SPSA Algorithms with Measurement Reuse

Mohammed Shahid Abdulla Shalabh Bhatnagar

IISc-CSA-TR-2006-8
http://archive.csa.iisc.ernet.in/TR/2006/8/

Computer Science and Automation
Indian Institute of Science, India

August 2006

Online-companion for similarly titled submission towards Winter Simulation Conference, 2006

SPSA Algorithms with Measurement Reuse

Mohammed Shahid Abdulla* Shalabh Bhatnagar

ABSTRACT

Five algorithms, all variants of Simultaneous Perturbation Stochastic Approximation (SPSA), are proposed. The original one-measurement SPSA uses an estimate of the gradient of objective function L containing an additional bias term not seen in two-measurement SPSA. As a result, the asymptotic covariance matrix of the iterate convergence process has a bias term. We propose a one-measurement algorithm that eliminates this bias, and has asymptotic convergence properties making for easier comparison with the two-measurement SPSA. The algorithm, under certain conditions, outperforms both forms of SPSA with the only overhead being the storage of a single measurement. We also propose a similar algorithm that uses perturbations obtained from normalized Hadamard matrices. The convergence w.p. 1 of both algorithms is established. We extend measurement reuse to design three second-order SPSA algorithms and sketch the convergence analysis. Finally, we present simulation results on an illustrative minimization problem.

1 INTRODUCTION

Simultaneous Perturbation Stochastic Approximation (SPSA) is an efficient parameter optimization method that operates under the constraint that only noisy measurements of the objective function L are available at each parameter iterate θ_k. First proposed in (Spall 1992), it involves making only two measurements of L at each update epoch k that are obtained by perturbing θ_k along random directions. A plethora of applications and enhancements of this technique can be found at (Spall 2001). A variant of SPSA that reduces the number of function measurements made at each iteration k from two to one and establishes the conditions under which a net lower number of observations suffice to attain the same Mean Square Error (MSE) is provided in (Spall 1997). However, an impediment in rapid convergence to θ^* is that the algorithm constructs a gradient estimate of L at θ_k that contains an additional error term over the scheme in (Spall 1992) and that contributes heavily to the bias in the estimate. A solution to this problem was proposed in (Bhatnagar, Fu, Marcus, and Wang 2003) in the simulation optimization setting, where the perturbation to θ_k in the one-simulation case is generated in a deterministic manner. While this algorithm performs considerably better in practice, the asymptotic convergence properties in the setting of (Spall 1992) and (Spall 1997) were derived in (Xiong, Wang, and Fu 2002) and found to be on par with those of one-measurement SPSA.

In this work, we first propose two first-order algorithms: one using randomly generated perturbations (cf. Section 2) and the other using deterministic perturbations (cf. Section 4). We show convergence w.p. 1 for both the algorithms. For the first algorithm, we also derive the asymptotic convergence properties and compare these with (Spall 1992) (cf. Section 3). Further, we design three second-order algorithms based on the measurement-storage concept in Section 5 A numerical example is used to justify our findings (cf. Section 6).

The general structure of gradient descent algorithms is as follows. Suppose $\theta_k \stackrel{\Delta}{=} (\theta_{k,1}, ..., \theta_{k,p})^T$ where $\theta_{k,i}$, $1 \leq i \leq p$, are the p components of parameter θ_k. Let $G_{k,l}$ be an estimate of the $l-$th partial derivative of the cost $L(\theta_k)$, $l \in \{1, 2, ..., p\}$. Then,

$$\theta_{k+1,l} = \theta_{k,l} - a_k G_{k,l}(\theta_k), 1 \leq l \leq p, k \geq 0, \quad (1)$$

where $\{a_k\}$ is a step-size sequence. In the following, we refer to the one-measurement form of SPSA as SPSA2-1R and the two-measurement form as SPSA2-2R following the convention of (Bhatnagar, Fu, Marcus, and Wang 2003). In such a convention, the 'R' refers to perturbations which are randomly obtained, in contrast to deterministic perturbations in §3. The trailing '1' in SPSA2-1R refers to the fact that at each iteration, the algorithm makes one measurement. The leading '2' stands for a variant of the algorithm that makes parameter updates at every epoch, in contrast to algorithms like SPSA1-2R which update the parameter after an (increasingly large) number of epochs. The current parameter estimate θ_k is perturbed with a vector $\Delta_k = (\Delta_{k,1}, ..., \Delta_{k,p})^T$ to produce $\theta_k^+ = \theta_k + c_k\Delta_k$, where c_k is a small step-size parameter that satisfies Assumption 1 (below) together with the step-size parameter a_k in (1). The gradient estimates $G_{k,l}(\theta_k)$ used in SPSA2-1R are: $G_{k,l}(\theta_k) = \frac{L(\theta_k^+)+\epsilon_k^+}{c_k\Delta_{k,l}}$

$$= \frac{L(\theta_k)}{c_k\Delta_{k,l}} + g_l(\theta_k) + \sum_{i=1,i\neq l}^{p} g_i(\theta_k)\frac{\Delta_{k,i}}{\Delta_{k,l}} + \frac{c_k^2\Delta_k^T H(\theta_k)\Delta_k}{2c_k\Delta_{k,l}} + \frac{\epsilon_k^+}{c_k\Delta_{k,l}} + \frac{c_k^3 L^{(3)}(\bar\theta_k)\Delta_k\otimes\Delta_k\otimes\Delta_k}{6c_k\Delta_{k,l}}. \quad (2)$$

We assume here that L is twice continuously differentiable with bounded third derivative. Note that $H(\theta_k)$ is the Hessian evaluated at θ_k and $L^{(3)}(\bar\theta_k)\Delta_k\otimes\Delta_k\otimes\Delta_k = \Delta_k^T(L^{(3)}(\bar\theta_k)\Delta_k)\Delta_k$ where $\bar\theta_k = \theta_k + \lambda_k c_k\Delta_k$ for some $0 \le \lambda_k \le 1$ and $L^{(3)}$ is the third derivative of objective function $L(\cdot)$, where $\otimes$ denotes the Kronecker product. Also, ϵ_k^+ corresponds to a random additive noise component signifying that noisy observations of the cost are available for each parameter update. Thus, $G_{k,l}$ is a random variable that we assume is measurable w.r.t. the $\sigma-$algebra $\mathcal{F}_k = \sigma(\theta_i, \Delta_i, 0 \le i \le k-1, \theta_k)$. In contrast, $g_l(\theta_k)$ is the $l-$th component of the derivative of $L(\cdot)$ at θ_k. The gradient of $L(\cdot)$ at θ_k is now defined as $g(\theta_k) = (g_l(\theta_k),\ 1 \le l \le p)^T$. Although not the current object of study, it is illustrative to observe that the estimate of SPSA2-2R needs two measurements of $L(\cdot)$ about θ_k: $G_{k,l}(\theta_k) = \frac{L(\theta_k^+)+\epsilon_k^+-L(\theta_k^-)-\epsilon_k^-}{2c_k\Delta_{k,l}}$. In the above, $G_{k,l}$ uses function measurements at both θ_k^+ and $\theta_k^-{=}\theta_k - c_k\Delta_k$ and the measurement noise values at these points are ϵ_k^+ and ϵ_k^-, respectively.

We retain all assumptions of (Spall 1997), most of which are carried over from (Spall 1992). As in (Spall 1997), the key assumption requires the measurement noise ϵ_k^+ to be mean 0: $E(\epsilon_k^+|\theta_k, \Delta_k) = 0,\ \forall k$, and $var(\epsilon_k^+) \to \sigma_\epsilon^2$, where σ_ϵ^2 is some finite constant. The step-size sequences used are of the form $a_k = ak^{-\alpha}$ and $c_k = ck^{-\gamma}$, respectively, where $k \ge 1$, $a, c > 0$, are given constants and with constraints on $0 < \gamma, \alpha \le 1$ such that the following assumption holds

Assumption 1 $\sum_{k=0}^{\infty} a_k = \infty$ *and* $\sum_{k=0}^{\infty}\frac{a_k^2}{c_k^2} < \infty$.

2 ALGORITHM SPSA2-1UR

The proposed algorithm also has a similar structure as SPSA2-1R and we call this algorithm SPSA2-1UR, the alphabet 'U' indicating 'unbiased'. We utilize the noisy measurement already made at θ_{k-1}^+, the storage of which results in unit space complexity.

Algorithm 1 (SPSA2-1UR)

$$\theta_{k+1,l} = \theta_{k,l} - a_k\frac{L(\theta_k^+)+\epsilon_k^+-L(\theta_{k-1}^+)-\epsilon_{k-1}^+}{c_k\Delta_{k,l}},$$

where $k \ge 0$, $1 \le l \le p$. We have in the above $G_{k,l}(\theta_k) = \frac{L(\theta_k)-L(\theta_{k-1})}{c_k\Delta_{k,l}} + g_l(\theta_k) + \sum_{i=1,i\neq l}^{p} g_i(\theta_k)\frac{\Delta_{k,i}}{\Delta_{k,l}}$

$$- \sum_{i=1}^{p}\frac{c_{k-1}}{c_k}g_i(\theta_{k-1})\frac{\Delta_{k-1,i}}{\Delta_{k,l}} + \frac{c_k^2\Delta_k^T H(\theta_k)\Delta_k - c_{k-1}^2\Delta_{k-1}^T H(\theta_{k-1})\Delta_{k-1}}{2c_k\Delta_{k,l}} + \frac{c_k^3 L^{(3)}(\bar\theta_k)\Delta_k\otimes\Delta_k\otimes\Delta_k}{6c_k\Delta_{k,l}} - \frac{c_{k-1}^3 L^{(3)}(\bar\theta_{k-1})\Delta_{k-1}\otimes\Delta_{k-1}\otimes\Delta_{k-1}}{6c_k\Delta_{k,l}} + \frac{\epsilon_k^+ - \epsilon_{k-1}^+}{c_k\Delta_{k,l}}.$$

Using a similar analysis as in Proposition 2 of (Spall 1992), we identify below the order of convergence of the bias to 0:

Lemma 1 *Suppose for each $k > K$ for some $K < \infty$, $\{\Delta_{k,i}\}$ are i.i.d., symmetrically distributed about 0, with $\Delta_{k,i}$ independent of θ_j, ϵ_j^+, $1 \le j < k$. Further let $|\Delta_{k,i}| \le \beta_0$ a.s., $E|\Delta_{k,i}^{-1}| \le \beta_1$, and L be thrice continuously differentiable with $|L_{i_1,i_2,i_3}^{(3)}| \le \beta_2$, $\forall i_1, i_2, i_3 \in \{1, 2, ..., p\}$, for some constants β_0, β_1, and β_2. Then, $E\{G_{k,l}(\theta_k)|\mathcal{F}_k\} = g_l(\theta_k) + O(c_k^2)$, $1 \le l \le p$, a.s.*

Proof: The bias vector $b_k(\theta) = (b_{k,1}(\theta), b_{k,2}(\theta), ..., b_{k,p}(\theta))^T$ is defined as:

$$b_k(\theta_k) = E\{G_k(\theta_k) - g(\theta_k)|\theta_k\}, \quad (3)$$

where $G_k(\theta_k) = (G_{k,1}(\theta_k), G_{k,2}(\theta_k), ..., G_{k,p}(\theta_k))^T$. Due to the mean-zero assumption on ϵ_k^+ and Δ_k^{-1} w.r.t $\mathcal{F}_k$, we have $E\{\frac{\epsilon_k^+-\epsilon_{k-1}^+}{\Delta_{k,l}}|\mathcal{F}_k\} = 0$. It is crucial here to note that, despite the previous equality, $E\{\epsilon_k^+ - \epsilon_{k-1}^+|\mathcal{F}_k, \Delta_k\} \neq 0$. Further, using the properties of independence, symmetry and finite inverse moments of perturbation vector elements (viz. $\Delta_{k,l}$, $\Delta_{k,i}$, and $\Delta_{k-1,i}$), observe that terms on the RHS of (3) have zero mean, with the exception of

$$b_{k,l}(\theta_k) = E\left\{\frac{c_k^3 L^{(3)}(\bar\theta_k)\Delta_k\otimes\Delta_k\otimes\Delta_k}{6c_k\Delta_{k,l}}|\mathcal{F}_k\right\}. \quad (4)$$

Observe that the bias term here is the same as for SPSA2-1R. The claim is now obtained by the arguments following Equation (3.1) in Lemma 1 of (Spall 1992). In particular, note that

$$|b_{k,l}(\theta_k)| \le \frac{\beta_2 c_k^2}{6}\sum_{i_1}\sum_{i_2}\sum_{i_3} E\left|\frac{\Delta_{k,i_1}\Delta_{k,i_2}\Delta_{k,i_3}}{\Delta_{k,l}}\right| \le \frac{\beta_2 c_k^2}{6}\left([p^3 - (p-1)^3]\beta_0^2 + (p-1)^3\beta_1\beta_0^3\right)$$

□

The relation in Lemma 1 is of value in establishing a form of asymptotic normality of the scaled iterate convergence process, see Section 3. It is interesting to note that Lemma 1 will not hold if normalized Hadamard matrix-based $\{\pm 1\}^p$−valued perturbations Δ_k that were first introduced in (Bhatnagar, Fu, Marcus, and Wang 2003) (Section 4 below explains this deterministic perturbation method in some detail). This is because there is no assurance that the term $\sum_{i=1}^{p} \frac{c_{k-1}}{c_k} g_i(\theta_{k-1}) \frac{\Delta_{k-1,i}}{\Delta_{k,l}}$ will average to 0 as $k \to \infty$, unlike the previous term $\sum_{i=1, i\neq l}^{p} g_i(\theta_k)\frac{\Delta_{k,i}}{\Delta_{k,l}}$. In such a case, a different method for unbiasing that does not use the immediate past measurement, in the spirit of Section 4 later, would be appropriate. A consequence of the a.s. convergence of the bias $b_k(\theta_k)$ is the strong convergence of the iterates θ_k to a local minimum θ^*. We now state Assumption A2 of (Spall 1992) (that is also applicable to the setting of (Spall 1997)):

Assumption 2 *$\exists \alpha_0, \alpha_1, \alpha_2 > 0$ and $\forall k$, $E{\epsilon_k^+}^2 \leq \alpha_0$, $EL^2(\theta_k^+) \leq \alpha_1$, and $\Delta_{k,l}^{-2} \leq \alpha_2$ a.s., for $1 \leq l \leq p$.*

While this does not entail any difference, observe that we use $\Delta_{k,l}^{-2} \leq \alpha_2$ a.s. instead of the original $E\{\Delta_{k,l}^{-2}\} \leq \alpha_2$ in (Spall 1992).

Lemma 2 *Under assumptions of (Spall 1997), as $k \to \infty$: $\theta_k \to \theta^*$ a.s.*

Proof: Follows almost verbatim as Proposition 1 of (Spall 1992). The only modifications are due to a different error process e_k, defined as $e_k(\theta_k) = G_k(\theta_k) - E(G_k(\theta_k)|\theta_k)$. We can thus rewrite recursion (1) as: $\theta_{k+1} = \theta_k - a_k(g(\theta_k) + b_k(\theta_k) + e_k(\theta_k))$. The claim is obtained if the following conditions are satisfied:

(a) $\|b_k(\theta_k)\| < \infty$, $\forall k$ and $b_k(\theta_k) \to 0$ a.s.

(b) $\lim_{k\to\infty} P(\sup_{m\geq k} \|\sum_{i=k}^{m} a_i e_i(\theta_i)\| \geq \eta) = 0$, for any $\eta > 0$.

where $\|\cdot\|$ represents the Euclidean norm in parameter space $\mathcal{R}^p$. Lemma 1 establishes (a) whilst for (b), notice that $\{\sum_{i=k}^{m} a_i e_i\}_{m\geq k}$ is a martingale sequence (since $E(e_{i+1}|\mathcal{F}_i) = 0$) and the martingale inequality gives: $P(\sup_{m\geq k} \|\sum_{i=k}^{m} a_i e_i(\theta_i)\| \geq \eta) \leq \eta^{-2} E\|\sum_{i=k}^{\infty} a_i e_i\|^2$

$$= \eta^{-2} \sum_{i=k}^{\infty} a_i^2 E\|e_i\|^2,$$

since $E(e_i^T e_j) = E(e_i^T E(e_j|\theta_j)) = 0$, $\forall j \geq i+1$. Further, for $1 \leq l \leq p$ using Hölder's inequality: $E\left(G_{i,l}^2(\theta_i)\right) \leq E(L(\theta_i^+) - L(\theta_{i-1}^+) + \epsilon_i^+ - \epsilon_{i-1}^+)^2 \cdot \frac{\|\Delta_{i,l}^{-2}(\omega)\|_\infty}{c_i^2} \leq 2(\alpha_1+\alpha_0)\alpha_2 c_i^{-2}$. Due to the mean-zero property of $e_{i,l}(\theta_i)$, we have $E\left(G_{i,l}^2(\theta_i)\right) = (g_l(\theta_i) + b_{i,l}(\theta_i))^2 + E(e_{i,l}^2(\theta_i))$, thus having $E(e_{i,l}^2(\theta_i)) \leq E(G_{i,l}^2(\theta_i))$, and resulting in $E\|e_i\|^2 \leq 2p(\alpha_1+\alpha_0)\alpha_2 c_i^{-2}$. The square summability of $\frac{a_k}{c_k}$, from Assumption 1, now establishes (b). □

3 ASYMPTOTIC NORMALITY AND COMPARISON

The results obtained so far aid us in establishing the asymptotic normality of a scaled iterate convergence process. We show that

$$k^{\frac{\beta}{2}}(\theta_k - \theta^*) \quad \xrightarrow{D} \quad N(\mu, P\tilde{M}_1 P^T)$$

as $k \to \infty$ where the indicated convergence is in distribution, $\beta = \alpha - 2\gamma > 0$ (given $3\gamma - \frac{\alpha}{2} \geq 0$), and the mean μ is the same as in SPSA2-2R (Spall 1992, Proposition 2) and SPSA2-1R. The orthogonal matrix P above is s.t. $P^T aH(\theta^*)P = \mathrm{Diag}(\{\lambda_l\}_{l=1}^p)$, $\lambda_1, ..., \lambda_p$ being the p eigen values of $aH(\theta^*)$. Unlike (Spall 1997), $\tilde{M}_1$ above does not have an $L^2(\theta^*)$ bias; however, it is scaled by a factor of 2. This factor arises due to the use of the additional noisy measurement $L(\theta_{k-1}^+) + \epsilon_k^+$ in (3). In particular, $\tilde{M}_1 = 2a^2c^{-2}\rho^2\sigma_\epsilon^2 \mathrm{Diag}(\{(2\lambda_l - \beta_+)^{-1}\}_{l=1}^p)$, where $\beta_+ = \beta$ if $\alpha = 1$ and 0 otherwise, and $E\Delta_{k,l}^{-2} \to \rho^2$. However, as confirmed by (Spall 2005), the M_1 in (5) of (Spall 1997) should be

$$M_1 = a^2c^{-2}\rho^2(\sigma_\epsilon^2 + L^2(\theta^*))\mathrm{Diag}(\{(2\lambda_l - \beta_+)^{-1}\}_{l=1}^p), \tag{5}$$

and not $a^2c^{-2}\rho^2(\sigma_\epsilon^2 \mathrm{Diag}(\{(2\lambda_l - \beta_+)^{-1}\}_{l=1}^p) + L^2(\theta^*)I)$ as printed, see Appendix 1 for a cursory derivation of M_1. Similarly, Appendix 2 establishes the form of $\tilde{M}_1$ and mean μ that we claim.

We compare the proposed SPSA2-1UR with SPSA2-2R in the number of measurements of cost function L, viz. $\tilde{n}_1$ and n_2, respectively. As in (Spall 1997), we consider the case $\alpha = 1$ and $\gamma = \frac{1}{6}$ (giving $\beta = \frac{2}{3}$) and $E((\epsilon_k^+ - \epsilon_k^-)^2|\theta_k, \Delta_k) = 2\sigma_\epsilon^2$, resulting in $M_2 = \frac{1}{2}a^2c^{-2}\rho^2\sigma_\epsilon^2 \mathrm{Diag}(\{(2\lambda_l - \beta_+)^{-1}\}_{l=1}^p)$ and $\tilde{M}_1 = 4M_2$. This gives us:

$$\frac{\tilde{n}_1}{n_2} \to \frac{1}{2}\left(\frac{4\mathrm{tr}PM_2P^T + \mu^T\mu}{\mathrm{tr}PM_2P^T + \mu^T\mu}\right)^{\frac{3}{2}}, \tag{6}$$

where tr stands for the trace of the matrix. The ratio above depends upon quantity μ and to achieve $\tilde{n}_1 < n_2$ we need that $\mu^T\mu > \left(\frac{4 - 2^{\frac{2}{3}}}{2^{\frac{2}{3}} - 1}\right)\mathrm{tr}PM_2P^T \approx 4.11\mathrm{tr}PM_2P^T$. We use n_1 to denote the number of measurements of L made by SPSA2-1R. In the special case $L(\theta^*) = 0$, it is shown in (Spall 1997, eq. (8)) that $n_1 < n_2$ when $\mu^T\mu > 0.7024\mathrm{tr}PM_2P^T$. While our result does not compare favorably, the advantage is that (6) holds for all values of $L(\theta^*)$.

The comparison with SPSA2-1R yields an interesting rule of thumb. Using D_λ to represent the diagonal matrix $\mathrm{Diag}(\{2\lambda_l - \beta_+\}_{l=1}^{p})$, we have:

$$\begin{aligned}\frac{\tilde{n}_1}{n_1} &\to \left(\frac{\mathrm{tr}P\tilde{M}_1P^T+\mu^T\mu}{\mathrm{tr}PM_1P^T+\mu^T\mu}\right)^{\frac{3}{2}} \\ &= \left(\frac{2a^2c^{-2}\rho^2\sigma_\epsilon^2\mathrm{tr}PD_\lambda P^T+\mu^T\mu}{a^2c^{-2}\rho^2(\sigma_\epsilon^2+L^2(\theta^*))\mathrm{tr}PD_\lambda P^T+\mu^T\mu}\right)^{\frac{3}{2}}.\end{aligned}$$

Irrespective of μ, D_λ, and P (quantities that may require substantial knowledge of the system), it suffices to have $L^2(\theta^*) > \sigma_\epsilon^2$ to ensure that $\frac{\tilde{n}_1}{n_1} \le 1$. The experimental results in Section 6 provide verification of these claims.

4 ALGORITHM SPSA2-1UH

We now propose a fast convergence algorithm by modifying SPSA2-1H of (Bhatnagar, Fu, Marcus, and Wang 2003, §3). The key departure in SPSA2-1H from gradient estimate (2) of SPSA2-1R is that perturbation vectors Δ_k are now deterministically obtained from normalized Hadamard matrices. The kind of matrices considered are the following: Let H_2 be a 2×2 matrix with elements $H_2(1,1) = H_2(1,2) = H_2(2,1) = 1$ and $H_2(2,2) = -1$. Likewise for any $q > 1$, let the block matrices $H_{2^q}(1,1)$, $H_{2^q}(1,2)$, and $H_{2^q}(2,1)$ equal $H_{2^{q-1}}$. Also, let $H_{2^q}(2,2) = -H_{2^{q-1}}$. For a parameter of dimension p, the dimension of the Hadamard matrix needed is 2^q where $q = \lceil \log_2(p+1)\rceil$. Next, p columns from the above matrix H_q are arbitrarily chosen from the $q-1$ columns that remain after the first column is removed. The latter column is removed as it does not satisfy a key property of the perturbation sequence. Each row of the resulting $q \times p$ matrix $\hat{H}$ is now used for the perturbation vector Δ_k in a cyclic manner, i.e. $\Delta_k = \hat{H}(k\%q+1)$, where % indicates the modulo operator. Though not shown here, the convergence of SPSA2-1H can be shown as a special case of Prop. 2.5 of (Xiong, Wang, and Fu 2002). The proposed algorithm, which we call SPSA2-1UH, has two steps:

Algorithm 2 (SPSA2-1UH)

1. *For* $k \ge 0$, $1 \le l \le p$,
$$\theta_{k+1,l} := \theta_{k,l} - a_k \frac{L(\theta_k^+)+\epsilon_k^+ - \bar{L}_k}{c_k\Delta_{k,l}}$$
2. *if* $k\%q = 0$, $\bar{L}_k := L(\theta_k^+) + \epsilon_k^+$ *else* $\bar{L}_{k+1} := \bar{L}_k$.

In the above, $\bar{L}_k$ changes only periodically in epochs of size q and the algorithm has a unit space requirement. Given index k, define $\bar{k} = \max\{m : m < k, m\%q = 0\}$. For SPSA2-1UH, (2) is now modified to:

$$\begin{aligned}G_{k,l}(\theta_k) &= \frac{L(\theta_k)-L(\theta_{\bar{k}})}{c_k\Delta_{k,l}} + g_l(\theta_k) \\ &\quad + \sum_{i=1,i\neq l}^{p}\frac{\Delta_{k,i}}{\Delta_{k,l}}g_i(\theta_k) - \sum_{i=1}^{p}\frac{c_{\bar{k}}}{c_k}\frac{\Delta_{\bar{k},i}}{\Delta_{k,l}}g_i(\theta_{\bar{k}}) \\ &\quad + O(c_k) + \frac{\epsilon_k^+ - \epsilon_{\bar{k}}^+}{c_k\Delta_{k,l}},\end{aligned}$$

where $O(c_k)$ contains higher order terms. Since $\Delta_{\bar{k}} = \underline{1}$, we have $\frac{\Delta_{\bar{k},i}}{\Delta_{k,l}} = \frac{1}{\Delta_{k,l}}$, $\forall 1 \le i,l \le p$ and $\forall k$. Therefore, it can be shown as in Lemma 3.5 of (Bhatnagar, Fu, Marcus, and Wang 2003) that the fourth term above averages to 0 over q steps as $k\to\infty$, thus settling the problem posed in §2. In passing, we also note that step 2 can be written as $k\%q = m$ for any given m s.t. $0 \le m \le q-1$.

4.1 Convergence Analysis

We can now formally establish convergence w.p. 1 of θ_k. The original SPSA2-1H algorithm can be expanded as follows

$$\begin{aligned}\theta_{k+1} &= \theta_k - a_k\Delta_k^{-1}\Delta_k^T g(\theta_k+\lambda_k c_k\Delta_k) \\ &\quad - \frac{a_k}{c_k}L(\theta_k)\Delta_k^{-1} - \frac{a_k}{c_k}\epsilon_k^+\Delta_k^{-1}, \qquad (7)\end{aligned}$$

where $0 \le \lambda_k \le 1$. Here Δ_k^{-1} is the vector $\Delta_k^{-1} = \left(\frac{1}{\Delta_{k,1}},\ldots,\frac{1}{\Delta_{k,p}}\right)^T$. This recursion is now presented in the manner of (Xiong, Wang, and Fu 2002, eq. (6)), with r_k, d_k and e_k^+ there replaced by Δ_k^{-1}, Δ_k and ϵ_k^+, respectively:

$$\begin{aligned}\theta_{k+1} &= \theta_k - a_k g(\theta_k) \\ &\quad - a_k\Delta_k^{-1}\Delta_k^T\{g(\theta_k+\lambda_k c_k\Delta_k) - g(\theta_k)\} \\ &\quad - a_k\{\Delta_k^{-1}\Delta_k^T - I\}g(\theta_k) - \frac{a_k}{c_k}L(\theta_k)\Delta_k^{-1} \\ &\quad - \frac{a_k}{c_k}\epsilon_k^+\Delta_k^{-1}.\end{aligned}$$

In the manner of (7), the SPSA2-1UH recursion is written as: $\theta_{k+1} = \theta_k - a_k\Delta_k^{-1}\Delta_k^T g(\theta_k+\lambda_k c_k\Delta_k) - \frac{a_k}{c_k}(L(\theta_k) - L(\theta_{\bar{k}}^+))\Delta_k^{-1} - \frac{a_k}{c_k}(\epsilon_k^+ - \epsilon_{\bar{k}}^+)\Delta_k^{-1}$ which can be expanded as $\theta_{k+1} = \theta_k - a_k g(\theta_k)$

$$\begin{aligned}&- a_k\Delta_k^{-1}\Delta_k^T\{g(\theta_k+\lambda_k c_k\Delta_k) - g(\theta_k)\} \\ &- a_k\{\Delta_k^{-1}\Delta_k^T - I\}g(\theta_k) \qquad (8) \\ &+ a_k\Delta_k^{-1}\Delta_{\bar{k}}^T\{g(\theta_{\bar{k}}+\lambda_{\bar{k}}c_{\bar{k}}\Delta_{\bar{k}}) - g(\theta_{\bar{k}})\} \\ &+ a_k\Delta_k^{-1}\Delta_{\bar{k}}^T g(\theta_{\bar{k}}) \qquad (9) \\ &- \frac{a_k}{c_k}\left(L(\theta_k) - L(\theta_{\bar{k}}) - \epsilon_k^+ + \epsilon_{\bar{k}}^+\right)\Delta_k^{-1} \qquad (10)\end{aligned}$$

However, we need to make a non-restrictive assumption:

Assumption 3 *The function g (cf. A1 of (Xiong, Wang, and Fu 2002)) is uniformly continuous.*

Theorem 1 *Under Assumptions from (Spall 1997) and 3, Algorithm 2 produces iterates θ_k s.t. $\theta_k \to \theta^*$ w.p. 1.*

Proof: We first show that terms in (9) and (10) are error terms in the nature of $e_i(\theta_i)$ in condition (b) of Lemma 2. In particular, we show that these satisfy the conditions (B1) and (B4) in (Xiong, Wang, and Fu 2002). We reproduce these two conditions for clarity:

(B1) $\lim_{n\to\infty}\left(\sup_{n\le k\le m(n,T)}\|\sum_{i=n}^{k} a_i e_i\|\right) = 0$, for some $T > 0$, where $m(n,T)\triangleq\max\{k : a_n+...+a_k\le T\}$.

(B4) There exist sequences $\{e_{1,n}\}$ and $\{e_{2,n}\}$ with $e_n = e_{1,n}+e_{2,n}$ for all n such that $\sum_{k=1}^{n} a_k e_{1,k}$ converges, and $\lim_{n\to\infty} e_{2,n}=0$.

Observe that due to $\lim_{k\to\infty} c_{\bar{k}}=0$ and the uniform continuity of g, $\Delta_{\bar{k}}^{-1}\Delta_{\bar{k}}^{T}\{g(\theta_{\bar{k}}+\lambda_{\bar{k}}c_{\bar{k}}\Delta_{\bar{k}})-g(\theta_{\bar{k}})\}$ satisfies (B4). Since $\lim_{k\to\infty}\theta_{\bar{k}}-\theta_{\overline{k+1}}=0$, $\Delta_{\bar{k}}^{-1}\Delta_{\bar{k}}^{T}g(\theta_{\bar{k}})$ satisfies (B1). This is shown by applying Lemma 2.2 of (Xiong, Wang, and Fu 2002) with the substitution $\{x_n\}$ where $x_n = 1\ \forall n\ge 1$, $\{\Delta_n^{-1}\Delta_{\bar{n}}\}$ and $\{g(\theta_{\bar{n}})\}$ for $\{c_n\}$, $\{r_n\}$, and $\{e_n\}$, respectively. We have $\forall k$, $\frac{|(L(\theta_k)-L(\theta_{\bar{k}}))-(L(\theta_{k+1})-L(\theta_{\overline{k+1}}))|}{c_k}\le\frac{|L(\theta_k)-L(\theta_{\bar{k}})|}{c_k}I_{k\%q=0}+\frac{|L(\theta_k)-L(\theta_{k+1})|}{c_k}$. We consider the first term on the RHS, the second follows similarly: $\frac{|L(\theta_k)-L(\theta_{\bar{k}})|}{c_k}I_{k\%q=0}\le\frac{M_0}{c_k}\sum_{m=\bar{k}}^{k-1}\|\theta_{m+1}-\theta_m\|I_{k\%q=0}+\frac{M_0}{c_k}\sum_{m=\bar{k}}^{k-1}\left(M_1\frac{a_m}{c_m}+M_2\frac{a_m}{c_m}|\epsilon_m^{+}|\right)I_{k\%q=0}\le\left(\frac{M_0M_1q}{c_k}\frac{a_{\bar{k}}}{c_{\bar{k}}}+\frac{M_0M_2}{c_k}\frac{a_{\bar{k}}}{c_{\bar{k}}}\sum_{m=\bar{k}}^{k-1}|\epsilon_m^{+}|\right)I_{k\%q=0}$, where M_0, M_1, and M_2 represent appropriate bounds. The summability of $\{\frac{a_k a_{\bar{k}}}{c_k c_{\bar{k}}}\}$ is obtained using Assumption 1 - implying that the LHS satisfies (B1). This fact is used when we apply Lemma 2.2 of (Xiong, Wang, and Fu 2002) again (with $\{\Delta_n^{-1}\}$, $\{L(\theta_n)-L(\theta_{\bar{n}})\}$ replacing $\{r_n\}$ and $\{e_n\}$, respectively, and $\{c_n\}$ as is) to see that $\frac{L(\theta_k)-L(\theta_{\bar{k}})}{c_k}\Delta_k^{-1}$ satisfies (B1). We now consider the last term viz. $\frac{\epsilon_k^{+}-\epsilon_{\bar{k}}^{+}}{c_k}\Delta_k^{-1}$. However, now the noise term $\bar{\epsilon}_{k,l}=\frac{\epsilon_k^{+}-\epsilon_{\bar{k}}^{+}}{\Delta_{k,l}}$ is not mean 0 w.r.t. $\mathcal{F}_k$ but letting $\tilde{k}=k+q$, $\forall k$ we see that $E(\bar{\epsilon}_{\tilde{k},l}|\mathcal{F}_k)=0$. This results in $\left\{\sum_{k=\tilde{n}}^{m}\frac{a_k}{c_k}\bar{\epsilon}_k\right\}_{m\ge\tilde{n}}$ being a martingale sequence w.r.t. $\mathcal{F}_n$, where we again utilize the inequality

$$P\left(\sup_{m\ge\tilde{n}}\left\|\sum_{k=\tilde{n}}^{m}\frac{a_k}{c_k}\bar{\epsilon}_k\right\|\ge\eta\right)\le\eta^{-2}\sum_{k=\tilde{n}}^{\infty}\left(\frac{a_k}{c_k}\right)^2 E\|\bar{\epsilon}_k\|^2,$$

the LHS modified to obtain $\le\eta^{-2}\sum_{k=\tilde{n}}^{\infty}\left(\frac{a_k}{c_k}\right)^2 E\|\bar{\epsilon}_k\|^2+P\left(\sup_{\tilde{n}>m\ge n}\left\|\sum_{k=n}^{m}\frac{a_k}{c_k}\bar{\epsilon}_k\right\|\ge\eta\right)$. The square summability of $\frac{a_k}{c_k}$ and boundedness of $\bar{\epsilon}_k$ result in quantities on the RHS vanishing as $n\to\infty$. The proof of Proposition 2.3 in (Xiong, Wang, and Fu 2002) handles the terms in the RHS of (8), thus resulting in the claim. □

5 SECOND-ORDER ALGORITHMS

We now propose three second order SPSA algorithms, all of which re-use noisy function measurements. The first algorithm - called 2SPSA-3UR since it is a modification of 2SPSA of (Spall 2000) - makes three measurements in the vicinity of each iterate θ_k and re-uses the current gradient estimate $G_k(\theta_k)$ to estimate the Hessian matrix $H_k(\theta_k)$ at θ_k. The second algorithm 2SPSA-2UR makes two measurements at θ_k and reuses the value $L(\theta_{k-1}^{+})$ in the Hessian matrix estimation. The third algorithm (2SPSA-1UR) makes one measurement per iteration. Second-order SPSA algorithms, which are stochastic analogs of the Newton-Raphson algorithm, are also proposed in (Spall 2000) and (Bhatnagar 2005). In particular, the three algorithms that we propose can be seen as modifications of the algorithms 3SA, 2SA and 1SA in (Bhatnagar 2005) with the following properties:

- Unlike the 2SPSA in (Spall 2000), all three algorithms 2SPSA-nUR $n = 1,2,3$ use an additional a_k-like step-size sequence $\{b_k\}$ (not to be confused with the bias term $b_k(\theta_k)$ in Lemma 1) in the recursion to compute H_k. Such an additional step-size $\{b_k\}$ is employed in all the four second-order SPSA algorithms described in (Bhatnagar 2005). The property of b_k relative to a_k is the well-known 'two-timescale' property: $\sum_k b_k=\infty$, $\sum_k b_k^2<\infty$ and $a_k=o(b_k)$. In (Bhatnagar 2005), it is observed in the algorithms that Hessian should be averaged and its inverse taken on a faster timescale. Thus on the timescale on which the parameter θ_k is updated, the Hessian updates H_k would be seen as having converged. This provides good algorithmic behavior.

- Similar to 2SPSA, we employ an auxiliary perturbation sequence $\{\tilde{\Delta}_k\}$ with the same properties as the original $\{\Delta_k\}$, although independently generated. There is also an associated scaling parameter $\{\tilde{c}_k\}$. We will also require an analog of Assumption 1: replace the pair (a_k,c_k) in Assumption 1 with the pairs $(a_k,\tilde{c}_k)$, $(b_k,\tilde{c}_k)$, (b_k,c_k).

- We use the 'unbiasing' concept by storing past or current measurements of L and gradient estimate G. However, unlike the unit storage overhead in SPSA2-1UR and SPSA2-1UH, this retention of the current estimate of gradient G arguably costs $O(p)$ in storage. It must be mentioned here that second-order algorithms of (Spall 2000) and (Bhatnagar 2005) do implicitly assume memory to store, manipulate and multiply Hessian estimates H_k - which are $O(p^2)$ data structures.

- The algorithm 2SPSA-1UR is distinct from the single-simulation algorithm 1SA proposed in (Bhatnagar 2005) due to the use of the sum $\Delta_k + \tilde{\Delta}_k$ as perturbation in the gradient estimate G_k.

5.1 2SPSA-3UR

The proposed algorithm can be viewed as a modification of the 3SA three-measurement algorithm of (Bhatnagar 2005). Also used in (Bhatnagar 2005), the function Γ used below maps from the set of general $p \times p$ matrices to the set of positive definite matrices. There are many possible candidates for such a Γ, as explained in §II-D of (Spall 2000) where the notation f_k is used.

$$\begin{aligned} \theta_{k+1} &= \theta_k - a_k H_k^{-1} G_k(\theta_k) \qquad (11) \\ H_k &= \Gamma(\bar{H}_k) \\ \bar{H}_k &= \bar{H}_{k-1} + b_{k-1}(\hat{H}_k - \bar{H}_{k-1}) \end{aligned}$$

where

$$\begin{aligned} \hat{H}_k &= \frac{1}{2}\left[\frac{\delta G_k^T}{c_k \Delta_k} + \left(\frac{\delta G_k^T}{c_k \Delta_k}\right)^T\right] \\ \delta G_k &= G_k^1(\theta_k^+) - G_k(\theta_k). \end{aligned}$$

There is a re-use of the current gradient estimate $G_k(\theta_k)$ in the latter recursion. This estimate is computed as in the algorithm SPSA2-2R. In addition to θ_k^+ and θ_k^-, we now employ the shorthand notation $\theta_k^{++} = \theta_k + c_k\Delta_k + \tilde{c}_k\tilde{\Delta}_k$. Similarly, we denote the measurement noise incurred at θ_k^{++} as ϵ_k^{++}.

$$\begin{aligned} G_k^1(\theta_k^+) &= \frac{\tilde{\Delta}_k^{-1}}{\tilde{c}_k}\left(L(\theta_k^{++}) + \epsilon_k^{++} - L(\theta_k^+) - \epsilon_k^+\right) \\ G_k(\theta_k) &= \frac{\Delta_k^{-1}}{2c_k}\left(L(\theta_k^+) + \epsilon_k^+ - L(\theta_k^-) - \epsilon_k^-\right) \end{aligned}$$

Appendix 3 contains the derivation regarding $E(\hat{H}_k|\mathcal{F}_k) = H(\theta_k) + O(c_k)$. The convergence analysis of $\theta_k \to \theta^*$ proceeds as in (Bhatnagar 2005), outlines of which we explain here. We construct a time-axis using the step-size b_k: assume that $t(n) = \sum_{m=0}^{n} b_m$ and define a function $H(\cdot)$ as $H(t(k)) = H_k$ with linear interpolation between $[t(k), t(k+1))$. Similarly define function $\theta(\cdot)$ by setting $\theta(t(k)) = \theta_k$ and linear interpolation on the interval $[t(k), t(k+1)]$. Let $T > 0$ be a given scalar and next define a sequence $\{T_k\}$ as $T_0 = 0 = H(0)$ and $T_k = min\{t(m)|t(m) \geq T_{k-1} + T\}$. Then $T_k = t(m_k)$ for some m_k and $T_k - T_{k-1} = T$. Now define functions $\bar{H}(\cdot)$ and $\bar{\theta}(\cdot)$ as $\bar{H}(T_k) = H(t_{m_k}) = H_k$ and $\bar{\theta}(T_k) = \theta(t_{m_k}) = \theta_k$, and for $t \in [T_k, T_{k+1}]$, the evolution is according to the system of ODEs:

$$\begin{aligned} \dot{\bar{H}}_{i,j}(t) &= \nabla^2_{i,j} L(\bar{\theta}(t)) - \bar{H}_{i,j}(t) \\ \dot{\bar{\theta}}(t) &= 0, \end{aligned}$$

where $\nabla^2_{i,j}$ indicates $\frac{\partial^2 L(\bar{\theta})}{\partial\theta_i \partial\theta_j}$. One can now show as in Lemma A.8 of (Bhatnagar 2005) that $\lim_{k\to\infty} \sup_{t\in[T_k,T_{k+1}]} \|H(t) - \bar{H}(t)\| = 0$ and $\lim_{k\to\infty} \sup_{t\in[T_k,T_{k+1}]} \|\theta(t) - \bar{\theta}(t)\| = 0$. Recursion (11) can now be shown to asymptotically track the trajectories of the ODE

$$\dot{\theta}(t) = -H(\theta(t)) \nabla L(\theta(t))$$

in a similar manner as above on the slower timescale $\{a_k\}$ as in Theorem 3.1 of (Bhatnagar 2005).

5.2 2SPSA-2UR

The proposed algorithm is a modification of the gradient-free four-measurement algorithm 2SPSA which was described in (Spall 2000). In that article (cf. footnote 6), the SPSA2-1R analog of 2SPSA was not considered due to the inherent variability of both estimates G_k and H_k if the one-measurement form of SPSA were to be used. We employ the technique of the proposed SPSA2-1UR to unbias the estimates, and thus reduce the number of function measurement required from 4 to 2. The family of recursions (11) is retained with the differences that

$$\begin{aligned} G_k(\theta_k) &= \frac{\tilde{\Delta}_k^{-1}}{\tilde{c}_k}\left(L(\theta_k^{++}) + \epsilon_k^{++} - L(\theta_k^+) - \epsilon_k^+\right) \\ \hat{H}_k &= \frac{1}{2}\left[\frac{\delta G_k^T}{c_k \Delta_k} + \left(\frac{\delta G_k^T}{c_k \Delta_k}\right)^T\right] \\ \delta G_k &= G_k^1(\theta_k^+) - \tilde{G}_k^1(\theta_k), \end{aligned}$$

followed by a correction of the diagonal terms in $\hat{H}_k$:

$$\hat{H}_k(i,i) := \hat{H}_k(i,i) + \frac{L(\theta_k^+) + \epsilon_k^+ - L(\theta_{k-1}^+) - \epsilon_{k-1}^+}{c_k^2} \qquad (12)$$

where the measurement at θ_{k-1}^+ is reused. This correction assumes that Δ_k and $\tilde{\Delta}_k$ are both Bernoulli distributed over $\{+,-\}$, however, a similar corrective term can be derived for other classes of perturbations. Appendix 4 derives the steps leading to this correction. In the above, $\tilde{G}_k^1(\theta_k) = \frac{L(\theta_k^+) + \epsilon_k^+ - L(\theta_{k-1}^+) - \epsilon_{k-1}^+}{c_k}\Delta_k^{-1}$ and $G_k^1(\theta_k^+)$ is as in 2SPSA-3UR. Note that $\tilde{G}_k^1(\theta_k)$ is precisely the gradient estimate in the algorithm SPSA2-1UR of Section 2. Also, $G_k(\theta_k) = G_k^1(\theta_k^+)$ indicating that a re-use of the gradient estimate G_k is being made to compute the Hessian estimate $\hat{H}_k$. Here, G_k is computed using a one-sided difference just as in 2SA of (Bhatnagar 2005). Such an estimate still uses 2 measurements, and is different from the one-measurement form of G_k as in SPSA2-1R or the unbiased G_k of SPSA2-1UR proposed in Section 2.

While we do not show a detailed convergence analysis, we observe the following arguments.

$$\begin{aligned} & E(G_k^1(\theta_k^+) - \tilde{G}_k^1(\theta_k)|\theta_k, \Delta_k) \\ &= E(G_k^1(\theta_k^+)|\theta_k, \Delta_k) - E(\tilde{G}_k^1(\theta_k)|\theta_k, \Delta_k) \\ &= g(\theta_k^+) - \frac{L(\theta_k^+) - L(\theta_{k-1}^+) - \epsilon_{k-1}^+}{c_k}\Delta_k^{-1}. \end{aligned}$$

We make the claim that, modulo the correction (12) in the diagonal elements,

$$E\left(\frac{g(\theta_k^+)-\frac{L(\theta_k^+)-L(\theta_{k-1}^+)-\epsilon_{k-1}^+}{c_k}\Delta_k^{-1}}{c_k}(\Delta_k^{-1})^T|\mathcal{F}_k\right)$$

$$= H(\theta_k)+O(c_k),$$

the proof being in Appendix 4. Note in the above RHS that $H(\theta_k)$ is the Hessian at θ_k and the trailing term corresponds to a matrix with an induced norm bounded above by $O(c_k)$. We can also write the above as: $E\left(\frac{\delta G_{k,i}}{c_k\Delta_{k,j}}|\mathcal{F}_k\right) = H_{i,j}(\theta_k)+O(c_k), 1\le i,j\le p$. The convergence analysis uses the ODE technique discussed in the previous subsection, and since G_k is the same as algorithm 2SA of (Bhatnagar 2005), convergence of θ_k is assured using Theorem 3.3 of (Bhatnagar 2005).

The proof of convergence of the proposed algorithm can also be obtained in a manner similar to that of Theorems 1a and 2a of (Spall 2000). Note that (Spall 2000) uses the step-size $b_{k+1} = \frac{1}{k+1}$. Our algorithm is applicable for more general step-sizes as long as the requirement $a_k = o(b_k)$ is met.

An additional possibility in 2SPSA-2UR is to use a the gradient estimate $G_{k,l} = \frac{L(\theta_k^+)+\epsilon_k^+-L(\theta_{k-1}^+)-\epsilon_{k-1}^+}{c_k}\Delta_k^{-1}$ as in (3). The convergence can be shown by modifying (49) of (Bhatnagar 2005) using $\frac{L(\theta_k^+)-L(\theta_{k-1}^+)}{c_k}\Delta_k^{-1}$ (as in SPSA2-1UR) instead of the original. The advantage in such an estimate of G_k is that a less noisy set of terms are shown equal to zero in their mean by the property of conditional expectation. In particular, the terms in the G_k estimate of 2SA involve both Δ_k and $\tilde{\Delta}_k$ whilst the modified G_k of 2SPSA-2UR (as in (3)) has only Δ_k.

5.3 2SPSA-1UR

Here we propose a one-measurement algorithm which is also a second-order SPSA algorithm as the two algorithms of the previous sub-sections are. The method has similarities with algorithm 1SA of (Bhatnagar 2005), the crucial difference lying in design of the perturbations Δ_k and $\tilde{\Delta}_k$ and consequently, the estimation of $G_{k,l}$. We will require that $\Delta_k + \tilde{\Delta}_k \neq 0$ w.p. 1 and that $c_k = \tilde{c}_k$. The random variable $\Delta_k + \tilde{\Delta}_k$ would then be symmetric, mean-0 and bounded away from 0. This can be ensured, for example, by using Bernoulli-distributed random variables $\Delta_k \in \{-1,+1\}$ and $\tilde{\Delta}_k \in \{-0.5, 0.5\}$. The only measurement made at each iteration k is at θ_k^{++}. We retain the algorithm in (11), except that $G_k^1(\theta_k^+)$ and $G_k(\theta_k)$ are identified as below:

$$\begin{aligned} G_k^1(\theta_k^+) &= \frac{\tilde{\Delta}_k^{-1}}{\tilde{c}_k}\left(L(\theta_k^{++})+\epsilon_k^{++}-L(\theta_{k-1}^{++})-\epsilon_{k-1}^{++}\right) \\ G_k(\theta_k) &= \frac{(\Delta_k+\tilde{\Delta}_k)^{-1}}{c_k}\cdot\left(L(\theta_k^{++})+\epsilon_k^{++}\right. \\ &\quad \left.-L(\theta_{k-1}^{++})-\epsilon_{k-1}^{++}\right) \end{aligned}$$

We identify the changes required using results associated with 1SA of (Bhatnagar 2005). First, Proposition A.12 of (Bhatnagar 2005) is retained almost verbatim except for corresponding notation changes. In particular, the original statement that shows

$$\left|E\left[\left(\frac{L(\theta_k^{++})}{c_k^2}\tilde{\Delta}_k^{-1}\right)^T\Delta_k^{-1}|\mathcal{F}_k\right]-H(\theta_k)\right| \quad \rightarrow \quad 0,$$

is replaced by the analogous

$$|E\left[\left(\frac{L(\theta_k^{++})}{c_k^2}\tilde{\Delta}_k^{-1}-\frac{L(\theta_{k-1}^{++})}{c_{k-1}^2}\tilde{\Delta}_{k-1}^{-1}\right)^T\Delta_k{}^{-1}|\mathcal{F}_k\right]$$

$$-H(\theta_k)| \quad \rightarrow \quad 0.$$

An elementary modification of the proof of Theorem 3.4 in (Bhatnagar 2005) suffices to show convergence of the parameter θ_k. We modify (51) of (Bhatnagar 2005) to place the following term as LHS: $\frac{L(\theta_k^{++})-L(\theta_{k-1}^{++})}{c_k}(\Delta_k + \tilde{\Delta}_k)^{-1}$.

A key point to note with regard to the current algorithm is that it is possible to avoid storage of the size$-p$ vector $\frac{L(\theta_{k-1}^{++})}{c_{k-1}^2}\tilde{\Delta}_{k-1}^{-1}$ above by storing the noisy function measurement $L(\theta_{k-1}^{++})+\epsilon_k^{++}$ (done otherwise to unbias the $G_{k,l}$ measurement in 2SPSA-1UR) and then generate the vector $\tilde{\Delta}_{k-1}$ afresh at each step k. However, the variability produced using the actual estimate may be lesser. Further, it is interesting to note that the storage of the size$-p$ vector can be separated into two parts, the first involving unit storage for $L(\theta_{k-1}^{++})+\epsilon_k^{++}$. The second part involves storing the auxiliary perturbation $\tilde{\Delta}_{k-1}$, which in the Bernoulli case is a mere two-symbol string of size p that yields to efficient compression. Thus the $O(p)$ space requirement is conservative.

6 NUMERICAL EXAMPLE

We first compare algorithm SPSA2-2R of (Spall 1992) with the proposed SPSA2-1UR using the setting of (Spall 1997). In particular, the objective function used is

$$L_b(\theta) = b+\theta^T\theta+0.1\sum_{i=1}^{5}\theta_i^3+0.01\sum_{i=1}^{5}\theta_i^4,$$

with $\theta^* = 0$ and $L_b(\theta^*) = 0$ for all b. We keep $b = 0$ for comparison with SPSA2-2R and change to 0.1 for comparison with SPSA2-1R. We use $a = c = 1$, $\alpha = 6\gamma = 1$ and $\theta_0 = 0.1\underline{1}$ (i.e., the vector with 0.1

in all its components) in all the experiments. Assume that ϵ_k^+ are i.i.d., mean-zero, Gaussian random variables with variance σ_ϵ^2. The formula for asymptotic normality derived previously lets us consider two cases for the observation noise:

1. $\sigma_\epsilon = 0.1$ where $\frac{\tilde{n}_1}{n_2} \to 1.30$, and
2. $\sigma_\epsilon = 0.07$ where $\frac{\tilde{n}_1}{n_2} \to 0.93$, respectively.

Each run of the SPSA2-2R algorithm is for 2000 iterations, thus making 4000 observations of the objective function. Table 1 summarizes the results, the mean square error (MSE) obtained being over 100 runs of each algorithm. The MSE values for SPSA2-2R are less when compared to SPSA2-1UR, the proportion being 0.93 and 0.92, respectively for the two cases. However, this ratio improves if we use the SPSA2-1UH algorithm, which we compare with the analogous SPSA2-2H algorithm in Table 2.

Table 1: Mean Square Error and No. of Iterations

	$\sigma_\epsilon = 0.1$		$\sigma_\epsilon = 0.07$	
Algorithm	MSE	Iter.	MSE	Iter.
SPSA2-2R	0.0135	2000	0.0130	2000
SPSA2-1UR	0.0145	5200	0.0144	3600

While we have no asymptotic normality results for SPSA2-1UH, the performance obtained is better than that of SPSA2-1UR. We also observe the performance of SPSA2-1UR vis-a-vis SPSA2-1R in Table 3. Possibly due to the larger number of iterations required to achieve asymptotic normality, the MSE is always higher. A notable change in the behaviour of SPSA2-1R is the higher MSE when $b = 0.1$. This is due to the $L^2(\theta^*)$ bias term in (7). Note that we use $\sigma_e = 0.1$ in both the above comparisons.

We compare the second-order algorithms on the same setting. For algorithms 2SPSA-3UR and 2SPSA-2UR, we use $\tilde{\Delta}_{k,i} \in \{+1, -1\}$ whereas for 2SPSA-1UR we used $\tilde{\Delta}_{k,i} \in \{+0.5, -0.5\}$. The step-size $\tilde{c}_k$ was the same as c_k, whilst $b_k = \frac{1}{k^{0.55}}$. We used a similar projection operator $\Gamma(\cdot)$ as in the experiments of (Bhatnagar 2005), i.e., choose the diagonal elements $\bar{H}_k(i,i)$, $1 \leq i \leq p$ of the Hessian estimate and truncate to interval $[0.1, 10.0]$. An upper bound of 10 on $H_k(i,i)$ was justified since typically two-timescale algorithms are known to perform better with an additional averaging on the faster timescale, where $L >> 1$ measurements are made. Since recourse to multiple measurements is ruled out in this setting, we chose to prune the fluctuations in these diagonal terms $\bar{H}_k(i,i)$.

We compare 2SPSA-3UR with the four-measurement 2SPSA of (Spall 2000) to obtain the results in Table 4. We run both algorithms in such a manner that the number of function evaluations is the same, viz. 4000. The convergence of the bias (of $\hat{H}_k$) in 2SPSA-3UR is $O(c_k)$, resulting in problems establishing any asymptotic normality results. As a consequence, there is no clear set of parameters for which 2SPSA-3UR would outperform 2SPSA. This slower order of convergence may also be responsible for the poor performance of the algorithm. The experiments do indicate an occasional disconnect between finite-time performance of the second-order algorithms vis-a-vis the robust convergence behaviour expected from a Newton-Raphson analog. We chose this numerical setting to compare the proposed algorithms with those in the literature. The work (Zhu and Spall 2002) explores both finite-time performance and a computationally-efficient second-order SPSA algorithm. The difference with (Zhu and Spall 2002) would lie in choosing the Γ operator of (11). This is an issue identified in (Bhatnagar 2005), from where we chose the 3SA, 2SA and 1SA algorithms for modification. Table 5 compares performance of 2SPSA-2UR w.r.t. 2SA of (Bhatnagar 2005). The algorithms are more or less on par with each other. Table 6 compares the proposed 2SPSA-1UR algorithm with 1SA of (Bhatnagar 2005), the former being seen to perform better.

Table 2: Mean Square Error and No. of Iterations

	$\sigma_\epsilon = 0.1$		$\sigma_\epsilon = 0.07$	
Algorithm	MSE	Iter.	MSE	Iter.
SPSA2-2H	0.0133	2000	0.0127	2000
SPSA2-1UH	0.0109	5200	0.0109	3600

Table 3: Mean Square Error and No. of Iterations

	$b = 0$		$b = 0.1$	
Algorithm	MSE	Iter.	MSE	Iter.
SPSA2-1R	0.0443	4000	0.0492	4000
SPSA2-1UR	0.0132	6000	0.0147	4000

Table 4: Comparison of 2SPSA-3UR

	$\sigma_\epsilon = 0.1$		$\sigma_\epsilon = 0.07$	
Algorithm	MSE	Iter.	MSE	Iter.
2SPSA	0.037	1000	0.039	1000
2SPSA-3UR	0.078	1333	0.073	1333

7 FUTURE DIRECTIONS

The asymptotic convergence properties of SPSA2-1H have been theoretically shown to be on par with SPSA2-1R in Proposition 2.5 of (Xiong, Wang, and Fu 2002). Yet, it is unclear why SPSA2-1H performs better in practice and this represents an avenue for future investigation. Also of interest is the possibility of reducing the scale factor 2 in the asymptotic covariance matrix

Table 5: Comparison of 2SPSA-2UR

	$\sigma_\epsilon = 0.1$		$\sigma_\epsilon = 0.07$	
Algorithm	MSE	Iter.	MSE	Iter.
2SA	0.076	2000	0.077	2000
2SPSA-2UR	0.072	2000	0.078	2000

Table 6: Comparison of 2SPSA-1UR

	$\sigma_\epsilon = 0.1$		$\sigma_\epsilon = 0.07$	
Algorithm	MSE	Iter.	MSE	Iter.
1SA	0.854	4000	0.851	4000
2SPSA-1UR	0.655	4000	0.630	4000

$\tilde{M}_1$ using an average of past measurements $L(\theta_{k-j})$, $j > 1$. Whether online function regression mechanisms will serve as a 'critic' to speed up SPSA gradient descent by yielding an approximation of the objective function remains to be seen. Such an arrangement would place the resulting algorithm in-between the accepted forms of 'gradient-free' and 'gradient-based' methods. Further, in line with the asymptotic normality results of both first and second order SPSA algorithms, work such as (Konda and Tsitsiklis 2004) that identifies rate of convergence of two-timescale recursions should be useful.

ACKNOWLEDGMENTS

This work was supported in part by Grant no. SR/S3/EE/43/2002-SERC-Engg from the Department of Science and Technology, Government of India. We thank Dr. James C. Spall of Applied Physics Laboratory, Johns Hopkins University for his kind help in identifying the correct expression for the covariance matrix M_1 of the algorithm SPSA2-1R.

REFERENCES

Bhatnagar, S. 2005. Adaptive multivariate three–timescale stochastic approximation algorithms for simulation based optimization. *ACM Transactions on Modeling and Computer Simulation* 15 (1): 74–107.

Bhatnagar, S., M. Fu, S. Marcus, and I.-J. Wang. 2003. Two–timescale simultaneous perturbation stochastic approximation using deterministic perturbation sequences. *ACM Transactions on Modeling and Computer Simulation* 13 (4): 180–209.

Konda, V. R., and J. N. Tsitsiklis. 2004. Convergence rate of linear two-time-scale stochastic approximation. *Annals Of Applied Probability* 14 (2): 796–819.

Spall, J. C. 1992. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control* 37 (1): 332–341.

Spall, J. C. 1997. A one–measurement form of simultaneous perturbation stochastic approximation. *Automatica* 33 (1): 109–112.

Spall, J. C. 2000. Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Transactions on Automatic Control* 45 (10): 1839–1853.

Spall, J. C. 2001. Selected references on theory, applications, and numerical Analysis. URL `http://www.jhuapl.edu/SPSA/Pages/References-List_Ref.htm`, Accessed August 2006.

Spall, J. C. 2005. personal communication. dated 5 Oct.

Xiong, X., I.-J. Wang, and M. C. Fu. 2002. Randomized-direction stochastic approximation algorithms using deterministic sequences. In *Proceedings of the 2002 Winter Simulation Conference*, 285–291. Institute of Electrical and Electronics Engineers, Inc.

Zhu, X., and J. C. Spall. 2002. A modified second-order SPSA optimization algorithm for finite samples. *International Journal of Adaptive Control and Signal Processing* 16:397–409.

Appendix 1

The following is the algebra to show that $M_1 = a^2c^{-2}\rho^2(\sigma_\epsilon^2 + L^2(\theta^*))\mathrm{Diag}(\{(2\lambda_l - \beta_+)^{-1}\}_{l=1}^p)$ (cf. (3.5) of (Spall 1992) for notation and derivation upto this point):

$E(V_k{V_k}^T|\mathcal{F}_k)$

$$
\begin{aligned}
&= k^{-2\gamma}E\left\{\left(\frac{L(\theta_k^+)+\epsilon_k^+}{ck^{-\gamma}}(\Delta_k)^{-1} - [g(\theta_k)+b_k(\theta_k)]\right)\right.\\
&\qquad \left.\cdot\left(\frac{L(\theta_k^+)+\epsilon_k^+}{ck^{-\gamma}}(\Delta_k)^{-1} - [g(\theta_k)+b_k(\theta_k)]\right)^T|\mathcal{F}_k\right\}\\
&= k^{-2\gamma}E\left\{\Delta_k^{-1}(\Delta_k^{-1})^T\frac{L^2(\theta_k^+)}{c^2k^{-2\gamma}}|\mathcal{F}_k\right\}\\
&\quad + 2k^{-2\gamma}E\left\{\Delta_k^{-1}(\Delta_k^{-1})^T\frac{\epsilon_k^+L(\theta_k^+)}{c^2k^{-2\gamma}}|\mathcal{F}_k\right\}\\
&\quad -k^{-2\gamma}E\left\{[g(\theta_k)+b_k(\theta_k)](\Delta_k^{-1})^T\frac{L(\theta_k^+)}{ck^{-\gamma}}|\mathcal{F}_k\right\}\\
&\quad + k^{-2\gamma}E\left\{\Delta_k^{-1}(\Delta_k^{-1})^T\frac{{\epsilon_k^+}^2}{c^2k^{-2\gamma}}|\mathcal{F}_k\right\}\\
&\quad -k^{-2\gamma}E\left\{[g(\theta_k)+b_k(\theta_k)](\Delta_k^{-1})^T\frac{\epsilon_k^+}{ck^{-\gamma}}|\mathcal{F}_k\right\}
\end{aligned}
$$

$$- k^{-2\gamma}E\left\{\Delta_k^{-1}[g(\theta_k)+b_k(\theta_k)]^T\frac{L(\theta_k^+)}{ck^{-\gamma}}|\mathcal{F}_k\right\}$$
$$-k^{-2\gamma}E\left\{\Delta_k^{-1}[g(\theta_k)+b_k(\theta_k)]^T\frac{\epsilon_k^+}{ck^{-\gamma}}|\mathcal{F}_k\right\}$$
$$+ k^{-2\gamma}[g(\theta_k)+b_k(\theta_k)][g(\theta_k)+b_k(\theta_k)]^T. \tag{13}$$

Only the first, fourth and eighth terms in (13) remain, since the rest cancel as a result of independent, zero mean ϵ_k^+ and $(\Delta_k)^{-1}$. In the expansion (13), note that the terms:

$$k^{-2\gamma}E\left\{\Delta_k^{-1}(\Delta_k^{-1})^T\frac{L^2(\theta_k+\bar{\Delta}_k)}{c^2k^{-2\gamma}}|\mathcal{F}_k\right\}$$
$$k^{-2\gamma}E\left\{\Delta_k^{-1}(\Delta_k^{-1})^T\frac{{\epsilon_k^+}^2}{c^2k^{-2\gamma}}|\mathcal{F}_k\right\}$$
$$k^{-2\gamma}[g(\theta_k)+b_k(\theta_k)][g(\theta_k)+b_k(\theta_k)]^T$$

converge to $c^{-2}\rho^2L^2(\theta^*)$, $c^{-2}\rho^2\sigma_\epsilon^2$, and 0, respectively as $k\to\infty$. The mean zero property of ϵ_k^+ is used for the second term. Further, $\Phi=-aI$, resulting in:

$$M_1(i,i)=a^2c^{-2}\rho^2(\sigma_\epsilon^2+L^2(\theta^*))(2\lambda_i-\beta_+)^{-1},$$

and thus validating (5).

Appendix 2

Now to proceed with computation of $\tilde{M}_1$, arranging terms together similar to (13), one obtains:

$$\begin{aligned}
&E(V_kV_k^T|\mathcal{F}_k)\\
&= k^{-2\gamma}E\left\{(\frac{L(\theta_k^+)+\epsilon_k^+-L(\theta_{k-1}^+)-\epsilon_{k-1}^+}{ck^{-\gamma}}(\Delta_k)^{-1}\right.\\
&\quad -[g(\theta_k)+b_k(\theta_k)])\\
&\quad \cdot(\frac{L(\theta_k^+)+\epsilon_k^+-L(\theta_{k-1}^+)-\epsilon_{k-1}^+}{ck^{-\gamma}}(\Delta_k)^{-1}\\
&\quad \left. -[g(\theta_k)+b_k(\theta_k)])^T|\mathcal{F}_k\right\}\\
&= k^{-2\gamma}E\left\{\Delta_k^{-1}(\Delta_k^{-1})^T\frac{(L(\theta_k^+)-L(\theta_{k-1}^+))^2}{c^2k^{-2\gamma}}|\mathcal{F}_k\right\}\\
&\quad +2k^{-2\gamma}E\left\{\Delta_k^{-1}(\Delta_k^{-1})^T\frac{\epsilon_k^+-\epsilon_{k-1}^+}{ck^{-\gamma}}\right.\\
&\quad \left.\cdot\frac{L(\theta_k^+)-L(\theta_{k-1}^+)}{ck^{-\gamma}}|\mathcal{F}_k\right\}\\
&\quad +k^{-2\gamma}E\left\{\Delta_k^{-1}(\Delta_k^{-1})^T\frac{(\epsilon_k^+-\epsilon_{k-1}^+)^2}{c^2k^{-2\gamma}}|\mathcal{F}_k\right\}\\
&\quad +k^{-2\gamma}[g(\theta_k)+b_k(\theta_k)][g(\theta_k)+b_k(\theta_k)]^T.
\end{aligned}\tag{14}$$

Of the above terms, the third term on the RHS of the second equality in (14) can be seen to converge to $2c^{-2}\sigma_\epsilon^2\rho^2$. All the other terms above converge to 0, giving us $\tilde{M}_1$ as claimed. To obtain μ, observe that the bias term $b_{k,l}(\theta_k)$ is the same as for SPSA-1, which is used in the special case $3\gamma-\frac{\alpha}{2}=0$ in Proposition 2 of (Spall 1992)).

Appendix 3

We claim that $\frac{1}{c_k\Delta_{k,j}}\left(g_i(\theta_k^+)-\frac{L(\theta_k^+)-L(\theta_k^-)}{2c_k\Delta_{k,i}}\right) = H_{i,j}(\theta_k)+O(c_k)$ where we use the knowledge that $g_i(\theta_k^+)$

$$= g_i(\theta_k)+c_k\sum_{m=1}^{p}H_{i,m}(\theta_k)\Delta_{k,m}+O(c_k^2)$$

and $\frac{L(\theta_k^+)-L(\theta_k^-)}{2c_k\Delta_{k,i}}$

$$= g_i(\theta_k)+\sum_{m=1,m\neq i}^{p}g_m(\theta_k)\frac{\Delta_{k,m}}{\Delta_{k,i}}+O(c_k^2)$$

substituting both the above into the original expression, we have

$$H_{i,j}(\theta_k)+\sum_{m\neq j}H_{i,m}(\theta_k)\frac{\Delta_{k,m}}{\Delta_{k,j}}$$
$$-\frac{1}{c_k}\sum_{m=1,m\neq i}^{p}g_m(\theta_k)\frac{\Delta_{k,m}}{\Delta_{k,i}\Delta_{k,j}}+O(c_k),$$

where the expectation of the second and third terms are zero.

Appendix 4

Subject to corrections (12) in the diagonal elements, we claim that $\frac{1}{c_k\Delta_{k,j}}\left(g_i(\theta_k^+)-\frac{L(\theta_k^+)-L(\theta_{k-1}^+)-\epsilon_{k-1}^+}{c_k\Delta_{k,i}}\right) = H_{i,j}(\theta_k)+O(c_k)$ where we use the expression for $g_i(\theta_k^+)$ from Appendix 3 and, employing the shorthand $\tilde{L}(\theta_{k-1}^+)=L(\theta_{k-1}^+)+\epsilon_{k-1}^+$, observe that

$$\begin{aligned}
\frac{L(\theta_k^+)-\tilde{L}(\theta_{k-1}^+)}{c_k\Delta_{k,i}} &= \frac{L(\theta_k)-\tilde{L}(\theta_{k-1}^+)}{c_k\Delta_{k,i}}+g_i(\theta_k)\\
&\quad+\sum_{m=1,m\neq i}^{p}g_m(\theta_k)\frac{\Delta_{k,m}}{\Delta_{k,i}}\\
&\quad+\frac{c_k\Delta_k^TH(\theta_k)\Delta_k}{2\Delta_{k,i}}+O(c_k^2).
\end{aligned}$$

substituting which, we have

$$
\begin{aligned}
= \quad & -\frac{L(\theta_k)-\tilde{L}(\theta_{k-1}^{+})}{c_k^2\Delta_{k,i}\Delta_{k,j}} + H_{i,j}(\theta_k) \\
& + \sum_{m\neq j} H_{i,m}(\theta_k)\frac{\Delta_{k,m}}{\Delta_{k,j}} \\
& -\frac{1}{c_k}\sum_{m=1,m\neq i}^{p} g_m(\theta_k)\frac{\Delta_{k,m}}{\Delta_{k,i}\Delta_{k,j}} \\
& -\frac{\Delta_k^T H(\theta_k)\Delta_k}{2\Delta_{k,i}\Delta_{k,j}} + O(c_k).
\end{aligned}
$$

The first and the second-last terms above play a critical role when $i = j$, whilst all other terms have mean 0.

In this connection, assume that $\Delta_{k,i}$ are Bernoulli distributed RVs and $\Delta_{k,i}^2 = 1$, resulting in $E(\Delta_k^T H(\theta_k)\Delta_k|\mathcal{F}_k) = \sum_{i=1}^{p} H_{i,i}(\theta_k)$. Now note that due to the correction term that we add in (12), $E(-\frac{L(\theta_k)-\tilde{L}(\theta_{k-1}^{+})}{c_k^2} + \frac{L(\theta_k^{+})-\tilde{L}(\theta_{k-1}^{+})}{c_k^2}|\mathcal{F}_k)$ reduces to $E(\frac{L(\theta_k^{+})-L(\theta_k)}{c_k^2}|\mathcal{F}_k)$. This in turn equals $\frac{1}{2}E(\Delta_k^T H(\theta_k)\Delta_k) + O(c_k)$, canceling the second-last term in the RHS above.