SIAM J. CONTROL AND OPTIMIZATION
Vol. 30, No. 4, pp. 838-855, July 1992

006

ACCELERATION OF STOCHASTIC APPROXIMATION BY AVERAGING*

B. T. POLYAK† AND A. B. JUDITSKY‡

Abstract. A new recursive algorithm of stochastic approximation type with the averaging of trajectories is investigated. Convergence with probability one is proved for a variety of classical optimization and identification problems. It is also demonstrated for these problems that the proposed algorithm achieves the highest possible rate of convergence.

Key words. stochastic approximation, recursive estimation, stochastic optimization, optimal algorithms

AMS(MOS) subject classifications. 62L20, 93E10, 93E12

1. Introduction. The methods of stochastic approximation originate in the works [29], [12] and are currently well studied [5], [21], [14], [16], [40]. These methods are widely applied in problems of adaptation, identification, estimation, and stochastic optimization [36], [37], [1], [8]-[10], [18]. The optimal versions (algorithms having the highest rate of convergence) of these methods have been developed as well [34], [38], [6], [27], [28]. However, the application of these optimal methods requires a large amount of a priori information. For example, the matrix $\nabla^2 f(x^*)$ must be known in the problem of stochastic optimization (here x^* is the minimum point of $f(x)$).

The new way of developing optimal algorithms that does not require such information is based on the idea of averaging the trajectories. It was proposed independently by Polyak [24] and Ruppert [32]. In the latter work, the linear algorithm for the one-dimensional case was considered, and asymptotic normality of the procedure was proved. Polyak [24] studies multidimensional problems and nonlinear algorithms. He has demonstrated the mean square convergence for these methods. In this paper we consider the same framework as in [24], but we demonstrate the asymptotic normality of the estimates. The use of essentially new techniques in the proofs allows us to substantially weaken the conditions of the theorems. Moreover, we prove the statements on almost sure convergence.

The idea of using averaging to accelerate stochastic approximation algorithms appeared in the 1960s (see [36] and the references therein). Afterward, the result was that the hopes associated with this method could not be realized; see, for instance, [23], where it was proved that usual averaging methods are not optimal for linear problems. Nevertheless, the processes with averaging were proposed and studied in the vast variety of papers [11], [14], [20], [13], [33], [4]. The essential advancement [24], [32] was reached on the basis of the paradoxical idea: a slow algorithm having less than optimal convergence rate must be averaged.

The paper is organized as follows. In § 2 the linear case is discussed (i.e., linear equation and linear algorithm). The formulation of the result and proofs are the most clear for that problem. Then in § 3 the general problem of stochastic approximation is studied. The general result obtained is then applied to the unconstrained stochastic optimization problem and to the problem of estimation of linear regression parameters.

2. Linear problem. We want to find x^*, which solves the following equation:

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

Here $b \in R^N$, $x \in R^N$, and $A \in R^{N \times N}$. The sequence $(y_t)_{t \geq 1}$ is observed, where $y_t = Ax_{t-1} - b + \xi_t$. Here $Ax_{t-1} - b$ is a prediction residual and ξ_t is a random disturbance.

* Received by the editors July 30, 1990; accepted for publication (in revised form) June 24, 1991.

† Institute for Control Sciences, Profsoyuznaya 65, 117806, Moscow, Russia.

‡ Institut de Recherche en Informatique et Systemes Aleatoires (IRISA), 35042 Rennes, France.

To obtain the sequence of estimates $(\bar{x}_t)_{t\geq 1}$ of the solution x^* of (1), the following recursive algorithm will be used:

$$x_t = x_{t-1} - \gamma_t y_t, \qquad y_t = Ax_{t-1} - b + \xi_t,$$
$$\bar{x}_t = \frac{1}{t}\sum_{i=0}^{t-1} x_i. \tag{2}$$

x_0 is an arbitrary (nonrandom) point in R^N.

Let us suppose that the following assumptions hold.

Assumption 2.1. The matrix $-A$ is Hurwitz, i.e., $\operatorname{Re}\lambda_i(A) > 0$. (Here $\lambda_i(A)$ are the eigenvalues of the matrix A.)

Assumption 2.2. Coefficients $\gamma_t > 0$ satisfy either

$$\gamma_t \equiv \gamma, \qquad 0 < \gamma < 2\left(\min_i \operatorname{Re}\lambda_i(A)\right)^{-1} \tag{3}$$

or

$$\gamma_t \to 0, \qquad \frac{\gamma_t - \gamma_{t+1}}{\gamma_t} = o(\gamma_t). \tag{4}$$

Commentary. Condition (4) for $\gamma_t \to 0$ is the requirement on γ_t to decrease sufficiently slow. For example, the sequences $\gamma_t = \gamma t^{-\alpha}$ with $0 < \alpha < 1$ satisfy this restriction, but the sequence $\gamma_t = \gamma t^{-1}$ does not.

We assume a probability space with an increasing family of Borel fields $(\Omega, \mathfrak{F}, \mathfrak{F}_t, P)$. Suppose that ξ_t is a random variable, adopted to $\mathfrak{F}_t$.

Assumption 2.3. ξ_t is martingale-difference process, i.e., $E(\xi_t | \mathfrak{F}_{t-1}) = 0$;

$$\sup_t E(|\xi_t|^2 | \mathfrak{F}_{t-1}) < \infty \quad \text{a.s.}$$

(Here $|\cdot|$ is a Euclidean norm in R^N.)

Assumption 2.4. The following limit exists:

$$\lim_{C\to\infty} \overline{\lim_{t\to\infty}} E(|\xi_t|^2 I(|\xi_t| > C) | \mathfrak{F}_{t-1}) \overset{p}{=} 0.$$

(Here $I(A)$ is the characteristic function of a set A.)

Assumption 2.5. The following hold:

(a) $\lim_{t\to\infty} E(\xi_t \xi_t^T | \mathfrak{F}_{t-1}) \overset{p}{=} S > 0;$

(b) $\lim_{t\to\infty} E\xi_t \xi_t^T = S > 0.$

The notation $S > 0$ means that a matrix S is symmetrical and positive definite.

THEOREM 1. (a) *Let Assumptions* 2.1-2.4, 2.5(a) *be satisfied. Then*

$$\sqrt{t}(\bar{x}_t - x^*) \xrightarrow{D} N(0, V);$$

i.e., the distribution of normalized error $\sqrt{t}(\bar{x}_t - x^*)$ *is asymptotically normal with zero mean and the covariance matrix*

$$V = A^{-1} S (A^{-1})^T. \tag{5}$$

(b) *If Assumptions* 2.1-2.3, 2.5(b) *are satisfied, then*

$$\lim_{t\to\infty} Et(\bar{x}_t - x^*)(\bar{x}_t - x^*)^T = V.$$

(c) *Let Assumptions* 2.1-2.3 *be satisfied and let* $(\xi_t)_{t\geq 1}$ *be mutually independent and identically distributed. Then*

$$\bar{x}_t - x^* \to 0 \quad \text{a.s.}$$

The proofs of the theorems in this paper are in the Appendix.

Part (b) of the theorem was developed in [24], for the case of independent disturbances. Note that Assumption 2.2 on γ_t is significant. If the sequence $\gamma_t = \gamma t^{-1}$ is chosen for algorithm (2) (as is often done for methods using averaging), then the rate of convergence decreases [23].

It was shown in [26] that in the case of independent noises,

$$E(\hat{x}_t - x^*)(\hat{x}_t - x^*)^T \geq t^{-1}V + o(t^{-1})$$

for all linear recursive estimates $\hat{x}_t$. This asymptotic rate of convergence is achieved by the algorithm

$$x_t = x_{t-1} - t^{-1}A^{-1}y_t. \tag{6}$$

Method (2) provides the same rate of convergence as the optimal linear algorithm. The advantage of this method is that it does not require any knowledge about A and does not use matrix-valued γ_t. Several versions of algorithm (6) use an estimate of matrix A^{-1}, instead of the true value [22], [31]. The significant advantage of these procedures is that they require only the nonsingularity of A (compare to the rather restrictive Assumption 2.1).

3. Nonlinear problem. For nonlinear problems, consider the classical problem of stochastic approximation [21]. Let $R(x): R^N \to R^N$ be some unknown function. Observations y_t of the function are available at any point $x_{t-1} \in R^N$ and contain the following random disturbances ξ_t:

$$y_t = R(x_{t-1}) + \xi_t.$$

The problem is finding the solution x^* of the equation $R(x) = 0$ by using the observations y_t under the assumption that a unique solution exists.

To solve the problem, we use the following modification of algorithm (2):

$$\begin{aligned} &x_t = x_{t-1} - \gamma_t y_t, \qquad y_t = R(x_{t-1}) + \xi_t, \\ &\bar{x}_t = \frac{1}{t}\sum_{i=0}^{t-1} x_i, \qquad x_0 \in R^N. \end{aligned} \tag{7}$$

The first equation in (7) defines the standard stochastic approximation process. Let the following assumptions be fulfilled.

Assumption 3.1. There exists a function $V(x): R^N \to R^1$ such that for some $\lambda > 0$, $\alpha > 0$, $\varepsilon > 0$, $L > 0$, and all $x, y \in R^N$, the conditions $V(x) \geq \alpha|x|^2$, $|\nabla V(x) - \nabla V(y)| \leq L|x-y|$, $V(x^*) = 0$, $\nabla V(x-x^*)^T R(x) > 0$ for $x \neq x^*$ hold true. Moreover, $\nabla V(x-x^*)^T R(x) \geq \lambda V(x)$ for all $|x-x^*| \leq \varepsilon$.

Assumption 3.2. There exists a matrix $G \in R^{N\times N}$ and $K_1 < \infty$, $\varepsilon > 0$, $0 < \lambda \leq 1$ such that

$$|R(x) - G(x-x^*)| \leq K_1|x-x^*|^{1+\lambda}, \tag{8}$$

for all $|x-x^*| \leq \varepsilon$ and $\operatorname{Re}\lambda_i(G) > 0$, $i = \overline{1, N}$.

Assumption 3.3. $(\xi_t)_{t\geq 1}$ is a martingale-difference process, defined on a probability space $(\Omega, \mathfrak{F}, \mathfrak{F}_t, P)$, i.e., $E(\xi_t|\mathfrak{F}_{t-1})=0$ almost surely, and for some K_2

$$E(|\xi_t|^2|\mathfrak{F}_{t-1})+|R(x_{t-1})|^2\leq K_2(1+|x_{t-1}|^2) \quad \text{a.s.}$$

for all $t\geq 1$. The following decomposition takes place:

$$\xi_t=\xi_t(0)+\zeta_t(x_{t-1}), \tag{9}$$

where

$$E(\xi_t(0)|\mathfrak{F}_{t-1})=0 \quad \text{a.s.},$$

$$E(\xi_t(0)\xi_t^T(0)|\mathfrak{F}_{t-1})\xrightarrow{P} S \quad \text{as } t\to\infty;\ S>0,$$

$$\sup_t E(|\xi_t(0)|^2 I(|\xi_t(0)|>C)|\mathfrak{F}_{t-1})\xrightarrow{P} 0 \quad \text{as } C\to\infty;$$

and, for all t large enough,

$$E(|\zeta_t(x_{t-1})|^2|\mathfrak{F}_{t-1})\leq\delta(x_{t-1}) \quad \text{a.s.}$$

with $\delta(x)\to 0$ as $x\to 0$.

Assumption 3.4. It holds that $(\gamma_t-\gamma_{t+1})/\gamma_t=o(\gamma_t)$, $\gamma_t>0$ for all t;

$$\sum_{t=1}^{\infty}(1+\lambda)/\gamma_t^2 t^{-1/2}<\infty. \tag{10}$$

Commentary. Assumption 3.4, when compared to Assumption 3.2 of Theorem 1, not only restricts the rate of decrease of the coefficients γ_t from above, but it forces the coefficients to decrease not very slowly. Thus, if $\lambda=1$ in (8), then the sequence $\gamma_t=\gamma t^{-\alpha}$ satisfies this condition only for $\frac{1}{2}<\alpha<1$.

THEOREM 2. *If Assumptions* 3.1-3.4 *are satisfied, then* $\bar{x}_t\to x^*$ *almost surely, and*

$$\sqrt{t}(\bar{x}_t-x^*)\xrightarrow{D} N(0, V).$$

Here

$$V=G^{-1}S(G^{-1})^T. \tag{11}$$

A proposition similar to Theorem 2 has been stated in [32] for the one-dimensional case. It is well known (see, for example, [6], [21]) that the stochastic approximation algorithm obtains the maximum rate of convergence if it has the form

$$x_t=x_{t-1}-t^{-1}R'(x^*)^{-1}y_t.$$

For that method, $\sqrt{t}(x_t-x^*)\xrightarrow{D} N(0, V)$; here V is the same as in (11). The algorithm, however, could not be realized in that form (the matrix $R'(x^*)$ is unknown). There are some implementable versions of the optimal algorithm [38], [22], [7], [2], [31], but all of them utilize an estimate of the matrix $R'(x^*)$ and usually require additional observations. Algorithm (7) achieves the same optimal rate of convergence and has smaller computational complexity. We must repeat here the comment that already appears at the end of § 2: several procedures that use the estimate of the matrix $R'(x^*)$ [22], [31] do not require the assumption that $\operatorname{Re}\lambda_i(R'(x^*))>0$.

4. Stochastic optimization. Consider the problem of searching for the minimum x^* of the smooth function $f(x)$, $x\in R^N$. The values of the gradient $y_t=\nabla f(x_{t-1})+\xi_t$

containing random noise ξ_t are available at an arbitrary point x_{t-1} of R^N. To solve this problem, we use the following algorithm of the form (7):

$$x_t = x_{t-1} - \gamma_t \varphi(y_t), \qquad y_t = \nabla f(x_{t-1}) + \xi_t,$$
$$\bar{x}_t = \frac{1}{t}\sum_{i=0}^{t-1} x_i, \qquad x_0 \in R^N. \tag{12}$$

Let the following assumptions be fulfilled.

Assumption 4.1. Let $f(x)$ be a twice continuously differentiable function and $lI \leqq \nabla^2 f(x) \leqq LI$ for all x and some $l > 0$ and $L > 0$; here I is the identity matrix.

Assumption 4.2. $(\xi_t)_{t \geqq 1}$ is the sequence of mutually independent and identically distributed random variables $E\xi_1 = 0$.

Assumption 4.3. It holds that $|\varphi(x)| \leqq K_1(1+|x|)$.

Assumption 4.4. The function $\psi(x) = E\varphi(x+\xi_1)$ is defined and has a derivative at zero, $\psi(0) = 0$ and $x^T\psi(x) > 0$ for all $x \neq 0$. Moreover, there exist ε, $K_2 > 0$, $0 < \lambda \leqq 1$, such that

$$|\psi'(0)x - \psi(x)| \leqq K_2|x|^{1+\lambda}$$

for $|x| < \varepsilon$.

Assumption 4.5. The matrix function $\chi(x) = E\varphi(x+\xi_1)\varphi(x+\xi_1)^T$ is defined and is continuous at zero.

Assumption 4.6. The matrix $-G = -\psi'(0) = \nabla^2 f(x^*)$ is Hurwitz, i.e., $\operatorname{Re} \lambda_i(G) > 0$, $i = \overline{1, N}$.

Assumption 4.7. It holds that $(\gamma_t - \gamma_{t+1})/\gamma_t = o(\gamma_t)$, $\gamma_t > 0$ for all t;

$$\sum_{t=1}^{\infty} \gamma_t^{(1+\lambda)/2} t^{-1/2} < \infty.$$

The following theorem is a simple corollary of Theorem 2.

THEOREM 3. *Let Assumptions* 4.1–4.6 *be fulfilled. Then* $\bar{x}_t \to x^*$ *almost surely and* $\sqrt{t}(\bar{x}_t - x^*) \xrightarrow{D} N(0, V)$, *where* $V = G^{-1}\chi(0)(G^{-1})^T$.

The above conditions concerning the function, noises, and score function φ are close to those of [27]. We can find results on the mean square convergence of algorithm (12) under more restrictive conditions (than those of Theorem 3) in [24].

Suppose that disturbance ξ_1 possesses a continuously differentiable density p_ξ and that there exists a finite Fisher information matrix

$$J(p_\xi) = \int (\nabla p_\xi \nabla^T p_\xi) p_\xi^{-1}\, dy.$$

Let us choose the function φ according to the density p_ξ

$$\varphi(x) = -J^{-1}(p_\xi)\nabla \ln p_\xi(x). \tag{13}$$

In this case, we obtain

$$V = \nabla^2 f(x^*)^{-1} J(p_\xi)^{-1} \nabla^2 f(x^*)^{-1}.$$

Let us compare the proposed algorithm to the asymptotically optimal form of the stochastic optimization algorithm [27]

$$x_t = x_{t-1} - t^{-1} B\varphi(y_t), \tag{14}$$

where

$$B = \nabla^2 f(x^*)^{-1} \quad \text{and} \quad \varphi(y) = -J^{-1}(p_\xi)\nabla \ln p_\xi(y).$$

The value of the matrix $\nabla^2 f(x^*)$ is employed in algorithm (14). There exist some implementable versions of the algorithm [39], where an estimate of the matrix is used instead of the true value. Meanwhile, algorithms (12), (13) achieve the same rate of convergence as the optimal unimplementable algorithm (14). Therefore the algorithm with averaging is optimal in this situation in the same sense as in the other problems discussed. Note that this property of optimality corresponds not only to the class of stochastic approximation recursive algorithms, but to a wider class of methods of searching for a minimum point [19].

5. Estimation of regression parameters. Assume that the random variables $x_t \in R^N$, $y_t \in R^1$ are observed in successive instants $t = 1, 2, \cdots$, where

$$y_t = x_t^T \theta + \xi_t. \tag{15}$$

Here $\theta \in R^N$ is an unknown parameter and ξ_t is a random noise. We use the following two-step algorithm to produce the sequence of estimates $(\bar{\theta}_t)_{t \geq 1}$ of the parameter θ:

$$\begin{aligned} \theta_t &= \theta_{t-1} + \gamma_t \varphi(y_t - \theta_{t-1}^T x_t) x_t, \\ \bar{\theta}_t &= \frac{1}{t} \sum_{i=0}^{t-1} \theta_i, \qquad \theta_0 \in R^N. \end{aligned} \tag{16}$$

Suppose the following assumptions hold true.

Assumption 5.1. Let $(\xi_t)_{t \geq 1}$ be a sequence of mutually independent and identically distributed random variables $E\xi_1 = 0$, $E\xi_1^2 < \infty$.

Assumption 5.2. Let $(x_t)_{t \geq 1}$ be a sequence of mutually independent and identically distributed random variables $E|x_1|^4 < \infty$, $Ex_1 x_1^T = B$, $B > 0$. Sequences $(\xi_t)_{t \geq 1}$ and $(x_t)_{t \geq 1}$ are mutually independent.

Assumption 5.3. There exists K_1 such that $|\varphi(x)| \leqq K_1(1 + |x|)$ for all $x \in R^N$.

The functions $\psi(x) = E\varphi(x + \xi_1)$, $\chi(x) = E\varphi^2(x + \xi_1)$ are defined under Assumptions 5.1–5.3. Now we state restrictions on ψ, χ.

Assumption 5.4. It holds that $\psi(0) = 0$, $x\psi(x) > 0$ for all $x \neq 0$, $\psi(x)$ has a derivative at zero, and $\psi'(0) > 0$. Moreover, there exist $K_2 < \infty$ and $0 < \lambda \leqq 1$ such that

$$|\psi(x) - \psi'(0)x| \leqq K_2 |x|^{1+\lambda}.$$

Assumption 5.5. The function $\chi(x)$ is continuous at zero.

Assumption 5.6. It holds that $(\gamma_t - \gamma_{t+1})/\gamma_t = o(\gamma_t)$, $\gamma_t > 0$ for all t;

$$\sum_{t=1}^{\infty} \gamma_t^{(1+\lambda)/2} t^{-1/2} < \infty.$$

THEOREM 4. *Assume that Assumptions* 5.1–5.6 *hold. Then, for algorithm* (16), *the following properties hold true*: $\bar{\theta}_t \to \theta$ *almost surely and* $(\bar{\theta}_t - \theta)\sqrt{t} \xrightarrow{D} N(0, V)$, *where*

$$V = B^{-1} \frac{\chi(0)}{\psi'^2(0)}.$$

The problem of the mean square convergence of method (16) is discussed in [24]. Note that conditions of Theorem 4 are similar to the conditions of standard results for this problem [27]. If ξ_1 possesses a continuously differentiable density function p_ξ, then the optimal algorithm proposed in the latter paper has the following form:

$$\begin{aligned} \theta_t &= \theta_{t-1} + \Gamma_t \varphi(y_t - \theta_{t-1}^T x_t) x_t, \\ \Gamma &= B^{-1} t^{-1}, \qquad \varphi(x) = -J(p_\xi)^{-1} p_\xi'(x)/p_\xi(x). \end{aligned} \tag{17}$$

For method (17),

$$(\theta_t - \theta)\sqrt{t} \xrightarrow{D} N(0, V), \qquad V = J(p_\xi)^{-1}B^{-1}.$$

Since the matrix B is unknown, algorithm (17) is unimplementable. Nevertheless, it is possible to use, instead of B, its estimate

$$B_t = \left(t^{-1}\sum_{k=1}^{t} x_k x_k^T\right)^{-1}. \tag{18}$$

In particular, for linear algorithms (i.e., for Gaussian noises) methods (17), (18) coincide with the recursive MLS algorithm. It follows from Theorem 4 that if we choose

$$\varphi(x) = -J(p_\xi)^{-1}p'_\xi(x)/p_\xi(x)$$

for algorithm (16), then the rate of convergence is equal to $V = J(p_\xi)^{-1}B^{-1}$. So the asymptotical rates of convergence of (16) and (17) coincide.

Appendix. Proofs of the theorems consist of the sequence of propositions followed by their proofs. Everywhere in the following, we use the notation $\Delta_t = x_t - x^*$ for an error of the first equation of the algorithm, and $\bar{\Delta}_t = \bar{x}_t - x^*$ for an estimation error. Nonrandom constants that are unimportant will be denoted by the symbols K and α. All relations between random variables are supposed to be true almost surely (unless declared otherwise).

The two matrix lemmas below will be useful in later developments.

Let $(X_j^t)_{t \geqq j}$, $(\bar{X}_j^t)_{t \geqq j}$ be the sequences of matrices, $\bar{X}_j^t$, $X_j^t \in R^{N \times N}$, determined by the following recursive relations:

$$X_j^{t+1} = X_j^t - \gamma_t A X_j^t, \qquad X_j^j = I,$$

$$\bar{X}_j^t = \gamma_j \sum_{i=j}^{t-1} X_j^i. \tag{A1}$$

and $\phi_j^t = A^{-1} - \bar{X}_j^t$.

LEMMA 1. *Let the following hold*:

(i) *Assumption* 2.2 *of Theorem* 1 *holds*;

(ii) $\operatorname{Re}\lambda_i(A) > 0$, $i = \overline{1, N}$.

Then there is constant $K < \infty$ *such that for all* j *and* $t \geqq j$

$$\|\phi_j^t\| \leqq K, \tag{A2}$$

$$\lim_{t\to\infty} \frac{1}{t}\sum_{j=0}^{t-1} \phi_j^t = 0. \tag{A3}$$

Proof of Lemma 1.

Part 1. Proposition of the lemma is true with $\gamma_t \equiv \gamma$.

Proof. We obtain from (A1) that

$$X_j^t = (I - \gamma A)^{t-j} \qquad X_j^j = I,$$

$$\bar{X}_j^t = \gamma(I + (I - \gamma A) + \cdots + (I - \gamma A)^{t-j}) = A^{-1} - (I - \gamma A)^{t-j+1}A^{-1}.$$

The eigenvalues of the matrix $I - \gamma A$ are $\lambda_i(I - \gamma A) = I - \gamma\lambda_i(A)$ and $|\lambda_i(I - \gamma A)| < 1$. So

$$\lim_{t\to\infty} (I - \gamma A)^t = 0;$$

hence (A2) holds, and

$$\frac{1}{t}\sum_{j=0}^{t-1}\phi_j^t=\frac{1}{t}\sum_{j=0}^{t-1}(I-\gamma A)^{t-j+1}A^{-1}=\frac{1}{t}\sum_{k=2}^{t+1}(I-\gamma A)^k A^{-1}\to 0 \quad \text{as } t\to\infty.$$

Part 2. It holds that $t\gamma_t\to\infty$.

Proof. Let us define $\alpha_t=1/t\gamma_t$. Then

$$\alpha_{t+1}=(t+1)^{-1}\gamma_{t+1}^{-1}=\frac{1}{t+1}(\gamma_t^{-1}+o(1))=\alpha_t\frac{t}{t+1}+o(1)\frac{1}{t+1}$$

$$=\alpha_t\left(1-\frac{1}{t+1}\right)+\frac{o(1)}{t+1},$$

where $o(1)\to 0$ as $t\to\infty$. Since $\sum_{t=1}^{\infty}1/(t+1)=\infty$, we obtain that $\alpha_t\to 0$. □

Part 3. There are $\alpha>0$ and $K<\infty$ such that for all j and $t\geqq j$

$$\|X_j^t\|\leqq K\exp\left(-\alpha\sum_{i=j}^{t-1}\gamma_i\right).$$

Proof. From assumption (ii) of the lemma and from the Lyapunov theorem, we have that there exists the solution $V=V^T>0$ of the Lyapunov equation $A^TV+VA=I$.

Define $L=\max\lambda_i(V)$, $l=\min\lambda_i(V)$, $U_t=(X_j^t)^TVX_j^t$. Then

$$\begin{aligned}U_{t+1}&=(X_j^t)^T(I-\gamma_tA)^TV(I-\gamma_tA)X_j^t\\&=U_t-\gamma_t(X_j^t)^T(A^TV+VA)X_j^t+\gamma_t^2(X_j^t)^TA^TVAX_j^t.\end{aligned}\tag{A4}$$

Note that $(X_j^t)^TX_j^t\geqq(1/L)(X_j^t)^TVX_j^t$ and $(X_j^t)^TA^TVAX_j^t\leqq c(X_j^t)^TVX_j^t$, where $c=(\|A\|^2L)/l$. Then, for t sufficiently large and some $\lambda>0$, we get from (A4) that

$$U_{t+1}\leqq U_t\left(1-\frac{1}{L}\gamma_t+c\gamma_t^2\right)\leqq(1-\lambda\gamma_t)U_t\leqq e^{-\lambda\gamma_t}U_t.$$

Thus $U_t\leqq U_j\exp(-\lambda\sum_{i=j}^{t-1}\gamma_i)$. However,

$$\|U_t\|\geqq l\|X_j^t\|^2 \quad\text{and}\quad \|U_j\|\leqq L\|X_j^j\|^2=L;$$

so we obtain that

$$\|X_j^t\|\leqq\sqrt{\frac{L}{l}}\exp\left(-\frac{\lambda}{2}\sum_{i=j}^{t-1}\gamma_i\right).$$ □

Part 4. Equations (A2) and (A3) hold.

Proof. Summing the first equation of (A1) from j to t, we have that

$$X_j^t=X_j^j-A\sum_{i=j}^{t-1}\gamma_iX_j^i=I-A\sum_{i=j}^{t-1}\gamma_iX_j^i.\tag{A5}$$

Let us consider the sum in the right-hand side of (A5). Summing by parts, we get that

$$\sum_{i=j}^{t-1}\gamma_iX_j^i=\gamma_j\sum_{i=j}^{t-1}X_j^i+\sum_{i=j}^{t-1}(\gamma_i-\gamma_j)X_j^i=\bar{X}_j^t+S_j^t.$$

Let us estimate S_j^t. By using the result of Part 3, we obtain that

$$\begin{aligned}\|S_j^t\|&\leqq\left\|\sum_{i=1}^{t}\left[\sum_{k=j}^{i-1}(\gamma_{k+1}-\gamma_k)\right]X_j^i\right\|\leqq\sum_{i=j}^{t}\sum_{k=j}^{i-1}\gamma_ko(\gamma_k)\|X_j^i\|\\&\leqq o(\gamma_j)\sum_{i=j}^{t}m_j^i\,e^{-\lambda m_j^i}=o(\gamma_j)\sum_{i=j}^{t}\frac{m_j^i\,e^{-\lambda m_j^i}(m_j^i-m_j^{i-1})}{\gamma_i},\end{aligned}\tag{A6}$$

where $m_j^i = \sum_{k=j}^{i} \gamma_k$. From Part 2, it follows that $j\gamma_j \leqq Ki\gamma_i$ for i sufficiently large. Since $m_j^i = \sum_{k=j}^{i} \gamma_k \geqq \mu(\ln(i/j))$, we can estimate $1/\gamma_i$ as

$$\frac{1}{\gamma_i} \leqq K\frac{i}{j\gamma_j} \leqq \frac{K}{\gamma_j}\exp\left(\frac{m_j^i}{\mu}\right)$$

for μ arbitrarily large. Finally, we have from (A6) that

$$\|S_j^t\| \leqq \frac{Ko(\gamma_j)}{\gamma_j}\sum_{i=j}^{t} m_j^i\, e^{-\lambda m_j^i}(m_j^i - m_j^{i-1}) \equiv \frac{Ko(\gamma_j)}{\gamma_j}\int_0^{\infty} m\, e^{-\lambda m}\, dm\, \varepsilon_j$$

such that, for all $t \geqq j$,

$$\lim_{j\to\infty} \varepsilon_j = 0. \tag{A7}$$

Recall that since $\bar{X}_j^t + S_j^t = A^{-1} - A^{-1}X_j^t$ (see (A5)), we have, by the definition of ϕ_j^t, that

$$\phi_j^t = S_j^t + A^{-1}X_j^t.$$

From Part 3, however, we have that $\|X_j^t\| \leqq K$; thus we obtain (A2) from (A7).

Since $\|X_j^t\| \leqq K\ \exp(-\mu(\ln(t/j))) = K(j/t)^{\mu}$ for μ arbitrarily large, we get that

$$\frac{1}{t}\sum_{j=j_0}^{t-1} \|X_j^t\| \leqq K(\mu+1)^{-1}$$

for j_0 large enough. Note that, for some K,

$$\frac{1}{t}\sum_{j=0}^{t-1}\|X_j^t\| = \frac{1}{t}\sum_{j=0}^{j_0}\|X_j^t\| + \frac{1}{t}\sum_{j=j_0+1}^{t-1}\|X_j^t\| \leqq \frac{1}{t}\sum_{j=0}^{j_0} K + \frac{1}{t}\sum_{j=j_0+1}^{t-1}\|X_j^t\|.$$

For arbitrary $\varepsilon > 0$, we can choose μ and $j_0(\mu)$ such that

$$\frac{1}{t}\sum_{j=j_0+1}^{t-1}\|X_j^t\| \leqq K(\mu+1)^{-1} \leqq \varepsilon/2.$$

Then, choosing t sufficiently large, we get that $1/t\sum_{j=0}^{j_0} K \leqq \varepsilon/2$. Hence $1/t\sum_{j=0}^{t-1}\|X_j^t\| \leqq \varepsilon$. Moreover, from (A7), we have that

$$\lim_{t\to\infty}\frac{1}{t}\sum_{j=0}^{t-1}\|S_j^t\| = 0.$$

Hence, from the inequality above, we obtain (A3).

This completes the proof of Lemma 1. □

Note that we can get from (2) the following equation for the error $\bar{\Delta}_t$ of the algorithm:

$$\begin{aligned} &\Delta_t = \Delta_{t-1} - \gamma_t(A\Delta_{t-1} + \xi_t), \qquad \Delta_0 = x_0 - x^*, \\ &\bar{\Delta}_t = \frac{1}{t}\sum_{i=0}^{t-1}\Delta_i. \end{aligned} \tag{A8}$$

The next lemma states a convenient representation for the solution of system (A8).

LEMMA 2. *Let the statements of Lemma* 1 *be fulfilled. Then*

$$\sqrt{t}\,\bar{\Delta}_t = \frac{1}{\sqrt{t}\gamma_0}\alpha_t\Delta_0 + \frac{1}{\sqrt{t}}\sum_{j=1}^{t-1} A^{-1}\xi_j + \frac{1}{\sqrt{t}}\sum_{j=1}^{t-1} w_j^t\xi_j, \tag{A9}$$

where α_t, $w_j^t \in R^{N\times N}$ *are such that* $\|\alpha_t\| \leqq K$, $\|w_j^t\| \leqq K$ *for some* $K < \infty$, *and*

$$\frac{1}{t}\sum_{j=1}^{t-1}\|w_j^t\| \to 0 \quad \text{as } t\to\infty.$$

Proof of Lemma 2. From the first equation of (A8), we have that

$$\Delta_t = \prod_{j=1}^{t} (I - \gamma_j A)\Delta_0 + \sum_{j=1}^{t} \prod_{i=j+1}^{t} (I - \gamma_i A)\gamma_j \xi_j$$

(Set $\prod_{i=n+1}^{n} (I - \gamma_i A) = I$). Then we get for the error of the algorithm

$$\bar{\Delta}_t = \frac{1}{t} \sum_{j=0}^{t-1} \prod_{i=1}^{j} (I - \gamma_i A)\Delta_0 + \frac{1}{t} \sum_{k=1}^{t-1} \sum_{j=1}^{k} \left[\prod_{i=j+1}^{k} (I - \gamma_i A) \right] \gamma_j \xi_j$$

$$= \frac{1}{t} \sum_{j=0}^{t-1} \prod_{i=1}^{j} (I - \gamma_i A)\Delta_0 + \frac{1}{t} \sum_{j=1}^{t-1} \left[\sum_{k=j}^{t-1} \prod_{i=j+1}^{k} (I - \gamma_i A) \right] \gamma_j \xi_j.$$

Set

$$\alpha_j^t = \gamma_j \sum_{i=j}^{t-1} \prod_{k=j+1}^{i} (I - \gamma_k A),$$

$\alpha_t = \alpha_0^t$, and $w_j^t = \alpha_j^t - A^{-1}$. Then

$$\bar{\Delta}_t = \frac{1}{t\gamma_0} \alpha_t \Delta_0 + \frac{1}{t} \sum_{j=1}^{t-1} A^{-1}\xi_j + \frac{1}{t} \sum_{j=1}^{t-1} w_j^t \xi_j.$$

Note that from (A1) we obtain that $X_j^t = \prod_{i=j}^{t} (I - \gamma_i A)$ and $\bar{X}_j^t = \gamma_j \sum_{i=j}^{t} \prod_{k=j}^{i} (I - \gamma_k A)$.

Thus, from Lemma 1, we get that

$$\lim_{t\to\infty} \frac{1}{t} \sum_{j=0}^{t-1} \|w_j^t\| = 0, \qquad \|w_j^t\| \leqq K, \quad \|\alpha_t\| \leqq K. \qquad \square$$

Proof of Theorem 1.

Part 1. Proposition (a) of the theorem holds.

Proof. We obtain from (A9) that

$$\sqrt{t}\,\bar{\Delta}_t = I^{(1)} + I^{(2)} + I^{(3)}, \tag{A10}$$

where

$$I^{(1)} = \frac{1}{\sqrt{t}\,\gamma_0} \alpha_t \Delta_0,$$

$$I^{(2)} = \frac{1}{\sqrt{t}} \sum_{j=1}^{t-1} A^{-1}\xi_j,$$

$$I^{(3)} = \frac{1}{\sqrt{t}} \sum_{j=1}^{t-1} w_j^t \xi_j.$$

Note that, since $\|\alpha_t\| \leqq K$, $I^{(1)} \to 0$ in mean square. By Lemma 2 for $I^{(3)}$, we get that

$$E\left| \frac{1}{\sqrt{t}} \sum_{j=1}^{t-1} w_j^t \xi_j \right|^2 \leqq \frac{K}{t} \sum_{j=1}^{t-1} \|w_j^t\|^2 \leqq \frac{K}{t} \sum_{j=1}^{t-1} \|w_j^t\| \to 0 \quad \text{as } t \to \infty,$$

so $|I^{(3)}| \to 0$. We must demonstrate that the central limit theorem for martingales can be employed for $I^{(2)}$ (see, for example, Theorem 5.5.11 in [17]). We have, for a sufficiently large constant C, that

$$\overline{\lim_{t\to\infty}} \frac{1}{t} \sum_{j=1}^{t-1} E(|A^{-1}\xi_j|^2 I(|A^{-1}\xi_j| > C) \,|\, \mathfrak{F}_{j-1})$$

$$\leqq K^2 \overline{\lim_{t\to\infty}} \frac{1}{t} \sum_{j=1}^{t-1} E(|\xi_j|^2 I(|\xi_j| > CK^{-1}) \,|\, \mathfrak{F}_{j-1}) = \ell(C).$$

According to Assumption 2.4, $\ell(C) \xrightarrow{P} 0$ as $C \to \infty$. Thus the Lindeberg condition is fulfilled. By Assumption 2.5(a), we get that

$$\frac{1}{t}\sum_{j=1}^{t-1} A^{-1}E(\xi_j\xi_j^T|\mathfrak{F}_{j-1})(A^{-1})^T \xrightarrow{P} V.$$

Thus all the conditions of Theorem 5.5.11 [17] are fulfilled.

Part 2. Proposition (b) of the theorem holds.

Proof. We have from (A10) that

$$tE\bar{\Delta}_t\bar{\Delta}_t = EI^{(2)}(I^{(2)})^T + \varepsilon_t.$$

As in the proof of Part 1, we obtain from Lemma 2 that $\varepsilon_t \to 0$ as $t \to \infty$. Then

$$\lim_{t\to\infty} tE\bar{\Delta}_t\bar{\Delta}_t^T = \lim_{t\to\infty}\frac{1}{t}\sum_{j=1}^{t-1} A^{-1}E\xi_j\xi_j^T(A^{-1})^T$$

$$= \lim_{t\to\infty}\frac{1}{t}\sum_{j=1}^{t-1} A^{-1}S(A^{-1})^T = V.$$

Part 3. Proposition (c) of the theorem holds.

Proof. To simplify notation, we suppose that $\xi_t \in R^1$ (the proof for N-dimensional case is completely analogous). Let us again use decomposition (A10). We immediately get from Lemma 2 that $I^{(1)}/\sqrt{t} \to 0$. Next, by the law of large numbers (see, e.g., [35]), we get that $I^{(2)}/\sqrt{t} \to 0$. Let us evaluate the last term of (A10). Define the random sequence $(\bar{\xi}_t)_{t\geq 1}$ by the following equation:

$$\bar{\xi}_t = \begin{cases} \xi_t, & \text{if } |\xi_t| \leqq t^{3/4}, \\ 0, & \text{if } |\xi_t| > t^{3/4}. \end{cases}$$

By the Chebyshev inequality, we get that

$$P(|\xi_t| > t^{3/4}) \leqq E|\xi_t|^2 t^{-3/2} \leqq Kt^{-3/2}.$$

Then $\sum_{i=1}^{\infty} P(|\xi_i| > i^{3/4}) < \infty$ and $P\{|\xi_t| > t^{3/4} \text{ infinitely often}\} = 0$. Since w_j^t are uniformly bounded, it suffices to demonstrate that

$$\frac{1}{t}\sum_{j=1}^{t-1} w_j^t\bar{\xi}_j = t^{-1}S_t \to 0.$$

Note that $E\xi_t = 0$. Thus

$$|E\bar{\xi}_t| = E\xi_t I(|\xi_t| > t^{3/4}) \leqq (E\xi_t^2)^{1/2}(P(|\xi_t| > t^{3/4}))^{1/2}$$

$$\leqq Kt^{-3/4}.$$

Then we have that

$$S_t^4 = \left(\sum_{j=0}^{t-1} w_j^t\bar{\xi}_j\right)^4 = \sum_{j=0}^{t-1}(w_j^t)^4\bar{\xi}_j^4 + K\sum_{\substack{i,j \\ i<j}}^{t-1}(w_j^t)^2(w_i^t)^2\bar{\xi}_i^2\bar{\xi}_j^2$$

$$+ K\sum_{\substack{i\neq j \\ i\neq k \\ j<k}}^{t-1}(w_i^t)w_j^t w_k^t\bar{\xi}_i^2\bar{\xi}_j\bar{\xi}_k$$

$$+ K\sum_{i<j<k<1}^{t-1} w_i^t w_j^t w_k^t w_1^t\bar{\xi}_i\bar{\xi}_j\bar{\xi}_k\bar{\xi}_1$$

$$+ K\sum_{i\neq j}^{t-1} w_i^t(w_j^t)^3\bar{\xi}_i\bar{\xi}_j^3 = \sum_{i=1}^{5} I_t^{(i)}.$$

Note that

$$EI_t^{(1)} = KE \sum_{j=0}^{t-1} \bar{\xi}_j^4 \leqq Kt^{3/2} \sum_{j=0}^{t-1} E\bar{\xi}_j^2 \leqq Kt^{5/2}.$$

For $I_t^{(5)}$, we have that

$$|EI_t^{(5)}| \leqq \left| E \sum_{i \neq j}^{t-1} K\bar{\xi}_i \bar{\xi}_j^3 \right| \leqq 2 \left| \sum_{i>j}^{t-1} KE\bar{\xi}_j^3 E\bar{\xi}_i \right|$$

$$\leqq K \sum_{i>j}^{t-1} j^{3/4} E\bar{\xi}_j^2 i^{-3/4} \leqq K \sum_{i>j}^{t-1} E\bar{\xi}_j^2 \leqq Kt^2.$$

By the same arguments, we get that

$$|EI_t^{(2)}| \leqq Kt^2, \quad |EI_t^{(3)}| \leqq Kt^{3/2}, \quad |EI_t^{(4)}| \leqq Kt.$$

Therefore we obtain that

$$t^{-4} ES_t^4 \leqq Kt^{-4}(t^{5/2} + t^2 + t^{3/2} + t) \leqq Kt^{-3/2}.$$

By the Chebyshev inequality, we get that

$$\sum_{t=1}^{\infty} P(|t^{-1}S_t| > \delta) \leqq \sum_{t=1}^{\infty} (t\delta)^{-4} ES_t^4 \leqq K \sum_{t=1}^{\infty} t^{-3/2} < \infty.$$

Hence $t^{-1}S_t \to 0$. □

Proof of Theorem 2. Let Δ_t be the error of the first equation of (7). Define the function $\bar{R}(x): R^N \to R^N$ by the equation $\bar{R}(x) = R(x - x^*)$.

Part 1. It holds that $V(\Delta_t) \to V(\omega)$, where $V(\omega)$ is bounded.

Proof. The increment of the function $V_t = V(\Delta_t)$ on one step of algorithm (7) is given by

$$V_t \leqq V_{t-1} - \gamma_t \nabla V_{t-1}^T \bar{R}(\Delta_{t-1}) - \gamma_t \nabla V_{t-1}^T \xi_{t-1}(\nabla_{t-1})$$

$$+ \frac{L}{2} \gamma_t^2 |\bar{R}(\Delta_{t-1}) + \xi_t(\Delta_{t-1})|^2$$

compare with [25, p. 55]. Taking the expectation, conditioned to $\mathfrak{F}_{t-1}$, by Assumptions 3.2 and 3.3 for some suitable K, we obtain that

$$\begin{aligned} E(V_t | \mathfrak{F}_{t-1}) &\leqq V_{t-1} - \gamma_t \nabla V_{t-1}^T \bar{R}(\Delta_{t-1}) + \gamma_t^2 K(|\Delta_{t-1}|^2 + 1) \\ &\leqq V_{t-1}(1 + \gamma_t^2 K) + \gamma_t^2 K - \gamma_t \nabla V_{t-1}^T \bar{R}(\Delta_{t-1}). \end{aligned} \tag{A11}$$

From Assumption 3.6 and from Part 2 of Lemma 1, we have that $\sum_{t=1}^{\infty} \gamma_t = \infty$. It can be simply recognized from Assumption 3.6 that $\sum_{t=1}^{\infty} \gamma_t^2 < \infty$; so we obtain by the Robbins–Siegmund theorem [30], that $V_t \to V(\omega)$.

Since $V \geqq \alpha |\Delta|^2$ for some $\alpha > 0$, we have from Part 1 of the proof that $P(\sup_t |\Delta_t| < \infty) = 1$. Thus, for every $\varepsilon > 0$, there exists some $R < \infty$ such that

$$P\left(\sup_t |\Delta_t| \leqq R \right) \geqq 1 - \varepsilon. \tag{A12}$$

Define the stopping time $\tau_R = \inf\{t \geqq 1: |\Delta_t| > R\}$.

Part 2. It holds that $E|\Delta_t|^2 I(\tau_R > t) \leqq K\gamma_t$.

Proof. On $\{\tau_R > t\}$ we have from (A11) that

$$\begin{aligned} E(V_t I(\tau_R > t) | \mathfrak{F}_{t-1}) &\leqq E(V_t I(\tau_R > t-1) | \mathfrak{F}_{t-1}) \\ &\leqq [V_{t-1}(1 + \gamma_t^2 K) - \alpha \gamma_t V_{t-1}] I(\tau_R > t-1) + \gamma_t^2 K \end{aligned} \tag{A13}$$

for some K, $\alpha > 0$. Taking the expectation, we obtain from (A13) that

$$EV_t I(\tau_R > t) \leqq EV_{t-1} I(\tau_R > t-1)(1 - \gamma_t \alpha + K\gamma_t^2) + K\gamma_t^2.$$

Finally, by Lemma 2.1.26 [3], we obtain that

$$EV_t I(\tau_R > t) \leqq K\gamma_t. \tag{A14}$$

Note that almost sure convergence of the algorithm follows from Parts 1 and 2. Let us define the process $\bar{\Delta}_t^1$ by the following equations:

$$\Delta_t^1 = \Delta_{t-1}^1 - \gamma_t G \Delta_{t-1}^1 + \gamma_t \xi_t, \qquad \Delta_1^0 = \Delta_0,$$

$$\bar{\Delta}_t^1 = \frac{1}{t} \sum_{i=0}^{t-1} \Delta_i^1.$$

Let us demonstrate, that for the process $\bar{\Delta}_t^1$, all the properties to be proved follow from Theorem 1.

Part 3. It holds that

$$\lim_{C \to \infty} \overline{\lim_{t \to \infty}} E(|\xi_t|^2 I(|\xi_t| > C) | \mathfrak{F}_{t-1}) \overset{P}{=} 0.$$

Proof. By decomposition (9), we have that

$$I(|\xi_t| > C) \leqq I\left(|\zeta_t(\Delta_{t-1})| > \frac{C}{2}\right) + I\left(|\xi_t(0)| > \frac{C}{2}\right);$$

so

$$\begin{aligned} &E(|\xi_t|^2 I(|\xi_t| > C) | \mathfrak{F}_{t-1}) \\ &\quad \leqq 2E\left(|\zeta_t(\Delta_{t-1})|^2 I\left(|\zeta_t(\Delta_{t-1})| > \frac{C}{2}\right) \middle| \mathfrak{F}_{t-1}\right) \\ &\qquad + 2E\left(|\xi_t(0)|^2 I\left(|\xi_t(0)| > \frac{C}{2}\right) \middle| \mathfrak{F}_{t-1}\right) \\ &\quad \leqq 2\delta(\Delta_{t-1}) + E\left(|\xi_t(0)|^2 I\left(|\xi_t(0)| > \frac{C}{2}\right) \middle| \mathfrak{F}_{t-1}\right) = I_1 + I_2. \end{aligned}$$

Then $I_2 \to 0$ as $t \to \infty$ and $C \to \infty$ by Assumption 2.3; $I_1 \to 0$, since Δ_t converges to zero.

Therefore all the conditions of proposition (a) of Theorem 1 hold for the process $\bar{\Delta}_t^1$.

We demonstrate the proximity of the processes $\bar{\Delta}_t^1$ and $\bar{\Delta}_t$. Set $\delta_t = \bar{\Delta}_t^1 - \bar{\Delta}_t$; then for δ_t we obtain the equation (compare with (A9))

$$\begin{aligned} \sqrt{t}\, \delta_t &= \frac{1}{\sqrt{t}\, \gamma_0} \alpha_t \Delta_0 + \frac{1}{\sqrt{t}} \sum_{j=1}^{t-1} (G^{-1} + w_j^t)(\bar{R}(\Delta_j) - G\Delta_j) \\ &= I_t^{(1)} + I_t^{(2)}. \end{aligned}$$

Part 4. It holds that $\delta_t \sqrt{t} \to 0$ as $t \to \infty$.

Proof. From Lemma 2 we immediately get that $I_t^{(1)} \to 0$ as $t \to \infty$. Next, due to Assumption 2.2 and Lemma 2, we get that

$$\begin{aligned} I_t^{(2)} &\leqq \sum_{i=0}^{\infty} \frac{1}{i^{1/2}} |(G^{-1} + w_j^t)(\bar{R}(\Delta_i) - G\Delta_i)| \\ &\leqq K \sum_{i=0}^{\infty} \frac{1}{i^{1/2}} |\bar{R}(\Delta_i) - G\Delta_i| \\ &\leqq K \sum_{i=0}^{\infty} \frac{|\Delta_i|^{1+\lambda}}{i^{1/2}}. \end{aligned}$$

Thus we obtain from (A14) and Assumption 2.4 that

$$\sum_{i=0}^{\infty} \frac{E(|\Delta_i|^{1+\lambda} I(\tau_R > t))}{i^{1/2}} \leqq \sum_{i=0}^{\infty} \frac{K\gamma_i^{(1+\lambda)/2}}{i^{1/2}} < \infty;$$

so

$$\sum_{i=0}^{\infty} \frac{|\Delta_i|^{1+\lambda} I(\tau_R > t)}{i^{1/2}} < \infty.$$

Since

$$\left\{ \sum_{i=0}^{\infty} \frac{|\Delta_i|^{1+\lambda}}{i^{1/2}} < \infty \right\} \supseteq \left\{ \sup_i |\Delta_i| \leqq R \right\} \cap \left\{ \sum_{i=0}^{\infty} \frac{|\Delta_i|^{1+\lambda} I(\tau_R > t)}{i^{1/2}} < \infty \right\},$$

we have, by (A12), that

$$P\left\{ \sum_{i=0}^{\infty} \frac{|\Delta_i|^{1+\lambda}}{i^{1/2}} < \infty \right\} \geqq 1 - \varepsilon.$$

By the arbitrary choice of ε in (A12),

$$\sum_{i=0}^{\infty} \frac{|\Delta_i|^{1+\lambda}}{i^{1/2}} < \infty.$$

Hence, by the Kronecker lemma,

$$I_t^{(2)} = \frac{1}{\sqrt{t}} \sum_{j=1}^{t-1} \|G^{-1} + w_j'\| \, |\bar{R}(\Delta_j) - G\Delta_j| \to 0.$$

So the processes $\bar{\Delta}_t^1$ and $\bar{\Delta}_t$ are asymptotically equivalent.

This completes the proof of Theorem 2. □

Proof of Theorem 3. Let us check whether the assumptions of Theorem 2 are fulfilled. For that purpose, we transform the first equation of algorithm (12) in the following way:

$$\begin{aligned} x_t &= x_{t-1} - \gamma_t \psi(\nabla f(x_{t-1})) + \gamma_t(\psi(\nabla f(x_{t-1})) - \varphi(\nabla f(x_{t-1}) + \xi_t)) \\ &= x_{t-1} - \gamma_t R(x_{t-1}) + \gamma_t \xi_t (x_{t-1} - x^*); \end{aligned} \tag{A15}$$

here

$$\begin{aligned} \xi_t(x_{t-1} - x^*) &= \psi(\nabla f(x_{t-1})) - \varphi(\nabla f(x_{t-1}) + \xi_t), \\ R(x_{t-1}) &= \psi(\nabla f(x_{t-1})). \end{aligned} \tag{A16}$$

From Assumption 3.4 we have that $R^T(x)\nabla f(x) > 0$ for all $x \neq 0$. Let $f(x^*) = 0$ for the sake of simplicity. It follows from Assumptions 3.1 and 3.4 that there exist $\alpha > 0$, $\alpha' > 0$, $\varepsilon > 0$ such that

$$R^T(x)\nabla f(x) \geqq \alpha |\nabla f(x)|^2 \geqq \alpha' f(x)$$

for all $|x - x^*| \leqq \varepsilon$; hence $f(x)$ is a Lyapunov function for (A15), and all corresponding conditions of Theorem 2 are fulfilled.

So we obtain by Assumption 3.4 that

$$\begin{aligned}|R(x)-G(x-x^*)| &= |\psi(\nabla f(x))-\psi'(0)\nabla^2 f(x^*)(x-x^*)| \\ &\leqq |\psi(\nabla f(x))-\psi'(0)\nabla f(x)| \\ &\quad + |\psi'(0)\nabla f(x)-\psi'(0)\nabla^2 f(x^*)(x-x^*)| \\ &\leqq K|\nabla f(x)|^{1+\lambda} + \|\psi'(0)\|\,|\nabla f(x)-\nabla^2 f(x^*)(x-x^*)| \\ &\leqq K|x-x^*|^{1+\lambda} + K|x-x^*|^2 \leqq K|x-x^*|^{1+\lambda}.\end{aligned}$$

Hence Assumption 3.2 of Theorem 2 is fulfilled. Next, again using the notation Δ_t for the error of the first equation of (12), we note that $\xi_t(\Delta_{t-1})$ is a martingale-difference process and that

$$E|\xi_t(\Delta_{t-1})|^2 \leqq K(1+|\Delta_{t-1}|^2).$$

So, as concluded in the proof of the Theorem 2 (see Parts 1 and 2), $\Delta_t \to 0$ and

$$E|\Delta_t|^2 I(t \leqq \tau_R) \leqq K\gamma_t. \tag{A17}$$

Then, from (A17) by Assumptions 3.5 and 3.4, we have that

$$\begin{aligned}&|E(\xi_t(\Delta_{t-1})\xi_t(\Delta_{t-1})^T \,|\, \mathfrak{F}_{t-1}) - \chi(0)| \\ &\quad \leqq K|\chi(\Delta_{t-1})-\chi(0)| + K|\Delta_{t-1}|^2 \to 0.\end{aligned}$$

Next, we obtain that

$$\begin{aligned}&E(|\xi_t(\Delta_{t-1})|^2 I(|\xi_t(\Delta_{t-1})| > C)\,|\,\mathfrak{F}_{t-1}) \\ &\quad \leqq KE(|\xi_t|^2 I(|\xi_t(\Delta_{t-1})| > C)\,|\,\mathfrak{F}_{t-1}) + K|\Delta_{t-1}|^2.\end{aligned}$$

From the definition (A16) by Assumption 3.3, we get that

$$I(|\xi_t(\Delta_{t-1})| > C) \leqq I(|\Delta_{t-1}| > KC) + I(|\xi_t| > KC);$$

so

$$\begin{aligned}&E(|\xi_t(\Delta_{t-1})|^2 I(|\xi_t(\Delta_{t-1})| > C)\,|\,\mathfrak{F}_{t-1}) \\ &\quad \leqq o(1) + KE(|\xi_t|^2 I(|\xi_t| > KC)\,|\,\mathfrak{F}_{t-1}) \to 0 \quad \text{as } t \to \infty.\end{aligned}$$

(Here $o(1) \to 0$ as $t \to \infty$.) This means that Assumption 3.3 of Theorem 2 holds. Therefore all conditions of the proposition of Theorem 2 are fulfilled, and the matrix V is defined by the equation

$$V = G^{-1}\chi(0)G^{-1} = (\psi'(0)\nabla^2 f(0))^{-1}\chi(0)(\psi'(0)\nabla^2 f(0))^{-1}. \qquad \square$$

Proof of Theorem 4. Let $\Delta_t = \theta_t - \theta^*$ be an error of the first equation in (16). Denote by $\mathfrak{F}_t$ the minimum σ-algebra generated by disturbances and inputs until the time t: $\mathfrak{F}_t = \sigma(\xi_1, x_1, \cdots, \xi_t, x_t)$. Let $R(\Delta) = E\psi(\Delta^T x_1)x_1$. We obtain the following equation for Δ_t:

$$\begin{aligned}\Delta_t &= \Delta_{t-1} - \gamma_t E\psi(\Delta_{t-1}^T x_t)x_t \\ &\quad + \gamma_t(E\psi(\Delta_{t-1}^T x_t)x_t - \varphi(\Delta_{t-1}^T x_t + \xi_t)x_t) \\ &= \Delta_{t-1} - \gamma_t R(\Delta_{t-1}) + \gamma_t \varepsilon_t,\end{aligned} \tag{A18}$$

where $\varepsilon_t = R(\Delta_{t-1}) - \varphi(\Delta_{t-1}^T x_t + \xi_t)x_t$.

We check the fulfillment of the assumptions of Theorem 2 in that case. Assumption 5.4 implies that $\Delta^T R(\Delta) > 0$ for all $\Delta \neq 0$ and $R(\Delta) = 0$ for $\Delta = 0$; so $V(\Delta) = |\Delta|^2$ is the

Lyapunov function for (A18). Hence Assumption 3.1 of Theorem 2 is satisfied. Next, for some K,

$$|R(\Delta)-\psi'(0)B\Delta| \leqq KE|\Delta^T x_1|^{1+\lambda} \leqq K(\Delta^T E x_1 x_1^T \Delta)^{(1+\lambda)/2} \leqq K|\Delta|^{1+\lambda},$$

and, again, Assumption 3.2 of Theorem 2 holds. Since ε_t is a martingale-difference process and

$$E(|\varepsilon_t|^2 | \mathfrak{F}_{t-1}) \leqq K(|\Delta_{t-1}|^2+1)$$

for some K, we obtain that (see Parts 1 and 2 of the proof of Theorem 2) $|\Delta_t| \to 0$ and

$$E|\Delta_t|^2 I(t \leqq \tau_R) \leqq K\gamma_t.$$

By Assumption 5.5, we have that

$$|E(\varepsilon_t \varepsilon_t^T | \mathfrak{F}_{t-1}) - \chi(0)B| = |E(\chi(\Delta_{t-1}^T x_t) x_t x_t^T | \mathfrak{F}_{t-1}) - \chi(0)B| \xrightarrow{P} 0.$$

We must demonstrate that

$$\sup_t E(|\varepsilon_t|^2 I(|\varepsilon_t| > C) | \mathfrak{F}_{t-1}) \xrightarrow{P} 0 \quad \text{as } C \to \infty.$$

It follows from Assumption 5.3 that

$$\begin{aligned} I(|\varepsilon_t| > C) &\leqq I\left(|\varphi(\Delta_{t-1}^T x_t + \xi_t) x_t| > \frac{C}{2}\right) + I\left(|R(\Delta_{t-1})| > \frac{C}{2}\right) \\ &\leqq I\left(|\Delta_{t-1}||x_t|^2 > K\frac{C}{2}\right) + I\left(|\xi_t||x_t| > K\frac{C}{2}\right) + I\left(|R(\Delta_{t-1})| > \frac{C}{2}\right) \\ &\leqq I\left(|\Delta_{t-1}| > K\sqrt{\frac{C}{2}}\right) + 2I\left(|x_t|^2 > K\sqrt{\frac{C}{2}}\right) + I\left(|\xi_t| > \frac{C}{2}\right) \\ &= I_t^{(1)} + I_t^{(2)} + I_t^{(3)}. \end{aligned}$$

So we obtain that

$$\begin{aligned} &E(|\varepsilon_t|^2 I(|\varepsilon_t| > C | \mathfrak{F}_{t-1}) \\ &\quad \leqq KE((|\Delta_{t-1}|^2|x_t|^4 + |\xi_t|^2|x_t|^2)(I_t^{(1)} + I_t^{(2)} + I_t^{(3)}) | \mathfrak{F}_{t-1}) \\ &\quad \leqq KI_t^{(1)}|\Delta_{t-1}|^2 + KI_t^{(1)} \\ &\qquad + KE(|x_t|^2 I_t^{(2)}) + KE(|\xi_t|^2 I_t^{(3)}) = I_1 + I_2 + I_3 + I_4. \end{aligned}$$

Since $\Delta_t \to 0$, $I_1 \to 0$ and $I_2 \to 0$ as $C \to \infty$ and $t \to \infty$. From Assumption 5.1 we get that $I_4 \to 0$. By the Chebyshev inequality, we get that

$$\begin{aligned} I_3 &\leqq K(E|x_t|^4)^{1/2} P^{1/2}(|x_t|^2 > \sqrt{C}) \\ &\leqq K\frac{(E|x_t|^4)^{1/2}}{\sqrt{C}} \to 0 \quad \text{as } C \to \infty. \end{aligned}$$

So Assumption 3.3 of Theorem 2 is fulfilled. Therefore all the conditions of Theorem 2 are fulfilled under the assumptions of Theorem 4. Finally, we obtain for the matrix V that

$$V = (\psi'(0)B)^{-1}\chi(0)B(\psi'(0)B)^{-1}. \qquad \square$$

REFERENCES

[1] M. A. AIZERMAN, E. M. BRAVERMAN, AND L. I. ROZONOER, *The Method of Potential Functions in the Machine Learning Theory*, Nauka, Moscow, 1970. (In Russian.)

[2] A. M. BENDERSKIJ AND M. B. NEVEL'SON, *Multidimensional asymptotically optimal stochastic approximation procedure*, Problems Inform. Transmission, 17 (1982), pp. 423–434.

[3] A. BENVENISTE, M. METIVIER, AND P. PRIOURET, *Algorithmes Adaptatifs et Approximations Stochastiques (Théorie et Applications)*, Masson, Paris, 1987.
[4] A. BERMAN, A. FEUER, AND E. WAHNON, *Convergence analysis of smoothed stochastic gradient-type algorithm*, Internat. J. Systems Sci., 18 (1987), pp. 1061–1078.
[5] YU. M. ERMOL'EV, *Stochastic Programming Methods*, Nauka, Moscow, 1976. (In Russian.)
[6] V. FABIAN, *Asymptotically efficient stochastic approximation: The* RM *case*, Ann. Statist., 1 (1973), pp. 486–495.
[7] ———, *On asymptotically efficient recursive estimation*, Ann. Statist., 6 (1978), pp. 854–866.
[8] V. N. FOMIN, *Recursive Estimation and Adaptive Filtering*, Nauka, Moscow, 1984. (In Russian.)
[9] K. S. FU, *Sequential Methods in Pattern Recognition and Machine Learning*, Academic Press, New York, London, 1968.
[10] G. S. GOODWIN AND K. S. SIN, *Adaptive Filtering, Prediction and Control*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
[11] A. M. GUPAL AND L. G. BAJENOV, *Stochastic analog of the conjugate gradients method*, Cybernetics, N1 (1972), pp. 125–126. (In Russian.)
[12] E. KIEFER AND J. WOLFOVITZ, *Stochastic estimation of the maximum of a regression function*, Ann. Math. Statist., 23 (1952), pp. 462–466.
[13] I. P. KORNFEL'D AND SH. E. SHTEINBERG, *Estimation of the parameters of linear and nonlinear systems using the method of averaged residuals*, Automat. Remote Control, 46 (1986), pp. 966–974.
[14] A. P. KOROSTELEV, *Stochastic Recurrent Procedures*, Nauka, Moscow, 1981. (In Russian.)
[15] ———, *On multi-step stochastic optimization procedures*, Automat. Remote Control, 43 (1982), pp. 606–611.
[16] H. J. KUSHNER AND D. S. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer, New York, 1978.
[17] R. SH. LIPTZER AND A. N. SHIRYAEV, *Martingale Theory*, Nauka, Moscow, 1986. (In Russian.)
[18] L. LJUNG AND T. SÖDERSTRÖM, *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA, 1983.
[19] A. V. NAZIN, *Informational bounds for gradient stochastic optimization and optimal implemented algorithms*, Automat. Remote Control, 50 (1989), pp. 520–531.
[20] A. S. NEMIROVSKIJ AND D. B. YUDIN, *Complexity of Problems and Effectiveness of Optimization Methods*, Nauka, Moscow, 1980. (In Russian.)
[21] M. B. NEVEL'SON AND R. Z. KHAS'MINSKIJ, *Stochastic Approximation and Recursive Estimation*, American Mathematical Society, Providence, RI, 1973.
[22] ———, *Adaptive Robbins-Monro procedure*, Automat. Remote Control, 34 (1974), pp. 1594–1607.
[23] B. T. POLYAK, *Comparison of convergence rate for single-step and multi-step optimization algorithms in the presence of noise*, Engrg. Cybernet., 15 (1977), pp. 6–10.
[24] ———, *New stochastic approximation type procedures*, Avtomat. i Telemekh., N7 (1990), pp. 98–107. (In Russian.); translated in Automat. Remote Control, 51 (1991), to appear.
[25] ———, *Introduction to Optimization*, Optimization Software, New York, 1987.
[26] B. T. POLYAK AND YA. Z. TSYPKIN, *Attainable accuracy of adaptation algorithms*, in Problems of Cybernetics. Adaptive Systems, Nauka, Moscow, 1976, pp. 6–19. (In Russian.)
[27] ———, *Adaptive estimation algorithms (convergence, optimality, stability)*, Automat. Remote Control, 40 (1980), pp. 378–389.
[28] ———, *Optimal pseudogradient adaptation algorithms*, Automat. Remote Control, 41 (1981), pp. 1101–1110.
[29] H. ROBBINS AND S. MONROE, *A stochastic approximation method*, Ann. Math. Statist., 22 (1951), pp. 400–407.
[30] H. ROBBINS AND D. SIEGMUND, *A convergence theorem for nonnegative almost supermartingales and some applications*, in Optimizing Methods in Statistics, J. S. Rustaji, ed., Academic Press, New York, 1971, pp. 233–257.
[31] D. RUPPERT, *A Newton-Rafson version of the multivariate Robbins-Monro procedure*, Ann. Statist., 13 (1985), pp. 236–245.
[32] ———, *Efficient estimators from a slowly convergent Robbins-Monro process*, Tech. Report No. 781, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1988.
[33] A. RUSZYNSKI AND W. SYSKI, *Stochastic approximation method with gradient averaging for unconstrained problems*, IEEE Trans., AC-28 (1983), pp. 1097–1105.
[34] D. T. SAKRISON, *Stochastic approximation: A recursive method for solving regression problems*, in Advances in Communication Theory and Applications, 2, A. V. Balakrishnan, ed., Academic Press, New York, London, 1966, pp. 51–106.
[35] A. N. SHIRYAEV, *Probability*, Nauka, Moscow, 1980. (In Russian.)

[36] YA. Z. TSYPKIN, *Adaptation and Learning in Automatic Systems*, Academic Press, New York, London, 1971.
[37] ———, *Foundations of Informational Theory of Identification*, Nauka, Moscow, 1984. (In Russian.)
[38] J. H. VENTER, *An extension of the Robbins-Monro procedure*, Ann. Math. Statist., 38 (1967), pp. 181–190.
[39] M. L. VIL'K AND S. V. SHIL'MAN, *Convergence and optimality of implementable, adaptation algorithms (informational approach)*, Problems Inform. Transmission, 20 (1985), pp. 314–326.
[40] M. WASAN, *Stochastic Approximation*, Cambridge University Press, London, 1970.