New stochastic approximation type procedures

Article · January 1990

CITATIONS
83

READS
547

1 author:

Boris T. Polyak
Institute of Control Sciences
246 PUBLICATIONS **7,389** CITATIONS

AURCAT 51(7) 937-1008 (1990)

ISSN 0005-1179

Vol. 51, No. 7, Part 2, July, 1990

December 20, 1990

AUTOMATION AND REMOTE CONTROL

АВТОМАТИКА и ТЕЛЕМЕХАНИКА
(AVTOMATIKA I TELEMEKHANIKA)

TRANSLATED FROM RUSSIAN

CONSULTANTS BUREAU, NEW YORK

AUTOMATION AND REMOTE CONTROL

A translation of *Avtomatika i Telemekhanika*

December 20, 1990

Volume 51, Number 7, Part 2 July, 1990

CONTENTS

The Russian press date (podpisano k pechati) of this issue was 6/6/1990. Publication therefore did not occur prior to this date, but must be assumed to have taken place reasonably soon thereafter.

NEW METHOD OF STOCHASTIC APPROXIMATION TYPE

B. T. Polyak

UDC 519.245

An iterative method of solution of problems of stochastic optimization and parameter estimation is considered that combines two processes. The first is a stochastic approximation process with a constant or slowly decreasing step. The second involves averaging this process, so that it is possible to obtain the asymptotically highest feasible convergence rate. The advantage of this method over conventional methods consists in the low a priori information requirements and the absence of operations with matrices.

1. INTRODUCTION

Stochastic approximation methods are widely used in optimization, estimation, and pattern recognition. The modifications of these methods proposed in [1] and [2] have the highest possible asymptotic convergence rate (various extensions and applications of the former can be found in [3] and [4]). However such optimal methods require a very large amount of a priori information; thus in stochastic optimization problems we must know the matrix $\nabla^2 f(x^*)$ (x^* being the sought point of minimum of the function $f(x)$).

In this paper we suggest another technique of construction of optimal methods. It is related to the idea of averaging the paths specified by a simple (and certainly nonoptimal) algorithm of stochastic approximation. It is found that such a combined algorithm does not require any a priori information, it is very simple, and yet it has an asymptotically optimal convergence rate. Let us note that the very idea of averaging in stochastic approximation methods is not new; it has been put forward in [5-15] in diverse forms and for different purposes. However the hopes that averaging will speed up the convergence (expressed as early as in [5]) were not fulfilled (see [6, 9]). In our method the speeding-up effect is achieved by virtue of the fact that the main process is "slower" than the optimal process.

2. LINEAR CASE

Let us consider a linear problem for which the results are simplest. It is required to solve the system

$$Ax=b, \tag{1}$$

where $b \in R^N$, $x \in R^N$, and A is an $(N \times N)$-dimensional matrix. It is assumed that at a point x_k we measure the variable $y_k = Ax_k - b + \xi_k$, where $Ax_k - b$ is the mismatch, and ξ_k is random noise. Let the following assumptions be satisfied.

1. The matrix A is Hurwitzian, i.e., $\mathrm{Re}\,\lambda_j > 0$ for all the eigenvalues of A.

2. The noises ξ_k are uncorrelated, with $M\xi_k = 0$, $M\xi_k\xi_k^T = S > 0$ (the notation $S > 0$ signifies that the matrix S is positive definite, and M denotes the mean).

We shall consider the following iterative method of finding the point $x^* = A^{-1}b$:

$$\begin{aligned} x_{k+1}&=x_k-\gamma_k y_k, \quad y_k=Ax_k-b+\xi_k, \\ \hat{x}_{k+1}&=\hat{x}_k+1/(k+1)\,(x_k-\hat{x}_k), \quad k=0, 1, \ldots \end{aligned} \tag{2}$$

Here the process for x_k is an ordinary stochastic approximation process, and $\hat{x}_k=\sum_{i=0}^{k-1} x_k/k$, i.e., the points $\hat{x}_k$ are the averaged values of the iterations x_j, $j = 0, 1, \ldots, k-1$. With regard to the initial values x_0 and the step lengths γ_k we shall assume the following.

Institute of Control Problems, Moscow. Translated from Avtomatika i Telemekhanika, No. 7, pp. 98-107, July, 1990. Original article submitted February 6, 1989.

0005-1179/90/5107-0937$12.50

3. The point x_0 is either deterministic and arbitrary, or random with $\mathbf{M}|x_0 - x^*|^2 < \infty$ (here and below we denote by $|x|$ the Euclidean norm in R^N).

4. Either

$$\gamma_k \equiv \gamma, \quad 0 < \gamma < \bar{\gamma} = \min_j 2 \operatorname{Re} \lambda_j |\lambda_j|^{-2}, \tag{3}$$

or

$$\gamma_k / \gamma_{k+1} = 1 + o(\gamma_k), \quad \gamma_k \to 0, \quad \sum_{k=0}^{\infty} \gamma_k = \infty. \tag{4}$$

The condition (4) is satisfied, for example, by the sequences $\gamma_k = \gamma_k^{-\alpha}$, $0 < \alpha < 1$, but not by $\gamma_k = \gamma k^{-1}$. In other words, the γ_k's must decrease more slowly than $1/k$.

THEOREM 1. Let the conditions 1-4 be satisfied. For the algorithm 2 we then have

$$V_k \triangleq \mathbf{M}(\hat{x}_k - x^*)(\hat{x}_k - x^*)^T = k^{-1} A^{-1} S (A^{-1})^T + o(k^{-1}). \tag{5}$$

All the theorems are proved in the Appendix.

It has been proved in [16] that for linear algorithms of the form

$$x_{k+1} = x_k - \Gamma_k y_k, \quad y_k = A x_k - b + \xi_k$$

we have for any choice of the matrices Γ_k a lower bound $U_k \triangleq M(x_k - x^*)(x_k - x^*)^T \geqslant (U_0^{-1} + k A^T S^{-1} A)^{-1}$, i.e., $U_k \geq k^{-1} A^{-1} S (A^{-1})^T + o(k^{-1})$, this asymptotic rate being achieved for an optimal algorithm with $\Gamma_k = k^{-1} A^{-1}$. Thus the method (2) yields the same convergence rate as an optimal linear algorithm. In contrast to the latter, it does not presuppose that the matrix A is known, and in general it does not use any operations with matrices.

Let us also note that the estimate (5) is the same for any procedures of selection of γ_k that satisfy the conditions (3) or (4). If we consider instead of (2) a more general method

$$x_{k+1} = x_k - \gamma_k \Gamma y_k, \quad \hat{x}_{k+1} = \hat{x}_k + (x_k - \hat{x}_k)/(k+1),$$

where the matrix Γ is not singular and $-\Gamma A$ is Hurwitzian, then it is easy to see that its convergence rate is also expressed by (5), i.e., it is also independent of Γ.

If we take $\gamma_k = \gamma k^{-1}$ in the method (2) (as is usually the case in methods with averaging), then the convergence rate of the process decreases and the bound (5) is not reached. This was noted as early as in [6] and [9], i.e., the use of averaging (in such a form) cannot speed up the convergence.

3. STOCHASTIC OPTIMIZATION

Now let us consider nonlinear problems and algorithms. We shall seek the minimum of a smooth function $f(x)$, $x \in R^N$, and measure at some point x_k the value of the gradient $y_k = \nabla f(x_k) + \xi_k$ that contains random noise ξ_k. The stochastic approximation method goes over into the gradient method

$$x_{k+1} = x_k - \gamma_k y_k, \quad y_k = \nabla f(x_k) + \xi_k,$$

and (as is shown in [2]) the asymptotically optimal method contains a nonlinear gradient transformation

$$x_{k+1} = x_k - \gamma_k B \varphi_*(y_k), \quad B = \nabla^2 f(x^*)^{-1}, \quad \varphi_*(y) = -J(p)^{-1} \nabla \ln p(y), \tag{6}$$

where $p(y)$ is the distribution density of the noise ξ_k; $J(p) = \int \frac{\nabla p \nabla p^T}{p} dy$ is Fisher's information matrix; x^* is a point of minimum of $f(x)$. This algorithm requires that we know $\nabla^2 f(x^*)$; there exist realizable versions of this algorithm in which this matrix can be estimated [17], but they are fairly cumbersome.

Let us use a counterpart of method (2):

$$x_{k+1}=x_k-\gamma_k\varphi(y_k), \quad y_k=\nabla f(x_k)+\xi_k,$$
$$\hat{x}_{k+1}=\hat{x}_k+(x_k-\hat{x}_k)/(k+1), \quad k=0, 1, \ldots \tag{7}$$

We shall make the following assumptions.

1. The function f(x) is twice differentiable, $||\nabla^2 f(x)|| \le L$, $\ell > 0$. Then it has a unique point of minimum x^*.

2. The noises ξ_k are uncorrelated and equally distributed with a density p, and there exists a J(p), $0 < J(p) < \infty$.

3. Let us denote $\psi(a) \triangleq \mathbf{M}\phi(a+\xi))$ (i.e., $\psi(a)=\int \varphi(a+y)p(y)dy$); then $\psi(0) = 0$, $(\psi(a), a) \ge \alpha|a|^2$, $\alpha > 0$ for any $a \in R^N$. Similarly, $\psi_2(a) \triangleq \mathbf{M}\varphi(a+\xi)\varphi(a+\xi)^T$, $\psi_2(0)>0$, $\psi_4(a) \triangleq \mathbf{M}|\varphi(a+\xi)|^4 \le c(1+|a|^4)$ (in general, c stands here for different constants), $||\psi_2(a) - \psi_2(0)|| \le c|a|^2$. Moreover, $\psi(a)$ is differentiable at the origin and $|\psi(a) - \psi'(0)a| \le c|a|^2$.

4. $\gamma_k/\gamma_{k+1}=1+o(\gamma_k)$, $\gamma_k=o(k^{-1/3})$, $\sum_{k=0}^{\infty}\gamma_k=\infty$.

These conditions are satisfied, for example, for $\gamma_k = \gamma_k^{-\alpha}$, $2/3 < \alpha < 1$. However, they are more rigorous than the conditions (4), where the γ_k's could tend to zero as slowly as desired.

THEOREM 2. Under the above assumptions we have

$$V_k \triangleq \mathbf{M}(\hat{x}_k-x^*)(\hat{x}_k-x^*)^T=k^{-1}BDB+o(k^{-1}),$$
$$B=\nabla^2 f(x^*)^{-1}, \quad D=\psi'(0)^{-1}\psi_2(0)(\psi'(0)^{-1})^T. \tag{8}$$

If we take

$$\varphi_*(y)=-J(p)^{-1}\nabla \ln p(y) \tag{9}$$

and this ϕ_* satisfies condition 3, then D = J(p).

For the algorithm (7), (9) we thus have $V_k = k^{-1}\nabla^2 f(x^*)^{-1}J(p) \times \nabla^2 f(x^*)^{-1} + o(k^{-1})$; but this asymptotic convergence rate corresponds to the optimal algorithm (6). Moreover (see [18]), this rate cannot be exceeded for any procedure of finding the point x^* that is suitable for an approprirate class of functions f. Thus the simple algorithm (7), (9) has the highest possible asymptotic convergence rate. Let us also note that the condition $\gamma_k = o(k^{-2/3})$ can apparently be relaxed to $\gamma_k = o(k^{-1/2})$ (see the proof of the theorem).

4. PARAMETER ESTIMATION

Let us consider a linear regression model

$$y_k=x_k^T\theta^*+\xi_k, \tag{10}$$

where the $y_k \in R^1$ are outputs, the $x_k \in R^N$ are inputs, the $\theta^* \in R^N$ are unknown parameters, and the ξ_k are noises. The gradient method of parameter estimation has the form

$$\theta_{k+1}=\theta_k-\gamma_k x_k\varepsilon_k, \quad \varepsilon_k=x_k^T\theta_k-y_k,$$

whereas according to the recursive method of least squares we have

$$\theta_{k+1}=\theta_k-\Gamma_k x_k\varepsilon_k, \quad \Gamma_{k+1}=\Gamma_k-\frac{\Gamma_k x_k x_k^T\Gamma_k}{1+x_k^T\Gamma_k x_k}. \tag{11}$$

The estimate (11) is the best in the class of linear estimates (and if the ξ_k's are normally distributed, then it is also the best in any class of estimates). However the method (11) has a number of disadvantages from a computational point of view (such as the need to store matrices Γ_k which might be ill-conditioned, and therefore errors could accumulate, etc.).

Let us construct a method with averaging:

$$\theta_{k+1}=\theta_k-\gamma_k x_k\varepsilon_k, \quad \varepsilon_k=x_k^T\theta_k-y_k,$$
$$\hat{\theta}_{k+1}=\hat{\theta}_k+(\theta_k-\hat{\theta}_k)/(k+1), \quad k=0,1,\ldots \tag{12}$$

Now let us formulate the assumptions of the problem (they are similar to those introduced in [1]).

1. The inputs x_k are independent and equally distributed, with $\mathbf{M}x_kx_k^T = B > 0$, $\mathbf{M}|x_k|^4 < \infty$.

2. The noises ξ_k are uncorrelated and independent of x_j, with $\mathbf{M}\xi_k = 0$, $\mathbf{M}\xi_k^2 = \sigma^2$.

3. $\gamma_k/\gamma_{k+1}=1+o(\gamma_k)$, $\gamma_k\to 0$, $\sum_{k=0}^{\infty}\gamma_k=\infty$.

THEOREM 3. According to the method (12) we have

$$V_k \triangleq \mathbf{M}(\hat{\theta}_k-\theta^*)(\hat{\theta}_k-\theta^*)^T=k^{-1}\sigma^2B^{-1}+o(k^{-1}). \tag{13}$$

The same asymptotic convergence rate is characteristic for the method of least squares (11), and also for the optimal linear method [1]

$$\theta_{k+1}=\theta_k-k^{-1}B^{-1}x_k\varepsilon_k. \tag{14}$$

On the other hand the method (12) is simpler than (11) and (14). Let us note that Theorem 1 cannot be used for proving (13), i.e., if we introduce $2f(\theta) = \mathbf{M}(y - x^T\theta)^2 = \sigma^2 + (\theta - \theta^*)^TB(\theta - \theta^*)$, then $\nabla f(\theta) = B(\theta - \theta^*)$, and $x_k\varepsilon_k$ can be regarded as a realization of $\nabla f(\theta_k)$: $x_k\varepsilon_k = \nabla f(\theta_k) + \eta_k$, however the noise $\eta_k = \xi_kx_k + (B - x_kx_k^T)(\theta_k - \theta^*)$ depends on the point θ_k (which is not assumed in Theorem 1). Another consequence of this is the fact that in general we cannot take $\gamma_k \equiv \gamma$ in (12) (we can merely prove that $V_k = k^{-1}C + o(k^{-1})$, but $C \geq \sigma^2B^{-1}$).

It is well known [1] that if the noise ξ_k has a non-Gaussian distribution, then an asymptotically optimal algorithm contains a nonlinear transformation of the mismatch:

$$\theta_{k+1}=\theta_k-k^{-1}B^{-1}x_k\varphi_*(\varepsilon_k), \quad \varphi_*(y)=-J(p)^{-1}(\ln p(y))', \tag{15}$$

where $p(y)$ is the density of the noise ξ_k; $J(p) = \int((p')^2/p)dy$. A similar method with averaging has the form

$$\theta_{k+1}=\theta_k-\gamma_kx_k\varphi_*(\varepsilon_k), \quad \hat{\theta}_{k+1}=\hat{\theta}_k+(\theta_k-\hat{\theta}_k)/(k+1)$$

and under natural assumptions it converges at the same rate as (15). However we shall not dwell on this here.

5. SINGULAR AND NONSMOOTH CASE

Again let us consider the stochastic optimization problem, but without assuming strong convexity and smoothness of $f(x)$. Let $f(x)$ be a convex function on R^N, let Q be a convex closed bounded set in R^N, and let $\partial f(x)$ be a subgradient [19] of f at the point x. We shall seek the minimum of $f(x)$ on Q, when we measure at a point x_k the quantity $y_k = \partial f(x_k) + \xi_k$. Let us consider a counterpart of the method (7):

$$x_{k+1}=P_Q(x_k-\gamma_ky_k), \quad y_k=\partial f(x_k)+\xi_k,$$
$$\hat{x}_{k+1}=\hat{x}_k+(x_k-\hat{x}_k)/(k+1), \quad k=0,1,\ldots \tag{16}$$

Here P_Q is the operator of projection on Q. With regard to the noises ξ_k we shall assume that they are uncorrelated, with $\mathbf{M}\xi_k = 0$, $\mathbf{M}|\xi_k|^2 \leq \sigma^2$.

THEOREM 4. Let $\gamma_k = \gamma k^{-1/2}$. Then

$$v_k \triangleq \mathbf{M}(f(\hat{x}_k)-f^*)=O(k^{-1/2}\ln k), \quad f^*=\min_{x\in Q} f(x). \tag{17}$$

A lower bound has been obtained in [12] for the quantities $\mathbf{M}(f(\tilde{x}_k) - f^*)$ for any methods of construction of the estimates $\tilde{x}_k$; it is equal to $O(k^{-1/2})$. Thus the method (16) yields an optimal asymptotic convergence rate to within a logarithmic factor. In contrast to the previous versions of the averaging method, the choice of the sequence γ_k is fixed in (16). This result shows that if we use a method that $\gamma_k = \gamma k^{-1/2}$, then it operates for both smooth and nonsmooth problems, as well as for strongly convex and for singular functions. In other words, our method is robust with respect to the a priori information, as well as the function to be minimized.

Let us also note that the method proposed in [12] is also a method with averaging that yields an optimal convergence rate for the singular case; it differs, however, from (16).

6. NONASYMPTOTIC BEHAVIOR OF METHOD

All the results obtained above are asymptotic. To what extent can we rely on these estimates for finite k? Let us examine this question for several particular cases.

Let us consider the linear problem (1) and the method (2) with $\gamma_k \equiv \gamma$. Let A be a symmetry matrix, with $\ell > 0$ and $L \geq \ell$ being its smallest and its largest eigenvalues; condition (3) goes over into $0 < \gamma < 2/L$. Let us consider the case $\xi_k \equiv 0$, i.e., absence of noise, and let $x^* = 0$. Then $x_k = (I - \gamma A)^k x_0$, $\hat{x}_k = k^{-1}\gamma^{-1}A^{-1}(I - (I - \gamma A)^k)x_0 \approx k^{-1}\gamma^{-1}A^{-1}x_0$. Thus, $|\hat{x}_k| \approx k^{-1}\gamma^{-1} \times ||A^{-1}||\,|x_0|$. If we take $\gamma = 1/L$ (when $x_k \to 0$ in the fastest way), then $|\hat{x}_k|^2 \leq k^{-2}\mu^2|x_0|^2 + o(k^{-2})$, $\mu = L/\ell$. The term which depends on the initial approximation behaves similarly also for $\xi_k \neq 0$. Therefore $V_k = k^{-1}A^{-1}SA^{-1} + F_k$, $||F_k|| \leq k^{-2}\mu^2\,|x_0|^2 + o(k^{-2})$. Hence the asymptotic expression (5) is true only for $k \gg c\mu^2|x_0|^2$, $c = ||A^{-1}SA^{-1}||^{-1}$. In other words, the smaller the error of the initial approximation and the less conditioned the matrix A, and the larger the measurement noise, the earlier will the asymptotic behavior begin.

Similar results can be obtained for other manners of selection of γ_k, and also for the method (12) of estimation of regression parameters. The situation is a bit more complicated in nonlinear problems. Thus according to the method (7), the asymptotic behavior begins the later, the more the function ϕ (or ψ) differs from a linear function, and f from a quadratic function.

For some simple problems, we can write down estimates of the method with averaging also for finite k (and not only in the asymptotic case). Let us consider the estimation of the parameter θ^* on the basis of direct measurements:

$$y_k = \theta^* + \xi_k, \quad k = 0, 1, \ldots$$

Our method (for $\gamma_k \equiv \gamma$) has the form

$$\theta_{k+1} = \theta_k - \gamma(\theta_k - y_k), \quad 0 < \gamma < 1, \quad \hat{\theta}_k = \frac{1}{k}\sum_{j=0}^{k-1}\theta_j. \tag{18}$$

Let $u_0 = \mathbf{M}(\theta_0 - \theta^*)^2$, $v_k = \mathbf{M}(\hat{\theta}_k - \theta^*)^2$. Direct estimates yield $v_k = \sigma^2 k^{-1} + ck^{-2} + o(k^{-2})$, $c = (1 - \gamma)\gamma^{-1}(u_0(\gamma^{-1} - 1) - (3 - \gamma)(2 - \gamma)^{-1}\sigma^2)$, so that for $\gamma = 1/2$ we have $v_k = \sigma^2 k^{-1} + (u_0 - 5\sigma^2/3)k^{-2} + o(k^{-2})$. If we compare this with the accuracy of the arithmetic mean $\bar{\theta}_k = k^{-1}\sum_{j=0}^{k-1} y_j$, $u_k = \mathbf{M}(\bar{\theta}_k - \theta^*)^2$, then $u_k = \sigma^2 k^{-1}$ and the estimate (18) can be better (for $u_0 < 5\sigma^2/3$), or worse than $\bar{\theta}_k$. This is quite natural, since it matters in (18) how the initial approximation θ_0 has been selected, and if it is close to θ^*, then the accuracy of (18) will be higher. In fact, this difference applies only to terms of the order of k^{-2}; the principal term $\sigma^2 k^{-1}$ of the accuracy estimate is the same in both methods.

7. PRACTICAL REALIZATION OF ALGORITHMS

In using these methods for solving practical problems, there arise a number of questions.

The first of them involves the method of selection of γ_k in the main procedure. As we noted above, asymptotic estimates of the convergence rate are robust with respect to this selection; however the behavior of the method in the initial iterations is strongly dependent on it. We used the following approach. At first we calculate with $\gamma_k \equiv \gamma$. This γ is selected in such a way that the process does not diverge (this manifests itself in a systematic increase of the variables $|y_k|$ in stochastic optimization, and of the variables $|\varepsilon_k|$ in the case of estimation). After the process arrives in a neighborhood of the solution (criteria for this are described below), we establish that $\gamma_k = \gamma(k - k_0)^{-1/2}$, $k > k_0$. Here k_0 is the time of arrival in a neighborhood of the solution. As we mentioned above, the procedure of selection of $\gamma_k = O(k^{-1/2})$ satisfies the conditions of Theorems 1 and 3 (see also the remark following Theorem 2), and it yields good results even for nonsmooth and singular problems (Theorem 4).

The second question involves the choice of the time of going over to the averaging procedure. The effect of a poor initial approximation can be quite strong, and the beginning of the asymptotic behavior will be delayed. Therefore it makes sense to start the averaging procedure only after the process enters a neighborhood of the solution, i.e., for $k > k_0$.

Thus we are realizing the following algorithm (for definiteness, for problem (1)):

$$x_{k+1}=x_k-\gamma y_k, \quad k\leqslant k_0,$$

$$x_{k+1}=x_k-\gamma_k y_k, \quad \hat{x}_{k+1}=\hat{x}_k+(x_k-\hat{x}_k)/(k-k_0), \quad k>k_0$$

(here we omit the method of selection of γ). It remains to indicate the method of selection of the time k_0.

For stochastic optimization problems ($y_k = \nabla f(x_k) + \xi_k$), a criterion that we have reached a neighborhood of x^* is the occurrence of sign changes in the variables $\zeta_k = (y_k, y_{k+1})$. More precisely, if during the last ten iterations we find no less than three negative values of ζ_k (i.e., $\sum_{j=k-9}^{k} \operatorname{sign} \zeta_j < 4$), then this k is taken as a k_0. For parameter estimation problems we used another criterion. The averaging process is carried out simultaneously with the process for θ_k. After every 20 iterations we check the quantity $|\theta_k - \bar{\theta}_k|$; if it is large compared to the mean value $|\theta_j - \theta_{j+1}|$ after these iterations, then the averaging process is "renewed" (i.e., we begin anew after this k); but if it is small, then we take $k = k_0$.

The results of calculations performed for problems of stochastic optimization and parameter estimation with the aid of such algorithms showed that the process enters fairly rapidly a neighborhood of the solution, and after that the behavior of the approximations can be adequately described by asymptotic formulas. To sum up, a scalar algorithm of estimation of regression operates just as efficiently as the matrix recursive method of least squares.

APPENDIX

1. We shall present several lemmas on the behavior of matrix sequences. Everywhere below, A is a Hurwitz matrix, $\gamma_k \geq 0$, $\gamma_k \to 0$, $\Sigma\gamma_k = \infty$, and all the matrices (A, S_k, R_k, etc.) have dimension $N \times N$.

LEMMA 1 [20]. If

$$S_{k+1}=S_k-\gamma_k A S_k-\gamma_k S_k A^T+F_k, \quad \|F_k\|=o(\gamma_k)+o(\gamma_k)\|S_k\|, \tag{A.1}$$

then $S_k \to 0$.

LEMMA 2 [20]. If

$$R_{k+1}=(I-\gamma_k A)R_k+G_k, \quad \|G_k\|=o(\gamma_k)+o(\gamma_k)\|R_k\|, \tag{A.2}$$

then $R_k \to 0$.

LEMMA 3. If

$$U_{k+1}=U_k-\gamma_k A U_k-\gamma_k U_k A^T+\gamma_k \nu_k S+H_k, \quad S>0, \quad \nu_k\to 0, \quad \nu_k/\nu_{k+1}=1+o(\gamma_k), \quad \|H_k\|=o(\gamma_k)\nu_k+o(\gamma_k)/\|U_k\|, \tag{A.3}$$

then $U_k = \nu_k U + o(\nu_k)$, where U is a solution of Lyapunov's equation $AU + UA^T = S$.

Proof. Let $S_k = U_k/\nu_k - U$. Then $U + S_{k+1} = (\nu_k/\nu_{k+1})(U + S_k - \gamma_k AU - \gamma_k UA^T - \gamma_k AS_k - \gamma_k S_k A^T + \gamma_k C) + H_k/\nu_{k+1}$, and we obtain (A.1) for S_k.

LEMMA 4. Let ν_k and H_k be the same as in Lemma 3, with

$$R_{k+1} = (I - \gamma_k A)R_k + \gamma_k \nu_k C + H_k, \tag{A.4}$$

then $R_k = \nu_k A^{-1} C + o(\nu_k)$.

The proof is the same as for Lemma 3.

LEMMA 5. Let

$$V_{k+1} = (1 - 2(k+1)^{-1})V_k + (k+1)^{-2}C + H_k, \quad \|H_k\| = o(k^{-2}) + o(k^{-2})\|V_k\|. \tag{A.5}$$

Then $V_k = k^{-1}C + o(k^{-1})$.

The proof involves a change of variables $S_k = k^{-1}V_k - C$ and the use of Lemma 1 with $A = -I$ and $\gamma_k = k^{-1}$.

2. Proof of Theorem 1. From (2) we obtain

$$x_{k+1} - x^* = (I - \gamma_k A)(x_k - x^*) - \gamma_k \xi_k, \tag{A.6}$$

for $U_k \triangleq M(x_k - x^*)(x_k - x^*)^T$ we therefore obtain

$$U_{k+1} = (I - \gamma_k A)U_k(I - \gamma_k A)^T + \gamma_k^2 S. \tag{A.7}$$

If (4) holds, then we can apply Lemma 3 with $\nu_k = \gamma_k$, and therefore $U_k = \gamma_k U + o(\gamma_k)$.

By multiplying (A.6) and $\hat{x}_{k+1} - x^* = (1 - \mu_k)(\hat{x}_k - x^*) + \mu_k(x_k - x^*)$, $\mu_k \triangleq (k+1)^{-1}$ and taking the average, we obtain for $R_k \triangleq M(x_k - x^*)(\hat{x}_k - x^*)^T$ the expression

$$R_{k+1} = (I - \gamma_k A)(1 - \mu_k)R_k + \mu_k(I - \gamma_k A)U_k. \tag{A.8}$$

In the case of (4), by using the asymptotic formula for U_k and the relation $\mu_k = o(\gamma_k)$, we obtain (A.4) with $\nu_k = \mu_k$ and $C = U$. Hence $R_k = \mu_k A^{-1}U + o(\mu_k)$, $R_k + R_k^T = \mu_k A^{-1}S(A^{-1})^T + o(\mu_k)$ (by virtue of the definition of U). For $V_k \triangleq M(\hat{x}_k - x^*)(\hat{x}_k - x^*)^T$ we have

$$V_{k+1} = (1 - \mu_k)^2 V_k + \mu_k^2 U_k + (1 - \mu_k)\mu_k(R_k + R_k^T) = (1 - 2\mu_k)V_k + \mu_k^2 A^{-1}S(A^{-1})^T + H_k, \quad \|H_k\| = o(\mu_k^2) + o(\mu_k^2)\|V_k\|. \tag{A.9}$$

From Lemma 5 we obtain (5).

Now let $\gamma_k \equiv \gamma$ and let (3) be satisfied. It then follows [20] from (A.7) that $U_k \to U$, where U is a solution of the equation $U = (I - \gamma A)U(I - \gamma A)^T + \gamma^2 S$, i.e., $AU + UA^T - \gamma AUA^T = \gamma S$. By virtue of (A.8) we have $R_{k+1} = (I - \gamma A)R_k + \mu_k(I - \gamma A)U + H_k$, $\|H_k\| = O(\mu_k)\|R_k\| + o(\mu_k)$. Hence $R_k = \mu_k \gamma^{-1} A^{-1}(I - \gamma A)U + o(\mu_k)$, $R_k + R_k^T = \mu_k(A^{-1}S(A^{-1})^T - U) + o(\mu_k)$, and by virtue of (A.9) we again arrive at (5).

3. The proof of Theorem 2 will be carried out in several stages.

a) Estimation of $f_k \triangleq M(f(x_k) - f^*)$, $f^* = \min f(x)$. By virtue of a Lipschitz condition on $\nabla f(x)$ we have [19]

$$f(x_{k+1}) \leq f(x_k) - \gamma_k(\nabla f(x_k), \varphi(y_k)) + L\gamma_k^2|\varphi(y_k)|^2/2. \tag{A.10}$$

Let us take the conditional mean with respect to ξ_k and use condition 3:

$$M(f(x_{k+1}) - f^* | x_k) \leq f(x_k) - f^* - \gamma_k \alpha |\nabla f(x_k)|^2 + c\gamma_k^2(1 + |\nabla f(x_k)|^2) \leq (f(x_k) - f^*)(1 - 2\alpha l\gamma_k + o(\gamma_k)) + c\gamma_k^2,$$

since $|\nabla f(x)|^2 \leq 2L(f(x) - f^*)$, $|\nabla f(x)|^2 \geq 2l(f(x) - f^*)$ [19].

Hence $f_{k+1} \leq (1 - c\gamma_k + o(\gamma_k))f_k + c\gamma_k^2$, and in conjunction with the conditions $\gamma_k/\gamma_{k+1} = 1 + o(\gamma_k)$, $\gamma_k \to 0$, $\Sigma\gamma_k = \infty$ this yields $f_k = O(\gamma_k)$.

b) Estimation of $g_k \triangleq M(f(x_k) - f^*)^2$. By squaring (A.10) and by taking the conditional mean with respect to ξ_k and using the estimates for ψ_2 and ψ_4, we obtain

$$M((f(x_{k+1})-f^*)^2|x_k) \leq (f(x_k)-f^*)^2(1-c\gamma_k+o(\gamma_k))+c(f(x_k)-f^*)\gamma_k^2+o(\gamma_k^3),$$

$$g_{k+1} \leq g_k(1-c\gamma_k+o(\gamma_k))+c\gamma_k^3+o(\gamma_k^3),$$

in the last inequality we used the estimate $f_k = O(\gamma_k)$. Hence $g_k = O(\gamma_k^2)$.

c) Estimation of $M|x_k - x^*|^j$, $j = 2, 3, 4$ and $v_k \triangleq M|\hat{x}_k - x^*|^2$. Since $|\hat{x} - x^*|^2 \leq (2/\ell)\cdot(f(x) - f^*)$, it follows that $M|x_k - x^*|^j = O(\gamma_k^{j/2})$, $j = 2, 3, 4$.

By virtue of the definition of $\hat{x}_{k+1}$ we have

$$v_{k+1} \leq (1-\mu_k)^2 v_k+\mu_k^2 u_k+2\mu_k(1-\mu_k)u_k^{1/2}v_k^{1/2},$$

$$u_k \triangleq M|x_k-x^*|^2, \quad \mu_k=1/(k+1).$$

Hence for $s_k = v_k^{1/2}$: $s_{k+1} \leqslant (1-\mu_k)s_k+\mu_k u_k^{1/2} = (1-\mu_k)s_k+\mu_k O(\gamma_k^{1/2})$, $s_k = O(\gamma_k^{1/2})$, $v_k = O(\gamma_k)$.

Thus $v_k = O(\gamma_k)$ (this is a crude estimate, and we shall see below that $v_k = O(\mu_k)$).

d) Estimation of $U_k \triangleq M(x_k - x^*)(x_k - x^*)^T$. We have

$$M((x_{k+1}-x^*)(x_{k+1}-x^*)^T|x_k) = (x_k-x^*)(x_k-x^*)^T-\gamma_k(x_k-x^*)\psi^T(\nabla f(x_k))-\gamma_k\psi(\nabla f(x_k))(x_k-x^*)^T+\gamma_k^2\psi_2(\nabla f(x_k)).$$

Since $\psi(a)=\psi'(0)a+r$, $|r|\leqslant c|a|^2$, $\psi_2(a)=\psi_2(0)+R$, $\|R\|\leqslant c|a|^2$, then $U_{k+1}=U_k-\gamma_k AU_k-\gamma_k U_k A^T+\gamma_k^2 S+F_k$, where $A=\psi'(0)\nabla^2 f(x^*)$, $S=\psi_2(0)$, $\|F_k\|=O(\gamma_k M|x_k-x^*|^3+\gamma_k^2 M|x_k-x^*|^2)=o(\gamma_k^2)$. It follows from Lemma 1 that $U_k = \gamma_k U + o(\gamma_k)$, where U is a solution of the equation $AU + UA^T = S$.

e) Estimation of $R_k \triangleq M(x_k - x^*)(\hat{x}_k - x^*)^T$. As above,

$$M((x_{k+1}-x^*)(\hat{x}_{k+1}-x^*)^T|x_k) = (x_k-x^*)(\hat{x}_k-x^*)(1-\mu_k)+$$
$$+(x_k-x^*)(x_k-x^*)^T\mu_k-\gamma_k\psi(\nabla f(x_k))((\hat{x}_k-x^*)(1+\mu_k)+(x_k-x^*)\mu_k)^T,$$
$$R_{k+1}=R_k(1-\mu_k)+\mu_k U_k-\gamma_k A R_k(1-\mu_k)+\gamma_k G,$$

$$\|G_k\|=O(M|r|\,|\hat{x}_k-x^*|+\mu_k M|r|\,|x_k-x^*|)=O((M|r|^2 M|\hat{x}_k-x^*|^2)^{1/2}+$$

$$+\mu_k M|x_k-x^*|^3)=O((M|x_k-x^*|^4 v_k)^{1/2}+\mu_k O(\gamma_k^{3/2}))=O(\gamma_k^{3/2}),$$

$$r \triangleq \psi(\nabla f(x_k))-\psi'(0)\nabla f(x_k), \quad |r|\leqslant c|x_k-x^*|^2.$$

Since $\mu_k = o(\gamma_k)$, it follows that $R_{k+1}=(I-\gamma_k A)R_k+\mu_k\gamma_k U+H_k$, $\|H_k\|=\mu_k o(\gamma_k)+o(\gamma_k^{5/2})$.

If $\gamma_k = o(k^{-2/3})$, i.e., $o(\gamma_k^{5/2}) = \mu_k o(\gamma_k)$, then we can use Lemma 4, and therefore $R_k = \mu_k A^{-1}U + o(\mu_k)$. Hence $R_k + R_k^T = \mu_k A^{-1}S(A^{-1})^T + o(\mu_k)$.

f) Estimation of $V_k \triangleq M(\hat{x}_k-x^*)(\hat{x}_k-x^*)^T$. For V_k we have (A.9), and Lemma 5 yields (8). The optimization of D with respect to ϕ has been carried out in [1], and an optimal ϕ is expressed by (9).

Let us note that if instead of the crude estimate $v_k = O(\gamma_k)$ we use in (d) the correct (not yet proved) estimate $v_k = O(\mu_k)$, then instead of the condition $\gamma_k = o(k^{-2/3})$, we obtain $\gamma_k = o(k^{-1/2})$. For realizing such a method of proof, we must analyze simultaneously (and not successively) the equations for R_{k+1} and V_{k+1}; this, however, is quite cumbersome.

4. Proof of Theorem 3. For $U_k \triangleq M(\theta_k - \theta^*)(\theta_k - \theta^*)^T$ we obtain

$$U_{k+1}=U_k-\gamma_k BU_k-\gamma_k U_k B+\gamma_k^2\sigma^2 B+F_k,$$
$$\|F_k\|=\gamma_k^2\|Mx_k x_k^T(\theta_k-\theta^*)(\theta_k-\theta^*)^T x_k x_k^T\|\leqslant\gamma_k^2\|U_k\|M|x_k|^4,$$

and by virtue of Lemma 3 we then have $U_k = \gamma_k\sigma^2 I/2 + o(\gamma_k)$. Next, $R_k \triangleq M(\theta_k - \theta^*)(\hat{\theta}_k - \theta^*)^T$ satisfies the relation $R_{k+1}=(I-\gamma_k B)R_k(1-\mu_k)+\mu_k(I-\gamma_k B)U_k$, $\mu_k=1/(k+1)$, whence $R_k = \mu_k\sigma^2 B^{-1}/2 + o(\mu_k)$.

Finally, $V_{k+1}=(1-\mu_k)^2V_k+(1-\mu_k)\mu_k(R_k+R_k^T)+\mu_k^2U_k$ and $V_k = \mu_k\sigma^2B^{-1} + o(\mu_k)$.

Proof of Theorem 4. Let x^* be a point of minimum of $f(x)$ on Q. By virtue of a property of the projection operator we then have

$$|x_{k+1}-x^*|^2\leq|x_k-\gamma_k y_k-x^*|^2,$$
$$\mathbf{M}(|x_{k+1}-x^*|^2|x_k)\leq|x_k-x^*|^2-2\gamma_k(f(x_k)-f^*)+c\gamma_k^2,$$

since $(\partial f(x_k), x_k - x^*) \geq f(x) - f^*$, $|\partial f(x)| \leq c$ for $x \in Q$. By denoting $u_k\triangleq\mathbf{M}|x_k-x^*|^2$, $f_k\triangleq\mathbf{M}(f(x^k)-f^*)$, we obtain $u_{k+1} \leq u_k - 2\gamma_k f_k + c\gamma_k^2$,

$$0\leq u_k\leq u_0-2\sum_{j=0}^{k-1}\gamma_jf_j+c\sum_{j=0}^{k-1}\gamma_j^2,\quad\sum_{j=0}^{k-1}\gamma_jf_j\leq\left(u_0+c\sum_{j=0}^{k-1}\gamma_j^2\right)\Big/2.$$

On the other hand $\hat{x}_k=k^{-1}\sum_{j=0}^{k-1}x_j$, and by virtue of the convexity of f we obtain $f(\hat{x}_k)\leq k^{-1}\sum_{j=0}^{k-1}f(x_j)$.

By taking the average of this inequality, we have for $v_k \triangleq \mathbf{M}(f(\hat{x}_k) - f^*)$ the expression

$$v_k\leq k^{-1}\sum_{j=0}^{k-1}f_j\leq\gamma^{-1}k^{-1}\sum_{j=0}^{k-1}\gamma_jf_j\leq\gamma^{-1}k^{-1/2}\left(u_0+c\sum_{j=0}^{k-1}\gamma_j^2\right)\Big/2=O(k^{-1}\ln k).$$

LITERATURE CITED

1. B. T. Polyak and Ya. Z. Tsypkin, "Adaptive estimation algorithms (convergence, optimality, stability)," Avtomat. Telemekh., No. 3, 71-84 (1979).
2. B. T. Polyak and Ya. Z. Tsypkin, "Optimal pseudogradient adaptation algorithms," Avtomat. Telemekh., No. 8, 74-84 (1980).
3. Ya. Z. Tsypkin, Foundations of Information Theory of Identification [in Russian], Nauka, Moscow (1984).
4. Ya. Z. Tsypkin and A. S. Poznyak, "Recursive optimization algorithms under uncertainty," Itogi Nauki Tekh., Tekh. Kibern., 16, VINITI, Moscow (1983), pp. 3-70.
5. K. S. Fu and Z. J. Nikolic, "On some reinforcement techniques and their relations to the stochastic approximation," IEEE Trans. Automat. Control, 11, No. 4, 361-363 (1966).
6. Ya. Z. Tsypkin, Foundations of the Theory of Learning Systems [in Russian], Nauka, Moscow (1970).
7. A. M. Gupal and L. G. Bazhenov, "A stochastic counterpart of the conjugate gradient method," Kibernetika, No. 1, 125-126 (1972).
8. S. V. Shil'man and V. E. Semenchuk, "Search for extremum of functions on the basis of an a priori Markov model of gradients," Izv. Akad. Nauk SSSR, Tekh. Kibern., No. 5, 25-31 (1974).
9. B. T. Polyak, "Comparison of convergence rate of single-step and multistep optimization algorithms in the presence of noise," Izv. Akad. Nauk SSSR, Tekh. Kibern., No. 1, 9-12 (1977).
10. I. P. Devyaterikov and A. I. Koshlan', "A method of solution of stochastic optimization problems with constraints," Avtomat. Telemekh., No. 5,99-104 (1988).
11. A. P. Korostelev, "On multistep stochastic optimization procedures," Avtomat. Telemekh., No. 5, 82-90 (1981).
12. A. S. Nemirovskii and D. B. Yudin, Complexity of Problems and Effectiveness of Optimization Methods [in Russian], Nauka, Moscow (1979).
13. I. P. Kornfel'd and Sh. E. Shteinberg, "Estimation of parameters of linear and nonlinear stochastic systems by the method of averaging of the deviations," Avtomat. Telemekh., No. 8 (1985).
14. A. Ruszyunski and W. Syski, "Stochastic approximation method with gradient averaging for unconstrained problems," IEEE Trans. Automat. Control, 28, No. 12, 1097-1105 (1983).
15. A. Berman, A. Feuer, and E. Wahnon, "Convergence analysis of smoothed stochastic gradient-type algorithm," Int. J. Syst. Sci., 18, No. 6, 1061-1078 (1987).
16. B. T. Polyak and Ya. Z. Tsypkin, "Potential possibilities of adaptation algorithms," in: Problems of Cybernetics, Adaptive Systems [in Russian], Science Council "Kibernetika," Academy of Sciences of the USSR, Moscow (1976), pp. 6-19.

17. M. L. Vil'k and S. V. Shil'man, "Convergence and optimality of realizable adaptation algorithms (the information-theoretical approach)," Probl. Peredachi Inf., No. 3, 80-88 (1985).
18. A. V. Nazin, "Information inequalities in gradient stochastic optimization problem and optimal realizable algorithms," Avtomat. Telemekh., No. 4, 127-138 (1989).
19. B. T. Polyak, Introduction to Optimization [in Russian], Nauka, Moscow (1983).
20. B. T. Polyak, "Convergence and convergence rate of iterative stochastic algorithms. II. The linear case," Avtomat. Telemekh., No. 4, 101-107 (1977).