Finally, sketches were made of the characters which were incorrectly identified. A sample of 24 of these, from a total of 56, is shown in Fig. 8. Two 5's contain full stops. Two squashed patterns, a 6 and a 9, were thought to be zeros. An extension of the preprocessor to enable angularities to be described may be the best way to deal with such mistakes. Many of the patterns are badly distorted and are usually easy to reject. This recreation method has been found to be a useful tool to identify the deficiencies of the system and to plan its improvements.

VIII. Conclusions

A complete system for handwritten numeral recognition has been evaluated. The preprocessing part of the system has been designed as a hardware device. The classifying system employs a simple statistical decision. The form of this is such that it might easily be turned into a set of linear discriminant functions and realized as a parallel network, giving a high speed for the total system. Methods used in the study can be interpreted visually, and this facilitates rapid improvements to the system.

The scheme shows promise for extension to more difficult problems in character recognition. A useful step will be to associate a set of edge segments which come from the same region of constant density. In this way several characters within the same picture field would be identifiable. Preprocessor extensions to the descriptions of angularities would be useful in that many nonnumeric pattern analysis schemes (e.g., [10]) use angularities as their starting points. This would open up a new type of decision logic to complement the existing statistical scheme.

Acknowledgment

The authors would like to thank the Post Office of Great Britain for providing the raw data used in the experiment.

References

[1] W. H. Highleyman, "An analog method for character recognition," *IRE Trans. Electron. Comput.*, vol. EC-10, Sept. 1961, pp. 502–512.
[2] R. Bakis, N. Herbst, and G. Nagy, "An experimental study of machine recognition of hand-printed numerals," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-4, July 1968, pp. 119–132.
[3] H. Genchi, K. Mori, S. Watanabe, and S. Katsuragai, "Recognition of handwritten numerical characters for automatic letter sorting," *Proc. IEEE*, vol. 56, Aug. 1968, pp. 1292–1301.
[4] E. C. Greanias, P. F. Meagher, R. J. Norman, and P. Essinger, "The recognition of handwritten numerals by contour analysis," *IBM J. Res. Develop.*, vol. 7, Jan. 1963, pp. 14–21.
[5] T. Marill and D. M. Green, "Statistical recognition functions and the design of pattern recognizers," *IRE Trans. Electron. Comput.*, vol. EC-9, Dec. 1960, pp. 472–477.
[6] R. G. Casey, "Moment normalisation of handprinted characters," *IBM J. Res. Develop.*, vol. 14, Sept. 1970, pp. 548–557.
[7] C. K. Chow, "An optimum character recognition system using decision functions," *IRE Trans. Electron. Comput.*, vol. EC-6, Dec. 1957, pp. 247–254.
[8] B. J. Vieri and M. J. Gollifer, "An on-line television camera," in *Industrial Measurement Techniques for On-line Computers*, Inst. Elec. Eng., Conf. Publ. 43, June 1968.
[9] G. Sebestyen and J. Edie, "An algorithm for non-parametric pattern recognition," *IEEE Trans. Electron. Comput.*, vol. EC-15, Dec. 1966, pp. 908–915.
[10] W. F. Miller and A. C. Shaw, "Linguistic methods in picture processing—a survey," in *1968 Fall Joint Computer Conf.*, pp. 279–290.

A Stochastic Approximation Method

NARESH K. SINHA, SENIOR MEMBER, IEEE, AND MICHAEL P. GRISCIK, MEMBER, IEEE

***Abstract*—A new algorithm for stochastic approximation has been proposed, along with the assumptions and conditions necessary for convergence. It has been proved by two different methods that the algorithm converges to the sought value in the mean-square sense and with probability one. The rate of convergence of the new algorithm is shown to be better than two existing algorithms under certain conditions. Results of simulation have been given, making a realistic comparison between the three algorithms.**

I. Introduction

STOCHASTIC approximation may be defined as a scheme for successive approximation of a sought quantity when the observations involve random errors due to the stochastic nature of the problem. It can be applied to any problem that can be formulated as some form of regression in which repeated observations are made. Compared to conventional methods such as maximum likelihood estimation it has the following advantages.

1) Only a small interval of data needs processing.

2) Only simple computations are required, even when the actual functional dependence of the regression function on the parameters of interest is nonlinear.

3) *A priori* knowledge of the process statistics is not necessary, nor is the detailed knowledge of the functional relationship between the desired parameters and the observed data. The only requirements are that the regression function satisfy certain regularity conditions and that the regression problem have a unique solution. The method can be made asymptotically efficient when the *a priori* knowledge concerning the statistics and the functional relationship is available.

Applications of stochastic approximation algorithms have been proposed in adaptive and learning control [1], system

Manuscript received December 21, 1971; revised May 9, 1971. This work was supported by the National Research Council of Canada.

N. K. Sinha is with the Department of Electrical Engineering, McMaster University, Hamilton, Ont., Canada.
M. P. Griscik is with Procter and Gamble of Canada, Ltd., Hamilton, Ont., Canada.

identification [2], adaptive communication, and pattern recognition [3].

Major contributions to the area of stochastic approximation have been made by Robbins and Monro [4], Kiefer and Wolfowitz [5], and Dvoretzky [6]. Essentially the Robbins–Monro procedure finds the root of a regression function whose form is unknown but which can be observed and sampled. The Kiefer–Wolfowitz procedure is a method for finding the extremum, maximum or minimum, of a regression function, given only pertinent random observations. Dvoretzky formulated and proved a general theorem to this effect, showing both convergence with probability one and in the mean-square sense for iterative solution algorithms convergent on regression functions obtained from random observations.

Dvoretzky's work represents a major contribution to the mathematical structure of stochastic approximation theory. Practical applications, however, are concerned with the rate of convergence and their dependence on the parameters of the recursive solution algorithm. In the area of control theory two stochastic approximation algorithms have been developed and investigated by Fu *et al.* [7]. Further improvement in the rate of convergence is desirable to reduce the number of iterations and the overall time to get a result with a given level of confidence.

The objective of this paper is to show the development of a new algorithm which has a considerably faster rate of convergence than the two proposed in [7]. It will be shown that the new algorithm converges to the true value regardless of the starting point. The conditions and limitations on the convergence will be made explicit in the proof. Having proven convergence, comparison will be made between the two existing algorithms and the new algorithm.

II. General Stochastic Approximation Procedure

A general algorithm of stochastic approximation of the form

$$\hat{x}_{n+1} = \hat{x}_n + \gamma_{n+1}\{f(\boldsymbol{r}_{n+1}) - \hat{x}_n\}, \qquad n = 1,2,\cdots \tag{1}$$

is said to be of the Dvoretzky type, where

$\hat{x}_n$	nth estimate of x;
s	true value of the parameter being estimated;
r_n	nth observation taken and used to calculate a function;
$\boldsymbol{r}_{n+1}$	superscript T denoting transposition, $= [r_1, r_2\cdots,r_{n+1}]^T$;
$f(\boldsymbol{r}_{n+1})$	scalar functional of $\boldsymbol{r}_{n+1}$; and
γ_n	gain sequence.

It is required that the function $f(r_n)$ be such that its expectation

$$E[f(\boldsymbol{r}_n)] = x. \tag{2}$$

The first algorithm in [7] uses

$$f_1(\boldsymbol{r}_n) = r_n \tag{3}$$

where the observations

$$r_n = x + \xi_n \tag{4}$$

ξ_n being a random component of zero-mean noise, and

$$\gamma_n = 1/(n + \alpha) \tag{5}$$

where α is constant. The selection of the constant α is arbitrary, but if *a priori* statistics are known and the process is known to have a noise component of finite variance, then

$$\alpha = V_0^2/\sigma^2 \tag{6}$$

where σ^2 is the variance of the distribution and V_0^2 is the initial value of the expected mean-square error. Selecting γ_n and α as in (5) and (6), respectively, gives the algorithm (3) the best convergence.

The second algorithm in [7] uses

$$f_2(\boldsymbol{r}_n) = \frac{1}{n}\sum_{i=1}^{n} r_i \tag{7}$$

where the samples r_n are of the same form as (4), but the gain sequence is given by

$$\gamma_n = \frac{n}{n(n+1)/2 + \alpha} \tag{8}$$

for the best convergence, where α is again dependent upon *a priori* statistics and defined by (6).

The convergence of both the algorithms has been proved [7] using Dvoretzky's theorem.

III. The New Algorithm

It is proposed that the function f be chosen as

$$f_3(\boldsymbol{r}_{n+1}) = |\hat{R}_{r_{n+1}}(l)|^{1/2} \tag{9}$$

where $\hat{R}_{r_{n+1}}(l)$ is an estimate of the sample autocorrelation function of the samples $r_1,r_2,\cdots,r_{n+1}$, with $l - 1 \leq n$, and where $|\hat{R}_{r_{n+1}}(l)|$ is the absolute value of $\hat{R}_{r_{n+1}}(l)$.

The convergence of this algorithm will now be proved. An alternative proof based on Dvoretzky's theorem is given in the Appendix. From the definition the estimate of the sample autocorrelation.

$$\hat{R}_{r_n}(l) = E[r_n r_{n+l}] \tag{10}$$

and recalling (4), one may write

$$\begin{aligned}\hat{R}_{r_n}(l) &= E[(x + \xi_n)(x + \xi_{n+l})] \\ &= x^2 + |\hat{R}_{\xi_n}(l)|\end{aligned} \tag{11}$$

since x is a constant and ξ_n and ξ_{n+l} are elements of a zero-mean random process. Hence (1) may be written as

$$\hat{x}_{n+1} = \hat{x}_n + \gamma_{n+1}\{[x^2 + |\hat{R}_{\xi_{n+1}}(l)|]^{1/2} - \hat{x}_n\}. \tag{12}$$

Subtracting x from both sides of (12) and rearranging terms,

$$(\hat{x}_{n+1} - x) = (1 - \gamma_{n+1})(\hat{x}_n - x) + \gamma_{n+1}\psi_{n+1} \tag{13}$$

where

$$\psi_{n+1} = x\left\{\left[1 + \frac{|\hat{R}_{\xi_{n+1}}(l)|}{x^2}\right]^{1/2} - 1\right\}. \tag{14}$$

If (13) is iterated, then an expansion can be developed in terms of the initial mean-square error. After $n + 1$ itera-

tions, the following equation is obtained:

$$(\hat{x}_{n+1} - x) = (1 - \gamma_{n+1})(1 - \gamma_n)\cdots(1 - \gamma_2)(1 - \gamma_1) \cdot (\hat{x}_0 - x) + (1 - \gamma_{n+1})(1 - \gamma_n)\cdots (1 - \gamma_2)\gamma_1\psi_1 + (1 - \gamma_{n+1})(1 - \gamma_n)\cdots (1 - \gamma_3)\gamma_2\psi_2 + \cdots + (1 - \gamma_{n+1})\gamma_n\psi_n + \gamma_{n+1}\psi_{n+1}. \quad (15)$$

Equation (15) may be written into the following closed form:

$$(\hat{x}_{n+1} - x) = \left\{\prod_{i=1}^{n+1}(1 - \gamma_i)\right\}(\hat{x}_0 - x) + \sum_{i=1}^{n+1}\gamma_i\psi_i \cdot \prod_{j=i+1}^{n+1}(1 - \gamma_j) \quad (16)$$

where all void products are taken as unity. Squaring (16) and substituting for ψ_i from (14) gives, after some rearrangement,

$$\begin{aligned}(\hat{x}_{n+1} - x)^2 &= (\hat{x}_0 - x)^2\left\{\prod_{i=1}^{n+1}(1 - \gamma_i)\right\}^2 \\ &+ x^2\left[\sum_{i=1}^{n+1}\gamma_i\prod_{j=i+1}^{n+1}(1 - \gamma_j)\right]^2 \\ &+ \left[\sum_{i=1}^{n+1}\gamma_i\{x^2 + |\hat{R}_{\xi_i}(l)|\}^{1/2}\cdot\prod_{j=i+1}^{n+1}(1 - \gamma_j)\right]^2 \\ &+ 2(\hat{x}_0 - x)\left\{\prod_{i=1}^{n+1}(1 - \gamma_i)\right\} \\ &\cdot\left[\sum_{i=1}^{n+1}\gamma_i\{x^2 + |\hat{R}_{\xi_i}(l)|\}^{1/2}\cdot\prod_{j=i+1}^{n+1}(1 - \gamma_j)\right] \\ &- 2x(\hat{x}_0 - x)\left\{\prod_{i=1}^{n+1}(1 - \gamma_i)\right\}\left[\sum_{i=1}^{n+1}\gamma_i\prod_{j=i+1}^{n+1}(1 - \gamma_j)\right] \\ &- 2x\left\{\sum_{i=1}^{n+1}\gamma_i\prod_{j=i+1}^{n+1}(1 - \gamma_j)\right\} \\ &\cdot\left[\sum_{i=1}^{n+1}\gamma_i\{x^2 + |\hat{R}_{\xi_i}(l)|\}^{1/2}\cdot\prod_{j=i+1}^{n+1}(i - \gamma_j)\right]. \quad (17)\end{aligned}$$

In general the autocorrelation function of the noise may be written as

$$\hat{R}_{\xi_k}(l) = E[\xi_k\xi_{k+l}] = \sum_i \sigma_i^2 e^{-\alpha_i|l|} \quad (18)$$

where $|l|$ represents the magnitude of l. As the spectrum of the noise becomes wider, i.e., as it approaches white noise, the autocorrelation function becomes an impulse,

$$\hat{R}_{\xi_k}(l) = \sigma^2\,\delta(l) \quad (19)$$

and for $l > 1$,

$$\hat{R}_{\xi_k}(l) \to 0. \quad (20)$$

Hence, for $l > 1$, (17) reduces to

$$(\hat{x}_{n+1} - x)^2 = (\hat{x} - x_0)^2 \cdot \left\{\prod_{i=1}^{n+1}(1 - \gamma_i)\right\}^2. \quad (21)$$

Note that (21) is valid not only for white noise, but also for colored noise if l is taken sufficiently large.

Therefore, for the mean-square error to become zero regardless of the starting value and for the algorithm to converge, it is necessary and sufficient that

$$\lim_{n\to\infty}\prod_{i=1}^{n+1}(1 - \gamma_i) \to 0. \quad (22)$$

The selection of a γ-sequence that satisfies (22) is arbitrary within the restrictions of the limit given in the preceding. One may therefore try to select the sequence in such a manner as to maximize the rate of convergence of the algorithm. An attempt to obtain the optimum γ-sequence using the discrete maximum principle did not prove fruitful, however, due mainly to the fact that the resulting equations were quite formidable, and also a knowledge of the initial error as well as some statistical properties of the noise were required.

It was therefore decided arbitrarily to select two γ-sequences which appeared good and satisfied (22). These are

$$\gamma_n = 1/n \quad (23)$$

and

$$\gamma_n = 1 - (1/n). \quad (24)$$

It will now be shown that the new algorithm (9) provides an unbiased estimate. By definition, an estimator $\hat{x}_n$ of x is unbiased if

$$E[\hat{x}_n] = x. \quad (25)$$

Again using the assumption that the noise has a fairly wide spectrum, for large l, one may approximate (11) as

$$\hat{R}_{r_{n+1}}(l) = x^2. \quad (26)$$

Using this approximation in (16) and taking the expectations of both sides, it is seen that

$$E[\hat{x}_{n+1} - x] = 0. \quad (27)$$

Thus it can be said that with a choice of the γ-sequence satisfying (22) and sufficiently large l, the algorithm converges and gives an unbiased estimate of x. In practice ξ_i is often a white-noise sequence, and it may be permissible to make l as small as 2 or 3.

IV. Comparison of the Three Algorithms

The rate of convergence of the various algorithms would depend considerably upon the initial guess, as well as the γ-sequence selected for each case. One way to make a fair comparison would be to plot the normalized expected mean-square error as a function of the number of iterations n. This is particularly convenient as an analytical expression can easily be obtained for this quantity for each of the three algorithms. It has been shown [7] that for the first algorithm of Fu *et al.*, the expected mean-square error after $n + 1$ iterations V_{n+1}, is given by

$$V_{n+1}^2 = (1 - \gamma_{n+1})V_n^2 + \gamma_{n+1}^2\sigma^2 \quad (28)$$

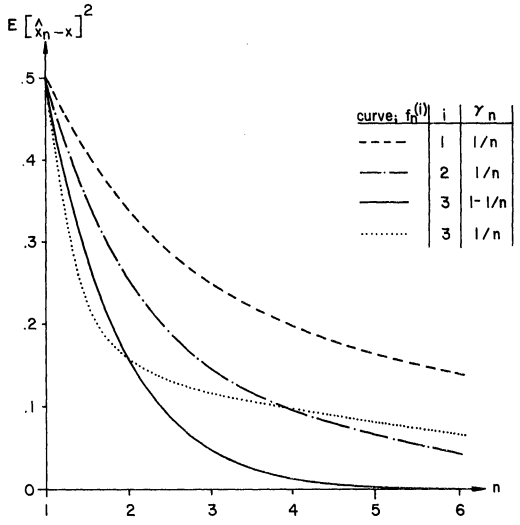

Fig. 1. Normalized expected mean-square error as function of n. ($\sigma^2 = 1$).

where σ^2 is the variance of the noise. Hence a curve of V_n may be plotted against n, assuming that V_0 and σ are equal to one. This would be the plot of the normalized expected mean-square error against the number of iterations.

The expected mean-square error for the second algorithm of [7] is given by

$$V_{n+1}^2 = (1 - \gamma_{n+1})^2 V_n^2 + \gamma_{n+1}^2(\sigma^2/n). \tag{29}$$

The expected mean-square error after $n + 1$ iterations as a function of the initial error for the new algorithm is obtained from (21) and may be written as (subject to the approximation involved):

$$V_{n+1}^2 = V_0^2 \cdot \left\{\prod_{i=1}^{n+1} (1 - \gamma_i)\right\}^2. \tag{30}$$

A plot of these three equations against n for the γ-sequence $\gamma_n = 1/n$ is shown in Fig. 1. In addition a plot for the sequence $\gamma_n = 1 - (1/n)$ is also shown for the new algorithm. It will be seen that the performance of the new algorithm is better than of the two older ones, particularly if the sequence $\gamma_n = 1 - (1/n)$ is used. It may be noted that for the first two algorithms the sequence $\gamma_n = 1/n$ may be regarded as the optimal in the absence of *a priori* statistical information.

V. Simultaneous Results

To test the relative effectiveness of the various algorithms a simple stochastic approximation problem was simulated on the CDC 6400 digital computer at the McMaster University Computation Center. A zero-mean pseudo-random noise sequence ξ_n with a given value σ of the standard deviation was generated and added to a constant x in order to obtain the set of observations r_n as defined in (4). Each of the three algorithms based its sample information on these r_n at every n and calculated an estimate $\hat{x}_n$, the first estimate being taken as the first sample r_1. The calculations were made for different values of the rms

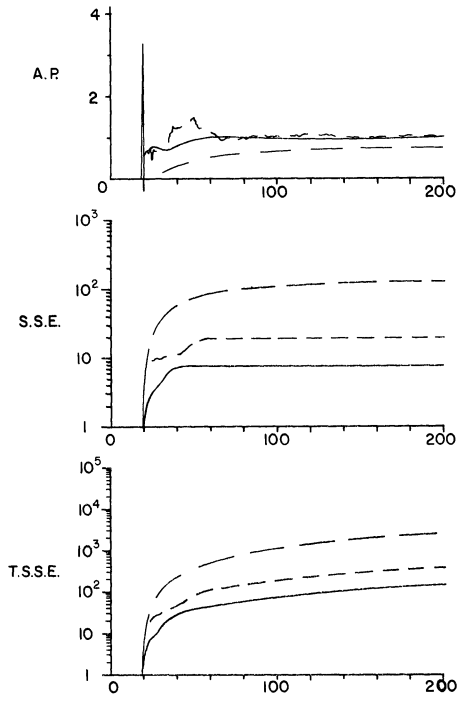

Fig. 2. AP, SSE, and TSSE as function of n for $\Lambda = 1.50$. ----first algorithm of [7]; — —second algorithm of [7]; ——new algorithm.

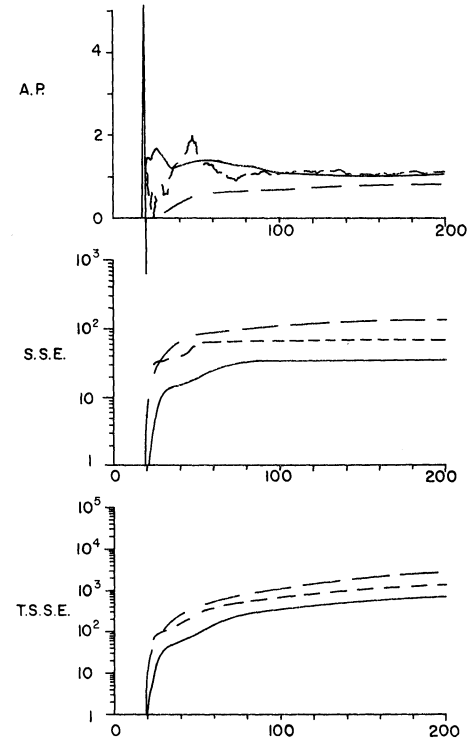

Fig. 3. AP, SSE, and TSSE as function of n for $\Lambda = 3.0$. ----first algorithm of [7]; — —second algorithm of [7]; ——new algorithm.

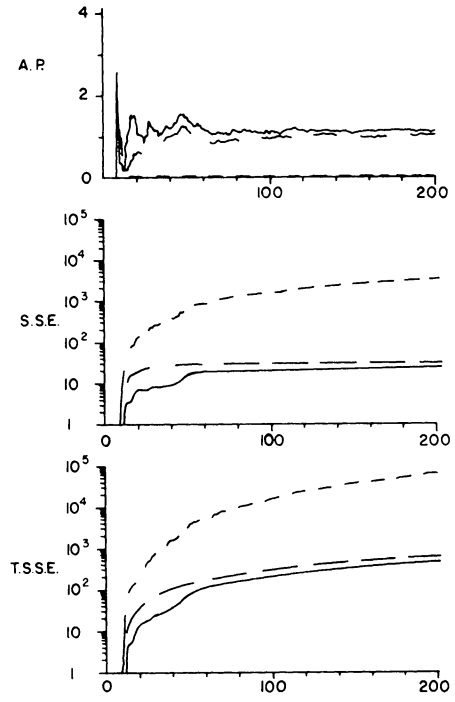

Fig. 4. AP, SSE, and TSSE as function of n for $\Lambda = 2.0$ ----first algorithm of [7]; — —second algorithm of [7]; ——new algorithm.

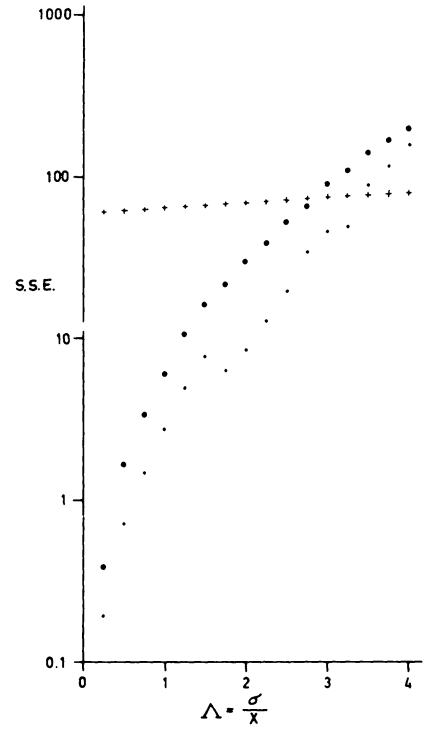

Fig. 5. Sum of square errors (SSE) as function of Λ. + + + + second algorithm of [7]; · · · ·new algorithm with $r_n = 1/n$; ∘∘∘∘ new algorithm with $r_n = 1 - 1/n$.

noise-to-signal ratio Λ, defined as

$$\Lambda = \sigma/x \tag{31}$$

and from these the approximation trajectory (AP), i.e., the path of the successive approximation x_{n_n}, normalized with respect to x, was calculated for each algorithm. To obtain suitable measures for comparison, along with these were computed the sample-square error (SSE) and the time sample-square error (TSSE) defined as

$$\text{SSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x)^2 \tag{32}$$

and

$$\text{TSSE} = \frac{1}{n}\sum_{i=1}^{n} i(\hat{x}_i - x)^2. \tag{33}$$

For $l = 20$, $\gamma_n = 1/n$, and $\Lambda = 1.5$, the AP, SSE, and TSSE for the three algorithms are shown in Fig. 2, and these have been repeated in Fig. 3 with $\Lambda = 3$. Fig. 4 shows a comparison with the γ-sequence $\gamma_n = 1 - (1/n)$ using $\Lambda = 2$, and $l = 10$. It will be seen that the results obtained from the new algorithm are better in each case.

It should be noted, however, that for both γ-sequences, there is a noise level, that is, a value of Λ at which the new algorithm has the same sum of square errors as the second algorithm in [7]. This value of Λ is called the value of equi-utility for the two algorithms being compared. Evidently, this depends on the γ-sequence chosen, as well as the values of l and n. Fig. 5 shows the plot of the SSE against Λ for $l = 10$, $n = 100$, and the two γ-sequences mentioned earlier. It will be seen that the new algorithm is better for values of Λ less than or equal to 3.

VI. Conclusion

A new algorithm has been proposed for stochastic approximation, and its convergence has been established. It has been shown that the normalized mean-square error as a function of the number of iterations decreases at a much faster rate with the new algorithm than with the existing algorithms. Results of simulation indicate that the new algorithm gives a smaller sum of square errors than the second algorithm in [7] as long as the rms noise-to-signal ratio is less than 3.

Appendix
Alternative Proof Based on Dvoretzky's Theorem

Since the new algorithm is of the Dvoretzky type, the property of convergence can be proved using his theorem. First a statement of Dvoretzky's theorem will be given, outlining the conditions and the results. Then it will be shown that the algorithm satisfies all the requirements of the theorem.

Theorem

Let α_n, β_n, and γ_n, $n = 1,2,\cdots$, be nonnegative real numbers satisfying the following conditions:

$$\lim_{n\to\infty} \alpha_n = 0 \tag{34}$$

$$\sum_{n=1}^{\infty} \beta_n < \infty \tag{35}$$

$$\sum_{n=1}^{\infty} \gamma_n = \infty. \tag{36}$$

Let θ be a real number and T_n be measurable transformations satisfying

$$|T_n(\rho_1,\rho_2,\cdots,\rho_n) - \theta| \leq \max [\alpha_n,(1 + \beta_n)|\rho_n - \theta| - \gamma_n] \tag{37}$$

for all real $\rho_1,\rho_2,\cdots,\rho_n$. Let X_1 and Y_n be random variables and define

$$X_{n+1} = T_n[X_1,X_2,\cdots,X_n] + Y_n \tag{38}$$

for $n > 1$, satisfying the conditions

$$E[X_1^2] < \infty \tag{39}$$

$$\sum_{n=1}^{\infty} E[Y_n^2] < \infty \tag{40}$$

$$E[Y_n \mid X_1,X_2,\cdots,X_n] = 0 \tag{41}$$

with probability one. Then as $n \to \infty$, X_n approaches θ with probability one, and also in the mean-square sense, that is,

$$P[\lim_{n\to\infty} X_n = \theta] = 1 \tag{42}$$

$$\lim_{n\to\infty} E[(X_n - \theta)^2] = 0. \tag{43}$$

An extension, six generalizations, and two corollaries were also proved by Dvoretzky, but they will not be needed here and hence have not been stated.

Application to the New Algorithm

From (13) in Section III the new algorithm may be written as

$$(\hat{x}_{n+1} - x) = (1 - \gamma_{n+1})(\hat{x}_n - x) + \gamma_{n+1}\psi_{n+1} \tag{44}$$

where, as in (14),

$$\psi_{n+1} = \{x^2 + |\hat{R}_{\xi_{n+1}}(l)|\}^{1/2} - x. \tag{45}$$

Consider the transformation

$$\hat{x}_n - x = W_n. \tag{46}$$

Applying this transformation to (45) gives

$$W_{n+1} = (1 - \gamma_{n+1})W_n + \gamma_{n+1}\psi_{n+1}. \tag{47}$$

Equation (47) is a zero-seeking algorithm and is analogous to the case when $\theta = 0$ in (42) and (43). Thus, if the conditions of the theorem are satisfied by (47), the convergence of W_{n+1} to zero with probability one as $n \to \infty$ implies that $\hat{x}_n$ will converge to x in the same manner as well as in the mean-square sense.

Let

$$\alpha_n = 1/n \tag{48}$$

$$\beta_n = 1/n^2 \tag{49}$$

$$\gamma_n = \frac{1}{n + K} \quad \text{or} \quad 1 - \frac{1}{n + K} \tag{50}$$

where K is a nonnegative real constant. It can easily be shown that these sequences satisfy the conditions (34), (35), and (36), respectively.

Identifying $T_n(\rho_1,\cdots,\rho_n)$ with $(1 - \gamma_{n+1})W_n$ and ρ_n with W_n, condition (37) would be satisfied if

$$(1 - \gamma_{n+1})|W_n| \leq \max [\alpha_n,(1 + \beta_n - \gamma_n)|W_n|]. \tag{51}$$

(Although (51) is slightly different from (37), it has been shown as a stronger condition in one of Dvoretzky's generalizations.)

Since $1 - \gamma_{n+1}$ is positive for all n for both the γ-sequences specified in (50), the inequality (51) is equivalent to

$$(\beta_n + \gamma_{n+1} - \gamma_n)|W_n| \geq 0 \tag{52}$$

and, it is always satisfied by the inequality since

$$\beta_n + \gamma_{n+1} - \gamma_n = \frac{K^2 + K + 2nK + n}{n^2(n + K + 1)(n + K)} > 0 \tag{53}$$

when

$$\gamma_n = \frac{1}{n + K}$$

and

$$\beta_n + \gamma_{n+1} - \gamma_n = \frac{1}{n^2} + \frac{1}{n + K} - \frac{1}{n + K + 1} > 0 \tag{54}$$

when

$$\gamma_n = 1 - \frac{1}{n + K}$$

Now, identifying (38) with (47) yields

$$Y_n = \gamma_{n+1}\psi_{n+1} \tag{55}$$

where ψ_{n+1} was defined in (45). Hence

$$\begin{aligned} Y_n^2 &= \gamma_n^2 \cdot [\{x^2 + |\hat{R}_{\xi_{n+1}}(l)|\}^{1/2} - x]^2 \\ &= \gamma_n^2[2x^2 + |\hat{R}_{\xi_{n+1}}(l)| - 2x\sqrt{x^2 + |\hat{R}_{\xi_{n+1}}(l)|}]. \end{aligned} \tag{56}$$

Therefore,

$$\begin{aligned} \sum_{n=1}^{\infty} E[Y_n^2] &= E\left[\sum_{n=1}^{\infty} Y_n^2\right] \\ &= E\left[\sum_{n=1}^{\infty} \gamma_n^2\{2x^2 + |\hat{R}_{\xi_{n+1}}(l)| \right. \\ &\qquad \left. - 2x\sqrt{x^2 + |\hat{R}_{\xi_{n+1}}(l)|}\}\right]. \end{aligned} \tag{57}$$

As x is a constant, and as $R_{\xi_{n+1}}(l)$ is finite (except when $l = 0$) and decreasing with l, and because of the nature of γ_n, it can be shown that the right-hand side of (57) is finite. Thus condition (40) is satisfied.

Also, from (55) and (45)

$$E[Y_n \mid X_1,X_2,\cdots,X_n] = \gamma_n \cdot E \cdot [\{x^2 + |R_{\xi_{n+1}}(l)|\}^{1/2} - x]. \tag{58}$$

If l is taken sufficiently large, $|\hat{R}_{\xi_{n+1}}(l)| \to 0$. Hence the right-hand side of (58) approaches zero, satisfying condition (41) with probability one. Thus all the conditions of the

theorem are satisfied proving that the algorithm converges to the correct value in the mean-square sense and with probability one.

References

[1] J. Sklansky, "Learning systems for automatic control," *IEEE Trans. Automat. Contr.*, vol. AC-11, Jan. 1966, pp. 6–19.
[2] G. N. Saridis, Z. Z. Nikolic, and K. S. Fu, "Stochastic approximation algorithms for system identification, estimation and decomposition of mixtures," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-5, Jan. 1969, pp. 8–15.
[3] D. B. Cooper, "Adaptive pattern recognition and signal detection using stochastic approximation," *IEEE Trans. Electron. Comput.* (Corresp.), vol. EC-13, June 1964, pp. 306–307.
[4] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, 1951, pp. 400–407.
[5] J. Kiefer and J. Wolfowitz: "Stochastic estimation of the maximum of a regression function," *Ann. Math. Statist.*, vol. 23, 1952, pp. 462–466.
[6] A. Dvoretzky, "On stochastic approximation," in *Proc. 3rd Berkeley Symp. on Mathematical Statistics and Probability*, vol 1. Los Angeles: Univ. of California Press, 1956, pp. 39–56.
[7] K. S. Fu, Z. Z. Nicolic, T. Y. Chien, and W. G. Wee, "On the stochastic approximation and related learning techniques," Purdue Univ., Lafayette, Ind., Rep. TR-EE66, Apr. 1966.
[8] Y. T. Chien and K. S. Fu, "On Bayesian learning and stochastic approximation," *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-3, June 1967, pp. 28–38.

A User-Oriented Evaluation of a Time-Shared Computer System

SAKARI T. JUTILA, MEMBER, IEEE, AND GIORA BARAM, STUDENT MEMBER, IEEE

Abstract—**Criteria for the evaluation of computer systems are traditionally computer science oriented. This exploratory investigation is concerned with evaluations of users' satisfaction with a time-shared computer system. The first part of the investigation is an experiment indicating that the type of programming language, i.e., Basic versus Fortran, is the relevant significant factor in determining the learning rates for the use of the system. The second part of the investigation is devoted to the study of an additive users' utility function. The variables of this function are the waiting time in the queue for a time-shared computer terminal and the users' problem solving time (users' turnaround time at the terminal). It was found that users became rapidly dissatisfied if these times exceeded 10–15 min. A probability density function for the users' turnaround time at the terminal was obtained in the setting of an industrial laboratory. It turned out to be nearly exponential with a mean of about 12 min.**

Introduction

THERE HAS BEEN extensive research to develop techniques for the evaluation of computer and information systems and to set up criteria for the selection of such systems. The main emphasis, however, has been in the direction of computer sciences and not in the direction of users' satisfaction with various computer systems [1]–[24]. Relatively few recent articles [22]–[24] start diverting from a pure computer science approach. Sackman and Gold [22] emphasized the lack of experimental methods for measuring human behavior in man–computer communications: "experimental skills, statistical know-how, the mysteries of human learning and motivation, and the vagaries of human individual differences are not part of the legacy of the computer sciences."

Lesser and Ralston [20] explain the reason for developing a regional computation center, including a time-shared system, on a brute force trial and error process, rather than on the basis of a comprehensive experimental study: "there is little knowledge about the operation and management of such a service, with regard in particular to economics, reliability, customer acceptance, and variety and nature of user access." One might add that this situation is still a typical one.

Recent studies by Nickerson *et al.* [25] and Carbonel *et al.* [26] treat the man–computer interaction from the users' point of view. However, no specific empirical results are given. Independently from the aforementioned investigations, this exploratory study presents actual empirical results from such a point of view.

Experimental Determination of Learning Rates in Use of Time-Shared Computer System

An important aspect of users' satisfaction with any computer system is the rate at which they learn to use the system effectively. If the process of learning is slow and tedious, users tend to feel dissatisfied. If the learning process is rapid and relatively uncomplicated, the users tend to feel satisfied. An experimental design was developed to detect the predominant factors affecting the users' learning rates for a particular type of application utilizing a time-shared computer system. Thereafter, learning curves of users were estimated in terms of the predominant factors affecting the desirability of a time-shared computer system.

Manuscript received July 6, 1970; revised May 10, 1971. This work was supported in part by the General Electric Co.

The authors are with the Department of Operations Analysis, College of Business Administration, and the Department of Industrial Engineering, College of Engineering, University of Toledo, Toledo, Ohio.