Automation and Remote Control, Vol. 64, No. 2, 2003, pp. 252–262. Translated from Avtomatika i Telemekhanika, No. 2, 2003, pp. 88–99.
Original Russian Text Copyright © 2003 by Granichin.

Optimal Convergence Rate of the Randomized Algorithms of Stochastic Approximation in Arbitrary Noise

O. N. Granichin

St. Petersburg State University, St. Petersburg, Russia

Received August 29, 2002

Abstract—Multidimensional stochastic optimization plays an important role in analysis and control of many technical systems. To solve the challenging multidimensional problems of optimization, it was suggested to use the randomized algorithms of stochastic approximation with perturbed input which are not just simple, but also provide consistent estimates of the unknown parameters for observations in "almost arbitrary" noise. Optimal methods of choosing the parameters of algorithms were motivated.

1. INTRODUCTION

Let us consider by way of example the problem of determining the stationary point θ^* (of local minimum or maximum) of some function $f(\cdot)$, provided that for each value of $\theta \in \mathbb{R}$, that is, input to the algorithm or controllable variable, one observes the random variable

$$Y(\theta) = f(\theta) + V$$

which is the value of the function $f(\cdot)$ polluted by noise at the point θ. To solve this problem under some additional constraints, J. Kiefer and J. Wolfowitz [1] proved that the recurrent sequence obeying the rule (algorithm)

$$\widehat{\theta}_n = \widehat{\theta}_{n-1} - \alpha_n \frac{Y(\widehat{\theta}_{n-1} + \beta_n) - Y(\widehat{\theta}_{n-1} - \beta_n)}{2\beta_n},$$

where $\{\alpha_n\}$ and $\{\beta_n\}$ are some given decreasing numerical sequences with certain characteristics, converges to the point θ^*.

The requirement that the observation noise be conditional centered is the main condition which constrains its properties and usually is assumed to be satisfied. It can be formulated as follows. For the statistics

$$z(\theta, \beta) = \frac{Y(\theta + \beta) - Y(\theta - \beta)}{2\beta}$$

whose sampled values are precisely observed or calculated and for small β, the conditional expectation is close to the derivative of the function $\mathrm{E}\{z(\theta, \beta) \mid \theta\} \approx f'(\theta)$.

Behavior of the sequence of estimates determined by the algorithm of stochastic approximation (SA) depends on the choice of the observed statistic functions $z(\theta, \beta)$. In some applications, information about the statistic characteristics of the measurement errors may be insufficient or they may be defined by a deterministic function which the experimenter does not know. In this case, one encounters appreciable difficulties in motivating applicability of the conventional Kiefer–Wolfowitz (KW) procedure whose estimate often does not converge to the desired point. However, this does not suggest that in dealing with these problems one must abandon the easily representable SA algorithms. Let us assume that the function $f(\cdot)$ is twice continuously differentiable and given is

0005-1179/03/6402-0252$25.00 © 2003 MAIK "Nauka/Interperiodica"

an observed realization of the Bernoulli sequence of independent random variables $\{\Delta_n\}$ that are equal to ±1 with the same probability and not correlated at the nth step with the observation errors. We modify the KW procedure using the randomized statistics $\widetilde{z}(\theta, \beta, \Delta) = z(\theta, \beta, \Delta)$. By expanding the function $f(\theta)$ by the Taylor formula and making use of the fact that Δ_n and the observation noise are not correlated, we obtain for this new statistics that

$$\mathrm{E}\{\widetilde{z}(\theta,\beta,\Delta) \mid \theta\} = f'(\theta) + \mathrm{E}\left\{\frac{1}{\Delta}V\right\} + \mathcal{O}(\beta) = f'(\theta) + \mathcal{O}(\beta).$$

If the values of the numerical sequence $\{\beta_n\}$ in the algorithm tend to zero, then in the limit it coincides "in the mean" with the value of the derivative of $f(\cdot)$. A simpler statistics

$$\overline{z}(\theta,\beta,\Delta) = \frac{\Delta}{\beta}Y(\theta + \beta\Delta)$$

that at each iteration (step) uses only one observation has the same characteristics. The observation sequence can be enriched by adding to the algorithm and observation channel a new random process $\{\Delta_n\}$ called the simultaneous trial perturbation. The trial perturbation in essence is an exciting action because it is used mostly to make the observations nondegenerate.

In the multidimensional case of $\theta \in \mathbb{R}^r$, the conventional KW procedure based on finite-difference approximations at each iteration of the function gradient vector makes use of $2r$ observations (two observations for approximation of each component of the r-dimensional gradient vector) to construct the estimate sequence. The randomized statistics $\widetilde{z}(\theta,\beta,\Delta)$ and $\overline{z}(\theta,\beta,\Delta)$ admit a computationally simpler procedure of generalization to the multidimensional case which at each iteration employs only two or one measurement(s) of the function. Let $\{\Delta_n\}$ be an r-dimensional Bernoulli random process. Then,

$$\widetilde{z}(\theta,\beta,\Delta) = \begin{pmatrix} \dfrac{1}{\Delta^{(1)}} \\ \dfrac{1}{\Delta^{(2)}} \\ \vdots \\ \dfrac{1}{\Delta^{(r)}} \end{pmatrix} \frac{Y(\theta+\beta\Delta) - Y(\theta-\beta\Delta)}{2\beta},$$

and for $\overline{z}(\theta,\beta,\Delta)$ the formula is the same as in the scalar case. J.C. Spall [2] proposed to use the statistics $\widetilde{z}(\theta,\beta,\Delta)$ and demonstrated that for large n the probabilistic distribution of appropriately scaled estimation errors is approximately normal. He used the formula obtained for the asymptotic error variance and a similar characteristic of the conventional KW procedure to compare two algorithms. It was found out that, all other things being equal, the new algorithm has the same convergence rate as the conventional KW procedure. Despite the fact that in the multidimensional case appreciably less (by the factor of r for $n \to \infty$) observations are used. The present author was the first to use the statistics $\overline{z}(\theta,\beta,\Delta)$ in the scalar case [3] for constructing a sequence of consistent estimates for almost arbitrary observation errors.

Convergence rate of the estimates of the SA algorithms seems to be the main stimulus to modify the original algorithms. The properties of estimates of the conventional KW procedure and some of its generalizations were studied in detail in many works [4–11]. The estimate convergence rate depends on the smoothness of $f(\cdot)$. If it is twice differentiable, then the rms error of the conventional KW algorithm decreases as $\mathcal{O}\left(n^{-\frac{1}{2}}\right)$; if it is thrice differentiable, as $\mathcal{O}\left(n^{-\frac{2}{3}}\right)$ [4]. V. Fabian [12] modified the KW procedure by using, besides an approximation of the first derivative, higher-order finite-difference approximations with certain weights. If the function $f(\cdot)$ has ℓ continuous derivatives, then the Fabian algorithm provides the rms convergence rate of the order $\mathcal{O}\left(n^{-\frac{\ell-1}{\ell}}\right)$ for odd ℓ's. In computational terms, Fabian's algorithm is overcomplicated, the number of observations at one iteration growing rapidly with smoothness and dimensionality; also, at each step one has to invert a matrix. The asymptotic rms convergence rate can be increased without increasing the number of measurements of the function at each iteration if in problems with sufficiently smooth functions $f(\cdot)$ the randomized SA algorithms are used. For the case where some generalized measure of smoothness of $f(\cdot)$ is equal to γ ($\gamma = \ell+1$ if all partial derivatives of orders up to ℓ inclusive satisfy the Lipschitz condition), B.T. Polyak and A.B. Tsybakov [13] proposed to use a statistics of the form

$$\widetilde{z}_\gamma(\theta,\beta,\Delta) = \mathrm{K}(\Delta)\frac{Y(\theta+\beta\Delta)-Y(\theta-\beta\Delta)}{2\beta},$$
$$\overline{z}_\gamma(\theta,\beta,\Delta) = \frac{1}{\beta}\mathrm{K}\,(\Delta)Y(\theta+\beta\Delta),$$

where $\mathrm{K}\,(\cdot)$ is some vector function with finite support (differentiable kernel) determined by means of the orthogonal Legendre polynomials of a degree smaller than γ. Two corresponding randomized algorithms provide the rms convergence rate $\mathcal{O}\left(n^{-\frac{\gamma-1}{\gamma}}\right)$ of the estimate sequence. The same paper demonstrated that for a wide class of algorithms this convergence rate is optimal in some asymptotically minimax sense, that is, cannot be improved either for any other algorithm or for any other admissible rule of choosing the measurement points. For odd ℓ, this fact was earlier established by H.F. Chen [14].

Polyak and Tsybakov [13] also noted that the algorithm with one measurement asymptotically behaves worse than that with two measurements. As will be shown below, this is not quite true if one compares the algorithms in terms of the number of iterations multiplied by the number of measurements. Additionally, in many applications such as optimization of the real-time systems, the dynamic processes underlying the mathematical model can be too fast to enable two successive measurements. In some problems, at one step of the algorithm it is merely impossible to make two measurements such that the observation errors are uncorrelated with Δ_n at both points $\widehat{\theta}_{n-1}+\beta_n\Delta_n$ and $\widehat{\theta}_{n-1}-\beta_n\Delta_n$, which is one of the main conditions for applicability of the algorithm.

The reader is referred to [15] for the conditions for convergence of the randomized SA algorithms for "almost arbitrary" noise and for the corresponding references. Deterministic analysis of convergence of the randomized algorithms of stochastic approximation was done in [16, 17]. Convergence rate of the algorithms was considered also in [18, 19].

2. FORMULATION OF THE PROBLEM AND BASIC ASSUMPTIONS

Let $F(w,\theta) : \mathbb{R}^p \times \mathbb{R}^r \to \mathbb{R}^1$ be a function differentiable with respect to the second argument, $x_1, x_2, \ldots$ be the sequence of the measurement points (plan of observation) chosen by the experimenter where at each time instant $n = 1, 2, \ldots$ the value of the function $F(w_n, \cdot)$ is observable with additive noise v_n:

$$y_n = F(w_n, x_n) + v_n,$$

where $\{w_n\}$ is a noncontrollable sequence of random variables ($w_n \in \mathbb{R}^p$) having identical and, generally speaking, unknown distribution $\mathrm{P}_w\,(\cdot)$ with finite support.

Formulation of the problem. Needed is to construct from the observations $y_1, y_2, \ldots$ the sequence of estimates $\{\widehat{\theta}_n\}$ of the unknown vector θ^* minimizing the function

$$f(\theta) = \int\limits_{\mathbb{R}^p} F(w,\theta)\mathrm{P}_w(dw)$$

of the kind of mean-risk functional. The same formulation was considered in [15] which also aimed at optimizing the convergence rate of the sequence of estimates.

To formulate the basic assumptions, we make use of the notation $\|\cdot\|$ and $(\cdot,\cdot)$ for the Euclidean norm and the scalar product in $\mathbb{R}^r$, respectively.

(**A.1**) The function $f(\cdot)$ is strongly convex, that is, has a single minimum in $\mathbb{R}^r$ at some point $\theta^* = \theta^*(f(\cdot))$ and

$$(x-\theta^*, \nabla f(x)) \geq \mu\|x-\theta^*\|^2, \quad \forall x \in \mathbb{R}^r$$

with some constant $\mu > 0$.

(**A.2**) The Lipschitz condition for the gradient of the function $f(\cdot)$

$$\|\nabla f(x) - \nabla f(\theta)\| \leq A\|x-\theta\|, \quad \forall x,\ \theta \in \mathbb{R}^r$$

with some constant $A > \mu$.

(**A.3**) The function $f(\cdot) \in C^\ell$ is ℓ times continuously differentiable and all its partial derivatives of the order up to ℓ inclusive satisfy on $\mathbb{R}^r$ the Hölder condition of the order ρ, $0 < \rho \leq 1$:

$$\left| f(x) - \sum_{|\overline{\ell}|\leq\ell} \frac{1}{\overline{\ell}!} D^{\overline{\ell}} f(\theta)(x-\theta)^{\overline{\ell}} \right| \leq M\|x-\theta\|^\gamma,$$

where $\gamma = \ell + \rho \geq 2$, M is some constant, $\overline{\ell} = \left(\ell^{(1)}, \ldots, \ell^{(r)}\right)^{\mathrm{T}} \in \mathbb{N}^r$ is a multiindex, $\ell^{(i)} \geq 0$, $i = 1, \ldots, r$, $|\overline{\ell}| = \ell^{(1)} + \ldots + \ell^{(r)}$, $\overline{\ell}! = \ell^{(1)}! \ldots \ell^{(r)}!$, $x \in \mathbb{R}^r$, $x^{\overline{\ell}} = \left(x^{(1)}\right)^{\ell^{(1)}} \ldots \left(x^{(r)}\right)^{\ell^{(r)}}$, $D^{\overline{\ell}} = \partial^{|\overline{\ell}|}/\left(\partial x^{(1)}\right)^{\ell^{(1)}} \ldots \left(\partial x^{(r)}\right)^{\ell^{(r)}}$. For $\gamma = 2$, we assume that $M = A/2$.

3. TRIAL PERTURBATION AND THE BASIC ALGORITHMS

Let Δ_n, $n = 1, 2, \ldots$ be the observed sequence of mutually independent and identically distributed random variables in $\mathbb{R}^r$ which below is referred to as the simultaneous trial perturbation. All components of the vector Δ_n are mutually independent and have identical scalar distribution function $\mathrm{P}_\Delta(\cdot)$ with finite support.

We fix some initial vector $\widehat{\theta}_0 \in \mathbb{R}^r$, choose the positive numbers α and β and two scalar bounded functions (kernels) $K_0(\cdot)$ and $K_1(\cdot)$ satisfying

$$\begin{aligned} &\int uK_0(u)\mathrm{P}_\Delta(du) = 1, \quad \int u^k K_0(u)\mathrm{P}_\Delta(du) = 0, \quad k = 0, 2, \ldots, \ell, \\ &\int K_1(u)\mathrm{P}_\Delta(du) = 1, \quad \int u^k K_1 \mathrm{P}_\Delta(du) = 0, \quad k = 1, \ldots, \ell-1. \end{aligned} \tag{1}$$

Two algorithms are proposed for constructing the sequence of measurement points $\{x_n\}$ and estimates $\{\widehat{\theta}_n\}$. At each step (iteration), the first algorithm uses two observations

$$\begin{cases} x_{2n} = \widehat{\theta}_{n-1} + \beta n^{-\frac{1}{2\gamma}}\Delta_n, \quad x_{2n-1} = \widehat{\theta}_{n-1} - \beta n^{-\frac{1}{2\gamma}}\Delta_n \\ y_{2n} = F(w_{2n}, x_{2n}) + v_{2n}, \quad y_{2n-1} = F(w_{2n-1}, x_{2n-1}) + v_{2n-1} \\ \widehat{\theta}_n = \widehat{\theta}_{n-1} - \alpha n^{-1+\frac{1}{2\gamma}} \mathrm{K}\left(\Delta_n\right) \dfrac{y_{2n} - y_{2n-1}}{2}, \end{cases} \tag{2}$$

and the second algorithm, one observation

$$\begin{cases} x_n = \widehat{\theta}_{n-1} + \beta n^{-\frac{1}{2\gamma}}\Delta_n, \quad y_n = F(w_n, x_n) + v_n \\ \widehat{\theta}_n = \mathcal{P}_{\Theta_n}\left(\widehat{\theta}_{n-1} - \alpha n^{-1+\frac{1}{2\gamma}} \mathrm{K}\left(\Delta_n\right) y_n\right). \end{cases} \tag{3}$$

In both algorithms, $\mathrm{K}(\cdot)$ is the vector function with the components obeying

$$\mathrm{K}^{(i)}(x) = K_0(x^{(i)}) \prod_{j \neq i} K_1(x^{(j)}), \quad i, j = 1, \ldots, r, \quad x \in \mathbb{R}^r. \tag{4}$$

In algorithm (3), $\mathcal{P}_{\Theta_n}(\cdot)$, $n = 1, 2, \ldots$, denotes the operators of projection on some convex closed bounded subsets $\Theta_n \subset \mathbb{R}^r$ containing the point θ^* starting from some $n \geq 1$. If the bounded closed convex set Θ including the point θ^* is known in advance, then one can assume that $\Theta_n = \Theta$. Otherwise, the sets $\{\Theta_n\}$ can dilate to infinity.

We follow V.Ya. Katkovnik [7] and Polyak and Tsybakov [13] and show one of the possible methods of constructing the kernels $K_0(\cdot)$ and $K_1(\cdot)$ satisfying conditions (1). We note that in the scalar case one function $K_0(x)$ suffices to define $\mathrm{K}(x)$. Let $\{p_m(\cdot)\}_{m=0}^{\ell}$ be a system of polynomials that is defined on the support of the distribution $\mathrm{P}_\Delta(\cdot)$ and is orthogonal to the measure generated by it. By analogy with the proofs of [13], one can readily verify that the functions

$$K_0(u) = \sum_{m=0}^{\ell} a_m p_m(u), \quad a_m = p_m'(0) \Big/ \int p_m^2(u) \mathrm{P}_\Delta(du),$$

$$K_1(u) = \sum_{m=0}^{\ell-1} b_m p_m(u), \quad b_m = p_m(0) \Big/ \int p_m^2(u) \mathrm{P}_\Delta(du)$$

satisfy conditions (1).

It was proposed in [13] to take a distribution that is uniform over the interval $\left[-\frac{1}{2}, \frac{1}{2}\right]$ as the probabilistic distribution of the components of the trial perturbation $\mathrm{P}_\Delta(\cdot)$ and to construct over this interval the kernels $K_0(\cdot)$ and $K_1(\cdot)$ by the orthogonal Legendre polynomials. In this case, we obtain $K_0(u) = 12u$ and $K_1(u) = 1$ for the initial values of $\ell = 1, 2$, that is, $2 \leq \gamma \leq 3$, $K_0(u) = 5u(15 - 84u^2)$ and $K_1(u) = 9/4 - 15u^2$ for the subsequent values of $\ell = 3, 4$, that is, $3 < \gamma \leq 5$, and for $|u| > 1/2$ both functions are equal to zero.

Attention to a type of probabilistic distributions of the trial perturbation that is more general than that considered in [13] is due to the fact that in practice the statement of the problem itself defines a certain type of the distribution of the trial perturbation $\mathrm{P}_\Delta(\cdot)$ that can be conveniently modelled, whereas in some cases it is more suitable to use distributions only from some narrow fixed class. The possibility of choosing among various systems of orthogonal polynomials enables one to get estimation of appropriate quality for the same asymptotic order of the convergence rate.

4. CONVERGENCE RATE

We denote by $\mathbb{W} = \sup(\mathrm{P}_w(\cdot)) \subset \mathbb{R}^p$ the support of the distribution $\mathrm{P}_w(\cdot)$, by $\mathcal{F}_{n-1}$ the σ-algebra of probabilistic events generated by the random variables $\widehat{\theta}_0, \widehat{\theta}_1, \ldots, \widehat{\theta}_{n-1}$ obtained using algorithm (2) (or (3)):

$$\overline{v}_n = v_{2n} - v_{2n-1}, \quad \overline{w}_n = \begin{pmatrix} w_{2n} \\ w_{2n-1} \end{pmatrix}, \quad d_n = 1$$

if algorithm (2) is used or

$$\overline{v}_n = v_n, \quad \overline{w}_n = w_n, \quad d_n = \mathrm{diam}\,(\Theta_n)$$

if the estimates are constructed by algorithm (3); here $\mathrm{diam}\,(\cdot)$ is the Euclidean diameter of the set.

The following theorem establishes the sufficient conditions for optimality of the asymptotic convergence rate of algorithms (2) and (3).

Theorem 1. *If the following conditions are satisfied:*
(**1**) *for the functions $K_0(\cdot)$, $K_1(\cdot)$, and $\mathrm{P}_\Delta(\cdot)$;*
(**A.1, 3**) *for $\gamma \geq 2$, $\alpha\beta > \dfrac{\gamma-1}{2\mu\gamma}$ for the function $f(\theta) = \mathrm{E}\{F(w,\theta)\}$;*
(**A.2**) *for the functions $F(w,\cdot)$ $\forall w \in \mathbb{W}$;*
for any $\theta \in \mathbb{R}^r$ the functions $F(\cdot,\theta)$ and $\nabla_\theta F(\cdot,\theta)$ are uniformly bounded on $\mathbb{W}$;
$d_n n^{-1+\frac{1}{2\gamma}} \to 0$ for $n \to \infty$;
for any $n \geq 1$, the random vectors $\overline{w}_n$ and Δ_n are independent of $\overline{v}_1, \ldots, \overline{v}_n$, $\overline{w}_1, \ldots, \overline{w}_{n-1}$ and the random vector Δ_n is independent of $\overline{w}_n$;

$$\mathrm{E}\{(v_{2n} - v_{2n-1})^2/2\} \leq \sigma_2^2 \quad (\mathrm{E}\{v_n^2\} \leq \sigma_1^2),$$

then for $n \to \infty$

$$\mathrm{E}\left\{\left\|\widehat{\theta}_n - \theta^*\right\|^2\right\} = \mathcal{O}\left(n^{-\frac{\gamma-1}{\gamma}}\right)$$

is asymptotically satisfied for the rms convergence rate of the sequence of estimates $\{\widehat{\theta}_n\}$ generated by algorithm (2) (*or* (3)).

Theorem 1 is proved in the Appendix.

The present author proved [15] that if the additional condition $\sum\limits_n n^{-2+1/\gamma}\mathrm{E}\ \{\overline{v}_n^2 \mid \mathcal{F}_{n-1}\} < \infty$ is satisfied with unit probability, then the sequence of estimates converges with probability one: $\widehat{\theta}_n \to \theta^*$ for $n \to \infty$.

The resulting estimates of the order of the rms convergence rate are optimal. As was shown in [13], for the class of all functions satisfying conditions (**A.1**)–(**A.3**), there exists no algorithm which is asymptotically more efficient in some minimax sense. For a special case, the rules for choosing the corresponding optimal randomized algorithms estimating the unknown parameters of a linear regression can be found in [20].

In Theorem 1, the observation noise v_n can be conditionally called "almost arbitrary" noise because it can be nonrandom, but unknown and bounded or be a realization of some stochastic process with arbitrary structure of dependences. In particular, there is no need to assume something about the dependence between $\overline{v}_n$ and $\mathcal{F}_{n-1}$ in order to prove Theorem 1.

The condition for independence of the observation noise of the trial perturbation can be relaxed. It suffices to require that for $n \to \infty$ the conditional mutual correlation between $\overline{v}_n$ and $\mathrm{K}(\Delta_n)$: $\|\mathrm{E}\{\overline{v}_n \mathrm{K}(\Delta_n) \mid \mathcal{F}_{n-1}\}\| = \mathcal{O}\left(n^{-1+\frac{1}{2\gamma}}\right)$ tends to zero with unit probability.

For convenient notation of the formulas below, we define the constant χ as equal to one if the observation noise $\{v_n\}$ is independent and centered or two in all other cases. We denote $\widehat{K} = \int \|\mathrm{K}(x)\|^2 \mathrm{P}(dx)$, $\overline{K} = \int \|x\|^\gamma \|\mathrm{K}(x)\| \mathrm{P}(dx)$, $\mathrm{P}(x) = \prod\limits_{i=1}^r \mathrm{P}_\Delta(x^{(i)})$. The quantities $\widehat{K}$ and $\overline{K}$ are finite owing to boundedness of the vector function $\mathrm{K}(\cdot)$ and finiteness of the support of the distribution $\mathrm{P}_\Delta(\cdot)$. The optimal values of the parameters

$$\alpha^* = 1/(\mu\beta^*), \quad \beta^* = \left(2\chi(\nu_1 + \sigma_i^2/i)\widehat{K}\right)^{\frac{1}{2\gamma}} \left(\sqrt{\gamma(\gamma-1)}M\overline{K}\right)^{-\frac{1}{\gamma}}$$

and quantitative estimates of the asymptotic convergence rate of algorithms (2) and (3)

$$\mathrm{E}\left\{\left\|\widehat{\theta}_n-\theta^*\right\|^2\right\}\le n^{-\frac{\gamma-1}{\gamma}}\varkappa_i\widehat{K}^{\frac{\gamma-1}{\gamma}}\overline{K}^{\frac{2}{\gamma}}+o\left(n^{-\frac{\gamma-1}{\gamma}}\right),$$

$$\varkappa_i=\gamma^{\frac{1+\gamma}{\gamma}}(\gamma-1)^{\frac{1-\gamma}{\gamma}}\mu^{-2}\left(\chi(\nu_i+\sigma_i^2/i)\right)^{\frac{\gamma-1}{\gamma}}M^{\frac{2}{\gamma}},\quad i=1,2,$$

where $\nu_1=\sup\limits_{w\in\mathbb{W}}\left(F(w,\theta^*)+\frac{1}{2}(\nabla_\theta F(w,\theta^*))^2\right)^2$ corresponds to the one-measurement algorithm (3) and $\nu_2=\sup\limits_{w_1,w_2\in\mathbb{W}}\left(2|F(w_1,\theta^*)-F(w_2,\theta^*)|+(\nabla_\theta F(w_1,\theta^*))^2+(\nabla_\theta F(w_2,\theta^*))^2\right)^2/8$, to (2), will be established when proving Theorem 1 for $M>0$. If $F(w,x)=f(x)$, then, as can be seen from the comparison of $\varkappa_1$ and $\varkappa_2$, the asymptotic convergence rate for the iterations of the estimate sequence as generated by the two-observation algorithm (2) is always superior to that of algorithm (3). For comparison with regard for the number of observations, the advantage of algorithm (2) is no more indisputable. By comparing $\varkappa_1$ and $2\varkappa_2$, one can readily see that for $2^{\frac{1}{\gamma-1}}\sigma_2^2-\sigma_1^2>\nu_1-2^{\frac{1}{\gamma-1}}\nu_2$ and with account for the number of measurements the asymptotic convergence rate is better for algorithm (3) than (2).

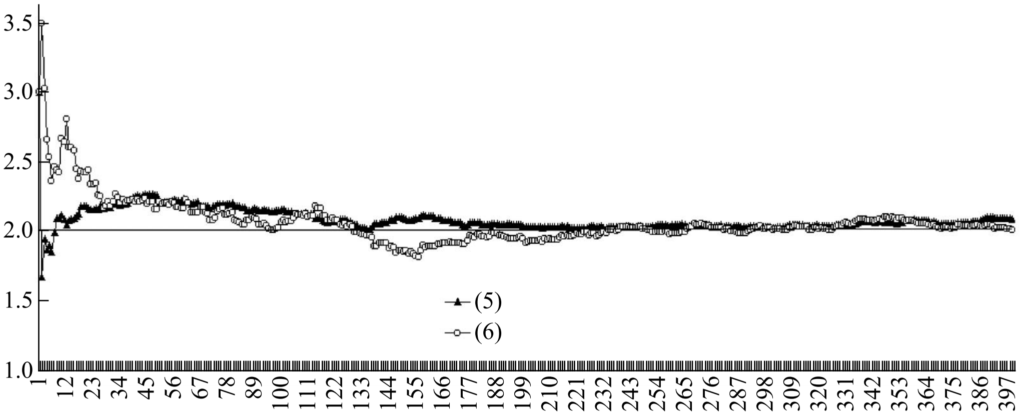
In the scalar case, for $F(w,x)=f(x)$ with the simultaneous trial perturbation $\{\Delta_n\}$ generated by independent and identically distributed random variables from the interval $\left[-\frac{1}{2},\frac{1}{2}\right]$ and independent centered observation noise $\{v_n\}:\mathrm{E}\{v_n^2\}\le\sigma_v^2$, $\gamma=2$, and $\mathrm{K}(x)=12x$, $|x|\le1/2$, we obtain on the strength of the established estimates of the convergence rate for algorithm (2) that

$$\mathrm{E}\left\{\left\|\widehat{\theta}_n-\theta^*\right\|^2\right\}\le\frac{9A\sigma_v}{4\sqrt{3}\,\mu^2}n^{-1/2}+o(n^{-1/2}),\quad\alpha^*=\frac{1}{\mu\beta^*},\quad\beta^*=\frac{4\sqrt{\sigma_v}}{\sqrt{A}\sqrt[4]{3}},$$

and for (3), $\mathrm{E}\left\{\left\|\widehat{\theta}_n-\theta^*\right\|^2\right\}\le4.5\sqrt{f(\theta^*)^2+\sigma_v^2}/(\sqrt{6}\,\mu^2)n^{-1/2}+o(n^{-1/2})$. Hence, if $f(\theta^*)^2<\sigma_v^2$, then algorithm (3) is preferable.

5. EXAMPLE

Several computer simulations were run to illustrate the typical behavior of the estimates of the randomized SA algorithms. Consideration was given, in particular, to the elementary example of [17]. Let $F(w,x)=f(x)=x^2-4x-2$. This function is smooth and has a single minimum for $x=2$. Its constants A and μ are equal to two. The function was measured in unknown additive



Trajectories of the estimates of algorithms (5) and (6).

noise that is uniformly distributed over the interval $\left[-\frac{1}{2}, \frac{1}{2}\right]$: $\mathrm{E}\{v_n^2\} = 1/12$. The figure depicts typical trajectories of estimates for two algorithms of the type of (2) corresponding to $\gamma = 2$

$$\begin{cases} y_{2n-i} = f\left(\widehat{\theta}_{n-1} + (-1)^i \frac{2}{\sqrt{3}} n^{-\frac{1}{4}} \Delta_n\right) + v_{2n-i}, \quad i = 0, 1 \\ \widehat{\theta}_n = \widehat{\theta}_{n-1} - 1.5\sqrt{3}\, n^{-\frac{3}{4}} \Delta_n (y_{2n} - y_{2n-1}) \end{cases} \tag{5}$$

and $\gamma = 5$

$$\begin{cases} y_{2n-i} = f\left(\widehat{\theta}_{n-1} + (-1)^i 2 n^{-\frac{1}{10}} \Delta_n\right) + v_{2n-i}, \quad i = 0, 1 \\ \widehat{\theta}_n = \widehat{\theta}_{n-1} - \frac{5}{16} n^{-\frac{9}{10}} (15\Delta_n - 84\Delta_n^3)(y_{2n} - y_{2n-1}). \end{cases} \tag{6}$$

The initial approximation $\widehat{\theta}_0 = 3$. Both algorithms used realizations of independent random variable uniformly distributed over the interval $\left[-\frac{1}{2}, \frac{1}{2}\right]$ as the simultaneous trial perturbation.

6. CONCLUSIONS

Retention of simplicity, operability, and optimal convergence rate with increased dimensionality of the vector of estimated parameters is a striking characteristic of the considered randomized SA algorithms. Since this is not accompanied by an increase in the number of measurements required for each iteration, application of these algorithms in the systems with infinite-dimensional and distributed parameters can be an important further step of their development. To the author's opinion, the replacement in the multidimensional case of numerous finite-difference approximations of the gradient of objective function by only one or two measurements at randomly chosen points is intuitively much closer to the behavioral model of highly organized living systems. It seems that the algorithms of this kind could be naturally used of design the artificial intelligence systems.

ACKNOWLEDGMENTS

The author would like to thank B.T. Polyak for his help and valuable remarks about the present paper.

APPENDIX

Proof of Theorem 1. Let us first consider algorithm (3). For sufficiently great n for which $\theta^* \in \Theta_n$, one can readily obtain by using the property of projection that

$$\left\|\widehat{\theta}_n - \theta^*\right\|^2 \leq \left\|\widehat{\theta}_{n-1} - \theta^* - \alpha n^{-1+\frac{1}{2\gamma}} \mathrm{K}(\Delta_n) y_n\right\|^2.$$

By applying to this inequality the operation of conditional expectation relative to the σ-algebra $\mathcal{F}_{n-1}$, we obtain

$$\begin{aligned} \mathrm{E}\left\{\left\|\widehat{\theta}_n - \theta^*\right\|^2 \mid \mathcal{F}_{n-1}\right\} &\leq \left\|\widehat{\theta}_{n-1} - \theta^*\right\|^2 \\ -2\alpha n^{-1+\frac{1}{2\gamma}} &\left(\widehat{\theta}_{n-1} - \theta^*, \mathrm{E}\left\{y_n \mathrm{K}(\Delta_n) \mid \mathcal{F}_{n-1}\right\}\right) \\ +\alpha^2 n^{-2+\frac{1}{\gamma}} &\mathrm{E}\left\{y_n^2 \|\mathrm{K}(\Delta_n)\|^2 \mid \mathcal{F}_{n-1}\right\}. \end{aligned} \tag{7}$$

One can readily get from (4) and condition (1) that $\int \mathrm{K}(x)\mathrm{P}(dx) = 0$. Therefore, by virtue of independence of Δ_n with v_n, we obtain $\mathrm{E}\{v_n\mathrm{K}(\Delta_n) \mid \mathcal{F}_{n-1}\} = 0$ and, consequently,

$$\mathrm{E}\{y_n\mathrm{K}(\Delta_n) \mid \mathcal{F}_{n-1}\} = \iint F\left(w, \widehat{\theta}_{n-1} + \beta n^{-\frac{1}{2\gamma}}x\right)\mathrm{K}(x)\mathrm{P}(dx)\mathrm{P}_w(dw).$$

We note that also by virtue of (4) and (1)

$$\beta^{-1}n^{\frac{1}{2\gamma}}\int\sum_{|\overline{\ell}|\le\ell}\frac{1}{\overline{\ell}!}D^{\overline{\ell}}f(\widehat{\theta}_{n-1})\beta^{|\overline{\ell}|}n^{-\frac{|\overline{\ell}|}{2\gamma}}x^{\overline{\ell}}\mathrm{K}(x)\mathrm{P}(dx) = \nabla f(\widehat{\theta}_{n-1}).$$

We obtain from the definition of the function $f(\cdot)$ that

$$\beta^{-1}n^{\frac{1}{2\gamma}}\mathrm{E}\{y_n\mathrm{K}(\Delta_n) \mid \mathcal{F}_{n-1}\} = \nabla f(\widehat{\theta}_{n-1}) + \beta^{-1}n^{\frac{1}{2\gamma}}\int\Bigg(f\left(\widehat{\theta}_{n-1} + \beta n^{-\frac{1}{2\gamma}}x\right)$$
$$-\sum_{|\overline{\ell}|\le\ell}\frac{1}{\overline{\ell}!}D^{\overline{\ell}}f\left(\widehat{\theta}_{n-1}\right)\beta^{|\overline{\ell}|}n^{-\frac{|\overline{\ell}|}{2\gamma}}x^{\overline{\ell}}\Bigg)\mathrm{K}(x)\mathrm{P}(dx).$$

If condition (**A.3**) is satisfied, then the following inequality is valid:

$$\left|\int\left(f\left(\widehat{\theta}_{n-1} + \beta_n x\right) - \sum_{|\overline{\ell}|\le\ell}\frac{1}{\overline{\ell}!}D^{\overline{\ell}}f\left(\widehat{\theta}_{n-1}\right)\beta^{|\overline{\ell}|}n^{-\frac{|\overline{\ell}|}{2\gamma}x^{\overline{\ell}}}\right)\mathrm{K}(x)\mathrm{P}(dx)\right|$$
$$\le M\int\left\|x\beta n^{-\frac{1}{2\gamma}}\right\|^{\gamma}\|\mathrm{K}(x)\|\mathrm{P}(dx) \le M\overline{K}\beta^{\gamma}n^{-\frac{1}{2}}.$$

By using the inequality

$$\left\|\widehat{\theta}_{n-1} - \theta^*\right\| \le \frac{\varepsilon^{-1}n^{-\frac{\gamma-1}{2\gamma}}M\overline{K}\beta^{\gamma-1} + \varepsilon\left(n^{-\frac{\gamma-1}{2\gamma}}M\overline{K}\beta^{\gamma-1}\right)^{-1}\|\theta_{n-1} - \theta^*\|^2}{2},$$

which is valid for any $\varepsilon > 0$, and substituting the above relations into the second term of the right-hand side of (7), we successively obtain from condition (**A.1**) that

$$\mathrm{E}\left\{\left\|\widehat{\theta}_n - \theta^*\right\|^2 \mid \mathcal{F}_{n-1}\right\} \le \left\|\widehat{\theta}_{n-1} - \theta^*\right\|^2 - 2\alpha\beta n^{-1}\left(\widehat{\theta}_{n-1} - \theta^*, \nabla f\left(\widehat{\theta}_{n-1}\right)\right)$$
$$+2\alpha\beta n^{-1-\frac{\gamma-1}{2\gamma}}M\overline{K}\beta^{\gamma-1}\left\|\widehat{\theta}_{n-1} - \theta^*\right\| + \alpha^2 n^{-2+\frac{1}{\gamma}}\mathrm{E}\{y_n^2\|\mathrm{K}(\Delta_n)\|^2 \mid \mathcal{F}_{n-1}\}$$
$$\le \left\|\widehat{\theta}_{n-1} - \theta^*\right\|^2\left(1 - \alpha\beta(2\mu - \varepsilon)n^{-1}\right) + n^{-2+\frac{1}{\gamma}}\Bigg(\alpha\beta^{2\gamma-1}\varepsilon^{-1}M^2\overline{K}^2$$
$$+\alpha^2\widehat{K}\chi\Bigg(\iint F\left(w, \widehat{\theta}_{n-1} + \beta n^{-\frac{1}{2\gamma}}x\right)^2\mathrm{P}(dx)\mathrm{P}_w(dw) + \mathrm{E}\{v_n^2 \mid \mathcal{F}_{n-1}\}\Bigg)\Bigg).$$

As it was the case with the proof of Theorem 1 in [15], it is possible to obtain from condition (**A.2**) that

$$\left|F\left(w, \widehat{\theta}_{n-1} + \beta n^{-\frac{1}{2\gamma}}x\right)\right| \le \sqrt{\nu_1} + (2A+1)\left(\left\|\widehat{\theta}_{n-1} - \theta^*\right\|^2 + \left\|\beta n^{-\frac{1}{2\gamma}}x\right\|^2\right)$$

uniformly over $w \in \mathbb{W}$. We take this fact in account and conclude by taking the unconditional expectation and bearing in mind the condition of theorem for $\{d_n\}$ that

$$\mathrm{E}\left\{\left\|\widehat{\theta}_n - \theta^*\right\|^2\right\} \le \mathrm{E}\left\{\left\|\widehat{\theta}_{n-1} - \theta^*\right\|^2\right\}\left(1 - \psi n^{-1} + o(n^{-1})\right) + C_1 n^{\frac{1}{\gamma}-2} + o\left(n^{\frac{1}{\gamma}-2}\right),$$

where $\psi = \alpha\beta(2\mu - \varepsilon)$, $C_1 = \alpha\beta^{2\gamma-1}\varepsilon^{-1}M^2\overline{K}^2 + \alpha^2\widehat{K}\chi(\nu_1 + \sigma_1^2)$. According to the Chung lemma (see [8], p. 51), if $\psi > (\gamma - 1)/\gamma$, then

$$n^{1-\frac{1}{\gamma}}\mathrm{E}\left\{\left\|\widehat{\theta}_n - \theta^*\right\|^2\right\} \le C_1\left(\alpha\beta(2\mu - \varepsilon) - \frac{\gamma - 1}{\gamma}\right)^{-1} + o(1). \qquad (8)$$

It remains to note that since $\varepsilon > 0$ is arbitrary, satisfaction of the inequality $\psi > (\gamma - 1)/\gamma$ is equivalent to the condition $2\mu\alpha\beta > (\gamma - 1)/\gamma$.

The proof for algorithm (2) differs only in some technicalities. In particular, one needs to use the estimate

$$\frac{1}{2}\left(F(w_1, \theta + x) - F(w_2, \theta - x)\right)^2 \le \nu_2 + 2(2A + 1)^2\left(\|x\|^2 + \|\theta - \theta^*\|^2\right)^2$$

which is uniform in $w \in \mathbb{W}$ and can be easily obtained if the conditions of Theorem 1 are satisfied.

The proof provides an asymptotic estimate of the rms convergence rate that is similar to (8), but has the constant $C_2 = \alpha\beta^{2\gamma-1}\varepsilon^{-1}M^2\overline{K}^2 + \alpha^2\widehat{K}\chi(\nu_2 + \sigma_2^2/2)$ instead of C_1. It is of interest to note the relationship between the constants C_1 and $C_2 : C_1 = C_2 + \widehat{K}\alpha^2\chi(\nu_1 + \sigma_1^2 - \sigma_2^2/2 - \nu_2)$. If $F(w, x) = f(x)$, then the asymptotic convergence rate in iterations of the algorithm using two observations is always superior to that of algorithm (3). In the general case, one has to compare $\nu_1 + \sigma_1^2$ and $\nu_2 + \sigma_2^2/2$.

The right-hand side of (8) is a function of α, β, and ε. Optimization in these parameters provides the values of α^*, β^*, and $\varepsilon^* = 2\mu/\gamma$. Similarly, by optimizing in α, β, and ε the upper bound of the asymptotic convergence rate for the estimates of algorithm (2), one establishes the optimal values of its parameters, which completes the proof of Theorem 1 and remark to it.

REFERENCES

1. Kiefer, J. and Wolfowitz, J., Statistical Estimation on the Maximum of a Regression Function, *Ann. Math. Statist.*, 1952, vol. 23, pp. 462–466.

2. Spall, J.C., Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation, *IEEE Trans. Autom. Control*, 1992, vol. 37, pp. 332–341.

3. Granichin, O.N., Stochastic Approximation with Input Perturbation in Dependent Observation Noise, *Vestn. LGU*, 1989, vol. 1, no. 4, pp. 27–31.

4. Wazan, M.T., *Stochastic Approximation*, Cambridge: Cambridge Univ. Press, 1969. Translated under the title *Stokhasticheskaya approksimatsiya*, Moscow: Mir, 1972.

5. Nevel'son, M.B. and Khas'minskii, R.Z., *Stokhasticheskaya approksimatsiya i rekurrentnoe otsenivanie* (Stochastic Approximation and Recurrent Estimation), Moscow: Nauka, 1972.

6. Ermol'ev, Yu.M., *Metody stokhasticheskogo programmirovaniya* (Methods of Stochastic Programming), Moscow: Nauka, 1976.

7. Katkovnik, V.Ya., *Lineinye otsenki i stokhasticheskie zadachi optimizatsii* (Linear Estimates and Stochastic Problems of Optimization), Moscow: Nauka, 1976.

8. Polyak, B.T., *Vvedenie v optimizatsiyu*, Moscow: Nauka, 1983. Translated under the title *Introduction to Optimization*, New York: Optimization Software, 1987.

9. Fomin, V.N., *Rekurrentnoe otsenivanie i adaptivnaya fil'tratsiya* (Recurrent Estimation and Adaptive Filtration), Moscow: Nauka, 1984.

10. Mikhalevich, V.S., Gupal, A.M., and Norkin, V.I., *Metody nevypukloi optimizatsii* (Methods of Nonconvex Optimization), Moscow: Nauka, 1987.

11. Kushner, H.J. and Yin, G.G., *Stochastic Approximation Algorithms and Applications*, New York: Springer, 1997.

12. Fabian, V., Stochastic Approximation of Minima with Improved Asymptotic Speed, *Ann. Math. Statist.*, 1967, vol. 38, pp. 191–200.

13. Polyak, B.T. and Tsybakov, A.B., Optimal Orders of Accuracy of the Searching Algorithms of Stochastic Approximation, *Probl. Peredachi Inf.*, 1990, no. 2, pp. 45–53.

14. Chen, H.F., Lower Rate of Convergence for Locating a Maximum of a Function, *Ann. Statist.*, 1988, vol. 16, pp. 1330–1334.

15. Granichin, O.N., Randomized Algorithms of Stochastic Approximation in Arbitrary Noise, *Avtom. Telemekh.*, 2002, no. 2, pp. 44–55.

16. Wang, I.-J. and Chong, E., A Deterministic Analysis of Stochastic Approximation with Randomized Directions, *IEEE Trans. Autom. Control*, 1998, vol. 43, pp. 1745–1749.

17. Chen, H.F., Duncan, T.E., and Pasik-Duncan, B., A Kiefer–Wolfowitz Algorithm with Randomized Differences, *IEEE Trans. Autom. Control*, 1999, vol. 44, no. 3, pp. 442–453.

18. Polyak, B.T. and Tsybakov, A.B., On Stochastic Approximation with Arbitrary Noise (the KW Case), in *Topics in Nonparametric Estimation*, *Adv. Sov. Math., Am. Math. Soc.*, Khasminskii, R.Z., Ed., 1992, no. 12, pp. 107–113.

19. Gerencser, L., Convergence Rate of Moments in Stochastic Approximation with Simultaneous Perturbation Gradient Approximation and Resetting, *IEEE Trans. Autom. Control*, 1999, vol. 44, pp. 894–905.

20. Granichin, O.N., Estimating the Parameters of a Linear Regression in Arbitrary Noise, *Avtom. Telemekh.*, 2002, no. 1, pp. 30–41.

This paper was recommended for publication by B.T. Polyak, a member of the Editorial Board