International Institute for Applied Systems Analysis
IIASA www.iiasa.ac.at

Stochastic Quasi-Gradient Algorithms with Adaptively Controlled Parameters

Uryasev, S.P.

IIASA Working Paper

WP-86-032

July 1986

Working Paper

STOCHASTIC QUASI-GRADIENT ALGORITHMS WITH ADAPTIVELY CONTROLLED PARAMETERS

S.P. Urjas'ev

July 1986
WP-86-32

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

STOCHASTIC QUASI-GRADIENT ALGORITHMS WITH ADAPTIVELY CONTROLLED PARAMETERS

S.P. Urjas'ev

July 1986
WP-86-32

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
2361 Laxenburg, Austria

FOREWORD

The paper deals with choosing stepsize and other parameters in stochastic quasi-gradient methods for solving convex problems of stochastic optimization. The principal idea of methods consists in using random estimates of gradients of the objective function to search for the point of extremum. To control algorithm parameters the iterative adaptive procedures are suggested which are quasi-gradient algorithms with respect to parameters. The convergence is proved and the estimates of the rate of convergence of such algorithms are given. The results of computations for several stochastic optimization problems are considered. The paper is part of the research on numerical techniques for stochastic optimization conducted in the Adaptation and Optimization project of the System and Decision Sciences program.

Alexander B. Kurzhanski
Chairman
System and Decision Sciences Program

CONTENTS

STOCHASTIC QUASI-GRADIENT ALGORITHMS WITH ADAPTIVELY CONTROLLED PARAMETERS

S.P. Urjas'ev

Institute of Cybernetics
Academy of Sciences of the Ukr.SSR
252207 Kiev 207, USSR

INTRODUCTION

The paper is devoted to the development of iterative non-monotine optimization algorithms for problems of convex stochastic optimization with and without constraints. Most problems under discussion feature the lack of complete information about objective and constraint functions and their derivatives as well as non-smooth nature of these functions. The central idea of the discussed numerical methods, called the stochastic quasi-gradient methods, consists in the use of random directions instead of precise values of gradients. The random directions are statistical estimates of gradients (stochastic quasi-gradients). The definition of a stochastic quasi-gradient was introduced in the work by Ju. M. Ermoliev, Z.V. Nekrylova [1] and then this concept was developed in works by Ju. M. Ermoliev (see, e.g., [2], [3]).

Stochastic approximation algorithms (which stem from the work by H. Robbins and S. Monro [4]) and many random search algorithms which are represented in the work by L.A. Rastrigin [5] and others are special cases of stochastic quasi-gradient algorithms.

Adaptive procedures are offered and studied through the use of which the parameters of the algorithms discussed in this paper are controlled and practical characteristics of these algorithms are improved. By the adaptivity is here meant the dependence of these parameters upon the process trajectory in distinction to program procedures where parameters depend upon the number of iteration only. In particular, step size control and stopping criteria are suggested for the stochastic algorithms. It should be emphasized that it is just the aspects which are most difficult and problematic in the numerical implementation of these methods.

The main point of the suggested approach consists in the following. Almost each iteration algorithm has some parameters to be controlled. Usually there are also criteria which define the quality of the chosen controls. But it is difficult to satisfy these criteria in practice (to find optimal control) because these quality criteria are difficult to compute. Nevertheless it is possible to vary these criteria by parameters and to calculate their gradients or stochastic quasi-gradients. The obtained gradients (quasi-gradients) may be used in construction of recurrence procedures to modify these parameters. In such approach several gradient procedures operate in the algorithm – in the main space and with respect to the algorithm parameters. I.e., the adaptation of the algorithm parameters occurs.

The following designations will be used:

R^n is an n-dimensional Euclidean space;

$\langle \cdot, \cdot \rangle$ is an inner product in R^n;

$\|\cdot\|$ is a norm in R^n;

$\partial f(x)$ is a subdifferential of the convex function $f: R_n \to R$ at point x, i.e.

$$\partial f(x) = \{g \in R^n : f(y) \geq f(x) + \langle g, y - x \rangle \text{ for all } y \in R^n\};$$

$(\Omega, \tilde{f}, P)$ is a probability space on which all random values are defined;

w is an elementary event belonging to the set Ω;

a.s. means "almost surely";

$E\xi$ is a mathematical expectation for the random value ξ;

$E[\xi / \tilde{f}_s]$ or

$E_s \xi$ is a conditional mathematical expectation with respect to the δ-field $\tilde{f}$;

$\Pi_x(\cdot)$ is a projection on a convex closed set $X \subset R^n$.

1. QUASI-GRADIENT ALGORITHM WITH ADAPTIVE PARAMETER CONTROL, NON-FORMAL DESCRIPTION AND A BRIEF REVIEW OF RESULTS

Here we shall consider the problem of minimizing a convex (possibly non-smooth) function $f(x)$

$$f(x) \to \min_{x \in X}$$

where X is a convex compact subset of R^n. In the considered class of problems in-

stead of exact values of gradients or generalized gradients of the function $f(x)$, the vectors are known which are statistical estimates of these quantities while 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

$$f(x) = E_w \varphi(x, w) = \int_{w \in \Omega} \varphi(x, w) P(dw)$$

Considering that under the most general assumptions the generalized differential of the convex function $f(x)$ is calculated by the formula [6]

$$\partial f(x) = \int_{w \in \Omega} \partial_x \varphi(x, w) P(dw) \ ,$$

$\partial_x \varphi(x, w)$ is then a set of vectors being the statistical estimates of gradients of the function $f(x)$.

EXAMPLE *A Random Location Equilibrium Problem.*

The classic formulation of Weber problem is as follows: given are n points w_i, $i = 1, \ldots, n$ in two-dimensional Euclidean space R^2, it is required to find a point $x \in R^2$ such that a sum of distances to all points $w_i \in R^2$, $i = 1, \ldots, n$ is minimal. In the generalized statement [7] each point w_i $i = 1, \ldots, n$ is assumed to be a random value specified by some probability measure $\Theta_i(w)$ on R^2. The problem consists in finding the point $x \in R^2$ which minimizes the sum of mathematical expectations for distances from the point x to points w_i, $i = 1, \ldots, n$

$$f(x) = \sum_{i=1}^{n} \beta_i \int_{R^2} \int \| x - w \| \Theta_i(dw) = \int_{R^2} \int \| x - w \| \sum_{i=1}^{n} \beta_i \Theta_i(dw) =$$

$$\int_{R^2} \int \sum_{i=1}^{n} \beta_i \| x - w \| \Theta(dw), \qquad \Theta = \sum_{i=1}^{n} \beta_i \Theta_i \Big/ \sum_{i=1}^{n} \beta_i$$

where $\beta_i > 0$, $i = 1, \ldots, n$.

The random function

$$\xi(x, w) = \begin{cases} \dfrac{x - w}{\| x - w \|} \sum_{i=1}^{n} \beta_i & \text{for } x \neq w, \\ 0 & \text{for } x = w, \end{cases}$$

where the random value w is specified by the probability measure Θ may be taken as statistical estimate of generalized gradients i.e.

$\int\int_{R^2} \xi(x, w)\, \Theta(dw) \in \partial f(x)$ is satisfied.

1.1. Description of the Algorithm and Various Approaches to Step Size Control

An unknown point of the minimum of the convex function $f(x)$ on the set X is estimated by the recurrent sequence [2]

$$x^{s+1} = \Pi_X(x^s - \rho_s \xi^s),\ s = 0, 1, \ldots \tag{1}$$

where X is a convex closed set in R^n; ξ^s is a stochastic quasi-gradient, i.e. a conditional mathematical expectation for this vector satisfies the relation

$$E_s \xi^s \in \partial f(x) + b^s\ ;$$

σ-field $\tilde{f}_s$ is specified by random vectors $(x^0, \xi^0, x^1, \xi^1, \ldots, x^s)$; ρ_s, $s = 0, 1, \ldots$ is some sequence of random values or random matrices $n \times n$.

The algorithm suggested by H. Robbins and S. Monro [4] for estimation of the root of a regression function (for $X = R^1$) is a special case of the algorithm (1). The algorithm suggested for optimization problems in [8] by H. Kiefer and J. Wolfowitz is also a special case of the algorithm (1). In his work the gradient estimate is taken as finite-difference approximations with random estimates of the objective function. Further this approach was developed in works of many authors (see, e.g., [9–11]) with the assumption of smoothness of the objective function and the absence of the dependence of parameters ρ_s, $s = 0, 1, \ldots$ on the process trajectory. Such parameter controls we will call program controls, i.e., the step size is equal to the constant or decreases monotonically by a pre-specified rule depending upon the number of iteration s.

Algorithms of the type (1) for optimization of different classes of non-smooth functions are described in [2], [12], [13]. Various approaches to estimating the rate of convergence of schemes of type (1) with program step size controls are presented fairly completely in [10], [11], [14–17]. In practice it may occur that the initial step size is chosen small and to increase the rate of convergence it should be enlarged. The program step size control, naturally, does not take into account such situation though from the viewpoint of the asymptotic rate of convergence this control may be ideal. That is why adaptive step size controls taking into account the behavior of the objective function are necessary which would enlarge

the step size far from the extremum if it is small and decrease it near the point of the minimum. R.J.-B. Wets proposed the stochastic quasi-Newton method [25], then these results were developed by A. Gaivoronski. Below we consider the different approach.

The first step in this direction was made by H. Kesten [18] who suggested the program-adaptive control. He suggested to choose in scheme (1), as step sizes ρ_s, a prespecified sequence $\{a_k\}$ which satisfies conditions

$$\sum_{0}^{\infty} a_k = \infty; \; a_k > 0, \; k = 0,1,...$$

but the step should be changed not at each iteration but only in the case when $<\xi^s, \xi^{s-1}> < 0$.

In theoretical studies on substantiation of stochastic quasi-gradient methods conducted at V.M. Glushkov Institute of Cybernetics (see, e.g., the generalizing work [2]) step sizes ρ_s are assumed to be dependent upon the process trajectory $(x^0, \ldots, x^s)$. In application studies beginning in 1967 heuristic adaptive procedures were used to control step sizes in the algorithm (1). At each iteration the unbiased estimate z_s of the objective function $f(x^s)$ is assumed to be known; denote

$$T_s = \frac{1}{k} \sum_{j=s-k+1}^{s} z_j$$

The value T_s may be used for dialogue or program step size control [2] (see also [19]). For example, the step size may be chosen according to the rule

$$\rho_{s+1} = \begin{cases} \rho_s / 2 & \text{if } |T_{s-k} - T_s| \le \delta, \\ \rho_s & \text{otherwise} \end{cases}$$

A. M. Gupal and F. Mirzoahkmedov [20] suggested to change the step size ρ_s according to the norm of vectors v^s

$$v^{s+1} = v^s + a_s(\xi^s - v^s), \; 0 < a_s \le 1, \; v^0 = 0 \; ,$$

which are convex combinations of previous stochastic quasi-gradients ξ^i, $i = 0,1,\ldots,s$. For stochastic problems of quadratic programming the step size controls which are the development of H. Kesten scheme were suggested and justified by G. Pflug [21].

The failing of the above-listed step size controls, except for dialogue ones, consists in a high dependence of the efficiency of algorithm operation upon the value of the initial step size ρ_0 since the step size can only decrease during the iteration process. In the scheme suggested and substantiated by the author in [22], [23] the step size not only can decrease but also increase. In the next section it is shown that this rule is a result of using the stochastic quasi-gradient algorithm to control this parameter.

2. USE OF STOCHASTIC QUASI-GRADIENT ALGORITHMS FOR STOCHASTIC ALGORITHM PARAMETER CONTROLS

Parameter controls in stochastic algorithms is usually difficult because of the absence of objective function values since only statistical estimates of these values are available. This circumstance does not make possible, for example, the realization of efficient procedure of search for the function minimum along some chosen direction. The suggested approach consists in using the gradient algorithms for parameter controls. To use such procedures there is no need for additional computations of the objective function or its gradients.

2.1. Step Size Control for Stochastic Quasi-Gradient Algorithm [22], [23]

When constructing adaptive step size control for the algorithm (1) we assume that the algorithm trajectory belongs to the interior of the admissible domain and $b^s = 0$, $s = 0, 1, \ldots$ i.e. $E_s \xi^s \in \partial f(x^s)$. In the algorithm (1) it is natural to take step sizes ρ_s as the point of minimum of the function $\Phi_s(\rho)$ with respect to ρ where

$$\Phi_s(\rho) = E_s f(x^s - \rho \xi^s) \ .$$

Usually it is difficult to calculate the values of the function $\Phi_s(\rho)$. Let us differentiate the function $f(x^s - \rho \xi^s)$ with respect to ρ at point ρ_s

$$\partial_\rho f(x^s - \rho_s \xi^s) = - \{ \langle y, \xi^s \rangle : y \in \partial f(x^{s+1}) = \partial f(x^s - \rho_s \xi^s) \} \ .$$

Since

$$\partial \Phi_s(\rho_s) = E_s \partial f(x^s - \rho_s \xi^s) \ ,$$

then $-E_s \langle \xi^{s+1}, \xi^s \rangle \in \partial \Phi_s(\rho_s)$.

To modify the step size ρ_s we may use the following gradient procedure

$$\rho_{s+1} = \rho_s + \lambda_s \langle \xi^{s+1}, \xi^s \rangle = \rho_s - \frac{\lambda_s}{\rho_s} \langle \xi^{s+1}, \Delta x^{s+1} \rangle ,$$

$$\lambda_s > 0, \; s = 0, 1, \ldots$$

where $\Delta x^{s+1} = x^{s+1} - x^s$.

To facilitate the proof of the algorithm convergence we rewrite the last relation in the form

$$\rho_{s+1} = \min(\bar{\rho}, \; \alpha_s^{-\langle \xi^{s+1}, \Delta x^{s+1} \rangle - \delta \rho_s}), \; \alpha_s > 1, \; \delta > 0 , \qquad (2)$$

the constant $\bar{\rho}$ bounds the step size above.

Note that the exponent is supplemented with the additional term $-\delta \rho_s$ which decreases the step size ρ_s. Here δ is some sufficiently small constant, therefore the additional decrease of the step size occurs in the case when the value $\langle \xi^{s+1}, \xi^s \rangle$ is sufficiently close to zero and is comparable to the value δ. The formula (2) may be interpreted in the following manner. The value $\langle \xi^{s+1}, \Delta x^{s+1} \rangle$ gives some information about whether the minimum of the function $\Phi_s(\rho)$ with respect to ρ was passed through at the iteration or not. If $-\langle \xi^{s+1}, \Delta x^s \rangle > 0$ then with a high probability the minimum was not passed through and the step size increases due to the member $\langle \xi^{s+1}, \Delta x^{s+1} \rangle$, otherwise the step size decreases. In [23] the Cesàro convergence of the algorithm (1), (2) was proved, i.e., the convergence to the optimal set with probability 1 of the sequence

$$\bar{x}^s = \sum_{l=0}^{s} \rho_l x^l \Big/ \sum_{l=0}^{s} \rho_l ,$$

which is a convex combination of the trajectory points [16]. In this paper the convergence of the algorithm (1), (2) with probability 1 is proved, and the asymptotic estimate of the rate of the algorithm convergence for the case of twice differentiable function $f(x)$ is obtained.

2.2. Stochastic Quasi-Gradient Algorithm with Variable Metric

Algorithms of the type (1) in the case when the function is ill-conditioned have the low practical rate of convergence. This forces to use more complex variants of the algorithms. In non-linear programming a wide spectrum of algorithms is developed, called the algorithms of variable metric [24] which successfully operate in such situations. In the given case, however, the direct use of these algorithms is impossible because only statistical estimates of values of the objective function and of its gradients are known.

Let it be required to minimize a convex possibly non-smooth function $f(x)$ specified on the space R^n. Stochastic quasi-gradients of the function are known. Approximations of the extremum point are considered by the rule

$$x^{s+1} = x^s - H^s \xi^s, \; s = 0,1,... \tag{3}$$

where H^s, $s = 0,1,...$ is a sequence of $n \times n$ random square matrices; ξ^s, $s = 0,1,...$ is a sequence of stochastic quasi-gradients, i.e. $E_s \xi^s \in \partial f(x^s)$, here δ-field $\tilde{f}$ is specified by random values $(x^0, \xi^0, H^0, x^1, \xi^1, H^1, \ldots, x^s)$. The matrix H^s is modified at each step in the following manner

$$H^s = Q^{s-1} H^{s-1} \; ,$$

where Q^s, $s = 0,1,...$ is a sequence of square matrices.

Denote $\varphi_s(Q) = f(x^s - QH^s \xi^s)$. The matrix Q^s at the iteration s can be chosen from the condition of the minimum of the following function of $n \times n$ variables

$$\Phi_s(Q) = E_s \varphi_s(Q) \; .$$

However this problem by complexity is equivalent to the source problem.

We calculate the stochastic quasi-gradient of the function $\Phi_s(Q)$ at point I where I is a unitary matrix.

We differentiate the function $\varphi_s(Q)$ in a generalized sense with respect to Q at the point I

$$\partial_Q \varphi_s(Q) = -\{y \bar{\xi}^s \bar{H}^s : y \in \partial f(x^{s+1})\}$$

Here $\bar{H}^s$, $\bar{\xi}^s$ denote the transposed matrix H^s and the vector column ξ^s.

Since

$$\partial_Q \Phi_s(Q) = E_s \partial_Q \varphi_s(Q) = -E_s\{y \bar{\xi}^s \bar{H}^s : y \in \partial f(x^{s+1})\}$$

then

$$-E_s \xi^{s+1} \bar{\xi}^s \bar{H}^s \in \partial_Q \Phi_s(Q) \ .$$

As a matrix Q^s we may take the matrix which is formed when executing one step from the point I in the direction of the stochastic quasi-gradient $\xi^{s+1} \bar{\xi}^s \bar{H}^s$, i.e.

$$Q^{s+1} = I + \gamma_s \xi^{s+1} \bar{\xi}^s \bar{H}^s \ ,$$

where γ_s is a positive scalar.

Then we may rewrite the formula for the matrix modification in the following manner

$$H^{s+1} = (I + \gamma_s \xi^{s+1} \bar{\xi}^s \bar{H}^s) H^s \ .$$

Note that the last formula is close to the method with dilatation of the space along the generalized gradient suggested by N. Z. Shor [26, p. 92] with V. A. Skokov modification.

2.3. Algorithm with the Averaging of Stochastic Quasi-Gradients

The algorithm with the averaging of stochastic quasi-gradients was considered by many authors [27], [12], [28–30]. The advantage of this algorithm consists in the ease of its realization and also in a higher efficiency for the ill-conditioned functions as compared to the stochastic quasi-gradient algorithm (1). The drawback of this algorithm consists in its "inertial motion", i.e., the direction of movement changes weakly from iteration to iteration, therefore the algorithm for simple functions may be less efficient than the algorithm (1).

Using the suggested approach the authors of [30] developed the recurrence schemes for modification of two parameters of the algorithm: step size and aggregation coefficient. This made it possible to increase the practical rate of the algorithm convergence far from the extremum, leaving without changes the local rate of convergence of classical methods. The algorithm convergence was proved and the asymptotic rate of convergence was given.

Let us consider the minimization problem of convex possibly non-smooth function $f(x)$ on the convex compact subset X of the space R^n. Stochastic gradients of the function $f(x)$ are known.

The algorithm generates sequences of random directions d^s and points $x^s \in R^n$, $s = 0, 1, \ldots$ according to formulas

$$d^s = (\xi^s + i_s \gamma_s d^{s-1}) / (1 + \gamma_s) \ , \tag{4}$$

$$x^{s+1} = \begin{cases} \Pi_X (x^s - \rho_s (1 + \gamma_s) d^s) \ , & \text{if} \quad \rho_s (1 + \gamma_s) \|d^s\| \le t \ , \\ \Pi_X (x^s - t d^s / \|d^s\|) \ , & \text{if} \quad \rho_s (1 + \gamma_s) \|d^s\| > t \ . \end{cases} \tag{5}$$

Here ξ^s is a stochastic quasi-gradient, i.e., $E_s \xi^s \in \partial f(x^s)$ where δ-field $\tilde{f}_s$ is generated by random values $(x^0, \xi^0, \ldots, x^{s-1}, x^s)$; ρ_s is a positive step size; γ_s is a positive aggregation coefficient; $i_s \in \{0, 1\}$ is a reset coefficient; $t \in (0; +\infty)$ is a constant.

At the initial point $x^0 \in X$ we assume $d^{-1} = 0$. From (4) it follows that the direction d^s is a convex combination of zero vector and stochastic subgradients ξ^i, $i = 0, \ldots, s$.

The reset coefficient is defined in the following manner:

$$i_s \in \{0, 1\} \quad \text{if} \quad \|\xi^{s-1}\| \le \delta \ ,$$

$$i_s = 0 \quad \text{if} \quad \|\xi^{s-1}\| > \delta \ ,$$

where δ is some fixed threshold.

To construct recurrence relations of modification of parameters ρ_s, γ_s we assume that the algorithm operates in the interior of the admissible domain X and $t = +\infty$. For the given x^{s-1}, d^{s-1} and $\lambda \ge 0$ we consider regularized function which characterizes the quality of th chosen parameters ρ and γ

$$\varphi_s(\rho, \gamma) = f(x^s(\rho, \gamma, \xi^{s-1})) - f(x^{s-1}) + \frac{1}{2}\lambda \|x^s(\rho, \gamma, \xi^{s-1}) - x^{s-1}\|^2 \ ,$$

where

$$x^s(\rho, \gamma, \xi^{s-1}) = x^{s-1} - \rho(\xi^{s-1} + i_{s-1}\gamma d^{s-1})$$

is defined by relations (4), (5). Values ρ_{s-1} and γ_{s-1} may be chosen from the condition of the minimum of the function $\Phi_s(\rho, \gamma) = E_{s-1}\varphi_s(\rho, \gamma)$. However, the program realization of such search at each iteration is difficult. We differentiate in the generalized sense the function $\varphi_s(\rho, \gamma)$ at the point ρ_{s-1}, γ_{s-1}. After simple

transformations we obtain

$$\partial \varphi_s(\rho_{s-1}, \gamma_{s-1}) = \{(\tilde{u}, \tilde{v}) : \tilde{u} = \frac{1}{\rho_{s-1}}[\langle g^s, \Delta x^s \rangle + \lambda \|\Delta x^s\|^2] ,$$

$$\tilde{v} = \frac{\rho_{s-1} i_{s-1}}{\rho_{s-2}(1+\gamma_{s-2})}[\langle g^s, \Delta x^{s-1} \rangle + \lambda \langle \Delta x^s, \Delta x^{s-1} \rangle], \; g^s \in \partial f(x^s)\} ,$$

where $\Delta x^s = x^s - x^{s-1}$. Taking into account the designations

$$u_s = \langle \xi^s, \Delta x^s \rangle + \lambda \|\Delta x\|^2 ,$$

$$v_s = i_{s-1}(\langle \xi^s, \Delta x^{s-1} \rangle + \langle \Delta x^s, \Delta x^{s-1} \rangle) ,$$

we have

$$E_{s-1}\begin{bmatrix} \frac{1}{\rho_{s-1}} u_s \\ \frac{\rho_{s-1}}{\rho_{s-2}(1+\gamma_{s-2})} v_s \end{bmatrix} \in \partial \Phi_s(\rho_{s-1}, \gamma_{s-1}) .$$

Thus the vector (U_s, v_s) may be interpreted as a stochastic quasi-gradient of the function Φ_s at the point $(\rho_{s-1}, \gamma_{s-1})$ with the accuracy up to positive multipliers.

Similarly to relation (2) the vector (u_s, v_s) was used in [31] for construction of the rule for calculation of the step size

$$\rho_0 > 0 ,$$

$$\rho_s = \min\{\bar{\rho}, \rho_{s-1} \exp[\min(\eta, -\alpha u_s - j_s \delta \rho_{s-1})]\} , \qquad (6)$$

where $\bar{\rho} > 0$, $\eta > 0$, $\alpha > 0$, $\lambda \geq 0$ are fixed parameters, the coefficient j_s in the last relation is calculated by the formula

$$j_s \in \{0, 1\} \quad \text{if} \quad \|\Delta x^s\| \geq \Delta_{\min} ,$$

$$j_s = 1 \quad \text{if} \quad \|\Delta x^s\| < \Delta_{\min} ,$$

$\Delta_{\min}$ is a small positive value.

The formula for calculation of aggregation coefficients γ_s is written similarly

$$\gamma_0 = \gamma_1 > 0 ,$$

$$\gamma_s = \min\{\bar{\gamma}, \gamma_{s-1} \exp(-\beta v_s - j_s \varkappa \gamma_{s-1})\} , \qquad (7)$$

$$\bar{\gamma} > 0, \; \Lambda > 0 .$$

In relations (6), (7) the additional members $j_s \delta \rho_{s-1}$, $j_s \Delta \gamma_{s-1}$ increase the rate of the decrease of coefficients ρ_s, γ_s in the case when the values u_s and v_s are close to zero.

The considered approach may be applied to other algorithms, stochastic and non-stochastic, in which the parameters control is required. The author suggested and theoretically substantiated adaptive step size controls for the stochastic Arrow-Hurwicz algorithm of search for saddle points of convex-concave functions [32] and for the gradient algorithm of search for Nash equilibrium in non-cooperative many-person games [33].

3. CONVERGENCE AND RATE OF CONVERGENCE OF STOCHASTIC QUASI-GRADIENT ALGORITHM

We will prove the convergence with probability 1 of the stochastic quasi-gradient algorithm (1) with step size control (2) to the extremal set of the convex function and estimate its asymptotic rate of convergence for twice differentiable functions.

We show that the sequence of step sizes chosen according to (2) satisfies the classical conditions

$$\rho_s > 0,\ s = 0,1,\ldots \ a.s.;\ \sum_0^\infty \rho_s = \infty \ a.s.;\ \sum_0^\infty E\rho_s^2 < \infty \ a.s.$$

Note that classical theorems about convergence for the algorithm (1) with step size control (2) cannot be used (see, e.g., [2]) because it is usually assumed that the step size ρ_s depends only on random vectors $(x^0, \ldots, x^s)$, in the given case this condition is broken since the step size ρ_s depends also on ξ^s.

Let us consider the problem of minimization of the convex function $f(x)$ on a convex compact subset $X \in R^n$. We use the stochastic quasi-gradient algorithm (1) with step size control (2) and $a_s = a > 1$, $s = 0, 1,\ldots$ for the search for the optimum of the function $f(x)$, i.e.,

$$x^{s+1} = \Pi_X(x^s - \rho_s \xi^s),\ s = 0,1,\ldots\ , \tag{8}$$

$$\rho_{s+1} = \min(\bar{\rho},\ \rho_s a^{-<\xi^{s+1},\ \Delta x^{s+1}> - \delta\rho_s}),\ s = 0,1,\ldots\ , \tag{9}$$

$$E_s \xi^s \in \partial f(x^s) + b^s,\ s = 0,1,\ldots\ ,$$

where δ-field $\tilde{f}$ is specified by random values $(x^0, \xi^0, \ldots, x^{s-1}, \xi^{s-1}, x^s)$. Denote $\varphi^s = \xi^s - \hat{f}^x(x^s)$, where $\hat{f}_x(x^s) \in \partial f(x^s)$, $E_s \xi^s = \hat{f}_x(x^s) + b^s$.

3.1. The Algorithm convergence

We will prove the convergence with probability 1 of the process (8), (9) to the extremal set of the function $f(x)$ on the admissible set X.

THEOREM 1 *Let $f(x)$ be a convex (possible non-smooth) function specified on the convex compact subset X of the space R^n, the function $f(x)$ satisfies the Lipscitz condition on X. If*

$$\max_{x, y \in X} \|x - y\| = C_1 \ , \tag{10}$$

$$\|\xi^s\| \le C_2 \ a.s., \ s = 0, 1, \ldots \ , \tag{11}$$

$$b^s \to 0 \ a.s. \ , \tag{12}$$

$$\delta \ge 2 \ln_a [E_s a^{|\varphi^s| C_2}] \ a.s., \ s = 0, 1, \ldots \ , \tag{13}$$

then with probability 1 all accumulation points of the sequence $\{x^s\}$ specified by relation (8), (9) belong to the set

$$X^* = \{x \in X : f(x) = \min_{y \in x} f(y)\} \ .$$

PROOF Prior to proving the principal assertion of the theorem let us set several properties of step sizes ρ_s, $s = 0, 1, \ldots$.

LEMMA 1 [23].

$$\sum_{s=0}^{\infty} \rho_s = \infty \ a.s.$$

PROOF Suppose the opposite, i.e., that there exists such constant K for which the probability of event

$$A = \left\{ w : \sum_{s=0}^{\infty} \rho_s \le K \right\}$$

is more than zero $P(A) > 0$. From relations (9), (11) it is easy to obtain

$$\rho_{s+1} \ge \min(\bar{\rho}, \ \rho_s a^{-|\xi^{s+1}| |\Delta x^{s+1}| - \delta \rho_s}) \ge$$

$$\min(\bar{\rho},\ \rho_s a^{-c_2^2\rho_s-\delta\rho_s}) \geq \min(\bar{\rho},\ \rho_s a^{-C_3\rho_s})\ a.s.$$

where $C_3 = (C_2^2 + \delta) > 0$.

For elementary events $w \in A$ from the last estimate we have

$$\rho_{s+1} \geq \min(\bar{\rho},\ \rho_0 a^{-C_2\sum_0^s \rho_i}) \geq (\bar{\rho},\ \rho_0 a^{-C_2 K}) > 0\ .$$

The obtained lower bound for the step size ρ_{s+1} is inconsistent with the relation $\sum_0^\infty \rho_s \geq K$. The lemma is proved.

LEMMA 2

$$\sum_{s=0}^{\infty} E\rho_s^2 < \infty\ \ a.s.$$

PROOF Taking into account (8), (9), (11), the definition of the gradient of the convex function and properties of the projection operation we obtain

$$\rho_{s+1} \leq \rho_s a^{-\langle \xi^{s+1},\, \Delta x^{s+1}\rangle - \delta\rho_s} =$$

$$\rho_s a^{-\langle f_x(x^{s+1}),\Delta x^{s+1}\rangle - \langle \zeta^{s+1},\, \Delta x^{s+1}\rangle - \delta\rho_s} \leq$$

$$\rho_s a^{f(x^s) - f(x^{s+1}) - \langle \zeta^{s+1},\, \Delta x^{s+1}\rangle - \delta\rho_s} \leq$$

$$\rho_s a^{f(x^s) - f(x^{s+1}) + |\zeta^{s+1}||\xi^s|\rho_s - \delta\rho_s}\ . \qquad (14)$$

Hence

$$\rho_{s+1} a^{f(x^{s+1})} \leq \rho_s a^{f(x^s)} a^{(|\zeta^{s+1}|C_2 - \delta)\rho_s}\ \ a.s.$$

Since according to (13)

$$E_{s+1} a^{|\zeta^{s+1}|C_2 - \delta/2} \leq 1\ \ a.s.$$

from the last inequality we obtain

$$E_{s+1}\rho_{s+1} a^{f(x^{s+1})} \leq \rho_s a^{f(x^s) - \frac{\delta}{2}\rho_s} E_{s+1} a^{(|\zeta^{s+1}|C_2 - \delta/2)\rho_s} \leq$$

$$\rho_s a^{f(x^s) - \frac{\delta}{2}\rho_s}\ \ a.s.$$

Since $0 \le \rho_s \le \bar{\rho}$ then for $\alpha = (1 - a^{-\frac{\delta}{2}\rho_s})\bar{\rho}^{-1} > 0$ the relation $1 - \alpha\rho_s \ge a^{-\frac{\delta}{2}\rho_s}$ is fulfilled.

By substituting this estimate into the previous inequality and by introducing the designation $\mu_s = \rho_s a^{f(x^s)}$ we have

$$E_{s+1}\mu_{s+1} \le \mu_s a^{-\frac{\delta}{2}\rho_s} \le \mu_s - \alpha\rho_s^2 a^{f(x^s)} \quad a.s.$$

By taking the mathematical expectation from both sides of the inequality we obtain

$$E\mu_{s+1} \le E\mu_s - \alpha C_4 E\rho_s^2 \le E\mu_0 - \alpha C_4 \sum_0^s E\rho_s^2 ,$$

where $C_4 = \inf_{x \in X} a^{f(x)}$. Since $E\mu_{s+1} - E\mu_0 \ge$ const then the last estimate results in the assertion of the lemma.

COROLLARY [23].

$$\rho_s \to 0 \quad a.s.$$

LEMMA 3 [23].

$$\rho_{s-1}/\rho_s \to 1 \quad a.s.$$

PROOF Since $\rho_s \to 0$ a.s. then

$$\rho_s a^{-<\xi^{s+1}, \Delta x^{s+1}> - \delta\rho_s} \to 0 \quad a.s.$$

From the relation (9) it follows that for almost each elementary event $w \in \Omega$ there may be found the number $s(w)$ such that for $s > S(w)$

$$\rho_{s+1} = \rho_s a^{-<\xi^{s+1}, \Delta x^{s+1}> - \delta\rho_s} . \tag{15}$$

Since $\rho_s \to 0$ a.s., then

$$-<\xi^{s+1}, \Delta x^{s+1}> - \delta\rho_s \to 0 \quad a.s.$$

and the assertion of the lemma results from the relation (15).

To prove the main assertion of the theorem we use the conditions of convergence of stochastic programming algorithms [13] with insignificant modifications.

THEOREM 2 *Let the random process* $\{x^s(w)\}$ *and a set of solutions* $X^* \in R^n$ *be such that:*

C1. Almost for all subsequences $\{x^{n_k}(w)\}$ such that $\lim_{k \to \infty} x^{n_k}(w) \in X^*$ the relation

$$\lim_{k \to \infty} \|x^{n_k}(w) - x^{n_k+1}(w)\| = 0$$

holds.

C2. There exists a compact set X such that

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

C3. If there exists such event $B \subset \Omega$ that $P(B) > 0$ and for all $w \in B$ there exists a subsequence $\{x^{s_k}(w)\}$, $x^{s_k}(w) \to x'(w) \overline{\in} X^*$ then for almost all $w \in B$ there exists such $\varepsilon_0(w) > 0$ that for all k and $0 < \varepsilon \leq \varepsilon_0(w)$

$$m_k(w) = \inf_{m > s_k} \{m : \|x^m(w) - x'(w)\| > \varepsilon\} < 0 \ . \tag{16}$$

C4. There exists a continuous function $W(x)$ such that for $w \in B$

$$\lim_{k \to \infty} W(x^{m_k}(w)) < W(x'(w)) \ .$$

C5. The function $W(x)$ takes on X^* at most countable number of values.

Then the limit of any convergent subsequence belongs to the set X^* almost for all w.

We assume

$$W(x) = \min_{y \in X^*} \|x - y\|^2, \ U_\varepsilon(x) = \{y \in R^n : \|y - x\| \leq \varepsilon\} \ ,$$

$$f^* = \min_{x \in X} f(x), \ x_s^* = arg \min_{y \in X^*} \|x^s - y\|, \ \eta^s = \xi^s - \hat{f}_x(x^s) - b^s \ .$$

We test the satisfiability of conditions C1–C5.

The condition C1 is satisfied obviously, since by virtue of the corollary of the lemma 2

$$\|\Delta x^{s+1}\| \to 0 \quad a.s.$$

The condition C2 is satisfied by virtue of theorem 1.

The condition C5 is satisfied since the function $W(x)$ is a constant on the set X^*.

We test the condition C3. Let the probability of event B is more than zero, $w \in B$ and $x^{s_k}(w) \to x'(w) \overline{\in} X^*$.

For the brevity we will omit the argument w. If the condition (16) for the given w is not satisfied then there may be found arbitrarily small ε and number s_k such that for $s > s_k$ valid is $x^s \in U_\varepsilon(x')$.

For $s > s_k$

$$W(x^{s+1}) \le \|x_s^* - x^s + \rho_s \xi^s\|^2 = W(x^s) + 2\rho_s < \xi^s, x_s^* - x^s > +$$

$$\rho_s^2 \|\xi^s\|^2 = W(x^s) + 2\rho_s < \hat{f}_x(x^s), x_s^* - x^s > + 2\rho_s < b^s, x_s^* - x^s > +$$

$$2\rho_s < \eta^s, x_s^* - x^* > + C_2^2 \rho_s^2 \le W(x^s) + 2\rho_s (f^* - f(x^s)) +$$

$$2\rho_s \|b^s\| C_1 + 2\rho_{s-1} < \eta^s, x_s^* - x^s > + 2(\rho_s - \rho_{s-1}) < \eta^s, x_s^* - x^s > +$$

$$C_2^2 \rho_s^2 \le w(x^{s_k}) + 2 \sum_{l=s_k}^{s} \rho_l (f^* - f(x^l) + x_l) + \gamma_s \ , \qquad (16')$$

where

$$\Lambda_s = \|b^s\| C_1 + (1 - \rho_{s-1}/\rho_s) < \eta^s, x_s^* - x^s > \ ,$$

$$\gamma_s = 2 \sum_{l=s_k}^{s} \rho_{l-1} < \eta^l, x_l^* - x^l > + C_2^2 \sum_{l=s_k}^{s} \rho_l^2 \ .$$

By virtue of conditions (10, (11) the scalar product $< \eta^s, x_s^* - x^s >$ is bounded, therefore, taking into account (12) and lemma 3, we have

$$\Lambda_s \to 0 \quad a.s.$$

From lemma 2 it follows that the martingale series $\sum_{l=s_k}^{s} \rho_{l-1} < \eta^l, x_l^* - x^l >$ is convergent a.s., therefore

$$|\gamma_s| < C \quad a.s.$$

the constant C here depends upon w. Consequently, for sufficiently large numbers s and small ε

$$f^* - f(x^s) + \Lambda_s < -\delta,\ \delta > 0 \ .$$

Taking into account lemma 1 we obtain that beginning with some number S_k for sufficiently large numbers s the estimate

$$W(x^{s+1}) \le W(x^{s_k}) - \delta \sum_{l=s_k}^{S} \rho_l \tag{17}$$

holds.

Passing to the limit $s \longrightarrow \infty$ we obtain the contradiction with boundedness $W(x)$ on the closed bounded set $U_\varepsilon(x')$. The contradiction proves C3.

We will prove C4. Since $\|\Delta x^s\| \longrightarrow 0$ a.s. then by constructing index m_k beginning with some number k valid is $\| \sum_{s_k}^{m_k-1} \rho_l \xi^l \| > \varepsilon/2$. By virtue of condition (11) we obtain

$$\sum_{s_k}^{m_k-1} \rho_l \ge \varepsilon/(2C_2) \ .$$

Substituting the last estimate in (17) for $s = m_k - 1$ we have

$$w(x^{m_k}) \le w(x^{s_k}) - \frac{\delta\varepsilon}{2C_2} \ .$$

Since $w(x^{s_k}) \longrightarrow w(x')$ then for sufficiently large k

$$w(x^{m_k}) \le w(x') - \frac{\delta\varepsilon}{3C_2}$$

The last inequality proves C4.

3.2. Asymptotic Properties of Step Sizes

We now study adymptotic properties of a sequence of step sizes ρ_s, $s = 0, 1, \ldots$ for the case of twice continuously-differentiable function. These results will be used for obtaining asymptotic rate of convergence of the algorithm (8)–(9).

LEMMA 4 *Let for the sequence* $\{x^s\}$ *specified by relations (8)–(9) valid be all conditions of theorem 1 and* $b^s = 0$; $s = 0,1,\ldots$ *the function* $f(x)$ *be twice continuously-differentiable on the open set containing* X, *then*

$$\rho_s = \frac{1}{(s+1)\delta \ln a} + o\left[\frac{1}{s+1}\right] \quad a.s.$$

PROOF Denote $\tau_s = \ln_a[(s+1)\rho_s]$, $s = 0,1,\ldots$. According to the corollary of lemma 2 for sufficiently large numbers s from relation (8) we obtain

$$\rho_s = \rho_{s-1} a^{-\langle \xi^s, \Delta x^s \rangle - \delta \rho_{s-1}} \quad a.s.$$

Consequently

$$\tau_s = \tau_{s-1} + \ln_a(1 + \frac{1}{s}) - \langle \xi^s, \Delta x^s \rangle - \delta \rho_{s-1} = \tau_{s-1} +$$

$$\frac{1}{s \ln a}(1 - \ln(a)\delta a^{\tau_{s-1}}) + \ln_a(1 + \frac{1}{s}) - \frac{1}{s \ln a}$$

$$- \langle \eta^s, \Delta x^s \rangle - \langle \nabla f(x^s), \Delta x^s \rangle \qquad (18)$$

It is obvious that the series $\sum_{s=1}^{\infty} [\ln_a(1 + \frac{1}{s}) - \frac{1}{s \ln a}]$ is convergent. From lemma 2 the convergence of the martingale series $\sum_{s=1}^{\infty} \langle \eta^s, \Delta x^s \rangle$ follows a.a. Since the function $f(x)$ is twice continuously differentiable, then $\langle \nabla f(x^s), \Delta x^s \rangle = f(x^s) - f(x^{s-1}) + \Psi_s \|\Delta x^s\|^2$ where Ψ_s is uniformly bounded for all s. The equality

$$\sum_{s=1}^{k} \langle \nabla f(x^s), \Delta x^s \rangle = f(x^k) - f(x^0) + \sum_{s=1}^{k} \Psi_s \|\Delta x^s\|^2$$

is satisfied. The function $f(x)$ is bounded on the compact set X, the series $\sum_{s=1}^{\infty} \Psi_s \|\Delta x^s\|^2$ is convergent a.s. by virtue of lemma 2, therefore the series $\sum_{s=1}^{\infty} \langle \nabla f(x^s), \Delta x^s \rangle$ is also convergent. The relation (18) then may be rewritten as follows

$$\tau_s = \tau_{s-1} + \frac{1}{s \ln a}(1 - \delta \ln(a) a^{\tau_{s-1}}) + t_s$$

where the series $\sum_{s=1}^{\infty} t_s$ is convergent a.s. The last formula is the Robbins-Monro

algorithm for solution of equation $1 - \delta \ln(a) a^z = 0$. Using standard results about convergence of stochastic approximation algorithms (see, e.g. [11]), we obtain

$$\tau_s \to \ln_a \left(\frac{1}{\delta \ln a}\right) \quad a.s.$$

Q.E.D.

3.3. Rate of Algorithm Convergence

For the case of twice continuously-differentiable function we estimate the asymptotic rate of convergence of the algorithm (8)–(9) in non-stochastic case, i.e., for $\xi^s = \nabla f(x^s)$, $s = 0, 1, \ldots$.

THEOREM 3 *Let all conditions of theorem 1 hold,* $\xi^s = \nabla f(x^s)$, $s = 0, 1, \ldots$, *the function* $f(x)$ *be twice continuously differentiable and*

$$f(x) \geq f(x^*) + B\|x^* - x\|^2, \; B > 0 \tag{19}$$

where x^* *is a unique point of minimum of the function* $f(x)$ *on the set* X, $\ln(a)(C_2^2 + \delta)/2B < 1$. *Then*

$$\|x^* - x^s\|^2 \leq O\left[\frac{1}{s}\right] \; s = 0, 1, \ldots$$

PROOF We use the following lemma to prove the theorem.

LEMMA 5 [*14*]. *Let there be a sequence* v_s, $s = 0, 1, \ldots$, *and*

$$v_{s+1} \leq (1 - \nu_s) v_s + \nu_s \mu_s, \; 0 \leq \nu_s \leq 1, \; 0 \leq \mu_s \leq \mu \; , \tag{20}$$

If

$$\beta_s = \left[\frac{\mu_s}{\mu_{s+1}} - 1\right] \frac{1}{\nu_s}, \; s = 0, 1, \ldots; \; \overline{\lim_{s \to \infty}} \beta_s \leq \beta < 1 \tag{21}$$

then $v_s \leq \dfrac{\mu_s}{1 - \beta} + o(\mu_s)$.

From the estimate (16') and condition (19) we have

$$W(x^{s+1}) \leq W(x^s) + 2\rho_s (f^* - f(x^s)) + C_2^2 \rho_s^2 \leq$$

$$W(x^s) - 2\rho_s B W(x^s) + C_2^2 \rho_s^2 = (1 - 2\rho_s B) W(x^s) + C_2^2 \rho_s^2 \; .$$

Denote $v_s = 2\rho_s B$, $v_s = W(x^s)$, $\mu_s = C_2^2 \rho_s / 2B$. The condition (20) of lemma 5 is satisfied obviously. We test the condition (21). According to the corollary of lemma 2 from the relation (9) we have for sufficiently large numbers

$$\frac{\rho_s}{\rho_{s+1}} = a^{<\xi^{s+1}, \Delta x^{s+1}> + \delta\rho_s} = 1 + \ln(a)(\Lambda_s + \delta)\rho_s + o(\rho_s) ,$$

where $|\Lambda_s| \le C_2^2$.

Consequently

$$\beta_s = \left[\frac{\mu_s}{\mu_{s+1}} - 1\right]\frac{1}{v_s} = \left[\frac{\rho_s}{\rho_{s+1}} - 1\right]\frac{1}{2\rho_s B} = (\ln(a)(\Lambda_s + \delta)\rho_s +$$

$$o(\rho_s))/(2\rho_s B) \le \ln(a)(C_2^2 + \delta)/(2B) + o(1) \quad a.s.$$

and $\overline{\lim_{s \to \infty}} \beta_s \le \ln(a)\,(C_2^2 + \delta)/(2B)$ $a.s.$ by the condition of the theorem.

Conditions of lemma 5 are tested, therefore

$$v_s = W(x^s) \le \frac{C_2^2 \rho_s}{2B - \ln(a)(C_2^2 + \delta)} + o(\rho_s) = O(\frac{1}{s})$$

since $\rho_s = O(\frac{1}{s})$.

4. ON PROGRAM REALIZATION OF STOCHASTIC QUASI-GRADIENT ALGORITHM

Program realization of algorithms in practice usually requires the introduction of some heuritic elements improving the algorithm operation.

Theorem 1 is proved provided that in step size controls (2)

$$a_s = \text{const}, \; s = 0, 1, \ldots$$

This may result in very speedy change of th step size ρ_s at each iteration. In program realization of the algorithm it is desirable to normalize the exponent in relation (9) to some value z_s which is the averaging of the value $|<\xi^{s+1}, \Delta x^{s+1}>|$. The averaging is made by the following recurrent formula

$$z_s = z_{s-1} + (|T_s| - z_{s-1})D, \; z_{s-1} = 0, \; T_s = <\xi^{s+1}, \Delta x^{s+1}> . \qquad (22)$$

It is desirable to set some threshold coefficients which limit the maximal change of the step size ρ_s. In numerical experiments the author used the following step size

rule [23]

$$\tilde{\rho}_{s+1} = \rho_s a^{\frac{T_s}{z_s}} \begin{cases} 1, & \text{if } T_s > 0, \\ U, & \text{if } T_s \le 0 \end{cases}, \tag{23}$$

$$\rho_{s+1} = \begin{cases} \rho_s 3, & \text{if } \tilde{\rho}_{s+1}/\rho_s > 3, \\ \rho_s / 4, & \text{if } \tilde{\rho}_{s+1}/\rho_s < 1/4, \\ \tilde{\rho}_{s+1}, & \text{otherwise} \end{cases}. \tag{24}$$

The recommended values of parameters are

$$a = 2,\ U = 0.8,\ D = 0.2 .$$

In relation (23) the additional reduction of the step size occurs only if the value T_s is negative. Results of computation experiments show that the scheme (8), (22), (23), (24) rapidly leads to the point of the extremum if the objective function is ot ill-conditioned, i.e., for non-"ravine" functions. In case when the function $f(x)$ is very "ravine" the algorithm gets stack "at the bottom of the ravine". This difficulty may be overcome by using more complex algorithms which employ matrices of space dilatation (3). In practice, the scaling procedure suggested by Saridis [34] for stochastic approximation algorithms proved to be efficient for such functions.

This procedure contains changes taking into account the projection operation and adaptive step size control.

$$x^{s+1} = \Pi_X(x^s - \rho_s H^s \xi^s),$$

$$H^{s+1} = \begin{bmatrix} h_1(s+1) & & 0 \\ & \ddots & \\ 0 & & h_n(s+1) \end{bmatrix}, H^0 = \begin{bmatrix} 1/n & & 0 \\ & \ddots & \\ 0 & & 1/n \end{bmatrix},$$

$$h_i(s+1) = \alpha h_{i(s)} + \lambda_i(s+1)(1-\alpha),$$

$$\lambda_i(s+1) = \begin{cases} 0 & \text{if } \xi_i^{s+1}(x_i^s - x_i^{s+1}) \le 0,\ k(s+1) \ne 0 \\ \dfrac{1}{k(s+1)} & \text{if } \xi_i^{s+1}(x_i^s - x_i^{s+1}) > 0,\ k(s+1) \ne 0 \\ \dfrac{1}{n} & \text{if } k(s+1) = 0 \end{cases}$$

where $k(s+1),\ 0 \le k(s+1) \le n$ is the quantity of numbers i for which $\xi_i^{s+1}(x_i^s - x_i^{s+1}) > 0$; n is the dimension of the space to which the set X belongs;

$$z_s = z_{s-1} + (|T_s| - z_{s-1})D,\ T_s = \langle H^{s+1}\xi^{s+1}, \Delta x^{s+1} \rangle,\ z_{-1} = 0 .$$

For this scheme the step size control is the same as in the previous case, i.e., (23), (24).

The recommended values of parameters are

$$a = 2,\ U = 0.8,\ D = 0.2,\ \alpha = 0.5\ .$$

Note that the considered schemes have the natural criterion of the break of iteration process. In the neighborhood of extremum the value $\|\Delta x^{s+1}\|$ becomes small and tends to zero. Therefore for the break we may use the following averaged value Q_s obtained as follows:

$$Q_s = Q_{s-1} + (\|\Delta x^{s+1}\| - Q_{s-1})D,\ Q_{-1} = 0\ .$$

If $Q_s = \varepsilon_B$, the process is broken. Here ε_B is some positive constant which characterizes the required precision of solution.

We give the results of computation experiments for scheme (8), (22), (23), (24).

EXAMPLE 1 The following problem statement arises in solving multi-list inventory problem [23].

Let us consider the problem

$$f(x) = E\sum_{i=1}^{5}\max\{a_i(x_i - \Theta_i),\ b_i(\Theta_i - x_i)\} \to \min$$

$$\begin{cases} x_1 + x_2 + 2x_3 + 3x_4 + x_5 = 200, \\ x_1 \le 50, \\ x_2 \le 07, \\ x_3 \le 07, \\ x_4 \le 08, \\ x_5 \le 25, \end{cases}$$

$$x_i \ge 0,\ i = 1, \ldots, 5\ .$$

Here Θ_i are random values uniformly distributed on intervals $[A_i, B_i]$, $i = 1, \ldots, 5$. Vectors $a = (a_1, \ldots, a_5)$, $B = (B_1, \ldots, B_5)$, $A = (A_1, \ldots, A_5)$, $B = (B_1, \ldots, B_5)$ are defined as follows:

$$a = (1, 0, 3, 1, 2)\ ,\ b = (3, 4, 1, 2, 3)\ ,$$

$$A = (0, 0, 0, 0, 0)\ ,\ B = (60, 15, 17, 90, 40)\ .$$

Analytical form of the function is as follows:

$$f(x) = \frac{1}{3}x_1^2 + \frac{2}{15}x_2^2 + \frac{2}{17}x_3^2 + \frac{1}{60}x_4^2 + \frac{1}{16}x_5^2 -$$

$$3x_1 - 4x_2 - x_3 - 2x_4 - 3x_5 + 278.5 \ .$$

Analytical form of the function $f(x)$ is used only for obtaining explicit solution by one of the methods of quadratic programming

$$f(x^*) = 98.10089, \ x^* = (41.88057, 7.00000, 2.48092, 41.27456, 22.33456) \ .$$

The stochastic quasi-gradient is computed by the formula

$$\xi^s = (\xi_1^s, \ldots, \xi_5^s), \ \xi_i^s = \begin{cases} a_i, & \text{if } x_i^s \geq \Theta_i^s, \\ -B_i, & \text{if } x_i^s < \Theta_i^s, \end{cases} \ i = 1, \ldots, 5 \ .$$

To solve this problem the scheme (8), (22), (23), (24) was used with parameters $a = 1.5$, $U = 0.0$, $D = 0.25$, $\rho_0 = 1$. Initial approximation is $x^0 = (0, 0, 0, 0, 0)$, $f(x^0) = 278.5$. At the 91st iteration the step size $\rho_{91} = 0.15$. The results of averaged values of coordinates and of the functions at 91st to 100th iteration are as follows:

$$\bar{x}_i = \frac{1}{10} \sum_{s=91}^{100} x_i^s, \ i = 1, \ldots, 5; \ \bar{x}_1 = 40.5485, \ \bar{x}_2 = 6.9981$$

$$\bar{x}_3 = 2.4381, \ \bar{x}_4 = 42.2561, \ \bar{x}_5 = 20.3561 \ ,$$

$$\frac{1}{10} \sum_{s=91}^{100} f(x^s, \Theta^s) = 97.4185 \ .$$

Note that to obtain the final result it is desirable to average solution approximations with respect to last iterations.

EXAMPLE 2 *Random location equilibrium problem* [7]. The calculations are performed by the author together with N. Roenko. This problem has been considered in section 1. It consists in minimization of the function

$$f(x) = \sum_{i=1}^{n} \beta_i \int\int_{R^2} \|x - w\| \Theta_i(dw) \to \min_{x \in R^2}$$

The number of points to be located in $n = 30$, probability measures Θ_i, $i = 1, \ldots, n$ are bivariate normal density function whose means and standard deviations are generated randomly in the range 0–20. The weights β_i are also generated randomly in the range 0–10. Data are given in the Table.

To solve the problem the scheme (8), (22), (23), (24) with parameters $\alpha = 2$, $U = 0.8$, $D = 0.2$, $\rho_0 = 1$ was used. The exact value of the point of extremum is $x^* =$ (8.36, 9.36). The initial approximation is $x^0 =$ (41, 87). The results of averaging of approximations x^s from the 51st to 60th iteration are

$$\frac{1}{10}\sum_{51}^{60} x^s = (9.1,\ 10.2)\ ,$$

from the 191st to 200th iteration

$$\frac{1}{10}\sum_{191}^{200} x^s = (8.9,\ 9.0)\ .$$

With initial approximation $x^0 = (54,\ 30)$ the following solution approximations

$$\frac{1}{10}\sum_{21}^{30} x^s = (8.0,\ 10.1)\ ;\ \frac{1}{10}\sum_{191}^{200} x^s = (7.9,\ 9.7)$$

are obtained.

The results of numerical experiments show that approximations of solutions sufficiently quickly fall into the neighborhood of solution and after this the accuracy of approximation is not practically improved.

It should be noted that this effect is connected with asymetry of generators of random numbers rather than with the choice of step size control.

The suggested approach has some advantages as compared to [7] because to realize the computation process it is not necessary to integrate complex functions.

Table 1

x_1 means are	3.02	6.07	9.77	16.26	6.12	14.80	7.24	7.52	15.91	13.57
	2.08	12.70	0.16	15.78	3.95	11.89	4.68	6.11	9.19	11.56
	12.43	19.98	15.33	18.20	7.84	1.16	4.54	17.48	10.78	1.45
x_2 means are	7.63	6.62	15.40	10.83	4.85	17.14	2.20	9.30	17.30	14.60
	5.68	4.77	19.10	17.17	0.80	10.82	11.48	18.99	0.36	2.52
	10.00	1.93	11.39	16.41	16.21	2.09	16.69	8.70	12.04	2.93
x_1 devs. are	18.65	18.95	0.45	13.50	17.55	1.12	18.42	1.59	15.65	9.49
	19.13	18.19	19.56	19.14	11.93	7.26	1.72	11.37	7.09	16.05
	15.62	4.31	15.44	1.40	5.82	8.56	16.72	5.29	10.36	12.49
x_2 devs. are	3.77	15.79	8.68	6.29	7.97	9.23	5.81	3.17	17.91	7.02
	16.27	15.08	5.12	6.11	1.55	19.25	8.24	17.78	13.48	9.80
	5.49	15.13	7.07	16.83	15.86	9.90	19.44	16.65	0.37	15.31
Weights are	8.50	9.48	6.03	8.16	9.05	1.80	8.17	7.57	3.43	9.62
	2.87	3.77	4.34	4.88	0.11	2.13	7.75	1.64	5.74	6.12
	4.57	4.45	2.95	0.17	7.53	9.39	7.38	1.15	2.09	7.20

REFERENCES

1 Ju. M. Ermoliev, Z.V. Nekrylova. On Some Stochastic Optimization Methods. Kibernetika, 1966, No. 6 (in Russian).

2 Ju. M. Ermoliev. Methods of Stochastic Programming. Nauka, Moscow, 1976, 240 p. (in Russian).

3 Ju. M. Ermoliev. Stochastic Quasi-Gradient Methods and their Applications to Systems Optimization. Stochastics, 1983, No. 4.

4 H. Robbins, S. Monro. Stochastic Approximation Methods. Ann. Math. Statist., 1951, 22, 400–407.

5 L.A. Rastrigin. Theory of Statistical Search Methods. Nauka, 1968 (in Russian).

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

7 N. Katz, L. Cooper. An Always-Convergent Numerical Scheme for a Random Locational Equilibrium Problem. SIAM J. Numer. Anal. 1974, Vol. 11, No. 4, September, 683–692.

8 H. Kiefer, J. Wolfowitz. Stochastic Estimation of the Maximum of a Regression Function. Ann. Math. Statist., 1952, 23, 462–466.

9 J.R. Blum. Ann. Math. Statist., 1954, 25, 737–744.

10 J. Sacks. Asymptotic Distribution of Stochastic Approximation Procedure. Ann. Math. Statist., 1954, 25, 737–744.

11 M.B. Nevelson, R.Z. Khasminski. Stochastic Approximation and Recursive Estimation. Nauka, Moscow, 1972 (in Russian).

12 A.M. Gupal. Stochastic Methods for Solving Non-Smooth Extremal Problems. Naukova Dumka, Kiev, 1979 (in Russian).

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

14 B.T. Poljak. Convergence and Rate of Convergence of Iterative Stochastic Algorithms. I. General Case. Avtomatika i Telemekhanika. 1976, No. 12, 83–94 (in Russian).

15 Ju.M. Ermoliev, Ju.M. Kaniovskij. Asymptotic properties of Some Methods of Stochastic Programming with Constant Step size. Zhurn. vych. mat. i mat. fiziki, 1979, No. 2, 356–366 (in Russian).

16 A.S. Nemirovskij, D.B. Judin. Complexity of Problem and Efficiency of Optimization Methods. Nauka, Moscow, 1979, 384 p. (in Russian).

17 H.J. Kushner. Asymptotic Behavior of Stochastic Approximation and Large Deviations. Divisions of Appl. Math. and Eng. Lefschetz Center for Dyn. Syst. Brown Univ. Providence, Rhode Island, 1983, 27 p.

18 H. Kesten. Accelerated Stochastic Approximation. Ann. Math. Statist. 1958, 29, 41–59.

19 Ju. M. Ermoliev, G. Leonarki, J. Vira. The Stochastic Quasi-Gradient Methods Applied to a Facility Location Model. Working paper, WR-81-14, 1981, Laxenburg, Austria, International Institute for Applied Systems Analysis.

20 A.M. Gupal, F. Mirzoakhmedov. On One Method of Step Size Control in Stochastic Programming Methods. Kibernetika, 1978, No. 1, 133–134 (in Russian).

21 G. Pflug. On the Determination of the Step Size in Stochastic Quasi-Gradient Methods. Working paper, 1983, May, Laxenburg, Austria, International Institute for Applied Systems Analysis, 24 p.

22 S.P. Urjas'ev. A Step Size Rule for Direct Methods of Stochastic Programming. Kibernetika (Kiev), 1980, No. 6, 96–98 (in Russian).

23 F. Mirzoakhmedov, S.P. Urjas'ev. Adaptive Step Size Control for Stochastic Optimization Algorithm. – Zhurn. vych. mat. i mat. fiziki., 1983, No. 6, 1314–1325 (in Russian).

24 D. Himmelblau. Applied Nonlinear Programming. McGrow-Hill Book Company, 1972.

25 R.J.-B. Wets. Modeling and solution strategies for unconstraned stochastic optimization problems. Annals of Operations Research 1984, No. 1, 3–22.

26 N.Z. Shor. Minimization Methods for Non-Differentialbe Functions. Springer-Verlag, 1985, 162 p.

27 A.M. Gupal, L.T. Bazhenov. A Stochastic Analog of the Methods of Conjugate Gradients. Kibernetika, 1972, 124–126, (in Russian).

28 A.P. Korostelev. On Multi-Step Procedure of Stochastic Optimization. Avtomatika i Telemekhanika, 1981, 82–90 (in Russian).

29 H.J. Kushner, Hai-Huang. Asymptotic Properties of Stochastic Approximation with Constant Coefficients. SIAM Journal on Control and Optimization, 1981, 19, 87–105.

30 N.D. Chepurnoj. One Step Size Control in Stochastic Method of Minimization of Non-Smooth Functions. Kibernetika, 1982, No. 4, 127–129 (in Russian).

31 A. Ruszczynski, W. Syski. A Method of Aggregate Stochastic Sub-Gradients with On-Line Stepsize Rules for Convex Stochastic Programming Problem. Mathematical Programming Study (to appear).

32 S.P. Urjas'ev. Arrow-Hurwicz Algorithm with Adaptively Controled Step Sizes. In: Operations Research and AMS, 1984, 24, 3–11 (in Russian).

33 Ju.M. Ermoliev, S.P. Urjas'ev. On Search for Equilibrium by Nash in Many-Person Games. Kibernetika, 1982, No. 3, 85–88 (in Russian).

34 G.M. Saridis. Learning Applied to Successive Approximation Algorithms. IEEE Trans. Syst. Sci. Cybern. 1970, Vol. SSC-6, Apv. 97–103.