SG algorithms (to contrast with 2SPSA and 2SG).

2SPSA Setting: From Theorem 3a, the root-mean-squared (rms) error from the asymptotic distribution of the normalized error $k^{\beta/2}(\hat{\theta}_k - \theta^*)$ is

$$\text{rms}_{\text{2SPSA}}(a, \alpha, c, \gamma) = [\mu^T\mu + \text{trace}(\Omega)]^{1/2} \tag{4.4}$$

where the arguments emphasize the dependence on the gain sequence coefficients (coefficient A, of course, does not affect the asymptotic distribution). (Under some additional conditions—e.g., Gerencsér [12]—the asymptotic distribution-based rms error in (4.4) is equal to $[E(\|k^{\beta/2}(\hat{\theta}_k - \theta^*)\|^2)]^{1/2}$.) To analyze the asymptotic efficiency, we compare $\text{rms}_{\text{2SPSA}}$ with a corresponding quantity based on standard 1SPSA. Let $\text{rms}_{\text{1SPSA}}(a, \alpha, c, \gamma)$ denote the rms error from the asymptotic distribution for 1SPSA, as given, e.g., in Spall [33, Prop. 2].

Dippon and Renz [7] pursue a line of reasoning close to that above in comparing the iterate averaging version of SPSA with optimal versions of 1SPSA. In particular, the rms error in (4.4) with $a = \alpha = 1, \gamma = 1/6$ is identical to the rms error for iterate averaging (note that $\alpha = 1, \gamma = 1/6$ is asymptotically optimal for both 1SPSA and 2SPSA since they maximize the rate of convergence $k^{-\beta/2}$ under the constraints on a_k, c_k). Then, based on Dippon and Renz [7, expressions (5.2) and (5.3)] and assuming the same number of iterations in both 1SPSA and 2SPSA, we have

$$\frac{\text{rms}_{\text{2SPSA}}(1, 1, c, \frac{1}{6})}{\min\limits_{3a > 1/\lambda_{\min}} \text{rms}_{\text{1SPSA}}(a, 1, c, \frac{1}{6})} < 2 \qquad \forall c > 0 \tag{4.5a}$$

$$\frac{\min\limits_{c>0} \text{rms}_{\text{2SPSA}}(1, 1, c, \frac{1}{6})}{\min\limits_{2a > 1/\lambda_{\min}} \min\limits_{c>0} \text{rms}_{\text{1SPSA}}(1, 1, c, \frac{1}{6})} < 2 \tag{4.5b}$$

where $\lambda_{\min}$ is the minimum eigenvalue of $H(\theta^*)$. The interpretation of (4.5a), (4.5b) is as follows. From (4.5a), we know that, for any common value of c, the asymptotic rms error of 2SPSA is less than twice that of 1SPSA with an optimal a (even when c is chosen optimally for 1SPSA). Expression (4.5b) states that, if we optimize only c for 2SPSA, while optimizing both a and c for 1SPSA, we are still guaranteed that the asymptotic rms error for 2SPSA is no more than twice the optimized rms error for 1SPSA. Another interesting aspect of 2SPSA is the relative robustness apparent in (4.5a), (4.5b) given that the optimal a for 1SPSA will not typically be known in practice. For certain suboptimal values of a in 1SPSA, the rms error can get very large whereas simply choosing a = 1 for 2SPSA provides the factor-of-2 guarantee mentioned above.

Although (4.5a), (4.5b) suggest that the 2SPSA approach yields a solution that is quite good, one might wonder if a true optimal solution is possible. Dippon and Renz [7, pp. 1817–1818] pursue this issue, and provide an alternative to

$H(\theta^*)^{-1}$ as the limiting weighting matrix for use in an SA form such as (2.1a). Unfortunately, this limiting matrix has no closed-form solution, and depends on the third derivatives of $L(\theta)$ at θ^*, and furthermore, it is not apparent how one would construct an adaptive matrix (analogous to $\overline{H}_k$) that would converge to this optimal limiting matrix. Likewise, the optimal c for 2SPSA is typically unavailable in practice since it also depends on the third derivatives of $L(\theta)$.

Expressions (4.5a), (4.5b) are based on an assumption that 1SPSA and 2SPSA have used the same number of iterations. This is a reasonable basis for a core comparison since the "cost" of solving for the optimal 1SPSA gains is unknown. However, a more conservative representation of relative efficiency is possible by considering only the direct number of loss measurements, ignoring the extra cost for optimal gains in 1SPSA. In particular, 1SPSA uses two loss measurements per iteration and 2SPSA uses four measurements per iteration. Hence, with both algorithms using the same number of loss measurements, the corresponding upper bounds to the ratios in (4.5a), (4.5b) (reflecting the ratio of rms errors as the common number of loss measurements gets large) would be $2^{4/3} \approx 2.52$, an increase from the bound of 2 under a common number of iterations. This bound's likely excessive conservativeness follows from the fact that the cost of solving for the optimal gains in 1SPSA is being ignored. Note that, for other adaptive approaches that are also asymptotically normally distributed, the same relative cost analysis can be used. Hence, for example, with the Fabian [10] approach using $O(p^2)$ measurements per iteration to generate the Hessian estimate, the corresponding upper bounds would be of magnitude $O(p^{2/3})$, bounds that, unlike the bounds for 2SPSA, increase with problem dimension.

2SG Setting: The SG case is more straightforward than the above. By minimizing the asymptotic rms error, it is well known that the optimal gain is $H(\theta^*)^{-1}/k$ (e.g., Wei [43], Ruppert [28], and Kushner and Yin [15, p. 289]). Then $\text{rms}_{\text{2SG}}(a,\alpha) = \text{trace}(\Omega')^{1/2} = a[\text{trace}(H(\theta^*)^{-1}\Sigma H(\theta^*)^{-1})/(2a-\beta_+)]^{1/2}$, as derived from (4.3) in Theorem 3b. Setting $a = \alpha = 1$ yields an rms error for the adaptive algorithm (2.1a), (2.1b) that is identical to that obtained by using the idealized optimal gain $H(\theta^*)^{-1}/k$. In particular, $\text{rms}_{\text{2SG}}(1,1) = [\text{trace}(H(\theta^*)^{-1}\Sigma H(\theta^*)^{-1})]^{1/2}$. Hence, rms ratios for SG analogous to (4.5a), (4.5b) (rms for 2SG over the optimal RMS for 1SG) have the value 1. As in 2SPSA above, this ratio is for a common number of iterations. If this ratio is expressed based on a common number of gradient measurements (reflecting the fact that three gradient measurements per iteration are used for 2SG versus one gradient measurement per iteration for 1SG), then the ratio of asymptotic rms errors for 2SG over 1SG is $\sqrt{3} \approx 1.73$. Although this ratio is likely to be overly conservative since it ignores the cost of solving for the optimal gains in 1SG, it is enlightening relative to fundamental limits. Also, in parallel with the analysis for 2SPSA, the ratio based on using one of the previous adaptive approaches (e.g., Wei [43] or Ruppert [27]) instead of 2SG shows the detrimental effects of increasing p. In particular, with the Wei or Ruppert adaptive approaches using $2p$ gradient measurements per iteration to generate the Hessian estimate, the corresponding upper bounds to the ratio would be equal to $(2p+1)^{1/2}$, bounds that, unlike the bounds for 2SG (i.e., which are equal to $\sqrt{3}$), increase with problem dimension.

TABLE I
NORMALIZED LOSS VALUES FOR 1SPSA AND 2SPSA WITH σ = 0.001; 90% CONFIDENCE INTERVAL SHOWN IN $[\cdot]$

No. of loss measurements	1SPSA	1SPSA with iterate averaging	2SPSA
2000	0.0046 [0.0040, 0.0052]	0.0047 [0.0040, 0.0054]	0.0023 [0.0021, 0.0025]
10,000	0.0023 [0.0021, 0.0025]	0.0023 [0.0021, 0.0025]	8.6×10^{-4} [7.6×10^{-4}, 9.6×10^{-4}]

V. NUMERICAL STUDIES

This section compares 2SPSA and 2SG with their corresponding first-order "standard" forms (1SPSA and 1SG). Numerical studies on other functions are given in Spall [35], Luman [18], and Vande Wouwer *et al.* [40]. The loss function considered here is a fourth-order polynomial with p = 10, significant variable interaction, and highly skewed level surfaces (the ratio of maximum to minimum eigenvalue of $H(\theta^*)$ is approximately 65). Gaussian noise is added to the $L(\cdot)$ or $g(\cdot)$ evaluations as appropriate. MATLAB software was used to carry out this study. The loss function is

$$L(\theta) = \theta^T A^T A\theta + 0.1 \sum_{i=1}^{p} (A\theta)_i^3 + 0.01 \sum_{i=1}^{p} (A\theta)_i^4 \quad (5.1)$$

where $(\cdot)_i$ represents the ith component of the argument vector (as in Section III) and A is such that pA is an upper triangular matrix of ones. The minimum occurs at $\theta^* = 0$ with $L(\theta^*) = 0$. The noise in the loss function measurements at any value of θ is given by $[\theta^T, 1]z$ where $z \sim N(0, \sigma^2 I_{11\times 11})$ is independently generated at each θ. This is a relatively simple noise structure representing the usual scenario where the noise values in $y(\cdot)$ depend on θ (and are therefore dependent over iterations); the z_{11} term provides some degree of independence at each noise contribution, and ensures that $y(\cdot)$ always contains noise of variance at least σ^2 (even if θ = 0). Guidelines 1), 2), 4), and 5) from Section II-D were applied here.

A fundamental philosophy in the comparisons below is that the loss function and gradient measurements are the dominant cost in the optimization process; the other calculations in the algorithms are considered relatively unimportant. This philosophy is consistent with most complex stochastic optimization problems where the loss function or gradient measurement may represent a large-scale simulation or a physical experiment. The relatively simple loss function here, of course, is merely a proxy for the more complex functions encountered in practice.

2SPSA Versus 1SPSA Results: Spall [37] provides a thorough numerical study based on the loss function (5.1). Three noise levels were considered: σ = 0.10, 0.001, and 0. The results here are a condensed study based on the same loss function. Table I shows results for the low-noise (σ = 0.001) case. The table shows the mean terminal loss value after 50 independent experiments, where the values are normalized (divided) by

TABLE II
NORMALIZED LOSS VALUES AND 90% CONFIDENCE INTERVALS FOR 1SG AND 2SG WITH σ = 0.10

1SG	1SG with iterate averaging	2SG: "High cost" loss measurements	2SG: "Low cost" loss measurements
1.1×10^{-4} $[1.0\times10^{-4}, 1.2\times10^{-4}]$	1.1×10^{-4} $[1.0\times10^{-4}, 1.2\times10^{-4}]$	7.4×10^{-4} $[6.3\times10^{-4}, 8.5\times10^{-4}]$	2.2×10^{-5} $[1.7\times10^{-5}, 2.7\times10^{-5}]$

$L(\hat{\theta}_0)$. Approximate 90% confidence intervals are shown below each mean loss value. Relative to guideline 4), the gains a_k, c_k, and $\tilde{c}_k$ decayed at the rates $1/k^{0.602}$, $1/k^{0.101}$, and $1/k^{0.101}$, respectively. These decay rates are approximately the slowest allowed by the theory and are slower than the asymptotically optimal values discussed in Section IV (which do not tend to work as well in finite-sample practice). Three separate algorithms are shown: basic 1SPSA with the coefficients of the slowly decaying gains mentioned above chosen empirically according to Spall [36], the same 1SPSA algorithm but with final estimate taken as the iterate average of the last 200 iterations, and 2SPSA. Additional study details are as in Spall [37].

We see that 2SPSA provides a considerable reduction in the loss function value for the same number of measurements used in 1SPSA.[7] Based on the numbers in the table together with supplementary studies, we find that 1SPSA needs approximately five–ten times the number of function evaluations used by 2SPSA to reach the levels of accuracy shown. The behavior of iterate averaging was consistent with the discussion in Section I in which the 1SPSA iterates had not yet settled into bouncing roughly uniformly around the solution. Numerical studies in Spall [37] show that 2SPSA outperforms 1SPSA even more strongly in the noise-free (σ = 0) case for this loss function, but that it is inferior to 1SPSA in the high-noise (σ = 0.10) case. However, Spall [37] presents a study based on a larger number of loss measurements (i.e., more asymptotic) where 2SPSA outperforms 1SPSA in the high-noise case. In addition, studies by other authors (e.g., simulation-based optimization in Luman [18], or neural network-based training in Vande Wouver *et al.* [40]) show that, for other loss functions, 2SPSA can outperform 1SPSA in high-noise settings with only a moderate number of loss measurements. The Luman [18] study is one where the transform invariance property of second-order algorithms is particularly useful given the large scaling differences among the elements of θ.

2SG Versus 1SG Results: We also examined the ASP approach as it applies in the SG setting with loss function (5.1) and the noise model above. Given this model, the noise in the gradient measurements is independently $N(0, \sigma^2 I_{10\times10})$ distributed. Consistent with the theory in Section IV, this study uses the asymptotically optimal $a_k = 1/k$ form for the gain (and from Theorem 2b, we chose c_k to correspondingly have the form $c/k^{0.499}$). Although this eases the implementation of the algorithm (since the critical gain sequence $\{a_k\}$ no longer has to be empirically determined), it likely limits the performance of the algorithm for the finite samples of interest [the gains for 1SG, on the other hand, were approximately optimized numerically as in the 2SPSA versus 1SPSA study, and used the slower decay form $a/(k+A)^{0.501}$]. *Hence, the results presented here should be considered a conservative representation of possible performance for 2SG.* Aside from this asymptotically optimal choice of gain, the same experimental setup reported in Spall [37] was used. The comparison between algorithms in the SG case is complicated by the mix of both loss and gradient measurements used in the algorithms, and the need to compare accuracy for the same overall "cost" of the optimization (as mentioned above, only the loss and gradient measurements are considered relevant to the cost here).

We report results in Table II for the high-noise (σ = 0.10) case. 1SG (unaveraged and averaged) used only gradient measurements, while 2SG used gradient measurements and [for the blocking step 5b)] loss measurements. All results are based on 5000 "gradient equivalents" for the algorithm budgets (so that 5000 iterations of 1SG is the same number of iterations as 1SPSA with 10 000 loss measurements). A gradient equivalent represents either a gradient measurement or some number of loss measurements. We consider two cases, one where the cost of a loss measurement is so high that it is undesirable to invoke blocking step 5b) due to the relatively high noise levels, and another case where the cost is negligible compared to a gradient measurement. In the former ("high-cost") setting, 2SG used three gradient measurements and no loss measurements at each iteration. In the latter ("low-cost") setting, it was assumed that one could obtain enough loss measurements so that, at a cost equivalent to one gradient measurement, one could effectively average out the noise in the loss values used in the blocking step 5b). The 2SG approach is inferior when the loss measurements are costly, and superior when the loss measurements are significantly cheaper than the gradient measurements.

Other studies have been conducted with 2SG. For example, Vande Wouver *et al.* [40] show an approximate order of magnitude reduction (relative to 1SG/backpropagation) in loss value in a neural-network training problem. For loss function (5.1), the performance of 2SG relative to 1SG improves when σ gets smaller. In fact, in the no-noise (σ = 0) setting (such as in system identification applications where one has exact information about the gradient of the loss function) with only 500 gradient equivalents (versus 5000 above), 2SG produces loss values of order 10^{-8}, about two orders of magnitude lower than those resulting from 1SG; the relative disparity between 2SG and 1SG grows even larger as the number of gradient equivalents gets larger.

[7]It was also found that, if the iterates were constrained to lie in some hypercube around θ^* (as required, e.g., in genetic algorithms), then all values in Table I will be reduced, sometimes by several orders of magnitude. Such prior information can be valuable at speeding convergence.

VI. CONCLUDING REMARKS

This paper has presented a general adaptive second-order SA approach that has a simple structure and is efficient in high-dimensional problems. The approach applies in either the gradient-free (Kiefer–Wolfowitz) setting where only noisy loss function evaluations are available or the stochastic gradient-based/root-finding (Robbins–Monro) setting where noisy gradient evaluations are available. This adaptive simultaneous perturbation algorithm is based on the principle of changing of all the parameters in the problem simultaneously in constructing gradient and Hessian estimates. In high-dimensional problems of practical interest, such simultaneous changes admit an efficient implementation by greatly reducing the number of loss function evaluations or gradient evaluations required to carry out the optimization process relative to conventional "one-at-a-time" changes. The ASP algorithm is composed of two parallel recursions: one a direct SA analog of the Newton–Raphson algorithm of deterministic optimization, and the other a sample mean calculation of per-iteration Hessian estimates formed using the simultaneous perturbation principle. The simple form for the Hessian estimate seems to obviate the claims of Schwefel [31, p. 76], Polyak and Tsypkin [26], or Yakowitz *et al.* [44] that few practical algorithms exist for estimating the Hessian in recursive optimization.

We establish conditions for the a.s. convergence of the θ and Hessian estimates from the parallel recursions. This allows us to establish the asymptotic normality of the θ estimate in both the gradient-free and stochastic gradient-based settings. In turn, the asymptotic normality provides the mechanism for analyzing the efficiency of the ASP approach. It is shown that the ASP algorithm has the same limiting efficiency that an SA algorithm would have if the true Hessian were known; this is a nearly optimal algorithm in the gradient-free case, and an optimal algorithm in the gradient-based case. Some numerical analysis illustrates the efficiency improvement possible in finite samples relative to conventional first-order approaches, with the advantage in the example here being larger in lower noise environments. (These numerical studies also illustrate some limitations of iterate averaging as a means for obtaining efficient algorithms in finite-sample practice.) Numerical studies of ASP by others on different problems have validated the efficiency of the approach for practical low- and high-noise settings.

The ASP method illustrates both the benefits and potential dangers of second-order approaches. Although ASP is a *relatively* simple adaptive approach, and the theory and numerical experience point to the improvements possible, one should be careful in implementation, and beware of potential divergence. Most of the care in implementation is devoted to choosing the important algorithm coefficients; there are generally more coefficients to choose than in the first-order algorithms (although fewer than certain other stochastic optimization methods such as the various genetic algorithms; further, the effort can be reduced by simply using the asymptotically optimal or near-optimal $a_k = 1/k$ for the important "gain" sequence if the initial condition is sufficiently close to the optimum). In addition, it is important to monitor the algorithm or implement the "blocking" procedures described in Section II-D to guard against wild steps during the iteration process. This problem seems inherent in second-order approaches, both deterministic (à la Newton–Raphson) and stochastic. Nevertheless, with the appropriate care, the adaptive approach is relatively easy to implement, and can offer impressive gains in efficiency.

APPENDIX A
PROOFS OF CONVERGENCE RESULTS IN SECTION III AND ASYMPTOTIC DISTRIBUTION RESULTS IN SECTION IV

Proof of Lemma (Sufficient Conditions for C.5 and C.7)

C.7 is used in the proofs of Theorems 1a and 1b only to ensure that $P\left(\limsup_{k\to\infty} \|\tilde{\theta}_k\| = \infty\right) = 0$. Given the boundedness of $\hat{\theta}_k$, this condition becomes superfluous. Regarding C.5, the boundedness condition together with the facts that $a_k/c_k^2 \to 0$ and $c_k^2 \overline{\overline{H}}_k^{-1} \to 0$ (C.6) imply that, for some $0 < \rho' < \rho, a_k|\overline{g}_{ki}(\hat{\theta}_k)| \leq \rho'$ a.s. for all k sufficiently large. From the basic recursion, $\tilde{\theta}_{k+1,i} = \tilde{\theta}_{ki} - a_k\overline{g}_{ki}(\hat{\theta}_k) - a_k e_{ki}$, where $e_k = G_k(\hat{\theta}_k) - \overline{g}_{ki}(\hat{\theta}_k)$. But $a_k e_k \to 0$ a.s. by the martingale convergence theorem (see (8) and (9) in Spall and Cristion [38]). Since $|\tilde{\theta}_{ki}| \geq \rho > \rho'$, we know that sign $\tilde{\theta}_{ki}$ = sign $\tilde{\theta}_{k+1,i}$ for all k sufficiently large, implying that sign $g_i(\hat{\theta}_k)$ = sign $g_i(\hat{\theta}_{k+1})$ a.s. Assumption (3.1) completes the proof of sufficiency for C.5. Q.E.D.

Proof of Theorem 1a (2SPSA)

The proof will proceed in three parts. Some of the proof closely follows that of the proposition in Spall and Cristion [38], in which case the details will be omitted here, and the reader will be directed to that reference. However, some of the proof differs in nontrivial ways due to, among other factors, the need to explicitly treat the bias in the gradient estimate $G_k(\cdot)$. First, we will show that $\tilde{\theta}_k \equiv \hat{\theta}_k - \theta^*$ does not diverge in magnitude to ∞ on any set of nonzero measure. Second, we will show that $\tilde{\theta}_k$ converges a.s. to some random vector, and third, we will show that this random vector is the constant 0, as desired. Equalities hold a.s. where relevant.

Part 1: First, from C.0, C.2, and C.3, it can be shown in the manner of Spall [33, Lemma 1] that, for all k sufficiently large,

$$E(G_k(\hat{\theta}_k)|\hat{\theta}_k) = g(\hat{\theta}_k) + b_k \qquad \text{(A1)}$$

where $c_k^{-2}\|b_k\|$ is uniformly bounded a.s. Using C.6, we know that $\overline{\overline{H}}_k^{-1}$ exists a.s., and hence we write $M_k \equiv a_k \overline{\overline{H}}_k^{-1}(g(\hat{\theta}_k) + b_k)$. Then, as in the proposition of Spall and Cristion [38], C.1, C.2, and C.6, and Holder's inequality imply, via the martingale convergence theorem,

$$\tilde{\theta}_{k+1} + \sum_{j=0}^{k} M_j \xrightarrow{\text{a.s.}} X \qquad \text{(A2)}$$

where X is some integrable random vector.

Let us now show that $P\left(\limsup_{k\to\infty} \|\tilde{\theta}_k\| = \infty\right) = 0$. Since the arguments below apply along any subsequence, we will, for ease of notation and without loss of generality, consider the event $\{\|\tilde{\theta}_k\| \to \infty\}$. We will show that this event has probability 0 in a modification to the arguments in [38, proposition] (which is a multivariate extension to scalar arguments in Blum [2], and Evans and Weber [8]). Furthermore, suppose that the limiting quantity of the unbounded elements is $+\infty$ (trivial modifications cover a limiting quantity including $-\infty$ limits). Then, as shown in [38, proposition], the event of interest $\{\|\tilde{\theta}_k\| \to \infty\}$ has probability 0 if

$$\left\{\{\tilde{\theta}_{ki} \le \rho'(\tau, S)\,\forall i \in S, \tilde{\theta}_{ki} \le \tau\,\forall i \notin S, k \ge K(\tau, S)\}\right.$$
$$\left.\cap \limsup_{k\to\infty}\{M_{ki} < 0\,\forall i \in S\}\right\} \tag{A3a}$$

and

$$\left\{\{\tilde{\theta}_{ki} \to \infty\,\forall i \in S \cap \liminf_{k\to\infty}\{M_{ki} < 0\,\forall i \in S\}^c\right\} \tag{A3b}$$

both have probabilities 0 for all τ, S, and $\rho'(\tau, S)$ as defined in C.7, where $K(\tau, S) < \infty$ and the superscript c denotes set complement.

For event (A3a), we know that there exists a subsequence $\{k_0, k_1, k_2, \cdots\}, k_0 \ge K(\tau, S)$ such that $\{\tilde{\theta}_{k_j i} \ge \rho'(\tau, S)\,\forall i \in S\} \cap \{M_{k_j i} < 0\,\forall i \in S\}$ is true. Then, from C.6 and (A1),

$$\sum_{i\in S} \tilde{\theta}_{k_j i}(\overline{g}_{k_j i}(\hat{\theta}_{k_j}) + o(1)) < 0 \quad \text{a.s.} \tag{A4}$$

for all k_j. By C.4, $\tilde{\theta}_{k_j}^T \overline{g}_{k_j}(\hat{\theta}_{k_j}) \ge \rho\|\tilde{\theta}_{k_j}\|$ a.s. which, by C.7, implies, for all j sufficiently large,

$$\sum_{i\in S} \tilde{\theta}_{k_j i}\overline{g}_{k_j i}(\tilde{\theta}_{k_j})$$
$$\ge \frac{\rho}{2}\|\tilde{\theta}_{k_j}\| \ge \left(\frac{\rho}{2}\right) \dim(S)\rho'(\tau, S) \ge \frac{\rho\tau}{2} \quad \text{a.s.} \tag{A5}$$

since $\rho'(\tau, S) \ge \tau$ and $\dim(S) \ge 1$. Taken together, (A4) and (A5) imply that, for each sample point (except possibly on a set of measure 0), the event in (A3a) has probability 0. Now, consider the second event (A3b). From (A2), we know that, for almost all sample points, $\sum_{k=0}^{\infty} M_{ki} \to -\infty\,\forall i \in S$ must be true. But this implies from C.5 and the above-mentioned uniformly bounded decaying bias (b_k) that for no $i \in S$ can $M_{ki} \ge 0$ occur i.o. However, at each k, the event $\{M_{ki} < 0\,\forall i \in S\}^c$ is composed of the union of $2^{\dim(S)} - 1$ events, each of which has $M_{ki} \ge 0$ for at least one $i \in S$. This, of course, requires that $M_{ki} \ge 0$ i.o. for at least one $i \in S$, which creates a contradiction. Hence, the probability of the event in (A3b) is 0. This completes Part 1 of the proof.

Part 2: To show that $\hat{\theta}_k$ converges a.s. to a unique (finite) limit, we show that

$$P\left(\liminf_{k\to\infty} \tilde{\theta}_{ki} < a' < b' < \limsup_{k\to\infty} \tilde{\theta}_{ki}\right) = 0 \qquad \forall i \tag{A6}$$

for any $a' < b'$. This result follows exactly as in the proof of Part 2 of the proposition in Spall and Cristion [38].

Part 3: Let us now show that the unique finite limit from Part 2 is 0. From (A2) and the conclusion of Part 1, we have $\limsup_{n\to\infty} |\sum_{k=0}^{\infty} M_{ki}| < \infty$ a.s. $\forall i$. Then the result to be shown follows if

$$P\left(\lim_{k\to\infty} \tilde{\theta}_k \ne 0, \left\|\sum_{k=0}^{\infty} M_k\right\| < \infty\right) = 0. \tag{A7}$$

Suppose that the event in the probability of (A7) is true, and let $I \subseteq \{1, 2, \cdots, p\}$ represent those indexes i such that $\tilde{\theta}_{ki} \nrightarrow 0$ as $k \to \infty$. Then, by the convergence in Part 2, there exists (for almost any sample point in the underlying sample space) some $0 < a' < b' < \infty$ and $K(a', b') < \infty$ (dependent on sample point) such that $\forall k \ge K, 0 < a' \le |\tilde{\theta}_{ki}| \le b' < \infty$ when $i \in I (I \ne \emptyset)$ and $|\tilde{\theta}_{ki}| < a'$ when $i \in I^c$. From C.4, it follows that

$$\sum_{k=K+1}^{n} a_k \sum_{i\in I} \tilde{\theta}_{ki}\overline{g}_{ki}(\hat{\theta}_k) \ge a'\rho \sum_{k=K+1}^{n} a_k. \tag{A8}$$

But since C.5 implies that $\overline{g}_{ki}(\hat{\theta}_k)$ can change sign only a finite number of times (except possibly on a set of sample points of measure 0), and since $|\tilde{\theta}_{ki}| \le b'$, we know from (A8) that, for at least one $i \in I$,

$$\limsup_{n\to\infty} \left|\frac{\rho a' \sum_{k=K+1}^{n} a_k}{\sum_{k=K+1}^{n} a_k\overline{g}_{ki}(\hat{\theta}_k)}\right| < \infty. \tag{A9}$$

Recall that $a_k\overline{g}_k(\hat{\theta}_k) = M_k - a_k\overline{\overline{H}}_k^{-1}b_k$ and $b_k = O(c_k^2)$ a.s. Hence, from C.6, we have $\overline{\overline{H}}_k^{-1}b_k = o(1)$. Then by (A9), $|\sum_{k=K+1}^{\infty} M_{ki}| = \infty$. Since, for the $a' < b'$ above, there exists such a K for each sample point in a set of measure one (recalling that $\hat{\theta}_k$ converges a.s. by Part 2), we know from the above discussion that there also exists an $i \in I$ (i possibly dependent on the sample point) such that $|\sum_{k=K+1}^{\infty} M_{ki}| = \infty$. Since I has a finite number of elements, $|\sum_{k=0}^{\infty} M_{ki}| = \infty$ with probability >0 for at least one i. However, this is inconsistent with the event in (A7), showing that the event does, in fact, have probability 0. This completes Part 3, which completes the proof. Q.E.D.

Proof of Theorem 1b (2SG)

The initial martingale convergence arguments establishing the 2SG analog to (A2) are based on C.0′–C.2′ and C.6. Although there is no bias in the gradient measurement, C.4 and C.7 still work together to guarantee that the elements potentially diverging [in the arguments analogous to those surrounding (A3a), (A3b)] asymptotically dominate the product $\tilde{\theta}_{k_j}^T \overline{g}_{k_j}(\hat{\theta}_{k_j})$. As in the Proof of Theorem 1a, this sets up a contradiction. The remainder of the proof follows exactly as in Parts 2 and 3 of the Proof of Theorem 1a, with some of the arguments made easier since $b_k \equiv 0$. Q.E.D

Proof of Theorem 2a (2SPSA)

First, note that the conditions subsume those of Theorem 1a; hence, we have a.s. convergence of $\hat{\theta}_k$. By C.8, we have $E((c_k\tilde{c}_k)^2\|\hat{H}_k\|^2)$ uniformly bounded $\forall k$. Hence, by the additional assumption introduced in C.1″ (beyond that in C.1), the martingale convergence result in, say, Laha and Rohatgi [16, p. 397], yields

$$\frac{1}{n+1}\sum_{k=0}^{n} (\hat{H}_k - E(\hat{H}_k|\hat{\theta}_k)) \to 0 \quad \text{a.s.} \tag{A10}$$

as $n \to \infty$. By (2.4), conditions C.3′, C.8, and C.9 imply $\forall \ell$ (A11) shown at the bottom of the page, where $L_{hij}^{(3)}$ represents the third derivative of L w.r.t. the hth, ith, and jth elements of θ; $\overline{\theta}_k^{\pm}$ are points on the line segments between $\hat{\theta}_k \pm c_k \Delta_k + c_k \tilde{\Delta}_k$ and $\hat{\theta}_k \pm c_k \Delta_k$; and we used the fact that $E(\tilde{\Delta}_{ki} \tilde{\Delta}_{kj} / \tilde{\Delta}_{k\ell}) = 0 \; \forall i, j, k$, and ℓ (implied by C.9 and the Cauchy–Schwarz inequality).

Let

$$B_{k\ell} = \frac{1}{6} E\left[\tilde{\Delta}_{k\ell}^{-1} \sum_{h,i,j} (L_{hij}^{(3)}(\overline{\theta}_k^{+}) - L_{hij}^{(3)}(\overline{\theta}_k^{-})) \cdot \tilde{\Delta}_{kh} \tilde{\Delta}_{ki} \tilde{\Delta}_{kj} \,\middle|\, \hat{\theta}_k, \Delta_k \right].$$

By C.3′ (bounding the difference in $L_{hij}^{(3)}$ terms) and C.9 in conjunction with the Cauchy–Schwarz inequality and C.1″ ($\tilde{c}_k = O(c_k)$), we have $B_{k\ell}/c_k$ uniformly bounded (in $\hat{\theta}_k, \Delta_k$) for all k sufficiently large. Hence, from (A11) the ℓmth element of $\hat{H}_k$ satisfies

$$\begin{aligned} E(\hat{H}_{k,\ell m} | \hat{\theta}_k) &= E\left(\frac{G_{k\ell}^{(1)}(\hat{\theta}_k + c_k \Delta_k) - G_{k\ell}^{(1)}(\hat{\theta}_k - c_k \Delta_k)}{2 c_k \Delta_{km}} \,\middle|\, \hat{\theta}_k \right) \\ &= E\left(\frac{g_\ell(\hat{\theta}_k + c_k \Delta_k) - g_\ell(\hat{\theta}_k - c_k \Delta_k) + \tilde{c}_k^2 B_{k\ell}}{2 c_k \Delta_{km}} \,\middle|\, \hat{\theta}_k \right) \\ &= E\left(\frac{2 c_k [\partial g_\ell / \partial \theta^T]_{\theta = \hat{\theta}_k} \Delta_k + O(c_k^3)}{2 c_k \Delta_{km}} \,\middle|\, \hat{\theta}_k \right) \\ &= H_{\ell m}(\hat{\theta}_k) + O(c_k^2) \end{aligned} \tag{A12}$$

where the $O(c_k^3)$ term in the third line of (A12) encompasses both $\tilde{c}_k^2 B_{k\ell}$ and the uniformly bounded contributions due to $\partial^2 g_\ell / \partial \theta^T \partial \theta^T$ in the remainder terms of the expansion of $g_\ell(\hat{\theta}_k + c_k \Delta_k) - g_\ell(\hat{\theta}_k - c_k \Delta_k)$ (so $O(c_k^3)/c_k^3$ is uniformly bounded, allowing the use of C.9 and the Cauchy–Schwarz inequality in producing the $O(c_k^2)$ term in the last line of (A12)).

Then, by (A12), the continuity of H near $\hat{\theta}_k$, and the fact that $\hat{\theta}_k \to \theta^*$ a.s. (Theorem 1a), the principle of Cesaro summability implies

$$\begin{aligned} \frac{1}{n+1} \sum_{k=0}^{n} E(\hat{H}_k | \hat{\theta}_k) &= \frac{1}{n+1} \sum_{k=0}^{n} (H(\hat{\theta}_k) + O(c_k^2)) \to H(\theta^*) \text{ a.s.} \end{aligned} \tag{A13}$$

Given that $\overline{H}_k = (n+1)^{-1} \sum_{k=0}^{n+1} \hat{H}_k$, (A.10) and (A13) then yield the result to be proved. Q.E.D.

Proof of Theorem 2b (2SG)

Since the conditions subsume those of Theorem 1b, we have $\hat{\theta}_k \to \theta^*$ a.s. Analogous to (A10), C.1‴ and C.8′ yield a martingale convergence result for the sample mean of $\hat{H}_k - E(\hat{H}_k | \hat{\theta}_k)$. Then, given the boundedness of the third derivatives of $L(\theta)$ near $\hat{\theta}_k$ for all k, the Cauchy–Schwarz inequality and C.8′, C.9′ imply that $E(\hat{H}_k | \hat{\theta}_k) = H(\hat{\theta}_k) + O(c_k^2)$. By $\hat{\theta}_k \to \theta^*$ a.s., the Cesaro summability arguments in (A13) yield the result to be proved. Q.E.D.

Proof of Theorem 3a (2SPSA)

Beginning with the expansion $E(G_k(\hat{\theta}_k) | \hat{\theta}_k) = H(\overline{\theta}_k)(\hat{\theta}_k - \theta^*) + b_k$, where $\overline{\theta}_k$ is on the line segment between $\hat{\theta}_k$ and θ^* and the bias b_k is defined in (A1), the estimation error can be represented in the notation of [9] (also [28]) as

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

where

$$\begin{aligned} \Gamma_k &= a \overline{\overline{H}}_k^{-1} H(\overline{\theta}_k) \\ \Phi_k &= -a \overline{\overline{H}}_k^{-1} \\ V_k &= k^{-\gamma} [G_k(\hat{\theta}_k) - E(G_k(\hat{\theta}_k) | \hat{\theta}_k)] \end{aligned}$$

and $T_k = -a k^{\beta/2} b_k$. The proof follows that of Spall [33, Prop. 2] closely, which shows that the three sufficient conditions for asymptotic normality, in Fabian [9, (2.2.1)–(2.2.3)], hold. By the convergence of $\hat{\theta}_k$, it is straightforward to show a.s. convergence of T_k to 0 if $3\gamma - \alpha/2 > 0$ or to T in (4.2) if $3\gamma - \alpha/2 = 0$. The mean expression μ then follows directly from Fabian

$$\begin{aligned} &E[G_{k\ell}^{(1)}(\hat{\theta}_k \pm c_k \Delta_k) | \hat{\theta}_k, \Delta_k] \\ &= E\left[\frac{\tilde{c}_k g(\hat{\theta}_k \pm c_k \Delta_k)^T \tilde{\Delta}_k + \frac{\tilde{c}_k^2}{2} \tilde{\Delta}_k^T H(\hat{\theta}_k \pm c_k \Delta_k) \tilde{\Delta}_k + \frac{\tilde{c}_k^3}{6} \sum_{h,i,j} L_{hij}^{(3)}(\overline{\theta}_k^{\pm}) \tilde{\Delta}_{kh} \tilde{\Delta}_{ki} \tilde{\Delta}_{kj}}{\tilde{c} \tilde{\Delta}_{k\ell}} \,\middle|\, \hat{\theta}_k, \Delta_k \right] \\ &= g_\ell(\hat{\theta}_k \pm c_k \Delta_k) + \frac{1}{6} \tilde{c}_k^2 E\left[\tilde{\Delta}_{k\ell}^{-1} \sum_{h,i,j} L_{hij}^{(3)}(\overline{\theta}_k^{\pm}) \tilde{\Delta}_{kh} \tilde{\Delta}_{ki} \tilde{\Delta}_{kj} \,\middle|\, \hat{\theta}_k, \Delta_k \right] \end{aligned} \tag{A11}$$

[9] and the convergence of $\overline{H}_k$ (and hence $\overline{\overline{H}}_k^{-1}$ by C.11 and the existence of $H(\theta^*)^{-1}$). Further, as in Spall [33, Prop. 2], $E(V_k V_k^T|\hat{\theta}_k)$ is a.s. convergent by C.2 and C.10, leading to the covariance matrix Ω. This shows Fabian [9, (2.2.1) and (2.2.2)]. The final condition [9, (2.2.3)] follows as in Spall [33, Prop. 2] since the definition of V_k is identical in both standard SPSA and 2SPSA. Q.E.D.

Proof of Theorem 3b (2SG)

Analogous to the Proof of Theorem 3a, the estimation error can be represented as

$$\hat{\theta}_{k+1} - \theta^* = (I - k^{-\alpha}\Gamma_k)(\hat{\theta}_k - \theta^*) + k^{-\alpha}\Phi_k e_k$$

where $\Gamma_k = a\overline{\overline{H}}_k^{-1} H(\overline{\theta}_k)$ and $\Phi_k = -a\overline{\overline{H}}_k^{-1}$. Conditions (2.2.1) and (2.2.2) of Fabian [9] follow immediately by the smoothness of $L(\theta)$ (from C.3′), the convergence of $\hat{\theta}_k$ and $\overline{H}_k$, and C.12. Condition (2.2.3) of Fabian follows by Holder's inequality and C.2′, C.3′. Q.E.D.

Appendix B
Interpretation of Regularity Conditions

This Appendix provides comments on some of the conditions of ASP relative to other adaptive SA approaches. In the confines of a short discussion, it is obviously not possible to provide a detailed discussion of all conditions of all known adaptive approaches. Nevertheless, we hope to convey a flavor of the relative nature of the conditions.

As discussed in Section III, some of the conditions of ASP depend on $\hat{\theta}_k$ itself, creating a type of circularity (i.e., direct conditions on the quantity being analyzed). This circularity has been discussed elsewhere (see Section III and Kushner and Clark [14, pp. 40–41]) since other SA algorithms also have $\hat{\theta}_k$-dependent conditions. Some of the ASP conditions can be eliminated or simplified if the conditions of the lemma in Section III hold. The foremost lemma condition is that $\hat{\theta}_k$ be uniformly bounded. Of course, this uniformly bounded condition is itself a circular condition, but it helps to simplify the other conditions of the theorems that are dependent on $\hat{\theta}_k$ since the $\hat{\theta}_k$ dependence can be replaced by an assumption that these other conditions hold uniformly over all θ in the bounded set guaranteed to contain $\hat{\theta}_k$ (e.g., the current assumption C.3 that $g(\theta)$ be twice continuously differentiable in neighborhoods of estimates $\hat{\theta}_k$ can be replaced by an assumption that $g(\theta)$ is twice continuously differentiable on some bounded set known to contain $\hat{\theta}_k$). If the lemma applies, condition C.5 (on the i.o. behavior of $\hat{\theta}_k$) is unnecessary.

In showing convergence and asymptotic normality, one might wonder whether other adaptive algorithms could avoid conditions that depend on $\hat{\theta}_k$, and avoid alternative conditions that are similarly undesirable. Based on currently available adaptive approaches, the answer appears to be "no." As an illustration, let us analyze one of the more powerful results on adaptive algorithms, the result in Wei [43]. Wei's results are multivariate generalizations of results in Nevel'son and Has'minskii [23, ch. 7] and Venter [41]. The Wei [43] approach is restricted to the SG/root-finding setting as opposed to the more general setting for ASP that encompasses both gradient-free and SG/root finding. The approach is based on $2p$ measurements of $g(\theta)$ at each iteration to estimate the Jacobian (Hessian) matrix. Some of the conditions in Wei [43] are similar to conditions for ASP (e.g., decaying gain sequences and smoothness of the functions involved), while other conditions are more stringent (the restriction to only the root-finding setting and the requirement for i.i.d. measurement noise). There are also conditions in ASP that are not required in Wei [43], principally those associated with "nice" behavior of the user-specified $\overline{\overline{H}}_k$ (bounded moments, etc.), the steepness conditions C.4 and C.7 (similar to standard conditions in some other adaptive approaches, e.g., Ruppert [27]), and limits on the amount of bouncing in "big steps" around θ^* (the i.o. condition C.5). An additional key assumption in Wei [43] is the symmetric function condition on the Jacobian (or Hessian) matrix:

$$H(\theta)H(\theta')^T + H(\theta')H(\theta)^T > 0 \qquad \forall\, \theta, \theta'. \tag{B1}$$

This, unfortunately, is a stringent condition that may be easily violated. In the optimization case (where H is a Hessian), this condition may fail even for benign (e.g., convex) loss functions. Consider, for example, a case with $\theta = (x, y)^T$ and a simple convex loss function $L(\theta) = x^4 + x^2 + y^2 + xy$. Letting $\theta = (0, 0)^T$ and $\theta' = (2, 0)^T$, we have

$$H(\theta)H(\theta')^T + H(\theta')H(\theta)^T = \begin{bmatrix} 202 & 56 \\ 56 & 10 \end{bmatrix}$$

which is not positive definite, violating condition (B1). Aside from the fact that this condition may be easily violated, it is also generally impossible to check in practice because it requires knowledge of the true $H(\theta)$ over the whole domain; this, of course, is the very quantity that is being estimated! The requirement for such prior knowledge is also apparent in other adaptive approaches discussed in Section I, e.g., Ruppert [27] and Fabian [10]. Given the above, it is clear that neither ASP nor Wei [43] (nor others) have uniformly "easier" conditions for their respective approaches.

The inherent difficulty in establishing theoretical properties of adaptive approaches comes from the need to couple the estimates for the parameters of interest and for the Hessian/Jacobian matrix. This tends to lead to nontrivial regularity conditions, as seen in the $\hat{\theta}_k$-dependent conditions of ASP and in the stringent conditions that have appeared in the literature for other approaches. There appear to be no easy conditions for establishing rigorous properties of adaptive algorithms. However, given that all of these approaches have a strong intuitive appeal based on analogies to deterministic optimization, the needs of practical users will focus less on the nuances of the regularity conditions and more on the cost of implementation (e.g., the number of function measurements needed), the ease of implementation, and the practical performance.

Acknowledgment

The author appreciates the insightful comments of Dr. J. Maryak on a draft version of the paper, and the suggestions of Dr. I.-J. Wang on the proof of the lemma in Section III.

REFERENCES

[1] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*. New York: Springer-Verlag, 1990.
[2] J. R. Blum, "Approximation methods which converge with probability one," *Ann. Mat. Statist.*, vol. 25, pp. 382–386, 1954.
[3] R. W. Brennan and P. Rogers, "Stochastic optimization applied to a manufacturing system operation problem," in *Proc. Winter Simulation Conf.*, C. Alexopoulos, K. Kang, W. R. Lilegdon, and D. Goldsman, Eds., 1995, pp. 857–864.
[4] H. F. Chen, T. E. Duncan, and B. Pasik-Duncan, "A stochastic approximation algorithm with random differences," in *Proc. 13th IFAC World Congr.*, vol. H, 1996, pp. 493–496.
[5] D. C. Chin, "A more efficient global optimization algorithm based on Styblinski and Tang," *Neural Networks*, vol. 7, pp. 573–574, 1994.
[6] ——, "Comparative study of stochastic algorithms for system optimization based on gradient approximation," *IEEE Trans. Syst., Man, Cybern. B*, vol. 27, pp. 244–249, 1997.
[7] J. Dippon and J. Renz, "Weighted means in stochastic approximation of minima," *SIAM J. Contr. Optimiz.*, vol. 35, pp. 1811–1827, 1997.
[8] S. N. Evans and N. C. Weber, "On the almost sure convergence of a general stochastic approximation procedure," *Bull. Australian Math. Soc.*, vol. 34, pp. 335–342, 1986.
[9] V. Fabian, "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] ——, "On asymptotically efficient recursive estimation," *Ann. Stat.*, vol. 6, pp. 854–866, 1978.
[12] L. Gerencsér, "Rate of convergence of moments for a simultaneous perturbation stochastic approximation method for optimization," *IEEE Trans. Automat. Contr.*, vol. 44, pp. 894–905, 1999.
[13] C. Kao, W. T. Song, and S.-P. Chen, "A modified quasi-Newton method for optimization in simulation," *Int. J. Oper. Res.*, vol. 4, pp. 223–233, 1997.
[14] H. J. Kushner and D. S. Clark, *Stochastic Approximation for Constrained and Unconstrained Systems*. New York: Springer-Verlag, 1978.
[15] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*. New York: Springer-Verlag, 1997.
[16] R. G. Laha and V. K. Rohatgi, *Probability Theory*. New York: Wiley, 1979.
[17] L. Ljung, "Applications to adaptation algorithms," in *Stochastic Approximation and Optimization of Random Systems*, L. Ljung, G. Pflug, and H. Walk, Eds. Basel: Birkhauser, 1992, pp. 95–113.
[18] R. R. Luman, "Upgrading complex systems of systems: A CAIV methodology for warfare area requirements allocation," *Mil. Oper. Res.*, vol. 5, no. 2, pp. 73–75, 2000.
[19] O. Macchi and E. Eweda, "Second order convergence analysis of stochastic adaptive linear filtering," *IEEE Trans. Automat. Contr.*, vol. AC-28, pp. 76–85, 1983.
[20] Y. Maeda and R. J. P. De Figueiredo, "Learning rules for neuro-controller via simultaneous perturbation," *IEEE Trans. Neural Networks*, vol. 8, pp. 1119–1130, 1997.
[21] J. L. Maryak, "Some guidelines for using iterate averaging in stochastic approximation," in *Proc. IEEE Conf. Decision Contr.*, 1997, pp. 2287–2290.
[22] M. Metivier and P. Priouret, "Applications of a Kushner and Clark lemma to general classes of stochastic algorithms," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 140–151, 1984.
[23] M. B. Nevel'son and R. Z. Has'minskii, *Stochastic Approximation and Recursive Estimation*. Providence, RI: Amer. Math. Soc., 1976.
[24] G. Pflug, "Applicational aspects of stochastic approximation," in *Stochastic Approximation and Optimization of Random Systems*, L. Ljung, G. Pflug, and H. Walk, Eds. Basel: Birkhauser, 1992, pp. 53–93.
[25] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM J. Contr. Optimiz.*, vol. 30, pp. 838–855, 1992.
[26] B. T. Polyak and Ya. Z. Tsypkin, "Optimal and robust methods for stochastic optimization," *Nova J. Math., Game Theory, Algebra*, vol. 6, pp. 163–176, 1997.
[27] D. Ruppert, "A Newton–Raphson version of the multivariate robbins-Monro procedure," *Ann. Stat.*, vol. 13, pp. 236–245, 1985.
[28] ——, "Stochastic approximation," in *Handbook of Sequential Analysis*, B. K. Ghosh and P. K. Sen, Eds. New York: Marcel Dekker, 1991, pp. 503–529.
[29] P. Sadegh, "Constrained optimization via stochastic approximation with a simultaneous perturbation gradient approximation," *Automatica*, vol. 33, pp. 889–892, 1997.
[30] P. Sadegh and J. C. Spall, "Optimal random perturbations for multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Trans. Automat. Contr.*, vol. 43, pp. 1480–1484, 1998. (corrections to references, vol. 44, p. 231, 1999).
[31] H.-P. Schwefel, *Evolution and Optimum Seeking*. New York: Wiley, 1995.
[32] J. C. Spall, "A stochastic approximation algorithm for large-dimensional systems in the Kiefer–Wolfowitz setting," in *Proc. IEEE Conf. Decision Contr.*, 1988, pp. 1544–1548.
[33] ——, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 322–341, 1992.
[34] ——, "A one-measurement form of simultaneous perturbation stochastic approximation," *Automatica*, vol. 33, pp. 109–112, 1997.
[35] ——, "Accelerated second-order stochastic optimization using only function measurements," in *Proc. IEEE Conf. Decision Contr.*, 1997, pp. 1417–1424.
[36] ——, "Implementation of the simultaneous perturbation algorithm for stochastic optimization," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 34, pp. 817–823, 1998.
[37] ——, "Adaptive stochastic approximation by the simultaneous perturbation method," in *Proc. IEEE Conf. Decision Contr.*, 1998, pp. 3872–3879.
[38] J. C. Spall and J. A. Cristion, "Model-free control of nonlinear stochastic systems with discrete-time measurements," *IEEE Trans. Automat. Contr.*, vol. 43, pp. 1198–1210, 1998.
[39] 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.
[40] A. Vande Wouwer, C. Renotte, M. Renotte, and M. Remy, "On the use of simultaneous perturbation stochastic approximation for neural network training," in *Proc. Amer. Contr. Conf.*, 1999, pp. 388–392.
[41] J. H. Venter, "An extension of the Robbins–Monro algorithm," *Ann. Math. Stat.*, vol. 38, pp. 181–190, 1967.
[42] H. Walk, "Foundations of stochastic approximation," in *Stochastic Approximation and Optimization of Random Systems*, L. Ljung, G. Pflug, and H. Walk, Eds. Basel: Birkhauser, 1991, vol. 37, pp. 2–51.
[43] C. Z. Wei, "Multivariate adaptive stochastic approximation," *Ann. Stat.*, vol. 15, pp. 1115–1130, 1987.
[44] S. Yakowitz, P. L'Ecuyer, and F. Vazquez-Abad, "Global stochastic optimization with low-dispersion point sets," Oper. Res., 2000, to be published.
[45] G. G. Yin and Y. Zhu, "Averaging procedures in adaptive filtering: an efficient approach," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 466–475, 1992.

James C. Spall (S'82–M'83–SM'90) joined The Johns Hopkins University, Applied Physics Laboratory in 1983 and was appointed to the principal Professional Staff in 1991. He also teaches in the Johns Hopkins School of Engineering and is Chairman of the Applied and Computational Mathematics Program. Dr. Spall has published many articles in the areas of statistics and control and holds two U.S. patents. In 1990, he received the Hart Prize as principal investigator of the most outstanding independent Research and Development project at JHU/APL. He was an Associate Editor for the IEEE TRANSACTIONS ON AUTOMATIC CONTROL and is a Contributing Editor for the CURRENT INDEX TO STATISTICS.

Dr. Spall is a fellow of the engineering honor society Tau Beta Pi.