

Convergence Analysis for Feedback- and Weighting-Based Jacobian Estimates in the Adaptive Simultaneous Perturbation Algorithm

James C. Spall (james.spall@jhuapl.edu)

The Johns Hopkins University
 Applied Physics Laboratory
 11100 Johns Hopkins Road
 Laurel, Maryland 20723-6099 U.S.A.

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. In a recent paper, Spall (2006) 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 the formation of an optimal weighting of “per-iteration” Jacobian estimates. This paper provides a formal convergence analysis for the algorithm introduced in Spall (2006). In particular, we present conditions for the almost sure convergence of the Jacobian estimates with the feedback and weighting. We also develop results for the rate of convergence in both the noisy and noise-free settings.

Keywords— Stochastic optimization; Jacobian matrix; root-finding; stochastic approximation; simultaneous perturbation stochastic approximation (SPSA); adaptive estimation.

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. Spall (2006) presents an approach for accelerating the convergence of SA algorithms through two enhancements to the adaptive simultaneous perturbation SA (SPSA) approach in Spall (2000). This adaptive 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) 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. This paper introduces the basic ideas associated with the two enhancements and presents a small-scale numerical study.

Acknowledgments—This work was partially supported by U.S. Navy Contract N00024-03-D-6606. Some details have been eliminated from this paper to meet CDC length constraints. In particular, a numerical study illustrating relative convergence rates is available from the author upon request.

The basic problem of interest will be the root-finding problem. That is, for a function $\mathbf{g}(\boldsymbol{\theta}): \mathbb{R}^p \rightarrow \mathbb{R}^p$, $p \geq 1$, we are interested in finding a point $\boldsymbol{\theta}$ satisfying $\mathbf{g}(\boldsymbol{\theta}) = \mathbf{0}$. 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 $\mathbf{g}(\boldsymbol{\theta}) = \partial L / \partial \boldsymbol{\theta} = \mathbf{0}$. Let $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ be a point satisfying $\mathbf{g}(\boldsymbol{\theta}) = \mathbf{0}$. The stochastic setting here allows for the use of only noisy values of \mathbf{g} and the estimation (versus exact calculation) of the associated $p \times p$ Jacobian matrix $\mathbf{H} = \mathbf{H}(\boldsymbol{\theta}) \equiv \partial \mathbf{g}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^T$. Note that the Jacobian matrix is a Hessian matrix of L when \mathbf{g} represents the gradient of L . As described in Spall (2000), simultaneous perturbation ideas that are used for gradient estimation in Spall (1992) can also be used for the per-iteration Jacobian matrix estimation as part of an adaptive stochastic approximation algorithm.

A number of others have looked at ways of enhancing the convergence of SA, including adaptive methods for Jacobian estimation (e.g., Fabian, 1971; Ruppert, 1985; and Wei, 1987) and iterate averaging (e.g., Polyak and Juditsky, 1992). A relatively recent review of such methods is in Spall (2003, Sect. 4.5); as discussed in the review, these methods are typically costly in number of measurements needed and may not yield the desired improvements in practical efficiency for the primary parameters of interest $\boldsymbol{\theta}$. There are also 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, 1983; Yin and Zhu, 1992; and Kushner and Yin, 2003, pp. 8–10); while these are more efficient than the general adaptive approaches mentioned above, they are more restricted in their range of application. This motivates the need for theoretically sound and practically efficient methods for Jacobian estimation, as considered here and in Spall (2006).

II. SUMMARY OF ALGORITHM FORM AND PER-ITERATION JACOBIAN (HESSIAN) ESTIMATE

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:

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

$$\bar{\mathbf{H}}_k = (1 - w_k) \bar{\mathbf{H}}_{k-1} + w_k (\hat{\mathbf{H}}_k - \hat{\Psi}_k), \quad k = 0, 1, 2, \dots, \quad (2.1b)$$

where a_k is a non-negative scalar gain coefficient, $\mathbf{G}_k(\hat{\boldsymbol{\theta}}_k)$ is

some unbiased or nearly unbiased estimate of $\mathbf{g}(\hat{\boldsymbol{\theta}}_k)$, $\mathbf{f}_k: \mathbb{R}^{p \times p} \rightarrow \{\text{invertible } p \times p \text{ matrices}\}$ is a mapping designed to cope with possible noninvertibility of $\bar{\mathbf{H}}_k$, $0 \leq w_k \leq 1$ is a weight to apply to the new input to the recursion for $\bar{\mathbf{H}}_k$, $\hat{\mathbf{H}}_k$ is a per-iteration estimate of $\mathbf{H} = \mathbf{H}(\boldsymbol{\theta})$, and $\hat{\Psi}_k$ is the feedback-based adjustment that is aimed at improving the per-iteration estimate. The two recursions above are identical to those in Spall (2000) with the exception of the more general weighting w_k in the second recursion ($w_k = 1/(k+1)$ in Spall, 2000, equivalent to a recursive calculation of the sample mean of the per-iteration $\mathbf{H}(\boldsymbol{\theta})$ estimates) and the inclusion of the adjustment $\hat{\Psi}_k$. Note that at $k = 0$ in (2.1b), $\bar{\mathbf{H}}_{k-1} = \bar{\mathbf{H}}_{-1}$ may be used to reflect prior information on \mathbf{H} if $0 < w_0 < 1$; alternatively, $\bar{\mathbf{H}}_{-1}$ may be unspecified—and irrelevant—when $w_0 = 1$. Because $\hat{\mathbf{H}}_k$ is defined in Spall (2000), the essential aspects of the parallel recursions in (2.1a, b) that remain to be specified are w_k and $\hat{\Psi}_k$.

Given that $\bar{\mathbf{H}}_k$ may not be invertible (especially for small k), a simple mapping \mathbf{f}_k is to add a matrix $\delta_k \mathbf{I}_p$ to $\bar{\mathbf{H}}_k$, where $\delta_k > 0$, $\delta_k \rightarrow 0$, and \mathbf{I}_p is a $p \times p$ identity matrix. In the case of optimization, where $\mathbf{g}(\boldsymbol{\theta})$ is a gradient and $\mathbf{H}(\boldsymbol{\theta})$ is a Hessian matrix, one may also wish to impose the requirement that the Hessian estimates be symmetric. (Bhatnagar, 2005, discusses Hessian estimation without imposing symmetry at each iteration.) In this case $\mathbf{f}_k: \mathbb{R}^{p \times p} \rightarrow \{\text{symmetric positive definite } p \times p \text{ matrices}\}$. Given that $\bar{\mathbf{H}}_k$ is forced to be symmetric, one useful form for \mathbf{f}_k when p is not too large is to take \mathbf{f}_k such that $\bar{\mathbf{H}}_k = (\bar{\mathbf{H}}_k \bar{\mathbf{H}}_k + \delta_k \mathbf{I}_p)^{1/2}$, 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.

Let us now present the basic per-iteration Jacobian estimate $\hat{\mathbf{H}}_k$, as given in Spall (2000). 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 $\Delta_k \equiv [\Delta_{k1}, \Delta_{k2}, \dots, \Delta_{kp}]^T$ be a user-generated mean-zero random vector with finite inverse moments; further conditions on c_k , Δ_k , and other relevant quantities are given in Spall (2000). These conditions are close to those of basic SPSA in Spall (1992) (e.g., Δ_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). Conditions for the gain sequences are given in Spall (2000).

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

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

where $\delta \mathbf{G}_k = \mathbf{G}_k^{(1)}(\hat{\boldsymbol{\theta}}_k + c_k \Delta_k) - \mathbf{G}_k^{(1)}(\hat{\boldsymbol{\theta}}_k - c_k \Delta_k)$, and,

depending on the setting, the function $\mathbf{G}_k^{(1)}$ may or may not be the same as the function \mathbf{G}_k introduced in (2.1a). In particular, when forming a simultaneous perturbation (or even finite difference) estimate for $\mathbf{g}(\boldsymbol{\theta})$ based on values of the loss function $L(\boldsymbol{\theta})$, there are advantages to using a *one-sided* gradient approximation in order to reduce the total number of function evaluations (vs. the standard two-sided form that would typically be used to construct \mathbf{G}_k). This is referred to as the 2SPSA (2nd-order SPSA) setting in Spall (2000). In the root-finding case, it is assumed that direct unbiased measurements of $\mathbf{g}(\boldsymbol{\theta})$ are available (e.g., Spall, 2003, Chap. 5), implying that $\mathbf{G}_k^{(1)} = \mathbf{G}_k$.

III. CHARACTERIZATION OF ERROR IN JACOBIAN ESTIMATE AND CALCULATION OF FEEDBACK TERM

The feedback method below rests on an error analysis for the elements of the estimate $\hat{\mathbf{H}}_k$. We present a summary here in support of the results to follow; greater detail is in Spall (2006). Subsection III.A considers the case where $\mathbf{G}_k^{(1)}$ is formed from possibly noisy values of L ; Subsection III.B considers the case where $\mathbf{G}_k^{(1)}$ is formed from possibly noisy values of \mathbf{g} . Subsection III.C presents the feedback term. The probabilistic big- O terms appearing below are to be interpreted in the a.s. sense (e.g., $O(c_k^2)$ implies a function bounded that is a.s. bounded when divided by c_k^2 , $c_k \rightarrow 0$); all associated equalities hold a.s.

A. Error for Estimates Based on Measurements of L

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

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

with $\tilde{\Delta}_k = [\tilde{\Delta}_{k1}, \tilde{\Delta}_{k2}, \dots, \tilde{\Delta}_{kp}]^T$ generated in the same statistical manner as Δ_k , but independently of Δ_k (in particular, choosing $\tilde{\Delta}_{ki}$ as independent Bernoulli ± 1 random variables is a valid—but not necessary—choice), and with \tilde{c}_k satisfying conditions similar to c_k (Spall, 2000).

Let us define

$$\mathbf{D}_k = \Delta_k [\Delta_{k1}^{-1}, \Delta_{k2}^{-1}, \dots, \Delta_{kp}^{-1}] - \mathbf{I}_p,$$

together with a corresponding matrix $\tilde{\mathbf{D}}_k$ based on replacing all Δ_{ki} in \mathbf{D}_k with the corresponding Δ_{ki} (\mathbf{I}_p is the $p \times p$ identity matrix). From (2.2) and (3.1), the term dependent on the noises ε_k is $O(\tilde{c}_k^{-1} c_k^{-1})$, implying

$$\hat{\mathbf{H}}_k = \mathbf{H}(\hat{\boldsymbol{\theta}}_k) + \Psi_k^{(L)}(\mathbf{H}(\hat{\boldsymbol{\theta}}_k)) + O(\tilde{c}_k) + O(\tilde{c}_k^{-1} c_k^{-1}) + \frac{O(\tilde{c}_k^3)}{c_k}, \quad (3.2)$$

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

$$\Psi_k^{(L)}(\mathbf{H}) = \frac{1}{2} \left[\tilde{\mathbf{D}}_k \mathbf{H} \mathbf{D}_k + \tilde{\mathbf{D}}_k \mathbf{H} + \mathbf{H} \mathbf{D}_k \right] + \frac{1}{2} \left[\tilde{\mathbf{D}}_k \mathbf{H} \mathbf{D}_k + \tilde{\mathbf{D}}_k \mathbf{H} + \mathbf{H} \mathbf{D}_k \right]^T. \quad (3.3)$$

(The superscript L in $\Psi_k^{(L)}$ represents the dependence of this form on L measurements for creating the \mathbf{H} estimate, to be contrasted with $\Psi_k^{(g)}$ in the next subsection, which is dependent on \mathbf{g} measurements.) Note that the $O(\tilde{c}_k)$ and $O(\tilde{c}_k^3)/c_k$ terms in (3.2) are both due to third-order effects in L (the $O(\tilde{c}_k)$ term is a mean-zero term while the $O(\tilde{c}_k^3)/c_k$ term is a non-zero bias effect).

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

We now consider the case where direct (but possibly noisy) values of \mathbf{g} are available. Hence, direct measurements $\mathbf{Y}_k(\boldsymbol{\theta}) = \mathbf{g}(\boldsymbol{\theta}) + \mathbf{e}_k(\boldsymbol{\theta})$ are used for \mathbf{G}_k in (2.1a) and for $\mathbf{G}_k^{(1)}$ in $\delta \mathbf{G}_k$ appearing in (2.2), where \mathbf{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 Subsection III.A as a consequence of having the direct measurements of \mathbf{g} . As in Subsection III.A, it is sufficient to work with the first line of (2.2) in characterizing the error for the second line (relevant for the Hessian estimation here). The noise contribution in this case is $O(c_k^{-1})$. Hence, based on a Taylor expansion of the two gradient expressions entering $\delta \mathbf{G}_k$, we find

$$\hat{\mathbf{H}}_k = \mathbf{H}(\hat{\boldsymbol{\theta}}_k) + \Psi_k^{(g)}(\mathbf{H}(\hat{\boldsymbol{\theta}}_k)) + O(c_k^2) + O(c_k^{-1}), \quad (3.4)$$

where from (2.2) and (3.10)

$$\Psi_k^{(g)}(\mathbf{H}) \equiv \begin{cases} \mathbf{H} \mathbf{D}_k & \text{for Jacobian or} \\ \frac{1}{2} \mathbf{H} \mathbf{D}_k + \frac{1}{2} \mathbf{D}_k^T \mathbf{H} & \text{for Hessian.} \end{cases} \quad (3.5)$$

C. Feedback-Based Estimate of \mathbf{H} Matrix

Using the analysis in Subsections III.A and III.B, let us consider the form for $\hat{\Psi}_k \neq \mathbf{0}$ through the use of feedback, as discussed in this subsection. If \mathbf{H} were known, setting $\hat{\Psi}_k$ equal to $\Psi_k^{(L)}(\mathbf{H}(\hat{\boldsymbol{\theta}}_k))$ would leave only the unavoidable errors due to the noise and the bias at each iteration, where $\Psi_k^{(L)}$ represents either $\Psi_k^{(L)}$ or $\Psi_k^{(g)}$, as appropriate (expressions (3.3) and (3.5), respectively). Unfortunately, of course, this relatively simple modification cannot be implemented because we do not know \mathbf{H} !

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

$$\hat{\Psi}_k \equiv \begin{cases} \Psi_k^{(L)}(\bar{\mathbf{H}}_{k-1}) & \text{when } L \text{ measurements used,} \\ \Psi_k^{(g)}(\bar{\mathbf{H}}_{k-1}) & \text{when } \mathbf{g} \text{ measurements used.} \end{cases}$$

IV. OPTIMAL WEIGHTING WITH NOISY MEASUREMENTS

The results in this short section are available in more complete form in Spall (2006); this summary is given here for the sake of providing essential information needed to properly interpret the convergence results in the remainder of this paper. As discussed above, the second way in which the accuracy of the \mathbf{H} estimate may be improved is through the optimal selection of weights w_k in (2.1b). We consider separately below the cases where $\mathbf{G}_k^{(1)}$ is formed from noisy values of L and noisy values of \mathbf{g} . The optimal weights w_k derived here assume that the noise contributions are nontrivial in the sense that $\text{var}[\varepsilon_k^{(+)} + \tilde{\varepsilon}_k^{(+)} - \varepsilon_k^{(-)} - \tilde{\varepsilon}_k^{(-)}] \geq \rho$ for all k with L measurements and $\text{cov}[\mathbf{e}_k^{(+)} - \mathbf{e}_k^{(-)}] \geq \rho \mathbf{I}_p$ for all k with \mathbf{g} measurements, where $\rho > 0$. (Section VII provides detailed treatment for the noise-free cases: $\varepsilon_k^{(\pm)} = \tilde{\varepsilon}_k^{(\pm)} = 0$ and $\mathbf{e}_k^{(\pm)} = \mathbf{0}$.)

Consider first the case of L measurements. Given that the noise terms $\varepsilon_k^{(\pm)}$ and $\tilde{\varepsilon}_k^{(\pm)}$ are uncorrelated across iterations (e.g., $\text{cov}(\varepsilon_j^{(+)}, \tilde{\varepsilon}_k^{(+)}) = 0$ and $\text{cov}(\varepsilon_j^{(+)}, \varepsilon_k^{(+)}) = 0$ for $j \neq k$), we may find the weights w_k that minimize the variance of the elements in $\bar{\mathbf{H}}_n$. It is fairly straightforward to find the solution via the method of Lagrange multipliers:

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

Now consider the case of direct noisy values of \mathbf{g} . Following the logic above, the optimal w_k is:

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

V. CONVERGENCE THEORY WITH NOISY MEASUREMENTS

Much of the convergence and efficiency analysis in Spall (2000) will hold verbatim in analyzing the enhanced form here. In particular, under conditions for Theorems 1a and 1b in Spall (2000), it is known that $\hat{\boldsymbol{\theta}}_k \rightarrow \boldsymbol{\theta}^*$ a.s. in the setting

of either L measurements or \mathbf{g} measurements. On the other hand, because the recursion (2.1b) differs from Spall (2000) due to the weighting and feedback, it is necessary to make some changes to the arguments showing convergence of $\bar{\mathbf{H}}_k$. Let $\mathfrak{I}_k = \{\hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_1, \dots, \hat{\boldsymbol{\theta}}_k; \hat{\mathbf{H}}_0, \hat{\mathbf{H}}_1, \dots, \hat{\mathbf{H}}_{k-1}\}$ ($\mathfrak{I}_0 = \{\hat{\boldsymbol{\theta}}_0\}$).

This section gives analogues to Theorems 2a and 2b in Spall (2000), showing convergence of $\bar{\mathbf{H}}_k$ with L measurements and with \mathbf{g} measurements, respectively. 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 C.1'' to C.1'''' and C.8 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 $0 < \gamma \leq 1/4$.

C.8''. For some $\rho > 0$ and all k, ℓ, m , the following hold

a.s.: $E\left[y(\hat{\theta}_k \pm c_k \Delta_k + c_k \tilde{\Delta}_k)^2 / (\Delta_{k\ell} \tilde{\Delta}_{km})^2 \mid \mathfrak{F}_k\right] \leq \rho$,

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

$E\left[(\tilde{\varepsilon}_k^{(\pm)} - \varepsilon_k^{(\pm)})^2 / (\Delta_{k\ell} \tilde{\Delta}_{km})^2 \mid \mathfrak{F}_k\right] \leq \rho$, and

$E\left(\left\|\bar{\mathbf{H}}_k\right\|^2 \mid \mathfrak{F}_k\right) \leq \rho$. (Note that the first two bounds are similar to the bounds in C.2 in Spall 2000, but are neither necessary nor sufficient for C.2.)

Theorem 1 (2SPSA setting). Suppose only noisy measurements of L are used to form \mathbf{G}_k and $\mathbf{G}_k^{(1)}$ (see (3.1)).

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 (2000).

Then, $\bar{\mathbf{H}}_k \rightarrow \mathbf{H}(\theta^*)$ a.s. as $k \rightarrow \infty$.

Proof. First, note that the conditions subsume those of Theorem 1a in Spall (2000) (C.0–C.7); hence we have a.s. convergence of $\hat{\theta}_k$ to θ^* . We first use a convergence result in Chow and Teicher (1988, p. 249) to establish the convergence for a particular sum of martingale differences. We then use the martingale difference conclusion to establish the result to be proved on convergence of $\bar{\mathbf{H}}_k$.

Let us first show that:

$$\sum_{k=0}^n \frac{\tilde{c}_k^2 c_k^2 \{\hat{\mathbf{H}}_k - \hat{\Psi}_k - E(\hat{\mathbf{H}}_k \mid \mathfrak{F}_k)\}}{\sum_{i=0}^n \tilde{c}_i^2 c_i^2} \rightarrow \mathbf{0} \text{ a.s.} \quad (5.1)$$

Let \hat{h}_k and $\hat{\psi}_k$ represent corresponding (arbitrary) elements of $\hat{\mathbf{H}}_k$ and $\hat{\Psi}_k$, respectively. Note that $\sum_{k=0}^n \tilde{c}_k^2 c_k^2 [(\hat{h}_k - \hat{\psi}_k - E(\hat{h}_k \mid \mathfrak{F}_k))]$ is a martingale with bounded second moments for all n . Hence, from a martingale convergence result in Chow and Teicher (1988, p. 249), (5.1) will be true if

$$\sum_{k=0}^n \frac{\tilde{c}_k^4 c_k^4 E\left\{[(\hat{h}_k - \hat{\psi}_k - E(\hat{h}_k \mid \mathfrak{F}_k))]^2 \mid \mathfrak{F}_k\right\}}{\left(\sum_{i=0}^n \tilde{c}_i^2 c_i^2\right)^2} < \infty \text{ a.s.} \quad (5.2)$$

Because $E(\hat{\psi}_k \mid \mathfrak{F}_k) = 0$, the conditional expectation appearing in the numerator of (5.2) satisfies

$$\begin{aligned} E\left\{[(\hat{h}_k - \hat{\psi}_k - E(\hat{h}_k \mid \mathfrak{F}_k))]^2 \mid \mathfrak{F}_k\right\} &\leq E\{(\hat{h}_k - \hat{\psi}_k)^2 \mid \mathfrak{F}_k\} \\ &\leq 2\left[E(\hat{h}_k^2 \mid \mathfrak{F}_k) + E(\hat{\psi}_k^2 \mid \mathfrak{F}_k)\right] \\ &= 2E(\hat{h}_k^2 \mid \mathfrak{F}_k) + O(1) \\ &= O(\tilde{c}_k^{-2} c_k^{-2}) \text{ a.s.}, \end{aligned}$$

where the two equalities follow by C.8''. Hence, from C.1''''', the left-hand side of (5.2) is given by

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

The above expression indicates that (5.2) is satisfied, in turn indicating that (5.1) holds.

Let us analyze $E(\hat{\mathbf{H}}_k \mid \mathfrak{F}_k)$ as appears in (5.1) to show convergence of $\bar{\mathbf{H}}_k$. It is sufficient to work with the Jacobian form in the first line of (2.2). Expanding the right-hand side of (3.1), the bias in the ij th component of $\hat{\mathbf{H}}_k$ is

$$E\left(\frac{1/6 \tilde{c}_k^3 [L'''(\bar{\theta}_k^{(+)}) - L'''(\bar{\theta}_k^{(-)})][\tilde{\Delta}_k \otimes \tilde{\Delta}_k \otimes \tilde{\Delta}_k]}{\tilde{c}_k c_k \tilde{\Delta}_{ki} \Delta_{kj}} \mid \mathfrak{F}_k\right).$$

Using C.1'''' and C.3', $\|L'''(\bar{\theta}_k^{(+)}) - L'''(\bar{\theta}_k^{(-)})\| = O(c_k)$ 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

$$E(\hat{\mathbf{H}}_k \mid \mathfrak{F}_k) = \mathbf{H}(\hat{\theta}_k) + O(c_k^2) \text{ a.s.} \quad (5.3)$$

From (5.3), the continuity of \mathbf{H} at all $\hat{\theta}_k$, and the a.s. convergence of $\hat{\theta}_k$ to θ^* ,

$$\begin{aligned} \sum_{k=0}^n \frac{\tilde{c}_k^2 c_k^2 E(\hat{\mathbf{H}}_k \mid \mathfrak{F}_k)}{\sum_{i=0}^n \tilde{c}_i^2 c_i^2} &= \sum_{k=0}^n \frac{\tilde{c}_k^2 c_k^2 [\mathbf{H}(\hat{\theta}_k) + O(c_k^2)]}{\sum_{i=0}^n \tilde{c}_i^2 c_i^2} \\ &= \sum_{k=0}^n \frac{\tilde{c}_k^2 c_k^2 [\mathbf{H}(\theta^*) + o(1)]}{\sum_{i=0}^n \tilde{c}_i^2 c_i^2} \\ &\rightarrow \mathbf{H}(\theta^*) \text{ a.s.} \end{aligned} \quad (5.4)$$

as $n \rightarrow \infty$, where the result follows by the fact that the denominator $\sum_{i=0}^n \tilde{c}_i^2 c_i^2 \rightarrow \infty$ (from C.1'''''). Given that $\bar{\mathbf{H}}_n = \sum_{k=0}^n \tilde{c}_k^2 c_k^2 (\hat{\mathbf{H}}_k - \hat{\Psi}_k) / \sum_{i=0}^n \tilde{c}_i^2 c_i^2$, (5.1) and (5.4) together yield the result to be proved. *Q.E.D.*

We now show convergence of $\bar{\mathbf{H}}_k$ in the root-finding case with only \mathbf{g} measurements; this result is an analogue of Theorem 2b in Spall, 2000. Following the pattern above, C.1'''' and C.8' in Spall (2000) 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.:

$$E\left[\left\|\mathbf{g}(\hat{\theta}_k \pm c_k \Delta_k) / \Delta_{k\ell}\right\|^2 \mid \mathfrak{F}_k\right] \leq \rho, E\left[\left\|(e_k^{(+)} - e_k^{(-)}) / \Delta_{k\ell}\right\|^2 \mid \mathfrak{F}_k\right]$$

$$\leq \rho, E\left[(e_k^{(+)} - e_k^{(-)}) / \Delta_{k\ell} \mid \mathfrak{F}_k\right] = \mathbf{0}, \text{ and } E\left(\left\|\bar{\mathbf{H}}_k\right\|^2 \mid \mathfrak{F}_k\right) \leq \rho,$$

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

Theorem 2 (root-finding setting). Suppose noisy measurements of \mathbf{g} are used to form \mathbf{G}_k . 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 (2000). Then, $\bar{\mathbf{H}}_k \rightarrow \mathbf{H}(\theta^*)$ a.s. as $k \rightarrow \infty$.

Proof. First, note that the conditions subsume those of Theorem 1b in Spall (2000) (C.0', C.1', and C.2', and C.3–C.7); hence we have a.s. convergence of $\hat{\theta}_k$ to θ^* . Following the steps in the proof of Theorem 1, let us first show that:

$$\sum_{k=0}^n \frac{c_k^2 \{ \hat{\mathbf{H}}_k - \hat{\Psi}_k - E(\hat{\mathbf{H}}_k | \mathfrak{I}_k) \}}{\sum_{i=0}^n c_i^2} \rightarrow \mathbf{0} \text{ a.s.} \quad (5.5)$$

From the martingale convergence result in Chow and Teicher (1988, p. 249), (5.5) will be true if

$$\sum_{k=0}^n \frac{c_k^4 E \{ [(\hat{h}_k - \hat{\psi}_k - E(\hat{h}_k | \mathfrak{I}_k))]^2 | \mathfrak{I}_k \}}{(\sum_{i=0}^n c_i^2)^2} < \infty \text{ a.s.} \quad (5.6)$$

(\hat{h}_k and $\hat{\psi}_k$ are as defined below (5.1)). Hence, by (C.8'''),

$$\begin{aligned} E \{ [(\hat{h}_k - \hat{\psi}_k - E(\hat{h}_k | \mathfrak{I}_k))]^2 | \mathfrak{I}_k \} \\ \leq E \{ (\hat{h}_k - \hat{\psi}_k)^2 | \mathfrak{I}_k \} = O(c_k^{-2}) \text{ a.s.} \end{aligned}$$

Hence, from C.1''''', the left-hand side of (5.6) is given by

$$\begin{aligned} \sum_{k=0}^n \frac{c_k^4 O(c_k^{-2})}{(\sum_{i=0}^n c_i^2)^2} &= O \left(\frac{\int_1^n x^{-2\gamma} dx}{\left(\int_1^n x^{-2\gamma} dx \right)^2} \right) \\ &= \begin{cases} O(n^{2\gamma-1}) \text{ a.s. if } 0 < \gamma < 1/2, \\ O(1/\log n) \text{ a.s. if } \gamma = 1/2. \end{cases} \end{aligned}$$

The above expression indicates that (5.6) is satisfied, in turn indicating that (5.5) holds. By the boundedness of the third derivative of \mathbf{g} (see C.3'), $E(\hat{\mathbf{H}}_k | \mathfrak{I}_k) = \mathbf{H}(\hat{\boldsymbol{\theta}}_k) + O(c_k^2)$ a.s. Then, analogous to (5.4),

$$\sum_{k=0}^n \frac{c_k^2 E(\hat{\mathbf{H}}_k | \mathfrak{I}_k)}{\sum_{i=0}^n c_i^2} = \sum_{k=0}^n \frac{c_k^2 [\mathbf{H}(\hat{\boldsymbol{\theta}}_k) + O(c_k^2)]}{\sum_{i=0}^n c_i^2} \rightarrow \mathbf{H}(\boldsymbol{\theta}^*) \text{ a.s.} \quad (5.7)$$

as $n \rightarrow \infty$, where the result follows by the fact that the denominator $\sum_{i=0}^n c_i^2 \rightarrow \infty$ (from C.1'''''). Thus, (5.5) and (5.7) together yield the result to be proved. *Q.E.D.*

Spall (2000) includes an asymptotic distribution theory for $\hat{\boldsymbol{\theta}}_k$, finding 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, 2003, 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. The improved Jacobian estimation above would not alter these asymptotic accuracy results, as the Spall (2000) 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.

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, 2000) 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 $\{\Delta_k\}$ and $\{\tilde{\Delta}_k\}$ are each identically distributed across k . Note that feedback (Subsection 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 $\{\Delta_k\}$ and $\{\tilde{\Delta}_k\}$ implies that the variance of an individual element in the summands $\hat{\mathbf{H}}_k$ is asymptotic to $K\tilde{c}_k^{-2}c_k^{-2}$ for large k and some constant $K > 0$ (see Subsection III.A). Hence, under the conditions on c_k and \tilde{c}_k given in Theorem 1, the variance of an individual element in a simple average form for $\bar{\mathbf{H}}_n$ is given by

$$\begin{aligned} \frac{1}{n^2} \sum_{k=0}^n O(\tilde{c}_k^{-2}c_k^{-2}) &\sim \frac{1}{n^2} K \int_0^n x^{4\gamma} dx \\ &= \begin{cases} K(4\gamma+1)^{-1} n^{4\gamma-1} & \text{if } 0 < \gamma < 1/4, \\ K/2 & \text{if } \gamma = 1/4, \end{cases} \quad (6.1) \end{aligned}$$

where “ \sim ” denotes “asymptotic to” (note that for $\gamma = 1/4$, the variance does *not* go to zero, consistent with the lack of convergence associated with Theorem 2a in Spall, 2000). For the weighted average case (Section IV), the corresponding variance of an individual element in $\bar{\mathbf{H}}_n$ is

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

The root-finding setting follows the line of reasoning above. Here, the variance of an individual element in the summands $\hat{\mathbf{H}}_k$ is asymptotic to $K'c_k^{-2}$ for large k and some constant $K' > 0$ (see Subsection III.B). Hence, under the conditions on c_k in Theorem 2, the variance of an individual element in a simple average form for $\bar{\mathbf{H}}_n$ is given by

$$\begin{aligned} \frac{1}{n^2} \sum_{k=0}^n O(c_k^{-2}) &\sim \frac{1}{n^2} K' \int_0^n x^{2\gamma} dx \\ &= \begin{cases} K'(2\gamma+1)^{-1} n^{2\gamma-1} & \text{if } 0 < \gamma < 1/2, \\ K/2 & \text{if } \gamma = 1/2. \end{cases} \quad (6.3) \end{aligned}$$

For the weighted average case (Section IV), the corresponding variance of an individual element in $\bar{\mathbf{H}}_n$ is

$$\sum_{k=0}^n \frac{c_k^4 O(c_k^{-2})}{\left(\sum_{i=0}^n c_i^2\right)^2} \sim \frac{K' \int_0^n x^{-2\gamma} dx}{\left(\int_0^n x^{-2\gamma} dx\right)^2} \\ = \begin{cases} K'(1-2\gamma)n^{2\gamma-1} + o(n^{2\gamma-1}) \text{ a.s. if } 0 < \gamma < 1/2, \\ K'/\log n + o(1/\log n) \text{ a.s. if } \gamma = 1/2. \end{cases} \quad (6.4)$$

Given the above, Table 1 shows the asymptotic ratio of variances for several popular values of γ . 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.

TABLE 1. Asymptotic ratio of variances of Jacobian estimate: Unweighted to weighted. (Note: $\gamma = 0.101$ Is a popular practical choice in SPSA and 2SPSA settings and $\gamma = 1/6$ is asymptotically optimal for SPSA and 2SPSA [e.g., Spall, 2000, and Spall, 2003, Sect. 7.5]; N/A = not applicable [invalid γ for unweighted and weighted settings].)

γ	Ratio in 2SPSA setting	Ratio in root-finding 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	∞

VII. RATE OF CONVERGENCE OF JACOBIAN/HESSIAN ESTIMATES WITH NOISE-FREE MEASUREMENTS

While most of the applications for SA are in cases of minimization and/or root-finding in the presence of noisy L or \mathbf{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 (2001); some theory on convergence rates in the noise-free case is given in Gerencsér and Vago (2001). 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 \mathbf{H} recursions in (2.1a, b) 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).

In the case of noise-free measurements of L , for example, decaying gains satisfying different conditions than the standard SPSA conditions are given in Maryak and Chin (2001) to ensure global convergence with a possibly multimodal loss function; further constant gains $a_k = a$ are considered in Gerencsér and Vago (2001) when the loss function is quadratic. In the noise-free case of direct measurements of \mathbf{g} , constant gains $a_k = a$ may be used to ensure convergence of what is effectively a quasi-Newton-type algorithm.

We present below a rate of convergence result for the Jacobian estimates in the noise-free case. Theorem 3 considers the settings of L measurements under the restriction of quadratic loss functions; for that reason, this theorem is best interpreted as a “local” theorem pertaining to

losses that are at least approximately quadratic in the vicinity of $\boldsymbol{\theta}^*$. For convenience, let $\boldsymbol{\Lambda}_k = \tilde{\mathbf{H}}_k - \mathbf{H}(\boldsymbol{\theta}^*)$; we write $E(\boldsymbol{\Lambda}_k^T \boldsymbol{\Lambda}_k)$, but note $E(\boldsymbol{\Lambda}_k^T \boldsymbol{\Lambda}_k) = E(\boldsymbol{\Lambda}_k \boldsymbol{\Lambda}_k^T)$. As a consequence of the quadratic loss function, there are no restrictions on the c_k values in Theorem 3 (i.e., unlike Theorems 1 and 2 above, $\tilde{\mathbf{H}}_k$ has no $O(c_k^2)$ or other bias).

Theorem 3 (2SPSA setting). Suppose L is a quadratic function and only noise-free measurements of L are used to form \mathbf{G}_k and $\mathbf{G}_k^{(1)}$ (see (3.1)). Suppose $0 < w_0 \leq 1$ and $w_k = w/k^\delta$, $k = 1, 2, \dots$, where $1/2 < \delta < 1$ and $0 < w \leq 1$. Suppose that C.8'' in Section 5 and C.2 and C.9 from Spall (2000) hold (see the Appendix here; note that $\nu(\cdot) = L(\cdot)$ in the setting here) and that $\boldsymbol{\Delta}_k$ and $\tilde{\boldsymbol{\Delta}}_k$ are identically distributed at each k and across k . Further, suppose that $\mathbf{H}^* > \mathbf{0}$ and that \mathbf{f}_k in (2.1a) is such that $E\left(\left\|\tilde{\mathbf{H}}_k - \bar{\mathbf{H}}_k\right\|^2\right) = o\left(e^{-2wk^{1-\delta}/(1-\delta)}\right)$ and $\left\|\mathbf{f}_k(\mathbf{H}) - \mathbf{H}\right\|^2 / \left(1 + \|\mathbf{H}\|^2\right)$ is uniformly bounded over the set of symmetric \mathbf{H} in $\mathbb{R}^{p \times p}$. Then, $\text{trace}[E(\boldsymbol{\Lambda}_n^T \boldsymbol{\Lambda}_n)] = O\left(e^{-2wn^{1-\delta}/(1-\delta)}\right)$.

Proof. Available by request.

VIII. REFERENCES

- [1] Bhatnagar, S. (2005), “Adaptive Multivariate Three-Timescale Stochastic Approximation Algorithms for Simulation-Based Optimization,” *ACM Transactions on Modeling and Computer Simulation*, vol. 15, pp. 74–107.
- [2] Chow, Y. S. and Teicher, H. (1988), *Probability Theory: Independence, Interchangeability, and Martingales* (2nd ed.), Springer-Verlag, New York.
- [3] Fabian, V. (1971), “Stochastic Approximation,” in *Optimizing Methods in Statistics* (J. S. Rustagi, ed.), Academic Press, New York, pp. 439–470.
- [4] Gerencsér, L. and Vago, Z. (2001), “The Mathematics of Noise-Free SPSA,” *Proceedings of the IEEE Conference on Decision and Control*, 4–7 December 2001, Orlando, FL, pp. 4400–4405.
- [5] Kushner, H. J. and Yin, G. G. (2003), *Stochastic Approximation and Recursive Algorithms and Applications* (2nd ed.), Springer-Verlag, New York.
- [6] Macchi, O. and Eweda, E. (1983), “Second-Order Convergence Analysis of Stochastic Adaptive Linear Filtering,” *IEEE Transactions on Automatic Control*, vol. AC-28, pp. 76–85.
- [7] Maryak, J. L. and Chin, D. C. (2001), “Global Random Optimization by Simultaneous Perturbation Stochastic Approximation,” *Proceedings of the American Control Conference*, 25–27 June 2001, Arlington, VA, pp. 756–762.
- [8] Polyak, B. T. and Juditsky, A. B. (1992), “Acceleration of Stochastic Approximation by Averaging,” *SIAM Journal on Control and Optimization*, vol. 30, pp. 838–855.
- [9] Ruppert, D. (1985), “A Newton–Raphson Version of the Multivariate Robbins–Monro Procedure,” *Annals of Statistics*, vol. 13, pp. 236–245.
- [10] Spall, J. C. (1992), “Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation,” *IEEE Transactions on Automatic Control*, vol. 37, pp. 332–341.
- [11] Spall, J. C. (2000), “Adaptive Stochastic Approximation by the Simultaneous Perturbation Method,” *IEEE Transactions on Automatic Control*, vol. 45, pp. 1839–1853.
- [12] Spall, J. C. (2003), *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, Wiley, Hoboken, NJ.
- [13] Spall, J. C. (2006), “Feedback and Weighting Mechanisms for Improving Jacobian (Hessian) Estimates in the Adaptive Simultaneous Perturbation Algorithm,” *Proceedings of the American Control Conference*, Minneapolis, MN, 14–16 June 2006, pp. 3086–3091.
- [14] Yin, G. and Zhu, Y. (1992), “Averaging Procedures in Adaptive Filtering: An Efficient Approach,” *IEEE Transactions on Automatic Control*, vol. 37, pp. 466–475.
- [15] Wei, C. Z. (1987), “Multivariate Adaptive Stochastic Approximation,” *Annals of Statistics*, vol. 15, pp. 1115–1130.