Working Paper

A Stochastic Quasigradient Algorithm with Variable Metric

S.P. Uryas'ev

WP-89-98
December 1989

IIASA

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

Telephone: (0 22 36) 715 21 * 0 □ Telex: 079 137 iiasa a □ Telefax: (0 22 36) 71313

A Stochastic Quasigradient Algorithm with Variable Metric

S.P. Uryas'ev

WP-89-98
December 1989

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

Telephone: (0 22 36) 715 21 * 0 □ Telex: 079 137 iiasa a □ Telefax: (0 22 36) 71313

Foreword

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.

Alexander B. Kurzhanski
Chairman
System and Decision Sciences Program

Contents

A Stochastic Quasigradient Algorithm with Variable Metric

S.P. Uryas'ev

1 Introduction

Stochastic quasigradient (or stochastic approximation) algorithms are used for the optimization of quite general stochastic systems with smooth, nonsmooth, and infinite-dimensional objective functions, for distributed systems and others (see, for example, [3], [4], [6]–[9], [11]–[13], [16], [20]). The structure of such algorithms is simple, and at each iteration only few additional calculations are required. 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 quite 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], [10], [14], [17], [18] and [21]). In this paper, the variable metric algorithm for stochastic programming problems with a *nonsmooth* objective function is presented. Such algorithms were already proposed in [19].

2 Basic Idea of the Algorithm

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

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

where 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 assumptions, the generalized differential of the convex function $f(x)$ is calculated by the formula (see [15])

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

then $\partial_x \varphi(x, \omega)$ is a set of vectors being 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 $\rho_s, s = 0, 1, \ldots$ is a sequence of positive random scalar stepsizes; $H^s, s = 0, 1, \ldots$ is a sequence of $n \times n$ random square matrices; $\xi^s, s = 0, 1, \ldots$ is a sequence of stochastic quasigradients, i.e. 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 defined 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 iteration 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

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

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)):

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

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

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

One can see that

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

thus $-\rho_s \xi_0^s \xi^{s^T}$ 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 iteration $s-1$. To modify matrix H_0^s, we use the stochastic quasigradient method (see [3]):

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

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

$$\begin{aligned} H_{i+1}^s &= H_i^s + \lambda_i^s \xi_i^s \xi^{s^T}, \qquad i = 0, \ldots, i(s); \\ H_0^s &= H^{s-1} = H_{i(s-1)+1}^{s-1}, \end{aligned} \tag{6}$$

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

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

In formula (6), the matrix H is modified additively, but multiplicative variants of this algorithm also can be developed (see [19]).

3 Formal Description of the Algorithm and Necessary Conditions for Convergence

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

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

Algorithm (3), (6) can solve the optimization problem (1) without constraints. To simplify the convergence proof of the algorithm, we assume that some convex compact set $X \subset R^n$ 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 of 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 $\{\epsilon_s\}, s = 0, s, \ldots$ and $\{\lambda_{sl}\}$, $s = 0, 1, \ldots, l = 0, 1, \ldots$ 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 also could be written for these sequences, but we do not want to overload the convergence proof with them now. The positive value ϵ_s define $i(s)$ in the algorithm, iterations with respect to matrix are stopped if $\rho_s \sum_{l=0}^{i(s)-1} \lambda_{sl} \geq \epsilon_s$. To avoid misunderstandings, we present here a full formal description of the algorithm.

Algorithm 1

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

Step III Set $i = 0$.

Step IV Compute the point x_i^s

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

Step V Compute $H_{i+1}^s = H_i^s + \lambda_{si} \xi_i^s \xi^{s^T}$,

here ξ_i^s is a stochastic quasigradient at the point x_i^s.

Step VI If $i \geq 1$ and $\rho_s \sum_{l=0}^{i-1} \lambda_{sl} \geq \epsilon_s$ then $i(s) = i$; go to Step VIII.

Step VII Set $i = i + 1$ and return to Step IV.

Step VIII 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 IX Set $s = s + 1$ and return to Step II.

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 necessary conditions (see [20]) for convergence of stochastic algorithms. (These conditions are similar to the conditions in [12] but are more general.)

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

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

D2 $W : X \to R$ 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 $\epsilon(\omega) > 0$ a.s. there exists a subsequence $\{\nu_k(\omega)\}$ such that

$$W(x^\tau) \leq W(x'(\omega)) + \epsilon(\omega) \quad \text{for} \quad l_\kappa(\omega) \leq \tau \leq \nu_\kappa(\omega),$$

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

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

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

is satisfied.

Next is the theorem about these necessary conditions (see [20]).

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.*

4 Convergence of the Algorithm

Below we formulate the theorem on the convergence of algorithm 1.

Theorem 2 *Let $f : R^n \to R$ be a convex (possibly nonsmooth) function, X be a compact convex set such that $X^* \subset X \subset R^n$ 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 $\{\epsilon_s\}$ 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 parameters satisfy the conditions

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

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

$$\epsilon_s > 0, s = 0, 1, \dots , \tag{11}$$

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

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

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

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

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

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

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

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

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

Proof We use necessary 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 algorithm construction 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

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

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 $x^{l_\kappa(\omega)}(\omega) \to x'(\omega) \notin X^*\}$ be greater than zero. We shall omit the latter for the simplicity of argument ω. Steps IV, V and VI of the algorithm and conditions (9) and (10) of the theorem imply

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

Applying this estimate the proper amount of times we obtain

$$W(x^s_{i(s)}) \le W(x^s_0) + 2\rho_s\|\xi^s\|^2 \sum_{l=0}^{i(s)-1} \lambda_{sl}\langle x^*_{sl} - x^s_l, \xi^s_l\rangle + \rho_s C_2^6 \sum_{l=0}^{i(s)-1} \lambda^2_{sl} . \tag{21}$$

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

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

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

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

Substituting the two previous estimates into estimate (21) and (19)–(20) yields

$$\begin{aligned} W(x^s_{i(s)}) \le\ & W(x^s) + 2\rho_s\|x^*_s - x^s\|\|H^s_0\xi^s\| + \rho_s^2\|H^s_0\xi^s\|^2 + \\ & + 2\rho_s\|\xi^s\|^2 \sum_{l=0}^{i(s)-1} \lambda_{sl}(f^* - f(x^s_l)) + 2\rho_s\|\xi^s\|^2 \sum_{l=0}^{i(s)-1} \lambda_{sl}\langle x^*_{sl} - x^s_l, \eta^s_l\rangle + \\ & + \rho_s^2 C_2^6 \sum_{l=0}^{i(s)-1} \lambda^2_{sl} \le W(x^s) + 2\rho_s C_3\|H^s_0\xi^s\| + \\ & + \rho_s^2\|H_0^2\xi^s\|^2 + 2\rho_s\|\xi^s\|^2 \sum_{l=0}^{i(s)-1} \lambda_{sl}(f^* - f(x^s_l)) + \\ & + 2\rho_s\|\xi^s\|^2 \sum_{l=0}^{i(s)-1} \lambda_{sl}\langle x^*_{sl} - x^s_l, \eta^s_l\rangle + \rho_s^2 C_2^6 \Lambda . \end{aligned} \tag{22}$$

If $x^{m-1}_{i(m-1)} \in X$, then we have from the algorithm formulae

$$\begin{aligned} x^m &= x^{m-1}_{i(m-1)} = x^{m-1} - \rho_{m-1} H^{m-1}_{i(m-1)}\xi^{m-1} = \\ &= x^{m-1} - \rho_{m-1}(H^{m-1}_{i(m-1)-1} + \lambda_{m-1,i(m-1)-1}\xi^{m-1}_{i(m-1)-1}\xi^{m-1^T})\xi^{m-1} = \\ &= (x^{m-1} - \rho_{m-1}H^{m-1}_{i(m-1)-1}\xi^{m-1}) - \rho_{m-1}\lambda_{m-1,i(m-1)-1}\xi^{m-1}_{i(m-1)-1}\|\xi^{m-1}\|^2 = \\ &= x^{m-1}_{i(m-1)-1} - \rho_{m-1}\lambda_{m-1,i(m-1)-1}\xi^{m-1}_{i(m-1)-1}\|\xi^{m-1}\|^2 . \end{aligned} \tag{23}$$

Using this equality the proper amount of times we get

$$\begin{aligned} x^m &= x^{m-1}_0 - \rho_{m-1}\|\xi^{m-1}\|^2 \sum_{l=0}^{i(m-1)-1} \lambda_{m-1,l}\xi^{m-1}_l = \\ &= x^{m-1} - \rho_{m-1}H^{m-1}_0\xi^{m-1} - \rho_{m-1}\|\xi^{m-1}\|^2 \sum_{l=0}^{i(m-1)-1} \lambda_{m-1,l}\xi^{m-1}_l . \end{aligned}$$

If x^s, $x^s_{i(s)}, \ldots, x^{m-1}_{i(m-1)} \in X$, $m > s$ then again applying this formula for $m-1, \ldots, s+1$ we obtain

$$x^m = x^s - \sum_{\tau=s}^{m-1} \rho_\tau H^\tau_0\xi^\tau - \sum_{\tau=s}^{m-1} \rho_\tau\|\xi^\tau\|^2 \sum_{l=0}^{i(\tau)-1} \lambda_{\tau l}\xi^\tau_l . \tag{24}$$

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

$$\begin{aligned}
\|x^m - x^s\| &\le \sum_{\tau=s}^{m-1} \rho_\tau \|H_0^\tau \xi^\tau\| + \sum_{\tau=s}^{m-1} \rho_\tau \|\xi^\tau\|^2 \sum_{l=0}^{i(\tau)-1} \lambda_{\tau l} \|\xi_l^\tau\| \le \\
&\le \sum_{\tau=s}^{m-1} \left(\rho_\tau \|H_0^\tau \xi^\tau\| + C_2^3 \rho_\tau \sum_{l=0}^{i(\tau)-1} \lambda_{\tau l} \right) = \\
&= \sum_{\tau=s}^{m-1} \rho_\tau \sum_{l=0}^{i(\tau)-1} \lambda_{\tau l} \left(\rho_\tau \|H_0^\tau \xi^\tau\| \left(\rho_\tau \sum_{l=0}^{i(\tau)-1} \lambda_{\tau l} \right)^{-1} + C_2^3 \right) \le \\
&\le \sum_{\tau=s}^{m-1} \rho_\tau \sum_{l=0}^{i(\tau)-1} \lambda_{\tau l} (\rho_\tau \|H_0^\tau \xi^\tau\| \epsilon_\tau^{-1} + C_2^3) \,. \qquad (25)
\end{aligned}$$

It also follows from (23), (24) and (25) also that

$$\max_{s \le \tau \le m-1} \max_{0 \le l \le i(\tau)} \|x_l^\tau - x^s\| \le \sum_{\tau=s}^{m-1} \rho_\tau \sum_{l=0}^{i(\tau)-1} \lambda_{\tau l} (\rho_\tau \|H_0^\tau \xi^\tau\| \epsilon_\tau^{-1} + C_2^3) \,. \qquad (26)$$

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

$$x^{l_\kappa} \to x', W(x') \ge 0, U_\delta(x') \subset X \qquad \text{for} \quad \kappa \to \infty \,, \qquad (27)$$

where δ is some positive random value for almost all $\omega \in B$. Denote ϵ as some random value such that $0 < \epsilon < \sqrt{W(x')}$ for $\omega \in B$. We define the index subsequence $\{\nu_\kappa\}$ (this subsequence depends upon ω) such that

$$C_2^3 \sum_{\tau=l_\kappa}^{\nu_\kappa - 1} \epsilon_\tau \to q \stackrel{\text{def}}{=} 2^{-1} \min\{\epsilon, \delta\} \qquad \text{a.s.} \,, \qquad (28)$$

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

$$\sum_{\tau=l_\kappa}^{\nu_\kappa - 1} \rho_\tau \sum_{l=o}^{i(\tau)-1} \lambda_{\tau l} C_2^3 \to q \qquad \text{a.s.} \qquad (29)$$

Since $\rho_\tau \|H_0^\tau \xi^\tau\| \epsilon_\tau^{-1} \to 0$ a.s. for $\tau \to \infty$ (see condition (14)), then (26) and (29) imply

$$\overline{\lim_{\kappa \to \infty}} \max_{l_\kappa < \tau \le \nu_\kappa - 1} \max_{0 \le l \le i(\tau)} \|x_l^\tau - x^{l_\kappa}\| \le q \qquad \text{a.s.} \qquad (30)$$

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 $\cup_{2q}(x')$ for sufficiently large numbers κ (this κ depends upon ω). Since

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

then

$$X^* \cap U_{2q}(x') = \emptyset \,. \qquad (31)$$

It also follows from (27) 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, then (31) implies the existence of a random value $\alpha > 0$ a.s. such that

$$f^* - f(x_l^\tau) \le -\alpha \quad \text{for} \quad l_\kappa \le \tau \le \nu_\kappa - 1, 0 \le l \le i(\tau) \tag{32}$$

for sufficiently large κ. Applying inequality (22) the necessary amount of times with (32) we have

$$\begin{aligned} W(x^{\nu_\kappa}) &= W(x^{\nu_\kappa - 1}_{i(\nu_\kappa - 1)}) \le W(x^{\nu_\kappa - 1}) + 2\rho_{\nu_\kappa - 1} C_3 \|H_0^{\nu_\kappa - 1}\xi^{\nu_\kappa - 1}\| + \\ &+ \rho_{\nu_\kappa - 1}^2 \|H_0^{\nu_\kappa - 1}\xi^{\nu_\kappa - 1}\|^2 - 2\rho_{\nu_\kappa - 1}\|\xi^{\nu_\kappa - 1}\|^2 \sum_{l=0}^{i(\nu_\kappa - 1) - 1} \lambda_{\nu_\kappa - 1, l}\alpha + \\ &+ 2\rho_{\nu_\kappa - 1}\|\xi^{\nu_\kappa - 1}\|^2 \sum_{l=0}^{i(\nu_\kappa - 1) - 1} \lambda_{\nu_\kappa - 1, l}\langle x^*_{\nu_\kappa - 1, l} - x_l^{\nu_\kappa - 1}, \eta_l^{\nu_\kappa - 1}\rangle + \\ &+ \rho_{\nu_\kappa - 1}^2 C_2^6 \Lambda \le W(x^{l_\kappa}) + 2\sum_{\tau = l_\kappa}^{\nu_\kappa - 1} \rho_\tau C_3 \|H_0^\tau \xi^\tau\| + \sum_{\tau = l_\kappa}^{\nu_\kappa - 1} \rho_\tau^2 \|H_0^\tau \xi^\tau\|^2 - \\ &- 2\sum_{\tau = l_\kappa}^{\nu_\kappa - 1} \rho_\tau \alpha \|\xi^\tau\|^2 \sum_{l=0}^{i(\tau) - 1} \lambda_{\tau l} + 2\sum_{\tau = l_\kappa}^{\nu_\kappa - 1} \rho_\tau \|\xi^\tau\|^2 \sum_{l=0}^{i(\tau) - 1} \lambda_{\tau l}\langle x^*_{\tau l} - x_l^\tau, \eta_l^\tau\rangle + \\ &+ C_2^6 \Lambda \sum_{\tau = l_\kappa}^{\nu_\kappa - 1} \rho_\tau^2 \stackrel{\text{def}}{=} W(x^{l_\kappa}) + T_2 + T_3 + T_4 + T_5 + T_6 \ . \end{aligned} \tag{33}$$

We estimate the lower limit of the terms in inequality (23). For the second term we have (see (14) and (28))

$$\begin{aligned} \varliminf_{\kappa \to \infty} T_2 &= \varliminf_{\kappa \to \infty} 2\sum_{\tau = l_\kappa}^{\nu_\kappa - 1} \rho_\tau C_3 \|H_0^\tau \xi^\tau\| = \\ &= 2C_3 \varliminf_{\kappa \to \infty} \left(\sum_{\tau = l_\kappa}^{\nu_\kappa - 1} \epsilon_\tau\right) \left[\left(\sum_{\tau = l_\kappa}^{\nu_\kappa - 1} \epsilon_\tau\right)^{-1} \left(\sum_{\tau = l_\kappa}^{\nu_\kappa - 1} \rho_\tau \|H_0^\tau \xi^\tau\|\right)\right] = \\ &\le 2C_3 \varlimsup_{\kappa \to \infty} \left(\sum_{\tau = l_\kappa}^{\nu_\kappa - 1} \epsilon_\tau\right) \varliminf_{\kappa \to \infty} \left[\left(\sum_{\tau = l_\kappa}^{\nu_\kappa - 1} \epsilon_\tau\right)^{-1} \left(\sum_{\tau = l_\kappa}^{\nu_\kappa - 1} \rho_\tau \|H_0^\tau \xi^\tau\|\right)\right] = \\ &= 2C_3(C_2^{-3} q) \cdot 0 = 0 \qquad \text{a.s.} \end{aligned} \tag{34}$$

For term T_3

$$\begin{aligned} \varliminf_{\kappa \to \infty} T_3 &= \varliminf_{\kappa \to \infty} \sum_{\tau = l_\kappa}^{\nu_\kappa - 1} \rho_\tau^2 \|H_0^\tau \xi^\tau\|^2 \le \varliminf_{\kappa \to \infty} \left(\sum_{\tau = l_\kappa}^{\nu_\kappa - 1} \rho_\tau \|H_0^\tau \xi^\tau\|\right)^2 = \\ &= (2C_3)^{-2} \varliminf_{\kappa \to \infty} T_2^2 = 0 \qquad \text{a.s.} \end{aligned} \tag{35}$$

In view of algorithm step VI and the convexity of the function $\|\cdot\|^2$ for the fourth term in (33)

$$\begin{aligned}
\underline{\lim}_{\kappa\to\infty} T_4 &= \underline{\lim}_{\kappa\to\infty}\left(-2\alpha\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\rho_\tau\|\xi^\tau\|^2\sum_{l=0}^{i(\tau)-1}\lambda_{\tau l}\right)\le \\
&\le -2\alpha\,\overline{\lim}_{\kappa\to\infty}\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\epsilon_\tau\|\xi^\tau\|^2 = -2\alpha\,\overline{\lim}_{\kappa\to\infty}\left(\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\epsilon_\tau\right)\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\left(\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\epsilon_\kappa\right)^{-1}\epsilon_\tau\|\xi^\tau\|^2\le \\
&\le -2\alpha\,\overline{\lim}_{\kappa\to\infty}\left(\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\epsilon_\tau\right)\|\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\epsilon_\tau\left(\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\epsilon_\kappa\right)^{-1}\xi^\tau\|^2 = \\
&\le -2\alpha\,\underline{\lim}_{\kappa\to\infty}\left(\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\epsilon_\tau\right)^{-1}\underline{\lim}_{\kappa\to\infty}\|\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\epsilon_\tau\xi^\tau\|^2 = \\
&= -2\alpha C_2^{-3}q^{-1}\,\underline{\lim}_{\kappa\to\infty}\|\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\epsilon_\tau\xi^\tau\|^2 . \qquad (36)
\end{aligned}$$

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

$$\begin{aligned}
\underline{\lim}_{\kappa\to\infty}\|\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\epsilon_\tau\xi^\tau\|^2 &= \underline{\lim}_{\kappa\to\infty}\|\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\epsilon_\tau g(x^\tau)+\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\epsilon_\tau\eta^\tau\|^2 = \\
&= \underline{\lim}_{\kappa\to\infty}\|\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\epsilon_\tau g(x^\tau)\|^2 . \qquad (37)
\end{aligned}$$

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 $U_{2q}(x')\cap X^*=\emptyset$. Consequently, 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 get

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

Combining the last inequality with (36) and (37) we obtain

$$\underline{\lim}_{\kappa\to\infty} T_4\le -2\alpha C_2^{-3}q^{-1}\,\overline{\lim}_{\kappa\to\infty}\left(\gamma\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\epsilon_\tau\right)^2 = -2\alpha C_2^{-3}q^{-1}\gamma^2C_2^{-6}q^2<0 \qquad \text{a.s.} \qquad (38)$$

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

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

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

$$\underline{\lim}_{\kappa\to\infty} T_5=\lim_{\kappa\to\infty} T_5=2\,\underline{\lim}_{\kappa\to\infty}\sum_{\tau=l_\kappa}^{\nu_\kappa-1}\rho_\tau\|\xi^\tau\|^2\sum_{l=0}^{i(\tau)-1}\lambda_{\tau l}\langle x_{\tau l}^*-x_l^\tau,\eta_l^\tau\rangle=0 \qquad \text{a.s.} \qquad (39)$$

We have from condition (16) also that

$$\underline{\lim}_{\kappa\to\infty} T_6 = \underline{\lim}_{\kappa\to\infty} C_2^6\Lambda \sum_{\tau=l_\kappa}^{\nu_\kappa-1} \rho_\tau^2 = 0 \qquad \text{a.s.} \tag{40}$$

Taking the lower limit for (33) and using (34), (35) and (38)–(40),

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

This last inequality proves the necessary condition D3 for the subsequences, which satisfies condition (27).

Now let us consider the case with

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

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

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

We consider the following two possibilities:

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

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

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

2. There exists a number K such that $x^\tau \in X$ for $l_\kappa \le \tau \le \nu_\kappa$, $\kappa \ge K$. For this case, the proof of condition D3 coincides with the proof 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 conside the subsequence x^{s_κ} such that $x^{s_\kappa} \to x^*$, $x^* \in X^*$. It follows from estimate (25) that

$$\|x^{s_\kappa+1} - x^{s_\kappa}\| \le \rho_{s_\kappa} \sum_{l=0}^{i(s_\kappa)-1} \lambda_{s_\kappa l}(\rho_{s_\kappa}\|H_0^{s_\kappa}\xi^{s_\kappa}\|\epsilon_{s_\kappa}^{-1} + C_2^3) . \tag{41}$$

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

$$\rho_{s_\kappa}\|H_o^{s_\kappa}\xi^{s_\kappa}\|\epsilon_{s_\kappa}^{-1} \to 0 ,$$

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

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

then (41) implies

$$\|x^{s_\kappa+1} - x^{s_\kappa}\| \to 0 \tag{42}$$

for almost all ω such that $x^{s_\kappa} \to X^*$, $x^* \in X^*$. The function $W(x)$ is continuous, thus (42) proves condition D5.

All conditions D1–D5 are checked and the theorem is proved. □

5 References

1. Betro, B. and L. De Biase: A Newton-like Method for Stochastic Optimization. In "*Towards Global Optimization 2*", North-Holland Publishing Company, 1978, pp. 269–289.

2. Dennis, J.N. and J.J. Moré: Quasi-Newton Methods, Motivation and Theory, *SIAM Review*, **19**, 1977, pp. 46–89.

3. Ermoliev, Yu.M.: Stochastic Quasi-Gradient Methods and Their Applications to Systems Optimization. *Stochastics*, **4**, 1983, pp. 1–37.

4. Gaivoronski, A.: Implementation of Stochastic Quasigradient Methods. In: "*Numerical Techniques for Stochastic Optimization*" (Yu. Ermoliev, R.J-B Wets Eds.) Springer-Verlag, 1988, pp. 313–353.

5. Gerencser, L.: Strong Consistency Theorems Related to Stochastic Quasi-Newton Methods. In: "*Stochastic Optimization*", Springer-Verlag Lecture Notes in Control and Information Science, **81**, 1984.

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

7. Kushner, H.J. and G. Jin: Stochastic Approximation Algorithms for Parallel and Distributed Processing. *Stochastics*, **22**, 1987, pp. 219–250.

8. Ljung, L. and T. Söderström: Theory and Practice of Recursive Identification. MIT Press, 1983.

9. Marti, K.: Descent Stochastic Quasigradient Methods. In: "*Numerical Techniques for Stochastic Optimization*" (Yu. Ermoliev, R.J-B Wets eds.), Springer-Verlag, 1988, pp. 393–401.

10. McAllister, P.M.: Adaptive Approaches to Stochastic Programming, this volume.

11. Novikova, N.M.: Some Stochastic Programming Methods in Hilbert Space. *Zhurnal Vichislitelnoj Matematiki i Matematicheskoj Phisiki*, 12, **25**, 1985, pp. 1795–1813.

12. Nurminski, E.A.: Numerical Methods for Solving Deterministic and Stochastic Minimax Problems. Naukova Dumka, Kiev, 1979, 159 p. (in Russian).

13. Pflug, G.: Stepsize Rules, Stopping Times and Their Implementation in Stochastic Quasigradient Algorithms. In: "*Numerical Techniques for Stochastic Optimization*" (Yu. Ermoliev, R.J-B Wets Eds.) Springer-Verlag, 1988, pp. 353–373.

14. Polyak, B.T. and Ya.Z. Tsypkin: Adaptive estimation algorithms: convergence, optimality, robustness. Automation and Remote Control, **3**, 1979, pp. 71–84.

15. Rockafellar, R.T. and R.J-B Wets: On the Interchange of Subdifferentiation and Conditional Expectation for Convex Functionals. *Stochastics*, **7**, 1982, pp. 173–182.

16. Ruszczynski, A. and W. Syski: A Method of Aggregate Stochastic Sub-Gradients with On-Line Stepsize Rules for Convex Stochastic Programming Problem. *Mathematical Programming Study*, **28**, part II, 1985, pp. 113–131.

17. Saridis, G.S.: Learning Applied to Successive Approximation Algorithms. *IEEE, Trans. Syst. Sci. Cybern.*, 1970, **SSC-6**, Apr., pp. 97–103.

18. Sarkison, D.J.: The Use of Stochastic Approximation to Solve the System Identification Problem. *IEEE Transactions on Automatic Control*, **AC-12**, 1967, pp. 563–567.

19. Uryas'ev, S.P.: Stochastic Quasi-Gradient Algorithms with Adaptively Controlled Parameters. Working Paper, IIASA, WP-86-32, 1986.

20. Uryas'ev, S.P.: Adaptive Algorithms of Stochastic Optimization and Game Theory. Nauka, Moskow, 1990 (to appear in Russian).

21. Wets, R. J-B: Modeling and Solution Strategies for a Constrained Stochastic Optimization Problems. *Annals of Operation Research*, **1**, 1984, pp. 3-22.