A STOCHASTIC QUASIGRADIENT ALGORITHM WITH VARIABLE METRIC

S.P. URYS'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.

© J.C. Baltzer AG, Scientific Publishing Company
2. Basic idea of the algorithm

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

\[
f(x) \rightarrow \min_{x \in \mathbb{R}^n},
\]

(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

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

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])

\[
\partial f(x) = \int_{\omega \in \Omega} \partial_x \varphi(x, \omega)P(d\omega),
\]

(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 \textit{stochastic quasigradients} [3]. To solve problem (1), the following algorithm is used:

\[
x^{s+1} = x^s - \rho_s H^s \xi^s.
\]

(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. the conditional mathematical expectation

\[
E_s \xi^s \overset{\text{def}}{=} E(\xi^s/x^s) \in \partial f(x^s),
\]

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

\[
\Phi_s(H) = E_s f(x^s - \rho_s H^s \xi^s).
\]

(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

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

(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 $H_0$ (see formula (2)):

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

where $\xi^T$ is the transposed vector $\xi^s$. We denote $\xi_0^s$ as some stochastic quasigradient at the point $x_0^s = x^s - \rho_s H_0^T \xi^s$, i.e.

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

One can see that

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

thus, $-\rho_s \xi_0^s \xi_0^T$ is a stochastic quasigradient of the function $\Phi_s(H)$ at the point $H_0^T$.

We consider that the matrix $H_0^T$ is known from the previous iterations $s-1$. To modify matrix $H_0^T$, we use the stochastic quasigradient method (see [3]):

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

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

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

(6)

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

$$
E(\xi_i^T / x_i^T) \in \partial f(x_i^T), \quad x_i^T = x^s - \rho_s H_i^T \xi^s.
$$

(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 $X \subseteq \mathbb{R}^n$ is known in advance such that $X^* \subseteq X$. This is not a serious restriction, since in practical situations such a set is usually known. This set could be very large. If $x^* \notin X$, then we assume that the approximation $x^i$ 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 $\{\epsilon_s\}$, $s = 0, 1, \ldots$, and $\{\lambda_{si}\}$, $s = 0, 1, \ldots, i = 0, 1, \ldots$, 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 could also be written for these sequences, but we do not want to overload the convergence proof with them now. The positive value $\epsilon_s$ defines $i(s)$ in the algorithm; iterations with respect to the matrix are stopped if $\rho_s \sum_{i=0}^{s-1} \lambda_{si} \geq \epsilon_s$. To avoid misunderstandings, we present here a full formal description of the algorithm.

**ALGORITHM 1**

**Step 1** Initialization:

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

**Step 2** Set $H_0^T = H_i^T$.

**Step 3** Set $i = 0$.

**Step 4** Compute the point $x_i^f$:

$$x_i^f = x^T - \rho_s H_i^T \xi^T.$$ 

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

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

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

**Step 8** If $x_i^T \in X$, then $x^{s+1} = x_i^T$, $\xi^{s+1} = \xi_i^T$; 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 $X^*$:

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

\[ \{x^\epsilon(\omega)\} \subset X \text{ a.s.} \]

**D2** \( W : X \rightarrow \mathbb{R} \) 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 \( \{x^{\epsilon_k(\omega)}(\omega)\} \) 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 \( \{\epsilon_k(\omega)\} \) such that

\[ W(x^\epsilon_k) \leq W(x'(\omega)) + \epsilon(\omega) \quad \text{for} \quad l_k(\omega) \leq \tau \leq \epsilon_k(\omega), \]

\[ \lim_{k \to \infty} W(x^{\epsilon_k(\omega)}(\omega)) \overset{\text{def}}{=} W(\omega) < W(x'(\omega)). \]

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

**D5** For almost all subsequences \( \{x^{\epsilon_k(\omega)}(\omega)\} \) such that \( \lim_{k \to \infty} x^{\epsilon_k(\omega)}(\omega) = x^*(\omega), \)

\[ x^*(\omega) \in X^*, \]

the condition

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

is satisfied.

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

**THEOREM 1**

Let the stochastic sequence \( \{x^\epsilon(\omega)\} \) satisfy conditions D1–D5; then \( x^\epsilon(\omega) \rightarrow X^* \) a.s., i.e. \( d(x^\epsilon(\omega), X^*) \rightarrow 0 \) 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 \( \{x^{\epsilon_k(\omega)}(\omega)\} \) with
\[ d(x_{\kappa}(\omega), x^*) > \delta > 0 \quad \text{for} \quad \kappa = 0, 1, \ldots \]

Further, we shall omit the argument \( \omega \). Since \( X \) is a compact set, then without loss of generality we can consider that \( x^* \to x' \in X^* \). Condition D3 implies the existence of the subsequence \( \{x_{\kappa}\} \) such that \( \lim_{\kappa \to \infty} W(x_{\kappa}^*) = \overline{W} < W(x') \). 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 \( m_{\kappa} < q_{\kappa} \), \( W(x_{m_{\kappa}}^*) \leq a \), \( W(x_{q_{\kappa}}^*) \geq b \) for \( \kappa = 0, 1, \ldots \), and \( a < W(x_s^*) \) for \( s \) satisfying inequalities \( m_{\kappa} < s \leq q_{\kappa} \), \( \kappa = 0, 1, \ldots \) (see fig. 1). The set

![Figure 1](image_url)

\( W(x') \)
\( b \)
\( a \)
\( \overline{W} \)
\( m_{\kappa} \) \( q_{\kappa} \) \( m_{\kappa+1} \) \( q_{\kappa+1} \)

\( X \) 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),

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

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 \( v_{\kappa} \) such that \( W(x_{v_{\kappa}}^*) \leq W(\overline{x}) + \varepsilon \) for \( \tau \) satisfying inequality \( m_{\kappa} \leq \tau \leq v_{\kappa} \), and

\[
\lim_{\kappa \to \infty} W(x^*) < W(\overline{x})
\]

By the selection of \( m_{\kappa} \) and \( q_{\kappa} \), without loss of generality we can consider that \( m_{\kappa} < v_{\kappa} < q_{\kappa} \). Since \( a < W(x_t^*) \) for \( m_{\kappa} < t < q_{\kappa} \).
\( a \leq \lim_{k \to \infty} W(x^k) < W(x). \)

Also, \( W(x^k) \leq a \) implies \( W(x) \leq a \). This contradiction proves the theorem. \( \square \)

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

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

let the sequences \( \{\lambda_{zd}\} \) and \( \{\epsilon_i\} \) be given and let \( \{\rho_s\} \) be a random sequence such that \( \rho_s \) depends upon the random vectors

\[
(x^0, x_i^1, x_i^2, \ldots, x_i^s, x_i^{s+1}, \ldots, x_i^s, 0 \leq s \leq \tau, 0 \leq i \leq i(\tau));
\]

let the stochastic quasigradients and algorithm satisfy the conditions

\[
\|\epsilon_0^s\| \leq C_2 \quad \text{a.s.}, \tag{9}
\]

\[
\|\epsilon_i^s\| \leq C_2 \quad \text{a.s.} \quad i = 1, \ldots, i(s); \quad s = 0, 1, \ldots, \tag{10}
\]

\[
\epsilon_s > 0, \quad s = 0, 1, \ldots, \tag{11}
\]

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

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

\[
\rho_s \|H_s\| \epsilon_s \|\epsilon_s^{-1} \to 0 \quad \text{a.s. for } s \to \infty, \tag{14}
\]

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

\[
\sum_{s=0}^{\infty} \rho_s^2 \leq \infty \quad \text{a.s.} \tag{16}
\]

\[
\lambda_{zd} > 0, \quad s = 0, 1, \ldots; \quad l = 0, 1, \ldots, \tag{17}
\]
\[ \sum_{l=0}^{\infty} \lambda_{il} = \infty, \quad s = 0, 1, \ldots, \tag{18} \]
\[ \sum_{l=0}^{\infty} \lambda_{il}^2 \leq \Lambda = \text{const}, \quad s = 0, 1, \ldots. \tag{19} \]

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

\[ \eta^s = \xi^s - g(x^s), \quad \eta^i_t = \xi^s_t - g(x^s_t). \]

\[ U_\varepsilon(x) = \{y \in \mathbb{R}^s : \|y - x\| \leq \varepsilon\}. \]

\[ f^* = \min_{x \in \mathbb{R}^s} f(x), \quad C_3 = \max_{x, y \in X} \|x - y\|, \tag{20} \]

\[ x^*_s = \arg \min_{x \in X^*} \|x^s - y\|, \quad x^*_t = \arg \min_{y \in X^*} \|x^s_t - y\|. \]

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

**LEMMA 1**

The following inequality is true:

\[ W(x^s_{i+1}) \leq W(x^s_i) + 2\rho_s C_3 \|H_0^2 \xi^s\| + \rho_e^2 \|H_0^2 \xi^s\|^2 + 2\rho_s \xi^s \|\sum_{l=0}^{i(t)-1} \lambda_{il} (f^* - f(x^s_l)) + 2\rho_s \xi^s \| \sum_{l=0}^{i(t)-1} \lambda_{il} (x^*_s - x^s_l, \eta^s_t) + \rho_e^2 C_2^2 \Lambda. \tag{21} \]
Proof

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

\[
W(x_{i(i)}^*) = \| x_{i(i)}^* - x_{i(i)}^f \|^2
\]

\[
\leq \| x_{i(i)}^* - x_{i(i)}^f \|^2 = \| x_{i(i)}^* - x^f + \rho_s H_{i(i)}^f \xi_{i(i)}^f \|^2
\]

\[
= \| x_{i(i)}^* - x^f + \rho_s (H_{i(i)}^f - H_{i(i)}^f) \xi_{i(i)}^f + \lambda_{x,i(i)} \xi_{i(i)}^f \|^2
\]

\[
= \| x_{i(i)}^* - x^f + \rho_s \lambda_{x,i(i)} \xi_{i(i)}^f \|^2 + \rho_s \lambda_{x,i(i)} \xi_{i(i)}^f \|^2
\]

\[
= \| x_{i(i)}^* - x_{i(i)}^f \|^2 + \rho_s \lambda_{x,i(i)} \xi_{i(i)}^f \|^2
\]

\[
\leq W(x_{i(i)}^f) + 2 \rho_s \lambda_{x,i(i)} \xi_{i(i)}^f \|^2 \leq W(x_{i(i)}^f) + 2 \rho_s \lambda_{x,i(i)} \xi_{i(i)}^f \|^2 \leq W(x_{i(i)}^f) + 2 \rho_s \lambda_{x,i(i)} \xi_{i(i)}^f \|^2
\]

Applying this estimate the proper amount of times, we obtain

\[
W(x_{i(i)}) \leq W(x_0^f) + 2 \rho_s \xi_{i(i)}^f \sum_{l=0}^{i(i)-1} \lambda_{x,l} (x_{x,l}^* - x_{x,l}^f, \xi_{x,l}^f) + \rho_s C_2 \sum_{l=0}^{i(i)-1} \lambda_{x,l}^2.
\]  \hspace{1cm} (22)

Estimate \( W(x_0^f) \) is as follows:

\[
W(x_0^f) \leq \| x_0^* - x_0^f \|^2 = \| x_0^* - x^f + \rho_s H_0^f \xi_0^f \|^2
\]

\[
= \| x_0^* - x^f \|^2 + 2 \rho_s \| x_0^* - x^f, H_0^f \xi_0^f \|^2 + \rho_s^2 \| H_0^f \xi_0^f \|^2
\]

\[
\leq W(x^f) + 2 \rho_s \| x_0^* - x^f \| H_0^f \xi_0^f \|^2 + \rho_s^2 \| H_0^f \xi_0^f \|^2.
\]

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

\[
(x_{x,l}^* - x_{x,l}^f, \xi_{x,l}^f) = (x_{x,l}^* - x_{x,l}^f, g(x_{x,l}^f)) + (x_{x,l}^* - x_{x,l}^f, \eta_{x,l}^f) \leq f^* - f(x_{x,l}^f) + (x_{x,l}^* - x_{x,l}^f, \eta_{x,l}^f).
\]

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

\[
W(x_{i(i)}) \leq W(x^f) + 2 \rho_s \| x_{i(i)}^* - x^f \| H_0^f \xi_{i(i)}^f \|^2 + \rho_s^2 \| H_0^f \xi_{i(i)}^f \|^2
\]

\[
+ 2 \rho_s \xi_{i(i)}^f \sum_{l=0}^{i(i)-1} \lambda_{x,l} (f^* - f(x_{x,l}^f)) + 2 \rho_s \xi_{i(i)}^f \sum_{l=0}^{i(i)-1} \lambda_{x,l} (x_{x,l}^* - x_{x,l}^f, \eta_{x,l}^f)
\]
\[ + \rho_s^2 C_2^{i(\gamma)-1} \sum_{l=0}^{i(\gamma)-1} \lambda_{sf}^2 \leq W(x^s) + 2\rho_s C_3 \| H_0^s \xi^s \| \\
+ \rho_s^2 \| H_0^s \xi^s \|^2 + 2\rho_s \| \xi^s \|^2 \sum_{l=0}^{i(\gamma)-1} \lambda_{sf} (f^* - f(x_l^s)) \]
\[ + 2\rho_s \| \xi^s \|^2 \sum_{l=0}^{i(\gamma)-1} \lambda_{sf} (x_l^* - x_l^s, \eta_l^s) + \rho_s^2 C_2 \Lambda, \] (23)

which proves the statement of the lemma. \( \Box \)

**Lemma 2**

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

\[ \max_{s \leq \gamma \leq m-1} \max_{0 \leq i \leq i(\gamma)} \| x^s - x^\gamma \| \leq \sum_{\tau=s}^{m-1} \rho_\tau (\rho_\tau \| H_0^\tau \xi^\tau \| \xi^\tau + C_2^2). \] (24)

**Proof**

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

\[ x^m = x_{i(m-1)} = x_{i(m-1)} - \rho_{m-1} H_{i(m-1))} \xi^{m-1} \]
\[ = x^m - \rho_{m-1} (H_{i(m-1))} + \lambda_{m-1, i(m-1))} \xi^{m-1} \xi^{m-1}) \xi^{m-1} \]
\[ = (x^m - \rho_{m-1} H_{i(m-1))} - \rho_{m-1} \lambda_{m-1, i(m-1))} \xi^{m-1}) \xi^{m-1} \]
\[ = x_{i(m-1))} - \rho_{m-1} \lambda_{m-1, i(m-1))} \xi^{m-1} \xi^{m-1} \| \xi^{m-1} \|^2. \] (25)

Using this equality the proper amount of times, we obtain

\[ x^m = x_{0}^{m-1} - \rho_{m-1} \xi^{m-1} \| \xi^{m-1} \|^2 \sum_{l=0}^{i(m-1))} \lambda_{m-1, l} \xi^{m-1} \]
\[ = x^m - \rho_{m-1} H_0^{m-1} \xi^{m-1} - \rho_{m-1} \xi^{m-1} \| \xi^{m-1} \|^2 \sum_{l=0}^{i(m-1))} \lambda_{m-1, l} \xi^{m-1}. \]

Since \( x^s, x_{i(\gamma)}, \ldots, x_{i(m-1))} \in X, \ m > s, \) again applying this formula for \( m-1, \ldots, s+1, \) we obtain

\[ x^m = x^s - \sum_{\tau=s}^{m-1} \rho_\tau H_0^\tau \xi^\tau - \sum_{\tau=s}^{m-1} \rho_\tau \| \xi^\tau \|^2 \sum_{l=0}^{i(\gamma)-1} \lambda_{\tau, l} \xi^\tau. \] (26)
In view of conditions (9) and (10) of the theorem, step 6 of the algorithm, and the last equality, we can estimate

\[
\|x^m - x^r\| \leq \sum_{\tau=s}^{m-1} \rho_{\tau} \|H_0^r \xi_{\tau}^r\| + \sum_{\tau=s}^{m-1} \rho_{\tau} \|\xi_{\tau}^r\|^2 \sum_{l=0}^{i(\tau)-1} \lambda_{\tau,l} \|\xi_{\tau,l}^r\|
\]

\[
\leq \sum_{\tau=s}^{m-1} \left( \rho_{\tau} \|H_0^r \xi_{\tau}^r\| + C_2^3 \rho_{\tau} \sum_{l=0}^{i(\tau)-1} \lambda_{\tau,l} \right)
\]

\[
= \sum_{\tau=s}^{m-1} \rho_{\tau} \sum_{l=0}^{i(\tau)-1} \lambda_{\tau,l} \left( \rho_{\tau} \|H_0^r \xi_{\tau}^r\| \left( \rho_{\tau} \sum_{l=0}^{i(\tau)-1} \lambda_{\tau,l} \right)^{-1} + C_2^3 \right)
\]

\[
\leq \sum_{\tau=s}^{m-1} \rho_{\tau} \sum_{l=0}^{i(\tau)-1} \lambda_{\tau,l} (\rho_{\tau} \|H_0^r \xi_{\tau}^r\| \epsilon_{\tau}^{-1} + C_2^3).
\]  

(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 \(\{x^k\}\) with

\[
x^k \to x', \ W(x') \geq 0, \ U_{\delta}(x') \subset X,
\]

where \(\delta\) is some positive random value for almost all \(\omega \in B\).

LEMMA 3

If \(W(x') \geq 0, \ U_{\delta}(x') \subset X\), then for any random value \(\epsilon\) satisfying inequality \(0 < \epsilon < \sqrt{W(x')}\) there exists a subsequence \(\{v_k\}\) such that

\[
\lim_{k \to \infty} W(x'^k) < W(x') \quad \text{a.s.,}
\]

(29)

and

\[
\limsup_{k \to \infty} \max_{0 \leq \tau \leq \max \{v_k^{-1} \leq \tau \leq i(\tau)\}} \max_{0 \leq s \leq i(\tau)} \|x^s - x^k\| \leq q \quad \text{a.s.}
\]  

(30)

Proof

Denote \(\epsilon\) as some random value such that \(0 < \epsilon < \sqrt{W(x')}\) for \(\omega \in B\). We define the index subsequence \(\{v_k\}\) (this subsequence depends upon \(\omega\)) such that

\[
C_2^3 \sum_{\tau=v_k}^{v_k^{-1}} \epsilon_{\tau} \to q \quad \text{a.s.};
\]

(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,

\[ \sum_{\tau = l_{k}}^{v_{k} - 1} \rho_{\tau} \sum_{l = 0}^{i(\tau) - 1} \lambda_{ui} C_{2}^{3} \rightarrow q \text{ a.s.} \]  

(32)

Since \( \rho_{\tau} \| H_{0}^{\xi} \| e_{\tau}^{-1} \rightarrow 0 \) a.s. for \( \tau \to \infty \) (see condition (14)), lemma 2 and (32) imply inequality (30). From (30) and \( x^{(x_{k})} \to x' \) for \( \kappa \to \infty \), it follows that the approximations \( x_{i}^{(x_{k})} \), \( l_{k} \leq \tau \leq v_{k} - 1, 0 \leq l \leq i(\tau) \), belong to the set \( U_{2q}(x') \) for sufficiently large numbers \( \kappa \) (this \( \kappa \) depends upon \( \omega \)). Since

\[ 2q = \min\{c, \delta\} \leq \epsilon < \sqrt{W(x')} = \sqrt{\min_{y \in X} \| x' - y \|^{2}} = \min_{y \in X} \| x' - y \|, \]

then

\[ X^{\ast} \cap U_{2q}(x') = \emptyset. \]  

(33)

It also follows from (28) that

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

Since the points \( x^{(x_{k})}_{i} \) for \( l_{k} \leq \tau \leq v_{k} - 1, 0 \leq l \leq i(\tau) \), belong to the set \( U_{2q}(x') \) for sufficiently large \( \kappa \), (33) implies the existence of a random value \( \alpha > 0 \) a.s. such that

\[ f^{\ast} - f(x^{(x_{k})}_{i}) \leq -\alpha \quad \text{for} \quad l_{k} \leq \tau \leq v_{k} - 1, 0 \leq l \leq i(\tau), \]  

(34)

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

\[ W(x^{v_{k}}) = W(x^{v_{k} - 1}) \leq W(x^{v_{k} - 1}) + 2 \rho_{v_{k} - 1} C_{3} \| H_{0}^{v_{k} - 1} e_{v_{k} - 1} \| \]

\[ + \rho_{v_{k} - 1} H_{0}^{v_{k} - 1} e_{v_{k} - 1} \| \sum_{l = 0}^{i(v_{k} - 1) - 1} \lambda_{v_{k} - 1, l} \alpha \]

\[ + 2 \rho_{v_{k} - 1} \| e_{v_{k} - 1} \| \sum_{l = 0}^{i(v_{k} - 1) - 1} \lambda_{v_{k} - 1, l} (x^{(x_{k})}_{v_{k} - 1, l} - x^{(x_{k})}_{v_{k} - 1, l}) \]

\[ + \rho_{v_{k} - 1} C_{2}^{6} \Lambda \leq W(x^{v_{k}}) + 2 \sum_{\tau = l_{k}}^{v_{k} - 1} \rho_{\tau} C_{3} \| H_{0}^{\xi_{\tau}} \| \sum_{\tau = l_{k}}^{v_{k} - 1} \rho_{\tau} \| H_{0}^{\xi_{\tau}} \|. \]
\[-2 \sum_{\tau = l_k}^{v_k - 1} \rho_{\tau} \alpha \xi^{\tau} \| \xi^{\tau} \|^2 + \sum_{l=0}^{v_k-1} \lambda_{\tau} \| \xi^{\tau} \|^2 \sum_{l=0}^{v_k-1} \lambda_{\tau} (x^{\tau}_l - x^*_l, \eta^{\tau}_l) + C_2^6 \Lambda \sum_{\tau = l_k}^{v_k - 1} \rho_{\tau}^2 \overset{\text{def}}{=} W(x^k) + T_2 + T_3 + T_4 + T_5 + T_6. \tag{35}\]

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

\[
\lim_{K \to \infty} T_2 = \lim_{K \to \infty} 2 \sum_{\tau = l_k}^{v_k - 1} \rho_{\tau} C_3 \| H_0^T \xi^{\tau} \|
= 2C_3 \lim_{K \to \infty} \left( \sum_{\tau = l_k}^{v_k - 1} \varepsilon^{\tau}_t \right) \left( \sum_{\tau = l_k}^{v_k - 1} \rho_{\tau} H_0^T \xi^{\tau} \right)
= 2C_3 \lim_{K \to \infty} \left( \sum_{\tau = l_k}^{v_k - 1} \varepsilon^{\tau}_t \right) \lim_{K \to \infty} \left( \sum_{\tau = l_k}^{v_k - 1} \rho_{\tau} H_0^T \xi^{\tau} \right)
= 2C_3 (C_2^6 q) \cdot 0 = 0 \quad \text{a.s.} \tag{36}\]

For term $T_3$:

\[
\lim_{K \to \infty} T_3 = \lim_{K \to \infty} 2 \sum_{\tau = l_k}^{v_k - 1} \rho_{\tau}^2 \| H_0^T \xi^{\tau} \|^2 \leq \lim_{K \to \infty} \left( \sum_{\tau = l_k}^{v_k - 1} \rho_{\tau} H_0^T \xi^{\tau} \right)^2
= (2C_3)^{-2} \lim_{K \to \infty} T_2^2 = 0 \quad \text{a.s.} \tag{37}\]

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

\[
\lim_{K \to \infty} T_4 = \lim_{K \to \infty} \left( -2\alpha \sum_{\tau = l_k}^{v_k - 1} \rho_{\tau} \| \xi^{\tau} \|^2 \sum_{l=0}^{v_k-1} \lambda_{\tau} \right)
\leq -2\alpha \lim_{K \to \infty} \left( \sum_{\tau = l_k}^{v_k - 1} \varepsilon^{\tau}_t \right) \sum_{l=0}^{v_k-1} \lambda_{\tau} \left( \sum_{\tau = l_k}^{v_k - 1} \varepsilon^{\tau}_t \right)^{-1} \varepsilon^{\tau}_t \| \xi^{\tau} \|^2
\leq -2\alpha \lim_{K \to \infty} \left( \sum_{\tau = l_k}^{v_k - 1} \varepsilon^{\tau}_t \right) \left( \sum_{\tau = l_k}^{v_k - 1} \varepsilon^{\tau}_t \right)^{-1} \| \xi^{\tau} \|^2
\]
\[ \leq -2\alpha \lim_{\kappa \to \infty} \left( \sum_{\tau=1}^{\nu_{\kappa}-1} \varepsilon_\tau \right)^{-1} \lim_{\kappa \to \infty} \left\| \sum_{\tau=1}^{\nu_{\kappa}-1} \varepsilon_\tau x^\tau \right\|_2^2 \]

\[ = -2\alpha C_2^{-3} q^{-1} \lim_{\kappa \to \infty} \left\| \sum_{\tau=1}^{\nu_{\kappa}-1} \varepsilon_\tau x^\tau \right\|_2^2. \quad (38) \]

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

\[ \lim_{\kappa \to \infty} \left\| \sum_{\tau=1}^{\nu_{\kappa}-1} \varepsilon_\tau x^\tau \right\|_2^2 = \lim_{\kappa \to \infty} \left\| \sum_{\tau=1}^{\nu_{\kappa}-1} \varepsilon_\tau g(x^\tau) + \sum_{\tau=1}^{\nu_{\kappa}-1} \varepsilon_\tau \eta^\tau \right\|_2^2 \]

\[ = \lim_{\kappa \to \infty} \left\| \sum_{\tau=1}^{\nu_{\kappa}-1} \varepsilon_\tau g(x^\tau) \right\|_2^2. \quad (39) \]

For sufficiently large \( \kappa \), the points \( y^\tau, l_\kappa \leq \tau \leq \nu_\kappa - 1, 0 \leq l \leq i(\tau) \), belong to the convex set \( U_{2q}(x') \) and \( U_{2q}(x') \cap X^* = \emptyset \). 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

\[ \left\| \sum_{\tau=1}^{\nu_{\kappa}-1} \varepsilon_\tau g(x^\tau) \right\|_2^2 < \left( \| x^* - x' \|^{-1} \left\langle \sum_{\tau=1}^{\nu_{\kappa}-1} \varepsilon_\tau g(x^\tau), x^* - x' \right\rangle \right)^2 \]

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

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

\[ \lim_{\kappa \to \infty} T_4 \leq -2\alpha C_2^{-3} q^{-1} \lim_{\kappa \to \infty} \left( \gamma \sum_{\tau=1}^{\nu_{\kappa}-1} \varepsilon_\tau \right)^2 = -2\alpha C_2^{-3} q^{-1} \gamma^2 C_2^{-6} q^{-2} < 0 \text{ a.s.} \quad (40) \]

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

\[ \sum_{\tau=0}^{\infty} \sum_{l=0}^{\infty} \rho_l^2 \| e^\tau \|_4 \lambda_l^2 \leq C_2 \Lambda \sum_{\tau=0}^{\infty} \rho_l^2 < \infty \text{ a.s.;} \]

consequently, the martingale series
\[
\sum_{\tau=0}^{\infty} \rho_{\tau} \|z_\tau^*\|^2 \sum_{l=0}^{i(\tau)-1} \lambda_{l\tau}(x_{il}^* - x_{l\tau}^*, \eta_{l\tau}^*)
\]
is convergent. This fact implies
\[
\lim_{K \to \infty} T_5 = \lim_{K \to \infty} T_5 = 2 \lim_{K \to \infty} \sum_{l=1}^{v_{K-1}} \rho_{l \tau} \|z_\tau^*\|^2 \sum_{l=0}^{i(\tau)-1} \lambda_{l\tau}(x_{il}^* - x_{l\tau}^*, \eta_{l\tau}^*) = 0 \text{ a.s.} \quad (41)
\]
We have from condition (16) also that
\[
\lim_{K \to \infty} T_6 = \lim_{K \to \infty} C_2^6 \Lambda \sum_{l=1}^{v_{K-1}} \rho_{l \tau}^2 = 0 \text{ a.s.} \quad (42)
\]
Taking the lower limit for (35) and using (36), (37) and (40)–(42),
\[
\lim_{K \to \infty} W(x_{K}^*) \leq \lim_{K \to \infty} W(x^*) - 2\alpha \gamma^2 q C_2^{-9} < W(x^*) - 2\alpha \gamma^2 q C_2^{-9}.
\]
The last inequality implies (29) and the lemma is proved. \(\square\)

Lemma 3 proves condition D3 for the subsequences that satisfy condition (28). Now let us consider the case with
\[
x_{K}^* \to x', \quad x' \in \partial X,
\]
where \(\partial X\) is the boundary of the set \(X\). As in the previous case, we define the index subsequence \(\{v_k\}\) such that
\[
C_2^3 \sum_{l=1}^{v_{K-1}} \varepsilon_l \to q = 2^{-1} \varepsilon.
\]
We consider the following two possibilities:

(1) There exists an infinite subsequence \(\{\theta_m\}\) such that \(l_m \leq \theta_m < v_m, \ x^{\theta_m} \in X, \ x^{\theta_m}_{i(\theta_m)} \notin X, \ x^{\theta_m+1} = x^0\). Then condition (8) implies
\[
\lim_{m \to \infty} W(x^{\theta_m+1}) = \|x_0^* - x^0\| \leq C_1^2 \leq W(x')
\]
and subsequence \(\{x^{\theta_m+1}\}\) satisfies the sufficient condition D3.

(2) There exists a number \(K\) such that \(x^t \in X\) for \(l_k \leq t \leq v_k, \ k \geq K\). 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 $x^\omega$ such that $x^\omega \to x^*$, $x^* \in X^*$. It follows from estimate (27) that

$$
\|x^{i+1}_\omega - x^\omega\| \leq \rho \sum_{l=0}^{i-1} \lambda^l \left( \rho \|H_0 \xi e_{\omega, l} \| e_{\omega, l}^{-1} + \epsilon^3 \right).
$$

(43)

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

$$
\rho \|H_0 \xi e_{\omega, l} \| e_{\omega, l}^{-1} \to 0,
$$

$$
\left| \rho \sum_{l=0}^{i-1} \lambda^l e_{\omega, l} \right| \to 0,
$$

$$
\epsilon_{\omega, l} \to 0,
$$

(43) implies

$$
\|x^{i+1}_\omega - x^\omega\| \to 0
$$

(44)

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

References


