Proceedings of the 1995 Winter Simulation Conference
ed. C. Alexopoulos, K. Kang, W. R. Lilegdon, and D. Goldsman

STOCHASTIC VERSION OF SECOND-ORDER (NEWTON-RAPHSON) OPTIMIZATION USING ONLY FUNCTION MEASUREMENTS

James C. Spall

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

ABSTRACT

Consider the problem of loss function minimization when only (possibly noisy) measurements of the loss function are available. In particular, no measurements of the gradient of the loss function are assumed available (as required in the steepest descent or Newton-Raphson algorithms). Stochastic approximation (SA) algorithms of the multivariate Kiefer-Wolfowitz (finite-difference) form have long been considered for such problems, but with only limited success. The simultaneous perturbation SA (SPSA) algorithm has successfully addressed one of the major shortcomings of those finite-difference SA algorithms by significantly reducing the number of measurements required in many multivariate problems of practical interest. This SPSA algorithm displays the classic behavior of first-order search algorithms by typically exhibiting a steep initial decline in the loss function followed by a slow decline to the optimum. This paper presents a second-order SPSA algorithm that is based on estimating both the loss function gradient and inverse Hessian matrix at each iteration. The aim of this approach is to emulate the acceleration properties associated with deterministic algorithms of Newton-Raphson form, particularly in the terminal phase where the first-order SPSA algorithm slows down in its convergence. This second-order SPSA algorithm requires only *three* loss function measurements at each iteration, independent of the problem dimension. This paper includes a formal convergence result for this second-order approach.

1 INTRODUCTION

There has recently been a growing interest in recursive optimization algorithms of stochastic approximation (SA) form that do not depend on direct gradient information or measurements. Rather, these SA algorithms are based on an *approximation* to the p-dimensional (say) gradient formed from measurements of the objective function. This interest has been motivated by problems such as the adaptive control of complex processes, system optimization based on computer- and/or labor-intensive simulations, the training of recurrent neural networks, and the optimization of complex queuing and discrete-event systems. The principal advantage of algorithms that do not require direct gradient measurements is that they do not require knowledge of the functional relationship between the parameters being adjusted (optimized) and the objective (say, loss) function being minimized. Such a relationship, together with its gradient, can be notoriously difficult to develop in problem areas such as those mentioned above.

The simultaneous perturbation SA (SPSA) algorithm in Spall (1988, 1992), which is based on a highly efficient gradient approximation (requiring only two measurements of the loss function for any p), is one such gradient-free algorithm. Theory and examples in these references have shown that SPSA is generally capable of significantly reducing the total number of loss function measurements needed to achieve convergence over other standard SA algorithms (see also Chin 1994 and Spall and Cristion 1992, 1994). This paper extends the SPSA algorithm to include second-order (Hessian) effects with the aim of accelerating convergence in a stochastic analogue to the deterministic Newton-Raphson algorithm. Like the standard (first-order) SPSA algorithm, this second-order algorithm is simple to implement and requires only a small number—independent of p—of loss function measurements per iteration. In particular, only three measurements are required to estimate the loss-function gradient and inverse Hessian at each iteration. The results here represent an extension and enhancement (relative to the basic approach, theory, and numerical analysis) of the basic second-order idea introduced in Spall (1994).

We consider the problem of minimizing a (scalar) differentiable loss function $L(\theta)$,where $\theta \in \mathbb{R}^p$, $p \geq 1$. A typical example of $L(\theta)$ would be some measure of mean-

square error for the output of a process as a function of some design parameters θ. For most cases of practical interest, this is equivalent to finding the minimizing θ^* such that

$$g(\theta^*) \equiv \left.\frac{\partial L}{\partial \theta}\right|_{\theta=\theta^*} = 0. \tag{1}$$

It is assumed that measurements of $L(\theta)$ are available at various values of θ. These measurements may or may not include added random noise. No direct measurements (either with or without noise) of $g(\theta)$ are assumed available, such as are required in the well-known Robbins-Monro (1951) SA algorithm (which includes algorithms such as neural network back-propagation, infinitesimal perturbation analysis for discrete event systems, and steepest descent as special cases).

The standard first-order SA algorithms for estimating θ involve a simple recursion with, usually, a scalar gain and an approximation to the gradient based on the measurements of $L(\cdot)$. The SPSA algorithm mentioned above requires only two measurements of $L(\cdot)$ to form the gradient approximation, independent of p (versus $2p$ in the standard multivariate finite-difference approximation considered, e.g., in Sacks 1958, which extends the scalar algorithm of Kiefer and Wolfowitz 1952). Theory presented in Spall (1992) and Chin (1994) shows that for large p the SPSA approach can be much more efficient (in terms of total number of loss measurements to achieve effective convergence to θ^*) than the finite-difference approach in many cases of practical interest.

In extending SPSA to a second-order (accelerated) form, we outline in Section 2 how the gradient and inverse Hessian of $L(\cdot)$ can both be estimated on a per-iteration basis using only *three* measurements of $L(\cdot)$ (again, independent of p). With these estimates, we can then create an SA analogue to the Newton-Raphson algorithm (which, recall, is based on an update step that is negatively proportional to the inverse Hessian times the gradient).

Before presenting the approach, let us contrast it with other second-order SA approaches. Fabian (1971) forms estimates of the gradient and Hessian for a Newton-Raphson-type SA algorithm by using, respectively, a finite difference approximation and a set of differences of finite difference approximations. This leads to $O(p^2)$ measurements of $L(\cdot)$ per update of the θ estimate, which is extremely costly when p is large. Ruppert (1985) assumes that direct measurements of the gradient $g(\cdot)$ are available, as in Robbins-Monro. He then forms a Hessian estimate by taking a finite-difference of gradient measurements; hence $O(p)$ measurements of $g(\cdot)$ are required for each update step in estimating θ. This approach differs from ours in both its requirements to measure $g(\cdot)$ and in its large number of measurements required per iteration. A type of second-order convergence for SA is reported in Ruppert (1988) and Polyak and Juditsky (1992) based on the idea of iterate averaging. However, as discussed in Chin and Maryak (1995), this approach does not generally provide accelerated convergence in the SPSA setting. The algorithm here is in the spirit of adaptive (matrix) gain SA algorithms such as those considered in Benveniste, Metivier, and Priouret (1990, Chaps. 3–4) in that a matrix gain is estimated concurrently with an estimate of the parameters of interest. It differs, however, in that no direct observations of the gradient are assumed available.

2 THE APPROACH

The second-order SPSA approach is composed of two parallel recursions, one for θ and one for the upper triangular matrix square-root, say $S = S(\theta)$, of the Hessian of $L(\theta)$. (We estimate the square root to ensure that the inverse Hessian estimate used in the second-order SPSA recursion for θ is positive semidefinite.) The two recursions are, respectively,

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k(\hat{S}_k^T \hat{S}_k)^{-1} \hat{g}_k(\hat{\theta}_k) \tag{2a}$$

$$\hat{S}_{k+1} = \hat{S}_k - \tilde{a}_k \hat{G}_k(\hat{S}_k), \tag{2b}$$

where a_k, $\tilde{a}_k$ are non-negative scalar gain coefficients, $\hat{g}_k(\hat{\theta}_k)$ is the SP gradient approximation to $g_k(\hat{\theta}_k)$ (see Spall 1992), and $\hat{G}_k$ is an observation related to the gradient of a certain loss function (defined in Equation (3) below) with respect to S. Note that $\hat{S}_k^T \hat{S}_k$ (which depends on $\hat{\theta}_k$) represents an estimate of the Hessian matrix of $L(\hat{\theta}_k)$. Hence Equation (2a) is a stochastic analogue of the well-known Newton-Raphson algorithm of deterministic optimization. Since $\hat{g}_k(\hat{\theta}_k)$ has a known form, the parallel recursions in Equations (2a) and (2b) can be implemented once we specify $\hat{G}_k$, which is addressed below.

As discussed in Spall (1988,1992), the SP gradient approximation requires two measurements of $L(\cdot)$: $y_k^{(+)}$ and $y_k^{(-)}$. These represent measurements at design levels $\hat{\theta}_k + c_k\Delta_k$ and $\hat{\theta}_k - c_k\Delta_k$ respectively, where c_k is a positive scalar and Δ_k represents a user-generated random vector satisfying certain regularity conditions, e.g., Δ_k being a vector of independent Bernoulli ± 1 random variables satisfies these conditions but a vector of uniformly distributed random variables does not. (The term "SP" comes from the fact that all elements of $\hat{\theta}_k$ are perturbed simultaneously in forming $\hat{g}_k(\hat{\theta}_k)$, as opposed to the finite difference form, where they are perturbed one-at-a-time.) To perform one iteration of Equations (2a) and (2b), one additional measurement, say $y_k^{(0)}$, is required; this measurement represents an observation of $L(\cdot)$ at the nominal

design level $\hat{\theta}_k$. For these three measurements, we make the assumption that their corresponding noises $\epsilon_k^{(+)}$, $\epsilon_k^{(-)}$, and $\epsilon_k^{(0)}$ (e.g., $y_k^{(0)} = L(\hat{\theta}_k) + \epsilon_k^{(0)}$) satisfy a standard martingale difference condition for all k (see condition C.2 below). Note that there is no requirement that the noises be mutually or sequentially independent.

As a means of obtaining an estimate of the square-root Hessian, we introduce the following loss function to be minimized:

$$\tilde{L}(S|\theta) \equiv \tfrac{1}{2} E\left[\Delta_k^T S^T S \Delta_k - \Delta_k^T H(\theta) \Delta_k\right]^2, \tag{3}$$

where S is restricted to be in upper triangular form and the $\{\Delta_k\}$ are i.i.d. vectors (see also C.2 below). Note that $\tilde{L}(S|\theta)$ is minimized (= 0) at $S = H(\theta)^{1/2}$, where the exponent, 1/2, represents the upper triangular square root. In the usual way, the optimal S can be found as a solution to $G(S|\theta) \equiv \partial \tilde{L}(S|\theta)/\partial S = 0$. The classical Robbins-Monro SA algorithm can then be used to find S once we obtain an appropriate estimate of the matrix gradient $G(S|\theta)$.

To motivate the form for the estimate of $\partial \tilde{L}(S|\theta)/\partial S$ at $S = \hat{S}_k$, denoted $\hat{G}_k(\hat{S}_k|\hat{\theta}_k)$ in Equation (2b) above, note that

$$G(S|\theta) = E\left[\left(\Delta_k^T S^T S \Delta_k - \Delta_k^T H(\theta) \Delta_k\right) \frac{\partial \Delta_k^T S^T S \Delta_k}{\partial S}\right], \tag{4}$$

where the matrix gradient is with respect to only the upper triangular elements in S. Other elements are fixed at zero (this assumes for now the validity of the indicated interchange of derivative and integral, which will be justified in the proposition below). Then we let

$$\hat{G}_k(\hat{S}_k|\hat{\theta}_k) = \left[\Delta_k^T \hat{S}_k^T \hat{S}_k \Delta_k + \frac{y_k^{(+)} + y_k^{(-)} - 2y_k^{(0)}}{c_k^2}\right] \times \left.\frac{\partial \Delta_k^T S^T S \Delta_k}{\partial S}\right|_{\hat{S}_k}, \tag{5}$$

where from Graham (1981, p. 123)

$$\frac{\partial \Delta_k^T S^T S \Delta_k}{\partial S} = 2 S \Delta_k \Delta_k^T. \tag{6}$$

A simple second-order Taylor expansion of both $L(\hat{\theta}_k \pm c_k \Delta_k)$ about $\hat{\theta}_k$ provides the intuition as to why the form in Equation (5) is chosen as an estimate of the gradient in Equation (4). This claim is made rigorous in the Proposition below.

We now establish conditions such that the form for $\hat{G}_k$ will yield an a.s. convergent estimate for $\hat{\theta}_k$ and $\hat{S}_k$. The Proposition below rests on the following regularity conditions, most of which are natural extensions of the conditions in Spall (1992, Prop. 1) for the first-order SPSA algorithm.

C.0 For almost all $\hat{\theta}_k$, $L^{(3)}(\theta) \equiv \partial^3 L/\partial\theta^T \partial\theta^T \partial\theta^T$ is Lipschitz continuous in a neighborhood of $\hat{\theta}_k$, where the size and shape of the neighborhood are independent of k and the value $\hat{\theta}_k$.

C.1 The standard conditions on a_k, c_k in Spall (1992, cond. A1) hold, as well as

$$\tilde{a}_k > 0, \tilde{a}_k \to 0, \sum \tilde{a}_k = \infty, \sum_{k=0}^{\infty} \frac{\tilde{a}_k^2}{c_k^4} < \infty,$$

$$\text{and } \frac{\tilde{a}_k}{a_k} \to r > 0$$

C.2 $E(\epsilon_k^{(\cdot)} | \hat{\beta}_k, \Delta_k) = 0$, $E\epsilon_k^{(\cdot)2} \le const.$, $EL(\hat{\theta}_k \pm c_k \Delta_k)^2 \le const.$, and Δ_{ki} is symmetrically distributed (about 0) and bounded with $E\Delta_{ki}^{-2} \le const.$ $(i = 1, 2, \ldots, p)$

C.3 $\sup_k \|\hat{\theta}_k\| < \infty$ a.s. and $\sup_k \|\hat{S}_k^{\pm 1}\| < \infty$ a.s.

C.4 The pair $\{\theta^*, S^*\}$ is an asymptotically stable solution of the linked differential equations

$$\frac{d\{x_\theta(t), x_S(t)\}}{dt} = \left\{-(x_S^T x_S)^{-1} g(x_\theta), -rG(x_S|x_\theta)\right\}$$

where r is defined in C.1.

C.5 There exists a compact set contained within the domain of attraction (see, e.g., Lai 1985) associated with the linked differential equations in C.4 such that $\{\hat{\theta}_k, \hat{S}_k\}$ lies within the compact set infinitely often for almost all sample points.

Proposition. If conditions C.0–C.5 hold, then

$$\hat{\theta}_k \to \theta^* \text{ a.s.} \tag{7a}$$

$$\hat{S}_k \to S^* \text{ a.s.} \tag{7b}$$

Proof. Let $\hat{\beta} = (\hat{\theta}_k, \hat{S}_k)$. Note that Equations (2a) and (2b) can be written as one joint SA algorithm for $\hat{\beta}_k$ with gain a_k (so that $\tilde{a}_k/a_k$ appears before $\hat{G}_k$ to preserve algebraic equivalence to Equation 2b. Hence, by C.1 and C.3–C.5, the differential equation method of Kushner and Clark (1978, pp. 38–39), Metivier and Priouret (1984), or Lai (1985) can be employed. This method implies that Equations (7a) and (7b) will hold if the following two pairs of conditions hold:

(i) (a) $\|b_k(\hat{\beta}_k)\| \le const.\ \forall k, b_k \to 0$ a.s.

(b) $\|\tilde{b}_k(\hat{\beta}_k)\| \le const.\ \forall k, \tilde{b}_k \to 0$ a.s.

$$\text{(ii) (a)} \quad \lim_{k\to\infty} P\left(\sup_{m\geq k} \left\| \sum_{i=k}^{m} a_i c_i (\hat{\beta}_i) \right\| \geq \eta \right) = 0, \text{ for any } \eta > 0$$

$$\text{(b)} \quad \lim_{k\to\infty} P\left(\sup_{m\geq k} \left\| \sum_{i=k}^{m} \tilde{a}_i \tilde{c}_i (\hat{S}_i) \right\| \geq \eta \right) = 0,$$

where b_k and $\tilde{b}_k$ denote the (conditional on $\hat{\beta}_k$) bias terms for the estimates $(\hat{S}_k^T \hat{S}_k)^{-1} \hat{g}_k(\hat{\theta}_k)$ and $\hat{G}_k(\hat{S}_k|\hat{\theta}_k)$ respectively, and e_k, $\tilde{e}_k$ denote the corresponding error terms (analogous to Spall 1992, Section 3B).

First, (i)(a) follows from C.3 and the fact that the bias in $\hat{g}_k(\hat{\theta}_k)$ is uniformly bounded and $O(c_k^2)$. Now for (i)(b), we know by a straightforward Taylor expansion about $\hat{\theta}_k$ that

$$\tilde{b}_k(\hat{\beta}_k) = \frac{1}{6} c_k E\left\{ \left[\left(L^{(3)}(\bar{\theta}_k^{(+)}) - L^{(3)}(\bar{\theta}_k^{(-)}) \right) \Delta_k \otimes \Delta_k \otimes \Delta_k + \frac{\epsilon_k^{(+)} + \epsilon_k^{(-)} - 2\epsilon_k^{(0)}}{c_k^2} \right] \hat{S}_k \Delta_k \Delta_k^T \middle| \hat{\beta}_k \right\}$$

where $\bar{\theta}_k^{(\pm)}$ lies on the line segment between $\hat{\theta}_k$ and $\hat{\theta}_k \pm c_k \Delta_k$ and we employ Equation (5) for the contribution due to $\partial \Delta_k^T S^T S \Delta_k / \partial S$. By the Lipschitz continuity assumption in C.0, we know that $\| L^{(3)}(\bar{\theta}_k^{(+)}) - L^{(3)}(\bar{\theta}_k^{(-)}) \| \leq (const.) c_k$. Further, the mean 0 noise condition in C.2 removes the noise contribution after the "+" sign. Finally, C.3 guarantees the boundedness of $\hat{S}_k \Delta_k \Delta_k^T$. Hence, both parts of (i)(b) hold (with $\tilde{b}_k(\hat{\beta}_k) = O(c_k^2)$ a.s.).

For (ii)(a), we can follow arguments analogous to Spall (1992, Prop. 1) that employ the martingale inequality in Doob (1953, p. 315) or Kushner and Clark (1978, p. 27). Briefly, with $e_k = (\hat{S}_k^T \hat{S}_k)^{-1} (\hat{g}_k(\hat{\theta}_k) - E(\hat{g}_k(\hat{\theta}_k)|\hat{\theta}_k))$ we can use assumption C.3 (boundedness of $\hat{S}_k^{-1}$) together with the simple martingale arguments associated with Equation (3.4) in Spall (1992) to conclude that (ii)(a) is true.

Finally, for (ii)(b), we can follow the martingale ideas above and conclude that the result holds if

$$\lim_{k\to\infty} \sum_{i=k}^{\infty} \tilde{a}_i^2 E\|\tilde{e}_i\|^2 = 0 \tag{8}$$

where $\tilde{e}_i = \hat{G}_i(\hat{S}_i|\hat{\theta}_i) - E(\hat{G}_i(\hat{S}_i|\hat{\theta}_i)|\hat{\beta}_i)$. Then, invoking the boundedness assumptions in C.0, C.2, and C.3, straightforward algebra shows that $E\|\tilde{e}_k\|^2 = O(c_k^{-1})$. Hence, by C.1, Equation (8) holds. Q.E.D.

Remark 1. There are several simple variations on the form for $\hat{G}_k$ shown in Equation (5), which might enhance the performance in certain cases, particularly when the measurement noises might tend to be large. These involve taking additional $L(\cdot)$ measurements per iteration with the aim of more than proportionally reducing the number of iterations required to achieve effective (practical) convergence to θ^*. Two obvious variations are: (i) replace $2y_k^{(0)}$ in Equation (8) with the sum of two separate measurements of $L(\hat{\theta}_k)$ (so that Equations (2a) and (2b) now require *four* $L(\cdot)$ measurements per iteration instead of three), or (ii) average several values of $\hat{G}_k$ (and $\hat{g}_k(\hat{\theta}_k)$) based on separate sets of three (or four, as just discussed) measurements to form an input to the recursion (Equations (2a) and (2b)) that has lower noise effects. Spall (1992) examines such averaging in the context of standard SPSA, and finds that it can often be both theoretically and computationally effective at reducing the total number of $L(\cdot)$ measurements required to achieve effective convergence.

Remark 2. Although we estimate the square root of the Hessian to ensure a positive semidefinite matrix in the SA update of Equation (2a), the same basic idea can be used to estimate (say) the Hessian directly (making the obvious changes to Equations (3) through (6)). This would allow an examination of the loss surface to ensure that we are seeking a minimum (not, say, a saddle point) if this were a concern for a particular application. In fact, using the same three loss measurements, we could augment the recursions (Equations (2a) and (2b)) with an additional recursion (say Equation (2c)) to estimate the Hessian directly as a way of monitoring that the algorithm in Equations (2a) and (2b) is yielding a loss minimum.

Remark 3. Using standard numerical methods (e.g., Householder 1964, Chap. 5), it is possible to avoid explicitly calculating the matrix inverse shown in Equation (2a). This will yield significant computational savings at greater numerical stability.

3 SMALL-SCALE NUMERICAL STUDY

This section summarizes the results of a preliminary numerical study on the second-order SPSA algorithm of Equations (2a) and (2b). We will compare its performance with that of the standard first-order SPSA algorithm in Spall (1988, 1992). The loss function $L(\cdot)$ we consider is a fourth-order polynomial with significant interaction among the $p = 10$ elements in θ (i.e., the Hessian matrix has significant off-diagonal elements); this makes the loss function flat near θ^* and, consequently, the optimization problem challenging.

Tables 1 and 2 provide the results for this preliminary study, showing the ratio of the estimation error $\|\hat{\theta}_k - \hat{\theta}^*\|$ to the initial error $\|\hat{\theta}_0 - \hat{\theta}^*\|$ based on an average of five independent runs (the same $\hat{\theta}_0$ was used in all runs, and represents the standard Euclidean norm). 1SPSA and 2SPSA represent the first-order and second-order SPSA algorithms, respectively. Table 1 considers the case where there is no noise in the measurements of $L(\cdot)$, while Table 2 includes Gaussian measurement noise (with a one-sigma

value that ranges from 3 to over 100 percent of the $L(\theta)$ value as θ varies). The left-hand column represents the total number of measurements used (so with 3000 measurements, 1SPSA has gone through $k = 1500$ iterations while 2SPSA has gone through $k = 1000$ iterations). The first two results columns in the tables represent runs with the same SA gains a_k, c_k, tuned numerically to approximately optimize the performance of the 1SPSA algorithm (the gains satisfied the conditions in the Proposition). The third results column is based on a (numerical) recalibration of a_k, c_k to be approximately optimized for the 2SPSA algorithm (an identical $\bar{a}_k$ sequence was used for both 2SPSA columns). The results in both tables illustrate the performance of the second-order SPSA approach for a difficult-to-optimize (i.e., flat surface) function. As expected, we see that the ratios (for both 1SPSA and 2SPSA) tend to be lower in the no-noise case of Table 1. Further, we see that the 2SPSA algorithm provides solutions closer to θ^* both with and without optimal 2SPSA gains. An enlightening way to look at the numbers in the tables is to compare the number of measurements needed to achieve the same level of accuracy. We see that in the no-noise case (Table 1), the ratio of number of measurements for 2SPSA:1SPSA ranged from 1:2 to 1:50. In the noisy measurement case (Table 2), the ratios for 2SPSA:1SPSA ranged from 1:2 to 1:20. These ratios offer considerable promise for practical problems, where p is even larger (say, as in the neural network–based direct adaptive control method of Spall and Cristion 1992, 1994, where p can easily be of order 10^2 or 10^3). In such cases, other second-order techniques that require a growing (with p) number of function measurements are likely to become infeasible.

Table 1: Values of $\dfrac{\|\hat{\theta}_k - \theta^*\|}{\|\hat{\theta}_0 - \theta^*\|}$ with No Measurement Noise

Number of measurements	1SPSA	2SPSA w/1SPSA gains	2SPSA w/optimal gains
3,000	0.265	0.287	0.122
15,000	0.184	0.160	0.033
30,000	0.146	0.128	0.018

Table 2: Values of $\dfrac{\|\hat{\theta}_k - \theta^*\|}{\|\hat{\theta}_0 - \theta^*\|}$ with Measurement Noise

Number of measurements	1SPSA	2SPSA w/1SPSA gains	2SPSA w/optimal gains
3,000	0.273	0.292	0.243
15,000	0.184	0.163	0.103
30,000	0.146	0.141	0.097

There are several important practical concerns in implementing the 2SPSA algorithm. One, of course, involves the choice of SA gains. As in all SA algorithms, this must be done with some care to ensure good performance of the algorithm. Some theoretical guidance is provided in Fabian (1971) and Chin (1994), but we have found that empirical experimentation is more effective and easier. Another practical aspect involves the use of the Hessian estimate: in the studies here we found it more effective to *not* use the Hessian estimate for the first few (100) iterations in Equations (2a), i.e., still compute $\hat{S}_k$ but replace $\hat{S}_k\hat{S}_k^T$ in Equation (2a) with an identity matrix so that it then becomes the standard SPSA algorithm for the first few iterations. This allows the inverse Hessian estimate to improve while it really is not needed since $L(\cdot)$ is dropping quickly because of the characteristic steep initial decline of the standard SPSA algorithm.

4 CONCLUDING REMARKS

The second-order SPSA algorithm presented above offers considerable potential for accelerating the convergence of SA algorithms while only requiring loss function measurements (no gradient or higher derivative measurements are needed). Since it requires only three measurements per iteration to estimate both the gradient and Hessian—independent of problem dimension p—it does not impose a large requirement for data collection and/or computation as p gets large. Future work will focus on strengthening the theoretical basis for the approach along the lines of the efficiency analysis for 1SPSA in Spall (1992, Section 4) and on running a more sophisticated numerical study. Nevertheless, the approach as it currently stands seems powerful and relatively easy to apply for use in difficult stochastic optimization problems.

ACKNOWLEDGMENTS

This work was partially supported by the JHU/APL IRAD Progam and U.S. Navy contract N00039-95-C-0002. The author thanks Daniel C. Chin for computational assistance.

REFERENCES

Benveniste, A., M. Metivier, and P. Priouret. 1990. *Adaptive algorithms and stochastic approximation*. New York: Springer-Verlag.

Chin, D. C. 1994. Comparative study of stochastic gradient-free algorithms for system optimization. In *Proceedings of the American Control Conference*, 3070–3075.

Chin, D. C., and J. L. Maryak. 1995. A cautionary note on iterate averaging in stochastic approximation. Submitted to the 13th IFAC World Congress.

Doob, J. L. 1953. *Stochastic processes*. New York: Wiley.

Fabian, V. 1971. Stochastic approximation. In *Optimizing methods in statistics*, ed. J. J. Rustigi, 439–470. New York: Academic Press.

Graham, A. 1981. *Kronecker products and matrix calculus with applications*. New York: Wiley.

Householder, A. S. 1964. *The theory of matrices in numerical analysis*. New York: Dover.

Kiefer, J., and J. Wolfowitz. 1952. Stochastic estimation of a regression function. *Annals of Mathematical Statistics* 23: 462–466.

Kushner, H. J., and D. S. Clark. 1978. *Stochastic approximation methods for constrained and unconstrained systems*. New York: Springer-Verlag.

Lai, T. L. 1985. Stochastic approximation and sequential search for optimum. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, eds. L. M. Le Cam and R. A. Olshen, Vol. II, 557–577. Wadsworth, Belmont, California.

Metivier, M., and P. Priouret. 1984. Applications of a Kushner and Clark lemma to general classes of stochastic algorithms. *IEEE Transactions on Information Theory* 30: 140–151.

Polyak, B. T., and A. B. Juditsky. 1992. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization* 30: 838–855.

Robbins, H., and S. Monro. 1951. A stochastic approximation method. *Annals of Mathematical Statistics* 29: 400–407.

Ruppert, D. 1985. A Newton-Raphson version of the multivariate Robbins-Monro procedure. *Annals of Statistics* 13: 236–245.

Ruppert, D. 1988. Efficient estimators from a slowly converging Robbins-Monro process. Technical Report No. 781, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York. (See also Section 2.8 of Ruppert, D. 1991. Stochastic approximation. In *Handbook of sequential analysis*, eds. B. K. Ghosh and P. K. Sen, 503–529. New York: Marcel Dekker.)

Sacks, J. 1958. Asymptotic distributions of stochastic approximation procedures. *Annals of Mathematical Statistics* 26: 373–405.

Spall, J. C. 1988. A stochastic approximation algorithm for large-dimensional systems in the Kiefer-Wolfowitz setting. In *Proceedings of the IEEE Conference on Decision and Control*, 1544–1548.

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

Spall, J. C. 1994. A second-order stochastic approximation algorithm using only function measurements. In *Proceedings of the IEEE Conference on Decision and Control*, 2472–2477.

Spall, J. C., and J. A. Cristion. 1992. Direct adaptive control of nonlinear systems using neural networks and stochastic approximation. In *Proceedings of the IEEE Conference on Decision and Control*, 878–883.

Spall, J. C., and J. A. Cristion. 1994. Nonlinear adaptive control using neural networks: estimation with a smoothed simultaneous perturbation gradient approximation. *Statistica Sinica* 4: 1–27.

AUTHOR BIOGRAPHY

JAMES C. SPALL (e-mail: james.spall@jhuapl.edu) has been with The Johns Hopkins University Applied Physics Laboratory since 1983, where he is a project leader for several research efforts focusing on problems in statistical modeling and control. For the year 1990, he received the R. W. Hart Prize as principal investigator of the most outstanding Independent Research and Development project at JHU/APL. In 1991, he was appointed to the Principal Professional Staff of the laboratory. Dr. Spall has published numerous research papers in the areas of statistics and control, including articles on subjects such as Kalman filtering, optimization, parameter estimation, and neural networks. He also served as editor and coauthor of the book *Bayesian Analysis of Time Series and Dynamic Models*. He is a member of IEEE, the American Statistical Association, and Sigma Xi, and a fellow of the engineering honor society Tau Beta Pi.