Reprinted from PROCEEDINGS OF THE 1987 AMERICAN CONTROL CONFERENCE, Minneapolis, MN, June 10-12, 1987

TP1 - 4:00

A STOCHASTIC APPROXIMATION TECHNIQUE FOR GENERATING MAXIMUM LIKELIHOOD PARAMETER ESTIMATES†

James C. Spall

The Johns Hopkins University
Applied Physics Laboratory
Laurel, Maryland 20707

ABSTRACT

This paper shows how stochastic approximation (SA) can be used to construct maximum likelihood estimates of system parameters. The procedure described here relies on a derivative approximation other than the usual finite-difference approximation associated with a Kiefer-Wolfowitz SA procedure. This alternative derivative approximation requires fewer, by a factor equal to the dimension of the parameter vector being estimated, computations than the standard finite-difference approximation. Numerical evidence presented in the paper indicates that this SA procedure is, relative to a Kiefer-Wolfowitz procedure, most efficient when considering large-scale systems.

1. INTRODUCTION

Stochastic approximation (SA) is a widely applicable recursive technique for finding roots of equations. SA algorithms are guaranteed to converge under generally weaker conditions on the shape of the function being optimized and the choice of the initial condition than many other iterative techniques. SA has been used extensively in the field of control for dynamic model parameter estimation (e.g., Ljung and Soderstrom [1983], pp. 42-48, Saridis [1974], and El-Sherief and Sinha [1977]). The estimates so generated are not generally optimal in the usual statistical sense (that is, they are not necessarily minimum variance or unbiased). The purpose of this paper is to show how, in contrast to those suboptimal SA estimates, SA can be used to generate maximum likelihood estimates (MLEs) (which, of course, do have certain optimality properties) when it is either not feasible or very difficult to implement a steepest descent, Newton-Raphson, or scoring procedure. The key to the SA algorithm here is a gradient approximation that differs from the usual finite-difference approximation.

We will assume here that, as usual, the MLE is found by determining the (consistent) root of the score equation:

$$s(\theta) = \frac{\partial L(\theta)}{\partial \theta} = 0 , \qquad (1.1)$$

where $L(\theta)$ is the likelihood or log-likelihood function for the parameter vector $\theta \in R^p$. To implement a steepest descent, Newton-Raphson, or scoring technique, it is required that $s(\theta)$ and $\partial s/\partial \theta^T$ (Newton-Raphson) or $E(\partial s/\partial \theta^T)$ (scoring) be evaluated at different values of θ as the algorithm proceeds to convergence near the consistent root of (1.1), say θ^*. When it is difficult to evaluate these quantities (perhaps because each evaluation requires numerous Kalman filter runs as in Goodrich and Caines [1979] or Porter et al. [1983]), a Kiefer-Wolfowitz [1952] SA (KWSA) procedure can be used to estimate θ^*, which only requires that $L(\theta)$ be evaluated. The key quantity in the implementation of KWSA is the usual finite-difference approximation to the derivative, $s(\theta)$, at each value of θ arising in the iterations. Each such approximation requires $2p$ evaluations of $L(\theta)$.

The goal of this paper is to show how SA can be used to form MLEs with a derivative approximation *other* than the standard finite-difference approximation used in KWSA. In particular, this alternative derivative approximation requires only two evaluations of $L(\theta)$, instead of $2p$ evaluations. We will evaluate the trade-off between the reduced computation per iteration associated with the alternative derivative approximation and the (expected) increased number of iterations to converge (relative to KWSA). As discussed later, numerical experience indicates that this trade-off is favorable when p is moderately large.

Section 2 presents the SA algorithm of this paper with its alternative derivative approximation and associated regularity conditions. Section 3 presents two theorems that justify the use of the alternative derivative approximation given in Section 2. Section 4 gives several numerical studies that illustrate the effectiveness of the SA procedure here relative to a KWSA procedure. Section 5 contains some concluding remarks.

2. OVERVIEW OF SA TECHNIQUE FOR FINDING MLEs

This section is divided into two subsections. Subsection 2.1 is a brief description of the SA algorithm that is of interest here, together with the key regularity conditions that are to be satisfied. Subsection 2.2 discusses the estimate for $s(\theta)$ that will be used in the SA algorithm.

2.1 The SA Algorithm and Associated Regularity Conditions

As pointed out in Section 1, SA is an interative root-finding technique. Since we are interested in finding the root θ^* of $s(\theta) = 0$, we will discuss the SA algorithm in this context. SA applies when the function for which a zero is to be found ($s(\theta)$ our case) is not known precisely. Of course, in ML estimation, $s(\cdot)$ is known precisely for any fixed θ. Our goal here is to replace the precisely cal-

†This work has been partially supported by U.S. Navy Contract N00039-87-C-5301. The author is most grateful to D. C. Chin of JHU/APL for his assistance in producing the numerical results of Section 4.

culated $s(\cdot)$ (which, as mentioned in Section 1, can be computationally very burdensome) with a much simpler-to-compute approximation, $\hat{s}(\cdot)$. Relative to steepest descent or scoring, this will lower the per-iteration computational burden significantly, albeit at the likely expense of an increased number of iterations.

Letting $\hat{\theta}(k)$ denote the estimate for θ^* at the kth iteration, where $\hat{\theta}(k) \in \Lambda(k) \subseteq R^p$ a.s., the standard SA algorithm has the form

$$\hat{\theta}(k) = \hat{\theta}(k-1) - a(k-1)\,\hat{s}(\hat{\theta}(k-1)), \qquad (2.1)$$

where the gain sequence $\{a(k)\}$ satisfies the following conditions: $\lim_{k\to\infty} a(k) = 0$, $\Sigma_{k=0}^{\infty} a(k) = \infty$, $\Sigma_{k=0}^{\infty} a(k)^2 < \infty$. A simple example of an $a(k)$ that satisfies these conditions is $a(k) = 1/(k+1)^{\alpha}$, $\frac{1}{2} < \alpha \leq 1$. Note also the close relationship of (2.1) to the method of steepest descent, the difference being that in steepest descent $s(\cdot)$ replaces $\hat{s}(\cdot)$.

There are a number of techniques for accelerating the convergence of the standard SA algorithm given in (2.1),[1] but they will not be considered in detail in this paper. Rather we will focus on the performance of (2.1) with $\hat{s}(\cdot)$ as defined below, and contrast (in Section 4) this performance with that of the closely related KWSA and steepest descent algorithms. We believe, however, that this "baseline" study will provide insight into the potential usefulness of $\hat{s}(\cdot)$ as it might apply in an accelerated algorithm.

The main conditions on $\hat{s}(\cdot)$ that are generally imposed for the SA algorithm in (2.1) are, in terms of the error $e(\hat{\theta}(k)) \equiv \hat{s}(\hat{\theta}(k)) - s(\hat{\theta}(k))$,

$$E[e(\hat{\theta}(k))|\hat{\theta}(k)] = 0 \qquad \forall\, k, \qquad (2.2)$$

$$E[\,\|e(\hat{\theta}(k))\|_2^2|\hat{\theta}(k)] \leq c < \infty \qquad \forall\, k, \qquad (2.3)$$

$$\{e(\hat{\theta}(k))|\hat{\theta}(k)\}_{k=0}^{\infty} \quad \text{mutually independent,} \qquad (2.4)$$

where $\|\cdot\|_2$ denotes the L^2 (Euclidean) norm. Under these and the above-mentioned conditions on $a(k)$, together with certain other regularity conditions (see, e.g., Blum [1954] or Kushner and Clark [1978]), $\hat{\theta}(k) \overset{\text{a.s.}}{\to} \theta^*$ as $k \to \infty$. Note that all expectations and probabilities in this paper are conditioned on the data appearing in the likelihood function (x in $L(\theta) = L(\theta|x)$, say) being fixed.

2.2 The Estimate, $\hat{s}(\cdot)$

We now define our estimate for $s(\cdot)$. Let $\Delta(k) \in \Omega(k) \subseteq R^p$ a.s. be a vector of p mutually independent random variables $\{\Delta_1(k), \ldots, \Delta_p(k)\}$. Furthermore let $\{\Delta(k)\}_{k=1}^{\infty}$ be a mutually independent sequence. Our estimate, $\hat{s}(\cdot)$, at the kth iteration of the SA algorithm will then be

$$\hat{s}(\hat{\theta}(k)) = \begin{bmatrix} \dfrac{L[\hat{\theta}(k)+\Delta(k)] - L[\hat{\theta}(k)-\Delta(k)]}{2\Delta_1(k)} \\ \vdots \\ \dfrac{L[\hat{\theta}(k)+\Delta(k)] - L[\hat{\theta}(k)-\Delta(k)]}{2\Delta_p(k)} \end{bmatrix}. \qquad (2.5)$$

Note that this estimate differs from the usual finite difference gradient approximation arising in KWSA since the numerator is the same for all elements of the vector. Conditions under which $\hat{s}(\cdot)$ is an appropriate estimator for $s(\cdot)$ (i.e., satisfies (2.2)–(2.4)) are given below and in Section 3.

Consider conditions (2.3) and (2.4). (2.4) is clearly satisfied since the $\Delta(k)$ are mutually independent. Now let us consider the second-moment condition, (2.3). For now, let k be fixed (and thus the argument k is suppressed). Let $\Gamma \subseteq R^p$ be the subspace in which $\hat{\theta} \pm \Delta$ lies (a.s.). Then, assuming that $s(\theta)$ is continuous and bounded on Γ, the mean value theorem indicates that

$$L(\hat{\theta} \pm \Delta) = L(\hat{\theta}) + s(\bar{\theta}^{\pm})^T(\pm\Delta)$$

where $\bar{\theta}^{\pm}$ denotes a point on the line segment between $\hat{\theta}$ and $\hat{\theta} \pm \Delta$ (as appropriate). Thus the ith component of $\hat{s}$ satisfies

$$\hat{s}_i(\hat{\theta}) = \frac{[s(\bar{\theta}^+) - s(\bar{\theta}^-)]^T\Delta}{\Delta_i},$$

so $|\hat{s}_i(\hat{\theta})| \leq c' \Sigma_{j=1}^{p} |\Delta_j/\Delta_i|$ where $c' = 2\max_j \sup_{\theta\in\Gamma} |s_j(\theta)|$, which is well-defined (bounded) since $s(\theta)$ is bounded on Γ. Thus by the Minkowski inequality,

$$E(\|\hat{s}(\hat{\theta})\|_2^2|\hat{\theta}) \leq (c')^2 \sum_{i=1}^{p}\left\{\sum_{j=1}^{p}\left[E\left(\frac{\Delta_j}{\Delta_i}\right)^2\right]^{1/2}\right\}^2, \qquad (2.6)$$

which, of course, is bounded when $E(\Delta_j/\Delta_i)^2 \leq c'' < \infty$ for all i, j (e.g., when Δ_i are Bernoulli distributed with nonzero outcomes). Now, (2.3) is satisfied at any k if the bound in (2.6) is finite for the specified $\hat{\theta} = \hat{\theta}(k)$ (recall that $s(\hat{\theta}(k))$ is considered a constant relative to the measure $P(\cdot|\hat{\theta}(k))$. Thus (2.3) is satisfied if $s(\theta)$ is continuous (and bounded) on $\cup_{k=0}^{\infty} \Gamma(k)$ and if $E(\Delta_j(k)/\Delta_i(k))^2 \leq c'' < \infty\ \forall\, i,j = 1,2, \ldots, p;\ k = 1,2, \ldots, \infty$.

Section 3 is devoted to establishing conditions under which the unbiasedness condition, (2.2), is satisfied to within a certain small error term.

3. THE BIAS IN $\hat{s}(\cdot)$

3.1 Introduction

This section presents two theorems that give conditions under which $\hat{s}(\cdot)$ as given in (2.5) is an unbiased estimator of $s(\cdot)$ to within a certain $O(\delta^2)$ error term, where δ is some positive constant that can be made arbitrarily small subject to computer accuracy limitations in forming a 0/0 type quantity.[2] These theorems (and associated corollaries) are presented in Subsection 3.2. In Subsection

[1] The "second-order" methods for SA acceleration replace the gain $a(k)$ by a matrix related to the Hessian of $L(\theta)$—see, e.g., Ruppert [1985], Ljung and Soderstrom [1983], pp. 46-47. These accelerated SA procedures involve a trade-off between greater computational burden per iteration and fewer iterations, and tend to be especially effective near θ^* where the algorithm defined in (2.1) can be very slow. There are also non-second-order methods for increasing the convergence rate; e.g., Kesten [1958] gives an adaptive scheme for choosing the $a(k)$, and Koch and Spall [1986] present a "multistep" procedure that involves a certain reuse of data (which increases the effective convergence rate in terms of available data).

3.3, a brief discussion is included that compares the bias in $\hat{s}(\cdot)$ with that in the usual finite difference derivative approximation.

The following notation will be employed. $T^{(m)}(\hat{\theta} \pm \Delta)$ will denote an mth-order Taylor expansion of $L(\hat{\theta} \pm \Delta)$ about $\hat{\theta}$, while $\Delta^{[m]} = \Delta \otimes \Delta \otimes \ldots \otimes \Delta$ (mth-order Kronecker product). $M(\delta)$ and $\beta(m)$ will denote some functions of δ and m satisfying conditions given in the theorem statements. For any functions $f(\cdot)$ and $g(\cdot)$, $f(x) \sim g(x)$ implies that $f(x)/g(x) \to 1$ as x approaches a given limiting value. Finally, $L^{(m)} = \partial^m L/(\partial\theta^T)^m$ with individual components $L^{(m)}_{i_1 i_2 \ldots i_m} = \partial^m L/\partial\theta_{i_1}\partial\theta_{i_2}\ldots\partial\theta_{i_m}$ for $1 \le i_1, i_2, \ldots, i_m \le p$. Note that $L^{(m)} \in R^{p^m}$.

3.2 Theorems and Corollaries

Theorem 1 below presents a bound for the bias $E[e(\hat{\theta})|\hat{\theta}] = E[\hat{s}(\hat{\theta}) - s(\hat{\theta})|\hat{\theta}]$ under the assumption that $L(\theta)$ is of class $C^{(\infty)}$ in a region about $\hat{\theta}$. Two corollaries follow the theorem; they consider important special cases for the key conditions ((3.1) and (3.2)) in the theorem statement.

Theorem 1. Suppose that Δ_j is symmetrically distributed about 0, $|\Delta_j| \le \delta$ a.s., and $|\Delta_j|^{-1} \le M(\delta) = O(\delta^{-2})$ a.s. $(\delta \to 0)$ $\forall\, j = 1,2, \ldots, p$. Furthermore, suppose that $L^{(m)}(\hat{\theta} \pm \Delta)$ exists $\forall\, m = 1,2, \ldots,$ and

$$|L^{(m)}_{i_1 i_2 \ldots i_m}(\hat{\theta} \pm \Delta)| \le \beta(m) \tag{3.1}$$

a.s. on Ω for any $\hat{\theta} \in \Lambda$ where

$$\beta(m) = O(m!/(\delta p + \epsilon)^m) \qquad (m \to \infty) \tag{3.2}$$

for any $\epsilon > 0$. Then $\forall\, j = 1,2, \ldots, p$,

$$|E[e_j(\hat{\theta})|\hat{\theta}]| \le b_j(\delta)$$

$$\equiv \delta^{-1}\sum_{i=1}^{\infty}\left[(p\delta)^{2i+1} - ((p-1)\delta)^{2i+1}\right]\frac{\beta(2i+1)}{(2i+1)!} \tag{3.3a}$$

$$= O(\delta^2) \qquad (\delta \to 0)\ . \tag{3.3b}$$

Proof. (3.1) and (3.2) imply that the remainder of an mth-order Taylor expansion of $L(\hat{\theta} \pm \Delta)$ about $\hat{\theta}$ satisfies

$$|L(\hat{\theta} \pm \Delta) - T^{(m)}(\hat{\theta} \pm \Delta)| \le \frac{(p\delta)^{m+1}}{(m+1)!}\beta(m+1) \to 0$$

as $m \to \infty$, which implies that

$$\hat{s}_l(\hat{\theta}) = \frac{\sum_{i=0}^{\infty} L^{(2i+1)}(\hat{\theta})\Delta^{[2i+1]}/(2i+1)!}{\Delta_l}\ .$$

Now, since $E(\Delta_j/\Delta_l) = 0\ \forall\, j \ne l$,

$$E[\hat{s}_l(\hat{\theta})|\hat{\theta}] = s_l(\hat{\theta}) + E[e_l(\hat{\theta})|\hat{\theta}]\ ,$$

where

$$E[e_l(\hat{\theta})|\hat{\theta}] = E\left[\frac{\sum_{i=1}^{\infty} L^{(2i+1)}(\hat{\theta})\ \Delta^{[2i+1]}/(2i+1)!}{\Delta_l}\,\middle|\,\hat{\theta}\right]\ . \tag{3.4}$$

Note that for any $m \in \{3,5,7, \ldots\}$, $(p-1)^m$ (out of a total of p^m) terms will not contain a Δ_l. Thus, using (3.1) and the fact that the Δ_i's are symmetrically distributed, we know that for $1 \le i_1, i_2, \ldots, i_m \le p$,

$$E\left[\frac{L^{(m)}_{i_1,i_2,\ldots i_m}\ \Delta_{i_1}\Delta_{i_2}\ \ldots\ \Delta_{i_m}}{\Delta_l}\,\middle|\,\hat{\theta}\right]$$

$$\begin{cases} = 0 \text{ for } (p-1)^m \text{ terms} \\ \le \beta(m)\delta^{m-1} \text{ in magnitude for } p^m - (p-1)^m \text{ terms} \end{cases}$$

(where the point of evaluation $\hat{\theta}$ has been suppressed in $L^{(m)}$), which implies that[3]

$$|E(L^{(m)}\ \Delta^{[m]}/\Delta_l|\hat{\theta})| \le [p^m - (p-1)^m]\,\beta(m)\,\delta^{m-1}\ . \tag{3.5}$$

Now define

$$D(\delta) = M(\delta)\sum_{i=1}^{\infty}\beta(2i+1)\ (p\delta)^{2i+1}/(2i+1)!\ ,$$

which is finite by (3.2). Then from (3.1) and the fact that $|\Delta_l|^{-1} \le M(\delta)$, we know that

$$\left|\frac{\sum_{i=1}^{m} L^{(2i+1)}\ \Delta^{[2i+1]}/(2i+1)!}{\Delta_l}\right| \le D(\delta) \tag{3.6}$$

$\forall\, m = 1,2, \ldots$. Furthermore, L'Hopital's rule applied to $D(\delta)$ yields $D(\delta) = O(1)$ $(\delta \to 0)$, and so $D(\delta)$ is integrable ($\int_\Omega D(\delta)\, dP_\Delta = D(\delta)$) as $\delta \to 0$. Thus from (3.6) and Lebesgue's Dominated Convergence Theorem, the expectation and sum on the r.h.s. of (3.4) can be interchanged, which by (3.5) yields (3.3a).

Now, to show (3.3b), L'Hopital's rule can be applied to (3.3a), where we find that

$$b_l(\delta) \sim [p^3 - (p-1)^3]\beta(3)\,\delta^2/6 \tag{3.7}$$

$$= O(\delta^2)\ ,$$

which completes the proof since l was chosen arbitrarily. *Q.E.D.*

The corollaries below give closed-form expressions for the bound in (3.3a) for two important special cases. Both corollaries assume that all conditions of Theorem 1 hold.

Corollary 1-1. Suppose that

$$\beta(m) = c\rho^m \tag{3.8}$$

[2] It is also of interest to know the effect of the $O(\delta^2)$ bias on the convergence properties of $\hat{\theta}(k)$. The author has not had an opportunity to explore this issue thoroughly, but believes that a martingale difference sequence approach such as that of Solo [1982] might be useful in this regard. The numerical studies of Section 4 indicate that the bias appears to have negligible effect.

[3] Note that combinatorics arguments could significantly reduce the bound in (3.5) (and hence the bound in 3.3a)). These arguments would exploit the fact that all terms of the form $\Delta_{j_1}\Delta_{j_2}\ldots\Delta_{j_{m-1}} = \Delta_{i_1}\Delta_{i_2}\ldots\Delta_{i_m}/\Delta_l$ (i.e., at least one $i_j = l$) would have expectation 0 when any odd powers of Δ_{j_k} exist (i.e., if $m - 1 = 4$ and $\Delta_{j_1}\Delta_{j_2}\Delta_{j_3}\Delta_{j_4} = \Delta_1\Delta_2\Delta_p^2$). As of this writing, the author has not had an opportunity to explore this approach in detail.

for some $0 < c, \rho < \infty$. Then $\forall\, j = 1,2, \ldots, p$,

$$b_j(\delta) = c\delta^{-1}[\sinh(p\rho\delta) - \sinh((p-1)\rho\delta) - \rho\delta] \tag{3.9a}$$

$$\sim \frac{c[p^3 - (p-1)^3]\rho^3}{6}\delta^2 . \tag{3.9b}$$

Corollary 1-2. Suppose that

$$\beta(m) = c\rho^m m! \tag{3.10}$$

for some $0 < \rho < (\delta p)^{-1}$. Then $\forall\, j = 1,2, \ldots, p$,

$$b_j(\delta) = c\left[\frac{p^3\rho^3\delta^2}{1 - (p\rho\delta)^2} - \frac{(p-1)^3\rho^3\delta^2}{1 - ((p-1)\rho\delta)^2}\right] \tag{3.11a}$$

$$\sim c[p^3 - (p-1)^3]\rho^3\delta^2 . \tag{3.11b}$$

Proof of Corollaries. Clearly both (3.8) and (3.10) satisfy (3.2). The bounds in (3.9a) and (3.11a) follow immediately using (3.3a) and standard series expansions. The asymptotic ($\delta \to 0$) equivalents follow from (3.7).

Note. For large p, the bounds in (3.7b) and (3.9b) could be further simplified by noting that $p^3 - (p-1)^3 \sim 3p^2$.

Theorem 1 and its corollaries require that $L(\theta)$ be of class $C^{(\infty)}$. Theorem 2 below relaxes this condition at the expense of a condition that is more difficult to check regarding the remainder term in an rth-order Taylor expansion.

Theorem 2. Suppose that the conditions on Δ_j, $\forall\, j = 1,2, \ldots, p$, given in Theorem 1 are satisfied. Furthermore, suppose that for some odd integer $r \geq 3$ there exist $\beta(1), \beta(2), \ldots, \beta(r)$ and $0 < K < \infty$ such that for $\hat{\theta} \in \Lambda$,

$$L_{i_1 i_2 \ldots i_m}(\hat{\theta}) \leq \beta(m)$$

$$\beta(m) \leq K,\ m = 1,2, \ldots, r.$$

If

$$T^{(r)}(\hat{\theta} + \Delta) - L(\hat{\theta} + \Delta) = T^{(r)}(\hat{\theta} - \Delta) - L(\hat{\theta} - \Delta) \tag{3.12}$$

a.s. on Ω for the $\hat{\theta}$ above, then $\forall\, j = 1,2, \ldots, p$,

$$|E[e_j(\hat{\theta})|\hat{\theta}]| \leq b_j^{(r)}(\delta)$$
$$\equiv \delta^{-1}\sum_{i=1}^{(r-1)/2}[(p\delta)^{2i+1} - ((p-1)\delta)^{2i+1}]\frac{\beta(2i+1)}{(2i+1)!} = O(\delta^2) . \tag{3.13}$$

Note. The condition that r be odd is, in fact, not a restriction since any given even order term in $T^{(m)}(\hat{\theta} + \Delta)$ equals the same order term in $T^{(m)}(\hat{\theta} - \Delta)$ for any m.

Proof. From (3.12), arguments analogous to those leading to (3.4) imply that $E[e_l(\hat{\theta})|\hat{\theta}]$ is given by the expression in (3.4) with the upper limit of ∞ in the sum within $E[\cdot\,|\hat{\theta}]$ on the r.h.s. replaced by $(r-1)/2$. Then, using arguments identical to those following (3.4) in the proof of Theorem 1, (3.13) follows. Finally the fact that $b_j'(\delta) = O(\delta^2)$ follows by the bound on $\beta(m)$ and L'Hopital's rule. *Q.E.D.*

The corollary below considers an important special case for $\beta(m)$.

Corollary 2-1. Suppose that $\beta(m) = c\rho^m m!$ for any $0 < c, \rho < \infty$. Then $\forall\, j = 1,2, \ldots, p$,

$$b_j(\delta) = cp\rho\left[\frac{(p\rho\delta)^2 - (p\rho\delta)^{r+1}}{1 - (p\rho\delta)^2}\right]$$
$$-c(p-1)\rho\left[\frac{((p-1)\rho\delta)^2 - ((p-1)\rho\delta)^{r+1}}{1 - ((p-1)\rho\delta)^2}\right]$$
$$\sim c[p^3 - (p-1)^3]\rho^3\delta^2 .$$

Proof. The proof is straightforward using (3.13) and standard geometric series arguments.

3.3 Comparison with Bias in Usual Finite Difference Approximation

We close this section with a brief discussion of the bias in the usual discrete difference approximation to $s(\cdot)$, say $\tilde{s}(\cdot)$, that would be used in KWSA (see Section 1) and contrast this with the bias in $\hat{s}(\cdot)$. Here

$$\tilde{s}_i(\hat{\theta}) = \frac{L(\hat{\theta} + \Delta_i I_i) - L(\hat{\theta} - \Delta_i I_i)}{2\Delta_i}$$

for all $i = 1,2, \ldots, p$ where I_i is the ith column of a $p \times p$ identity matrix. Under the assumptions of Corollary 1-1 (the bound $\beta(m)$ being the usual sufficient condition for $L(\cdot)$ to be analytic), we find (using arguments identical to those of the proof of Theorem 1)

$$\tilde{s}_l(\hat{\theta}) = \frac{\sum_{i=0}^{\infty} L^{(2i+1)}(\hat{\theta})(\Delta_l I_l)^{[2i+1]}}{\Delta_l} .$$

We then obtain the following bound for $|E[\tilde{s}(\hat{\theta}) - s(\hat{\theta})|\hat{\theta}]|$:

$$c\delta^{-1}\sum_{i=1}^{\infty}\frac{(\rho\delta)^{2i+1}}{(2i+1)!} = c\delta^{-1}[\sinh(\rho\delta) - \rho\delta]$$
$$\sim \frac{c\rho^3}{6}\delta^2 .$$

Contrasting $c\rho^3\delta^2/6$ with the error bound in (3.9b), we see that the bias in $\tilde{s}(\cdot)$ is less than that of $\hat{s}(\cdot)$ by a factor of $p^3 - (p-1)^3 \sim 3p^2$. As mentioned in Section 1, however, $\tilde{s}(\cdot)$ requires p times more computations than $\hat{s}(\cdot)$. Also note that in light of footnote 3, the "real" bias in $\hat{s}(\cdot)$ is somewhat less than that given in (3.9b), while the bias bound given above for $\tilde{s}(\cdot)$ is somewhat closer to the "real" bias in $\tilde{s}(\cdot)$ since the issue in footnote 3 is not relevant for $\tilde{s}(\cdot)$. Thus the factor of $p^3 - (p-1)^3$ given above overestimates the relative differences in the biases of $\hat{s}(\cdot)$ and $\tilde{s}(\cdot)$.

The practical effects of using $\hat{s}(\cdot)$ and $\tilde{s}(\cdot)$ in a problem of finding MLEs will be illustrated in the next section.

4. NUMERICAL STUDIES

4.1 Introduction

This section presents several numerical studies that illustrate how (2.1) with the input given in (2.2) can be used to calculate MLEs. For convenience, we will refer to this algorithm as ADSA (Alternative Derivative SA). The goals of these studies are threefold:

1. To gain insight into how ADSA compares with KWSA (which is closely related to steepest descent (SD));
2. To examine the performance of ADSA as δ (see Subsection 3.2) and p (dimension of θ) vary;
3. To suggest ideas for practical implementation of ADSA.

Note that we are *not* comparing ADSA with scoring or Newton-Raphson. The reason for this was given in Subsection 2.1, where it was pointed out that there are second-order (and other) procedures for accelerating the convergence of the SA algorithm. These second-order algorithms are the SA analog of scoring or Newton-Raphson. The author believes, however, that the preformance of $\hat{s}(\cdot)$ (from (2.2)) in the first-order algorithm considered here (vis-à-vis KWSA) provides insight into the potential applicability of $\hat{s}(\cdot)$ in a second-order algorithm.

The MLE problem we consider pertains to the estimation of the "signal" covariance matrix in a signal-plus-noise problem with independent nonidentically distributed noise. In particular, we assume that data x_i distributed $N(0, \Sigma + P_i)$, $i = 1,2, \ldots, N$, are obtained, and Σ is to be estimated with the $\{P_i\}$ known. Such problems pertain, for example, to the estimation of initial state parameters in a Kalman filter model (see, e.g., Shumway, Olsen, and Levy [1981] or Haley, Garner, and Levine [1984]) and have a number of interesting characteristics (see, e.g., Smith [1985] or Spall [1986]). In the nonidentical P_i case, no closed-form solution to $s(\cdot) = 0$ exists, and thus a numerical algorithm must be applied.

For the numerical studies here, Σ will be a diagonal matrix and θ will represent the vector of diagonal elements, i.e., $\Sigma = \text{diag}[\theta_1, \theta_2, \ldots, \theta_p]$, and $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^T$. The form for the gain $a(k)$ appearing in (2.1) is

$$a(k) = \frac{A}{(k + 1)^{\alpha}},$$

where $A > 0$ and α is as given below. The $\{\Delta_j(k)\}$ will be generated as independent Bernoulli random variables with outcomes $\pm\delta$, $\text{Prob}(\delta) = \text{Prob}(-\delta) = \frac{1}{2}$, for all $j = 1, 2, \ldots, p$ and $k = 1, 2, \ldots$. In the KWSA algorithm, (2.1) will be used with $\tilde{s}(\cdot)$ replacing $\hat{s}(\cdot)$, where each of the p elements in $\tilde{s}(\cdot)$ is given by the finite difference approximation:

$$\tilde{s}_j(\tilde{\theta}(k)) = \frac{L(\tilde{\theta} + \Delta_j I_j/(k + 1)^{\gamma}) - L(\tilde{\theta} - \Delta_j I_j/(k + 1)^{\gamma})}{2\Delta_j/(k + 1)^{\gamma}}, \tag{4.1}$$

where $\gamma > 0$ and $\tilde{\theta}$ represents the KWSA estimate. (Note that this definition for $\tilde{s}$ differs by the factor of $(k + 1)^{\gamma}$ from that given in Subsection 3.3. The reason for not including $(k + 1)^{\gamma}$ in Subsection 3.3 was that the goal there was to compare $\hat{s}(\cdot)$ and $\tilde{s}(\cdot)$ for the same δ; the "effective δ" in (4.1) is $\delta/(k + 1)^{\gamma}$.)[4] Letting $\alpha = \frac{3}{4} + \epsilon, 0 < \epsilon \leq \frac{1}{4}$, and $\gamma = \frac{1}{4}$ satisfies the conditions given after (2.1) and in Kiefer and Wolfowitz [1952] or Blum [1954] for the KWSA algorithm. In the studies below, we set $\epsilon = 0.0001$ (we found that making α as near $\frac{3}{4}$ as possible tended to speed convergence).

Note that under the Bernoulli assumptions, the conditions in Theorem 1 of Section 3 on the $\Delta_i(k)$'s are satisfied. Moreover, based on calculations for the $p = 1$ case, it appears that the derivative bound in Corollary 1-2, (3.10), applies to this setting.[5] The author has not had the opportunity to carefully establish that the derivative bound of (3.10) holds in the $p \geq 2$ case but is confident that a bound analogous to that of the $p = 1$ case can be established since there is no essential difference in the form of $L(\cdot)$ (as shown in Spall [1985], the calculation of $L^{(m)}$ can get onerous for multivariate θ and $m \geq 3$).

There are two remaining subsections in this section. Subsection 4.2 considers the $p = 15$ case and compares ADSA with KWSA. Subsection 4.3 considers the $p = 3$ case and, in addition to comparing ADSA with KWSA, reports on several related sensitivity studies.

4.2 The p = 15 Case

Our goal here is to compare ADSA with KWSA for the case where $\theta \in R^{15}$. However, as can be seen from the form for $\tilde{s}(\cdot)$ (relative to $s(\cdot)$) and as will be illustrated in Subsection 4.3, the performance of KWSA is close to that of SD; thus, these studies can also be thought of as providing insight into the relative performance of ADSA and SD.

The data $\{x_i\}$ were generated using an IBM 3083 (and associated pseudorandom number generator) according to the $N(0, \Sigma + P_i)$ distribution with

$$\Sigma = 225I$$
$$N = 60$$
$$P_i = A_i A_i^T$$
$$A_i \in R^{15 \times 30}$$

with each element of A_i generated uniformly (and independently) on $(-1, -0.001) \cup (0.001, 1)$, and all A_i, $i = 1, 2, \ldots N$ independently generated. Note that with the P_i's generated in this manner, the P_i's will be nonidentical, with no particular pattern.

As a way to compare ADSA and KWSA, we will work with the quantity

$$\bar{L}(\theta) \equiv L(\theta^*) - L(\theta)$$

at values of θ corresponding to $\hat{\theta}$ (ADSA) and $\tilde{\theta}$ (KWSA). The value θ^* was determined by the application of a scoring algorithm (corroborated by the results of a Newton-Raphson algorithm). $\bar{L}(\cdot)$ is convenient to work with here in that it compensates for the fact that $L(\cdot)$ is flat in a wide range about θ^* relative to the baseline value $L(\theta^*)$ (= 6578.3) and is nonnegative with $\bar{L}(\theta^*) = 0$.

---

[4] Note that $\hat{s}(\cdot)$ in ADSA could have also included the $(k + 1)^{\gamma}$ damping of KWSA. We chose not to do this here so that the $O(\delta^2)$ bias in $\hat{s}(\cdot)$ would manifest itself without being damped (thus helping us see its effect). In fact, however, there was no difference in the performance of ADSA with and without $(k + 1)^{\gamma}$ damping for the one case where both approaches were compared—see Subsection 4.2.

[5] For the $p = 1$ case, $L(\theta) = -\frac{1}{2}\sum_{i=1}^{N}[\log(\theta + P_i) + x_i^2/(\theta + P_i)]$ + constant, from which it is easily derived that

$$|L^{m}(\theta)| \leq cm!\rho^{m},$$

where $c = \frac{1}{2}N \max_i[(\theta + P_i + x_i^2)/(\theta + P_i)]$ and $\rho = [\min_i(\theta + P_i)]^{-1}$. Thus (3.10) is applicable when $\theta + P_i > 0 \; \forall \; \theta \in \Gamma$.

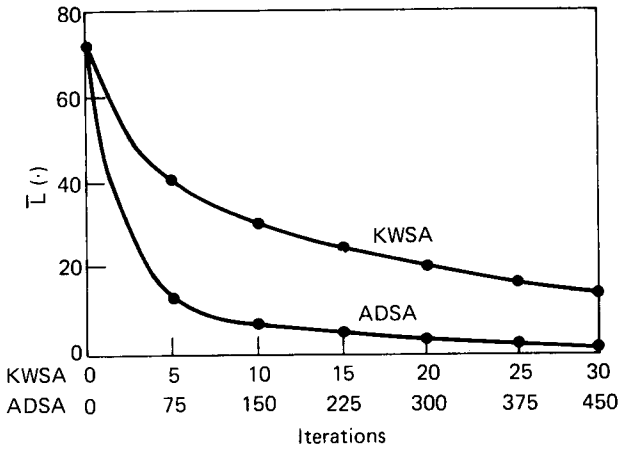

Fig. 1 Comparison of ADSA and KWSA for iteration pairs requiring equal number of L (·) evaluations.

Figure 1 contrasts the performance of ADSA and KWSA for the scenario described above. We took $\hat{\theta}(0) = \tilde{\theta}(0) = (400, 400, \ldots, 400)^T$ and generated the $\Delta_i(k)$'s for both algorithms from the same seed with $\delta = 0.01$. Also $A = 1000$. Since $p = 15$, each iteration of KWSA requires 30 evaluations of $L(\cdot)$, in contrast to the two evaluations required for ADSA. Thus, at each point along the horizontal axis the iterations indicated represent an equivalent number of $L(\cdot)$ evaluations (which represent essentially all of the computations required) for the ADSA and KWSA procedures.

Based on its ability to minimize $\bar{L}(\cdot)$ (i.e., maximize $L(\cdot)$), ADSA performs significantly better than KWSA at all iteration pairs considered. It also performs significantly better in terms of accuracy of the estimate. For example, at the terminal iteration pair in Fig. 1,

$$\frac{\|\hat{\theta}(450) - \theta^*\|_1}{\|\tilde{\theta}(30) - \theta^*\|_1} = \frac{339.3}{1074} = 0.316 ,$$

where $\|\cdot\|_1$ represents L^1 norm (for comparison, $\|\theta^*\|_1 = 2893.5$ and $\|\hat{\theta}(0) - \theta^*\|_1 = \|\tilde{\theta}(0) - \theta^*\|_1 = 3106.5$).

The ADSA algorithm was also tested under several scenarios different from that of Fig. 1. In one case $\hat{\theta}(0)$ was set very near θ^* with other SA parameters (A, α, δ) as in Fig. 1. The algorithm was then tested to make sure it would not diverge, given the realtively large values of $a(0)$, $a(1)$, etc. compared to the gain values near $a(450)$ in Fig. 1; no such divergence occurred and ADSA converged within a relatively few iterations. We also tested the algorithm with $\delta = 0.1$ and $\delta = 0.001$ (versus $\delta = 0.01$ above). The estimate (and, of course, the likelihood) values were the same for all iterations to within the four decimal points considered. Thus ADSA does not appear to be sensitive to δ, at least within the range considered. Finally, ADSA was run with the $(k + 1)^\gamma$ damping of KWSA (so $\hat{s}_i(\hat{\theta}(k)) = [L(\hat{\theta} + \Delta/(k + 1)^\gamma) - L(\hat{\theta} - \Delta/(k + 1)^\gamma)]/[2\Delta_i/(k + 1)^\gamma]$), and it was found (again) that there was no difference in the estimate or likelihood values for all iterations considered.

4.3 The p = 3 Case

A study similar to that above was performed for the $p = 3$ case to see if the dimension of θ had a significant effect on the relative performance of ADSA and KWSA. In this study the P_i matrices were generated in the same manner as above, and $\Sigma = \text{diag}\,(4, 9, 16)$. We let $N = 30$, $A = 1$, and, as above, took $\alpha = 0.7501$, $\gamma = 0.25$, and $\delta = 0.01$. The starting point for both ADSA and KWSA ($\hat{\theta}(0)$ and $\tilde{\theta}(0)$) was taken as (3, 8, 15).

The relative performance of ADSA and KWSA here was qualitatively the same as described in Subsection 4.2 for the $p = 15$ case. For example, at 180 ADSA and 60 KWSA interations ($180/60 = p$), we find that $\bar{L}(\hat{\theta}(180)) = 0.011$ versus $\bar{L}(\tilde{\theta}(60)) = 0.042$; likewise we find

$$\frac{\|\hat{\theta}(180) - \theta^*\|_1}{\|\tilde{\theta}(60) - \theta^*\|_1} = \frac{0.610}{1.25} = 0.49$$

(for comparison, $\|\theta^*\|_1 = 24.40$ and $\|\hat{\theta}(0) - \theta^*\|_1 = \|\tilde{\theta}(0) - \theta^*\|_1 = 3.21$). Similar behavior was seen for other iterations considered (≤ 180 for ADSA, ≤ 60 for KWSA).

We also compared the performance of KWSA and SD. The purpose of this was to gain insight into how representative the KWSA results are of SD (we would expect good agreement since $\bar{s}(\cdot) \approx s(\cdot)$ for small δ). For all iterations considered, the estimates of θ and corresponding values of $\bar{L}(\cdot)$ were close. For example, with $\tilde{\theta}_{SD}$ denoting the SD estimate, we have $\bar{L}(\tilde{\theta}_{SD}(60)) = 0.039$ versus $\bar{L}(\tilde{\theta}(60)) = 0.042$, and $\|\tilde{\theta}_{SD}(60) - \tilde{\theta}(60)\|_1 = 0.053$ relative to baseline values $\|\tilde{\theta}_{SD}(60)\|_1 = 25.27$ and $\|\tilde{\theta}(60)\|_1 = 25.29$. Thus, it appears that the relative performance of ADSA and KWSA given above is indicative of the relative performance of ADSA and SD.

One comment should be made regarding the $p = 3$ case. It sometimes happened that $L(\hat{\theta}(k)) < L(\hat{\theta}(k - 1))$ (i.e., $\bar{L}(\hat{\theta}(k)) > \bar{L}(\hat{\theta}(k - 1))$) for $k \leq 10$. We found that ADSA performed better when $\hat{\theta}$ was *not* updated when $L(\cdot)$ decreased. The results reported above reflect this fact (an iteration was still counted when no update occurred). This suggests that in the practical implementation of ADSA it might be valuable to evaluate $L(\cdot)$ at values of $\hat{\theta}$ for at least the first few iterations to ensure that the algorithm is proceeding in a "good" direction. This will increase the computational burden (albeit by a small amount) since ADSA will require three (instead of two) evaluations of $L(\cdot)$ during the first few iterations. Interestingly, in the $p = 15$ case, it never happened that $L(\cdot)$ decreased as $\hat{\theta}$ was updated, even for small k (the author believes that this is due to the fact that it is less likely that a specified large proportion of $\hat{s}_i(\cdot)$'s within $\bar{s}(\cdot)$ will be "bad" when p is large than when p is small.

5. CONCLUDING REMARKS

A procedure has been presented for using SA to find the root of the score equation that arises in maximum likelihood estimation. (The technique would also apply, of course, in finding zeros of gradients in other [non-MLE] settings.) The SA procedure here differs from the well-known Kiefer-Wolfowitz procedure in that a gradient approximation other than the usual finite difference approximation is used. This alternative derivative approximation requires fewer, by a factor equal to the dimension of the parameter vector being estimated, likelihood function evaluations. The technique was illustrated in a signal-plus-noise estimation problem and performed significantly better than KWSA in all cases considered.

The implementation discussed in this paper was restricted to a standard (i.e, "first-order") form for the SA algorithm. This is analogous to the steepest descent method. For the SA procedure to be a viable competitor to Newton-Raphson or scoring (which are generally faster than steepest descent), it would be required that an accelerated (i.e., "second-order") SA algorithm be used. Efforts will be taken in this direction by the author. A promising application for the present first-order form is to use it to bring the estimate

to within a neighborhood of the root and then use a scoring or Newton-Raphson procedure to complete the convergence to the root.

REFERENCES

Blum, J. R. [1954], "Multidimensional Stochastic Approximation Methods," *Ann. Math. Stat.*, 25, 737-744.

El-Sherief, H., and N. K. Sinha [1977], "A Nonparametric Stochastic Approximation Algorithm for On-Line Identification of Multivariate Systems," *Proc. Allerton Conf. on Communications, Control, and Computers,* 231-236.

Goodrich, R. L., and P. E. Caines [1979], "Linear System Identification from Nonstationary Cross-Sectional Data," *IEEE Trans. Auto. Control,* AC-24, 403-410.

Haley, D. R., J. P. Garner, and W. S. Levine [1984], "Efficient Maximum Likelihood Identification of a Positive Semi-Definite Covariance of Initial Population Statistics," *Proc. Am. Control Conf.*, 1085-1089.

Kesten, M. [1958], "Accelerated Stochastic Approximation," *Ann. Math. Stat.*, 29, 41-59.

Kiefer, J., and J. Wolfowitz [1952], "Stochastic Estimation of a Regression Function," *Ann. Math. Stat.* 23, 462-466.

Koch, M. I., and J. C. Spall [1986], "An Efficient Multistep Stochastic Approximation Algorithm," *Proc. Am. Control Conf.*, 1629-1632.

Kushner, H. J., and D. S. Clark [1978], *Stochastic Approximation Methods for Constrained and Unconstrained Systems,* Springer-Verlag, New York.

Ljung, L., and T. Soderstrom [1983], *Theory and Practice of Recursive Identification,* MIT Press, Cambridge, Mass.

Porter, D. W., M. D. Shuster, B. D. Gibbs, and W. S. Levine [1983], "A Partitioned Recursive Algorithm for the Estimation of Dynamical and Initial Condition Parameters from Cross-Sectional Data," *Proc. IEEE Conf. Decision Control,* 596-603.

Ruppert, D. (1985], "A Newton-Raphson Version of the Multivariate Robbins-Monro Procedure," *Ann. Stat.*, 13, 236-245.

Saridis, G. N. [1974], "Stochastic Approximation Methods for Identification and Control—A Survey," *IEEE Trans. Auto. Control,* AC-19, 798-809.

Shumway, R. H., D. E. Olsen, and L. J. Levy [1981], "Estimation and Tests of Hypotheses for the Initial Mean and Covariance in the Kalman Filter Model," *Commun. Stat.—Theor. Meth.*, A10, 1624-1641.

Solo, V. [1982], "Stochastic Approximation and the Final Value Theorem," *Stoch. Proc. Appl.*, 13, 139-156.

Smith, R. H. [1985], "Maximum Likelihood Mean and Covariance Matrix Estimation Constrained to General Positive Semi-Definiteness," *Commun. Stat.—Theor. Meth.*, 14, 2163-2180.

Spall, J. C. [1985], "A Closed Form Approximation to Implicity Defined Maximum Likelihood Estimates," *Proc. American Statistical Assoc., Business and Economics Statistics Section,* 195-200.

Spall, J. C. [1986], "An Approximation for Analyzing a Broad Class of Implicitly and Explicitly Defined Estimators," *Commun. Stat.—Theor. Meth.*, 15, 3747-3762.