Proceedings of the American Control Conference
Arlington, VA June 25-27, 2001

Matrix Conditioning and Adaptive Simultaneous Perturbation Stochastic Approximation Method

Xun Zhu
The Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road, Laurel, MD 20723-6099, U.S.A.
E-mail: xun.zhu@jhuapl.edu

Abstract — **This paper proposes a modification to the simultaneous perturbation stochastic approximation (SPSA) methods based on the comparisons made between the first order and the second order SPSA (1SPSA and 2SPSA) algorithms from the perspective of loss function Hessian. At finite iterations, the convergence rate depends on the matrix conditioning of the loss function Hessian. It is shown that 2SPSA converges more slowly for a loss function with an ill-conditioned Hessian than the one with a well-conditioned Hessian. On the other hand, the convergence rate of 1SPSA 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 at finite iterations which leads to significant improvements in its convergence rate in problems with an ill-conditioned Hessian matrix. Asymptotically, the efficiency analysis shows that M2SPSA is also superior to 2SPSA in terms of their convergence rate coefficients. It is shown that for the same asymptotic convergence rate, the ratio of the mean square errors for M2SPSA to 2SPSA is always less than one except for a perfectly conditioned Hessian.**

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 Kiefer-Wolfowitz (K-W) stochastic approximation method (Spall, 1992) whereas the second order SPSA (2SPSA) is a stochastic analogue of the deterministic Newton-Raphson algorithm (Spall, 2000). There have been several studies that compare the efficiency of 1SPSA with other stochastic approximation (SA) methods (e.g., Spall, 1992; Chin, 1997; Spall et al., 2000). 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 (2000) 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 ($\overline{a}_k = 1/k$ 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.

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. There are two approaches to compare different algorithms: theoretical and empirical. The theoretical approach attempts to discover the asymptotic convergence rate of an algorithm that will hold for general loss functions. On the other hand, the empirical approach generally assesses different algorithms based on a few selected examples. Our comparisons in this paper will focus on both the empirical results at finite iterations and the theoretical results on the asymptotic efficiency. 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. The asymptotic results cover most parameter domains for the gain sequence specification.

2. MATRIX CONDITIONING AND ITS RELATION TO 2SPSA

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

0-7803-6495-3/01/$10.00 © 2001 AACC

1SPSA (Spall 1992):

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

2SPSA (Spall 2000):

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

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

where a_k and $\overline{a}_k$ 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 of $c_k = c / k^{\gamma}$, $\hat{H}_k$ is the SP estimate of the Hessian matrix, and f_k maps the usual non-positive-definite $\overline{H}_k$ to a positive definite pxp matrix. Readers are referred to Spall (1992, 1998, 2000) 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 $\overline{H}_k$ to $\overline{\overline{H}}_k$ since the former is hardly positive definite in practice. We suggest the following approach that eliminates the non-positive-definiteness while preserving key spectral properties of $\overline{H}_k$. First, we compute the eigenvalues of $\overline{H}_k$ and to sort them into descending order:

$$\Lambda_k \equiv \mathrm{diag}[\lambda_1, \lambda_2, \ldots, \lambda_{q-1}, \lambda_q, \lambda_{q+1}, \ldots, \lambda_p] \tag{3}$$

where $\lambda_q > 0$ and $\lambda_{q+1} \leq 0$. Next, by considering that all the negative eigenvalues are unphysical and are caused by errors in $\overline{H}_k$ we replace them together with the smallest positive eigenvalue with a descending series of positive eigenvalues:

$$\hat{\lambda}_q = \varepsilon\lambda_{q-1}\,,\ \hat{\lambda}_{q+1} = \varepsilon\hat{\lambda}_q\,,\ \ldots,\ \hat{\lambda}_p = \varepsilon\hat{\lambda}_{p-1}\,, \tag{4}$$

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

$$\varepsilon = \left(\lambda_{q-1} / \lambda_1\right)^{q-2}\,. \tag{5}$$

Numerical experiments show 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. Equations (4) and (5) indicate that the spectral character of the existing positive eigenvalues as measured by the ratio of its maximum to minimum values is extrapolated to the rest of the matrix spectrum. The specification of (5) bears an ad hoc feature that is common in all the extrapolation techniques. Other forms of specifications such as $\varepsilon = \left(\lambda_{q-1} / \lambda_1\right)^{(q-2)/2}$ or $\varepsilon = 1$ would also effectively eliminate the non-positive-definiteness. Since the separating point between the positive and negative eigenvalues q slowly increases from 1 to p, we find numerically that the specification based on (5) yields relatively a faster convergence rate in most cases. Since $\overline{H}_k$ is symmetric it is orthogonally similar to the real diagonal matrix of its real eigenvalues (e.g., Horn and Johnson, 1985, p. 171)

$$\overline{H}_k = P\Lambda_k P^T\,, \tag{6}$$

where the orthogonal matrix P consists of all the eigenvectors of $\overline{H}_k$, which are usually derived together with the eigenvalues (e.g., Press, 1992, p. 460). Now, the mapping f_k can be expressed as

$$f_k(\overline{H}_k) = P\hat{\Lambda}_k P^T\,, \tag{7}$$

where $\hat{\Lambda}_k$ is the diagonal matrix Λ_k with part of its eigenvalues redefined according to (4). Since it is $\overline{\overline{H}}_k^{-1}$ that is used in the 2SPSA recursion (2a) the mapping (7) with the available eigenvectors of $\overline{H}_k$ also leads to an easy inversion of the estimated Hessian:

$$\overline{\overline{H}}_k^{-1} = P\hat{\Lambda}_k^{-1}P^T\,. \tag{8}$$

The 2SPSA (2) based on the mapping (7) makes the procedure of eliminating the non-positive-definiteness of $\overline{H}_k$ a precise one. It is noted that the key parameters needed for the mapping (ε and λ_{q-1}) are internally determined by $\overline{H}_k$ at each iteration. This is different from some other forms of f_k where an externally prescribed number series is needed.

2.2 Effect of matrix conditioning on 2SPSA

It is noted that the 2SPSA recursion (2a) involves computing the inverse matrix $\overline{\overline{H}}_k^{-1}$. The mapping f_k defined by (7) guarantees that $\overline{\overline{H}}_k$ is a nonsingular 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 $\overline{H}_k$ 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 (Spall, 2000):

$$\hat{\theta}_k \to \theta^*, \overline{H}_k(\hat{\theta}_k) \to H(\theta^*) \text{ (almost surely) as } k \to \infty . \quad (9)$$

Thus, the convergence rate of $\hat{\theta}_k$ at finite k should be related to that of $\overline{H}_k$. The recursion (2a) indicates a direct relation: the convergence rate of $\hat{\theta}_k$ is proportional to the convergence rate of $\overline{\overline{H}}_k^{-1}$. Therefore, the performance of 2SPSA will be sensitive to how the errors are amplified through the matrix inversion.

The amplification of errors in a matrix inversion can be quantitatively described by its matrix condition number κ with respect to a matrix norm (e.g., Horn and Johnson, 1985, p. 336)

$$\kappa(H) = \left\| H^{-1} \right\| \left\| H \right\| . \quad (10)$$

where $\|\bullet\|$ denotes matrix norm. For a symmetric Hessian matrix H with all positive eigenvalues, its condition number with respect to the spectral norm (κ_λ) is the ratio of the maximum eigenvalue to the minimum one (Horn and Johnson, 1985, p. 340)

$$\kappa_\lambda(H) = \lambda_{\max} / \lambda_{\min} . \quad (11)$$

It can be shown that when $\overline{H}_k$ only slightly deviates from the exact $H(\theta^*)$, the amplification of the errors through the matrix inversion is approximately proportional to the matrix condition number (Horn and Johnson, 1985, p. 336)

$$\frac{\left\| \overline{\overline{H}}_k^{-1} - H^{-1} \right\|}{\left\| H^{-1} \right\|} \leq \frac{\kappa(H)}{1 - \kappa(H)\left(\|\Delta H\| / \|H\| \right)} \cdot \frac{\|\Delta H\|}{\|H\|} ,$$

$$\text{if } \|\Delta H\| \left\| H^{-1} \right\| < 1 , \quad (12)$$

where $\Delta H = \overline{\overline{H}}_k - H$ is the perturbation to the exact Hessian. It is noted that depending on how $\overline{\overline{H}}_k$ is derived the perturbation matrix $\Delta H = \overline{\overline{H}}_k - H$ may also change with κ. Based on our analyses of (2a) and (12) we can conclude that the convergence rate of 2SPSA for an ill-conditioned Hessian of a greater $\kappa(H)$ will be slower than a well-conditioned Hessian of a smaller $\kappa(H)$. Since 1SPSA (1) does not work with matrix inverses, the additional errors introduced by matrix inversion that is directly connected 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., Spall, 2000; Luman, 2000). The underlying reason can be understood as follows: 1SPSA prescribes the gain series (a_k) in the whole iteration process whereas 2SPSA derives a generalized gain series ($\overline{a}_k \overline{\overline{H}}_k^{-1}$) 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 amplifies the errors inherited in $\overline{H}_k$ for a non-perfectly conditioned matrix (κ>1). To avoid computing inverse of an ill-conditioned and error-bearing matrix while still 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 (7) with $\overline{\Lambda}_k$ that contains constant diagonal elements

$$\hat{\theta}_{k+1} = \hat{\theta}_k - \overline{a}_k \overline{\lambda}_k^{-1} \hat{g}_k(\hat{\theta}_k) , \quad (13)$$

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

$$\overline{\lambda}_k = (\lambda_1 \lambda_2 \cdots \lambda_{q-1} \hat{\lambda}_q \hat{\lambda}_{q+1} \cdots \hat{\lambda}_p)^{1/p} . \quad (14)$$

The recursions (13) and (2b) together with (3)-(5) and (14) form a modified 2SPSA (M2SPSA) that takes advantage of both the well-conditioned 1SPSA and the higher order convergence rate of 2SPSA. Following Spall (2000), both gain series a_k and $\overline{a}_k$ are picked as proportional to $(k + A)^{-\alpha}$, $A \geq 0$, with α (= 0.602) near its theoretically allowed low value to achieve fast convergence with finite iterations. The proportionality coefficient a in 1SPSA depends on the individual loss function and is generally selected by a trial-and-error approach in practice (e.g., Spall, 1998). On the other hand, 2SPSA removes such an uncertainty in selecting the proportionality coefficient since the near-optimal selection of its $\overline{a}$ is 1 (Spall, 2000). The crucial property that a in 1SPSA is dependent on the individual loss function has been built into 2SPSA by its generalized gain series ($(k + A)^{-\alpha} \overline{\overline{H}}_k^{-1}$). From this perspective, our M2SPSA (13) can be considered as an extension of 1SPSA in which a is replaced by a scalar series $\overline{\lambda}_k^{-1}$ that depends on the individual loss function and varies with iteration.

3.2 Asymptotic efficiency analysis

To see further connections between M2SPSA and 1SPSA we note the asymptotic normality of $\hat{\theta}_k$ in 1SPSA (Spall, 1992)

$$k^{\beta/2}(\hat{\theta}_k - \theta^*) \xrightarrow{\text{dist}} N(\xi, \Sigma) \text{ as } k \to \infty , \quad (15)$$

where the mean ξ depends on the third derivatives of the loss function at θ^* and generally vanishes except for a special set of gain sequence. The covariance matrix Σ for $\alpha < 1$ is orthogonally similar to the diagonal matrix that is proportional to the inverse eigenvalues of the Hessian

$$\Sigma \sim aP^{*}\operatorname{diag}[\lambda_1^{*-1}, \lambda_2^{*-1}, \cdots, \lambda_p^{*-1}]P^{*T} . \qquad (16)$$

According to the eigenvalue perturbation theorem (Horn and Johnson, 1985, p. 365) the difference between λ_i (i = 1,2,...,p) at the kth iteration and λ_i^* in (16) is bounded by the difference in its Hessian

$$\left|\lambda_i - \lambda_i^*\right| \le \kappa(P^*)\left\|\overline{H}_k(\hat{\theta}_k) - H(\theta^*)\right\|, \quad i = 1,2,..., p. \qquad (17)$$

Since $\overline{H}_k(\hat{\theta}_k)$ converges almost surely to $H(\theta^*)$ and the mapping from $\overline{H}_k$ to $\overline{\overline{H}}_k$ defined by (7) preserves the matrix spectra we also have the following strong convergence for the eigenvalues of Hessian

$$\Lambda_k \to \Lambda^* = \operatorname{diag}[\lambda_1^*, \lambda_2^*, \cdots, \lambda_p^*],$$
$$\overline{\lambda}_k \to \overline{\lambda}^* \text{ (almost surely) as } k \to \infty , \qquad (18)$$

where $\overline{\lambda}^*$ is the geometric mean of all the eigenvalues of $H(\theta^*)$. Based on (15), (16) and (18) we conclude that the choice of $\overline{a}_k \overline{\lambda}_k^{-1}$ in M2SPSA can also be considered as a natural extension of 1SPSA with a sensible selection of a based on its asymptotic normality.

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 $\overline{a}_k \sim k^{-\alpha}$ and $c_k \sim k^{-\gamma}$. It is given by (Spall, 2000)

$$k^{\beta/2}(\hat{\theta}_k - \theta^*) \xrightarrow{\text{dist}} N(\mu, \Omega) \text{ as } k \to \infty, \qquad (19)$$

where $\beta = \alpha - 2\gamma$. The covariance matrix Ω is proportional to $H(\theta^*)^{-2} = P\Lambda^{*-2}P^T$ and the mean μ depends on both the gain sequence parameters and the third derivatives of the loss function at θ^*. The mean square error (MSE) of $\hat{\theta}_k - \theta^*$ for the asymptotic distribution (19) is (Spall 2000)

$$\text{MSE}_{2SPSA}(\alpha, \gamma) = k^{-\beta}[\mu^T \mu + \operatorname{trace}(\Omega)] . \qquad (20)$$

We first consider a special case of a diagonal Hessian with constant eigenvalues ($\lambda_i^* = \lambda = \overline{\lambda}^*$). It can be shown that the asymptotic normality of $\hat{\theta}_k$ in 2SPSA (Spall, 2000) is identical to that in 1SPSA (Spall 1992) when the optimal gain sequences are picked:

$$N(\mu, \Omega) = N(\xi, \Sigma) \text{ when } \overline{a}_k = 1/k \text{ and } a_k = 1/(k\lambda) . \quad (21)$$

Equation (21) suggests that the optimal MSE in 2SPSA can be achieved in 1SPSA by picking its proportionality coefficient a in such a way that $a = 1/\lambda$. 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 $\overline{a}_k \overline{\lambda}_k^{-1}$ ($= k^{-1}\lambda_k^{-1}$) in M2SPSA with (18) makes the optimal picking 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 (Spall 1992, 2000). The MSE for 2SPSA under this condition is inversely proportional to the sum of all the eigenvalues squared

$$\text{MSE}_{2SPSA}(\alpha, \gamma) = k^{-\beta} \operatorname{trace}\Omega$$
$$\propto k^{-\beta} \operatorname{trace}(\Lambda^{*-2}) = k^{-\beta} \sum_{i=1}^{p} \lambda_i^{*-2} . \qquad (22)$$

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

$$\text{MSE}_{M2SPSA}(\alpha, \gamma) = k^{-\beta} \operatorname{trace}\Sigma\big|_{a = 1/\overline{\lambda}^*}$$
$$\propto k^{-\beta}\overline{\lambda}^{*-1} \operatorname{trace}(\Lambda^{*-1}) = k^{-\beta}\overline{\lambda}^{*-1} \sum_{i=1}^{p} \lambda_i^{*-1} . \qquad (23)$$

Therefore, the ratio of MSEs for M2SPSA to 2SPSA is given by

$$\frac{\text{MSE}_{M2SPSA}(\alpha,\gamma)}{\text{MSE}_{2SPSA}(\alpha,\gamma)} = \frac{\left[\prod_{i=1}^{p} \lambda_i^{*-1}\right]^{1/p}}{\sqrt{\frac{1}{p}\sum_{i=1}^{p} \lambda_i^{*-2}}} \cdot \frac{\frac{1}{p}\sum_{i=1}^{p} \lambda_i^{*-1}}{\sqrt{\frac{1}{p}\sum_{i=1}^{p} \lambda_i^{*-2}}} \equiv R_0 \le 1 , \qquad (24)$$

where we have used a well-known relation in the last inequality of (24)

$$\text{(geometric mean)} \le \text{(arithmetic mean)}$$
$$\le \text{(root-mean-square)}. \qquad (25)$$

The equality in (24) holds only when all the eigenvalues are equal which corresponds to a perfectly conditioned Hessian of $\kappa_\lambda(H) = 1$. It is noted that the comparison between M2SPSA and 2SPSA has been made under the assumption of both MSEs having the same rate of convergence of $k^{-\beta}$.

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 $\overline{a} = 1$ can be accomplished in 1SPSA by adjusting its a to yield the same rate of convergence as 2SPSA (Spall 2000). The implementation of M2SPSA requires one to pick $a = 1/\bar{\lambda}$. This does not allow one to set $\alpha = 1$ in M2SPSA because such a setting will generally lead to a violation of the condition $2\min_i(\lambda_i/\bar{\lambda}) > \beta$ for 1SPSA (Spall 1992). A higher rate of convergence in M2SPSA can be achieved by choosing a different set of $\alpha_m < 1$ and $\gamma_m \neq \gamma$ from those for 2SPSA.

We now consider $3\gamma - \alpha/2 > 0$ when $\alpha = 1$ in 2SPSA. This setting again corresponds to $\mu = \xi = 0$ in 2SPSA and M2SPSA. It can be shown that given $\alpha = 1$ and $\gamma = (1/6) + \varepsilon$, ε>0, we can choose α_m and γ_m such that $1 - 3\varepsilon < \alpha_m < 1$ and $\gamma_m = (1/6) - \delta$ with $\delta < \varepsilon/2$ that satisfy the condition $3\gamma_m - \alpha_m/2 > 0$ and at the same time yield

$$\frac{\mathrm{MSE}_{M2SPSA}(\alpha_m, \gamma_m)}{\mathrm{MSE}_{2SPSA}(1, \gamma)} = \frac{k^{-\beta_m}}{k^{-\beta}} R_0 \to 0 \text{ as } k \to \infty, \quad (26)$$

where $\beta_m = \alpha_m - 2\gamma_m$. Similarly, given set of α_m and γ_m for M2SPSA we can also appropriately choose a different set of α and γ for 2SPSA that yields a better asymptotic MSE. In other words, there is no superiority of either one of M2SPSA and 2SPSA to the other in terms of the rate of convergence. The superiority of M2SPSA to 2SPSA shown by (24) only shows an improvement in the convergence rate coefficient.

Spall (2000) showed that by setting $\alpha = 1$ and $\gamma = 1/6$ an asymptotically optimal MSE can be achieved with a maximum rate of convergence of $k^{-\beta} = k^{-3/2}$ in both 1SPSA and 2SPSA. Since the setting of $\alpha = 1$ is not generally allowed in M2SPSA we can again show that the maximum rate of convergence of $k^{-3/2}$ can only be a supremum for M2SPSA. It should be pointed out that the optimal setting of $\alpha = 1$ and $\gamma = 1/6$ should always be understood in the limit process of $k \to \infty$. For any given $k < \infty$, one can always find $\gamma' = 1/6 + \varepsilon/2$ with $\varepsilon > 0$ (and $\mu = 0$) such that

$$\frac{MSE_{2SPSA}(1, \gamma')}{MSE_{2SPSA}(1, \frac{1}{6})} = \frac{k^{\varepsilon}\,\mathrm{trace}\,\Omega}{[\mu^T \mu + \mathrm{trace}\,\Omega]} < 1 \text{ for any } k < \infty. \quad (27)$$

Equation (27) coupled with (26) shows that by choosing ε such that

$$0 < \varepsilon < \frac{\log[\mu^T \mu + \mathrm{trace}\,\Omega] - \log[\mathrm{trace}\,\Omega]}{\log k} \quad (28)$$

we can construct a gain sequence for M2SPSA that can be infinitely close to the maximum rate of convergence of $k^{-3/2}$ achieved by 2SPSA and 1SPSA.

The relationships among 1SPSA, 2SPSA and M2SPSA can also be understood from a different perspective: 1SPSA (1) and M2SPSA (13) 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 eigen-direction (greater λ_i) requires a smaller step (~$1/\lambda_i$) to effectively reach the exact solution (e.g., Pierre, 1986, 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 (13) 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 convergence rate: the elimination of the matrix inverse that accelerates or the lacking of gradient sensitivity that decelerates. The asymptotic relation (24) provides a theoretical rationale of adopting M2SPSA over 2SPSA. The rationale of proposing M2SPSA at finite iterations is the fact 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., Pierre, 1986, p. 308). Further justification for M2SPSA over 2SPSA at finite iterations is given in the numerical experiments in the next section.

We have shown that the convergence rate of 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 (13) 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 convergence rate of 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 $p = 10$

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

The minimum occurs at $\theta^* = 0$ with $L(\theta^*) = 0$. A non-negative noise is added to the loss function to represent the measurement errors: $y(\theta) = L(\theta) + |N(0,\sigma^2)|$, where $N(0,\sigma^2)$ is a normal distribution with zero mean and σ^2 variance. Note that, consistent with the case here, the regularity conditions for 2SPSA in Spall (2000) do not require mean-zero noise. The matrix elements of the Hessian are specified according to

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

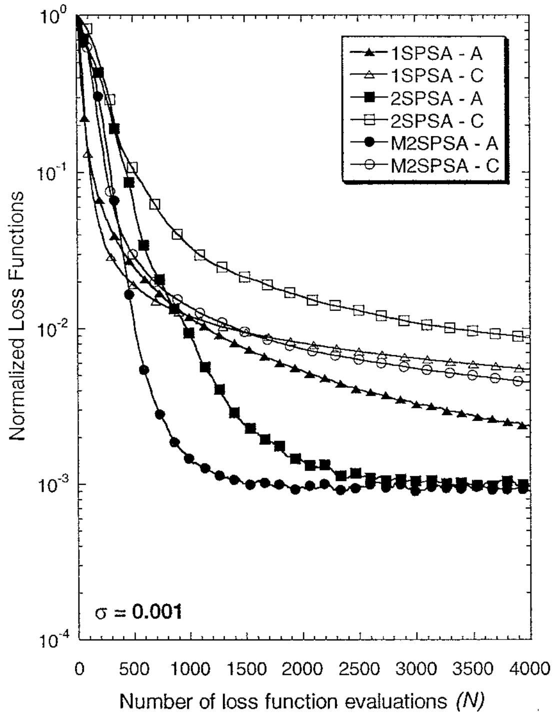

Figure 1. Normalized loss functions versus the number of loss function evaluations for 1SPSA (triangles), 2SPSA (squares), and M2SPSA (circles). The matrix condition numbers for Cases A (filled symbols) and C (open symbols) are 10 and 1000, respectively. The noise level $\sigma = 0.001$.

The following four cases are considered for numerical studies.

Case A: $\beta = 0.1291$, $\alpha = 1.1311$, $\kappa_{\lambda} = 10$; (31a)

Case B: $\beta = 0.2144$, $\alpha = 1.5416$, $\kappa_{\lambda} = 100$; (31b)

Case C: $\beta = 0.3941$, $\alpha = 1.9047$, $\kappa_{\lambda} = 1{,}000$; (31c)

Case D: $\beta = 0.7763$, $\alpha = 2.2597$, $\kappa_{\lambda} = 10{,}000$. (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 for different cases. Case D (with κ_{λ} = 10,000) is worse ill-conditioned than Case C, which in turn is worse ill-conditioned than Cases B and A.

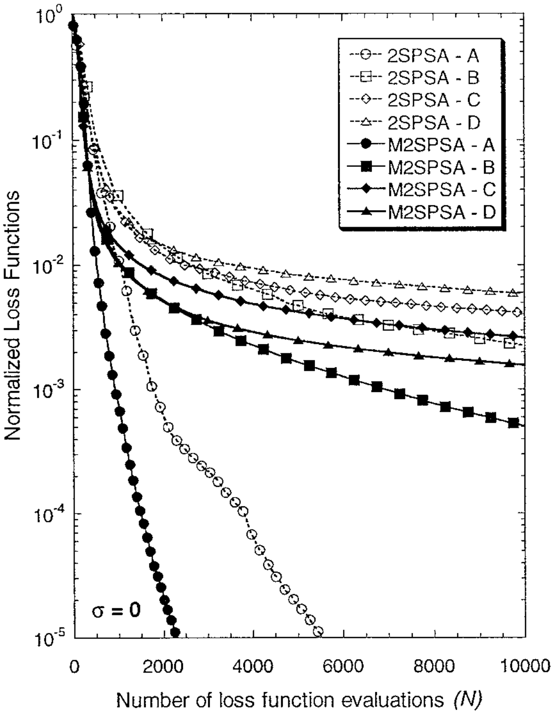

Figure 2. Normalized loss functions versus the number of loss function evaluations for 2SPSA (dashed lines) and M2SPSA (solid lines) and for all four cases of different matrix condition numbers. The noise level $\sigma = 0$.

Figure 1 shows the plots of averaged loss function versus the number of loss function evaluations (N) for two cases (A and C) with a noise level of $\sigma = 0.001$ after 50 independent experiments. All the loss functions are normalized to the initial $L(\hat{\theta}_1)$. We have followed the general guidance on picking gain series for 1SPSA (Spall, 1998). The figure shows that in the very early stage of iterations (say $N \le 400$) 1SPSA is better than both 2SPSA and M2SPSA since the estimated Hessian ($\overline{H}_k$) carries significant errors. As $\overline{H}_k$ becomes a better approximation of the real Hessian, 2SPSA based on (2a) and (8) outperforms 1SPSA in the chosen parameter setting when the matrix condition number is not extremely large. The results of Fig. 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 (13) and (14) is consistently better than 2SPSA in all cases, indicating a sound improvement of M2SPSA over 2SPSA based on the elimination of the matrix inversion errors. It is noted from Fig. 1 that the convergence rate of M2SPSA also depends on the matrix condition number that suggests a possible relation between errors in eigenvalue computation and matrix property such as its condition number. 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 Fig. 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.

5. CONCLUSIONS

We have made both empirical and theoretical comparisons between 1SPSA based on (1) and 2SPSA based on (2a) and (8) in the perspective of the loss function Hessian matrix. It is found that the additional errors introduced by matrix inversion in 2SPSA at finite iterations make the convergence rate more slowly for an ill-conditioned Hessian than a well-conditioned Hessian. On the other hand, the convergence rate of 1SPSA is less sensitive to the matrix conditioning of loss function Hessians. By analyzing the results for a special loss function Hessian, we show how the asymptotically optimal MSE for 2SPSA is only theoretically achievable by 1SPSA. To eliminate the errors introduced by matrix inversion we suggest a modification (13) to 2SPSA that replaces the Hessian inverse with a scalar inverse of the geometric mean of all the Hessian eigenvalues. At finite iterations, the newly introduced M2SPSA based on (13) and (14) consistently outperforms 2SPSA in the experimental cases representing a wide range of matrix conditioning. We also show how the asymptotically optimal MSE for 2SPSA might be practically achieved by M2SPSA. When the gain sequence does not take the optimal values, we show that the ratio of MSEs for M2SPSA to 2SPSA given by (24) is always less than one except for a perfectly conditioned Hessian.

ACKNOWLEDGMENT

The author would like to thank James C. Spall for introducing the SPSA method to him through the JHU/APL career enhancement course and for many insightful comments on this work. Discussions on the manuscript with Daniel C. Chin are also appreciated. This research was supported by the TIMED project sponsored by NASA under contract NAS5-97179 to the Johns Hopkins University Applied Physics Laboratory.

REFERENCES:

Chin, D. C., 1997: Comparative study of stochastic algorithms for system optimization based on gradient approximations. *IEEE Trans. Syst. Man Cybern, Part B*, **27**(2), 244-249

Horn, R. A. and C. R. Johnson, 1985: *Matrix analysis*. Cambridge University Press, Cambridge, 561 pp.

Luman, R. R., 2000: Upgrading complex systems of systems: A CAIV methodology for warfare area requirements allocation. *Mil. Oper. Res.*, **5**(2), 53-75.

Pierre, D. A., 1986: *Optimization Theory with Applications*. Dover Pub., Inc., New York, 612, pp.

Press, W. H., S. A. Teukolsky, and W. T. Vetterling, B. P. Flannery, 1992: *Numerical Recipes in Fortran. The Arts of Scientific Computing. 2nd Edition.* Cambridge University Press, 963 pp.

Spall, J. C., 1992: Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Autom. Control*, **37**, 332-341.

Spall, J. C., 1998: Implementation of the simultaneous perturbation algorithm for stochastic optimization. *IEEE Trans. Aerosp. Electron. Syst.*, **34(3)**, 817-823.

Spall, J. C., 2000: Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Trans. Autom. Control*, **45**, 1839-1853.

Spall, J. C., S. D. Hill, and D. R. Stark, 2000: Some theoretical comparisons of stochastic optimization approaches. *Proceedings of the American Control Conf.*, Chicago, IL, June 2000, pp. 1904-1908.