Optimal Random Perturbations for Stochastic Approximation using a Simultaneous Perturbation Gradient Approximation[1]

PAYMAN SADEGH*, and JAMES C. SPALL†

* Dept. of Mathematical Modeling, Technical University of Denmark, DK-2800 Lyngby, Denmark. e-mail: ps@imm.dtu.dk

† The Johns Hopkins University, Applied Physics Laboratory, Laurel, MD 20723-6099, USA. e-mail: james.spall@jhuapl.edu.

Abstract

The simultaneous perturbation stochastic approximation (SPSA) algorithm has recently attracted considerable attention for optimization problems where it is difficult or impossible to obtain a direct gradient of the objective (say, loss) function. The approach is based on a highly efficient simultaneous perturbation approximation to the gradient based on loss function measurements. SPSA is based on picking a simultaneous perturbation (random) vector in a Monte Carlo fashion as part of generating the approximation to the gradient. This paper derives the optimal distribution for the Monte Carlo process. The objective is to minimize the mean square error of the estimate. We also consider maximization of the likelihood that the estimate be confined within a bounded symmetric region of the true parameter. The optimal distribution for the components of the simultaneous perturbation vector is found to be a symmetric Bernoulli in both cases. We end the paper with a numerical study related to the area of experiment design.

1. Introduction

Consider the problem of determining the value of a p-dimensional parameter vector to minimize a loss function $L(\theta)$, where only measurements of the loss function are available (i.e., no gradient information is directly available). The simultaneous perturbation stochastic approximation (SPSA) algorithm has recently attracted considerable attention for challenging optimization problems of this type in application areas such as adaptive control, pattern recognition, discrete event systems, neural network training, and model parameter estimation, see, e.g., [1], [2], [3], [4], [5], and [6].

SPSA was introduced in [7] and more thoroughly analyzed in [8]. The essential feature of SPSA is its underlying gradient approximation that requires only two loss function measurements regardless of the number of parameters being optimized. Note the contrast of two function measurements with the $2p$ measurements required in classical finite difference based approaches (i.e., the Kiefer-Wolfowitz SA algorithm). Under reasonably general conditions, it was shown in [8] that the p-fold savings in function measurements per gradient approximation can translate directly into a p-fold savings in total number of measurements needed to achieve a given level of accuracy in the optimization process.

An essential part of the gradient approximation is a simultaneous (random) perturbation relative to the current estimate of θ. This perturbation is generated in a Monte Carlo fashion as part of the optimization process. Since the user has complete control over the perturbation distribution, there is strong reason to choose a distribution as a means of minimizing the number of (potentially costly) function measurements needed in the optimization process. These function measurements may involve physical experiments involving labor or material costs as well as computer related costs associated with simulations or data processing.

The aim of this paper is to determine the form of the optimal distribution for the simultaneous perturbations. This will involve both analytical analysis based on the asymptotic properties of the parameter iterate and numerical finite sample experimentation. The related objectives considered here are to minimize the mean square error of the estimate and to maximize the likelihood that the parameter iterate is restricted to a symmetric bounded region around the true parameter.

The rest of the paper is organized as follows. In Section 2, we briefly review the SPSA algorithm. Section 3 considers the choice of random perturbations. Section 4 presents a numerical example from the area of statistical experiment design. Section 5 offers concluding remarks.

2. Problem Formulation

Consider the problem of finding a root θ^* of $g(\theta) \equiv \partial L(\theta)/\partial\theta = 0$ for some differentiable loss function $L : \mathbb{R}^p \to \mathbb{R}$. In the case where the dependence of the loss function upon θ is unknown, but the loss function is observed in the presence of noise, an stochastic approximation (SA) algorithm of the generic Kiefer-Wolfowitz type

[1]The first author's work was partly supported by the Danish Research Academy, grant S950029, during his stay at JHU/APL. James Spall's work is supported by U.S. Navy contract N00039-95-C-0002 and the JHU/APL IRAD program.

0-7803-3835-9/97/$10.00 (c) 1997 AACC

(see [9]) is appropriate.

Let us now briefly review the SPSA algorithm (see [8]) for the problem posed above. Let $\hat{\theta}_k$ denote the estimate for θ at the kth iteration. The SPSA algorithm has the form

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{g}_k(\hat{\theta}_k)$$

where $\{a_k\}$ is a gain sequence and $\hat{g}_k(\hat{\theta}_k)$ is a simultaneous perturbation approximation to $g(\hat{\theta}_k)$ at iteration k. The simultaneous perturbation approximation is defined as follows. Let $\Delta_k \in \mathbb{R}^p$ be a vector of p mutually independent mean zero random variables $\{\Delta_{k1}, \Delta_{k2}, ..., \Delta_{kp}\}$. Consistent with the usual framework of stochastic approximations, we have noisy measurements of the loss function at specified "design levels". In particular, at the kth iteration

$$y_k^{(+)} = L(\hat{\theta}_k + c_k\Delta_k) + \epsilon_k^{(+)}$$
$$y_k^{(-)} = L(\hat{\theta}_k - c_k\Delta_k) + \epsilon_k^{(-)}$$

where $\{c_k\}$ is a gain sequence and $\epsilon_k^{(+)}$ and $\epsilon_k^{(-)}$ represent measurement noise terms. The basic simultaneous perturbation form for the estimate of $g(\cdot)$ at the kth iteration is then

$$\hat{g}_k(\hat{\theta}_k) = \begin{bmatrix} \frac{y_k^{(+)} - y_k^{(-)}}{2c_k\Delta_{k1}} \\ \vdots \\ \frac{y_k^{(+)} - y_k^{(-)}}{2c_k\Delta_{kp}} \end{bmatrix}. \tag{2.1}$$

Note that at each iteration, only *two* measurements are needed to form the estimate. To help mitigate noise effects in high noise environments, it is sometimes useful to consider averaging among gradient approximations, each generated as in Eq(2.1) based on a new pair of measurements that are conditionally (on $\hat{\theta}_k$) independent of the other measurement pairs; this is examined in [8] but will not be examined further here. Throughout the paper, we assume that:

A_1: $a_k = a/k^\alpha$, and $c_k = 1/k^\gamma$ where $a > 0$, $0 < \alpha \leq 1$, $\gamma > 0$, $\alpha - \gamma > 0.5$, $\alpha - 2\gamma > 0$, and $3\gamma - \alpha/2 \geq 0$ (since c_k and Δ_k always appear together as $c_k\Delta_k$, we fix the numerator in c_k to unity and let Δ_k vary freely).

A_2: $E\{\epsilon_k^{(+)} - \epsilon_k^{(-)} | \hat{\theta}_k, \Delta_k\} = 0$, and for some $\alpha_0, \delta > 0$ and $\forall k$, $E\{\epsilon_k^{\pm(2+\delta)}\} < \alpha_0$. Moreover, there is a σ^2 such that $E\{(\epsilon_k^{(+)} - \epsilon_k^{(-)})^2 | \hat{\theta}_k, \Delta_k\} \to \sigma^2$ as $k \to \infty$.

A_3: For all $k < \infty$, $\{\Delta_{ki}\}$ $(i = 1, ..., p)$ are i.i.d. and symmetrically distributed about 0 with $|\Delta_{ki}| \leq \alpha_0$ a.s. and $E|\Delta_{ki}^{-1}| \leq \alpha_1$ a.s. for some $\alpha_0, \alpha_1 > 0$. For some $\alpha_2, \alpha_3, \delta > 0$, it holds that $E\{|L(\hat{\theta}_k \pm c_k\Delta_k)|^{2+\delta}\} \leq \alpha_2$ and $E(\Delta_{ki}^{-2-\delta}) \leq \alpha_3$, $i = 1, ..., p$. Moreover, there are ρ^2, ξ^2 such that as $k \to \infty$, $E(\Delta_{ki}^2) \to \rho^2$ and $E(\Delta_{ki}^{-2}) \to \xi^2$ for all $i = 1, ..., p$.

A_4: $\sup_k ||\hat{\theta}_k|| < \infty$ a.s. where $||\cdot||$ denotes usual Euclidean norm.

A_5: θ^* is an asymptotically stable solution of the differential equation $dx/dt = -g(x)$.

A_6: Let $D(\theta^*) = \{x_0 : \lim_{t\to\infty} x(t|x_0) = \theta^*\}$ where $x(t|t_0)$ denotes solution to the differential equation of A_5 based on initial condition x_0. There exists a compact set $S \subset D(\theta^*)$ such that $\hat{\theta}_k \in S$ infinitely often for almost all $\hat{\theta}_k$.

A_7: For almost all $\hat{\theta}_k$, there is an open ball about $\hat{\theta}_k$ whose radius is independent of k or $\hat{\theta}_k$, where the third derivative of the loss function exists continuously and is uniformly bounded.

The reader is referred to [8] for remarks on the assumptions.

The problem of selecting random perturbations is formulated as selecting a sequence of probability distributions for Δ_{ki}, $k = 1, 2, ...$, each from the set of allowable probability distributions for the random perturbations (see A_3). The objective is to optimize a suitable criterion related to the parameter estimate.

For small k, the exact distribution of $\hat{\theta}_k$ is dependent upon the (unknown) joint probability distribution of the noise sequence. Therefore, we solve the optimal random perturbation problem using the asymptotic distribution of the estimate. It follows from Proposition 2 of [8] that as $k \to \infty$:

$$k^{\frac{\beta}{2}}(\hat{\theta}_k - \theta^*) \xrightarrow{\text{dist}} Z \sim N(\xi^2 d, \rho^2 D) \tag{2.2}$$

where β is a positive constant, and d and D are quantities *not* dependent upon the random perturbations. The matrix D depends on the Hessian of $L(\theta)$ at θ^* and σ^2, and d depends on the third order derivative of $L(\theta)$ at θ^*. Both d and D are dependent upon a, α, and γ. The reader is referred to [8] for the detailed forms of d and D.

From Eq(2.2), it is evident that the distribution of Z is affected by the random perturbations only through ρ^2 and ξ^2 (see A_3). Hence, using the asymptotic result for sufficiently large number of iterations, the problem simplifies to selection of a *single* probability distribution for Δ_{ki}, for *all* $k = 1, 2, ...$, optimizing some criterion related to Z.

3. Optimal Choice of Random Perturbations

As mentioned in the previous section, the analysis here is based on the asymptotic distribution of the parameter iterate; the authors are unaware of any corresponding finite sample result that would be useful in such calculations.

We consider the design of optimal perturbation distribution with the goal of minimizing the trace of mean square error of the estimate, and maximizing the probability of restricting the estimation error within some bounded symmetric about zero region, respectively.

First suppose that we seek a probability distribution that minimizes the expression $MSE \triangleq E\{\text{trace}[ZZ^\top]\}$. We refer to this criterion as the mean square error crite-

rion. Now, using Eq(2.2)

$$MSE = \rho^2 \text{trace}\{D\} + \xi^4 d^\top d. \tag{3.1}$$

Denote $K_1 = \text{trace}\{D\}$ and $K_2 = d^\top d$ (the numbers K_1 and K_2 do not depend upon the random perturbations). In the following, we let $Pr(.)$ denote probability.

Proposition 1 For all $k = 1, 2, ...,$ and $i = 1, ..., p$, the symmetric Bernoulli distribution

$$Pr(\Delta_{ki} = \pm(\frac{K_1}{2K_2})^{\frac{1}{6}}) = \frac{1}{2} \tag{3.2}$$

is the unique single allowable distribution for Δ_{ki}, minimizing the mean square error criterion.

PROOF: See [10].

From a practical point of view, Corollary 1 below is important in showing that a Bernoulli distribution with given ρ^2 and ξ^2 will always improve upon any other distribution with the the same ρ^2 and ξ^2. This result requires *no knowledge* of K_1 and K_2.

Corollary 1 For a given ρ^2 (or ξ^2), the Bernoulli distribution $\Delta_{ki} = \pm\rho$ (or $\Delta_{ki} = \pm\xi^{-1}$) provides a lower value of Eq(3.1) than any other distribution with the same ρ^2 (or ξ^2).

PROOF: Follows immediately from the necessity part of Proposition 1, see [10].

REMARK 1 To invoke the full optimality of the result in Proposition 1, we require knowledge of K_1 and K_2. This is analogous to the calculations for the optimal gain sequences of stochastic algorithms, see e.g. [11] and [12]. The result in Corollary 1 partially mitigates this situation in that it implies that no matter how a given perturbation distribution is determined, there is a Bernoulli distributions that yields a lower MSE, for any ρ (or ξ) of the given distribution. Another frequently encountered situation is the case where an implicit *a priori* model for $L(\theta)$ is given (i.e., it is only possible to compute $L(\theta)$ for each θ). In such cases, it is often difficult to accurately evaluate the second and third order derivatives or the noise variance σ^2 to determine K_1 and K_2. The following procedure may be useful in such situations. By applying SPSA to the available model using very large number of iterations K, we obtain the estimate $\hat{\theta}_K$ which we use as the true optimum in our calculations. We then obtain (rough) estimates of K_1 and K_2 using the given model and $\hat{\theta}_K$, and use Eq(3.2) to find an approximation to the optimal perturbation magnitude which will be used as an initial guess for a numerical search. Corollary 1 implies that the optimal perturbation distribution should be sought among symmetric Bernoulli distributions. We sample the Δ_{ki} from Bernoulli distributions with varying magnitudes around the initial guess. For each magnitude, we apply SPSA a number of times (cross sections), obtain $\hat{\theta}_k$ for each cross section to find $||\hat{\theta}_k - \hat{\theta}_K||^2$ where $k << K$ is some large iteration number of interest, and average over the computed values of $||\hat{\theta}_k - \hat{\theta}_K||^2$ to numerically evaluate the mean square error for each one of the Bernoulli distributions respectively. The numerical study of the paper illustrates such a procedure.

Now consider maximization of the likelihood of restricting the error Z within some bounded symmetric (about zero) region V_θ. A similar approach is pursued in [13] to determine the constants of a Robbins-Monro stochastic approximation algorithm. The optimality criterion is written as

$$J = Pr\{Z \in V_\theta\}. \tag{3.3}$$

An important special case is where V_θ is the closed unit ball. Then the criterion is $Pr\{||Z|| \leq A\}$, where as usual, $||\cdot||$ denotes Euclidean norm and A is a positive number chosen by the user. It reflects the user's tolerable amount of error.

For the probability criterion J, a result identical to Corollary 1 holds (see [10]). Numerical procedures for optimizing J, given an implicit *a priori* model for the loss function, are similar to the procedure described in Remark 1; they involve application of Bernoulli distributed perturbation sequences and numerical assessment of $Pr\{Z \in V_\theta\}$.

REMARK 2 Consider the degenerate case $d = 0$. This for example occurs when the third order derivatives of the loss function at θ^* are zero (see [8]). Then, clearly the optimal solution according to both the mean square error and probability criteria will be a distribution with $\rho \to 0$, forcing the covariance $\rho^2 D$ to zero. This implies that $\Delta_{ki} \to \pm\infty$ is the optimal choice for random perturbations. However, $\lim_{k\to\infty} c_k = 0$, meaning that it is not possible to draw any definitive conclusion about the optimal size of $c_k\Delta_k$ based on the asymptotic properties. In finite sample cases, c_k does not get infinitesimally small, and it is obviously not allowed to let $|\Delta_{ki}| \to \infty$, either. However, a practical guideline in $d = 0$ situations is to select the magnitude of Δ_{ki} as large as the algorithm does not go unstable. This example shows that the results based on the asymptotic distribution must be interpreted and used with some care in finite sample cases.

4. Numerical Study

In this section, we apply SPSA to a statistical experiment design problem for parameter estimation in a dynamic model, see e.g. [14]. Consider the following autoregressive model with exogenous inputs (ARX(2,1)):

$$y_t = h_1 y_{t-1} + h_2 y_{t-2} + u_t + e_t \tag{4.1}$$

where $\{u_t\}$ and $\{y_t\}$ are input and output sequences and $\{e_t\}$ is a sequence of mean zero i.i.d. Gaussian random variables. We assume that the input sequence is generated by a *finite* register with length 10, meaning that the input repeats periodically with cycle 10. We wish to compute

0-7803-3835-9/97/$10.00 (c) 1997 AACC

the input sequence parameter $(u_1, ..., u_{10})^\top$ which starting from zero initial condition minimizes

$$J_u = -E\{\log\det M_F\} + 0.5\sum u_t^2 \tag{4.2}$$

where

$$M_F = \begin{bmatrix} \sum_{i=n_1}^{n_2} y_{t-1}^2 & \sum_{i=n_1}^{n_2} y_{t-1}y_{t-2} \\ \sum_{i=n_1}^{n_2} y_{t-1}y_{t-2} & \sum_{i=n_1}^{n_2} y_{t-1}^2 \end{bmatrix}.$$

Notice that such a problem formulation implies that we deal with a static optimization problem and not a dynamic one since we consider the whole sequence of data $\{y_t\}$ in batch mode within the loss function and a fixed number of parameters independent of the size of the data set. We explain Eq(4.2) as follows. Assuming that we are interested in estimating $\Lambda = (h_1, h_2)^\top$, the basic least squares estimate is given by (see, e.g. [15])

$$\hat{\Lambda} = M_F^{-1} \begin{bmatrix} \sum_{i=n_1}^{n_2} (y_t - u_t)y_{t-1} \\ \sum_{i=n_1}^{n_2} (y_t - u_t)y_{t-2} \end{bmatrix}.$$

Hence, by selecting the input sequence to maximize the expected value of the (logarithm) of the determinant of M_F, we wish to avoid the problem of the singularity of M_F. Indeed, for large values of sample size, the matrix M_F is (approximately) proportional to Fisher's information matrix for the model given by Eq(4.1) (see [14], Chapter 6). Since the positive semi-definite matrix M_F is an increasing function of the input power $\sum u_t^2$, the second term of the criterion penalizes signals with large power. For a detailed treatment of the problem of input design for dynamic system identification, see [14], Chapter 6. In a large part of the literature on experiment design, the solution is obtained by assuming a model for the data and calculation of the information matrix as a function of input. Such models are often obtained through performing preliminary identification experiments. Here, we directly estimate the optimal inputs without requiring a preliminary identification stage.

Let us assume that the model parameters are given by $h_1 = 1.45$, $h_2 = -0.475$ (which correspond to poles 0.5 and 0.95), the standard deviation of e_t is 0.05, and the system is initially at rest. Note that these values are used for data generation purpose, and to (approximately) determine the optimal distribution of the random perturbations. The SPSA algorithm requires no knowledge of these values and the optimization may be carried out by real experimentations that involve exciting the system at initial rest by different inputs and output measurements to compute J_u. In the following, we select $n_1 = 9$, $n_2 = 64$ (see the definition of M_F below Eq(4.2)), $a_k = 0.1/k^{0.9}$, and $c_k = 1/k^{0.17}$.

We first apply SPSA with 50000 iterations and $\Delta_{ki} = \pm 0.1$ (Bernoulli distributed) in order to obtain an estimate of the (uncomputable) optimal sequence $\{u_t^*\}$ for later reference. This value will be used as the true optimum for the rest of the paper since the number of iterations for all later estimation is $1200 << 50000$. Then, we assess the second and third order derivatives of the loss function at the optimum, $\{u_t^*\}$, by numerical finite difference method for the noise free case. Also, we approximate σ^2 by simulation of 1000 realizations of $[\log\det(M_F)]$ at $\{u_t^*\}$. Inserting these estimates in Eq(3.2) yields the distribution $Pr(\Delta_{ki} = \pm 0.19) = \frac{1}{2}$. This distribution shall only be used as an initial guess for a numerical search to find the optimizer for the mean square error and probability criteria since only rough estimates of K_1 and K_2 (see Eq(3.2)) are available.

We apply Bernoulli distributions with magnitude of the outcome around 0.19, estimate the optimal input sequence 100 times, and assess the values of the mean square error and probability criteria numerically. The optimal distribution, according to both the mean square error and probability criteria, is found to be a ± 0.25 Bernoulli distribution. We use the same procedure as above to compare the optimal distribution against other choices of distribution. In Table 1, all the distributions correspond to Bernoulli distributed variables. The top row of the table provides the relevant Bernoulli distributions. For the probability criterion, we have chosen the special case below Eq(3.3) with $A = 4 \times 10^{-3}$. The results indicate that

	$\pm$ 0.15	$\pm$ 0.25	$\pm$ 0.4	$\pm$ 1
MSE	0.0063	0.0052	0.0073	0.1061
J	0.36	0.51	0.35	0.0

Table 1: Performance of SPSA under varying Bernoulli distributions

an inappropriate choice of random perturbations (e.g. ± 1 in this numerical study) would lead to very poor estimation properties.

We also apply a random variable uniformly distributed over $[-0.3, -0.2] \cup [0.2, 0.3]$. This choice is interesting since the distribution is continuous and its support includes the support points of the optimal Bernoulli (± 0.25). The numerical evaluations of MSE and J yield 0.0062 and 0.39, respectively, which are noticeably worse than the results for the optimal Bernoulli distribution.

Finally, notice that in Table 1, the number of iterations have been chosen relatively large (1200) in order to let the iterates reach the asymptotic condition. Therefore, we expect that the optimum should be sought among symmetric Bernoulli distributions (see Section 3). In order to investigate the performance of the asymptotic solution for small sample cases and large initial deviations from the true optimum, consider a case of 10 iterations with a 17.5% initial deviation for all components of $\{u_t\}$. We are particularly interested in numerically evaluating the performance of the (asymptotically) optimal Bernoulli distribution against other (symmetric) distributions that

0-7803-3835-9/97/$10.00 (c) 1997 AACC

contain more than two support points. Therefore, we use the MSE criterion to test the Bernoulli (± 0.25) distribution against two bimodal distributions. One is chosen to be a random variable uniformly distributed over $[-0.3, -0.2] \cup [0.2, 0.3]$. The other corresponds to a random variable triangular distributed over both $[0.2, 0.3]$ and $[-0.3, -0.2]$. The corresponding MSE values are 0.0756, 0.0789, 0.0764, respectively. This comparison indicates that the asymptotic solution may perform reasonably well even for very small sample sizes. Notice however that the solution to the random perturbation problem in small sample cases is an open question.

5. Concluding Remarks

The paper deals with the optimal choice of random perturbations for the SPSA algorithm. Since the user has full control over this choice, there is strong reason to pick this distribution wisely in order to reduce the overall costs of optimization. We have shown that for the mean square error and probability criteria, the optimal random perturbations should be sampled from a symmetric Bernoulli distribution. The choice of the optimal Bernoulli distribution (i.e. the magnitude of its outcome) is dependent upon the prior information about the loss function. However, in the usual case where such information is unavailable, this paper shows that the Bernoulli *distribution form* is the (asymptotically) optimal form *regardless* of the value of the variance of the perturbation distribution. This has significant practical implication as the perturbation distribution is typically determined based on small scale experimentation and/or limited prior knowledge about the form of the loss function. All the results are based on the asymptotic theory. Investigating the choice of random perturbations for finite sample cases is of significant theoretical and practical interest and represents a possible topic for future research on the subject.

References

[1] F. Rezayat. On the use of an SPSA-based model free controller in quality improvement. *Automatica*, 31:913–915, 1995.

[2] Y. Maeda, H. Hirano, and Y. Kanata. A learning rule of neural networks via simultaneous perturbation and its hardware implementation. *Neural Nets*, 8:251–259, 1995.

[3] S. D. Hill and M. C. Fu. Transfer optimization via simulation perturbation stochastic approximation. In *Proc. Winter Simulation Conference*, pages 242–249, 1995.

[4] G. Cauwenberghs. *Analog VLSI Autonomous Systems for Learning and Optimization*. PhD thesis, Dept of Electrical Engineering, California Institute of Technology, 1994.

[5] D. C. Chin. A more efficient global optimization based on Styblinski and Tang. *Neural Nets.*, 7:573–574, 1994.

[6] T. Parisini and A. Alessandri. Non-linear modeling and state estimation in a real power plant using neural networks and stochastic approximation. In *Proc. American Control Conference*, pages 1561–1567, 1995.

[7] J. C. Spall. A stochastic approximation technique for generating maximum likelihood parameter estimates. In *Proc. American Control Conference*, pages 1161–1167, 1987.

[8] J. C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.

[9] D. Ruppert. Kiefer-wolfowitz procedure. In S. Kotz and N. L. Johnson, editors, *Encyclopedia of Statistical Sciences*, pages 379–381. Wiley, 1983.

[10] P. Sadegh and J. C. Spall. Optimal random perturbations for stochastic approximation using a simultaneous perturbation gradient approximation. To appear in *IEEE Transactions on Automatic Control*. Manuscript can be provided upon request., 1997.

[11] V. Fabian. Stochastic approximation. In J. J. Rustagi, editor, *Optimizing Methods in Statistics*, pages 439–470. Academic, New York, 1971.

[12] D. C. Chin. Comparative study of stochastic algorithms for system optimization based on gradient approximations. *IEEE Transactions on Systems, Man, and Cybernetics*, 27, 1997. In press.

[13] S. V. Gusev and T. P. Krasulina. An algorithm for stochastic approximation with a preassigned probability of not exceeding a required threshold. *Journal of Computer and Systems Sciences International*, 33:39–41, 1995.

[14] G. C. Goodwin and R. L. Payne. *Dynamic System Identification: Experiment Design and Data Analysis*. Academic, New York, 1977.

[15] L. Ljung. *System Identification: Theory for the User*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1987.

0-7803-3835-9/97/$10.00 (c) 1997 AACC