Annals of Operations Research 39(1992)251–267

A STOCHASTIC QUASIGRADIENT ALGORITHM WITH VARIABLE METRIC

S.P. URYAS'EV

International Institute for Applied Systems Analysis, A-2361 Laxenburg, Austria

Abstract

This paper deals with a new variable metric algorithm for stochastic optimization problems. The essence of this is as follows: there exist two stochastic quasigradient algorithms working simultaneously – the first in the main space, the second with respect to the matrices that modify the space variables. Almost sure convergence of the algorithm is proved for the case of the convex (possibly *nonsmooth*) objective function.

1. Introduction

Stochastic quasigradient (or stochastic approximation) algorithms are used for the optimization of general stochastic systems with smooth, nonsmooth, and infinite-dimensional objective functions, for distributed systems and others (see, for example, [3,4,6–9,10–12,14,18]). The structure of the algorithms is simple, and each iteration requires only few additional calculations. However, the simplest variants of these algorithms have a significant drawback – a slow practical convergence rate for ill-conditioned functions. This fact is connected not only with randomness; for the deterministic case, the simple gradient algorithm is also inefficient for ill-conditioned functions. Variable metric algorithms are more complicated, but they have a considerably faster convergence rate. These algorithms are widely used for smooth deterministic optimization problems (see [2]). Several authors have generalized such algorithms for the stochastic case with a smooth objective function [1,5,8,15,16,20]. This paper presents the variable metric algorithm for stochastic programming problems with a *nonsmooth* objective function. To date, the author has not seen any strictly proved variable metric algorithm applicable to *nonsmooth* stochastic optimization problems in the literature. Without a strict proof of convergence, such algorithms were discussed in [17,18]. For nonsmooth deterministic objective functions, analogous algorithms (see [19]) showed a rather fast numerical rate of convergence for ill-conditioned problems. This stimulated the author to develop a variable metric method for stochastic optimization problems.

2. Basic idea of the algorithm

Here we consider the problem of minimizing a convex (possibly nonsmooth) function $f(x)$

$$f(x) \to \min_{x \in \mathbb{R}^n}, \tag{1}$$

where $\mathbb{R}^n$ is an n-dimensional Euclidean space. In the class of problems considered here, instead of exact values of gradients or generalized gradients of the function $f(x)$, vectors are known, which are statistical estimates of these quantities. (The exact values of the function and its gradients are very difficult to compute.) Such problems present themselves, for example, in the minimization of functions of the form

$$f(x) = E_\omega \varphi(x,\omega) = \int_{\omega \in \Omega} \varphi(x,\omega) P(\mathrm{d}\omega).$$

Here and below, we assume that all random values are given on the probability space $(\Omega, \mathcal{F}, P)$. Considering that, under the general assumption, the generalized differential of the convex function $f(x)$ is calculated by the formula (see [13])

$$\partial f(x) = \int_{\omega \in \Omega} \partial_x \varphi(x,\omega) P(\mathrm{d}\omega), \tag{2}$$

then $\partial_x \varphi(x, \omega)$ is a set of vectors which are the statistical estimates of gradients of the function $f(x)$. We call these estimates *stochastic quasigradients* [3]. To solve problem (1), the following algorithm is used:

$$x^{s+1} = x^s - \rho_s H^s \xi^s, \tag{3}$$

where ρ_s, $s = 0, 1, \ldots$, is a sequence of positive random scalar stepsizes; H^s, $s = 0, 1, \ldots$, is a sequence of $n \times n$ random square matrices; ξ^s, $s = 0, 1, \ldots$, is a sequence of stochastic quasigradients, i.e. the conditional mathematical expectation $E_s \xi^s$ is a generalized gradient

$$E_s \xi^s \stackrel{\text{def}}{=} E(\xi^s / x^s) \in \partial f(x^s),$$

where E_s is the conditional mathematical expectation with respect to the σ-field defiined by the random vector x^s. How can the matrix H^s be chosen? There exists the natural criterion function $\Phi_s(H)$:

$$\Phi_s(H) = E_s f(x^s - \rho_s H \xi^s), \tag{4}$$

which characterizes the quality of choice for matrix H at iterations s. The function $\Phi_s(H)$ is the mathematical expectation of the objective function f at the point x^{s+1}. The best matrix H at iteration s is a solution to the problem

$$\Phi_s(H) \to \min_{H \in \mathbb{R}^{n\times n}} . \tag{5}$$

Problem (5) is somewhat more complicated than problem (1). However, the optimal matrix H is not needed at each iteration; it is enough to find some updating rule. Let us differentiate the function $\Phi_s(H)$ at some point H_0^s (see formula (2)):

$$\partial_H \Phi_s(H_0^s) = E_s \partial_H f(x^s - \rho_s H_0^s \xi^s) = E_s\{-\rho_s y \xi^{s^{\mathrm{T}}}: y \in \partial_x f(x^s - \rho_s H_0^s \xi^s)\},$$

where $\xi^{s^{\mathrm{T}}}$ is the transposed vector ξ^s. We denote ξ_0^s as some stochastic quasigradient at the point $x_0^s \overset{\mathrm{def}}{=} x^s - \rho_s H_0^s \xi^s$, i.e.

$$E(\xi_0^s / x_0^s) = g(x_0^s) \in \partial f(x_0^s).$$

One can see that

$$E_s(-\rho_s \xi_0^s \xi^{s^{\mathrm{T}}}) = E_s(E(-\rho_s \xi_0^s \xi^{s^{\mathrm{T}}} / x_0^s)) = E_s(-\rho_s g(x_0^s) \xi^{s^{\mathrm{T}}}) \in \partial_H \Phi_s(H_0^s);$$

thus, $-\rho_s \xi_0^s \xi^{s^{\mathrm{T}}}$ is a stochastic quasigradient of the function $\Phi_s(H)$ at the point H_0^s. We consider that the matrix H_0^s is known from the previous iterations $s-1$. To modify matrix H_0^s, we use the stochastic quasigradient method (see [3]):

$$H_1^s = H_0^s - \lambda(-\rho_s \xi_0^s \xi^{s^{\mathrm{T}}}) = H_0^s + \lambda_0^s \xi_0^s \xi^{s^{\mathrm{T}}}, \quad \lambda_0^s = \lambda \rho_s .$$

Analogously, the next iteration can be done at the point H_1^s and so on. Let at s interations with respect to matrix H an amount $i(s) \geq 1$ iterations be made. Write this as follows:

$$H_{i+1}^s = H_i^s + \lambda_i^s \xi_i^s \xi^{s^{\mathrm{T}}}, \quad i = 0, \ldots, i(s); \tag{6}$$
$$H_0^s = H^{s-1} = H_{i(s-1)+1}^{s-1},$$

where ξ_i^s, $i = 0, \ldots, i(s)$, are stochastic quasigradients, i.e.

$$E(\xi_i^s / x_i^s) \in \partial f(x_i^s), \qquad x_i^s = x^s - \rho_s H_i^s \xi^s . \tag{7}$$

Thus, simultaneously we have two stochastic quasigradient algorithms – algorithm (3) in the main space and algorithm (6) with respect to the matrix. In formula (6), the matrix H is modified additively, but multiplicative variants of this algorithm can also be developed (see [19]).

3. Formal description of the algorithm and sufficient conditions for convergence

Define the optimal set X^* for problem (1) as follows:

$$X^* = \{x^* \in \mathbb{R}^n : f(x^*) = \min f(x)\}.$$

Algorithm (3), (6) can solve the optimization problem (1) without constraints. To simplify the convergence proof of the algorithm, we assume that some convex set $X \subset \mathbb{R}^n$ is known in advance such that $X^* \subset X$. This is not a serious restriction, since in practical situations such a set is usually known. This set could be very large. If $x^s \notin X$, then we assume that the approximation x^s is very far from the extremal set X^* and we restart the algorithm from the initial point x^0 with new initial algorithm parameters.

We also assume that the sequences $\{\varepsilon_s\}$, $s = 0, 1, \dots,$ and $\{\lambda_{sl}\}$, $s = 0, 1, \dots, l = 0, 1, \dots,$ are given before starting the algorithm. This predetermination is not very good from the practical point of view, but this can be relaxed later. Some adaptive formulae could also be written for these sequences, but we do not want to overload the convergence proof with them now. The positive value ε_s defines $i(s)$ in the algorithm; iterations with respect to the matrix are stopped if $\rho_s \sum_{l=0}^{i(s)-1} \lambda_{sl} \geq \varepsilon_s$. To avoid misunderstandings, we present here a full formal description of the algorithm.

ALGORITHM 1

Step 1 Initialization:

$s = 0$, $i = -1$, $x^0 = x_{init}$, $H_0^{-1} = I$ is the unit matrix; ξ^0 is a stochastic quasigradient at the point x^0.

Step 2 Set $H_0^s = H_{i+1}^{s-1}$.

Step 3 Set $i = 0$.

Step 4 Compute the point x_i^s:

$$x_i^s = x^s - \rho_s H_i^s \xi^s.$$

Step 5 Compute $H_{i+1}^s = H_i^s + \lambda_{si} \xi_i^s \xi^{s\,\mathrm{T}}$, where ξ_i^s is a stochastic quasigradient at the point x_i^s.

Step 6 If $i \geq 1$ and $\rho_s \sum_{l=0}^{i-1} \lambda_{sl} \geq \varepsilon_s$, then $i(s) = i$; go to step 8.

Step 7 Set $i = i + 1$ and return to step 4.

Step 8 If $x_{i(s)}^s \in X$, then $x^{s+1} = x_{i(s)}^s$, $\xi^{s+1} = \xi_{i(s)}^s$; otherwise $x^{s+1} = x^0$, $\xi^{s+1} = \xi^0$.

Step 9 Set $s = s + 1$ and return to step 2.

Let us define $d(x, X^*)$ as the distance between a point x and the set X^*:

$$d(x, X^*) = \min_{x^* \subset X^*} \|x - x^*\|.$$

To prove the convergence of algorithm 1, we shall use the following sufficient conditions (see [18]) for convergence of stochastic algorithms, which were used in [18] to prove convergence of different stochastic algorithms. These conditions are similar to the conditions in [11], but are more general. The sequence $\{x^s(\omega)\}$ described below in conditions D1–D5 could be an arbitrary sequence of random vectors and is not generated by algorithm 1.

D1 There exists a compact set $X \subset \mathbb{R}^n$ such that

$$\{x^s(\omega)\} \subset X \text{ a.s.}$$

D2 $W : X \to \mathbb{R}$ is a continuous function.

D3 If there exists an event $B \subset \Omega$ such that $P(B) > 0$ and for all $\omega \in B$ there exists a subsequence $\{x^{l_\kappa(\omega)}(\omega)\}$ convergent to a point $x'(\omega)$ with $d(x'(\omega), X^*) > 0$, then for any random value $\varepsilon(\omega) > 0$ a.s. there exists a subsequence $\{v_\kappa(\omega)\}$ such that

$$W(x^\tau) \le W(x'(\omega)) + \varepsilon(\omega) \quad \text{for } l_\kappa(\omega) \le \tau \le v_\kappa(\omega),$$

$$\underline{\lim}_{\kappa \to \infty} W(x^{v_\kappa(\omega)}(\omega)) \overset{\text{def}}{=} \overline{W}(\omega) < W(x'(\omega)).$$

D4 $(\overline{W}(\omega), W(x'(\omega))) \setminus W(X^*) \ne \varnothing$ for almost all $\omega \in B$, i.e. the open interval $(\overline{W}(\omega), W(x'(\omega)))$ does not belong to the set $W(X^*) \overset{\text{def}}{=} \{W(x^*) : x^* \in X^*\}$ for almost all $\omega \in B$.

D5 For almost all subsequences $\{x^{s_\kappa(\omega)}(\omega)\}$ such that $\lim_{\kappa\to\infty} x^{s_\kappa(\omega)}(\omega) = x^*(\omega)$, $x^*(\omega) \in X^*$, the condition

$$\max\left\{\left[W\left(x^{s_\kappa(\omega)+1}(\omega)\right) - W\left(x^{s_\kappa}(\omega)\right)\right], 0\right\} \to 0 \quad \text{for } \kappa \to \infty$$

is satisfied.

Next is the theorem above these sufficient conditions (see [18]).

THEOREM 1

Let the stochastic sequence $\{x^s(\omega)\}$ satisfy conditions D1–D5; then $x^s(\omega) \to X^*$ a.s., i.e. $d(x^s(\omega), X^*) \to 0$ a.s.

Theorem 1 is proved in [18]. Since this proof is not available in English, we briefly repeat it here.

Proof

Suppose that the statement of the theorem is not correct. This means that there exists an event $B \subset \Omega$ such that $P(B) > 0$ and for all $\omega \in B$ there exists a subsequence $\{x^{l_\kappa(\omega)}(\omega)\}$ with

$$d(x^{l_\kappa(\omega)}(\omega), X^*) \geq \delta > 0 \quad \text{for } \kappa = 0, 1, \ldots .$$

Further, we shall omit the argument ω. Since X is a compact set, then without loss of generality we can consider that $x^{l_\kappa} \to x' \notin X^*$. Condition D3 implies the existence of the subsequence $\{x^{\nu_n}\}$ such that $\underline{\lim}'_{n\to\infty} W(x^{\nu_n}) = \overline{W} < W(x')$. Since the interval $(\overline{W}, W(x'))$ does not belong to the set $W(X^*)$ (see D4), it is possible to choose numbers a, b such that $\overline{W} < a < b < W(x')$, $a \notin W(X^*)$. There exist subsequences $\{m_\kappa\}$, $\{q_\kappa\}$ such that $m_\kappa < q_\kappa$, $W(x^{q_\kappa}) \leq a$, $W(x^{m_\kappa}) \geq b$ for $\kappa = 0, 1, \ldots,$ and $a < W(x^s)$ for s satisfying inequalities $m_\kappa < s \leq q_\kappa$, $k = 0, 1, \ldots$ (see fig. 1). The set

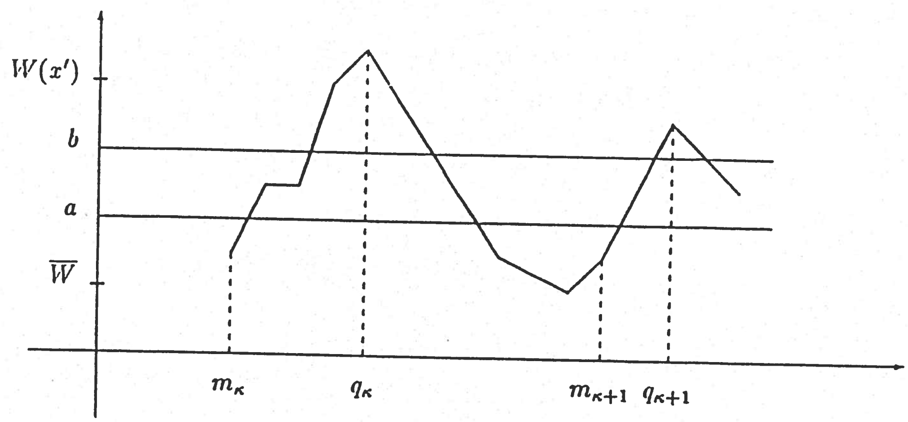

Fig. 1.

X is compact, and without loss of generality we can consider $\lim_{\kappa\to\infty} x^{m_\kappa} = \overline{x}$. Notice that $d(\overline{x}, X^*) > 0$. Otherwise (see D5),

$$0 \leq a - W(x^{m_\kappa}) \leq W(x^{m_\kappa+1}) - W(x^{m_\kappa})$$
$$= \max\left\{\left[W\left(x^{m_\kappa+1}\right) - W\left(x^{m_\kappa}\right)\right], 0\right\} \to 0 \quad \text{for } \kappa = 0, 1, \ldots .$$

Consequently, $W(\overline{x}) = a$ and $W(\overline{x}) = a \in W(X^*)$, but this contradicts the choice of a. Let us select $\varepsilon < b - a$. Since $d(\overline{x}, X^*) > 0$, condition D3 implies the existence of a subsequence ν_κ such that $W(x^\tau) \leq W(\overline{x}) + \varepsilon$ for τ satisfying inequality $m_\kappa \leq \tau \leq \nu_\kappa$, and

$$\underline{\lim}_{\kappa\to\infty} W(x^{\nu_\kappa}) < W(\overline{x}).$$

By the selection of m_κ and q_κ, without loss of generality we can consider that $m_\kappa < \nu_\kappa < q_\kappa$. Since $a \leq W(x^t)$ for $m_\kappa < t < q_\kappa$,

$$a \le \varliminf_{\kappa\to\infty} W(x^{\nu_\kappa}) < W(\bar{x}).$$

Also, $W(x^{m_\kappa}) \le a$ implies $W(\bar{x}) \le a$. This contradiction proves the theorem. □

4. Convergence of the algorithm

We now formulate the theorem on the convergence of algorithm 1.

THEOREM 2

Let $f: \mathbb{R}^n \to \mathbb{R}$ be a convex (possibly nonsmooth) function, X a compact convex set such that $X^* \subset X \subset \mathbb{R}^n$, and

$$\inf_{x \notin X, x^* \in X^*} \|x - x^*\| = C_1 > \min_{x^* \in X^*} \|x^0 - x^*\|; \tag{8}$$

let the sequences $\{\lambda_{sl}\}$ and $\{\varepsilon_s\}$ be given and let $\{\rho_s\}$ be a random sequence such that ρ_s depends upon the random vectors

$$\left(x^0, \xi^0; x_l^\tau, \xi_l^\tau, 0 \le \tau \le s-1,\ 0 \le l \le i(\tau)\right);$$

let the stochastic quasigradients and algorithm satisfy the conditions

$$\|\xi^0\| \le C_2 \quad \text{a.s.,} \tag{9}$$

$$\|\xi_i^s\| \le C_2 \quad \text{a.s.,} \quad i = 1, \ldots, i(s); \quad s = 0, 1, \ldots, \tag{10}$$

$$\varepsilon_s > 0, \qquad s = 0, 1, \ldots, \tag{11}$$

$$\sum_{s=0}^{\infty} \varepsilon_s^2 < \infty, \tag{12}$$

$$\sum_{s=0}^{\infty} \varepsilon_s = \infty, \tag{13}$$

$$\rho_s \|H_0^s \xi^s\| \varepsilon_s^{-1} \to 0 \quad \text{a.s.} \qquad \text{for} \quad s \to \infty, \tag{14}$$

$$\rho_s > 0 \quad \text{a.s.,} \qquad s = 0, 1, \ldots, \tag{15}$$

$$\sum_{s=0}^{\infty} \rho_s^2 \le \infty \quad \text{a.s.,} \tag{16}$$

$$\lambda_{sl} > 0, \qquad s = 0, 1, \ldots; \quad l = 0, 1, \ldots, \tag{17}$$

$$\sum_{l=0}^{\infty} \lambda_{sl} = \infty, \qquad s = 0, 1, \ldots, \tag{18}$$

$$\sum_{l=0}^{\infty} \lambda_{sl}^2 \le \Lambda = \text{const}, \quad s = 0, 1, \ldots. \tag{19}$$

Then almost surely all the accumulation points of the sequence $\{x^s\}$ generated by algorithm 1 belong to X^*.

Proof

We use sufficient conditions D1–D5 to prove the convergence of the algorithm. Define

$$W(x) = \min_{y \in X^*} \|x - y\|^2 = d^2(x, X^*).$$

Condition D1 is valid due to the construction of the algorithm and the compactness of the set X.
It is easy to see that the function $W(x)$ is continuous and consequently condition D2 holds.
Let us prove condition D3. Denote

$$\begin{gathered}\eta^s = \xi^s - g(x^s), \qquad \eta_i^s = \xi_i^s - g(x_i^s), \\ U_\varepsilon(x) = \{y \in \mathbb{R}^n : \|y - x\| \le \varepsilon\}, \\ f^* = \min_{x \in \mathbb{R}^n} f(x), \qquad C_3 = \max_{x, y \in X} \|x - y\|, \\ x_s^* = \arg\min_{y \in X^*} \|x^s - y\|, \; x_{si}^* = \arg\min_{y \in X^*} \|x_i^s - y\|.\end{gathered} \tag{20}$$

Let the probability of the event $B = \{\omega \in \Omega : \exists$ a subsequence $x^{l_\kappa(\omega)}(\omega)$ of the sequence $x^s(\omega)$ such that $x^{l_\kappa(\omega)}(\omega) \to x'(\omega) \notin X^*\}$ is greater than zero. We shall omit the latter for the simplicity of argument ω.

LEMMA 1

The following inequality is true:

$$\begin{aligned} W(x_{i(s)}^s) &\le W(x^s) + 2\rho_s C_3 \|H_0^s \xi^s\| + \rho_s^2 \|H_0^2 \xi^s\|^2 \\ &+ 2\rho_s \|\xi^s\|^2 \sum_{l=0}^{i(s)-1} \lambda_{sl}(f^* - f(x_l^s)) + 2\rho_s \|\xi^2\|^2 \sum_{l=0}^{i(s)-1} \lambda_{sl} \langle x_{sl}^* - x_l^s, \eta_l^s \rangle + \rho_s^2 C_2^6 \Lambda. \end{aligned} \tag{21}$$

Proof

Steps 4, 5 and 6 of the algorithm and conditions (9) and (10) of the theorem imply

$$
\begin{aligned}
W(x^s_{i(s)}) &= \| x^*_{s,i(s)} - x^s_{i(s)} \|^2 \\
&\le \| x^*_{s,i(s)-1} - x^s_{i(s)} \|^2 = \| x^*_{s,i(s)-1} - x^s + \rho_s H^s_{i(s)} \xi^s \|^2 \\
&= \| x^*_{s,i(s)-1} - x^s + \rho_s (H^s_{i(s)-1} + \lambda_{s,i(s)-1} \xi^s_{i(s)-1} \xi^{s^{\mathrm{T}}}) \xi^s \|^2 \\
&= \| x^*_{s,i(s)-1} - x^s + \rho_s H^s_{i(s)-1} \xi^s + \rho_s \lambda_{s,i(s)-1} \| \xi^s \|^2 \xi^s_{i(s)-1} \|^2 \\
&= \| x^*_{s,i(s)-1} - x^s_{i(s)-1} + \rho_s \lambda_{s,i(s)-1} \| \xi^s \|^2 \xi^s_{i(s)-1} \|^2 \\
&= \| x^*_{s,i(s)-1} - x^s_{i(s)-1} \|^2 + 2\rho_s \lambda_{s,i(s)-1} \| \xi^s \|^2 \langle x^*_{s,i(s)-1} - x^s_{i(s)-1}, \xi^s_{i(s)-1} \rangle \\
&\quad + \rho_s^2 \lambda^2_{s,i(s)-1} \| \xi^s \|^4 \| \xi^s_{i(s)-1} \|^2 \le W(x^s_{i(s)-1}) \\
&\quad + 2\rho_s \lambda_{s,i(s)-1} \| \xi^s \|^2 \langle x^*_{s,i(s)-1} - x^s_{i(s)-1}, \xi^s_{i(s)-1} \rangle + \rho_s^2 \lambda_{s,i(s)-1} C_2^6 .
\end{aligned}
$$

Applying this estimate the proper amount of times, we obtain

$$
W(x^s_{i(s)}) \le W(x^s_0) + 2\rho_s \| \xi^s \|^2 \sum_{l=0}^{i(s)-1} \lambda_{sl} \langle x^*_{sl} - x^s_l, \xi^s_l \rangle + \rho_s C_2^6 \sum_{l=0}^{i(s)-1} \lambda^2_{sl} . \tag{22}
$$

Estimate $W(x^s_0)$ is as follows:

$$
\begin{aligned}
W(x^s_0) &\le \| x^*_s - x^s_0 \|^2 = \| x^*_s - x^s + \rho_s H^s_0 \xi^s \|^2 \\
&= \| x^*_s - x^s \|^2 + 2\rho_s \langle x^*_s - x^s, H^s_0 \xi^s \rangle + \rho_s^2 \| H^s_0 \xi^s \|^2 \\
&\le W(x^s) + 2\rho_s \| x^*_s - x^s \| \, \| H^s_0 \xi^s \| + \rho_s^2 \| H^s_0 \xi^s \|^2 .
\end{aligned}
$$

Since the function $f(x)$ is convex, then, with designations (20), we obtain

$$
\langle x^*_{sl} - x^s_l, \xi^s_l \rangle = \langle x^*_{sl} - x^s_l, g(x^s_l) \rangle + \langle x^*_{sl} - x^s_l, \eta^s_l \rangle \le f^* - f(x^s_l) + \langle x^*_{sl} - x^s_l, \eta^s_l \rangle .
$$

Substituting the two previous estimates into estimate (22) and (19)–(20) yields

$$
\begin{aligned}
W(x^s_{i(s)}) &\le W(x^s) + 2\rho_s \| x^*_s - x^s \| \, \| H^s_0 \xi^s \| + \rho_s^2 \| H^s_0 \xi^s \|^2 \\
&\quad + 2\rho_s \| \xi^s \| \sum_{l=0}^{i(s)-1} \lambda_{sl} (f^* - f(x^s_l)) + 2\rho_s \| \xi^s \|^2 \sum_{l=0}^{i(s)-1} \lambda_{sl} \langle x^*_{sl} - x^s_l, \eta^s_l \rangle
\end{aligned}
$$

$$+\rho_s^2 C_2^6 \sum_{l=0}^{i(s)-1} \lambda_{sl}^2 \le W(x^s) + 2\rho_s C_3 \| H_0^s \xi^s \|$$

$$+\rho_s^2 \| H_0^s \xi^s \|^2 + 2\rho_s \| \xi^s \|^2 \sum_{l=0}^{i(s)-1} \lambda_{sl}(f^* - f(x_l^s))$$

$$+2\rho_s \| \xi^s \|^2 \sum_{l=0}^{i(s)-1} \lambda_{sl} \langle x_{sl}^* - x_l^s, \eta_l^s \rangle + \rho_s^2 C_2^6 \Lambda, \tag{23}$$

which proves the statement of the lemma. □

LEMMA 2

If $x^s, x_{i(s)}^s, \ldots, x_{i(m-1)}^{m-1} \in X$, $m > s$, then

$$\max_{s \le \tau \le m-1} \max_{0 \le l \le i(\tau)} \| x_l^\tau - x^s \| \le \sum_{\tau=s}^{m-1} \rho_\tau \sum_{l=0}^{i(\tau)-1,} \lambda_{\tau l} (\rho_\tau \| H_0^\tau \xi^\tau \| \varepsilon_\tau^{-1} + C_2^3). \tag{24}$$

Proof

Since $x_{i(m-1)}^{m-1} \in X$, we have from the algorithm formulae

$$\begin{aligned} x^m &= x_{i(m-1)}^{m-1} = x^{m-1} - \rho_{m-1} H_{i(m-1)}^{m-1} \xi^{m-1} \\ &= x^{m-1} - \rho_{m-1} (H_{i(m-1)-1}^{m-1} + \lambda_{m-1, i(m-1)-1} \xi_{i(m-1)-1}^{m-1} \xi^{m-1^{\mathrm{T}}}) \xi^{m-1} \\ &= (x^{m-1} - \rho_{m-1} H_{i(m-1)-1}^{m-1} \xi^{m-1}) - \rho_{m-1} \lambda_{m-1, i(m-1)-1} \xi_{i(m-1)-1}^{m-1} \| \xi^{m-1} \|^2 \\ &= x_{i(m-1)-1}^{m-1} - \rho_{m-1} \lambda_{m-1, i(m-1)-1} \xi_{i(m-1)-1}^{m-1} \| \xi^{m-1} \|^2. \end{aligned} \tag{25}$$

Using this equality the proper amount of times, we obtain

$$\begin{aligned} x^m &= x_0^{m-1} - \rho_{m-1} \| \xi^{m-1} \|^2 \sum_{l=0}^{i(m-1)-1} \lambda_{m-1,l} \xi_l^{m-1} \\ &= x^{m-1} - \rho_{m-1} H_0^{m-1} \xi^{m-1} - \rho_{m-1} \| \xi^{m-1} \|^2 \sum_{l=0}^{i(m-1)-1} \lambda_{m-1,l} \xi_l^{m-1}. \end{aligned}$$

Since $x^s, x_{i(s)}^s, \ldots, x_{i(m-1)}^{m-1} \in X$, $m > s$ (see statement of the lemma), again applying this formula for $m-1, \ldots, s+1$, we obtain

$$x^m = x^s - \sum_{\tau=s}^{m-1} \rho_\tau H_0^\tau \xi^\tau - \sum_{\tau=s}^{m-1} \rho_\tau \| \xi^\tau \|^2 \sum_{l=0}^{i(\tau)-1} \lambda_{\tau l} \xi_l^\tau. \tag{26}$$

In view of conditions (9) and (10) of the theorem, step 6 of the algorithm, and the last equality, we can estimate

$$\begin{aligned}
\|x^m - x^s\| &\le \sum_{\tau=s}^{m-1} \rho_\tau \| H_0^\tau \xi^\tau \| + \sum_{\tau=s}^{m-1} \rho_\tau \|\xi^\tau\|^2 \sum_{l=0}^{i(\tau)-1} \lambda_{\tau l} \|\xi_l^\tau\| \\
&\le \sum_{\tau=s}^{m-1} \left(\rho_\tau \| H_0^\tau \xi^\tau \| + C_2^3 \rho_\tau \sum_{l=0}^{i(\tau)-1} \lambda_{\tau l} \right) \\
&= \sum_{\tau=s}^{m-1} \rho_\tau \sum_{l=0}^{i(\tau)-1} \lambda_{\tau l} \left(\rho_\tau \| H_0^\tau \xi^\tau \| \left(\rho_\tau \sum_{l=0}^{i(\tau)-1} \lambda_{\tau l} \right)^{-1} + C_2^3 \right) \\
&\le \sum_{\tau=s}^{m-1} \rho_\tau \sum_{l=0}^{i(\tau)-1} \lambda_{\tau l} (\rho_\tau \| H_0^\tau \xi^\tau \| \varepsilon_\tau^{-1} + C_2^3).
\end{aligned} \tag{27}$$

The statement of the lemma follows from (25), (26) and (27). □

Let us consider the events $\omega \in B$ such that there exists a subsequence $\{x^{l_\kappa}\}$ with

$$x^{l_\kappa} \to x', \; W(x') \ge 0, \; U_\delta(x') \subset X, \tag{28}$$

where δ is some positive random value for almost all $\omega \in B$.

LEMMA 3

If $W(x') \ge 0$, $U_\delta(x') \subset X$, then for any random value ε satisfying inequality $0 < \varepsilon < \surd(W(x'))$ there exists a subsequence $\{v_\kappa\}$ such that

$$\varliminf_{\kappa\to\infty} W(x^{v_\kappa}) < W(x') \quad \text{a.s.}, \tag{29}$$

and

$$\varlimsup_{\kappa\to\infty} \max_{l_\kappa < \tau \le v_\kappa - 1} \max_{0 \le l \le i(\tau)} \|x_l^\tau - x^{l_\kappa}\| \le q \overset{\text{def}}{=} 2^{-1} \min\{\varepsilon, \delta\} \quad \text{a.s.} \tag{30}$$

Proof

Denote ε as some random value such that $0 < \varepsilon < \surd(W(x'))$ for $\omega \in B$. We define the index subsequence $\{v_\kappa\}$ (this subsequence depends upon ω) such that

$$C_2^3 \sum_{\tau=l_\kappa}^{v_\kappa - 1} \varepsilon_\tau \to q \quad \text{a.s.}; \tag{31}$$

the existence of this subsequence follows from the theorem conditions (11)–(13). In view of conditions (15)–(19) and step 6 of the algorithm,

$$\sum_{\tau=l_\kappa}^{\nu_\kappa-1} \rho_\tau \sum_{l=0}^{i(\tau)-1} \lambda_{\tau l} C_2^3 \to q \quad \text{a.s.} \tag{32}$$

Since $\rho_\tau \| H_0^\tau \xi^\tau \| \varepsilon_\tau^{-1} \to 0$ a.s. for $\tau \to \infty$ (see condition (14)), lemma 2 and (32) imply inequality (30). From (30) and $x^{l_\kappa} \to x'$ for $\kappa \to \infty$, it follows that the approximations x_l^τ, $l_\kappa < \tau \le \nu_\kappa - 1$, $0 \le l \le i(\tau)$ belong to the set $U_{2q}(x')$ for sufficiently large numbers κ (this κ depends upon ω). Since

$$2q = \min\{\varepsilon, \delta\} \le \varepsilon < \sqrt{W(x')} = \sqrt{\min_{y \in X^*} \| x' - y \|^2} = \min_{y \in X^*} \| x' - y \|,$$

then

$$X^* \cap U_{2q}(x') = \varnothing. \tag{33}$$

It also follows from (28) that

$$U_{2q}(x') \subset U_\delta(x') \subset X.$$

Since the points x_l^τ for $l_\kappa \le \tau \le \nu_\kappa - 1$, $0 \le l \le i(\tau)$, belong to the set $U_{2q}(x')$ for sufficiently large κ, (33) implies the existence of a random value $\alpha > 0$ a.s. such that

$$f^* - f(x_l^\tau) \le -\alpha \quad \text{for } l_\kappa \le \tau \le \nu_\kappa - 1, \ 0 \le l \le i(\tau), \tag{34}$$

for sufficiently large κ. Applying inequality (21) from lemma 11 the necessary amount of times with (34) we have

$$\begin{aligned} W(x^{\nu_\kappa}) = W\left(x_{i(\nu_\kappa-1)}^{\nu_\kappa-1}\right) &\le W(x^{\nu_\kappa-1}) + 2\rho_{\nu_\kappa-1} C_3 \| H_0^{\nu_\kappa-1} \xi^{\nu_\kappa-1} \| \\ &+ \rho_{\nu_\kappa-1}^2 \| H_0^{\nu_\kappa-1} \xi^{\nu_\kappa-1} \|^2 - 2\rho_{\nu_\kappa-1} \| \xi^{\nu_\kappa-1} \|^2 \sum_{l=0}^{i(\nu_\kappa-1)-1} \lambda_{\nu_\kappa-1,l}\, \alpha \\ &+ 2\rho_{\nu_\kappa-1} \| \xi^{\nu_\kappa-1} \|^2 \sum_{l=0}^{i(\nu_\kappa-1)-1} \lambda_{\nu_\kappa-1,l} \langle x_{\nu_\kappa-1,l}^* - x_l^{\nu_\kappa-1}, \eta_l^{\nu_\kappa-1} \rangle \\ &+ \rho_{\nu_\kappa-1}^2 C_2^6 \Lambda \le W(x^{l_\kappa}) + 2\sum_{\tau=l_\kappa}^{\nu_\kappa-1} \rho_\tau C_3 \| H_0^\tau \xi^\tau \| + \sum_{\tau=l_\kappa}^{\nu_\kappa-1} \rho_\tau^2 \| H_0^\tau \xi^\tau \|^2 \end{aligned}$$

$$-2\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\rho_\tau\alpha\|\xi^\tau\|^2\sum_{l=0}^{i(\tau)-1}\lambda_{\tau l}+2\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\rho_\tau\|\xi^\tau\|^2\sum_{l=0}^{i(\tau)-1}\lambda_{\tau l}\langle x_{\tau l}^*-x_l^\tau,\eta_l^\tau\rangle$$

$$+C_2^6\Lambda\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\rho_\tau^2\overset{\text{def}}{=}W(x^{l_\kappa})+T_2+T_3+T_4+T_5+T_6. \tag{35}$$

We estimate the lower limit of the terms in inequality (35). For the second term, we have (see (14) and (31))

$$\begin{aligned}\underline{\lim}_{\kappa\to\infty}T_2&=\underline{\lim}_{\kappa\to\infty}\,2\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\rho_\tau C_3\|H_0^\tau\xi^\tau\|\\
&=2C_3\underline{\lim}_{\kappa\to\infty}\left(\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\varepsilon_\tau\right)\left[\left(\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\varepsilon_\tau\right)^{-1}\left(\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\rho_\tau\|H_0^\tau\xi^\tau\|\right)\right]\\
&=2C_3\overline{\lim}_{\kappa\to\infty}\left(\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\varepsilon_\tau\right)\underline{\lim}_{\kappa\to\infty}\left[\left(\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\varepsilon_\tau\right)^{-1}\left(\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\rho_\tau\|H_0^\tau\xi^\tau\|\right)\right]\\
&=2C_3(C_2^{-3}q)\cdot 0=0\quad\text{a.s.}\end{aligned} \tag{36}$$

For term T_3:

$$\begin{aligned}\underline{\lim}_{\kappa\to\infty}T_3&=\underline{\lim}_{\kappa\to\infty}\,2\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\rho_\tau^2\|H_0^\tau\xi^\tau\|^2\le\underline{\lim}_{\kappa\to\infty}\left(\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\rho_\tau\|H_0^\tau\xi^\tau\|\right)^2\\
&=(2C_3)^{-2}\,\underline{\lim}_{\kappa\to\infty}T_2^2=0\quad\text{a.s.}\end{aligned} \tag{37}$$

In view of algorithm step 6 and the convexity of the function $\|\cdot\|^2$ for the fourth term in (35),

$$\begin{aligned}\underline{\lim}_{\kappa\to\infty}T_4&=\underline{\lim}_{\kappa\to\infty}\left(-2\alpha\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\rho_\tau\|\xi^\tau\|^2\sum_{l=0}^{i(\tau)-1}\lambda_{\tau l}\right)\\
&\le-2\alpha\,\overline{\lim}_{\kappa\to\infty}\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\varepsilon_\tau\|\xi^\tau\|^2=-2\alpha\,\overline{\lim}_{\kappa\to\infty}\left(\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\varepsilon_\tau\right)\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\left(\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\varepsilon_\kappa\right)^{-1}\varepsilon_\tau\|\xi^\tau\|^2\\
&\le-2\alpha\,\overline{\lim}_{\kappa\to\infty}\left(\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\varepsilon_\tau\right)\left\|\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\varepsilon_\tau\left(\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\varepsilon_\kappa\right)^{-1}\xi^\tau\right\|^2\end{aligned}$$

$$\le -2\alpha \varlimsup_{\kappa\to\infty} \left(\sum_{\tau=l_\kappa}^{\nu_\kappa-1} \varepsilon_\tau \right)^{-1} \varlimsup_{\kappa\to\infty} \left\| \sum_{\tau=l_\kappa}^{\nu_\kappa-1} \varepsilon_\tau \xi^\tau \right\|^2$$

$$= -2\alpha C_2^{-3} q^{-1} \varlimsup_{\kappa\to\infty} \left\| \sum_{\tau=l_\kappa}^{\nu_\kappa-1} \varepsilon_\tau \xi^\tau \right\|^2. \tag{38}$$

The martingale series $\sum_{\tau=0}^{\infty} \varepsilon_\tau \eta_\tau$ is convergent with conditions (11)–(13) and thus

$$\varlimsup_{\kappa\to\infty} \left\| \sum_{\tau=l_\kappa}^{\nu_\kappa-1} \varepsilon_\tau \xi^\tau \right\|^2 = \varlimsup_{\kappa\to\infty} \left\| \sum_{\tau=l_\kappa}^{\nu_\kappa-1} \varepsilon_\tau g(x^\tau) + \sum_{\tau=l_\kappa}^{\nu_\kappa-1} \varepsilon_\tau \eta^\tau \right\|^2$$

$$= \varlimsup_{\kappa\to\infty} \left\| \sum_{\tau=l_\kappa}^{\nu_\kappa-1} \varepsilon_\tau g(x^\tau) \right\|^2. \tag{39}$$

For sufficiently large κ, the points x_l^τ, $l_\kappa \le \tau \le \nu_\kappa - 1$, $0 \le l \le i(\tau)$, belong to the convex set $U_{2q}(x')$ and $U_{2q}(x') \cap X^* = \varnothing$. So, for properly small q there exists a positive random value $\gamma > 0$ a.s. such that $\langle g(x^\tau), x^* - x' \rangle > \gamma \|x^* - x'\| > 0$, $x^* \in X^*$. Further, we obtain

$$\left\| \sum_{\tau=l_\kappa}^{\nu_\kappa-1} \varepsilon_\tau g(x^\tau) \right\|^2 < \left(\|x^* - x'\|^{-1} \left\langle \sum_{\tau=l_\kappa}^{\nu_\kappa-1} \varepsilon_\tau g(x^\tau), x^* - x' \right\rangle \right)^2$$

$$= \left(\|x^* - x'\|^{-1} \sum_{\tau=l_\kappa}^{\nu_\kappa-1} \varepsilon_\tau \langle g(x^\tau), x^* - x' \rangle \right)^2 \ge \left(\gamma \sum_{\tau=l_\kappa}^{\nu_\kappa-1} \varepsilon_\tau \right)^2.$$

Combining the last inequality with (38) and (39), we obtain

$$\varlimsup_{\kappa\to\infty} T_4 \le -2\alpha C_2^{-3} q^{-1} \varlimsup_{\kappa\to\infty} \left(\gamma \sum_{\tau=l_\kappa}^{\nu_\kappa-1} \varepsilon_\kappa \right)^2 = -2\alpha C_2^{-3} q^{-1} \gamma^2 C_2^{-6} q^2 < 0 \quad \text{a.s.} \tag{40}$$

It follows from conditions (9), (16) and (19) that

$$\sum_{\tau=0}^{\infty} \sum_{l=0}^{\infty} \rho_\tau^2 \|\xi^\tau\|^4 \lambda_{\tau l}^2 \le C_2 \Lambda \sum_{\tau=0}^{\infty} \rho_\tau^2 < \infty \quad \text{a.s.;}$$

consequently, the martingale series

$$\sum_{\tau=0}^{\infty} \rho_\tau \|\xi^\tau\|^2 \sum_{l=0}^{i(\tau)-1} \lambda_{\tau l} \langle x_{\tau l}^* - x_l^\tau, \eta_l^\tau \rangle$$

is convergent. This fact implies

$$\varliminf_{\kappa\to\infty} T_5 = \lim_{\kappa\to\infty} T_5 = 2 \lim_{\kappa\to\infty} \sum_{\tau=l_\kappa}^{\nu_\kappa - 1} \rho_\tau \|\xi^\tau\|^2 \sum_{l=0}^{i(\tau)-1} \lambda_{\tau l} \langle x_{\tau l}^* - x_l^\tau, \eta_l^\tau \rangle = 0 \ \text{a.s.} \tag{41}$$

We have from condition (16) also that

$$\varliminf_{\kappa\to\infty} T_6 = \varliminf_{\kappa\to\infty} C_2^6 \Lambda \sum_{\tau=l_\kappa}^{\nu_\kappa - 1} \rho_\tau^2 = 0 \ \text{a.s.} \tag{42}$$

Taking the lower limit for (35) and using (36), (37) and (40)–(42),

$$\varliminf_{\kappa\to\infty} W(x^{\nu_\kappa}) \le \varliminf_{\kappa\to\infty} W(x^{l_\kappa}) - 2\alpha\gamma^2 q C_2^{-9} < W(x') - 2\alpha\gamma^2 q C_2^{-9}.$$

The last inequality implies (29) and the lemma is proved. □

Lemma 3 proves condition D3 for the subsequences that satisfy condition (28). Now let us consider the case with

$$x^{l_\kappa} \to x', \quad x' \in \partial X,$$

where ∂X is the boundary of the set X. As in the previous case, we define the index subsequence $\{\nu_\kappa\}$ such that

$$C_2^3 \sum_{\tau=l_\kappa}^{\nu_\kappa - 1} \varepsilon_\tau \to q = 2^{-1}\varepsilon.$$

We consider the following two possibilities:

(1) There exists an infinite subsequence $\{\theta_m\}$ such that $l_m \le \theta_m < \nu_m$, $x^{\theta_m} \in X$, $x_{i(\theta_m)}^{\theta_m} \notin X$, $x^{\theta_m+1} = x^0$. Then condition (8) implies

$$\varliminf_{m\to\infty} W(x^{\theta_m+1}) = \|x_0^* - x^0\|^2 < C_1^2 \le W(x')$$

and subsequence $\{x^{\theta_m+1}\}$ satisfies the sufficient condition D3.

(2) There exists a number K such that $x^\tau \in X$ for $l_k \le \tau \le \nu_\kappa$, $\kappa \ge K$. For this case, the proof of condition D3 coincides with the proof of lemma 3, where x' belongs to the interior of the set X.

This proves condition D3.

Condition D4 is valid because the function $W(x)$ is constant on X^*.

Let us prove the last condition D5. We consider the subsequence x^{s_κ} such that $x^{s_\kappa} \to x^*$, $x^* \in X^*$. It follows from estimate (27) that

$$\| x^{s_\kappa+1} - x^{s_\kappa} \| \le \rho_{s_\kappa} \sum_{l=0}^{i(s_\kappa)-1} \lambda_{s_\kappa l} (\rho_{s_\kappa} \| H_0^{s_\kappa} \xi^{s_\kappa} \| \varepsilon_{s_\kappa}^{-1} + C_2^3). \tag{43}$$

Since (see conditions (12), (14), (16) and (19))

$$\rho_{s_\kappa} \| H_0^{s_\kappa} \xi^{s_\kappa} \| \varepsilon_{s_\kappa}^{-1} \to 0,$$

$$\left| \rho_{s_\kappa} \sum_{l=0}^{i(s_\kappa)-1} \lambda_{s_\kappa l} - \varepsilon_{s_\kappa} \right| \to 0,$$

$$\varepsilon_{s_\kappa} \to 0,$$

(43) implies

$$\| x^{s_\kappa+1} - x^{s_\kappa} \| \to 0 \tag{44}$$

for almost all ω such that $x^{s_\kappa} \to X^*$, $x^* \in X^*$. The function $W(x)$ is continuous, thus (44) proves condition D5. All conditions D1–D5 are checked and the theorem is proved. □

References

[1] B. Betro and L. De Biase, A Newton-like method for stochastic optimization, in: *Towards Global Optimization 2* (North-Holland, 1978), pp. 269–289.

[2] J.N. Dennis and J.J. Moré, Quasi-Newton methods, motivation and theory, SIAM Rev. 19(1977) 46–89.

[3] Yu.M. Ermoliev, Stochastic quasi-gradient methods and their application to systems optimization, Stochastics 4(1983)1–37.

[4] A. Gaivoronski, Implementation of stochastic quasigradient methods, in: *Numerical Techniques for Stochastic Optimization*, ed. Yu.M. Ermoliev and R.J.-B. Wets (Springer, 1988), pp. 313–353.

[5] L. Gerencser, Strong consistency theorems related to stochastic quasi-Newton methods, in: *Stochastic Optimization*, Lecture Notes in Control and Information Science 81 (Springer, 1984).

[6] Yu.M. Kaniovski, P.S. Knopov and Z.V. Nekrylova, *Limit Theorems for Stochastic Programming Processes* (Naukova Dumka, Kiev, 1980), in Russian.

[7] H.J. Kushner and G. Jin, Stochastic approximation algorithms for parallel and distributed processing, Stochastics 22(1987)219–250.

[8] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification* (MIT Press, 1983).

[9] K. Marti, Descent stochastic quasigradient methods, in: *Numerical Techniques for Stochastic Optimization*, ed. Yu.M. Ermoliev and R.J.-B. Wets (Springer, 1988), pp. 393–401.

[10] N.M. Novikova, Some stochastic programming methods in Hilbert Space, Zh. Vichislitelnoj Mat. i Mat. Phis. 25(1985)1795–1813, in Russian [USSR Comput. Math. and Math. Phys. (in English)].

[11] E.A. Nurminski, *Numerical Methods for Solving Deterministic and Stochastic Minimax Problems* (Naukova Dumka, Kiev, 1979), in Russian.

[12] G. Pflug, Stepsize rules, stopping times and their implementation in stochastic quasigradient algorithms, in: *Numerical Techniques for Stochastic Optimization*, ed. Yu.M. Ermoliev and R.J.-B. Wets (Springer, 1988), pp. 353–373.

[13] R.T. Rockafellar and R.J.-B. Wets, On the interchange of subdifferentiation and conditional expectation for convex functionals, Stochastics 7(1982)173–182.

[14] A. Ruszczynski and W. Syski, A method of aggregate stochastic sub-gradients with on-line stepsize rules for convex stochastic programming problems, Math. Progr. Study 28(1985)113–131.

[15] G.S. Saridis, Learning applied to successive approximation algorithms, IEEE Trans. Syst. Sci. Cybern. SSC-6(1970)97–103.

[16] D.J. Sarkison, The use of stochastic approximation to solve the system identification problem, IEEE Trans. Auto. Control AC-12(1967)563–567.

[17] S.P. Uryas'ev, Stochastic quasi-gradient algorithms with adaptively controlled parameters, Working Paper WP-86-32, IIASA (1986).

[18] S.P. Uryas'ev, *Adaptive Algorithms of Stochastic Optimization and Game Theory* (Nauka, Moskow, 1990), in Russian.

[19] S.P. Uryas'ev, Adaptive variable metric algorithms for nondifferentiable optimization problems, in: *Analysis and Optimization of Systems*, ed. A. Bensoussan and J.L. Lions, Proc. 9th Int. Conf., Antibes (Springer, 1990), pp. 432–441.

[20] R.J.-B. Wets, Modeling and solution strategies for constrained stochastic optimization problems, Ann. Oper. Res. 1(1984)3–22.