Feedback and Weighting Mechanisms for Improving Jacobian Estimates in the Adaptive Simultaneous Perturbation Algorithm

James C. Spall, *Fellow, IEEE*

***Abstract*—It is known that a stochastic approximation (SA) analogue of the deterministic Newton-Raphson algorithm provides an asymptotically optimal or near-optimal form of stochastic search. However, directly determining the required Jacobian matrix (or Hessian matrix for optimization) 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 Jacobian matrix while concurrently estimating the primary parameters of interest. Relative to prior methods for adaptively estimating the Jacobian matrix, the paper introduces two enhancements that generally improve the quality of the estimates for underlying Jacobian (Hessian) matrices, thereby improving the quality of the estimates for the primary parameters of interest. The first enhancement rests on a feedback process that uses previous Jacobian estimates to reduce the error in the current estimate. The second enhancement is based on an optimal weighting of per-iteration Jacobian estimates. From the use of simultaneous perturbations, the algorithm requires only a small number of loss function or gradient measurements per iteration—independent of the problem dimension—to adaptively estimate the Jacobian matrix and parameters of primary interest.**

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

I. Introduction

STOCHASTIC approximation (SA) represents an important class of stochastic search algorithms for purposes of minimizing loss functions and/or finding roots of multivariate equations in the face of noisy measurements. This paper presents an approach for accelerating the convergence of SA algorithms through two enhancements—related to feedback and optimal weighting—to the adaptive simultaneous perturbation SA (SPSA) approach in Spall [17]. The adaptive SPSA algorithm is a stochastic analogue of the famous Newton-Raphson algorithm of deterministic nonlinear programming. Both enhancements are aimed at improving the quality of the estimates for underlying Jacobian (Hessian) matrices, thereby improving the quality of the estimates for the primary parameters of interest.

The first enhancement improves the quality of the Jacobian estimates through a feedback process that uses the previous Jacobian estimates to reduce the error. The second enhancement improves the quality via the formation of an optimal weighting of per-iteration Jacobian estimates. The simultaneous perturbation idea of varying all the parameters in the problem together (rather than one-at-a-time) (Spall [16]) is used to form the per-iteration Jacobian estimates. This leads to a more efficient adaptive algorithm than traditional finite-difference methods. The results apply in both the gradient-free optimization (Kiefer-Wolfowitz) and stochastic root-finding (Robbins-Monro) SA settings.

The basic problem of interest is the root-finding problem. That is, for a differentiable function $\boldsymbol{g}: \mathbb{R}^p \to \mathbb{R}^p, p \geq 1$, we are interested in finding a point $\boldsymbol{\theta}$ satisfying $\boldsymbol{g}(\boldsymbol{\theta}) = \mathbf{0}$. Of course, this problem is closely related to the optimization problem of minimizing a differentiable loss function $L = L(\boldsymbol{\theta})$ with respect to some parameter vector $\boldsymbol{\theta}$ via the equivalent problem of finding a point where $\boldsymbol{g}(\boldsymbol{\theta}) = \partial L/\partial \boldsymbol{\theta} = \mathbf{0}$. Let $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ be a point satisfying $\boldsymbol{g}(\boldsymbol{\theta}) = \mathbf{0}$. The stochastic setting here allows for the use of noisy values of $\boldsymbol{g}$ and the estimation (versus exact calculation) of the associated $p \times p$ Jacobian matrix $\boldsymbol{H} = \boldsymbol{H}(\boldsymbol{\theta}) \equiv \partial \boldsymbol{g}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}^T$. Note that the Jacobian matrix is a Hessian matrix of L when $\boldsymbol{g}$ represents the gradient of L. In this paper, let $\|\cdot\|$ denote the standard Euclidean vector norm or compatible matrix spectral norm (as appropriate).

Certainly others have looked at ways of enhancing the convergence of SA. A relatively recent review of many such methods is in Spall [18, Sect. 4.5]. In the optimization setting (using noisy measurements of L), Fabian [5] forms estimates of the gradient and Hessian by using, respectively, a finite-difference approximation and a set of differences of finite-difference approximations. This requires $O(p^2)$ loss function measurements for each update of the $\boldsymbol{\theta}$ estimate, which is extremely costly when p is large. For the root-finding setting, Ruppert [14] and Wei [19] develop stochastic Newton-like algorithms by forming Jacobian estimates via finite differences of $\boldsymbol{g}$ measurements. There are also numerous means for adaptively estimating a Jacobian (especially Hessian) matrix in special SA estimation settings where one has detailed knowledge of the underlying model (see, e.g., Macchi and Eweda [9] and Yin and Zhu [21]). While these are more efficient than the general adaptive approaches mentioned above, they are more restricted in their range of application. Bhatnagar [2] and Zhu and Spall [22] build on the adaptive SPSA approach in Spall [17], showing how some improvements are possible in special cases. Bhatnagar [3] develops several convolution-based (smoothed functional) methods for Hessian estimation in the context of simulation optimization.

Manuscript received March 09, 2007; revised November 01, 2007. First published May 27, 2009; current version published June 10, 2009. This paper was presented in part at the IEEE Conference on Decision and Control, 2006. This work was supported in part by the U.S. Navy under Contract N00024-03-D-6606. Recommended by Associate Editor, J.-F. Zhang.

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

Digital Object Identifier 10.1109/TAC.2009.2019793

Another approach aimed at achieving Newton-like convergence in a stochastic setting is iterate averaging (e.g., Polyak and Juditsky [12]; Kushner and Yin [7, Chap. 11]). While iterate averaging is conceptually appealing due to its ease of implementation, Spall [18, Sect. 4.5] shows that iterate averaging often does not produce the expected efficiency gains due to the lag in realizing an SA iteration process that bounces approximately uniformly around the solution. Hence, there is strong motivation to find theoretically justified and practically useful methods for building adaptive SA algorithms based on efficient estimates of the Jacobian matrix.

In particular, with adaptive SPSA applied to the optimization case, only *four* noisy measurements of the loss function L are needed at each iteration to estimate both the gradient and Hessian for any dimension p. In the root-finding case, *three* noisy measurements of the root-finding function $\boldsymbol{g}$ are needed at each iteration (for any p) to estimate the function and its Jacobian matrix. Although the adaptive SPSA method is a *relatively* simple 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 the choice of gain (step size) coefficients to avoid divergence. Other practical implementation suggestions are given in Spall [17, Sect. II.D]. The relatively simple form for the Jacobian estimate seems to address the criticisms of Schwefel [15, p. 76], Polyak and Tsypkin [13], and Yakowitz *et al.* [20] that few practical algorithms exist for estimating the Jacobian in recursive optimization.

The sections below describe the enhanced approach here, as well as the theory associated with convergence and efficiency. There is also a numerical study and, in the Appendix, a summary of certain key results from Spall [17].

II. Adaptive SPSA Algorithm and the Per-Iteration Jacobian (Hessian) Estimate

As with the original adaptive SPSA algorithm, the algorithm here has two parallel recursions, with one of the recursions being a stochastic version of the Newton–Raphson method for estimating $\boldsymbol{\theta}$ and the other being a weighted average of per-iteration (feedback-based) Jacobian estimates to form a best current estimate of the Jacobian matrix

$$\hat{\boldsymbol{\theta}}_{k+1} = \hat{\boldsymbol{\theta}}_k - a_k \overline{\overline{\boldsymbol{H}}}_k^{-1} \boldsymbol{G}_k(\hat{\boldsymbol{\theta}}_k), \overline{\overline{\boldsymbol{H}}}_k = \boldsymbol{f}_k(\overline{\boldsymbol{H}}_k) \tag{2.1a}$$

$$\overline{\boldsymbol{H}}_k = (1 - w_k)\overline{\boldsymbol{H}}_{k-1} + w_k(\hat{\boldsymbol{H}}_k - \hat{\boldsymbol{\Psi}}_k), k = 0, 1, 2, \ldots \tag{2.1b}$$

where a_k is a non-negative scalar gain coefficient, $\boldsymbol{G}_k(\hat{\boldsymbol{\theta}}_k)$ is some unbiased or nearly unbiased estimate of $\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)$, $\boldsymbol{f}_k : \mathbb{R}^{p\times p} \rightarrow \{\text{invertible } p \times p \text{ matrices}\}$ is a mapping designed to cope with possible noninvertibility of $\overline{\boldsymbol{H}}_k$, $0 \leq w_k \leq 1$ is a weight to apply to the new input to the recursion for $\overline{\boldsymbol{H}}_k$, $\hat{\boldsymbol{H}}_k$ is a per-iteration estimate of $\boldsymbol{H} = \boldsymbol{H}(\boldsymbol{\theta})$, and $\hat{\boldsymbol{\Psi}}_k$ is the feedback-based term that is aimed at improving the per-iteration estimate by removing the error in $\hat{\boldsymbol{H}}_k$ ($\hat{\boldsymbol{\Psi}}_k$ is defined in detail in Section III after some requisite analysis of the structure of $\hat{\boldsymbol{H}}_k$). The two recursions above are identical to those in Spall [17] with the exception of the more general weighting w_k in the second recursion ($w_k = 1/(k+1)$ in Spall [17], equivalent to a recursive calculation of the sample mean of the per-iteration $\boldsymbol{H}(\boldsymbol{\theta})$ estimates) and the inclusion of the feedback term $\hat{\boldsymbol{\Psi}}_k$. Note that at $k = 0$ in (2.1b), $\overline{\boldsymbol{H}}_{k-1} = \overline{\boldsymbol{H}}_{-1}$ may be used to reflect prior information on $\boldsymbol{H}$ if $0 < w_0 < 1$; alternatively, $\overline{\boldsymbol{H}}_{-1}$ may be unspecified—and irrelevant—when $w_0 = 1$ (we will also see that prior information may be folded in via $\hat{\boldsymbol{\Psi}}_0$). Because $\hat{\boldsymbol{H}}_k$ is defined in Spall [17], the essential aspects of the parallel recursions in ((2.1a), (2.1b)) that remain to be specified are w_k and $\hat{\boldsymbol{\Psi}}_k$.

Given that $\overline{\boldsymbol{H}}_k$ may not be invertible (especially for small k), a simple mapping $\boldsymbol{f}_k$ is to add a matrix $\delta_k \boldsymbol{I}_p$ to $\overline{\boldsymbol{H}}_k$, where $\delta_k > 0$, $\delta_k \rightarrow 0$, and $\boldsymbol{I}_p$ is a $p \times p$ identity matrix. While $\boldsymbol{H}(\boldsymbol{\theta})$ will not necessarily be symmetric in general root-finding problems, one may wish to impose the requirement that the Hessian estimates be symmetric in the case of optimization where $\boldsymbol{g}(\boldsymbol{\theta})$ is a gradient and $\boldsymbol{H}(\boldsymbol{\theta})$ is a Hessian matrix (Bhatnagar [2] discusses Hessian estimation without imposing symmetry at each iteration). In this case, $\boldsymbol{f}_k : \mathbb{R}^{p\times p} \rightarrow \{\text{symmetric positive definite } p \times p \text{ matrices}\}$. Given that $\overline{\boldsymbol{H}}_k$ is forced to be symmetric, one useful form for $\boldsymbol{f}_k$ when p is not too large is to take $\boldsymbol{f}_k$ such that $\overline{\overline{\boldsymbol{H}}}_k = (\overline{\boldsymbol{H}}_k^T \overline{\boldsymbol{H}}_k + \delta_k \boldsymbol{I}_p)^{1/2} = (\overline{\boldsymbol{H}}_k \overline{\boldsymbol{H}}_k + \delta_k \boldsymbol{I}_p)^{1/2}$, where the indicated square root is the (unique) positive definite square root (e.g., `sqrtm` in MATLAB) and $\delta_k > 0$ is some small number as above. Other forms for $\boldsymbol{f}_k$ may be useful as well (see Spall [17]).

Let us now present the basic per-iteration Jacobian estimate $\hat{\boldsymbol{H}}_k$, as given in Spall [17]. As with the basic first-order SPSA algorithm, let c_k be a positive scalar such that $c_k \rightarrow 0$ as $k \rightarrow \infty$ and let $\boldsymbol{\Delta}_k = [\Delta_{k1}, \Delta_{k2}, \ldots, \Delta_{kp}]^T$ be a user-generated mean-zero random vector with elements having finite *inverse* moments of order greater than 2; further conditions on c_k, $\boldsymbol{\Delta}_k$, and other relevant quantities are given in Spall [17] to guarantee convergence (and also discussed below in the context of the results here). These conditions are close to those of basic SPSA in Spall [16] (e.g., $\boldsymbol{\Delta}_k$ being a vector of independent Bernoulli ± 1 random variables satisfies the conditions on the perturbations, but a vector of uniformly or normally distributed random variables does not). Examples of valid gain sequences are given in Spall [17]; see also the numerical study in Section VIII below.

The formula for $\hat{\boldsymbol{H}}_k$ at each iteration is

$$\hat{\boldsymbol{H}}_k = \begin{cases} \frac{\delta \boldsymbol{G}_k}{2c_k}\left[\Delta_{k1}^{-1}, \Delta_{k2}^{-1}, \ldots, \Delta_{kp}^{-1}\right] & \text{for Jacobian } \textit{or} \\ \frac{1}{2}\left\{ \frac{\delta \boldsymbol{G}_k}{2c_k}\left[\Delta_{k1}^{-1}, \Delta_{k2}^{-1}, \ldots, \Delta_{kp}^{-1}\right] \right. & \\ \left. + \left(\frac{\delta \boldsymbol{G}_k}{2c_k}\left[\Delta_{k1}^{-1}, \Delta_{k2}^{-1}, \ldots, \Delta_{kp}^{-1}\right] \right)^T \right\} & \text{for Hessian} \end{cases} \tag{2.2}$$

where $\delta \boldsymbol{G}_k = \boldsymbol{G}_k^{(1)}(\hat{\boldsymbol{\theta}}_k + c_k \boldsymbol{\Delta}_k) - \boldsymbol{G}_k^{(1)}(\hat{\boldsymbol{\theta}}_k - c_k \boldsymbol{\Delta}_k)$, and, depending on the setting, the function $\boldsymbol{G}_k^{(1)}$ may or may not be the same as the function $\boldsymbol{G}_k$ introduced in (2.1a) (at a specified $\boldsymbol{\theta}$, $\boldsymbol{G}_k^{(1)}(\boldsymbol{\theta})$ will be either an approximation to $\boldsymbol{g}(\boldsymbol{\theta})$ or will be $\boldsymbol{g}(\boldsymbol{\theta})$ itself, as discussed below). Note that all elements of $\hat{\boldsymbol{\theta}}_k$ are varied simultaneously (and randomly) in forming $\hat{\boldsymbol{H}}_k$, as opposed to the finite-difference forms in, for example, Fabian [5] and Ruppert [14], where the elements of $\boldsymbol{\theta}$ are changed deterministically one at a time. When forming a simultaneous perturbation estimate for $\boldsymbol{g}(\boldsymbol{\theta})$ based on values of the loss function

$L(\boldsymbol{\theta})$, there are advantages to using a *one-sided* gradient approximation for $\boldsymbol{G}_k^{(1)}$ in order to reduce the total number of function evaluations (vs. the standard two-sided form that would typically be used to construct $\boldsymbol{G}_k$). This is the 2SPSA (2nd-order SPSA) setting in Spall [17]. In contrast, in the root-finding case, it is assumed that direct unbiased measurements of $\boldsymbol{g}(\boldsymbol{\theta})$ are available (e.g., Spall, 18, Chap. 5), implying that $\boldsymbol{G}_k^{(1)} = \boldsymbol{G}_k$.

The symmetrizing operation in the second part of (2.2) (the multiple 1/2 and the indicated sum) is convenient for the optimization case in order to maintain a symmetric Hessian estimate at each k. In the general root-finding case, where $\boldsymbol{H}(\boldsymbol{\theta})$ represents a Jacobian matrix, the symmetrizing operation should not typically be used (the first part of (2.2) applies).

The feedback and weighting methods below rest on an error analysis for the elements of the estimate $\hat{\boldsymbol{H}}_k$. Suppose that $\boldsymbol{g}$ is three-times continuously differentiable in a neighborhood of $\hat{\boldsymbol{\theta}}_k$. Then

$$E(\delta \boldsymbol{G}_k | \hat{\boldsymbol{\theta}}_k, \boldsymbol{\Delta}_k) = \boldsymbol{g}(\hat{\boldsymbol{\theta}}_k + c_k\boldsymbol{\Delta}_k) - \boldsymbol{g}(\hat{\boldsymbol{\theta}}_k - c_k\boldsymbol{\Delta}_k) + O\left(c_k^3\right) \quad (2.3)$$

where (2.3) follows easily (as in Spall [16, Lemma 1]) by a Taylor series argument when forming a simultaneous perturbation estimate for $\boldsymbol{g}(\boldsymbol{\theta})$ from measurements of the loss function $L(\boldsymbol{\theta})$ (the $O(c_k^3)$ term is the difference of the two $O(c_k^2)$ bias terms in the gradient estimate) and (2.3) is immediate (with $O(c_k^3) = \boldsymbol{0}$) when $\boldsymbol{G}_k^{(1)}$ and $\boldsymbol{G}_k$ represent direct unbiased measurements of $\boldsymbol{g}(\boldsymbol{\theta})$. Let δG_{ki} be the ith component of $\delta \boldsymbol{G}_k$. In the Jacobian case, (2.2) implies that the ijth element of $\hat{\boldsymbol{H}}_k$ is $\delta G_{ki}/(2c_k\Delta_{kj})$. Then, for any i, j, by an expansion of each of $\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k)$ as appear in (2.3)

$$E\left(\frac{\delta G_{ki}}{2c_k\Delta_{kj}}\middle| \hat{\boldsymbol{\theta}}_k, \boldsymbol{\Delta}_k\right) = H_{ij}(\hat{\boldsymbol{\theta}}_k) + \sum_{\ell \neq j} H_{i\ell}(\hat{\boldsymbol{\theta}}_k)\frac{\Delta_{k\ell}}{\Delta_{kj}} + O\left(c_k^2\right) \quad (2.4)$$

where H_{ij} denotes the ijth component of $\boldsymbol{H}$. In the case where exact $\boldsymbol{g}$ values are available (i.e., $\boldsymbol{G}_k^{(1)} = \boldsymbol{g}$, such as when $\boldsymbol{G}_k^{(1)}$ is an exact value of the gradient of a log-likelihood function), then $\delta G_{ki}/(2c_k\Delta_{kj})$ itself (without the conditional expectation) is equal to the right-hand side of (2.4). Because $E(\Delta_{k\ell}/\Delta_{kj}) = 0$ for all $j \neq \ell$ by the assumptions for $\boldsymbol{\Delta}_k$, it is known that the expectation of the second (summation) term on the right-hand side of (2.4) is 0 for all i, j, and k. Hence, $\hat{\boldsymbol{H}}_k$ is nearly unbiased with the bias disappearing at rate $O(c_k^2)$. Straightforward modifications to the above show the same for the Hessian estimate in (2.2) (i.e., second part in (2.2)).

Note that $\hat{\boldsymbol{H}}_k$ in (2.2) can be decomposed into four parts

$$\hat{\boldsymbol{H}}_k = \boldsymbol{H}(\hat{\boldsymbol{\theta}}_k) + \boldsymbol{\Psi}_k + O\left(c_k^2\right) + \textit{other error} \quad (2.5)$$

where $\boldsymbol{\Psi}_k$ is a $p \times p$ matrix of terms dependent on $\boldsymbol{H}(\hat{\boldsymbol{\theta}}_k)$, $\boldsymbol{\Delta}_k$, and, when only noisy L measurements are available, an additional perturbation vector $\tilde{\boldsymbol{\Delta}}_k$ associated with the creation of $\boldsymbol{G}_k^{(1)}$ and $\boldsymbol{G}_k$ (see Section III-A for the specific use of $\tilde{\boldsymbol{\Delta}}_k$). The $\boldsymbol{\Psi}_k$ term is described in detail in Section III. The summation-based term on the right-hand side of (2.4), or its equivalent in the Hessian estimation case, provides the foundation for $\boldsymbol{\Psi}_k$. The $O(c_k^2)$ term in (2.4) represents a bias in the $\boldsymbol{H}$ estimate. The *other error* term is derived from the differences $\boldsymbol{G}_k^{(1)}(\hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k) - \boldsymbol{g}(\hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k)$ (due to noise and, with 2SPSA, due to higher order effects in a Taylor expansion of L). The *other error* is identically zero when $\boldsymbol{G}_k^{(1)} = \boldsymbol{g}$. Note that $\boldsymbol{\Psi}_k$ represents the dominant error due to the simultaneous perturbations $\boldsymbol{\Delta}_k$ and, if relevant, $\tilde{\boldsymbol{\Delta}}_k$. Further, the specific form and notation for the $\boldsymbol{\Psi}_k$ term depends on whether the optimization or root-finding case is being considered ($\boldsymbol{\Psi}_k = \boldsymbol{\Psi}_k^{(L)}$ or $\boldsymbol{\Psi}_k = \boldsymbol{\Psi}_k^{(\boldsymbol{g})}$, respectively, in the notation of Sections III-A and III-B).

III. Error in Jacobian Estimate and Calculation of Feedback Term

This section characterizes the $\boldsymbol{\Psi}_k$ term in (2.5) as a vehicle towards creating the feedback term $\hat{\boldsymbol{\Psi}}_k$. Section III-A considers the case where $\boldsymbol{G}_k^{(1)}$ is formed from possibly noisy values of L; Section III-B considers the case where $\boldsymbol{G}_k^{(1)}$ is formed from possibly noisy values of $\boldsymbol{g}$. Section III-C presents $\hat{\boldsymbol{\Psi}}_k$. The probabilistic big-$O$ terms appearing below are to be interpreted in the almost surely (a.s.) sense (e.g., $O(c_k^2)$ implies a function that is a.s. bounded when divided by c_k^2, $c_k \to 0$); all associated equalities hold a.s.

A. Error for Estimates Based on Measurements of L

This subsection considers the problem of minimizing L; hence $\boldsymbol{H}$ represents a (symmetric) Hessian matrix and the estimate in the second part of (2.2) applies. When using only measurements of L as in the 2SPSA setting mentioned above (i.e., no direct measurements of $\boldsymbol{g}$), the core gradient approximation $\boldsymbol{G}_k(\hat{\boldsymbol{\theta}}_k)$ in (2.1a) uses two measurements, $y(\hat{\boldsymbol{\theta}}_k + c_k\boldsymbol{\Delta}_k)$ and $y(\hat{\boldsymbol{\theta}}_k - c_k\boldsymbol{\Delta}_k)$, representing noisy measurements of L at the two design levels $\hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k$, where c_k and $\boldsymbol{\Delta}_k$ are as defined above for $\hat{\boldsymbol{H}}_k$. These two measurements are used to generate $\boldsymbol{G}_k(\hat{\boldsymbol{\theta}}_k)$ in the conventional SPSA manner, in addition to being employed toward generating the one-sided gradient approximations $\boldsymbol{G}_k^{(1)}(\hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k)$ that form $\hat{\boldsymbol{H}}_k$. Two additional measurements $y(\hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k + \tilde{c}_k\tilde{\boldsymbol{\Delta}}_k)$ are used in generating one-sided approximations as follows:

$$\boldsymbol{G}_k^{(1)}(\hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k) = \frac{y(\hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k + \tilde{c}_k\tilde{\boldsymbol{\Delta}}_k) - y(\hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k)}{\tilde{c}_k} \times \begin{bmatrix} \tilde{\Delta}_{k1}^{-1} \\ \tilde{\Delta}_{k2}^{-1} \\ \vdots \\ \tilde{\Delta}_{kp}^{-1} \end{bmatrix} \quad (3.1)$$

with $\tilde{c}_k > 0$ satisfying conditions similar to c_k (although the numerical value of $\tilde{c}_k$ may be best chosen larger than c_k; see Spall [17]) and with the Monte-Carlo generated $\tilde{\boldsymbol{\Delta}}_k = [\tilde{\Delta}_{k1}, \tilde{\Delta}_{k2}, \ldots, \tilde{\Delta}_{kp}]^T$ being statistically independent of $\boldsymbol{\Delta}_k$ while satisfying the same regularity conditions as $\boldsymbol{\Delta}_k$ (with no loss in formal efficiency, the one-sided gradient approximation saves two loss measurements per iteration over the more conventional two-sided approximation by using the two $y(\hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k)$ values in *both* $\boldsymbol{G}_k(\hat{\boldsymbol{\theta}}_k)$ and $\boldsymbol{G}_k^{(1)}(\hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k)$). Although not required, it is usually convenient to generate $\tilde{\boldsymbol{\Delta}}_k$ and $\boldsymbol{\Delta}_k$ using the same distribution; in particular, choosing

the $\tilde{\Delta}_{ki}$ as independent Bernoulli ± 1 random variables is a valid—but not necessary—choice (note that the independence of $\tilde{\boldsymbol{\Delta}}_k$ and $\boldsymbol{\Delta}_k$ is important in establishing the "near-unbiasedness" of $\hat{\boldsymbol{H}}_k$, as shown below).

Suppose that L is four times continuously differentiable. Let $\varepsilon_k^{(\pm)}$ and $\tilde{\varepsilon}_k^{(\pm)}$ be the measurement noises: $\varepsilon_k^{(\pm)} \equiv y(\hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k) - L(\hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k)$ and $\tilde{\varepsilon}_k^{(\pm)} \equiv y(\hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k + \tilde{c}_k\tilde{\boldsymbol{\Delta}}_k) - L(\hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k + \tilde{c}_k\tilde{\boldsymbol{\Delta}}_k)$. Then, the ith component of $\boldsymbol{G}_k^{(1)}$ at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k$ is

$$\begin{aligned} G_{ki}^{(1)}(\boldsymbol{\theta}) &= \frac{L(\boldsymbol{\theta} + \tilde{c}_k\tilde{\boldsymbol{\Delta}}_k) - L(\boldsymbol{\theta}) + \tilde{\varepsilon}_k^{(\pm)} - \varepsilon_k^{(\pm)}}{\tilde{c}_k\tilde{\Delta}_{ki}} \\ &= \frac{\tilde{c}_k\boldsymbol{g}(\boldsymbol{\theta})^T\tilde{\boldsymbol{\Delta}}_k + \frac{1}{2}\tilde{c}_k^2\tilde{\boldsymbol{\Delta}}_k^T\boldsymbol{H}(\boldsymbol{\theta})\tilde{\boldsymbol{\Delta}}_k}{\tilde{c}_k\tilde{\Delta}_{ki}} \\ &\quad + \frac{\frac{1}{6}\tilde{c}_k^3 L'''\left(\overline{\boldsymbol{\theta}}_k^{(\pm)}\right)[\tilde{\boldsymbol{\Delta}}_k \otimes \tilde{\boldsymbol{\Delta}}_k \otimes \tilde{\boldsymbol{\Delta}}_k] + \tilde{\varepsilon}_k^{(\pm)} - \varepsilon_k^{(\pm)}}{\tilde{c}_k\tilde{\Delta}_{ki}} \end{aligned} \tag{3.2}$$

where $L'''(\boldsymbol{\theta}) = \partial^3 L/\partial\boldsymbol{\theta}^T\partial\boldsymbol{\theta}^T\partial\boldsymbol{\theta}^T$ denotes the $1 \times p^3$ row vector of all possible third derivatives of L, $\overline{\boldsymbol{\theta}}_k^{(\pm)}$ denotes a point on the line segment between $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + \tilde{c}_k\tilde{\boldsymbol{\Delta}}_k$ (the superscript in $\overline{\boldsymbol{\theta}}_k^{(\pm)}$ pertains to whether $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k$), and $\otimes$ denotes Kronecker product. It is sufficient to work with the first part of (2.2) (Jacobian) in characterizing the error for the second part (relevant for the Hessian estimation here), as the second part is trivially constructed from the first part. Substituting the expansion for $G_{ki}^{(1)}(\boldsymbol{\theta})$ in (3.2) into the first part of (2.2), the ijth component of $\hat{\boldsymbol{H}}_k$ is

$$\begin{aligned} &\frac{G_{ki}^{(1)}(\hat{\boldsymbol{\theta}}_k + c_k\boldsymbol{\Delta}_k) - G_{ki}^{(1)}(\hat{\boldsymbol{\theta}}_k - c_k\boldsymbol{\Delta}_k)}{2c_k\Delta_{kj}} \\ &= \frac{\left[\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k + c_k\boldsymbol{\Delta}_k) - \boldsymbol{g}(\hat{\boldsymbol{\theta}}_k - c_k\boldsymbol{\Delta}_k)\right]^T\tilde{\boldsymbol{\Delta}}_k}{2c_k\tilde{\Delta}_{ki}\Delta_{kj}} \\ &\quad + \frac{\tilde{c}_k\tilde{\boldsymbol{\Delta}}_k^T\left[\boldsymbol{H}(\hat{\boldsymbol{\theta}}_k + c_k\boldsymbol{\Delta}_k) - \boldsymbol{H}(\hat{\boldsymbol{\theta}}_k - c_k\boldsymbol{\Delta}_k)\right]^T\tilde{\boldsymbol{\Delta}}_k}{2c_k\tilde{\Delta}_{ki}\Delta_{kj}} \\ &\quad + \frac{\tilde{\varepsilon}_k^{(+)} - \varepsilon_k^{(+)} - \tilde{\varepsilon}_k^{(-)} + \varepsilon_k^{(-)}}{2\tilde{c}_k c_k\tilde{\Delta}_{ki}\Delta_{kj}} + \frac{O\left(\tilde{c}_k^2 c_k\right)}{c_k} \end{aligned} \tag{3.3}$$

where the probabilistic term $O(\tilde{c}_k^2 c_k)$ reflects the difference of third-order contributions in each of the two gradient approximations. The part inside the square brackets in the numerator of the first term on the right-hand side of (3.3) can be written as

$$\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k + c_k\boldsymbol{\Delta}_k) - \boldsymbol{g}(\hat{\boldsymbol{\theta}}_k - c_k\boldsymbol{\Delta}_k) = 2c_k\boldsymbol{H}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{\Delta}_k + O\left(c_k^3\right). \tag{3.4}$$

Hence, from (3.3) and (3.4), we have

$$\begin{aligned} &\frac{G_{ki}^{(1)}(\hat{\boldsymbol{\theta}}_k + c_k\boldsymbol{\Delta}_k) - G_{ki}^{(1)}(\hat{\boldsymbol{\theta}}_k - c_k\boldsymbol{\Delta}_k)}{2c_k\Delta_{kj}} \\ &= H_{ij}(\hat{\boldsymbol{\theta}}_k) + \frac{1}{\tilde{\Delta}_{ki}\Delta_{kj}}\sum_{\substack{\ell=1 \\ \ell m \neq ij}}^{p}\sum_{m=1}^{p} H_{\ell m}(\hat{\boldsymbol{\theta}}_k)\tilde{\Delta}_{k\ell}\Delta_{km} \\ &\quad + O\left(c_k^2\right) + O\left(\tilde{c}_k\right) + \frac{\tilde{\varepsilon}_k^{(+)} - \varepsilon_k^{(+)} - \tilde{\varepsilon}_k^{(-)} + \varepsilon_k^{(-)}}{2\tilde{c}_k c_k\tilde{\Delta}_{ki}\Delta_{kj}} + O\left(\tilde{c}_k^2\right). \end{aligned} \tag{3.5}$$

Note that the five expressions to the right of the first plus sign on the right-hand side of (3.5) represent the error in the estimate of $H_{ij}(\hat{\boldsymbol{\theta}}_k)$. The last four of these expressions either go to zero a.s. with k (the three big-O expressions) or are based on the noise terms, $\varepsilon_k^{(\pm)}$ and $\tilde{\varepsilon}_k^{(\pm)}$, which we control through the choice of the w_k (Sections IV and V). Hence, the focus in using feedback to improve the estimate for $\boldsymbol{H}$ will be on the first of the five error expressions (the double-sum-based expression).

Let us define $\boldsymbol{D}_k = \boldsymbol{\Delta}_k[\Delta_{k1}^{-1}, \Delta_{k2}^{-1}, \ldots, \Delta_{kp}^{-1}] - \boldsymbol{I}_p$, together with a corresponding matrix $\tilde{\boldsymbol{D}}_k$ based on replacing all Δ_{ki} in $\boldsymbol{D}_k$ with the corresponding $\tilde{\Delta}_{ki}$. Note that $\boldsymbol{D}_k$ is symmetric when the perturbations are independent, identically distributed (i.i.d.) Bernoulli. Then, absorbing the $O(\tilde{c}_k^2)$ term into $O(\tilde{c}_k)$, the matrix representation of (3.5) is

$$\begin{aligned} &\frac{\delta\boldsymbol{G}_k}{2c_k}\left[\Delta_{k1}^{-1}, \Delta_{k2}^{-1}, \ldots, \Delta_{kp}^{-1}\right] \\ &= \boldsymbol{H}(\hat{\boldsymbol{\theta}}_k) + \tilde{\boldsymbol{D}}_k^T\boldsymbol{H}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{D}_k + \tilde{\boldsymbol{D}}_k^T\boldsymbol{H}(\hat{\boldsymbol{\theta}}_k) + \boldsymbol{H}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{D}_k \\ &\quad + O\left(c_k^2\right) + O(\tilde{c}_k) + \frac{\tilde{\varepsilon}_k^{(+)} - \varepsilon_k^{(+)} - \tilde{\varepsilon}_k^{(-)} + \varepsilon_k^{(-)}}{2\tilde{c}_k c_k} \\ &\quad \times \begin{bmatrix} \tilde{\Delta}_{k1}^{-1} \\ \tilde{\Delta}_{k2}^{-1} \\ \vdots \\ \tilde{\Delta}_{kp}^{-1} \end{bmatrix}\left[\Delta_{k1}^{-1}, \Delta_{k2}^{-1}, \ldots, \Delta_{kp}^{-1}\right]. \end{aligned} \tag{3.6}$$

Given that the term dependent on the noises is $O(\tilde{c}_k^{-1}c_k^{-1})$, we have

$$\hat{\boldsymbol{H}}_k = \boldsymbol{H}(\hat{\boldsymbol{\theta}}_k) + \boldsymbol{\Psi}_k^{(L)}\left(\boldsymbol{H}(\hat{\boldsymbol{\theta}}_k)\right) + O\left(c_k^2\right) + O(\tilde{c}_k) + O\left(\tilde{c}_k^{-1}c_k^{-1}\right) \tag{3.7}$$

where from (2.2) (Hessian estimate in second part) and (3.6)

$$\begin{aligned} \boldsymbol{\Psi}_k^{(L)}(\boldsymbol{H}) &= \frac{1}{2}\left[\tilde{\boldsymbol{D}}_k^T\boldsymbol{H}\boldsymbol{D}_k + \tilde{\boldsymbol{D}}_k^T\boldsymbol{H} + \boldsymbol{H}\boldsymbol{D}_k\right] \\ &\quad + \frac{1}{2}\left[\tilde{\boldsymbol{D}}_k^T\boldsymbol{H}\boldsymbol{D}_k + \tilde{\boldsymbol{D}}_k^T\boldsymbol{H} + \boldsymbol{H}\boldsymbol{D}_k\right]^T. \end{aligned} \tag{3.8}$$

The superscript L in $\boldsymbol{\Psi}_k^{(L)}$ represents the dependence of this form on L measurements for creating the $\boldsymbol{H}$ estimate, to be contrasted with $\boldsymbol{\Psi}_k^{(\boldsymbol{g})}$ in the next subsection, which is dependent on $\boldsymbol{g}$ measurements.

B. Error for Estimates Based on Values of $\boldsymbol{g}$

We now consider the case where direct (but generally noisy) values of $\boldsymbol{g}$ are available. Hence, direct measurements $\boldsymbol{Y}_k(\boldsymbol{\theta}) = \boldsymbol{g}(\boldsymbol{\theta}) + \boldsymbol{e}_k(\boldsymbol{\theta})$ are used for $\boldsymbol{G}_k$ in (2.1a) and for $\boldsymbol{G}_k^{(1)}$ in $\delta\boldsymbol{G}_k$ appearing in (2.2), where $\boldsymbol{e}_k$ is a mean-zero noise term (not necessarily independent or identically distributed across k). The analysis in this case is easier than that in Section III-A as a consequence of having the direct measurements of $\boldsymbol{g}$. As in Section III-A, it is sufficient to work with the first part of (2.2) in characterizing the error for the second part (relevant for the

Hessian estimation here). Using the expansion in (3.4), the ijth component of the first part of (2.2) is

$$\begin{aligned}&\frac{G_{ki}^{(1)}(\hat{\boldsymbol{\theta}}_k + c_k\boldsymbol{\Delta}_k) - G_{ki}^{(1)}(\hat{\boldsymbol{\theta}}_k - c_k\boldsymbol{\Delta}_k)}{2c_k\Delta_{kj}}\\ &= \frac{g_i(\hat{\boldsymbol{\theta}}_k + c_k\boldsymbol{\Delta}_k) - g_i(\hat{\boldsymbol{\theta}}_k - c_k\boldsymbol{\Delta}_k)}{2c_k\Delta_{kj}} + \frac{e_{ki}^{(+)} - e_{ki}^{(-)}}{2c_k\Delta_{kj}}\\ &= H_{ij}(\hat{\boldsymbol{\theta}}_k) + \sum_{\ell\neq j} H_{i\ell}(\hat{\boldsymbol{\theta}}_k)\frac{\Delta_{k\ell}}{\Delta_{kj}} + O\left(c_k^2\right) + \frac{e_{ki}^{(+)} - e_{ki}^{(-)}}{2c_k\Delta_{kj}} \quad (3.9)\end{aligned}$$

where g_i is the ith term of $\boldsymbol{g}$ and the $e_{ki}^{(\pm)}$ represent the ith components of the noise vectors $\boldsymbol{e}_k^{(\pm)} \equiv \boldsymbol{e}_k(\hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k)$.

Note that the three expressions to the right of the first plus sign in the last equality of (3.9) represent the error in the estimate of $H_{ij}(\hat{\boldsymbol{\theta}}_k)$. The last two of these expressions either go to zero with k (the big-O expression) or are based on the noise terms $\boldsymbol{e}_{ki}^{(\pm)}$ that we control through the choice of the w_k (Section V). Hence, the focus in using feedback to improve the estimate for $\boldsymbol{H}$ will be on the first of the three error expressions (the summation-based expression).

The matrix representation of (3.9) is

$$\begin{aligned}\frac{\delta\boldsymbol{G}_k}{2c_k}\left[\Delta_{k1}^{-1}, \Delta_{k2}^{-1}, \ldots, \Delta_{kp}^{-1}\right] &= \boldsymbol{H}(\hat{\boldsymbol{\theta}}_k) + \boldsymbol{H}(\hat{\boldsymbol{\theta}}_k)\boldsymbol{D}_k\\ &+ O\left(c_k^2\right) + \frac{\boldsymbol{e}_k^{(+)} - \boldsymbol{e}_k^{(-)}}{2c_k}\left[\Delta_{k1}^{-1}, \Delta_{k2}^{-1}, \ldots, \Delta_{kp}^{-1}\right]. \quad (3.10)\end{aligned}$$

Analogous to (3.7), we have

$$\hat{\boldsymbol{H}}_k = \boldsymbol{H}(\hat{\boldsymbol{\theta}}_k) + \boldsymbol{\Psi}_k^{(\boldsymbol{g})}\left(\boldsymbol{H}(\hat{\boldsymbol{\theta}}_k)\right) + O\left(c_k^2\right) + O\left(c_k^{-1}\right) \quad (3.11)$$

where from (2.2) and (3.10), the error in the Jacobian estimate (non-symmetric) or Hessian estimate (symmetric) is

$$\boldsymbol{\Psi}_k^{(\boldsymbol{g})}(\boldsymbol{H}) \equiv \begin{cases} \boldsymbol{H}\boldsymbol{D}_k & \text{for Jacobian } \textit{or}\\ \frac{1}{2}\boldsymbol{H}\boldsymbol{D}_k + \frac{1}{2}\boldsymbol{D}_k^T\boldsymbol{H} & \text{for Hessian.}\end{cases} \quad (3.12)$$

C. Feedback Term for Estimating $\boldsymbol{H}$ Matrix

Using the analysis in Sections III-A and III-B, we now present the form for $\hat{\boldsymbol{\Psi}}_k$ through the use of feedback. If $\boldsymbol{H}$ were known, setting $\hat{\boldsymbol{\Psi}}_k$ equal to $\boldsymbol{\Psi}_k^{(\cdot)}(\boldsymbol{H}(\hat{\boldsymbol{\theta}}_k))$ would leave only the unavoidable errors due to the noise and the bias at each iteration, where $\boldsymbol{\Psi}_k^{(\cdot)}$ represents either $\boldsymbol{\Psi}_k^{(L)}$ or $\boldsymbol{\Psi}_k^{(g)}$, as appropriate (expressions (3.8) and (3.12), respectively). Unfortunately, of course, this relatively simple modification cannot be implemented because we do not know $\boldsymbol{H}$!

A variation on the idealized $\boldsymbol{H}$ estimate of the previous paragraph is to use *estimates* of $\boldsymbol{H}$ in place of the true $\boldsymbol{H}$. That is, the most recent *estimate* of $\boldsymbol{H}(\hat{\boldsymbol{\theta}}_k)$, as given by $\overline{\boldsymbol{H}}_{k-1}$ or $\overline{\overline{\boldsymbol{H}}}_{k-1}$, replaces $\boldsymbol{H}(\hat{\boldsymbol{\theta}}_k)$ in forming $\hat{\boldsymbol{\Psi}}_k$. Therefore, when using $\overline{\overline{\boldsymbol{H}}}_{k-1}$ as the estimate, the quantity $\hat{\boldsymbol{\Psi}}_k$ appearing in (2.1b) is given by

$$\hat{\boldsymbol{\Psi}}_k \equiv \begin{cases} \boldsymbol{\Psi}_k^{(L)}(\overline{\overline{\boldsymbol{H}}}_{k-1}) & \text{when } L \text{ measurements used}\\ \boldsymbol{\Psi}_k^{(\boldsymbol{g})}(\overline{\overline{\boldsymbol{H}}}_{k-1}) & \text{when } \boldsymbol{g} \text{ measurements used.}\end{cases} \quad (3.13)$$

Note that prior information on $\boldsymbol{H}$ (if available) may be conveniently incorporated into the search process via $\hat{\boldsymbol{\Psi}}_0$. In particular, to the extent that: (i) $\overline{\boldsymbol{H}}_k$ is negligibly different from $\overline{\overline{\boldsymbol{H}}}_k$, (ii) L is a quadratic function (and/or $\boldsymbol{g}$ is an affine function), and (iii) the noise level is negligible, then choosing $\overline{\overline{\boldsymbol{H}}}_{-1}$ close to $\boldsymbol{H}^* \equiv \boldsymbol{H}(\boldsymbol{\theta}^*)$ guarantees that $\overline{\boldsymbol{H}}_k$ will be close to $\boldsymbol{H}^*$ for all $k \geq 0$. Obviously, in practice, these idealized assumptions do not usually hold, but determining $\hat{\boldsymbol{\Psi}}_0$ based on a "good" value of $\boldsymbol{H}$ will, through the feedback process, typically enhance the quality of the estimate $\overline{\boldsymbol{H}}_k$ for subsequent k.

IV. Optimal Weighting With Noisy Measurements

A. General Form

As discussed above, the second way in which the accuracy of the $\boldsymbol{H}$ estimate may be improved is through the optimal selection of weights w_k in (2.1b). We consider separately below the cases where $\boldsymbol{G}_k^{(1)}$ is formed from noisy values of L and noisy values of $\boldsymbol{g}$. We restrict ourselves to a weighted, linear combination of $\hat{\boldsymbol{H}}_k - \hat{\boldsymbol{\Psi}}_k$ values, as represented in recursive form in (2.1b). Hence, the estimator in equivalent batch form for n total iterations is

$$\overline{\boldsymbol{H}}_n = \sum_{k=0}^{n} \omega_k^{(n)}(\hat{\boldsymbol{H}}_k - \hat{\boldsymbol{\Psi}}_k) \quad (4.1)$$

subject to $\omega_k^{(n)} \geq 0$ for all k and $\sum_{k=0}^{n}\omega_k^{(n)} = 1$ (note that the form in (4.1) is for analysis purposes only; the recursive form in (2.1b) is used in practice). It is straightforward to determine the $\omega_k^{(n)}$ once the weights w_k appearing in the recursion (2.1b) are specified (note that $w_0 = 1$); the converse is also true. In the 2SPSA setting (i.e., L measurements), the optimal weights w_k derived here assume that the noise contributions are nontrivial in the sense that $\text{var}[\varepsilon_k^{(+)} + \tilde{\varepsilon}_k^{(+)} - \varepsilon_k^{(-)} - \tilde{\varepsilon}_k^{(-)}]$ is asymptotically constant and strictly positive for all k. In the root-finding problem (i.e., direct $\boldsymbol{g}$ measurements), it is assumed that $\text{cov}[\boldsymbol{e}_k^{(+)} - \boldsymbol{e}_k^{(-)}]$ is asymptotically constant and positive definite for all k. (Section VII provides detailed treatment for the noise-free cases: $\varepsilon_k^{(\pm)} = \tilde{\varepsilon}_k^{(\pm)} = 0$ and $\boldsymbol{e}_k^{(\pm)} = 0$.) Further, it is assumed here that the perturbation vector sequences $\{\boldsymbol{\Delta}_k\}$ and $\{\tilde{\boldsymbol{\Delta}}_k\}$ are each identically distributed and mutually independent across k, with components $\Delta_{k\ell}$ and $\tilde{\Delta}_{k\ell}$ that are symmetrically distributed about 0.

B. Weights Using Measurements of L

The asymptotically optimal weighting is driven by the asymptotic variances of the elements in $\hat{\boldsymbol{H}}_k$. The variances of these elements are known to exist by Hölder's inequality when $\varepsilon_k^{(\pm)}$, $\tilde{\varepsilon}_k^{(\pm)}$, Δ_{ki}^{-1}, $\tilde{\Delta}_{ki}^{-1}$, and L at the relevant perturbations of $\boldsymbol{\theta}$ all have finite moments of order greater than 2 for all i and k. The dominant contributor to the asymptotic variance of each element is the $O(\tilde{c}_k^{-1}c_k^{-1})$ term on the right-hand side of (3.7), leading to a variance that is asymptotically proportional to $\tilde{c}_k^{-2}c_k^{-2}$ with constant of proportionality independent of k because $\text{var}[\varepsilon_k^{(+)} + \tilde{\varepsilon}_k^{(+)} - \varepsilon_k^{(-)} - \tilde{\varepsilon}_k^{(-)}]$ is asymptotically constant in k and because of the above-mentioned assumptions on the distributions

of $\{\boldsymbol{\Delta}_k\}$ and $\{\tilde{\boldsymbol{\Delta}}_k\}$. Further, the elements at an arbitrary position (say, the ijth) in the $O(\tilde{c}_k^{-1}c_k^{-1})$ matrices, as derived from the last term in (3.6), are uncorrelated across k by the independence assumptions on $\{\boldsymbol{\Delta}_k\}$ and $\{\tilde{\boldsymbol{\Delta}}_k\}$. Hence, from (4.1), the aim is to find the $\omega_k^{(n)}$ that minimize $\sum_{k=0}^n(\omega_k^{(n)})^2\tilde{c}_k^{-2}c_k^{-2}$ subject to the constraints on $\omega_k^{(n)}$ above. It is fairly straightforward to find the solution to this minimization problem (e.g., via the method of Lagrange multipliers), leading to optimal values $\omega_k^{(n)} = \tilde{c}_k^2c_k^2/\sum_{i=0}^n \tilde{c}_i^2c_i^2$ for all $0 \le k \le n$. This solution leads to the following weights for use in (2.1b):

$$w_k = \frac{\tilde{c}_k^2c_k^2}{\sum_{i=0}^k \tilde{c}_i^2c_i^2}. \tag{4.2}$$

C. Weights Using Measurements of $\boldsymbol{g}$

As in Section IV-B, the asymptotically optimal weighting is driven by the asymptotic variances of the elements in $\hat{\boldsymbol{H}}_k$. The dominant contributor to the asymptotic variance of each element is the $O(c_k^{-1})$ term on the right-hand side of (3.11), leading to a variance that is asymptotically proportional to c_k^{-2} with constant of proportionality independent of k because $\mathrm{cov}[\boldsymbol{e}_k^{(+)} - \boldsymbol{e}_k^{(-)}]$ is asymptotically constant and positive definite and because of the above-mentioned assumptions on the distributions of $\{\boldsymbol{\Delta}_k\}$. Further, the elements corresponding to an arbitrary position in the $O(c_k^{-1})$ matrices (based on the last term in (3.10)) are uncorrelated across k by the independence assumption on $\{\boldsymbol{\Delta}_k\}$. Hence, from (4.1), the aim is to find the $\omega_k^{(n)}$ that minimize $\sum_{k=0}^n(\omega_k^{(n)})^2c_k^{-2}$ subject to the constraints on $\omega_k^{(n)}$. The solution to this minimization problem is $\omega_k^{(n)} = c_k^2/\sum_{i=0}^n c_i^2$ for all $0 \le k \le n$, leading to the following weights for use in (2.1b):

$$w_k = \frac{c_k^2}{\sum_{i=0}^k c_i^2}. \tag{4.3}$$

V. Convergence Theory With Noisy Measurements

Some of the convergence and efficiency analysis in Spall [17] holds verbatim in analyzing the enhanced form here. In particular, under conditions for Theorems 1a and 1b in Spall [17] (see also the Appendix here), it is known that $\hat{\boldsymbol{\theta}}_k \to \boldsymbol{\theta}^*$ a.s. in the setting of either L measurements or $\boldsymbol{g}$ measurements. On the other hand, because the recursion (2.1b) differs from Spall [17] due to the weighting and feedback, it is necessary to make some changes to the arguments showing convergence of $\overline{\boldsymbol{H}}_k$ to $\boldsymbol{H}^*$. Let us define two sets for conditioning, $\Im_k^L$ and $\Im_k^{\boldsymbol{g}}$, as relate to the L measurement and $\boldsymbol{g}$ measurement cases, respectively. Namely, $\Im_k^L \equiv \{\hat{\boldsymbol{\theta}}_0, \varepsilon_j^{(\pm)}, \tilde{\varepsilon}_j^{(\pm)}, \boldsymbol{\Delta}_j, \tilde{\boldsymbol{\Delta}}_j, j = 0,1,\ldots,k-1\}$ and $\Im_k^{\boldsymbol{g}} \equiv \{\hat{\boldsymbol{\theta}}_0, \boldsymbol{e}_j^{(\pm)}, \boldsymbol{\Delta}_j, j = 0,1,\ldots,k-1\}$ are the sets generating $\hat{\boldsymbol{\theta}}_k$ and $\overline{\boldsymbol{H}}_{k-1}$ ($\Im_0^L = \Im_0^{\boldsymbol{g}} = \{\hat{\boldsymbol{\theta}}_0\}$). An intuitive interpretation of the convergence conditions is given in Spall [17], discussing how the conditions lead to relatively modest requirements in many practical applications. The interpretation in Spall [17] also applies here with obvious modifications for the conditions that are slightly changed in this paper.

This section gives analogues to Theorems 2a and 2b in Spall [17], showing convergence of $\overline{\boldsymbol{H}}_k$ with L measurements and with $\boldsymbol{g}$ measurements, respectively, based on the weights w_k in Section IV. Note that while these weights are asymptotically optimal with the variance of the noise contribution being asymptotically constant, the theorems below do not need to make this assumption. The theorems rely on Kronecker's Lemma (e.g., Chow and Teicher [4, pp. 114–115]): If $\{u_k\}$ and $\{v_k\}$ are sequences of real numbers such that $v_k > 0$, $\sum_{k=0}^n u_k/v_k$ converges (in n) to some finite value, and v_k diverges to ∞ monotonically, then $v_n^{-1}\sum_{k=0}^n u_k \to 0$ as $n \to \infty$. For the case with only L measurements, the conditions here are identical to the original Theorem 2a for 2SPSA with the exception of a slight modification of the original conditions C.1″ and C.8 to (respectively) C.1‴ to C.8″ below:

C.1‴:The conditions of C.1 hold plus $c_k = c/(k+1)^\gamma$, and $\tilde{c}_k = \tilde{c}/(k+1)^\gamma$, with $c > 0$, $\tilde{c} > 0$, and $0 < \gamma \le 1/4$.
C.8″: For some $\rho > 0$ and all k, ℓ, m, the following hold a.s.: $E[y(\hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k + c_k\tilde{\boldsymbol{\Delta}}_k)^2/(\Delta_{k\ell}\tilde{\Delta}_{km})^2|\Im_k^L] \le \rho$, $E[y(\hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k)^2/(\Delta_{k\ell}\tilde{\Delta}_{km})^2|\Im_k^L] \le \rho$, $E[(\varepsilon_k^{(\pm)} - \varepsilon_k^{(\pm)})^2/(\Delta_{k\ell}\tilde{\Delta}_{km})^2|\Im_k^L] \le \rho$, and $E(\|\overline{\overline{\boldsymbol{H}}}_k\|^2|\Im_k^L) \le \rho$. (Note that the first two bounds are similar to the bounds in C.2 in the Appendix, but are neither necessary nor sufficient for C.2.)

Theorem 1 (2SPSA Setting): Suppose only noisy measurements of L are used to form $\boldsymbol{G}_k$ and $\boldsymbol{G}_k^{(1)}$ (see (3.1)) and that $\hat{\boldsymbol{\Psi}}_k$ in (3.13) and w_k in (4.2) are used in the recursion (2.1b). Let conditions C.1‴ and C.8″ above hold together with conditions C.0, C.2, C.3′, C.4–C.7, and C.9 of Spall [17] (see the Appendix here). Then, $\overline{\boldsymbol{H}}_k \to \boldsymbol{H}^*$ a.s. as $k \to \infty$.

Proof: First, note that the conditions subsume those of Theorem 1a in Spall [17] (C.0–C.7); hence we have a.s. convergence of $\hat{\boldsymbol{\theta}}_k$ to $\boldsymbol{\theta}^*$. We first use Kronecker's Lemma (see above) to establish the convergence for a particular sum of martingale differences and then use this result to establish the convergence of $\overline{\boldsymbol{H}}_k$.

Let us first show that

$$\sum_{k=0}^n \frac{\tilde{c}_k^2c_k^2\left\{\hat{\boldsymbol{H}}_k - \hat{\boldsymbol{\Psi}}_k - E\left(\hat{\boldsymbol{H}}_k|\Im_k^L\right)\right\}}{\sum_{i=0}^n \tilde{c}_i^2c_i^2} \to \boldsymbol{0} \quad \text{a.s.} \tag{5.1}$$

where the existence of $E(\hat{\boldsymbol{H}}_k|\Im_k^L)$ is guaranteed by C.8″. Let $\hat{h}_k$ and $\hat{\psi}_k$ represent corresponding (arbitrary) elements of $\hat{\boldsymbol{H}}_k$ and $\hat{\boldsymbol{\Psi}}_k$, respectively. Because $E(\hat{\boldsymbol{\Psi}}_k|\Im_k^L) = \boldsymbol{0}$, we have that $\sum_{k=0}^n \tilde{c}_k^2c_k^2[\hat{h}_k - \hat{\psi}_k - E(\hat{h}_k|\Im_k^L)]$ is a martingale with bounded second moments for all n (it is not required that the moments be *uniformly* bounded); hence, the expression on the left-hand side of (5.1) is also a martingale. The term within the summands satisfies

$$\begin{aligned} E\left\{\left[\hat{h}_k - \hat{\psi}_k - E\left(\hat{h}_k|\Im_k^L\right)\right]^2\right\} &\le E\left\{E\left[(\hat{h}_k - \hat{\psi}_k)^2|\Im_k^L\right]\right\} \\ &\le 2E\left[E\left(\hat{h}_k^2|\Im_k^L\right) + E\left(\hat{\psi}_k^2|\Im_k^L\right)\right] \\ &= 2E\left[E\left(\hat{h}_k^2|\Im_k^L\right)\right] + O(1) \\ &= O\left(\tilde{c}_k^{-2}c_k^{-2}\right) \end{aligned} \tag{5.2}$$

where the two equalities follow by C.8″, (3.13), and the defining properties for the $\boldsymbol{\Delta}_k$ (same as for $\tilde{\boldsymbol{\Delta}}_k$) (see Section II).

We are now in a position to use Kronecker's Lemma (with $u_k = \tilde{c}_k^2 c_k^2[\hat{h}_k - \hat{\psi}_k - E(\hat{h}_k|\Im_k^L)]$ and $v_k = \sum_{i=0}^k \tilde{c}_i^2 c_i^2$) in conjunction with the martingale convergence theorem (e.g., Laha and Rohatgi [8, Theorem 6.2.1]) to show (5.1). Note that

$$E\left\{\left(\sum_{k=0}^{n}\frac{\tilde{c}_k^2 c_k^2\left[(\hat{h}_k-\hat{\psi}_k-E\left(\hat{h}_k|\Im_k^L\right)\right]}{\sum_{i=0}^k \tilde{c}_i^2 c_i^2}\right)^2\right\} = \sum_{k=0}^{n}\frac{\tilde{c}_k^4 c_k^4 E\left\{\left[\hat{h}_k-\hat{\psi}_k-E\left(\hat{h}_k|\Im_k^L\right)\right]^2\right\}}{\left(\sum_{i=0}^k \tilde{c}_i^2 c_i^2\right)^2} \quad (5.3)$$

where the equality follows by the fact that the summation argument (0 to n) inside the $(\cdot)^2$ on the left-hand side of (5.3) is a martingale (this argument differs from the corresponding scalar element in the left-hand side of (5.1) only in the summation limits for the denominator). The uniform boundedness (in n) of the right-hand side of (5.3) implies via Kronecker's Lemma that (5.1) holds. From C.1‴ and (5.2), each of the summands on the right-hand side of (5.3) are given by

$$\frac{\tilde{c}_k^4 c_k^4 O\left(\tilde{c}_k^{-2}c_k^{-2}\right)}{\left(\sum_{i=0}^k \tilde{c}_i^2 c_i^2\right)^2} = O\left(\frac{k^{-4\gamma}}{\left(\int_1^k x^{-4\gamma}dx\right)^2}\right) = \begin{cases} O(1/k^{2-4\gamma}) & \text{if } 0<\gamma<1/4 \\ O\left(1/\left[(\log k)^2 k\right]\right. & \text{if } \gamma = 1/4 \end{cases} \quad (5.4)$$

(large k), implying that (5.3) is bounded as $n \to \infty$ for all $0 < \gamma \leq 1/4$. By the martingale convergence theorem (e.g., Laha and Rohatgi [8, Theorem 6.2.1]), we therefore know that the above-mentioned martingale in the argument of the left-hand side of (5.3) converges a.s. to a random variable with finite second moment. Because this convergence holds for all elements of $\hat{\boldsymbol{H}}_k$ and $\hat{\boldsymbol{\Psi}}_k$, Kronecker's Lemma implies that (5.1) is true.

Let us analyze $E(\hat{\boldsymbol{H}}_k|\Im_k^L)$ as appears in (5.1) to show convergence of $\overline{\boldsymbol{H}}_k$. It is sufficient to work with the Jacobian form in the first part of (2.2). From (3.2), the bias in the ijth component of $\hat{\boldsymbol{H}}_k$ is

$$E\left(\frac{\frac{1}{6}\tilde{c}_k^3\left[L'''\left(\overline{\boldsymbol{\theta}}_k^{(+)}\right)-L'''\left(\overline{\boldsymbol{\theta}}_k^{(-)}\right)\right][\tilde{\boldsymbol{\Delta}}_k\otimes\tilde{\boldsymbol{\Delta}}_k\otimes\tilde{\boldsymbol{\Delta}}_k]}{\tilde{c}_k c_k \tilde{\Delta}_{ki}\Delta_{kj}}\middle|\Im_k^L\right).$$

Using C.1‴ and C.3′, $\|L'''(\overline{\boldsymbol{\theta}}_k^{(+)}) - L'''(\overline{\boldsymbol{\theta}}_k^{(-)})\| = O(c_k)$ a.s., with the implied constant in the big-O bound proportional to the magnitude of the uniformly bounded fourth derivative of L. Hence, by C.9, the above expectation exists and is $O(c_k^2)$ a.s., indicating that

$$E\left(\hat{\boldsymbol{H}}_k|\Im_k^L\right) = \boldsymbol{H}(\hat{\boldsymbol{\theta}}_k) + O\left(c_k^2\right) \quad \text{a.s.} \quad (5.5)$$

From (5.5), the continuity of $\boldsymbol{H}$ at all $\hat{\boldsymbol{\theta}}_k$, and the a.s. convergence of $\hat{\boldsymbol{\theta}}_k$ to $\boldsymbol{\theta}^*$

$$\sum_{k=0}^{n}\frac{\tilde{c}_k^2 c_k^2 E\left(\hat{\boldsymbol{H}}_k|\Im_k^L\right)}{\sum_{i=0}^n \tilde{c}_i^2 c_i^2} = \sum_{k=0}^{n}\frac{\tilde{c}_k^2 c_k^2\left[\boldsymbol{H}(\hat{\boldsymbol{\theta}}_k)+O\left(c_k^2\right)\right]}{\sum_{i=0}^n \tilde{c}_i^2 c_i^2} = \sum_{k=0}^{n}\frac{\tilde{c}_k^2 c_k^2\left[\boldsymbol{H}^*+o(1)\right]}{\sum_{i=0}^n \tilde{c}_i^2 c_i^2} \to \boldsymbol{H}^* \text{ a.s.} \quad (5.6)$$

as $n \to \infty$, where the result follows by the fact that the denominator $\sum_{i=0}^n \tilde{c}_i^2 c_i^2 \to \infty$ (from C.1⁗); the convergence of $\sum_{k=0}^n \tilde{c}_k^2 c_k^2 o(1)/\sum_{i=0}^n \tilde{c}_i^2 c_i^2$ to $\mathbf{0}$ in (5.6) follows from the Toeplitz Lemma (Laha and Rohatgi [8, p. 89]), with a trivial change in the Laha and Rohatgi proof from having n in the denominator to having $\sum_{i=0}^n \tilde{c}_i^2 c_i^2$. Given that $\overline{\boldsymbol{H}}_n = \sum_{k=0}^n \tilde{c}_k^2 c_k^2(\hat{\boldsymbol{H}}_k - \hat{\boldsymbol{\Psi}}_k)/\sum_{i=0}^n \tilde{c}_i^2 c_i^2$, (5.1) and (5.6) together yield the result to be proved. *Q.E.D.*

We now show convergence of $\overline{\boldsymbol{H}}_k$ in the root-finding case with direct (noisy) $\boldsymbol{g}$ measurements; this result is an analogue of Theorem 2b in Spall [17]. Following the pattern above, C.1‴ and C.8′ in Spall [17] are modified to a C.1⁗ and C.8‴:

C.1⁗: The conditions of C.1′ hold plus $c_k = c/(k+1)^\gamma$, with $c > 0$ and $0 < \gamma \leq 1/2$.

C.8‴: For some $\rho > 0$ and all k, ℓ, the following hold a.s.:

$$E\left[\left\|\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k \pm c_k\boldsymbol{\Delta}_k)/\Delta_{k\ell}\right\|^2|\Im_k^{\boldsymbol{g}}\right] \leq \rho$$

$$E\left[\left\|\left(\boldsymbol{e}_k^{(+)}-\boldsymbol{e}_k^{(-)}\right)/\Delta_{k\ell}\right\|^2|\Im_k^{\boldsymbol{g}}\right] \leq \rho$$

$$E\left[\left(\boldsymbol{e}_k^{(+)}-\boldsymbol{e}_k^{(-)}\right)/\Delta_{k\ell}|\Im_k^{\boldsymbol{g}}\right] = \mathbf{0}, \text{ and}$$

$$E\left(\left\|\overline{\overline{\boldsymbol{H}}}_k\right\|^2|\Im_k^{\boldsymbol{g}}\right) \leq \rho.$$

Theorem 2 (Root-Finding Setting): Suppose noisy measurements of $\boldsymbol{g}$ are used to form $\boldsymbol{G}_k$ and that $\hat{\boldsymbol{\Psi}}_k$ in (3.13) and w_k in (4.3) are used in the recursion (2.1b). Let conditions C.1⁗ and C.8‴ above hold together with C.0′, C.2′, C.3′, C.4–C.7, and C.9′ of Spall [17] (see the Appendix here). Then, $\overline{\boldsymbol{H}}_k \to \boldsymbol{H}^*$ a.s. as $k \to \infty$.

Proof: First, note that the conditions subsume those of Theorem 1b in Spall [17] (C.0′–C.2′ and C.3–C.7); hence, there is a.s. convergence of $\hat{\boldsymbol{\theta}}_k$ to $\boldsymbol{\theta}^*$. The proof then follows (5.1) to (5.4) in Theorem 1 (i.e., Kronecker's Lemma showing that the martingale convergence theorem applies) with c_k^2 replacing products $\tilde{c}_k^2 c_k^2$ (and, of course, conditioning now being based on $\Im_k^{\boldsymbol{g}}$ instead of $\Im_k^L$). Then, by the boundedness of the third derivative of $\boldsymbol{g}$ (see C.3′), $E(\hat{\boldsymbol{H}}_k|\Im_k^{\boldsymbol{g}}) = \boldsymbol{H}(\hat{\boldsymbol{\theta}}_k) + O(c_k^2)$ a.s., as in (5.5), leading to the convergence $\sum_{k=0}^n c_k^2 E(\hat{\boldsymbol{H}}_k|\Im_k^{\boldsymbol{g}})/\sum_{k=0}^n c_k^2 \to \boldsymbol{H}^*$ a.s. (analogous to (5.6)). The convergence $\overline{\boldsymbol{H}}_k \to \boldsymbol{H}^*$ a.s. as $k \to \infty$ then follows in a manner analogous to below (5.6). *Q.E.D.*

Spall [17] includes an asymptotic distribution theory for $\hat{\boldsymbol{\theta}}_k$ when a_k has the standard form: $a_k = a/(k+1)^\alpha$, $a > 0$ and $0 < \alpha \leq 1$. It is found that $k^{(\alpha-2\gamma)/2}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*)$ and $k^{\alpha/2}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^*)$ are asymptotically normal for the 2SPSA and root-finding settings, respectively, with (different) finite magnitude mean vectors and

covariance matrices. The conditions under which the asymptotic normality results hold are slightly beyond the conditions for convergence. While the *rates* of convergence (governed by the exponents $(\alpha-2\gamma)/2$ and $\alpha/2$) are identical to standard SA rates of convergence for first-order algorithms (e.g., Spall [18, Sects. 4.4 and 7.4]), the limiting mean vectors and covariance matrices are near-optimal (2SPSA) or optimal (root-finding) in a precise sense. Further, setting $a_k = 1/(k+1)$ is asymptotically near-optimal (2SPSA) or optimal (root-finding), which may be useful in practice to help reduce the tuning required for implementation (the need to choose a_k).

The improved Jacobian estimation above does not alter these asymptotic accuracy results, as the Spall [17] results are fundamentally based on the Jacobian matrix estimate achieving its limiting true value (to within a negligible error) during the search process. In practice, however, as a consequence of the Jacobian estimate reaching a nearly true value earlier in the recursive process, it would be expected that the (finite-sample) convergence accuracy in $\hat{\boldsymbol{\theta}}_k$ would improve when using the feedback and weighting above. This will be illustrated in the numerical results of Section VIII.

VI. Relative Accuracy of Jacobian Estimates with Noisy Measurements

It is fairly simple to compare the accuracy of the Jacobian estimates based on the optimal weightings above with the corresponding estimates based on simple averaging (as in Spall [17]) in the special case where the noise terms ε_k and e_{ki} (as appropriate) have constant (non-zero) variance (independent of k and $\boldsymbol{\theta}$) and the two perturbation vector sequences $\{\boldsymbol{\Delta}_k\}$ and $\{\tilde{\boldsymbol{\Delta}}_k\}$ are each identically distributed across k. Note that feedback (Section III-C) does not affect the results here, as the asymptotic variance of the Jacobian estimate is dominated by the noise contribution.

In the 2SPSA setting of only noisy loss measurements, the above assumption on the noise terms (constant variance) and sequences $\{\boldsymbol{\Delta}_k\}$ and $\{\tilde{\boldsymbol{\Delta}}_k\}$ implies that the variance of an individual element in the summands $\hat{\boldsymbol{H}}_k$ is asymptotic to $K\tilde{c}_k^{-2}c_k^{-2}$ for large k and some constant $K>0$ (see Section IV-B). Hence, under the conditions on c_k and $\tilde{c}_k$ given in Theorem 1 (e.g., $0<\gamma\leq 1/4$), the variance of an individual element in a simple average form for $\overline{\boldsymbol{H}}_n$ is given by

$$\frac{1}{n^2}\sum_{k=0}^{n} O\left(\tilde{c}_k^{-2}c_k^{-2}\right) \sim \frac{1}{n^2}K\int_0^n x^{4\gamma}dx = K(4\gamma+1)^{-1}n^{4\gamma-1} \tag{6.1}$$

where "$\sim$" denotes "asymptotic to" (note that for $\gamma=1/4$, the above asymptotic variance of $K/2>0$ does not go to 0, consistent with the lack of convergence associated with Theorem 2a in Spall [17]) . For the weighted average case (see Section IV-B), the corresponding variance of an individual element in $\overline{\boldsymbol{H}}_n$ is

$$\sum_{k=0}^{n}\frac{\tilde{c}_k^4c_k^4O\left(\tilde{c}_k^{-2}c_k^{-2}\right)}{\left(\sum_{i=0}^{n}\tilde{c}_i^2c_i^2\right)^2} \sim \frac{K\int_1^n x^{-4\gamma}dx}{\left(\int_1^n x^{-4\gamma}dx\right)^2} = \begin{cases} K(1-4\gamma)n^{4\gamma-1}+o(n^{4\gamma-1}) & \text{if } \gamma<1/4 \\ K/\log n + o(1/\log n) & \text{if } \gamma = 1/4. \end{cases} \tag{6.2}$$

TABLE I
Asymptotic Ratio of Variances of Elements in Jacobian Estimate as a Function of c_k (and $\tilde{c}_k$) Coefficient γ: Simple Average Over Weighted Average. Note: $\gamma=0.101$ is Popular Practical Choice in SPSA and 2SPSA Settings and $\gamma=1/6$ is Asymptotically Optimal for SPSA and 2SPSA (e.g., Spall [17] and Spall [18, Sect. 7.5]); N/A=Not Applicable (Invalid γ for Simple Average and Weighted Settings)

γ	Ratio in 2SPSA setting	Ratio in root-finding setting
0.101	1.20	1.04
1/6	1.80	1.13
0.24	12.76	1.30
0.25^-	∞	1.33
0.45	N/A	5.26
0.49	N/A	25.25
0.50^-	N/A	∞

The root-finding setting also follows the line of reasoning above. Here, the variance of an individual element in the summands $\hat{\boldsymbol{H}}_k$ is asymptotic to $K'c_k^{-2}$ for large k and some constant $K'>0$ (see Section IV-C). Hence, under the conditions on c_k in Theorem 2 (e.g., $0<\gamma\leq 1/2$), the variance of an individual element in a simple average form for $\overline{\boldsymbol{H}}_n$ is given by

$$\frac{1}{n^2}\sum_{k=0}^{n} O\left(c_k^{-2}\right) \sim \frac{1}{n^2}K'\int_0^n x^{2\gamma}dx = K'(2\gamma+1)^{-1}n^{2\gamma-1} \tag{6.3}$$

(analogous to (6.1), note that for $\gamma=1/2$, the asymptotic variance is $K'/2>0$, consistent with the lack of convergence associated with Theorem 2b in Spall [17]) . For the weighted average case (Section IV-C), the corresponding variance of an individual element in $\overline{\boldsymbol{H}}_n$ is

$$\sum_{k=0}^{n}\frac{c_k^4O\left(c_k^{-2}\right)}{\left(\sum_{i=0}^{n}c_i^2\right)^2} \sim \frac{K'\int_1^n x^{-2\gamma}dx}{\left(\int_1^n x^{-2\gamma}dx\right)^2} = \begin{cases} K'(1-2\gamma)n^{2\gamma-1}+o(n^{2\gamma-1}) & \text{if } \gamma<1/2 \\ K'/\log n + o(1/\log n) & \text{if } \gamma = 1/2. \end{cases} \tag{6.4}$$

Table I shows the asymptotic ratio of variances in the 2SPSA and root-finding cases. For the 2SPSA setting, these are computed by taking the ratio of the right-hand sides of (6.1) to (6.2) (yielding $1/[(4\gamma+1)(1-4\gamma)]$); for the root-finding case, it is (6.3) to (6.4) (yielding $1/[(2\gamma+1)(1-2\gamma)]$).

The table illustrates how the benefits of weighting grow with the value of γ in both the 2SPSA and root-finding settings. While expressions (6.2) and (6.4) suggest $\gamma\approx 0$ is optimal relative to the accuracy for the $\boldsymbol{H}$ estimate in terms of the convergence rate (in n), other values of γ may, in fact, be preferred. In particular, in contrast to only the convergence rate interpretation, the leading coefficients $K(1-4\gamma)$ and $K'(1-2\gamma)$ suggest preferred γ values near the *top* of the allowed range (indicating that, for finite n, an optimal γ to minimize the error in the $\boldsymbol{H}$ estimate may not be near 0). Further, using asymptotic normality of the $\boldsymbol{\theta}$ estimate (not explicitly considered here), $\gamma=1/6$ is asymptotically optimal in the case of 2SPSA in terms of minimizing the error in the $\boldsymbol{\theta}$ estimate, not the $\boldsymbol{H}$ estimate (see Spall [17, Sect. IV-B]); note also that asymptotic normality puts conditions on γ beyond those here, such as $\gamma>1/10$ (see Spall [18,

pp. 164 and 187]). Table I shows a range of values $\gamma > 1/10$ for the above reasons.

VII. Rate of Convergence of Jacobian/Hessian Estimates in Noise-Free Setting

While most applications of SA are for minimization and/or root-finding in the presence of noisy L or $\boldsymbol{g}$ measurements, the algorithms are sometimes used with perfect (noise-free) measurements. For example, SPSA is used for *global* optimization with noise-free (and noisy) measurements in Maryak and Chin [10]; some theory on convergence rates in the noise-free case is given in Gerencsér and Vago [6]. Many of the references at the SPSA web site www.jhuapl.edu/SPSA pertain to applications with noise-free measurements. Hence, there is some interest in the performance of the adaptive approach here with noise-free measurements. Although the general form for the $\boldsymbol{\theta}$ and $\boldsymbol{H}$ recursions in ((2.1a), (2.1b)) continue to apply, the values for a_k and w_k that are desirable (and possibly optimal) in the noisy case are not generally the preferred values in the noise-free case. In particular, the optimal weightings for w_k of Section IV are not recommended in the noise-free case (although, of course, convergence still holds due to the noise-free case being a special case of the noisy case). This section presents rate of convergence results for the Jacobian estimates in the noise-free case.

In the case of noise-free measurements of L, for example, decaying gains satisfying conditions that are different than the standard SPSA conditions are given in Maryak and Chin [10] to ensure global convergence with a generally multimodal loss function; further, constant gains $a_k = a$ are considered in Gerencsér and Vago [6] when the loss function is quadratic. In the noise-free case of direct measurements of $\boldsymbol{g}$, constant gains $a_k = a$ may be used to ensure convergence of what is effectively a quasi-Newton-type algorithm.

Theorems 3 and 4 below consider the settings of L measurements and $\boldsymbol{g}$ measurements, respectively. The theorems have the restriction of quadratic L and affine $\boldsymbol{g}$, respectively. As a consequence, there are no restrictions on the c_k values (i.e., unlike Theorems 1 and 2 above, $\hat{\boldsymbol{H}}_k$ has no $O(c_k^2)$ bias) and, because $\boldsymbol{H}(\boldsymbol{\theta})$ is constant, the results do not depend on the convergence of $\hat{\boldsymbol{\theta}}_k$ (so there are no explicit conditions on the a_k sequence). For this reason, the theorems are best interpreted in most practical problems as local results pertaining to the application of the algorithms when operating in the vicinity of $\boldsymbol{\theta}^*$. (We demonstrate in Section VIII that the implications of the theorems may be at least partially realized in non-quadratic L/non-affine $\boldsymbol{g}$ problems.) For convenience, let $\boldsymbol{\Lambda}_k = \overline{\boldsymbol{H}}_k - \boldsymbol{H}^*$, where $\boldsymbol{H}^* = \boldsymbol{H}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$ due to the quadratic/affine assumption. We write $E(\boldsymbol{\Lambda}_k^T\boldsymbol{\Lambda}_k)$, but note $E(\boldsymbol{\Lambda}_k^T\boldsymbol{\Lambda}_k) = E(\boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k)$ in the optimization (symmetric $\boldsymbol{H}$) case; also note that $\mathrm{trace}[E(\boldsymbol{\Lambda}_k^T\boldsymbol{\Lambda}_k)]$, as used below, corresponds to the expected value of the squared Frobenius (Euclidean) norm of $\boldsymbol{\Lambda}_k$ (i.e., $\mathrm{trace}[E(\boldsymbol{\Lambda}_k^T\boldsymbol{\Lambda}_k)] = E[\mathrm{trace}(\boldsymbol{\Lambda}_k^T\boldsymbol{\Lambda}_k)]$).

Theorem 3 (2SPSA setting): Suppose L is a quadratic function and only noise-free measurements of L are used to form $\boldsymbol{G}_k$ and $\boldsymbol{G}_k^{(1)}$ (see (3.1)). Suppose $0 < w_0 \leq 1$ and $w_k = w/k^\delta$, $k = 1, 2, \ldots, n$, where $1/2 < \delta < 1$ and $0 < w \leq 1$. Suppose that C.8″ in Section V and C.2 and C.9 from Spall [17] hold (see the Appendix here; note that $y(\cdot) = L(\cdot)$ in the setting here) and that Δ_k and $\tilde{\Delta}_k$ are identically distributed at each k and across k. Further, suppose that $\boldsymbol{H}^* > \boldsymbol{0}$ and that $\boldsymbol{f}_k$ in (2.1a) is such that $E(\|\overline{\overline{\boldsymbol{H}}}_k - \overline{\boldsymbol{H}}_k\|^2) = o(e^{-2wk^{1-\delta}/(1-\delta)})$ and $\|\boldsymbol{f}_k(\boldsymbol{H}) - \boldsymbol{H}\|^2/(1 + \|\boldsymbol{H}\|^2)$ is uniformly bounded with respect to k and the set of symmetric $\boldsymbol{H}$ in $\mathbb{R}^{p\times p}$. Then, $\mathrm{trace}[E(\boldsymbol{\Lambda}_n^T\boldsymbol{\Lambda}_n)] = O(e^{-2wn^{1-\delta}/(1-\delta)})$.

Proof: The proof is in three parts: (i) proof of the MSE convergence of $\overline{\boldsymbol{H}}_k$, (ii) derivation of a convenient representation of $\mathrm{trace}[E(\boldsymbol{\Lambda}_n^T\boldsymbol{\Lambda}_n)]$, and (iii) derivation of the main big-O result on rate of convergence.

Part (i): MSE convergence of $\overline{\boldsymbol{H}}_k$: Let us first express the recursion (2.1b) as the SA algorithm

$$\underline{\overline{\boldsymbol{H}}}_k = \underline{\overline{\boldsymbol{H}}}_{k-1} - w_k(\underline{\overline{\boldsymbol{H}}}_{k-1} - \underline{\hat{\boldsymbol{H}}}_k + \underline{\hat{\boldsymbol{\Psi}}}_k) \tag{7.1}$$

where the underbars denote that the unique elements of the associated matrix have been strung into a vector. The above recursion is associated with the root-finding equation $\underline{\boldsymbol{H}} - \underline{\boldsymbol{H}}^* = \boldsymbol{0}$. We first establish that the above recursion converges to $\underline{\boldsymbol{H}}^*$ in the mean-squared sense, and then (in part (iii)) use a resulting expression to show that this convergence must be at the rate in the theorem statement (note that mean-squared convergence results for SA are much less common than a.s. convergence results). Without loss of generality, suppose $w_0 = 1$ (so the $\overline{\boldsymbol{H}}_k$ recursion begins with $\overline{\boldsymbol{H}}_0 = \hat{\boldsymbol{H}}_0 - \hat{\boldsymbol{\Psi}}_0$).

From the quadratic assumption on L, (3.7) implies $\hat{\boldsymbol{H}}_k = \boldsymbol{H}^* + \boldsymbol{\Psi}_k(\boldsymbol{H}^*)$, where we have suppressed the superscript (L) in $\boldsymbol{\Psi}_k^{(L)}$. Hence, by the mutual independence of the sequences $\{\boldsymbol{\Delta}_k, \tilde{\boldsymbol{\Delta}}_k\}$ along k (from conditions C.2 and C.9) and the conditional boundedness of C.8″, recursion (7.1) defines a Markov process with mean-zero noise input (implying $E(\underline{\boldsymbol{H}}^* - \underline{\hat{\boldsymbol{H}}}_k + \underline{\hat{\boldsymbol{\Psi}}}_k|\Im_k^L) = E(\underline{\hat{\boldsymbol{\Psi}}}_k - \underline{\boldsymbol{\Psi}}_k(\boldsymbol{H}^*)|\Im_k^L) = \boldsymbol{0}$ a.s.). Then, given a Lyapunov function, Nevel'son and Has'minskii [11, Sect. 4.4] provide conditions for the mean-squared (m.s.) convergence of $\overline{\boldsymbol{H}}_k$ in (7.1). Consider the Lyapunov function $V(\underline{\boldsymbol{H}}) \equiv 1/2(\underline{\boldsymbol{H}} - \underline{\boldsymbol{H}}^*)^T(\underline{\boldsymbol{H}} - \underline{\boldsymbol{H}}^*)$. From [11, pp. 92–94], a set of conditions related to the differential generating operator for V (e.g., [11, p. 67, eqn. (5.6)]) may be applied to show the m.s. convergence *if*

$$\|\underline{\boldsymbol{H}} - \underline{\boldsymbol{H}}^*\|^2 + E\left(\|\underline{\boldsymbol{\Psi}}_k(\boldsymbol{f}_k(\boldsymbol{H})) - \underline{\boldsymbol{\Psi}}_k(\boldsymbol{H}^*)\|^2\right) \leq C\left(1 + \|\underline{\boldsymbol{H}}\|^2\right) \tag{7.2}$$

for all k and some $C > 0$. (This is a standard condition in SA; see, e.g., Spall [18, p. 106]. Note that (7.2) implies (4.13) in [11], which, in turn, yields the expansion at the top of p. 93 in [11] that is the basis of the Taylor expansion-based arguments in the next paragraph.) By the linearity of $\boldsymbol{\Psi}_k(\cdot)$, the matrix form of the argument within the second term on the left side of (7.2) is

$$\begin{aligned}\boldsymbol{\Psi}_k(\boldsymbol{f}_k(\boldsymbol{H})) - \boldsymbol{\Psi}_k(\boldsymbol{H}^*) &= \boldsymbol{\Psi}_k(\boldsymbol{f}_k(\boldsymbol{H}) - \boldsymbol{H}^*) \\ &= \boldsymbol{\Psi}_k(\boldsymbol{H} - \boldsymbol{H}^*) + \boldsymbol{\Psi}_k(\boldsymbol{f}_k(\boldsymbol{H}) - \boldsymbol{H}).\end{aligned}$$

We have $\|\boldsymbol{f}_k(\boldsymbol{H}) - \boldsymbol{H}\|^2/(1 + \|\boldsymbol{H}\|^2)$ uniformly bounded by assumption. Further, all second moments of elements in $\boldsymbol{D}_k$ and $\tilde{\boldsymbol{D}}_k$, as appear in $\boldsymbol{\Psi}_k(\cdot)$, are uniformly bounded by C.9. Hence, by an application of the triangle and Cauchy-Schwarz inequalities, the second term on the left-hand side of (7.2) is $O(\|\underline{\boldsymbol{H}}\|^2)$ (large $\|\underline{\boldsymbol{H}}\|$), indicating that the inequality in (7.2) is, in fact, true.

Because of the validity of (7.2), Nevel'son and Has'minskii [11, pp. 92–94], show that it is sufficient to have the differential generating operator applied to V (e.g., [11, p. 31, eqn. (1.7) and p. 67, eqn. (5.6)]) bounded above by some function, $-\beta_{k0}V(\underline{\underline{\boldsymbol{H}}})+\beta_{k1}$, for sequences β_{k0} and β_{k1} such that: (i) $0 \leq \beta_{k0} \leq 1$ for all k sufficiently large, (ii) $\sum_{k=0}^{\infty}\beta_{k0}=\infty$, (iii) $\beta_{k1} \to 0$, and (iv) $\beta_{k1}/\beta_{k0} \to 0$. Motivated by the right-hand side of (7.1) and the fact that $\hat{\boldsymbol{H}}_k = \boldsymbol{H}^* + \boldsymbol{\Psi}_k(\boldsymbol{H}^*)$, a Taylor expansion of $V(\underline{\underline{\boldsymbol{H}}} - w_k[\underline{\underline{\boldsymbol{H}}} - \underline{\underline{\boldsymbol{H}}}^* - \underline{\underline{\boldsymbol{\Psi}}}_k(\boldsymbol{H}^*) + \underline{\underline{\boldsymbol{\Psi}}}_k(\boldsymbol{f}_k(\boldsymbol{H}))])$ shows that the generating operator is bounded above by a term $-[w_k + O(w_k^2)]V(\underline{\underline{\boldsymbol{H}}}) + O(w_k^2)$ (see [11, p. 93]). With $\beta_{k0} = w_k + O(w_k^2)$ and $\beta_{k1} = O(w_k^2)$, it is clear from the conditions on w_k that the above conditions (i)–(iv) on the upper bound are satisfied. Hence, $E(\boldsymbol{\Lambda}_k^T\boldsymbol{\Lambda}_k) \to 0$.

Part (ii): Representation of $\mathrm{trace}[E(\boldsymbol{\Lambda}_n^T\boldsymbol{\Lambda}_n)]$*:* Let us first present a form for $\boldsymbol{\Lambda}_n$ that is convenient for representing $\mathrm{trace}[E(\boldsymbol{\Lambda}_n^T\boldsymbol{\Lambda}_n)]$ in a way that facilitates the analysis of its rate of convergence. Let $\boldsymbol{\Lambda}_k' = \overline{\overline{\boldsymbol{H}}}_k - \boldsymbol{H}^*$. From $\boldsymbol{H}^* = (1-w_k)\boldsymbol{H}^* + w_k\boldsymbol{H}^*$ and $\hat{\boldsymbol{H}}_k = \boldsymbol{H}^* + \boldsymbol{\Psi}_k(\boldsymbol{H}^*)$, (7.1) in matrix form implies

$$\begin{aligned}\boldsymbol{\Lambda}_k &= (1-w_k)\boldsymbol{\Lambda}_{k-1} + w_k(\hat{\boldsymbol{H}}_k - \hat{\boldsymbol{\Psi}}_k - \boldsymbol{H}^*)\\ &= (1-w_k)\boldsymbol{\Lambda}_{k-1} - w_k\boldsymbol{\Psi}_k(\overline{\overline{\boldsymbol{H}}}_{k-1}) + w_k\boldsymbol{\Psi}_k(\boldsymbol{H}^*)\\ &= (1-w_k)\boldsymbol{\Lambda}_{k-1} - w_k\boldsymbol{\Psi}_k\left(\boldsymbol{\Lambda}_{k-1}'\right)\end{aligned}$$

where the last equality follows from the form for $\boldsymbol{\Psi}_k(\cdot)$ in (3.8). Hence, solving for $\boldsymbol{\Lambda}_n$, $n \geq 1$, yields

$$\boldsymbol{\Lambda}_n = \left[\prod_{k=1}^{n}(1-w_k)\right]\boldsymbol{\Lambda}_0 - \sum_{k=1}^{n}\left[\prod_{j=k+1}^{n}(1-w_j)\right]w_k\boldsymbol{\Psi}_k\left(\boldsymbol{\Lambda}_{k-1}'\right) \quad \text{a.s.} \quad (7.3)$$

(note: $\Pi_{j=n+1}^{n}(1-w_j) = 1$ for all n).

Let us now characterize $\mathrm{trace}[E(\boldsymbol{\Lambda}_n^T\boldsymbol{\Lambda}_n)]$ using (7.3). From the mutual independence of the sequences $\{\boldsymbol{D}_k, \tilde{\boldsymbol{D}}_k\}$ along k, (7.3) represents a martingale difference sequence, leading to

$$E\left(\boldsymbol{\Lambda}_n^T\boldsymbol{\Lambda}_n\right) = \left[\prod_{k=1}^{n}(1-w_k)\right]^2 E\left(\boldsymbol{\Lambda}_0^T\boldsymbol{\Lambda}_0\right) + \sum_{k=1}^{n}\left[\prod_{j=k+1}^{n}(1-w_j)\right]^2 w_k^2 E\left[\boldsymbol{\Psi}_k\left(\boldsymbol{\Lambda}_{k-1}'\right)^T\boldsymbol{\Psi}_k\left(\boldsymbol{\Lambda}_{k-1}'\right)\right].$$

From (3.8), the product $E[\boldsymbol{\Psi}_k(\boldsymbol{\Lambda}_{k-1}')^T\boldsymbol{\Psi}_k(\boldsymbol{\Lambda}_{k-1}')]$ is formed by adding 36 (not necessarily unique or non-zero) matrix expressions; all of the non-zero matrices involve an expectation containing two $\boldsymbol{\Lambda}_{k-1}'$ values. From the independence of $\{\boldsymbol{D}_k, \tilde{\boldsymbol{D}}_k\}$ and $\boldsymbol{\Lambda}_{k-1}'$

$$\mathrm{trace}\left[E\left(\boldsymbol{\Lambda}_n^T\boldsymbol{\Lambda}_n\right)\right] = \left[\prod_{k=1}^{n}(1-w_k)\right]^2 \mathrm{trace}\left[E\left(\boldsymbol{\Lambda}_0^T\boldsymbol{\Lambda}_0\right)\right] + \sum_{k=1}^{n}\left[\prod_{j=k+1}^{n}(1-w_j)\right]^2 w_k^2\tau\left(E\left(\boldsymbol{\Lambda}_{k-1}'\otimes\boldsymbol{\Lambda}_{k-1}'\right)\right)$$

where $\tau(\cdot)$ represents the trace of a $p\times p$ linear (non-affine) transformation of the $p^2\times p^2$ matrix $E(\boldsymbol{\Lambda}_{k-1}'\otimes\boldsymbol{\Lambda}_{k-1}')$ (the mapping $\tau(\cdot)$ is not a function of k by the identical distribution assumption for $\boldsymbol{\Delta}_k$ and $\tilde{\boldsymbol{\Delta}}_k$ across k). Note that $1-w_k = e^{-w_k}(1-O(w_k^2))$, where the $O(w_k^2)$ term is strictly positive on $0<w_k<1$ by the convexity of e^{-w_k}. Letting $w_{\mathrm{sum}}(i,j) = \sum_{k=i}^{j}w_k$

$$\mathrm{trace}\left[E\left(\boldsymbol{\Lambda}_n^T\boldsymbol{\Lambda}_n\right)\right] = e^{-2w_{\mathrm{sum}}(1,n)}c_{0n}\mathrm{trace}\left[E\left(\boldsymbol{\Lambda}_0^T\boldsymbol{\Lambda}_0\right)\right] + e^{-2w_{\mathrm{sum}}(1,n)}\sum_{k=1}^{n}e^{2w_{\mathrm{sum}}(1,k)}c_{kn}w_k^2\tau\left(E\left(\boldsymbol{\Lambda}_{k-1}'\otimes\boldsymbol{\Lambda}_{k-1}'\right)\right) \quad (7.4)$$

where $c_{kn} = [\prod_{i=k+1}^{n}(1-O(w_i^2))]^2$, $k\geq 0$ ($c_{nn}=1$). By the facts that $0<w_k<1$ for all $k\geq 2$, and $\delta > 1/2$, the c_{kn} are uniformly bounded in magnitude (Apostol [1, p. 208]). Further, by the facts that the w_k are non-negative, monotonically decreasing, and $\sum_{k=i}^{j}w_k \to \infty$ as $j-i\to\infty$ $(i\geq 1)$

$$w_{\mathrm{sum}}(i,j) = \int_i^j \frac{w}{x^\delta}dx + O(1) = \left(\frac{w}{1-\delta}\right)(j^{1-\delta} - i^{1-\delta}) + O(1) \quad (7.5)$$

where the $O(1)$ term is uniformly bounded for all i, j. Then, from (7.4) and (7.5)

$$\mathrm{trace}\left[E\left(\boldsymbol{\Lambda}_n^T\boldsymbol{\Lambda}_n\right)\right] = e^{-2w_{\mathrm{sum}}(1,n)}c_{0n}\mathrm{trace}\left[E\left(\boldsymbol{\Lambda}_0^T\boldsymbol{\Lambda}_0\right)\right] + \bar{c}_n e^{-2wn^{1-\delta}/(1-\delta)}\sum_{k=1}^{n}e^{2wk^{1-\delta}/(1-\delta)} \times \frac{w^2}{k^{2\delta}}\tau\left(E\left(\boldsymbol{\Lambda}_{k-1}'\otimes\boldsymbol{\Lambda}_{k-1}'\right)\right) \quad (7.6)$$

where $\bar{c}_n$ is uniformly bounded in magnitude by the corresponding uniform boundedness of the c_{kn}, the $O(1)$ contributions as in (7.5), and the non-negativity of the summands in (7.4).

Part (iii): The big-O result on rate of convergence: We now demonstrate that $\mathrm{trace}[E(\boldsymbol{\Lambda}_n^T\boldsymbol{\Lambda}_n)] = O(e^{-2wn^{1-\delta}/(1-\delta)})$ by showing that a contradiction exists between the left- and right-hand sides of (7.6) for any slower rate of decay. Hence, suppose $\mathrm{trace}[E(\boldsymbol{\Lambda}_n^T\boldsymbol{\Lambda}_n)] = s(n)e^{-2wn^{1-\delta}(1-\delta)}$ for some function $s(n)$ satisfying $\limsup_{n\to\infty}s(n)=\infty$ and $s(n) = o(e^{2wn^{1-\delta}/(1-\delta)})$ (the latter because $E(\boldsymbol{\Lambda}_n^T\boldsymbol{\Lambda}_n)\to\boldsymbol{0}$). Note that for all $n<\infty$, $s(n)<\infty$ since $\mathrm{trace}[E(\boldsymbol{\Lambda}_n^T\boldsymbol{\Lambda}_n)] < \infty$ by (7.4) and $e^{-2wn^{1-\delta}/(1-\delta)} > 0$. Further, note that $E(\boldsymbol{\Lambda}_k'\otimes\boldsymbol{\Lambda}_k') = E(\boldsymbol{\Lambda}_k\otimes\boldsymbol{\Lambda}_k) + o(s(k)e^{-2wk^{1-\delta}/(1-\delta)})$ by an application of the triangle and Cauchy–Schwarz inequalities to the elements of $\boldsymbol{\Lambda}_k'\otimes\boldsymbol{\Lambda}_k' - \boldsymbol{\Lambda}_k\otimes\boldsymbol{\Lambda}_k$ together with the fact that $E(\|\overline{\overline{\boldsymbol{H}}}_k - \overline{\boldsymbol{H}}_k\|^2) = o(e^{-2wk^{1-\delta}/(1-\delta)})$. Hence, relative to the summands on the right-hand side of (7.6), note that $\tau(E(\boldsymbol{\Lambda}_{k-1}'\otimes\boldsymbol{\Lambda}_{k-1}')) = O(\mathrm{trace}[E(\boldsymbol{\Lambda}_{k-1}^T\boldsymbol{\Lambda}_{k-1})])$ by the fact that both representations represent weighted sums of all possible pairwise products of components of $\boldsymbol{\Lambda}_{k-1}$ to within the above "little-o" error decaying faster than $s(k-1)e^{-2w(k-1)^{1-\delta}/(1-\delta)}$. To establish the contradiction in (7.6), we consider two cases below: (i) The slow increase case

for $s(n)$, where $s(n)/n^{2\delta} = O(1)$, and (ii) the fast increase case, where $\limsup_{n\to\infty} s(n)/n^{2\delta} = \infty$.

Relative to the slow increase case (i), there exists a $0 < b \leq 2\delta$ and $0 < \varepsilon < 2\delta - 1$ such that

$$\frac{s(k)}{k^b} = O(1) \quad \text{and} \quad \limsup_{k\to\infty} \frac{s(k)}{k^{b-\varepsilon}} = \infty \tag{7.7}$$

($b-\varepsilon$ is not necessarily > 0). Hence, the right-hand side of (7.6), say $\mathrm{RHS}(n)$, satisfies

$$\begin{aligned}
\mathrm{RHS}(n) &= O(1)e^{-2w\mathrm{sum}(1,n)} + O(1)e^{-2wn^{1-\delta}/(1-\delta)} \sum_{k=1}^{n} \frac{s(k)}{k^b}\frac{w^2}{k^{2\delta-b}} \\
&= e^{-2wn^{1-\delta}/(1-\delta)}\left[O(1) + O(1)\sum_{k=1}^{n}\frac{1}{k^{2\delta-b}}\right] \\
&= e^{-2wn^{1-\delta}/(1-\delta)}\left[O(1) + O(1)\times\begin{Bmatrix} n^{b+1-2\delta} & \text{if } b \neq 2\delta-1 \\ \log n & \text{if } b = 2\delta - 1 \end{Bmatrix}\right]
\end{aligned}$$

where the first equality uses $\tau(E(\boldsymbol{\Lambda}_{k-1} \otimes \boldsymbol{\Lambda}_{k-1})) = O(\mathrm{trace}[E(\boldsymbol{\Lambda}_{k-1}^T\boldsymbol{\Lambda}_{k-1})])$ and $\mathrm{trace}[E(\boldsymbol{\Lambda}_k^T\boldsymbol{\Lambda}_k)] = s(k)e^{-2wk^{1-\delta}/(1-\delta)}$, the second equality uses (7.5) and the left-hand equation in (7.7), and the last equality uses the monotonicity of the summands (over 1 to n) from the second equality. From the second equation in (7.7), there exists a subsequence $\{k_1, k_2, \ldots, k_n\}$ such that $s(k_n)/k_n^{b-\varepsilon} \to \infty$ as $n \to \infty$. Hence, under the assumed form for $\mathrm{trace}[E(\boldsymbol{\Lambda}_{k_n}^T\boldsymbol{\Lambda}_{k_n})]$, the left-hand side of (7.6), say $\mathrm{LHS}(k_n)$, satisfies $\mathrm{LHS}(k_n) \geq Ck_n^{b-\varepsilon}e^{-2wk_n^{1-\delta}/(1-\delta)}$ for some $C > 0$ and all n sufficiently large. Because $0 < \varepsilon < 2\delta - 1$, we have $\mathrm{RHS}(k_n)/\mathrm{LHS}(k_n) \to 0$ as $n \to \infty$, violating the essential requirement that $\mathrm{LHS}(k) = \mathrm{RHS}(k)$ for all k. Hence, (7.6) leads to a contradiction in the slow increase case (i).

Relative to the fast increase case (ii), $\limsup_{n\to\infty} s(n)/n^{2\delta} = \infty$ and $s(n) < \infty$ for all $n < \infty$ imply that there exists a subsequence $\{k_1, k_2, \ldots, k_n\}$ such that $s(k_n)/k_n^{2\delta} \geq s(k)/k^{2\delta}$ for all $k \leq k_n$ (this is stronger than monotonically nondecreasing along the subsequence). Hence, for each k_n and some $C > 0$, the right-hand side of (7.6), satisfies

$$\begin{aligned}
&\mathrm{RHS}(k_n) = O(1)e^{-2w\mathrm{sum}(1,k_n)} \\
&+O(1)e^{-2wk_n^{1-\delta}/(1-\delta)}\left[\sum_{k=1}^{k_n} e^{2wk^{1-\delta}/(1-\delta)}\frac{w^2}{k^{2\delta}}\tau\left(E(\boldsymbol{\Lambda}_{k-1}\otimes\boldsymbol{\Lambda}_{k-1})\right)\right] \\
&\leq Ce^{-2wk_n^{1-\delta}/(1-\delta)}\left[1+\sum_{k=1}^{k_n}\frac{s(k)}{k^{2\delta}}\right] \\
&\leq Ce^{-2wk_n^{1-\delta}/(1-\delta)}\left[1+\frac{s(k_n)}{k_n^{2\delta-1}}\right].
\end{aligned}$$

In contrast, the left-hand side of (7.6) is $e^{-2wk_n^{1-\delta}/(1-\delta)}s(k_n)$, implying $\mathrm{RHS}(k_n)/\mathrm{LHS}(k_n) \to 0$ as $n \to \infty$ since $2\delta - 1 > 0$. Once again, this violates the essential requirement that $\mathrm{LHS}(k) = \mathrm{RHS}(k)$ for all k, indicating a contradiction. We have thus shown that (7.6) requires $\mathrm{trace}[E(\boldsymbol{\Lambda}_n^T\boldsymbol{\Lambda}_n)] = O(e^{-2wn^{1-\delta}/(1-\delta)})$. *Q.E.D.*

Theorem 4 below covers the noise-free root-finding setting. There are, of course, a number of well-known algorithms that apply in such a setting (e.g., conjugate gradient and quasi-Newton-type methods) and we make no claims here about the relative efficiency of these well-known methods and the adaptive SPSA method. Nevertheless, potential advantages of the method here are the avoidance of the line search that is embedded in some standard conjugate gradient and related methods and the ability to process either noisy or noise-free measurements within a data stream without fundamentally changing the algorithm.

Theorem 4 (Root-Finding Setting): Suppose $\boldsymbol{g}$ is an affine function and only noise-free measurements of $\boldsymbol{g}$ are used to form $\boldsymbol{G}_k$ and $\boldsymbol{G}_k^{(1)}$. Suppose $0 < w_0 \leq 1$ and $w_k = w/k^\delta$, $k = 1, 2, \ldots$, where $1/2 < \delta < 1$ and $0 <\leq 1$. Suppose that C.2′ and C.9′ from Spall [17] hold (see the Appendix here) and that the Δ_k are identically distributed across k. Further, suppose that $\boldsymbol{H}^* > \boldsymbol{0}$ and that $\boldsymbol{f}_k$ in (2.1a) is such that $E(\|\overline{\overline{\boldsymbol{H}}}_k - \overline{\boldsymbol{H}}_k\|^2) = o(e^{-2wk^{1-\delta}/(1-\delta)})$ and $\|\boldsymbol{f}_k(\boldsymbol{H}) - \boldsymbol{H}\|^2/(1+\|\boldsymbol{H}\|^2)$ is uniformly bounded with respect to k and the set of $\boldsymbol{H}$ in $\mathbb{R}^{p\times p}$ (the $\boldsymbol{H}$ are also symmetric in the optimization case). Then, $\mathrm{trace}[E(\boldsymbol{\Lambda}_n^T\boldsymbol{\Lambda}_n)] = O(e^{-2wn^{1-\delta}/(1-\delta)})$.

Proof: From the affine, noise-free assumption on $\boldsymbol{g}$, (3.11) implies $\hat{\boldsymbol{H}}_k = \boldsymbol{H}^* + \boldsymbol{\Psi}_k^{(\boldsymbol{g})}(\boldsymbol{H}^*)$. Then, the proof follows exactly as in the proof of Theorem 3 with the exception of using the simpler function $\boldsymbol{\Psi}_k^{(\boldsymbol{g})}$ (see (3.12)) in place of $\boldsymbol{\Psi}_k^{(L)}$. *Q.E.D.*

From the basic recursion (2.1b), it is apparent that the analyst must specify w_k. Theorems 3 and 4 suggest that the rate of convergence for $\overline{\boldsymbol{H}}_k$ in the noise-free case is maximized by choosing $\delta > 1/2$ arbitrarily close to 1/2.

VIII. Numerical Study

Consider the fourth-order loss function used in numerical demonstrations of Spall [17]

$$L(\boldsymbol{\theta}) = \boldsymbol{\theta}^T\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{\theta} + 0.1\sum_{i=1}^{p}(\boldsymbol{B}\boldsymbol{\theta})_i^3 + 0.01\sum_{i=1}^{p}(\boldsymbol{B}\boldsymbol{\theta})_i^4 \tag{8.1}$$

where $(\cdot)_i$ represents the ith component of the argument vector $\boldsymbol{B}\boldsymbol{\theta}$, and $\boldsymbol{B}$ is such that $p\boldsymbol{B}$ is an upper triangular matrix of 1's (so elements below the diagonal are zero). Let $p = 10$. The minimum occurs at $\boldsymbol{\theta}^* = \boldsymbol{0}$ with $L(\boldsymbol{\theta}^*) = 0$; all runs are initialized at $\hat{\boldsymbol{\theta}}_0 = [0.2, 0.2, \ldots, 0.2]^T$ (so $L(\hat{\boldsymbol{\theta}}_0) = 0.1565$). All iterates are constrained to be in $\boldsymbol{\theta} = [-10, 10]^{10}$. The mechanism for producing $\overline{\overline{\boldsymbol{H}}}_k$ discussed in Section II is used to ensure positive definiteness of the Hessian estimates with $\delta_k = 0.0001e^{-k}$ (other small, decaying sequences δ_k provide results indistinguishable from those here).

This study has two general parts. The first compares the standard adaptive SPSA method with the enhanced method when direct (noisy) measurements of $\boldsymbol{g}$ are available. This corresponds to the 2SG (second-order stochastic gradient) setting of Spall [17] (a special case of stochastic root-finding). The second part evaluates the numerical implications of Theorem 3 on noise-free SPSA, corresponding to the 2SPSA (second-order SPSA) setting of Spall [17]. Following practical guidelines in Spall [17, Sect. II.D], an iteration is blocked if $\boldsymbol{\theta}$ moves too far (an indication of algorithm instability); in particular if $\|\hat{\boldsymbol{\theta}}_{k+1} - \hat{\boldsymbol{\theta}}_k\| \geq 1.0$, then the step is blocked and $\hat{\boldsymbol{\theta}}_{k+1}$ is reset to $\hat{\boldsymbol{\theta}}_k$. We used the standard forms for the gain sequences a_k and c_k: $a_k = a/(k+1+A)^\alpha$ and $c_k = c/(k+1)^\gamma$, where a, c, α, and γ are strictly positive and the stability constant $A \leq 0$ (see Spall [18, Sects. 4.4, 6.6, or 7.5] for further discussion of these gain forms).

TABLE II
SAMPLE MEANS FOR TERMINAL VALUES OF NORMALIZED LOSS FUNCTIONS $[L(\hat{\boldsymbol{\theta}}_k) - L(\boldsymbol{\theta}^*)]/[L(\hat{\boldsymbol{\theta}}_0) - L(\boldsymbol{\theta}^*)]$; INDICATED P-VALUES ARE FOR DIFFERENCES BETWEEN SAMPLE MEANS OF NORMALIZED LOSS FUNCTIONS AND FOR DIFFERENCES BETWEEN SAMPLE MEANS OF HESSIAN ESTIMATION ERRORS $\|\overline{\overline{\boldsymbol{H}}}_k - \boldsymbol{H}^*\|$. N/A (NOT APPLICABLE) IS USED FOR P-VALUES >0.5 (I.E., WHEN THE SAMPLE MEAN FOR ENHANCED IS GREATER THAN THE SAMPLE MEAN FOR STANDARD)

Number of iterations	Sample mean for standard 2SG loss functions	Type of enhanced 2SG	Sample mean for enhanced 2SG loss functions	P-values for loss functions	P-values for Hessian estimates
2000	0.019	Feedback (F)	0.031	N/A	N/A
		Weighting (W)	0.028	N/A	N/A
		Joint F & W	**0.012**	**0.0061**	**1.2×10^{-8}**
10,000	0.015	Feedback (F)	0.010	0.085	0.30
		Weighting (W)	0.0053	0.0041	$< 1\times10^{-10}$
		Joint F & W	**0.0034**	**0.00049**	**$< 1\times10^{-10}$**

The results of the 2SG study are in Table II for 2000 and 10,000 iterations. The measurement noises for the underlying L measurements are dependent on $\boldsymbol{\theta}$ in the sense that $\varepsilon = \varepsilon(\boldsymbol{\theta}) = [\boldsymbol{\theta}^T, 1]\boldsymbol{V}$, where $\boldsymbol{V}$ is an i.i.d. vector (across L or $\boldsymbol{g}$ measurements) with distribution $N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{11})$. Hence, the distribution of the noise $\boldsymbol{e}$ in the gradient measurements used in the standard and enhanced 2SG algorithms is i.i.d. $N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{10})$. The noise level $\sigma = 0.05$ is relatively large when compared to the values of the elements in $\boldsymbol{g}(\boldsymbol{\theta})$; the values in $\boldsymbol{g}(\boldsymbol{\theta})$ range from 0.04 to 0.22 at $\hat{\boldsymbol{\theta}}_0$ to all zeroes at $\boldsymbol{\theta}^*$. In each row of the table, the standard and enhanced algorithms are run with the same gain sequence coefficients a, A, α, c, γ. The asymptotically optimal $\alpha = 1$ is used because $\hat{\boldsymbol{\theta}}_0$ is relatively close to $\boldsymbol{\theta}^*$; also $\gamma = 0.49$ is used because of the increased asymptotic accuracy of the Hessian estimate shown in Table I (relative to lower values of γ). We choose $c = 0.05$ and $A = 100$, consistent with guidelines in Spall [18, Sects. 4.4, 6.6, or 7.5] (as noted in Spall [17], it is possible that choosing c larger than one standard deviation may improve the accuracy of the Hessian estimate when the resulting reduction in noise more than compensates for the possible increase in bias). The critical step-size a is tuned to approximately optimize the *finite-sample* performance of the *standard* 2SG method given the choices for A, α, c, γ above; we let $a = 100$. Note that these gains a_k and c_k satisfy the convergence conditions for 2SG in Spall [17] (which apply to the enhanced 2SG as well).[1]

For each iteration count used here (2000 or 10,000), the table shows three implementations of the enhanced method: (i) "Feedback (F)," which uses the feedback form in Section III-C, but no optimal weighting (i.e., the simple average of $\hat{\boldsymbol{H}}_k$ in Spall [17] is used), (ii) "Weighting (W)," which uses the asymptotically optimal weights in Section IV-C, but no feedback, and (iii) "Joint F & W," which uses both feedback and optimal weighting. This joint version is the main intended implementation for the enhanced 2SG in practice.

The sample means for the normalized terminal loss values in Table II are formed from 50 independent runs. The indicated one-sided P-values are based on the two-sample t-test and represent the probability in a future experiment that the sample mean for the particular enhanced case is at least as much below the sample mean for the standard 2SG case as the enhanced case is below the standard 2SG case in the observed sample here under the null hypothesis that the true means are identical. The table shows P-values for the terminal loss function values and for the error in the final Hessian estimates. Separate statistical tests were also done with the distribution-free Wilcoxon rank test as a check on the t-test results here; the Wilcoxon P-values were similar to the values in the table.

In the main Joint F & W implementations for enhanced 2SG, the P-values are small for both loss values and Hessian estimates, consistent with the enhanced 2SG algorithm being statistically significantly better than the standard 2SG algorithm (i.e., rejecting the null hypothesis of equality of means). The results when using feedback or weighting individually were not as strong, and, in fact, they were sometimes inferior to those from standard 2SG (but recall that the gain values were not tuned for the enhanced implementations). It is interesting that the cumulative (joint) effect of feedback and weighting may be quite strong when the individual effects are relatively weak or even when the individual effects lead to no improvement (as with the 2000-iteration case). The greater improvement due to weighting (over feedback) is consistent with the high noise levels here (recall that feedback provides no *asymptotic* advantage in a noisy environment). An analyst might also be interested in the fraction of runs resulting in a better Hessian estimate in the enhanced 2SG case; it was found, in the joint F & W cases, that lower values of $\|\overline{\overline{\boldsymbol{H}}}_k - \boldsymbol{H}^*\|$ were produced in 44/50 runs with 2000 iterations and in 47/50 runs with 10,000 iterations. The computational times for enhanced 2SG (joint F & W) were negligibly greater (1.4 to 2.4 percent) than the times for standard 2SG. In addition, while not shown here, Spall [17] includes a brief comparison with iterate averaging on the same test function. Iterate averaging performed relatively poorly—

[1]As discussed in Spall [17], the asymptotically optimal values of a and α are both unity. At these values, the worth of exact Hessian information can be seen by comparing standard 2SG with an idealized (infeasible) algorithm based on $\boldsymbol{H}^*$ in place of $\overline{\overline{\boldsymbol{H}}}_k$ at all iterations. At 10,000 iterations, and with all else as in Table II except $A = 0$, the mean normalized loss is 0.060 for standard 2SG and 0.0008 for the algorithm with $\boldsymbol{H}^*$ (P-value $< 1 \times 10^{-10}$). However, at $a = A = 100$, as in Table II, the superior algorithm switches in the sense that the mean normalized loss is 0.0153 for standard 2SG and 0.0358 for the algorithm with $\boldsymbol{H}^*$ (P-value$= 2.7 \times 10^{-5}$). This is illustrative of the gap between finite sample results and asymptotic theory, a gap apparent even in deterministic optimization with the Newton-Raphson method (e.g., Spall [18, p. 28]).

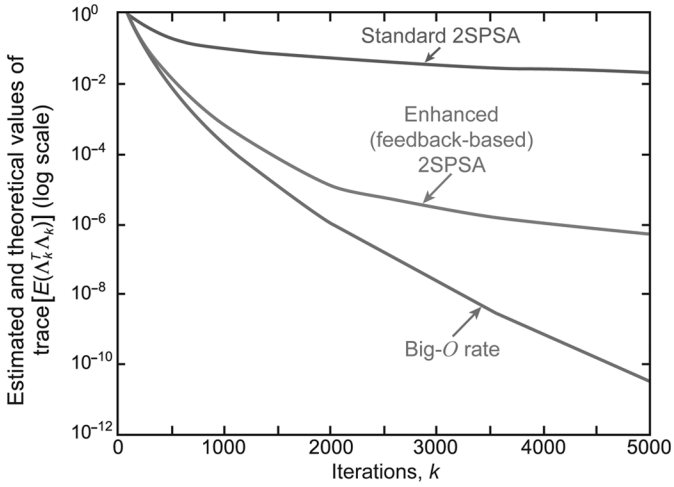

Fig. 1. Estimated and theoretical values of $\mathrm{trace}[E(\mathbf{\Lambda}_k^T\mathbf{\Lambda}_k)]$ as a function of k. The top two curves are based on numerical experiments (50 runs) for the given *fourth-order* loss function; the bottom curve is the idealized rate for quadratic loss functions from the big-O bound of Theorem 3.

significantly poorer than standard 2SG—for reasons discussed in Spall [18, pp. 117–119].

We now consider the noise-free 2SPSA setting and the implications of Theorem 3. Fig. 1 shows estimated values for $\mathrm{trace}[E(\mathbf{\Lambda}_k^T\mathbf{\Lambda}_k)]$ for the standard and enhanced (feedback-based) 2SPSA algorithms; the estimates are formed from a sample mean of 50 independent runs for each of the two curves. The third (lower) curve shows the theoretical big-O bound from Theorem 3. All curves are normalized to have a value of unity at $k = 100$ to help adjust for transient effects. We use $w_k = w/k^\delta$, $k \geq 1$, as in Theorem 1, with $\delta = 0.501$ (i.e., slightly over 1/2 per the allowable range for δ) and $w = 1/p = 0.1$ (note that the optimal w_k in (4.2) is not used here, as it is not aimed at the noise-free setting).

Fig. 1 shows that the $\mathrm{trace}[E(\mathbf{\Lambda}_k^T\mathbf{\Lambda}_k)]$ values in the enhanced implementation are much lower than the corresponding values in the standard 2SPSA averaging method, consistent with the improved rate of convergence predicted from Theorem 3. However, the specific numerical values are greater than those of the big-O bound. This is unsurprising in light of the formal restriction in Theorem 3 to quadratic loss functions (versus the fourth-order function here). We see that the practical convergence rate for the non-quadratic loss lies between the non-feedback rate (standard 2SPSA) and the idealized rate from the big-O bound.[2]

IX. Conclusions

This paper has presented an SA approach built on methods for efficient estimation of the Jacobian (Hessian) matrix. The approach here is motivated by the adaptive simultaneous perturbation SA method in Spall [17]. We introduced two significant enhancements of the adaptive SPSA method: a feedback process that reduces the error in the cumulative Jacobian estimate across iterations and an optimal weighting of per-iteration Jacobian estimates to account for noise. The feedback process is also useful in the noise-free setting, leading to a near-exponential decay in the error of the Jacobian estimates.

The convergence theory for the $\boldsymbol{\theta}$ iterate given in Spall [17] continues to hold in the enhanced Jacobian setting here. The theory for the Jacobian estimate, however, must be changed to accommodate the different algorithm forms. We establish conditions for the a.s. convergence of the Jacobian estimate and analyze rates of convergence in both the noisy and noise-free settings. In turn, the asymptotic normality of the $\boldsymbol{\theta}$ estimate in Spall [17] continues to hold, implying that for the optimization problem of minimizing L, the adaptive approach here is nearly asymptotically optimal in the gradient-free case (noisy measurements of L) and is asymptotically optimal in the stochastic gradient case with direct measurements of $\boldsymbol{g}$.

Although the method here is a *relatively* simple adaptive approach and the theory and numerical experience point to the improvements possible, one should, as in all stochastic algorithms, be careful in implementation. Such care extends to the choice of initial conditions and choice of algorithm coefficients (although the effort can be reduced by using the asymptotically near-optimal or optimal $a_k = 1/(k+1)$ for the important gain sequence if the initial condition is sufficiently close to the optimum). Nonetheless, the adaptive approach can offer impressive gains in efficiency.

Appendix

This appendix provides the regularity conditions for the a.s. convergence of $\hat{\boldsymbol{\theta}}_k$ to $\boldsymbol{\theta}^*$. These correspond to the conditions in Spall [17] for Theorems 1a (loss minimization based on noisy measurements of L) and 1b (root finding based on noisy measurements of $\boldsymbol{g}$), with slight modifications for clarification and for the slightly different notation of this paper. Let $\boldsymbol{\theta} = [t_1, t_2, \ldots, t_p]^T$, $\boldsymbol{\theta}^* = [t_1^*, t_2^*, \ldots, t_p^*]^T$, i.o. represent infinitely often, and $\overline{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k) = \overline{\overline{\boldsymbol{H}}}_k^{-1}\boldsymbol{g}(\hat{\boldsymbol{\theta}}_k)$ (with ith component $\overline{g}_{ki}(\hat{\boldsymbol{\theta}}_k)$):

C.0: $E(\hat{\varepsilon}_k^{(\pm)} - \varepsilon_k^{(\pm)} | \Im_k^L; \mathbf{\Delta}_k; \tilde{\mathbf{\Delta}}_k) = 0$ a.s. for all k.

C.1: $a_k, c_k > 0$; $a_k \to 0$, $c_k \to 0$ 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 δ, $\rho > 0$ and for all k, ℓ, $E[|y(\hat{\boldsymbol{\theta}}_k \pm c_k\mathbf{\Delta}_k + c_k\tilde{\mathbf{\Delta}}_k)/(\Delta_{k\ell}\tilde{\Delta}_{km})|^{2+\delta}] \leq \rho$, $E[|y(\hat{\boldsymbol{\theta}}_k \pm c_k\mathbf{\Delta}_k)/(\Delta_{k\ell}\tilde{\Delta}_{km})|^{2+\delta}] \leq \rho$, $|\Delta_{k\ell}| \leq \rho$, $\Delta_{k\ell}$ is symmetrically distributed about 0; and $\Delta_{k\ell}$ are mutually independent across k and ℓ.

C.3: For some $\rho > 0$ and almost all $\hat{\boldsymbol{\theta}}_k$, the function $\boldsymbol{g}(\cdot)$ is continuously twice differentiable with uniformly (in k) bounded second derivative for all $\boldsymbol{\theta}$ such that $\|\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}\| \leq \rho$.

C.4: For each $k \geq 1$ and all $\boldsymbol{\theta}$, there exists a $\rho > 0$ not dependent on k and $\boldsymbol{\theta}$, such that $(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T\overline{\boldsymbol{g}}_k(\boldsymbol{\theta}) \geq \rho\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|$.

C.5: For each $i = 1, 2, \ldots, p$, and any $\rho > 0$, $P(\{\overline{g}_{ki}(\hat{\boldsymbol{\theta}}_k) \geq 0 \text{ i.o.}\} \cap \{\overline{g}_{ki}(\hat{\boldsymbol{\theta}}_k) < 0 \text{ i.o.}\} | \{|\hat{\theta}_{ki} - t_i^*| \geq \rho \forall k\}) = 0$ (see the Lemma in Spall [17, p. 1845], for an easier-to-verify sufficient condition for C.5).

C.6: $\overline{\overline{\boldsymbol{H}}}_k^{-1}$ exists a.s. for all k, $c_k^2\overline{\overline{\boldsymbol{H}}}_k^{-1} \to \mathbf{0}$ a.s., and for some δ, $\rho > 0$, $E(\|\overline{\overline{\boldsymbol{H}}}_k^{-1}\|^{2+\delta}) \leq \rho$.

[2]We also considered the predictive performance of the big-O bound on a *quadratic* version of L, formed by simply truncating the third- and fourth-order parts of L in (8.1). The big-O bound accurately predicts the decay rate. For example, both the bound and the sample mean of $\mathrm{trace}[E(\mathbf{\Lambda}_k^T\mathbf{\Lambda}_k)]$ (50 runs) drop by a multiple of 3.3×10^{-5} when going from 2000 to 5000 iterations.

C.7: For any $\tau > 0$ and nonempty $S \subseteq \{1, 2, \ldots, p\}$, there exists a $\rho'(\tau, S) > \tau$ such that

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

for all $|t_i - t_i^*| < \tau$ when $i \notin S$ and $|t_i - t_i^*| \geq \rho'(\tau, S)$ when $i \in S$ (see the Lemma in Spall [17, p. 1845] for an easier-to-verify sufficient condition for C.7).

Theorem 1a—2SPSA (Spall [17]): Suppose only noisy measurements of L are used to form $\boldsymbol{G}_k$ and $\boldsymbol{G}_k^{(1)}$ (see (3.1)). Let conditions C.0 through C.7 hold. Then $\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^* \to \mathbf{0}$ a.s.

Theorem 1b below is the root-finding analogue of Theorem 1a on 2SPSA. We replace C.0, C.1, and C.2 with the following conditions:

C.0′: $E[\boldsymbol{e}_k(\hat{\boldsymbol{\theta}}_k) | \Im_k^{\boldsymbol{g}}, \boldsymbol{\Delta}_k] = \mathbf{0}$ a.s. for all k.

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

C.2′: For some δ, $\rho > 0$, $E(\|\boldsymbol{G}_k(\hat{\boldsymbol{\theta}}_k)\|^{2+\delta}) \leq \rho$ for all k.

Theorem 1b—Root-Finding (Spall [17]): Suppose direct measurements of $\boldsymbol{g}$ are used to form $\boldsymbol{G}_k$ and $\boldsymbol{G}_k^{(1)}$ (as in Section III-B). Let conditions C.0′ through C.2′ and C.3 through C.7 hold. Then $\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^* \to \mathbf{0}$ a.s.

In addition to the conditions above, the following three conditions from Spall [17] are used in Theorems 1–4 here:

C.3′: Change "twice differentiable" in C.3 to "three-times differentiable" with all else unchanged.

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

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

ACKNOWLEDGMENT

The author wishes to acknowledge the insightful comments of the reviewers on many parts of the paper and the comments of Dr. F. Torcaso (JHU) on the proof of Theorem 1.

REFERENCES

[1] T. M. Apostol, *Mathematical Analysis*, 2nd ed. Reading, MA: Addison-Wesley, 1974.

[2] S. Bhatnagar, "Adaptive multivariate three-timescale stochastic approximation algorithms for simulation-based optimization," *ACM Trans. Modeling Comput. Simul.*, vol. 15, pp. 74–107, 2005.

[3] S. Bhatnagar, "Adaptive Newton-based multivariate smoothed functional algorithms for simulation optimization," *ACM Trans. Modeling Comput. Simul.*, vol. 18, pp. 2:1–2:35, 2007.

[4] Y. S. Chow and H. Teicher, *Probability Theory: Independence, Interchangeability, and Martingales*, 2nd ed. New York: Springer-Verlag, 1988.

[5] V. Fabian, "Stochastic approximation," in *Optimizing Methods in Statistics*, J. S. Rustigi, Ed. New York: Academic Press, 1971, pp. 439–470.

[6] L. Gerencsér and Z. Vago, "The mathematics of noise-free SPSA," in *Proc. IEEE Conf. Decision Control*, Dec. 4–7, 2001, pp. 4400–4405.

[7] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed. New York: Springer-Verlag, 2003.

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

[9] O. Macchi and E. Eweda, "Second-order convergence analysis of stochastic adaptive linear filtering," *IEEE Trans. Automat. Control*, vol. AC-28, no. 1, pp. 76–85, Jan. 1983.

[10] J. L. Maryak and D. C. Chin, "Global random optimization by simultaneous perturbation stochastic approximation," *IEEE Trans. Automat. Control*, vol. 53, no. 3, pp. 780–783, Apr. 2008.

[11] M. B. Nevel'son and R. Z. Has'minskii, *Stochastic Approximation and Recursive Estimation*. Providence, RI: American Mathematical Society, 1973.

[12] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM J. Control Optim.*, vol. 30, pp. 838–855, 1992.

[13] B. T. Polyak and Y. Z. Tsypkin, "Optimal and robust methods for stochastic optimization," *Nova J. Math., Game Theory, Algebra*, vol. 6, pp. 163–176, 1997.

[14] D. Ruppert, "A Newton-Raphson version of the multivariate Robbins-Monro procedure," *Annals Statistics*, vol. 13, pp. 236–245, 1985.

[15] H.-P. Schwefel, *Evolution and Optimum Seeking*. New York: Wiley, 1995.

[16] J. C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Trans. Automat. Control*, vol. 37, no. 3, pp. 332–341, Mar. 1992.

[17] J. C. Spall, "Adaptive stochastic approximation by the simultaneous perturbation method," *IEEE Trans. Automat. Control*, vol. 45, no. 10, pp. 1839–1853, Oct. 2000.

[18] J. C. Spall, *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Hoboken, NJ: Wiley, 2003.

[19] C. Z. Wei, "Multivariate adaptive stochastic approximation," *Annals Statistics*, vol. 15, pp. 1115–1130, 1987.

[20] S. Yakowitz, P. L'Ecuyer, and F. Vazquez-Abad, "Global stochastic optimization with low-dispersion point sets," *Oper. Res.*, vol. 48, pp. 939–950, 2000.

[21] G. Yin and Y. Zhu, "Averaging procedures in adaptive filtering: An efficient approach," *IEEE Trans. Automat. Control*, vol. 37, no. 4, pp. 466–475, Apr. 1992.

[22] X. Zhu and J. C. Spall, "A modified second-order SPSA optimization algorithm for finite samples," *Int. J. Adaptive Control Signal Processing*, vol. 16, pp. 397–409, 2002.

James C. Spall (S'82–M'83–SM'90–F'03) is a member of the Principal Professional Staff at the Johns Hopkins University (JHU), Applied Physics Laboratory, Laurel, MD, and a Research Professor in the JHU Department of Applied Mathematics and Statistics, Baltimore, MD. He is also Chairman of the Applied and Computational Mathematics Program within the JHU Engineering Programs for Professionals. He has published many articles in the areas of statistics and control and holds two U.S. patents (both licensed) for inventions in control systems. He is the Editor and co-author for the book *Bayesian Analysis of Time Series and Dynamic Models* (New York: Marcel Dekker, 1988) and is the author of *Introduction to Stochastic Search and Optimization* (New York: Wiley, 2003).

Dr. Spall is a Senior Editor for the IEEE TRANSACTIONS ON AUTOMATIC CONTROL and a Contributing Editor for the *Current Index to Statistics*. He was the Program Chair for the 2007 IEEE Conference on Decision and Control.