INTERNATIONAL JOURNAL OF ADAPTIVE CONTROL AND SIGNAL PROCESSING
Int. J. Adapt. Control Signal Process. 2002; **16**:397–409 (DOI: 10.1002/acs.715)

A modified second-order SPSA optimization algorithm for finite samples

Xun Zhu*,† and James C. Spall

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

SUMMARY

We propose a modification to the simultaneous perturbation stochastic approximation (SPSA) methods based on the comparisons made between the first- and second-order SPSA (1SPSA and 2SPSA) algorithms from the perspective of loss function Hessian. At finite iterations, the accuracy of the algorithm depends on the matrix conditioning of the loss function Hessian. The error of 2SPSA algorithm for a loss function with an ill-conditioned Hessian is greater than the one with a well-conditioned Hessian. On the other hand, the 1SPSA algorithm is less sensitive to the matrix conditioning of loss function Hessians. The modified 2SPSA (M2SPSA) eliminates the error amplification caused by the inversion of an ill-conditioned Hessian. This leads to significant improvements in its algorithm efficiency in problems with an ill-conditioned Hessian matrix. Asymptotically, the efficiency analysis shows that M2SPSA is also superior to 2SPSA in a large parameter domain. It is shown that the ratio of the mean square errors for M2SPSA to 2SPSA is always less than one except for a perfectly conditioned Hessian or for an asymptotically optimal setting of the gain sequence. Copyright © 2002 John Wiley & Sons, Ltd.

1. INTRODUCTION

The recently developed simultaneous perturbation stochastic approximation (SPSA) method has found many applications in areas such as physical parameter estimation and simulation-based optimization. The novelty of the SPSA is the underlying derivative approximation that requires only two (for the gradient) or four (for the Hessian matrix) evaluations of the loss function regardless of the dimension of the optimization problem. There exist two basic SPSA algorithms that are based on the 'simultaneous perturbation' (SP) concept and that use only (noisy) loss function measurements. The first-order SPSA (1SPSA) is related to the

*Correspondence to: Xun Zhu, Applied Physics Laboratory, The John Hopkins University, 11100 Johns Hopkins Road, Laurel, MD 20723-6099, USA
†E-mail: xun.zhu@jhuapl.edu

Contract/grant sponsor: NSF
Contract/grant number: ATM-0091514
Contract/grant sponsor: NASA
Contract/grant number: NAS5-97179
Contract/grant sponsor: JHU/APL IRAD Program and U.S. Navy; Contract/grant number: N00024-98-D-8124

Kiefer–Wolfowitz (K–W) stochastic approximation (SA) method [1] whereas the second-order SPSA (2SPSA) is a stochastic analogue of the deterministic Newton–Raphson algorithm [2]. There have been several studies that compare the efficiency of 1SPSA with other stochastic approximation (SA) methods (e.g. References [1, 3, 4]). It is generally accepted that 1SPSA is superior to other first-order SA methods (such as the standard K–W method) due to its efficient estimator for the loss function gradient.

Spall [2] shows that a 'standard' implementation of 2SPSA achieves a nearly optimal asymptotic error, with the asymptotic root-mean-square error being no more than twice the optimal (but unachievable) error from an infeasible gain sequence depending on the third derivatives of the loss function. This appealing result for 2SPSA is achieved with a trivial gain sequence ($\bar{a}_k = 1/(k+1)$ in the notation below), which effectively eliminates the nettlesome issue of selecting a 'good' gain sequence. Because this result is asymptotic, however, performance in finite samples may sometimes be improved using other considerations.

Part of the purpose of this paper is to provide a comparison between 1SPSA and 2SPSA from the perspective of the conditioning of the loss function Hessian matrix. To achieve the objectivity of the comparison we also suggest a new mapping for implementing 2SPSA that eliminates the non-positive definiteness while preserving key spectral properties of the estimated Hessian. While the focus of this paper is finite-sample analysis, we are necessarily limited by the theory available for SA algorithms, almost all of which is asymptotic. For that reason, the discussion and rationale here will be a blend of static (finite-sample) results from matrix theory, asymptotic theory, and numerical analysis. The numerical examples illustrating the empirical results at finite iterations will be carefully chosen to represent a wide range of matrix conditioning for the loss function Hessians.

2. MATRIX CONDITIONING AND ITS RELATION TO 2SPSA

The SA algorithms are the general recursions for the estimate ($\hat{\theta}_k$) of a solution (θ^*) having dimension p. The core recursions for the SPSA algorithms are

1SPSA [1]:

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{g}_k(\hat{\theta}_k), \quad k = 0, 1, 2, \ldots \tag{1}$$

2SPSA [2]:

$$\hat{\theta}_{k+1} = \hat{\theta}_k - \bar{a}_k \bar{\bar{H}}_k^{-1} \hat{g}_k(\hat{\theta}_k), \quad \bar{\bar{H}}_k = f_k(\bar{H}_k) \tag{2a}$$

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

where a_k and $\bar{a}_k$ are the scalar gain series that satisfy certain SA conditions, $\hat{g}_k$ is the SP estimate of the loss function gradient that depends on the gain sequence c_k (representing a difference interval of the perturbations), $\hat{H}_k$ is the SP estimate of the Hessian matrix, and f_k maps the usual non-positive-definite $\bar{H}_k$ to a positive-definite $p \times p$ matrix. Let Δ_k be a user-generated mean-zero random vector of dimension p with its components being independent random variables. The ith element of the loss function gradient is given by [1]

$$(\hat{g}_k)_i = (2c_k\Delta_{ki})^{-1}[y(\hat{\theta}_k + c_k\Delta_k) - y(\hat{\theta}_k - c_k\Delta_k)], \quad i = 1, 2, \ldots, p \tag{3}$$

where Δ_{ki} is the ith component of the Δ_k, vector and $y(\theta)$ is the measurements of the loss function: $y(\theta) = L(\theta) + \text{noise}$, where θ is the parameter that has the true value of θ^*. It is noted that the 2SPSA form is a special case of the general adaptive SP method of Spall [2]. The general method can also be used in root-finding problems where $\bar{\boldsymbol{H}}_k$ represents an estimate of the associated Jacobian matrix.

The true Hessian matrix of the loss function $H(\theta)$ has its ijth element defined as $H_{ij} = \partial^2 L/\partial\theta_i\partial\theta_j$ and its value at the solution $H(\theta^*)$ denoted by H^*. The ijth element of the per-iteration estimate of H is given by [2]

$$(\hat{\boldsymbol{H}}_k)_{ij} = (4c_k\tilde{c}_k)^{-1}[(\Delta_{ki}\tilde{\Delta}_{kj})^{-1} + (\Delta_{kj}\tilde{\Delta}_{ki})^{-1}]\,[y(\hat{\theta}_k + c_k\Delta_k + \tilde{c}_k\tilde{\Delta}_k) - y(\hat{\theta}_k + c_k\Delta_k) - y(\hat{\theta}_k - c_k\Delta_k + \tilde{c}_k\tilde{\Delta}_k) + y(\hat{\theta}_k - c_k\Delta_k)], \quad i, j = 1, 2, \ldots, p \tag{4}$$

with the gain sequence $\tilde{c}_k$ satisfying conditions similar to c_k and with $\tilde{\Delta}_k = (\tilde{\Delta}_{k1}, \tilde{\Delta}_{k2}, \ldots, \tilde{\Delta}_{kp})^{\mathrm{T}}$ generated in the same statistical manner as Δ_k. It is noted that $\hat{\boldsymbol{H}}_k$ defined by (4) is a symmetric Hessian estimate that is convenient in an optimization application and is a crucial requirement for the new mapping f_k proposed in the following section. Readers are referred to Spall [2] for more detailed definitions and discussions on implementation aspects, including some possible forms for the mapping f_k.

2.1. A new form of mapping f_k for 2SPSA

One crucial aspect of implementing 2SPSA is to define the mapping f_k, from $\bar{\boldsymbol{H}}_k$ to $\bar{\bar{\boldsymbol{H}}}_k$ since the former is often non-positive definite in practice. It is noted that there are no simple and universal conditions that guarantee a matrix to be positively definite. The existence of a minimum(s) for a loss function based on the problem's physical nature guarantees that its Hessian should be positively definite. We suggest the following approach that eliminates the non-positive definiteness while preserving key spectral properties of $\bar{\boldsymbol{H}}_k$. This approach is motivated by finite-sample concerns, as we discuss below. First, we compute the eigenvalues of $\bar{\boldsymbol{H}}_k$ and sort them into descending order:

$$\Lambda_k \equiv \text{diag}[\lambda_1, \lambda_2, \ldots, \lambda_{q-1}, \lambda_q, \lambda_{q+1}, \ldots, \lambda_p] \tag{5}$$

where $\lambda_q > 0$ and $\lambda_{q+1} \leqslant 0$. For the sake of simplicity, we have omitted the index k for the individual eigenvalue λ_i that is a function of k. Next, we assume that the negative eigenvalues will not lead to a physically meaningful solution. They are either caused by errors in $\bar{\boldsymbol{H}}_k$ or are due to the fact that the iteration has not reached the neighborhood of θ^* where the loss function is locally quadratic. Therefore, we replace them together with the smallest positive eigenvalue with a descending series of positive eigenvalues:

$$\hat{\lambda}_q = \varepsilon\lambda_{q-1}, \hat{\lambda}_{q+1} = \varepsilon\hat{\lambda}_q, \ldots, \hat{\lambda}_p = \varepsilon\hat{\lambda}_{p-1} \tag{6}$$

where the adjustable parameter $0 < \varepsilon < 1$ can be specified based on the existing positive eigenvalues

$$\varepsilon = (\lambda_{q-1}/\lambda_1)^{q-2} \tag{7}$$

The purpose of having the smallest positive eigenvalue (λ_q) redefined is to avoid its possible near-zero value that would make the mapped matrix near singular. We let $\hat{\boldsymbol{\Lambda}}_k$ be the diagonal matrix Λ_k, with eigenvalues $\lambda_q, \ldots, \lambda_p$ replaced by $\hat{\lambda}_q, \ldots, \hat{\lambda}_p$ defined according to (6).

Equations (6) and (7) indicate that the spectral character of the existing positive eigenvalues as measured by the ratio of its maximum-to-minimum eigenvalues, whether it is wide or narrowly spread, is extrapolated to the rest of the matrix spectrum. The specification of (7) bears an ad hoc feature that is common in all extrapolation techniques (e.g. Reference [5, p. 99]). Other forms of specifications such as $\varepsilon = (\lambda_{q-1}/\lambda_1)^{(q-2)/2}$ or $\varepsilon = 1$ would also effectively eliminate the non-positive definiteness. Because the separating point between the positive and negative eigenvalues q slowly increases from 1 to p, we find numerically that the specification based on (7) yields relatively a faster convergence in most castes. Since $\bar{\boldsymbol{H}}_k$ is symmetric, it is orthogonally similar to the real diagonal matrix of its real eigenvalues (e.g. References [6, p. 171])

$$\bar{\boldsymbol{H}}_k = P_k \Lambda_k P_k^{\mathrm{T}} \tag{8}$$

where the orthogonal matrix P_k consists of all the eigenvectors of $\bar{\boldsymbol{H}}_k$, which are usually derived together with the eigenvalues (e.g. Reference [5, p. 460]). Now, the mapping f_k can be expressed as

$$f_k(\bar{\boldsymbol{H}}_k) = P_k \hat{\Lambda}_k P_k^{\mathrm{T}} \tag{9}$$

Since it is $\bar{\bar{\boldsymbol{H}}}_k^{-1}$ that is used in the 2SPSA recursion (2a) mapping (9) with the available eigenvectors of $\bar{\boldsymbol{H}}_k$ also leads to an easy inversion of the estimated Hessian:

$$\bar{\bar{\boldsymbol{H}}}_k^{-1} = P_k \hat{\Lambda}_k^{-1} P_k^{\mathrm{T}} \tag{10}$$

The 2SPSA based on mapping (9) makes the procedure of eliminating the non-positive definiteness of $\bar{\boldsymbol{H}}_k$ a precise one. It is noted that the key parameters needed for the mapping (ε and λ_{q-1}) are internally determined by $\bar{\boldsymbol{H}}_k$ at each iteration. This is different from some other forms of f_k where a user-specified coefficient is needed.

According to the perturbation theorem for the eigenvalues of a symmetric matrix the differences in eigenvalues are bounded by eigenvalues of the perturbation matrix [6, p. 367]

$$\lambda_p(\Delta \bar{\boldsymbol{H}}_k) \leqslant \lambda_i - \lambda_i^* \leqslant \lambda_1(\Delta \bar{\boldsymbol{H}}_k) \quad \text{for all} \quad i = 1, 2, \ldots, p \tag{11}$$

where λ_i^* denotes the eigenvalues of H^*. Furthermore, $\lambda_p(\Delta \bar{\boldsymbol{H}}_k)$ and $\lambda_1(\Delta \bar{\boldsymbol{H}}_k)$ are the minimum and maximum eigenvalues of the kth perturbation matrix $\Delta \bar{\boldsymbol{H}}_k = \bar{\boldsymbol{H}}_k - H^*$, respectively. Equation (11) suggests that the perturbation matrix will have greater impact on the smaller eigenvalues in terms of their fractional changes as $\bar{\boldsymbol{H}}_k$ converges (almost surely) to H^* (see conditions in Reference [2]). The numerical experiments confirm that large eigenvalues (e.g. λ_1, λ_2) quickly approach near-steady values in iterations whereas small eigenvalues (e.g. λ_q, λ_{q+1}) vary noticeably per iteration. Hence, the smallest positive eigenvalue (λ_q) has also been redefined at each iteration to avoid its possible near-zero value. When all the eigenvalues in (5) are positive and the smallest λ_p becomes stabilized, say empirically $\lambda_p > 0.1(\varepsilon \lambda_{p-1})$ with $\varepsilon = (\lambda_{p-1}/\lambda_1)^{p-2}$ or $\lambda_p > 0$ in 10 consecutive iterations, we set $\hat{\Lambda}_k = \Lambda_k$.

Specifically, $\bar{\boldsymbol{H}}_k$ asymptotically converges (almost surely) to a positively definite H^* so that $\lambda_p > 0$ as $k \to \infty$ (as shown in Reference [2]). Hence, we have $\hat{\Lambda}_k - \Lambda_\kappa \to 0$ since, asymptotically, elements of $\hat{\Lambda}_k$ are continuous functions of $\bar{\boldsymbol{H}}_k$. We first note that Λ_k is a continuous function of $\bar{\boldsymbol{H}}_k$. Therefore, $\Lambda_k \to \Lambda^*$ almost surely when $\bar{\boldsymbol{H}}_k \to H^*$, where Λ^* denotes all the eigenvalues of H^*. This follows from the basic property of continuous function for deterministic sequence (e.g. Reference [7, p. 70]). Both Λ_k and $\bar{\boldsymbol{H}}_k$ converge for almost all points in their underlying sample spaces. We further note that our mapping from Λ_k, to $\hat{\Lambda}_k$ defined by (6) and (7) is also a continuous function asymptotically. Here, we like to point out that the

mapping f_k defined by (9) preserves the key spectral characters such as the spread of those known positive eigenvalues λ_1/λ_{q-1}. Furthermore, as $k \to \infty$, any mapping for 2SPSA should preserve the complete spectral property of $\bar{\boldsymbol{H}}_k$. Therefore, the proposed mapping to a matrix in 2SPSA is different from the matrix regularization in an ill-posed inversion problem where the spectral property of an ill-conditioned matrix is changed to make the problem well posed (e.g. Reference [5, Chapter 18]).

2.2. *Effect of matrix conditioning on 2SPSA*

It is noted that the 2SPSA recursion (2a) effectively involves computing the inverse matrix $\bar{\bar{\boldsymbol{H}}}_k^{-1}$ (although, in implementation, the explicit inversion should be avoided using standard methods of linear algebra). The mapping f_k defined by (9) guarantees that $\bar{\bar{\boldsymbol{H}}}_k$ is a non-singular matrix. Our mapping procedure of replacing a possible near-zero λ_q with a better behaved $\hat{\lambda}_q$ also eliminates the possibility of a near-singular matrix. However, the elements of $\bar{\boldsymbol{H}}_k$ resulted from the SP approximation and imperfect measurements of the loss function are subject to errors. These errors will directly affect the computed matrix inverse. An underlying rationale for 2SPSA is the strong convergence of both $\hat{\theta}_k$ and its Hessian [2]

$$\hat{\theta}_k \to \theta^*,\ \bar{\boldsymbol{H}}_k(\hat{\theta}_k) \to H^* \text{ (almost surely) as } k \to \infty \tag{12}$$

Thus, the accuracy of $\hat{\theta}_k$ at finite k should be related to that of $\bar{\boldsymbol{H}}_k$. Recursion (2a) indicates a direct relation: the errors of $\hat{\theta}_k$ are proportional to those of $\bar{\bar{\boldsymbol{H}}}_k^{-1}$. Therefore, the performance of 2SPSA will be sensitive to how the errors are affected through the matrix inversion.

The magnitude of errors in a matrix inversion can be quantitatively described by the matrix condition number κ with respect to a matrix norm (e.g. [6, p. 336])

$$\kappa(H) = \left\|H^{-1}\right\| \|H\| \tag{13}$$

where $\|\bullet\|$ denotes arbitrary matrix norm. For a symmetric Hessian matrix H with all positive eigenvalues, its condition number with respect to the spectral norm (κ_λ) provides a straightforward illustration of the ill-conditioning for the computation of matrix inversion. The spectral condition number κ_λ is defined as the ratio of the maximum eigenvalue to the minimum one [6, p. 340]

$$\kappa_\lambda(H) = \lambda_{\max}/\lambda_{\min} \tag{14}$$

We use κ_λ in the numerical studies of Section 4. It can be shown that when $\bar{\bar{\boldsymbol{H}}}_k$ only slightly deviates from the exact H^*, the fractional error in an inverse matrix is approximately proportional to the matrix condition number [6, p. 336]

$$\frac{\|\bar{\bar{\boldsymbol{H}}}_k^{-1} - H^{*-1}\|}{\|H^{*-1}\|} \leqslant \frac{\kappa(H^*)}{1 - \kappa(H^*)(\|\Delta\bar{\bar{\boldsymbol{H}}}_k\| / \|H^*\|)} \cdot \frac{\|\Delta\bar{\bar{\boldsymbol{H}}}_k\|}{\|H^*\|} \quad \text{if } \|\Delta\bar{\bar{\boldsymbol{H}}}_k\| \, \|H^{*-1}\| < 1 \tag{15}$$

where $\Delta\bar{\bar{\boldsymbol{H}}}_k = \bar{\bar{\boldsymbol{H}}}_k - H^*$ is the perturbation from the exact Hessian. It is noted that depending on how $\bar{\boldsymbol{H}}_k$ is derived, the perturbation matrix $\Delta\bar{\boldsymbol{H}}_k$ may also change with κ. Based on our analyses of (2a) and (15) we can conclude that the errors of 2SPSA for an ill-conditioned Hessian of a greater $\kappa(H^*)$ will be greater than a well-conditioned Hessian of a smaller $\kappa(H^*)$. Since 1SPSA (1) does not work with matrix inversion, the additional error sensitivity introduced by matrix inversion that are directly related to $\kappa(H^*)$ will not exist in 1SPSA.

3. MODIFIED 2SPSA

3.1. Description of a modified 2SPSA (M2SPSA)

Several numerical studies have suggested that 2SPSA may outperform 1SPSA in practice (e.g. References [2, 8]). The underlying reason can be understood as follows: 1SPSA predetermines the gain series (a_k) for the whole iteration process whereas 2SPSA derives a generalized gain series $(\bar{a}_k \bar{\bar{H}}_k^{-1})$ that is adapted to near optimality at each iteration. However, based on our analyses in the last section, the inverse of the estimated Hessian generally introduces additional error sensitivity inherited in $\bar{H}_k$ for a non-perfectly conditioned matrix ($\kappa > 1$). To avoid computing the inverse of an ill-conditioned matrix while still approximately optimizing the gain series at each iteration we can modify the first recursion for 2SPSA (2a) by replacing $\hat{\Lambda}_k$ in the mapping f_k of (9) with $\bar{\Lambda}_k$ that contains constant diagonal elements

$$\hat{\theta}_{k+1} = \hat{\theta}_k - \bar{a}_k \bar{\lambda}_k^{-1} \hat{g}_k(\hat{\theta}_k) \tag{16}$$

where $\bar{\lambda}_k$ is the geometric mean of all the eigenvalues of $\bar{\bar{H}}_k$

$$\bar{\lambda}_k = (\lambda_1 \lambda_2 \cdots \lambda_{q-1} \hat{\lambda}_q \hat{\lambda}_{q+1} \cdots \hat{\lambda}_p)^{1/p} \tag{17}$$

Recursions (16) and (2b) together with (5)–(7) and (17) form a modified 2SPSA (M2SPSA) that takes advantage of both the well-conditioned 1SPSA and the internally determined gain sequence of 2SPSA. The proportionality coefficient a of $a_k\,(= a/(k+1+A)^\alpha,\ A \geqslant 0)$ in 1SPSA depends on the individual loss function and is generally selected by a trial-and-error approach in practice (e.g. Reference [9]). On the other hand, 2SPSA removes such an uncertainty in selecting its proportionality coefficient $\bar{a}$ of $\bar{a}_k\,(= \bar{a}/(k+1+A)^\alpha,\ A \geqslant 0)$ since the asymptotically near-optimal selection of $\bar{a}$ is 1 [2]. The crucial property that a in 1SPSA is dependent on the individual loss function has been built into 2SPSA by its generalized gain series $((k+1+A)^{-\alpha} \bar{\bar{H}}_k^{-1},\ A \geqslant 0)$. From this perspective, our M2SPSA (16) can be considered as an extension of 1SPSA in which a is replaced by a scalar series $\bar{\lambda}_k^{-1}$ that depends on the individual loss function and varies with iteration.

3.2. Asymptotic efficiency analysis

The strong convergence of $\hat{\theta}_k$ generally implies an asymptotic normal distribution. Spall [1, 2] established the asymptotic normal distributions for both 1SPSA and 2SPSA. Although our interests are mainly in finite samples, let us present the following asymptotic arguments as a way of relating to previous known results. Since the M2SPSA can also be considered as an extension of 1SPSA with a special gain series $\bar{\lambda}_k^{-1}$ the analysis of the asymptotic normality for 1SPSA can also be extended to M2SPSA. In this section, we first review the asymptotic normal distributions for 1SPSA and 2SPSA. Then, the asymptotic efficiency is compared for three different algorithms of lSPSA, 2SPSA, and M2SPSA.

3.2.1. Asymptotic normality of $\hat{\theta}_k$ in 1SPSA

Using Fabian's [10] result, Spall [1] established the following asymptotic normality of $\hat{\theta}_k$ in 1SPSA:

$$k^{\beta/2}(\hat{\theta}_k - \theta^*) \xrightarrow{\text{dist}} N(\xi, \Sigma) \quad \text{as } k \to \infty \tag{18}$$

where ξ and Σ are the mean vector and covariance matrix and $\beta/2$ characterizes the rate of convergence and is related to the parameters of gain sequences a_k and c_k. The mean ξ in (18) depends on the third derivatives of the loss function at θ^* and generally vanishes except for a special set of gain sequences.

The covariance matrix Σ for $\alpha<1$ is orthogonally similar to the diagonal matrix that is proportional to the inverse eigenvalues of the Hessian

$$\Sigma = \psi a P^{\mathrm{T}} \Lambda^{*-1} P \tag{19}$$

where P is orthogonal with $H^* = P\Lambda^* P^{\mathrm{T}}$, $\Lambda^* = \mathrm{diag}[\lambda_1^*, \lambda_2^*, \ldots, \lambda_p^*]$, and the coefficient of proportionality ψ depends on the statistical parameters in the algorithm [1]. Again, according to the eigenvalue perturbation theorem [6, p. 365] the difference between λ_i ($i = 1, 2, \ldots, p$) at the kth iteration and λ_i^* in (19) is bounded by the difference in its Hessian

$$|\lambda_i - \lambda_i^*| \leqslant \kappa_\lambda(P) \left\| \bar{\boldsymbol{H}}_k(\hat{\theta}_k) - H^* \right\|_2, \quad i = 1, 2, \ldots, p \tag{20}$$

where $\|\bullet\|_2$ denotes the spectral norm of a matrix [6, p. 295] that leads to the definition of spectral condition number (14). It is noted that $\bar{\boldsymbol{H}}_k(\hat{\theta}_k)$ converges almost surely to H^* and the mapping from $\bar{\boldsymbol{H}}_k$ to $\bar{\bar{\boldsymbol{H}}}_k$ defined by (9) preserves the matrix spectra. Furthermore, $\hat{\Lambda}_k - \Lambda_k \to 0$ as $k \to \infty$ and the calculation from $\bar{\boldsymbol{H}}_k$ to Λ_k is a continuous function, we also have the following strong convergence for the eigenvalues of Hessian:

$$\Lambda_k \to \Lambda^* = \mathrm{diag}[\lambda_1^*, \lambda_2^*, \ldots, \lambda_p^*], \bar{\lambda}_k \to \bar{\lambda}^* \text{ (almost surely) as } k \to \infty \tag{21}$$

where $\bar{\lambda}^*$ is the geometric mean of all the eigenvalues of H^*. Based on (18), (19) and (21) we conclude that the choice of $\bar{a}_k \bar{\lambda}_k^{-1}$ in M2SPSA can also be considered as a natural extension of 1SPSA with a sensible selection of a based on its asymptotic normality.

3.2.2. *Asymptotic normality of $\hat{\theta}_k$ in 2SPSA and M2SPSA*

To further illustrate the above point and compare M2SPSA with 2SPSA asymptotically, we consider the asymptotic normality of $\hat{\theta}_k$ for 2SPSA for the gain sequence of the form $\bar{a}_k \sim k^{-\alpha}$ and $c_k \sim \kappa^{-\gamma}$. The asymptotic distribution is given by [2]

$$k^{\beta/2}(\hat{\theta}_k - \theta^*) \xrightarrow{\text{dist}} N(\mu, \Omega) \text{ as } k \to \infty \tag{22}$$

where $\beta = \alpha - 2\gamma$. The covariance matrix Ω is proportional to $H^{*-2} = P\Lambda^{*-2}P^{\mathrm{T}}$ with the same coefficient of proportionality ψ as in (19), and the mean μ depends on both the gain sequence parameters and the third derivatives of the loss function at θ^*. The asymptotic mean square error (MSE) of $k^{\beta/2}(\hat{\theta}_k - \theta^*)$ in (22) is [2]

$$\mathrm{MSE}_{2\mathrm{SPSA}}(\alpha, \gamma) = \mu^{\mathrm{T}}\mu + \mathrm{trace}(\Omega) \tag{23}$$

We first consider a special case of a diagonal Hessian with constant eigenvalues ($\lambda_i^* = \lambda = \bar{\lambda}^*$). It can be shown that the asymptotic normality of $\hat{\theta}_k$, in 2SPSA [2] is identical to that in 1SPSA [1] when the following gain sequences are picked:

$$N(\mu, \Omega) = N(\xi, \Sigma) \quad \text{when } \bar{a}_k = \phi/(k+1) \text{ and } a_k = \phi/[(k+1)\lambda] \tag{24}$$

where the constant ϕ represents a common scale factor for the two gain sequences. The near-optimal selection of ϕ for 2SPSA is $\phi = 1$. Note that the true optimal selection of the gain is essentially infeasible as it depends on the third derivatives of the loss [2]. Equation (24) suggests

that the near-optimal MSE in 2SPSA can be achieved in 1SPSA by picking its proportionality coefficient a in such a way that $a = 1/\bar{\lambda}$. Since a in 1SPSA is externally prescribed, such an optimal picking of a is only theoretically possible. On the other hand, the internally determined gain sequence of $\bar{a}_k \bar{\lambda}_k^{-1}$ $(= k^{-1}\lambda_k^{-1})$ in M2SPSA with (21) makes the near-optimal picking for the special case of constant eigenvalues practically possible.

Next, we consider the specification of the gain sequence $\alpha < 1$ and $3\gamma - \alpha/2 > 0$ from which we have $\mu = \xi = 0$ [1, 2]. The asymptotic distribution-based MSE for 2SPSA under this condition is inversely proportional to the sum of all the eigenvalues squared

$$\mathrm{MSE}_{\mathrm{2SPSA}}(\alpha, \gamma) = \mathrm{trace}\,\Omega \propto \mathrm{trace}(\Lambda^{*-2}) = \sum_{i=1}^{p} \lambda_i^{*-2} \tag{25}$$

On the other hand, the MSE for M2SPSA can be derived by setting $a = 1/\bar{\lambda}^*$ in 1SPSA

$$\mathrm{MSE}_{\mathrm{M2SPSA}}(\alpha, \gamma) = \mathrm{trace}\,\Sigma|_{a=1/\bar{\lambda}^*} \propto \bar{\lambda}^{*-1}\,\mathrm{trace}\,(\Lambda^{*-1}) = \bar{\lambda}^{*-1} \sum_{i=1}^{p} \lambda_i^{*-1} \tag{26}$$

The constants of proportionality are related to c and to the variances of Δ_k and measurement noise [1,2]. They are identical in the present settings of (25) and (26). Therefore, the ratio of MSEs for M2SPSA to 2SPSA is given by

$$\frac{\mathrm{MSE}_{\mathrm{M2SPSA}}(\alpha, \gamma)}{\mathrm{MSE}_{\mathrm{2SPSA}}(\alpha, \gamma)} = \frac{[\prod_{i=1}^{p} \lambda_i^{*-1}]^{1/p}}{\sqrt{(1/p)\sum_{i=1}^{p} \lambda_i^{*-2}}} \cdot \frac{(1/p)\sum_{i=1}^{p} \lambda_i^{*-1}}{\sqrt{(1/p)\sum_{i=1}^{p} \lambda_i^{*-2}}} \equiv R_0 \leqslant 1 \tag{27}$$

where we have used a well-known relation in the last inequality of (27):

$$(\text{geometric mean}) \leqslant (\text{arithmetic mean}) \leqslant (\text{root-mean-square}) \tag{28}$$

Equality in (28) holds only when all the eigenvalues are equal which corresponds to a perfectly conditioned Hessian of $\kappa(H^*) = 1$. Since the ratio R_0 has been derived from the asymptotic MSEs the comparison between M2SPSA and 2SPSA has been made under the same rate of convergence.

Our third case in the asymptotic efficiency analysis is to consider $\alpha = 1$ when $3\gamma - \alpha/2 > 0$ (i.e. $\gamma > 1/6$) in 2SPSA. This setting again corresponds to $\mu = \xi = 0$ in 2SPSA and M2SPSA. It is possible for both 1SPSA and 2SPSA to set $\alpha = 1$ for their gain sequence selection. The near-optimal rate of convergence in 2SPSA by setting $\bar{a} = 1$ can be accomplished in 1SPSA by adjusting its a to yield the same rate of convergence as 2SPSA [2]. By setting $a = 1/\bar{\lambda}$ in 1SPSA for the implementation of M2SPSA we can again derive (27) that shows the superiority of M2SPSA to 2SPSA under the same rate of convergence. However, the above setting of $a = 1/\bar{\lambda}$ in 1SPSA is allowed only if the resulting condition in 1SPSA of $\min_i(\lambda_i/\bar{\lambda}) > \beta/2$ still holds [1]. When the above condition is violated while implementing M2SPSA for relatively large $\kappa_\lambda(H)$, the setting of $\alpha = 1$ in M2SPSA is excluded and we can no longer make a straight comparison of the asymptotic MSEs between 2SPSA and M2SPSA under the same rate of convergence. Under this circumstance, there is no superiority of either one of M2SPSA and 2SPSA to the other in terms of the efficiency or the rate of convergence. The superiority of M2SPSA to 2SPSA indicated by (27) only shows an improvement in the multiplier for the convergence rate (R_0) when the common convergence rate is sub-optimal.

Spall [2] showed that by setting $\alpha = 1$ and $\gamma = 1/6$ an asymptotically optimal MSE can be achieved with a maximum rate of convergence for the MSE of $\hat{\theta}_k$ of $k^{-\beta} = k^{-2/3}$ in both 1SPSA

and 2SPSA. We have already shown that in order to avoid the violation of the condition $\min_i(\lambda_i/\bar{\lambda}) > \beta/2$ the setting of $\alpha = 1$ (with $\beta \approx 2/3$) is often not allowed in M2SPSA. Neither is it possible to choose a different set of α_m and γ_m to yield $\beta_m = 2/3$ when $\gamma = 1/6$. Under this circumstance, the maximum rate of convergence of $k^{-2/3}$ for MSE cannot be achieved by M2SPSA. It is noted that the mapping f_k such as the one proposed in section *II.A* will leave the asymptotic $\bar{\boldsymbol{H}}_k$ unchanged (when we set $\hat{\Lambda}_k = \Lambda_k$) as $k \to \infty$. On the other hand, M2SPSA changes $\bar{\boldsymbol{H}}_k$ when its Λ_k is replaced by $\bar{\Lambda}_k$. The asymptotically unachievable optimal MSE is the price M2SPSA pays when it forces $\bar{\boldsymbol{H}}_k$ in 2SPSA to a different form of $\bar{\Lambda}_k$.

3.2.3. Efficiency of 1SPSA, 2SPSA, and M2SPSA

The relationships among 1SPSA, 2SPSA and M2SPSA can also be understood from a different perspective: 1SPSA (1) and M2SPSA (16) weight the different components of the estimated gradient $\hat{g}_k(\hat{\theta}_k)$ equally whereas 2SPSA (2a) weights them differently to account for different sensitivities of θ. A steeper eigendirection (greater λ_i) requires a smaller step ($\sim 1/\lambda_i$) to effectively reach the exact solution (e.g. Reference [11, p. 273]). Both 2SPSA and M2SPSA have captured the dependence of the step size on the overall sensitivities of θ at each iteration. From this perspective, 2SPSA and M2SPSA are superior to 1SPSA. However, M2SPSA (16) weights the different components of $\hat{g}_k(\hat{\theta}_k)$ equally with an averaged step ($\sim 1/\bar{\lambda}_k$), it has given up the further advantage of higher-order sensitivity of θ. Therefore, whether M2SPSA is better than 2SPSA or not at finite iterations is determined by the relative importance of two competing factors that influence the efficiency of the algorithm. The elimination of the matrix inverse reduces the magnitude of errors whereas the lack of gradient sensitivity may deteriorate the accuracy. It is noted that the asymptotic relation (27) only shows an improvement of M2SPSA over 2SPSA in terms of its rate coefficient. Both M2SPSA and 2SPSA have the same rate of convergence characterized by $k^{-\beta/2}$ as shown by (22).

The asymptotic relation (27) provides a theoretical rationale of considering M2SPSA over 2SPSA in practice although the maximum rate of convergence of $k^{-2/3}$ for MSE cannot be achieved for M2SPSA. Another rationale of proposing M2SPSA is that the amplification of errors in an ill-conditioned H^* through the matrix inversion is a well-established result whereas the efficiency of the gradient sensitivity through Newton–Raphson search only shows near the extreme point (θ^*) with a near-exact Hessian (e.g. Reference [11, p. 308]). Further support for M2SPSA over 2SPSA is given in the numerical experiments at finite iterations in the next section. Recall, however, that such justification for M2SPSA is restricted to the case where the gains are not asymptotically optimal in order to achieve fast convergence with finite iterations. For the asymptotic optimal gains ($\bar{a}_k \sim 1/k$, $c_k \sim 1/k^{1/6}$), 2SPSA is superior to M2SPSA except in the case where all eigenvalues of H^* are identical (where 2SPSA and M2SPSA are identical).

We have shown that the magnitude of errors in 2SPSA is dependent on the matrix conditioning of H^* due to two competing factors. Since both factors are strongly related to the same quantity of the matrix conditioning, the relative efficiency between M2SPSA and 2SPSA might be less dependent on specific loss functions. It is noted that replacement of the recursion (2a) by (16) eliminates the part of errors amplified by matrix inverse computation. It also removes the higher-order sensitivity of θ that too depends on the matrix conditioning. However, such a replacement does not necessarily suggest that the magnitude of errors in M2SPSA be independent on the matrix conditioning of H^* since the computation of $\bar{\lambda}_k$ is dependent on the matrix properties of H^*.

4. NUMERICAL COMPARISONS

To study the efficiencies of three SPSA algorithms (1SPSA, 2SPSA and M2SPSA) to the matrix conditioning of the loss function Hessian we consider here the simple quadratic loss function built on the prescribed Hessian with $p = 10$

$$L(\theta) = \tfrac{1}{2}\theta^{\mathrm{T}} H \theta \tag{29}$$

The minimum occurs at $\theta^* = 0$ with $L(\theta^*) = 0$. A Gaussian noise is added to the loss function to represent the measurement errors: $y(\theta) = L(\theta) + N(0, \sigma^2)$, where $N(0, \sigma^2)$ represents a random variable having a normal distribution with zero mean and σ^2 variance. The matrix elements of the Hessian are specified according to

$$(H)_{ij} = \beta \exp[-(i-j)^2/\alpha^2] \tag{30}$$

The following four cases are considered for numerical studies.

$$\textit{Case}\ \mathrm{A}: \ \beta = 0.1291,\ \alpha = 1.1311,\ \kappa_\lambda = 10 \tag{31a}$$

$$\textit{Case}\ \mathrm{B}: \ \beta = 0.2144,\ \alpha = 1.5416,\ \kappa_\lambda = 100 \tag{31b}$$

$$\textit{Case}\ \mathrm{C}: \beta = 0.3941,\ \alpha = 1.9047,\ \kappa_\lambda = 1000 \tag{31c}$$

$$\textit{Case}\ \mathrm{D}: \beta = 0.7763,\ \alpha = 2.2597,\ \kappa_\lambda = 10\,000 \tag{31d}$$

All four cases have the same geometric mean of eigenvalues of $\bar{\lambda} = 0.1$. In the above, we have also listed the matrix condition number with respect to the spectral norm as defined in (14) for different cases. Case D (with $\kappa_\lambda = 10\,000$) is worse ill-conditioned than Case C, which in turn is worse ill-conditioned than Cases B and A. Following the general guidance on picking gain series for 1SPSA and 2SPSA [1, 2] the gain sequences a_k, $\bar{a}_k$, c_k and $\tilde{c}_k$, are picked to satisfy standard SA conditions. We set the gain sequences $a_k = \bar{a}_k = a/(k+1+A)^\alpha$, $c_k = c/(k+1)^\gamma$ and $\tilde{c}_k = \tilde{c}/(k+1)^\gamma$ with $\alpha = 0.602$ and $\gamma = 0.101$ near their theoretically allowed low values that are intended to achieve fast convergence with finite iterations. The other constants are set to the following values: $a = 0.5$, $A = 1$, $c = 0.1$, and $\tilde{c} = 0.15$. We use (9) for the mapping for implementing 2SPSA.

Figure 1 shows the plots of averaged loss function versus the number of loss function evaluations (IV) for two cases (A and C) with a noise level of $\sigma = 0.001$ after 50 independent experiments. All the loss functions are normalized by the initial $L(\hat{\theta}_0)$ with each component of $\hat{\theta}_0$ being a uniformly distributed random variable over $(-1, 1)$. The figure shows that in the very early stage of iterations (say $N \leqslant 400$) 1SPSA is better than both 2SPSA and M2SPSA since the estimated Hessian ($\bar{H}_k$) carries significant errors. As $\bar{H}_k$ becomes a better approximation of the real Hessian, 2SPSA based on (2a) and (11) outperforms 1SPSA in the chosen parameter setting when the matrix condition number is not extremely large. The results of Figure 1 support our conjecture that larger matrix condition number yields a slower convergence rate for θ. On the other hand, 1SPSA is less sensitive to the condition number. Figure 1 also shows that M2SPSA based on (16) and (17) is consistently better than 2SPSA in all cases before the iterations reach the noise level, indicating a sound improvement of M2SPSA over 2SPSA based on the elimination of the matrix inversion errors. It is noted from Figure 1 that the convergence rate of M2SPSA also depends on the matrix condition number, which suggests a possible relation between errors in eigenvalue computation and matrix property such as its condition number.

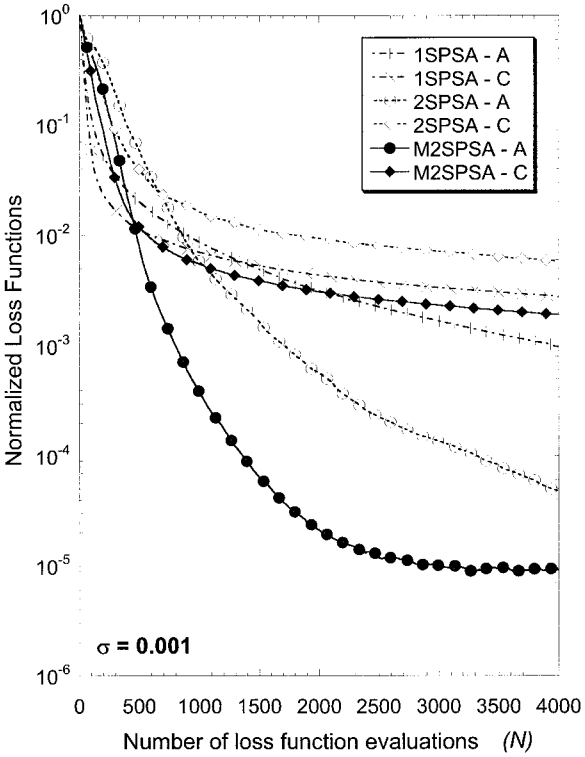

Figure 1. Normalized loss functions versus the number of loss function evaluations for 1SPSA (dash-dot lines), 2SPSA (dashed lines), and M2SPSA (solid lines). The matrix condition numbers for Cases A and C are 10 and 1000, respectively. The noise level $\sigma = 0.001$.

Similar results are obtained for the numerical experiments with a greater noise level of $\sigma = 0.01$ or a noise-free ($\sigma = 0$) setting.

In Figure 2, we show the comparison between 2SPSA and M2SPSA for all four cases of numerical experiments for the noise-free ($\sigma = 0$) setting for the loss function. Again, M2SPSA consistently outperforms 2SPSA in all the cases and the improvements become even more significant at large N. These results point to the strengths of M2SPSA in finite sample problems with gains decaying more slowly than the asymptotic optimal gains.

5. CONCLUSIONS

We have made both empirical and theoretical comparisons between 1SPSA based on (1) and 2SPSA based on (2a) and (10) in the perspective of the loss function Hessian matrix. It is found that the magnitude of errors introduced by matrix inversion in 2SPSA is greater for an ill-conditioned Hessian than a well-conditioned Hessian. On the other hand, the errors in 1SPSA are less sensitive to the matrix conditioning of loss function Hessians. To eliminate the errors introduced by the inversion of estimated Hessian ($\bar{\bar{H}}_k^{-1}$) we suggest a modification (16) to 2SPSA

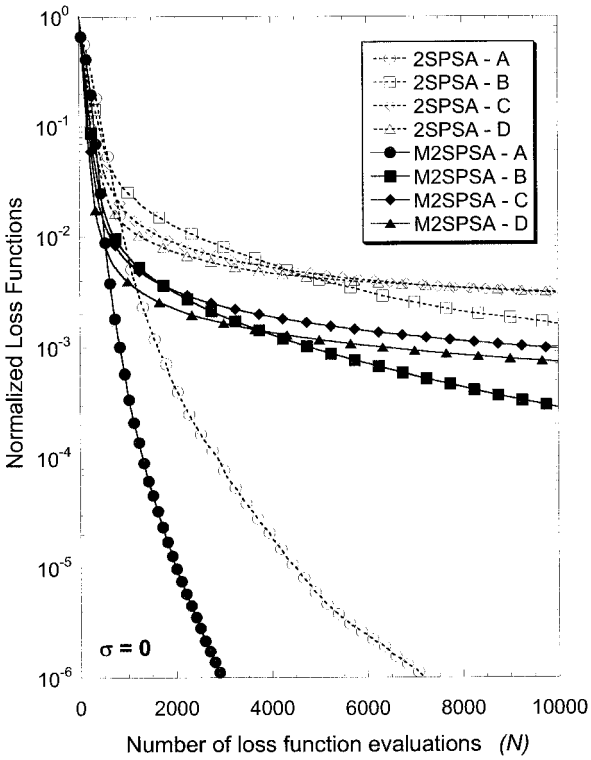

Figure 2. Normalized loss functions versus the number of loss function evaluations for 2SPSA (dashed lines) and M2SPSA (solid lines) and for all four cases of different matrix condition numbers. The noise level $\sigma = 0$.

that replaces $\bar{\bar{H}}_k^{-1}$ with a scalar inverse of the geometric mean of all the eigenvalues of $\bar{\bar{H}}_k$. At finite iterations, it is found that the newly introduced M2SPSA based on (16) and (17) consistently outperforms 2SPSA in the numerical experiments that represent a wide range of matrix conditioning. The asymptotic efficiency analysis shows that the ratio of the mean square errors for M2SPSA to 2SPSA is always less than unity except for a perfectly conditioned Hessian or for an asymptotically optimal setting of the gain sequence.

ACKNOWLEDGEMENTS

Discussions on the manuscript with Daniel C. Chin are appreciated. X.Z.'s research was supported by the NSF Grant ATM-0091514 and in part by the TIMED project sponsored by NASA under contract NAS5-97179 to the Johns Hopkins University Applied Physics Laboratory. J.C.S.'s research was supported by the JHU/APL IRAD Program and U.S. Navy contract N00024-98-D-8124.

REFERENCES

1. Spall JC. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control* 1992; **37**:332–341.

2. Spall JC. Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Transactions on Automatic Control* 2000; **45**:1839–1853.
3. Chin DC. Comparative study of stochastic algorithms for system optimization based on gradient approximations. *IEEE Transactions on Systems Man, and Cybernetus Part B*, 1997; **27**(2):244–249.
4. Spall JC, Hill SD, Stark DR. Some theoretical comparisons of stochastic optimization approaches. *Proceedings of the American Control Conference*, Chicago, IL, June 2000; 1904–1908.
5. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. *Numerical Recipes in Fortran. The Arts of Scientific Computing* (2nd edn). Cambridge University Press: Cambridge, MA, 1992.
6. Horn RA, Johnson CR. *Matrix Analysis*. Cambridge University Press: Cambridge, 1985.
7. Apostol TM. *Mathematical Analysis* (2nd edn). Addison-Wesley: Reading, MA, 1974.
8. Luman RR. Upgrading complex systems of systems: a CAIV methodology for warfare area requirements allocation. *Military Operations Research* 2000; **5**(2):53–75.
9. Spall JC. Implementation of the simultaneous perturbation algorithm for stochastic optimization. *IEEE Transactions on Aerospace and Electronic Systems* 1998; **34**(3):817–823.
10. Fabian V. On asymptotic normality in stochastic approximation. *Annals of Mathematics and Statistics* 1968; **39**:1327–1332.
11. Pierre DA. *Optimization Theory with Applications*. Dover Pub. Inc.: New York, 1986.

