STOCHASTIC QUASIGRADIENT METHODS AND THEIR APPLICATION IN SYSTEMS OPTIMIZATION

Yuri Ermoliev

January 1981
WP-81-2

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
A-2361 Laxenburg, Austria

ABSTRACT

This paper systematically surveys the basic direction of development of stochastic quasigradient methods which allow one to solve optimization problems without calculating the precise values of objective and constraints function (all the more of their derivatives). For deterministic nonlinear optimization problems these methods can be regarded as methods of random search. For the stochastic programming problems, SQG methods generalize the well-known stochastic approximation method for unconstrained optimization of the expectation of random functions to problems involving general constraints.

STOCHASTIC QUASIGRADIENT METHODS AND THEIR APPLICATION IN SYSTEMS OPTIMIZATION

Yuri Ermoliev

1. INTRODUCTION

The stochastic quasigradient (SQG) methods are stochastic algorithmic procedures for solving general constrained optimization problems with nondifferentiable, nonconvex functions, see [1] - [34]. There are two main reasons for interests in stochastic procedures of optimization. Firstly, deterministic processes are special cases of stochastic processes, therefore stochastic procedures of optimization give us a new opportunity to build more flexible and effective algorithms; secondly, a wide range of applied problems cannot be stated and solved within the framework of deterministic optimization techniques.

The SQG methods allow us to solve optimization problems with objective functions and constraints of such a complex nature that the calculation of the precise values of these function (all the more of their derivatives) is impossible. The main idea of these methods consists of using statistical estimates for the values of the functions and of their derivatives instead of their precise values. For the stochastic programming problems, SQG methods generalize the well-known stochastic approximation methods for unconstrained optimization of the expectation of random functions (see for instance [47] to problems involving general

constraints. For deterministic nonlinear programming problems SQG methods can be regarded as methods of random search [44]. Some extensions of stochastic approximation methods to problems with differentiable functions are made in [48] - [53], [66] B.T. Poljak (see survey [46]) proposed techniques for investigating the local convergence of the stochastic optimization process and proved some results concerning differentiable optimization with strong assumptions about the noise: the random disturbances in the function evaluations and of their derivatives are assumed to be independent from each other. Such cases can be regarded as methods of optimization in the presence of random noise. The paper [33] is a survey of SQG mthods for stochastic minimax problems.

The purpose of this paper is a systematic review of recent work concerning the main direction of development of SQG procedures, their applications and an overview of the key ideas involved in the proofs. During the course of writing this paper, several colleagues at IIASA read various draft versions and made many useful comments. I wish especially to thank R. Wets who read the paper and commented on it in detail. Several of his suggestions were included to eliminate misleading statements. The author is also grateful to M. Dempster for the proposition of writing this paper and numerous comments.

2. THE GENERAL IDEA OF SQG METHODS

Consider the problem of minimization:

$$F^0(x) = \min \tag{1}$$

subject to

$$F^i(x) \leqslant 0, i=\overline{1,m} \quad , \tag{2}$$

$$x \in X \leqslant R^n \quad . \tag{3}$$

To start with, let us assume that the functions $F^\nu(x), \nu = \overline{o,m}$ are convex. Then for every x we have the inequality

$$F^{\nu}(z) - F^{\nu}(x) \geq \langle \hat{F}^{\nu}_{x}(x), z - x \rangle, \ \forall z \quad , \tag{4}$$

where $\hat{F}^{\nu}_{x}(x)$ is a subgradient (generalized gradient). It should be noted that the notation $\hat{F}^{\nu}_{x}(x)$ for a subgradient used here is convenient in cases where a function depends on several groups of variables and the subgradient is to be taken with respect to one of them (this occurs in minimax problems, two-stage stochastic programming problems etc., which are considered later).

For such problems, a number of iterative deterministic methods are known. In these methods the sequence of approximate solutions $x^0, x^1, .., x^s, ...$ is created by means of precise evaluations of the function $F^{\nu}(x)$ and subgradient $\hat{F}^{\nu}_{x}(x)$ at each point $x=x^s$, $s=0,1,...$

In stochastic quasigradient methods, the sequence of approximates $x^0, x^1, ..., x^s, ...$ is constructed by using statistic estimates of the $F^{\nu}(x^s)$ and $\hat{F}^{\nu}_{x}(x^s)$. In SQG methods instead of the precise values of $\hat{F}^{\nu}(x^s)$, $\hat{F}^{\nu}_{x}(x^s)$, random numbers $\eta_{\nu}(s)$ and random vectors $\xi^{\nu}(s)$ are used such that the conditional mathematical expectation satisfy

$$E\{\eta_{\nu}(s) \,|\, x^0, .., x^s\} = F^{\nu}(x^s) + a_{\nu}(s) \quad , \tag{5}$$

$$E\{\xi^{\nu}(s) \,|\, x^0, .., x^s\} = \hat{F}^{\nu}_{x}(x^s) + b^{\nu}(s) \quad , \tag{6}$$

where the numbers $a_{\nu}(s)$ and the vectors $b^{\nu}(s)$ may depend on $(x^0, .., x^s)$. For exact convergence to an optimal solution, the values $a_{\nu}(s)$, $\| b^{\nu}(s) \|$ must be small (in a certain sense) when $s \to \infty$. At some time we must have that

$$a_{\nu}(s) \to 0, \| b^{\nu}(s) \| \to 0 \tag{5'}$$

directly or in such a way that (compare with (4)).

$$F^{\nu}(x^*) - F^{\nu}(x^s) \geq \langle E\{\xi^{\nu} \,|\, x^0, .., x^s\}, x^* - x^s \rangle + \gamma_{\nu}(s) \quad , \tag{6'}$$

where $\gamma_\nu(s) \to 0$ as $s \to \infty$ and x^* an optimal solution. The vector $\xi^s(s)$ is called a stochastic quasi-gradient when $b^\nu(s) \not\equiv 0$, or stochastic generalized gradient (stochastic gradient for differentialbe function $F^\nu(x)$) when $b^\nu(s) \equiv 0$. For a better grasp of these concepts, it is important to discuss some difficult problems and to see that usually $\eta_\nu(s), \xi^\nu(s)$ are easily calculted.

3. EXAMPLE PROBLEMS

3.1 A General Problem of Stochastic Programming

A rather general problem of the stochastic programming can be formulated as the minimization of

$$F^0(x) = Ef^0(x,w) \tag{7}$$

subject to

$$F^i(x) = Ef^i(x,w) \leqslant 0, \quad i=\overline{1,m}, \tag{8}$$

$$x \in X \subseteq R^n \quad , \tag{9}$$

where E is the operation of mathematical expectation with respect to some probability space $(\Omega; F; P)$; $f^\nu(x,w)$, $\nu=\overline{0,m}$, are random functions possessing all the properties necessary for the expressions (7) and (8) to be meaningful. For example, the constraints

$$P\{\sum_{j=1}^{n} a_{ij}(w)x_j \leqslant b_i(w)\} \geqslant p_i, i=\overline{1,m} \tag{10}$$

of the stochastic programs with chance-constraint would be of the type (8), if we assume that

$$f^i(x,w) = \begin{cases} p_i - 1, & \text{when } \sum_{j=1}^{n} a_{ij}(w)x_j \leqslant b_i(w) \\ p_i, & \text{when } \sum_{j=1}^{n} a_{ij}(w)x_j > b_i(w) \end{cases} . \qquad (10')$$

The problem (7) - (9) is more difficult than the common nonlinear programming problem. The main difficulty of this problem is that, as a rule, the calculation of exact values of the functions

$$F^\nu(x) = Ef^\nu(x,w) = \int f^\nu(x,w)P(dw)$$

is feasible only in exceptional cases for special types of probability measures $P(dw)$ and random functions $f^\nu(x,w)$, $\nu = \overline{0,m}$. For instance, to calculate the values of the constraint functions (10), it is necessary to find the probability of the event

$$\{w \mid \sum_{j=1}^{n} a_{ij}(w)x_j \leqslant b_i(w)\}$$

as a function of $x = (x_1,\ldots,x_n) \in X$. Generally speaking, this is possible only in rare cases, this distribution may depend dramatically on x (compare $x = (0,\ldots,0)$ and $x = (1,\ldots,1)$). The computing of the exact values of the functions $F^\nu(x)$ is out of question in those cases when the distribution $P(dw)$ is unknown and only some observations $w^0,w^1,\ldots,w^s,\ldots$ of the random element w is available at each iteration $s = 0,1,\ldots$ Such situations are typical in the optimization of systems when the values of the characteristics of the system output are obtained through real measurement or through Monte Carlo simulation.

For the stochastic problem (7) - (9), in practice it is only possible to calculate random realizations $f^\nu(x,w)$ of the functions $F^\nu(x)$. In such cases we can take

$$\eta_\nu(s) = f^\nu(x^s, w^s), \quad \nu = \overline{0,m} \quad ,$$

where the w^s results from mutually independent samples of w. We have

$$E\{\eta_\nu(s) \,|\, x^s\} = E\{f^\nu(x^s, w) \,|\, x^s\} = F^\nu(x^s), \quad \nu = \overline{0,m} \quad .$$

If the functions $F^\nu(x)$ have uniformly bounded second derivatives at $x \in \{x^s\}_{s=0}^{\infty}$ then for the random vectors

$$\xi^\nu(s) = \sum_{j=1}^{n} \frac{f^\nu(x^s + \Delta_s e^j, w^{sj}) - f^\nu(x^s, w^{s0})}{\Delta_s} \tag{11}$$

we would have

$$E\{\xi^\nu(s) \,|\, x^s\} = F^\nu_x(x^s) + b^\nu(s) \quad , \quad \|b^\nu(s)\| \leqslant \text{const.}\Delta_s \quad ,$$

where e^j is the unit vector on the j-th axis; $\Delta_s > 0$; $\{(w^{s0}, w^{s1}, \ldots, w^{sn})\}_{s=0}^{\infty}$ are a result of independent $s = 0,1,\ldots,$ samples of w (we could have $w^{s0} = s^{s1} = \ldots = w^{sn}$). For the vector

$$\xi^\nu(s) = (3r_s/2) \sum_{k=1}^{r_s} \frac{f^\nu(x^s + \Delta_s h^k, w^{sk}) - f^\nu(x^s, w^{s0})}{\Delta_s} h^k \quad , \tag{12}$$

where $r_s \geqslant 1, h^1, \ldots, h^{r_s}$ are observations of the random vector $h = (h_1, \ldots, h_n)$ whose components are independently and uniformly distributed over [-1,1]

$$E\{\xi^\nu_j(s) \,|\, x^s\} = (3r_s/2)\, E \sum_{k=1}^{r_s} \frac{F^\nu(x^s + \Delta_s h^k) - F^\nu(x^s)}{\Delta_s} h^k_j =$$

$$= (3r_s/2)\, E \sum_{k=1}^{r_s} F^\nu_{x_j}(x^s)\, h^k_i h^k_j + \Delta_s \alpha_i(s) = F_{x_j}(x^s) + \Delta_s \alpha_j(s) \quad .$$

Since the second derivatives of the functions $F^{\nu}(x)$ are bounded then $|\alpha_j(s)| <$ const. It is remarkable that independent of the dimensionality of the problem, the vector (12) can be found by calculating the functions $f^{\nu}(x,w)$ at $(r_s + 1)$ points only, $r_s \geqslant 1$.

3.2 Recourse Problems

The simplest well-known recourse problem (two-stage stochastic programming problem) may be formulated in the following way: to find a vector $x \geqslant 0$ such that the function

$$F^0(x) = Ef^0(x,w) \quad ,$$

$$f^0(x,w) = \langle c,x \rangle + \langle d,y(x,w) \rangle =$$

$$= \langle c,x \rangle + \min \{ \langle d,y \rangle \,|\, y \geqslant b - Ax \} \quad ,$$

has the minimum value, where all coefficient $w = (d,b,A,D)$ may be random variables.

Problems of this kind often appear in long-term planning. It is often necessary to choose a production plan or make some other decision which takes into account possible variations in the exogenous parameters and which are resilient to random variations of the initial data. For this purpose the notion of a correction y is introduced and the losses $\langle d,y \rangle$ connected with this correction have to be considered. An optimal long-term plan x should minimize the total expenditures of the plan's realization and its possible corrections. In a two-stage problem the long-term decision x is made in advance, before observation of w; a corrective solution y is derived from the known w and x.

The objective function $F^0(x)$ of this problem is a convex one, but in general nonsmooth, since the minimization operator is present under the integral sign. The random realization of $F^0(x^s)$, a statistical estimate of $F^0(x^s)$

$$\eta_0(s) = f^0(x^s,w^s) = \langle c,x^s \rangle + \langle d,y(x^s,w^s) \rangle$$

is calcutlated without any difficulties. To calculate $F^0(x)$ it is necessary to find the distribution of the $\langle d,y(x,w) \rangle$ as a function of x and then to compute the corresponding integral, which is possible only in rare cases. A stochastic estimate of a subgradient $F^0_x(x)$ at $x = x^s$ looks as follows:

$$\xi^0(s) = c + u(x^s,w^s)\ A(w^s) \quad ; \tag{13}$$

the w^s, $s = 0, 1, \ldots$ are mutually independent samples of w, and the $u(x^s,w^s)$ are a dual variables corresponding to a second-stage optimal plan $y(x^s,w^s)$. It can be shown that under any reasonable choice for the $u(x^s,\cdot)$, see [2], [5], we have that

$$E\{\xi^0(s) \,|\, x^s\} = \hat{F}^0_x(x^s) \quad .$$

3.3 The Stochastic Minimax Problems

Stochastic minimax problems are, at least formally, closely related to recourse problems, but their specific structures allows for a more detailed analysis. The objective function of the simplest stochastic minimax problem (see [1], [3], [5], [13] and [33]) takes on the form

$$F^0(x) = Ef^0(x,w) = E \max_{1 \leq i \leq m} \left[\sum_{j=1}^{n} a_{ij}(w)\ x_j + b_i(w) \right] . \tag{14}$$

Many inventory models are of this type: suppose that the decision about the stock-size x must be make before the information about the demand w is available, the optimal stock-size minimizes the expected cost, i.e.,

$$F^0(x) = cx + E \max \{\alpha(x-w), \beta(w-x)\} \quad , \tag{14'}$$

where c is the unit cost of the product (at delivery), α is the unit storage cost and β is the unit shortage cost.

A more general stochastic minimax problem is to minimize the objective

$$F^0(x) = E \max_{y \in Y} g(x,y,w) = E\, g(x,y(x,w),w) \qquad (15)$$

subject to the constrainsts (8) and (9). In this model, for decision making under uncertainty, the three variables x,y and w contribute to the evnetual choice of a decision. Naturally, the x are the decision variables themselves, the y variables are there to take into account the worst case whereas the w variables can be viewed as the states of nature with either a known a priori probability measure or one that can be obtained through Monte Carlo simulation. The criterion (15) is a mixture between a purely minimax one, such as

$$\max_{\substack{y \in Y \\ w \in \Omega}} g(x,y,w)$$

and the Bayesian criterion

$$E_{Y,\Omega} g(x,y,w) \quad ,$$

where some probability measure is assigned to Y, e.g., the uniform distribution if Y is bounded.

Here it is quite easy to obtain a stochastic estimate of the value of the objective function F^0 at any point x^s. For instance, if F^0 is given by (14)

$$\eta_0(s) = \max_{1 \leq i \leq m} \left[\sum_{j=1}^{n} a_{ij}(w^s) x_j^s + b_i(w^s) \right]$$

and more generally, when F^0 is defined by (15)

$$\eta_0(s) = g(x^s, y^s, w^s) \quad ,$$

where y^s is an approximation to $y(x^s,w^s)$ - a point that maximizes $g(x^s,\cdot,w^s)$ on Y - with

$$g(x^s,y(x^s,w^s),w^s) - g(x^s,y^s,w^s) \leqslant \varepsilon_s \quad , \tag{16}$$

where $\varepsilon_s \to 0$ as $s \to \infty$. A statistical estimate of the generalized gradient $\hat{F}^0_x(x^s)$ for the problem (15) takes on the form

$$\xi^0(s) = \hat{g}_x(x^s,y,w^s)\big|_{y = y(x^s,w^s)} \quad , \tag{17}$$

where $g(x,y,w)$ is assumed to be a convex function with respect to x. It is easy to show that

$$E\{\xi^0(s)\,|\,x^s\} = \hat{F}^0_x(x^s) \quad .$$

To see this, recall that $g(\cdot,y,w)$ is convex and thus

$$g(x,y(x,w^s),w^s) - g(x^s,y(x^s,w^s),w^s) \geqslant g(x,y(x^s,w^s),w^s) -$$
$$-g(x^s,y(x^s,w^s),w^s) \geqslant \langle \hat{g}_x(x^s,y(x^s,w^s),w^s),\, x - x^s\rangle =$$
$$\langle \xi^0(s), x-x^s\rangle \quad .$$

Taking conditional expectation on both side, we get

$$F^0(x) - F^0(x^s) \geqslant \langle E\{\xi^0(s)\,|\,x^s\},\, x - x^s\rangle \quad ,$$

from which the assertion follows. Instead of $y(x^s,w^s)$ we can use also y^s (see (16)). It is easy to see that

$$\xi^0(s) = \hat{g}_x(x^s,y,w^s)\big|_{y = y^s} \tag{18}$$

satisfy the conditions (6'). In (17) and (18) we can apply also the approximation (11) or (12) for computing the gradient g_x (for a differentiable function $g(\cdot,y,w^s)$).

3.4 Nonlinear Programming Problems, Optimization of Large-Scale Systems

If differentiable functions $F^{\nu}(x_1,\ldots,x_n)$ of linear programming problems have a great number of variables, then the calculation of gradient $F^{\nu}_x(.) = (F^{\nu}_{x_1},\ldots,F^{\nu}_{x_n})$ would require computing a great number of different derivatives $F^{\nu}_{x_i}$, $i=\overline{1,n}$. It can be shown that the random vector

$$\xi^{\nu}(s) = \sum_{k=1}^{r_s} \frac{F^{\nu}(x^s + \Delta_s h^k) - F^{\nu}(x^s)}{\Delta_s} h^k \quad ,$$

similar to the (12) is the stochastic quasigradient of $F^{\nu}(x)$ at $x = x^s$ and computing of this vector requires only the calculation of the function $F^{\nu}(x)$ in $(r_s + 1)$ points, $r_s \geq 1$, independent of the dimensionality of x.

It should be noted also that the recourse problem is strongly connected with large scale linear programming problems. For instance, if w has a discrete distribution, i.e., $w \in \{1,2,\ldots,N\}$ and $w = k$ with probability p_k, then the initial problem becomes

$$\langle c,x\rangle + \langle d(1),y(1)\rangle + \langle d(2),y(2)\rangle + \ldots + \langle d(N),Y(N)\rangle = \min$$

$$A(1)x + D(1)y(1) \geq b(1)$$

$$A(2)x + D(2)y(2) \geq b(2)$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$A(N)x + D(N)y(N) \geq b(N)$$

$$x \geq 0,\ y(1) \geq 0,\quad y(2) \geq 0,\ldots\ldots\ldots\ldots,\ y(N) \geq 0 \ ,$$

where y(k) is the correction of the plan x if w = k. The number N may be very large. If only the vector $b = (b_1,\dots,b_m)$ is random and each of the components has two independent outcomes, then $N = 2^m$. The use of the stochastic quasigradient (13) for solving such a problem allows us to solve extremely large-scale problems.

4. METHODS FOR CONVEX FUNCTIONS

4.1 The Projection Method

Suppose we have to minimize a convex continuous function $F^0(x)$ in $x \in X \leqslant R^n$, where X is a compact convex set such that a projection π_x on X can easily be calculated, e.g., $X = \{x \mid a \leqslant x \leqslant b\}$. Let X* be a set of optimal solutions. The method is defined by the relations:

$$x^{s+1} = \pi_x(x^s - \rho_s \xi^0(s)), s = 0,1,\dots \tag{19}$$

$$F^0(x^*) - F^0(x^s) \geqslant \langle E\{\xi^0(s) \mid x^0,\dots,x^s\}, x^* - x^s\rangle + \gamma_0(s), \tag{20}$$

where ρ_s is the step size, $\gamma_0(s)$ may depend on $(x^0,\dots,x^s), x^* \in X^*$. This method was proposed and studied in [1] - [3], [5]. If $\xi^0(s) = \hat{F}^0_x(x^s)$, we obtain the generalized gradient method which was suggested by Shor [36] and was studied by the author [37] and Poljak [38]. If $X = R^n$,

$$F^0(x) = Ef^0(x,w) \quad ,$$

$$\xi^0(s) = \sum_{j=1}^{n} \frac{f^0(x^s + \Delta_s e^j, w^{sj}) - f^0(x^s, w^{s0})}{\Delta_s} e^j \quad ,$$

then the method suggested by (19) corresponds to the well-known stochastic approximation methods which were developed by Robbins and Monro, Kiefer and Wolfowitz, Dvoretsky, Blum and others.

It was shown that under natural assumptions, that are also those of interest in practice, the iterative method defined by (19), converges to a set of minimum points of the original

problem with probability 1. The proof of this fact is based on the notion of a stochastic quasi-Feyer sequence [3]. A sequence $\{z^s\}_{s=0}^{\infty}$ is a Feyer sequence for a set $Z \subset R^n$ if [10]

$$\|z - z^{s+1}\| < \|z - z^s\| , \forall z \in Z .$$

A sequence of random vectors $\{z^s\}_{s=0}^{\infty}$ defined on a probability space $(\Theta, \mathcal{R}, \mu)$ is a stochastic quasi-Feyer sequence [3] for a set $z \subset R^n$, if $E\|z^0\|^2 < \infty$, and for any $z \in Z$

$$E\{\|z - z^{s+1}\|^2 (z^0,\ldots,z^s\} \leqslant \|z - z^s\|^2 + d_s, \quad s = 0,1,\ldots$$

$$d_s \geqslant 0, \quad \sum_{s=0}^{\infty} E\, d_s < \infty . \tag{21}$$

Theorem 1 [5, p.98]. If $\{z^s\}$ is a stochastic quasi-Feyer sequence for a set Z, then:

a) the sequence $\|z - z^{s+1}\|^2, s=0,1,$ converges with probability 1 for any $z \in Z \qquad E\|z - z^s\|^2 < C < \infty$,

b) the set of accumulation points of $\{z^s(\theta)\}$ is not empty for almost all θ,

c) if $z'(\theta), z''(\theta)$ are a two distinct accumulation points of the sequence $\{z^s(\theta)\}$ which do not belong to the set Z then Z lies in the hyperplan equidistant from the point $z'(\theta), z''(\theta)$.

The fact (a) would follow from convergence of super martingale.

$$\upsilon_s = \|z - z^s\|^2 + \sum_{k=s}^{\infty} d_k , \quad \upsilon_s \geqslant 0 ,$$

$$E\{\upsilon_{s+1} | \upsilon_s\} \leqslant \upsilon_s ,$$

if d_s independs on $(x^0,\ldots,x^s)$. The (c) follows from the equality

$$\|z - z'\|^2 - \|z - z''\|^2 = 2(z, z''-z') + \|z'\|^2 + \|z''\|^2 = 0 .$$

Consider now a simple version of the convergence theorem for the iterative procedure (19) to illustrate the techniques of proof.

Theorem 2. Assume that

a) $F^0(x)$ is a convex continuous function,

b) X is a convex compact set,

c) $E\{\|\xi^0\|^2 | x^0,\dots,x^s\} \leqslant$ const and also that the parameters ρ_s, d_s satisfy with probability 1 the conditions

$$\rho_s \geqslant 0, \sum_{s=0}^{\infty} \rho_s = \infty, \sum_{s=0}^{\infty} E\{\rho_s |\gamma_0(s)| + \rho_s^2\} < \infty \quad , \tag{22}$$

Then $\lim x^s \in X^*$ with probability 1.

Note that the requirements (b), (c) are not too stringent for most applications. In practice (c) is the consequence of (b) and finite distributions of random parameters. The condition (22) for the random vector $\xi^0(s)$ defined by (11) or (12) signifies that in (11) or (12) the step-size Δ_s of the finite difference approximations to the gradient and the step-size ρ_s used in the procedure described by (19), must be subjected to the conditions

$$\rho_s \geqslant 0, \sum_{s=0}^{\infty} \rho_s = \infty \quad , \quad \sum_{s=0}^{\infty} (\rho_s \Delta_s + \rho_s^2) < \infty \quad ,$$

when the parameters ρ_s, Δ_s are chosen independently of $(x^0,\dots,x^s)$.

Proof of Theorem 2:

The properties of the projection π_x yield for any $x^* \in X$

$$E\{\|x^* - x^{s+1}\|^2 | x^0,\dots,x^s\} \leqslant \|x^* - x^s\|^2 + 2\rho_s \langle E\{\xi^0(s) | x^0,\dots,x^s\}, x^* - x^s\rangle +$$

$$+ \rho_s E\{\|\xi^0(s)\|^2 | x^0,..,x^s\} \quad .$$

By the assumption (c) and (20) (taking into account that $F(x^*) - F(x^s) \leqslant 0$)

$$E\{\|x^* - x^{s+1}\|^2 | x^0, .., x^s\} \leqslant \|x^* - x^s\|^2 + C(\rho_s |\gamma_o(s)| + \rho_s^2) \quad ,$$

where C is a constant.

In view of (22) and by the definition (21), it means that $\{x^s\}$ is indeed a stochastic quasi-Feyer sequence for this set X^*. Consequently, the sequence $\|x^* - x^s\|^2$, $s = 0, 1, ..$ converges with probability 1 for any $x^* \in X^*$, the set of accumulation points of $\{x^s\}$ is not empty. If we show that one of the accumulation points of $\{x^s(\theta)\}$ belongs to X^* for almost all θ, then from assertion (c) of Theorem 1 would follow the convergence of $\{x^s\}$ with probability 1 to a point of X^*.

Consider the inequality

$$E\|x^* - x^{s+1}\|^2 \leqslant E\|x^* - x^0\|^2 + 2 E \sum_{k=0}^{s} \rho_k \langle E\{\xi^0(k) | x^0, .., x^k\}, x^* - x^k \rangle + C \sum_{k=0}^{s} E\rho_s^2 .$$

Due to the inequality (20)

$$E\|x^* - x^{s+1}\|^2 \leqslant E\|x^* - x^0\|^2 + 2 E \sum_{k=0}^{s} \rho_k (F^0(x^*) - F^0(x^s)) + C \sum_{k=0}^{s} E\{\rho_k |\gamma_o(k)| + \rho_k^2\}$$

from which we get

$$E \sum_{k=0}^{\infty} \rho_k (F^0(x^*) - F^0(x^k)) > \infty \quad .$$

Since

$$\sum_{k=0}^{\infty} \rho_k = \infty \text{ and } F^0(x^*) - F^0(x^k) \leqslant 0 \quad ,$$

there exists a subsequence x^{k_s} such that $F^0(x^*)-F(x^{k_s}) \to 0$, and this completes the proof.

The methods which we shall consider below, converge under conditions approximately analogous to those mentioned above. Theorem 2 establishes the convergence of the iterative procedure (19) with probability 1. Such a convergence is important in many applications. If $\gamma_0(s)\equiv 0$ and if instead of (22) only the conditions

$$\rho_s \uparrow 0, \quad \sum_{s=0}^{\infty} \rho_s = \infty$$

hold, then it can be shown [5], that

$$\inf_{x^*} E\|x^* - x^s\|^2 \to 0 \quad .$$

In [65] the following idea was proposed for estimating efficiently the vector

$$\bar{x}^s = \left(\sum_{k=0}^{s} \rho_k x^k\right)\left(\sum_{k=0}^{s} \rho_k\right)^{-1} \quad .$$

This depends on the parameters $\rho_k, \gamma_0(k)$. From the inequality

$$E\|x^* - x^{s+1}\|^2 \leqslant E\|x^* - x^0\|^2 + 2E \sum_{k=0}^{s} \rho_k (F^0(x^*) - F^0(x^k)) +$$

$$+ C \sum_{k=0}^{s} E\{\rho_k |\gamma_0(k)| + \rho_k^2\}$$

we have that

$$2E \sum_{k=0}^{s} \rho_k (F^0(x^k) - F^0(x^*)) \leqslant E\|x^* - x^0\|^2 + C \sum_{k=0}^{s} E\{\rho_k |\gamma_0(k)| + \rho_k^2\} \quad .$$

If the ρ_k are independent of $(x^0,\ldots,x^k)$, then

$$\left(\sum_{k=0}^{s} \rho_k\right)^{-1} E \sum_{k=0}^{s} \rho_k (F^0(x^k) - F^0(x^*)) \geqslant EF^0(\bar{x}^s) - F^0(x^*)$$

and we have such estimation

$$EF^0(\bar{x}^s) - F^0(x^*) \leqslant \left(2 \sum_{k=0}^{s} \rho_k\right)^{-1} (E\|x^* - x^\circ\|^2 + C \sum_{k=0}^{s} (\rho_k |\gamma_0(k)| + \rho_k^2)) \quad .$$

4.2 Penalty Function Methods

Constraints of type (2) of the general problem (1) - (3) can be taken into account by means of penalty functions and instead of the original problem, we can minimize a penalized function, for instance

$$\psi(x,c) = F^0(x) + c \sum_{i=1}^{m} \min\{0, F^i(x)\}$$

on the set X. A generalized gradient of $\psi(x,c)$ at $x = x^s$ is

$$\hat{F}_x^0(x^s) + c \sum_{i=1}^{m} \min\{0, F^i(x^s)\}\, \hat{F}_x^i(x^s) \quad .$$

If the exact values of $F^i(x^s), \hat{F}_x^0(x^s), \hat{F}_x^i(x^s)$ are known, then a deterministic generalized gradient procedure can be used for

minimizing $\psi(x,c)$. The penalty function methods for a problem with known values of the constraint functions $F^i(x^s)$ was considered in [48], [66]. In such cases the projection method (19) is applicable to minimizing $\psi(x,c)$. In general, if instead of the values $F^\nu(x^s), \hat{F}^\nu_x(x^s), \nu = \overline{0,m}$, only statistical estimations $\eta_\nu(s), \xi^\nu(s)$ are available, it is impossible to actually find $\min\{0, F^i(x^s)\}$. How to handle this situation was studied in [4]. Because of the inherent difficulties in estimating the subgradient of the function $\psi(x,c)$, we are led to the following variant of the iterative scheme studied in the previous section.

$$x^{s+1} = \pi_x(x^s - \rho_s[\xi^0(s) + c\sum_{i=1}^{m} \min\{0, \beta_i(s)\}\xi^i(s)]), \quad (23)$$

$$\beta_i(s+1) = \delta_s\eta_i(s) + (1 - \delta_s)\beta_i(s), \; i = \overline{1,m} \quad , \quad (24)$$

where δ_s is the step-size and

$$E\{\eta_i(s) | x^0,\ldots,x^s\} = F^i(x^s) + a_i(s) \quad ,$$

$$F^\nu(x^*) - F^\nu(x^s) \geqslant \langle E\{\xi^\nu(x^s) | x^0,..,x^s\} + \gamma_\nu(s) \quad .$$

For convergence with probability 1 of these kinds of procedures in addition to (22), we must demand that with probability 1

$$\delta_s > 0, \; \rho_s/\delta_s \to 0, \; \sum_{s=0}^{\infty}\sum_{i=1}^{m} E\{\rho_s|\gamma_i(s)| + \delta_s|a_i(s)|\} < \infty \quad .$$

It is worthwhile to note that the above mentioned method may not converge when $\delta_s \equiv 1$. i.e., for $\beta_i(s) \equiv \eta_i(s)$. If $\delta_s = 1/(s+1)$ then

$$\beta_i(s) = \frac{1}{s}\sum_{k=0}^{s} \eta_i(k) \quad .$$

The averaging procedure of the type (24) proved to be very useful of SQG methods. In particular, Gupal [8] has studied the method characterized by the relations:

$$x^{s+1} = \pi_X(x^s - \rho_s \zeta^s) \quad , \tag{25}$$

$$\zeta^s = \begin{cases} \xi^0(s), \text{ if } \beta_{i_s}(s) = \max\limits_{1 \leq i \leq m} \beta_i(s) \leq 0 \quad , \\ \xi^{i_s}(s), \text{ if } \beta_{i_s}(s) > 0 \quad . \end{cases}$$

The requirements for convergence of this method are similar to those for the method (23).

4.3 The Linearization Method

Let the function $F^0(x)$ have continuous derivatives. If $F^0(x^s)$ and $F^0_x(x^s)$ are known, then the standard linearization method is defined by the relations

$$x^{s+1} = x^s + \rho_s(\bar{x}^s - x^s) \quad ,$$

$$\langle F^0_x(x^s), \bar{x}^s \rangle = \min_{x \in X} \langle F^0_x(x^s), x \rangle \quad ,$$

$$F^0(x^{s+1}) = \min_{0 \leq \rho \leq 1} F^0(x^s + \rho(\bar{x}^s - x^s)) \quad .$$

The stochastic variant of this method has been studied in [6], [30] and is defined by the relations

$$x^{s+1} = x^s + \rho_s(\bar{x}^s - x^s) \quad , \tag{26}$$

$$\langle v^0(s), \bar{x}^s \rangle = \min_{x \in X} \langle v^0(s), x \rangle \quad ,$$

$$v^0(s+1) = \delta_s \xi^0(s) + (1 - \delta_s) v^0(s) \quad ,$$

where ρ_s, δ_s satisfy conditions similar to those of the previous section. Notice that if instead of $v^0(s)$ the vectors $\xi^0(s)$ are used ($\delta_s \equiv 1$) then, some simple examples show that the method may not converge.

4.4 The Lagrange Multiplier Method

The method is characterized by the relations

$$x^{s+1} = \pi_X(x^s - \rho_s[\xi^0(s) + \sum_{i=1}^{m} u_i^s \xi^i(s)]) \quad , \tag{27}$$

$$u_i^{s+1} = \max\{0, u_i(s) + \delta_s \eta_i(s)\}$$

and when $X = R^n$, $\delta_s \equiv \rho_s \equiv \text{const}, \xi^\nu(s) = F_x^\nu(x^s), \eta_i(s) = F^i(x^s)$, $1 = \overline{i,m}$, and the $f^\nu(x)$, $\nu = \overline{0,m}$ are smooth it is a deterministic algorithm proposed in [54]. The stochastic version of this method was studied in [1], [5], where it was proved that the $\min_{k \le s} F^0(x^k)$ to $\min F^0(x)$ converge with probability 1, provided that $F^0(x)$ is strictly convex and $\delta_s \equiv \rho_s$. The convergence for convex functions $F^0(x)$--not necessarily strictly convex--was studied in [21] with assumptions that $\rho_s/\delta_s \to 0$.

5. SQG Methods for Nonconvex Functions

The convergence of SQG methods for nonconvex differentiable functions was studied in [3], [5]. In [12] Nurminski considered the case at non-convex non-differentiable functions $F^\nu(x)$ satisfying the inequality

$$F^\nu(z) - F^\nu(x) \geqslant \langle \hat{F}_x^\nu(x), z - x \rangle + o(\|z - x\|) \quad .$$

Such functions are called weakly convex. The class of weakly convex functions includes convex functions as well as nonconvex differentiable. Moreover, the maximum of a collection of weakly convex functions is also a weakly convex. This case needs new techniques for proving the convergence [11] and later on this technique was widely used for proving the convergence of various algorithms (see [5], [10], [13]). This technique relies on arguing by contradiction.

Let us assume that X^* is a set of solutions, $\{x^s(\theta)\}$ is a random sequence of approximations. Then we obtain (see[5]) the following generalization of Nurminski's results [11].

Theorem 3 [5, p. 181]. Suppose that $X^* \subset R^n$ is closed and $\{x^s(\cdot)\}_{s=0}^{\infty}$ is a random sequence of vectors in R^n defined on a probability space $(\Theta, \mathcal{R}, \mu)$. Moreover, suppose that almost surely

1) for all s, $x^s(\theta) \in K(\theta)$ with $K(\theta)$ compact
2) for any subsequence $\{x^{s_k}(\cdot)\}$ with $\lim x^{s_k} = x'$
   a) if $x' \in X^*$, then $\|x^{s_k+1} - x^{s_k}\| \to 0$ as $k \to \infty$
   b) if $x' \notin X^*$, then for ε sufficiently small and for any s_k

$$\tau_k = \min \{s \mid_{s \geqslant s_k} \|x^{s_k} - x_s\| > \varepsilon\} < \infty \quad .$$

3) there exists a continuous function $V(x)$ such that the set $V(X^*)$ is at most countable and with probability 1

$$\lim_{k\to\infty} V(x^{\tau_k}) < \lim_{k\to\infty} V(x^{s_k}) \quad .$$

Then for almost all θ in Θ

$$V(x^s(\theta)) \to V(x^*(\theta))$$

and $x^*(\theta) \in X^*$.

The conditions of this theorem are similar to necessary and sufficient convergence conditions, proposed by Zangwill (see [69]). However, Zangwill's conditions are very difficult to verify for a nondescent procedure.

Conditions (2) of Theorem 3 prevent all sequence $\{x^s\}$ converge to limit point x', which does not belong to the set X^*. However, condition (2) alone does not prevent "cycling", i.e., such a behavior of $\{x^s\}$ that it will be visiting any neighborhood of $x' \notin X^*$ infinitely many times. To exclude such a case the condition (3) is imposed, which guarantees that the sequence $\{x^s\}$ will be leaving a neighborhood of x' with decreasing values of some Lypunov functions $V(x)$. Later we shall illustrate the use of this theorem.

Gupal [9], [10], [32] studied SQG methods with functions satisfying a local Lipschitz condition. This approach is based on the limit extremal problem idea [14], [43].

6. LIMIT EXTREMAL PROBLEMS, NONSTATIONARY OPTIMIZATION

Briefly, the essence of this idea is the following: suppose we have to minimize a function $f^0(x)$ of a rather complex nature, for example, it does not have continuous derivatives. Consider the sequence of the "good" functions $F^0(x,s)$, for instance smooth, converging to $f^0(x)$ for $s \to \infty$. Now consider the procedure

$$x^{s+1} = x^s - \rho_s F^0_x(x,s), s = 0,1,.... \qquad (28)$$

Under rather general conditions ($\rho_s \downarrow 0, \Sigma\rho_s = \infty$) it is possible to show (see [5], [17] and Theorem 4) that $F^0(x^s,s) \to \min f^0(x)$.

Often approximate the functions may have the form of mathematical expectations

$$F^0(x,s) = \int f^0(x+h)P_s(dh) = Ef^0(x+h(s)) \quad , \qquad (29)$$

where the measure $P_s(dw)$ for $s \to \infty$ is centered at the point 0. Hence instead of the procedure given by (28) that requires the exact value of the gradient of the mathematical expectation, we can use the ideas of the stochastic quasigradient methods.

For example, see [9], let h(s) be random vectors with independent components uniformly distributed on $[-\Delta_s/2, \Delta_s/2]$, $\Delta_s \to 0$ for $s \to \infty$, and suppose that $f^0(x)$ satisfies a local Lipschitz condition, then the function $F^0(x,s)$ is smooth and $F^0(x,s) \to f^0(x)$ uniformly on any bounded domain. Consider the stochastic procedure

$$x^{s+1} = x^s - (\rho_s/\Delta_s) \sum_{j=1}^{n} [f^0(\tilde{x}^s_1,...,x^s_j + \frac{\Delta_s}{2},...,\tilde{x}^s_n) - f^0(\tilde{x}^s_1,...,x^s_j - \frac{\Delta_s}{2},...,\tilde{x}^s_n)],$$

where the $\tilde{x}^s_i$ are independent random quantities uniformly

distributed on intervals $[x_i^s - \frac{r_s}{2}, x_i^s + \frac{r_s}{2}]$. It can be shown that $E\{\xi^0(s) \mid x^s\} = F_x^0(x^s,s)$,

where

$$\xi^0(s) = (1/\Delta_s) \sum_{j=1}^{n} [f^0(\tilde{x}_1^s,..,\tilde{x}_j^s + \frac{\Delta_s}{2},..,\tilde{x}_n^s) - f^0(\tilde{x}_1^s,..,x_j^s - \frac{\Delta_s}{2},..,\tilde{x}_n^s)]\, e^j \quad (30)$$

and $F_x^0(x^s,s)$ is the gradient of the function (29). The convergence of this method with probability 1 is then proved under assumptions that

$$\rho_s > 0, \sum_{k=0}^{\infty} \rho_k = \infty, \sum_{k=0}^{\infty} \rho_k^2 > \infty, \frac{\rho_k}{\Delta_k} \to 0, \Delta_k \to 0, \frac{\Delta_k - \Delta_{k+1}}{\Delta_k \rho_k} \to 0 \quad .$$

In [32] this method was modified to handle semicontinuous functions by smooth functions of the semicontinuous function $f^0(x)$ also being approximated

$$F^0(x,s) = E\, f^0(x+h(s)+t(s)) \quad ,$$

where $h(s),t(s)$ are random independnet vectors with independent components uniformly distributed $[-\Delta_s/2, \Delta_s/2]$. To illustrate the ideas involved in the proof of convergence results, let us consider the following simple case:

Theorem 4. Assume that:

a) $F^0(x,s), f^0(x)$ are convex continuous functions,

b) X is a convex compact set,

c) $F^0(x,s) \to f^0(x)$ uniformly in X,

d) $\| \hat{F}_x^0(x^s,s) \| \leqslant$ const,

$$x^{s+1} = \pi_x(x^s - \rho_s \hat{F}_x^0(x^s,s))$$

and the parameters ρ_s satisfy the conditions

$$\rho_s \downarrow 0, \sum_{s=0}^{\infty} \rho_s = \infty$$

Then $F^0(x^s,s) \to f^0(x^*) = \min f^0(x)$

Proof

The conditions 1,2(a) of Theorem 3 are fulfilled. It suffices to verify the conditions 2(b) and 3. Let $x^{s_k} \to x' \in X^*$, we need to show that $\tau_k < \infty$. We argue by contradiction, to suppose the contrary that $\tau_k = \infty$. For this purpose, we consider the function $V(x) = \min_{x^*} \|x^* - x\|^2$. We have that

$$V(x^{s+1}) = \min_{x^*} \|x^* - x^{s+1}\|^2 = \|x^*(s) - x^{s+1}\|^2 \leqslant V(x^s) +$$

$$2\rho_s \langle \hat{F}^0_x(x^s),\ x^*(s) - x^s \rangle + \rho_s^2 \|\hat{F}^0_x(x^s,s)\|^2 \ .$$

Since $x^{s_k} \to x' \bar{\in} X^*$ and $\|x^s - x^{s_k}\| < \varepsilon$ for sufficiently large s and any ε. Then there exists $\delta > 0$ such that

$$f^0(x^*) - f^0(x^s) < -\delta$$

and for $x^* \in X^*$ we have

$$\langle F^0_x(x^s,s), x^* - x^s \rangle \leqslant F^0(x^*s) - F^0(x^s,s) \leqslant F^0(x^*,s) - f^0(x^*) + f^0(x^s) -$$

$$F^0(x^s,s) < -\frac{\delta}{2} \ .$$

Therefore

$$V(x^{s+1}) \leqslant V(x^s) - \delta\rho_s + c\rho_s^2 = V(x^s) - \rho_s(\delta - c\rho_s) \leqslant = V(x^{s_k}) - \delta \sum_{e=s_k}^{s} \rho_e$$

and for a sufficiently large s, this contradicts the fact that $|V(x)| <$ const when $x \in X^*$. So, condition 2 is satisfied. Looking at condition 3, it is easy to realize that

$$V(x^{\tau_k}) \leqslant V(x^{s_k}) - \delta \sum_{s=s_k}^{s} \rho_s$$

Hence, in view of the properties of π_x ,

$$\varepsilon < \|x^{\tau_k} - x^{s_k}\| \leqslant \sum_{s=s_k}^{\tau_{k-1}} \|x^{s+1} - x^s\| \leqslant C \sum_{s=s_k}^{\tau_k - 1} \rho_s \quad ,$$

where C is a constant. Then

$$V(x^{\tau_k}) \leqslant V(x^{s_k}) - \frac{\varepsilon\delta}{C}$$

or equivalently

$$\overline{\lim}\, V(x^{\tau_k}) < \lim V(x^{s_k})$$

and this completes the proof.

This approach is very important in nonsmooth and particularly in discontinuous optimization. Thus in [30] it is shown that the general linearization scheme (26) may be used for optimizing a function that satisfies a local Lipschitz condition. The convergence with probability 1 of the following methods was investigated:

$$x^{s+1} = x^s + \rho_s(\bar{x}^s - x^s) \quad ,$$

$$\langle v^0(s), \bar{x}^s \rangle = \min \langle v^0(s), x \rangle \quad ,$$

$$v^0(s+1) = \delta_s \xi^0(s) + (1-\delta_s)\, v^0(s) \quad ,$$

where $\xi^0(s)$ is the vector (30), and

$$\rho_s \geqslant 0,\ \rho_s/\Delta_s\delta_s \to 0,\ \Delta_s \to 0 \quad ,$$

$$(\Delta_s - \Delta_{s+1})/\Delta_s\rho_s \to 0,\ \sum_{s=0}^{\infty} \rho_s = \infty, \sum_{s=0}^{\infty} (\rho_s/\Delta_s)^2 < \infty \quad .$$

The systematic study of methods for the solution of general limit extremal problems was undertaken by Verchenko [17]. The general problem was formulated as follows: given a sequence of functions $F^{\nu}(x,s) \to f^{\nu}(x)$, $\nu = \overline{0,m}$. It is necessary to find an optimal solution to the problem

$$\min \{f^0(x) \mid f^i(x) \leqslant 0, i=\overline{1,m}, x \in X\}$$

by using only information about values of the functions $F^{\nu}(x,s)$, $s = 0,1,..$, and their subgradients or statistical estimates of these quantities.

There may be several reasons for considering such problems. One of them - the idea mentioned above of approximating "bad" functions $f^{\nu}(x)$ by a sequence of "good" functions $F^{\nu}(x,s)$. Secondly, the functions $f^{\nu}(x)$ may be defined as $\lim_{s\to\infty} F^{\nu}(x,s)$ and it is very difficult to get an explicit expression for the limit functions. Thirdly, the $F^{\nu}(x,s)$ may be time dependent functions and at iteration s only information about $F^{\nu}(x,s)$ is accessible. The optimization problem with time-varying functions and known trend of the optimal solutions is considered in [55], [56] and [63]. The methods for solving the following general problem on nonstationary optimization were investigated in the articles [15] - [20]: to find a sequence $x^0, x^1, \ldots, x^s, \ldots$, such that

$$\lim_{s\to\infty} [F^0(x^s,s) - \phi(s)] = 0 \quad ,$$

where

$$\phi(s) = \min \{F^0(x,s) \mid F^i(x,s) \leqslant 0, i=\overline{1,m}, x \in X\} \quad .$$

7. APPLICATIONS OF SQG METHODS

The applications of SQG methods to long-term planning problems, optimization of probabilistic systems, decision-making under risk and uncertainty, identification and reliability of systems, inventory control, etc., were considered in [5] and [7]. In this part of the paper we sketch out some of them.

7.1 Optimization of Stochastic Systems

Taking into account the influence of uncertain random factors in optimization of systems leads to stochastic programming problems. The problem (7)- (9) is a model for stochastic systems optimization, when the decision (values to assign to the system parameters) x is chosen in advance, before the random factors w is realized. A stochastic model tends to take into account all possible eventualities for stabilizing the optimal solution with respect to perturbations of the data. There may also be a class of models, when the decision x is chosen only after an experiment over w is realized and x is based on the actual knowledge of the outcomes of this experiment. Such situations occur in real-time control and short-term planning. In practice, these problems are usually rediced to problems of the type (7) - (9) via decision rules.

The formulation of such models can be done - at least formally - in terms of decision function theory. Given probability space (Ω,A,P) of random parameters, the experiment maps (Ω,A,P) in the sample or outcome space. Let B be the subfield associated with this outcome space. If the events of B are to have any relevance as to which decision x is made, then x must depend on w and be a B-measurable function x(w). The problem is to find such B-measurable function x(w), which minimizes

$$F^0(x(w)) = Ef^0(x(w),w) \tag{31}$$

subject to

$$F^i(x(w)) = Ef^i(x(w),w) \leqslant 0, i = \overline{1,m} \quad , \tag{32}$$

$$x(w) \in X \quad . \tag{33}$$

The optimality conditions derived for this problem, in a form which is convenient, for application of SQG methods, have been treated in particular in [5], [7]. Under suitable hypotheses, an optimal solution x(w) is defined (for $X = R^n$) as a function satisfying the conditions: there exist B-measurable functions

$\lambda_\nu w \geqslant 0, \nu = \overline{0,m}$ such that

$$\sum_{\nu=0}^{m} \lambda_\nu E\{f_e^\nu(x,w)|B\} \geqslant 0 \quad ,$$

$$\lambda_i E\{f^i(x,w)|B\} = 0, i = \overline{1,m} \quad ,$$

for any vector $e = (e_1 \ldots, e_n)$, where $f_e^\nu(x,w)$ is the directional derivative. Such optimality conditions reduce the problem (31)-(33) with unknown B-measurable functions to the problem of the type (7)-(9) with $x \in R^n$ and with conditional mathematical expectations. There may be also a way of formulating the original problem directly as the problem of minimizing

$$F^0(x) = E\{f^0(x,w)|B\}$$

subject to

$$F^i(x) = E\{f^i(x,w)|B\} \leqslant 0, i = \overline{1,m} \quad ,$$

$$X \in X \subseteq R^n \quad .$$

The investingation of more general problems with unknown distributions belong to a given class and with associated (simple) numerical procedures that was considered in [5] and more systematically in [22].

In stochastic programming problems with $x \in R^n$, a SQG method can be used to obtain procedures similar to those of stochastic approximation [47], but for more general regression functions and with more general constraints. The problems solvable by stochastic approximation methods (see 4.1) occupy a place in the general range of stochastic programming problems comparable to the place occupied by problems requiring the determination of an unconditioned minimum of a smooth function in the range of nonlinear programming problems.

Consider some of the concrete SQG procedures. From (13) and the convergence of the procedure given by (19) we can obtain the following method for solving a recourse problem.

(i) For given x^s observe the random realization of b, d, A, D, which we note as B(s), D(s), A(s), D(s);

(ii) Solve the problem

$$\langle d^s, y \rangle = \min \quad ,$$

$$D(s)y \geqslant b(s) - A(s)x^s \quad ,$$

$$y \geqslant 0$$

and calculate the dual variables $u(x^s, w^s)$.

(iii) Get

$$\xi^0(s) = c + u(x^s, w^s)\ A(w^s)$$

and change x^s:

$$x^{s+1} = \pi_x(x^s - \rho_s \xi^0(s))$$

It is worthwhile to note that this method can be regarded as a stochastic iterative procedure for the decomposition of large scale problems (see 4.1). It is not difficult to obtain a similarly simple (implementable) procedure for solving other stochastic problems. For instance, by using (17) and (19) one obtains a SQG procedure for stochastic minimax problem (14):

(i) For given x^s observe the realizations $a_{ij}(w^s)$, $b_i(w^s)$.

(ii) Calculate

$$\xi^0(s) = a^{i_s} = (a_{i_s 1}, \ldots, a_{i_s n})$$

$$i_s \in \{k \mid \sum_{j=1}^{n} a_{kj}(w^s) x_j^s + b_k(w^s) = \max_i [\sum_{j=1}^{n} a_{ij}(w^s) x_j^s + b_i(w^s)]\} \ .$$

(iii) Change x^s

$$x^{s+1} = \pi_x(x^s - \rho_s \xi^0(s)) \quad .$$

In particular, in the simplest inventory problem (14') with $x \geqslant 0$

$$x^{s+1} = \max \{0, x^s - \rho_s \xi^0(s)\} \quad ,$$

$$\xi^0(s) = \begin{cases} \alpha, & \text{if } x^s \geqslant w^s \quad , \\ -\beta, & \text{if } x^s < w^s \quad . \end{cases}$$

The methods (23), (25), (26) and others allow us to solve a more difficult problem with constraints of type (8) or with so-called complex functions (see [5]) of the form

$$F^\nu(x) = Eq^\nu(Ef(x,w), x, w) \quad .$$

As an example of a complex criteria, we can consider the penalty function of the general stochastic problem (7)-(9)

$$Ef^0(x,w) + c \sum_{i=1}^{m} \min \{0, Ef^i(x,w)\} \quad ,$$

or the functions

$$E[f^0(x,w) - Ef^0(x,w)]^2 \quad .$$

The main idea of solving the problems with complex functions is similar to (23).

7.2 Multiobjective Problems: Optimization with a Preference Structure

Many complex decision problems involve multiple conflicting objectives. Generally, we cannot optimize several objectives simultaneously, for instance, minimize cost and at the same

time maximize benefits. It would be nice if we could find some function (utility function) that combines all objectives into a scale index of preferability. Then the problem of decision making can be put into the format of the standard optimization problem: to find $x \in X$ to optimize the utility function. The finding of a utility function may be a very difficult problem and often it is easy to have a preference ordering (preference structure) among feasible solutions $x \in X$ and deal with this structure directly to get the prefered solution. This ordering may be based on the decision maker's judgement or other rules, for instance lexicographic ordering. So let us assume that the decision maker has a preference structure at different points $x \in X$ and there exists a utility function (unknown) $U(x)$ such that

$$x' \sim x'' \Leftrightarrow U(x) = U(x''),\ x' > x'' \Leftrightarrow U(x') > U(x'') \quad .$$

Consider the procedure

$$x^{s+1} = \pi_x(x^s + \rho_s \xi^0(s)) \quad ,$$

$$\xi^0(s) = \begin{cases} h_s, & \text{if } x^s + \Delta_s h^s \gtrsim x^s \quad , \\ -h_s, & \text{if } x^s + \Delta_s h^s < x^s \quad , \end{cases}$$

where $h^0, h^1, \ldots, h^s, \ldots$ are the results of independent samples of the random vector $h = (h_1, \ldots, h_n)$ uniformly distributed over the unit sphere. It can be shown [7] that

$$E\{\xi^0(s) | x^s\} = \alpha \frac{U_x(x^s)}{\|U_x(x^s)\|} \quad ,$$

for differentiable $U(x)$, where α is positive number. Therefore, the convergence of this procedure follows from the general conditions of the procedure given by (19)(with small corrections). A series of similar procedures for general constrained problems was investigated in [68].

7.3 The Global Nondifferentiable Optimization Problem Arising from Linkage Systems

The presence of random disturbances in gradient type procedures:

$$x^{s+1} = x^s - \rho_s [f_x^0(x^s) + \varepsilon(s)] \quad ,$$

$$x^{s+1} = x^s - \rho_s [f_x^0(x^s + \varepsilon(s)] \quad ,$$

$$E\varepsilon(s) = 0, \ \rho_s \geqslant 0, \ \sum_{s=0}^{\infty} \rho_s = \infty, \ \sum_{s=0}^{\infty} \rho_s^2 < \infty \quad ,$$

(for ordinary problems of minimizing $f^0(x)$ without noise w) permits us to bypass stationary points, where $f_x^0(x^s) = 0$. Notice that

$$E\{f_x^0(x^s) + \varepsilon(s) | x^s\} = f_x^0(x^s) \quad ,$$

$$E\{f_x^0(x^s) + \varepsilon(s) | x^s\} = F_x^0(x^s, s) \quad ,$$

where $F_x^0(x^s, s)$ is the gradient of the function (29). An optimization problem becomes especially difficult when the objective function $f^0(x)$ possesses many local optima and has no continuous derivatives. A typical example of such a problem may be the following problem of linkage of systems (see [61]). The problem is defined as the opposite to decomposition. If in the decomposition problem one tries to subdivide the original model of the system into a number of small models of the subsystems, then in a linkage problem one must try to obtain a model of the whole system by concatenation of the models for subsystems.

Let us suppose that each model of a subsystem $k = \overline{1,N}$ (submodel) can be described by the minimization problem

$$\langle a(k), x(k) \rangle$$

subject to

$$A(k)x(k) \geqslant b(k) \quad ,$$

$$B(k)x(k) \geqslant y(k) \quad ,$$

$$x(k) \geqslant 0 \quad .$$

These models have exogenous variables $y(k), k = \overline{1,N}$ which describes interactions between subsystems. One can consider these variables as endogenous or as decision variables when these submodels are linked in a model for the whole system. Denote by $x(k,y)$ the solution of the k-th problem for given $y(k)$, $\phi_k(y) = \langle a(k), x(k,y) \rangle$. Then the problem of linkage is the problem of finding such $y = (y(1),\ldots,Y(N))$, which minimizes the objective function of the whole system

$$g(y) = \psi(\phi_1(y),\ldots,\phi_N(y))$$

for a feasible set of linking variables y. For instance

$$g(y) = \sum_{k=1}^{N} c_k \phi_k(y)$$

The functions $\phi_k(y)$ are nondifferentialbe piecewise linear convex functions and $g(y)$ would be also convex, if $\psi(v_1,..,v_N)$ is a convex differentiable function and $\psi'_{v_k} \geqslant 0$. If the ψ'_{v_k} are also allowed to be negative differentiable function with many local minima.

Random directions of search may be a simple method to construct nondifferentiable optimization descent procedures which are easy to use with a computer. One of them is as follows: from the point x^s, the direction of the descent is chosen at random and a motion is made in this direction with a certain step size.

However, such a descent method of pure random search may take a long time in finding the direction descent. For instance, the probability of a randomly chosen direction at $x = 0$, which would lead into the set $\{x=(x_1,..,x_n)|x_i<0, i=\overline{1,n}\}$ equal $1/2^n$. Such directions are descending for function $\max\limits_{1\le i\le n} x_i$ at $x = 0$.

To avoid those situations, two classes of deterministic methods were proposed based on the idea of a subgradient: descent methods (see the works Wolfe and Lemarechal in [39]) and nondescent methods [37], [38], and [40].

The first class of the methods yields a monotonic decrease of the objective function but has a complex logic and is sensitive to local minima.

The second class which generalizes gradient type procedures

$$x^{s+1} = x^s - \rho_s \hat{f}^0_x(x^s) \quad ,$$

$$\rho_s \downarrow 0, \quad \sum_{s=0}^{\infty} \rho_s = \infty \quad ,$$

does not result in a monotonic decrease of the objective function, but they are easy to use on the computer and they are less sensitive to local minima. Consideration of random disturbances in procedures of the type

$$x^{s+1} = x^s - \rho_s [\hat{f}^0_x(x^s) + \varepsilon(s)] \quad ,$$

or in a more effective way as in (30), make them still less senstive and permits to us to bypass even points of discontinuity, as mentioned above in section 6.

7.4 Systems Identification and Parameter Estimation

Determination of mathematical models of systems require determining the nominal parameter of systems. Problems of estimation of unknown system parameters and system identification

can often be formulated as stochastic programming problems. The SQG methods in such cases allow us to construct iterative procedures which can be performed on line and can use a priori information concerning the structure of the system for improving estimates. Let us consider some examples.

Many problems of statistical estimation deal with the problem of estimating the true value x^* of unknown parameters $x = (x_1,\ldots,x_n)$ from the elements of a sample $h^0,h^1,..,h^S,..$ assumed to have been drawn from a distribution function $H(y,x^*) = P\{h \leqslant y\}$. There may be different formulations of optimization problems (see [5], [28]) concerning such problems of estimation (it depends on our knowledge about $H(y,x^*)$).

There is no information about $H(y,x^*)$ except the sample $h^0,h^1,\ldots,h^S,\ldots$and $x^* = Eh$. Therefore the problem is to estimate x^*, where

$$h^s = x^* + \varepsilon(s),\ E\varepsilon(s) = 0,\ s = 0,1,\ldots.$$

The sought-for parameter x^* minimizes the function

$$F^0(x) = E\|x - h\|^2 \ ,$$

because $x^* = Eh$ satisfies the optimality conditions

$$F^0_{x_i} = 2(x_i - Eh_i) = 0,\ i = \overline{1,n} \ .$$

If a priori knowledge about the unknown x is introduced as $x \in X$, then from (19) we could obtain the following iterative procedure for finding x^* (with $\xi^0(s) = 2(x^s - h^s)$):

$$x^{s+1} = \pi_x(x^s - \rho_s(x^s - h^s)),\ s = 0,1,\ldots \qquad (34)$$

If $X = R^n$, $\rho_s = \frac{1}{2(s+1)}$, then

$$x^{s+1} = x^s - \frac{1}{s+1}(x^s - h^s) = \frac{1}{s+1}\sum_{k=1}^{s+1} h^k \qquad (35)$$

The estimation (35) is the sample mean. The advantages of the estimation (34) when compared to (35) are

a) possibilities of choosing ρ_s as a function of $(x^0,..x^s)$ in order to decrease the value of the objective function;

b) if $X \neq R^n$, then from (34) it follows that $x^s \in X$ for all $s = 0,1...$, whereas in (35) only $\lim x^s \in X$. Therefore the estimations from (34) must be better for small samples.

Problems of estimation of the moments

$$E\eta^e, E|\eta|^e, E(\eta - E\eta)^e \text{ , where } \eta^e = (\eta_1^e \ldots, \eta_n^e) \text{ ,}$$

$$|\eta|^e = (|\eta_1|^e,..,|\eta_n|^e), (\eta - E\eta)^e = ((\eta_1 - E\eta_1)^e,\ldots,(\eta_n - E\eta_n)^e)$$

may also be formulated as minimization problems

$$F_1^0(x) = E\|x - h^e\|^2, F_2^0(x) = E\|x - |h|^e\|^2 \text{ ,}$$

$$F_3^0(x) = E\|x - (\eta - E\eta)^e\|^2 \text{ .}$$

The stochastic gradients of these functions are:

$$\xi_1^0(s) = 2(x^s - (h^{s+1})e), \xi_2^0(s) = 2(x^s - |h^{s+1}|^e) \text{ ,}$$

$$\xi_3^0(s) = 2(x^s - \prod_{k=1}^{e}(h^{s+1} - h^{s+1+k})) \text{ .}$$

Suppose now that we have the information

$$x^* = Eh = t(z)\big|_{z = z^*} \text{ ,}$$

where $t(z)$ is a given function and z^* is an unknown vector. Then z^* minimizes the function

$$E\|t(z) - h\|^2 \quad .$$

If we have information about the density $p(y,x^*)$ of $H(y,x^*)$ with a measure $\mu(dy)$, then it could be shown that x^* maximizes the function

$$E\ln p(x,h) = \int \ln p(x,y)p(y,x^*)\mu(dy) \quad .$$

These problems are re-formulations of well-known principles for the least square i.e., minimization of the function

$$\frac{1}{N} \sum_{k=1}^{N} \|t(z) - h^k\|^2$$

and maximum likelihood, i.e., maximization of the function

$$\frac{1}{N} \sum_{k=1}^{N} \ln p(x,h^k) \quad .$$

It gives us a good opportunity to apply SQG methods.

The above mentioned problems are the problems of pure estimation. Very often the main reasons for estimation and identification are control or optimization. In such cases, it seems to be unnecessary to first determine a model (unknown parameters) and then design an optimization strategy based on this model. Why not use a procedure that directly solves an optimization problem and simultaneously extracts from the answers the information needed for estimation? Such kinds of procedures based on general ideas of nonstationary optimization, were considered in [20]. Let the model of the system be formulated as the problem of minimizing

$$F^0(x,z) \quad ,$$

where x is a control variables, $x \in X \subseteq R^n$, $z \in Z$ is a vector of unknown parameters. For a sequence of given approximation $x^0, x^1, \ldots$ there are available the observations $h^0, h^1, \ldots, h^s, \ldots$ of random vector h

$$Eh^s = g(x^s, z^*) ,$$

where $g(x,z)$ is known, z^* is the true value of z. If $g(x,z^*) \equiv z^*$, then we could consider the sequence of estimates z^s, such that $z^s \to z^*$ with probability 1 and the problem of simultaneous estimation z^* and optimization of the $F^0(x,z^*)$ becomes the limit extremal problem with time dependent function $F^0(x,z^s)$. When the values $\hat{F}^0_x(x,z^s)$ are known, then the procedure, mentioned above in section 6

$$x^{s+1} = \pi_x(x^s - \rho_s \hat{F}^0_x(x^s, z^s))$$

can be used for minimizing $F^0(x,z^*)$. In the general case we shall consider the procedure

$$x^{s+1} = \pi_x(x^s - \rho_s \xi^0(s))$$

$$E\{\xi^0(s) \mid x^0, z^0, .., x^s, z^s\} = \hat{F}^0_x(x^s, z^s) + a^0(s) \quad ,$$

simultaneously with the procedure of estimation

$$z^{s+1} = \pi_V(z^s - \delta_s \zeta(s)) \quad ,$$

$$E\{\zeta(s) \mid x^0, z^0, \ldots, x^s z^s\} = \phi_z(x^s, z^s) + d(s) \quad ,$$

$$\phi(x,z) = E\|h - g(x,z)\|^2 \quad .$$

Theorem 5. Let X,V be convex compact, $F^0(x,z)$ is a convex continuous function with respect to x, for all $z \in Z$; the function

$$[g(x,z) - g(x,z^*)]^T [g(x,z) - g(x,z^*)]$$

is convex with respect to z, for all $x \in X$, and there is the unique solution of equation

$$g(x^s,z) = g(x^s,z^*)$$

and with probability 1:

$$\rho_s \geq 0, \quad \delta_s \geq 0, \quad \sum_{s=0}^{\infty} \rho_s = \infty, \quad \sum_{s=0}^{\infty} \delta_s = \infty \quad ,$$

$$\sum_{s=0}^{\infty} (\rho_s \|a^0(s)\| + \delta_s \|d^s\| + \rho_s^2 + \delta_s^2) < \infty \quad ,$$

$$\|\xi^0(s)\| + \|\zeta(s)\| + \|a^0(s)\| + \|d(s)\| < \text{const} \quad .$$

Then

$$\lim F(x^s,z^s) \to \min \{F^0(x,z^*) \mid x \in X\} \quad .$$

The article [20] contains numerical results and similar theorems without assumptions of convexity, existence of a unique solution and stationarity of the models.

The dynamic aspects of systems identification were studied in [5], [29], and [34]. The problem was formulated for instance, as minimizing

$$F^0(x) = E \max_k \| z(k) - h(k) \|^2$$

subject to

$$z(k+1) = g(z(k),x,w,k) \quad ,$$

$$z(0) = z^0, \; k = 0,1,\ldots,N-1 \quad ,$$

where x are unknown parameters, $x \in X \subseteq R^n$, h(k) are observations of the trajectory. SQG methods for such and more general problems with differentiable and nondifferentiable criterias and constraints were studied in [5] and [34].

8. COMPUTATIONAL AMPLEMENTATION: AN EXAMPLE

The SQG methods have been applied to several problems (deterministic and stochastic), containing a great number of variables. One of the advantages of these methods is that a priori knowledge of the statistics is not necessary (this opens up the possibilities of on-line optimization), numerical stability (these algorithms work in the presence of noise). The behavior of SQG methods is unusual compared with deterministic methods. It converges to one of the solutions but this solution may be different for different realizations of the stochastic method. For a unique solution there may be different ways of approaching a neighborhood of this optimal solution. The process of optimization could hardly be done in one run. It was frequently useful to interfere manually, by choosing different starting values, and to change the parameters of the algorithm, when it is difficult to know if a local minimum had been achieved or not. Efficient optimization processes require interactive program packages to cover the whole range from data modification to simulation. The reason why interactive programs are so efficient for optimization is that optimization is always an iterative procedure.

The success of the application of SQG methods depends on the rules for choosing the parameters of the algorithms (random directions, step size). To demonstrate this, consider the results of the solution of the following stochastic facility location problem (see [59], [64]).

A set of places of residence for the users (demand points) is given and a set of possible locations for the facilities. The users of demand point $i = \overline{1,m}$ are choosing the facility $j = \overline{1,n}$ with probability p_{ij}. Let ε_{ij} be the random flow of users from demand point i to facility j

$$\sum_{j=1}^{n} \varepsilon_{ij} = a_i, \quad i = \overline{1,m}$$

where a_i is the random demand at point i. Determine the size x_j of the facility $i = \overline{1,n}$ in order to minimize the expenditures

$$F^0(x_1,\ldots,x_n) = \sum_{j=1}^{n} E \max \{\alpha_j(x_j - \sum_{i=1}^{m} \varepsilon_{ij}), \beta_j(\sum_{i=1}^{m} \varepsilon_{ij} - x_j)\}$$

subject to

$$0 \leq x_i \leq r_j \quad , \quad j = \overline{1,n}$$

The algorithm (19) with $\xi^0(s)$ as (17) takes the form

$$x_j^{s+1} = \max \{0, \min\{r_j, x_j^s - \rho_s \xi_j^0(s)\}\}$$

$$\xi_j^0(s) = \begin{cases} \alpha_j, & \text{if } x_j^s \geq \sum_{i=1}^{m} \varepsilon_{ij}^s \\ -\beta_j, & \text{if } x_j^s < \sum_{i=1}^{m} \varepsilon_{ij}^s \end{cases} .$$

Here ε_{ij}^s is an observation of the flow variables ε_{ij}

$$\sum_{j=1}^{n} \varepsilon_{ij}^s = a_i^s \quad , \quad i = \overline{1,m} \quad ,$$

where a_i^s are the observations of the demand.

From Theorem 2 it follows that ρ_s might be chosen adaptively as a function of the realization $(x^0,x^1,\ldots,x^s)$ or independently as $\rho_s = \frac{1}{s}$. The choice $\rho_s = \frac{1}{s}$ serves all realizations of the stochastic procedure and cannot be a good one. The nice ways of choosing ρ_s are the adaptive rules, which depend on each realization separately.

The step size adaptation was inserted into this algorithm by starting an optimization process with $\rho_s = C_0$ (or C_o/S) where C_0 is a relatively big number. By trial-and-error mechanism we can find C_0 with which the irregular behavior of the quantities

$$f^0(x^s,w^s) = \sum_{j=1}^{n} \max \left\{ \alpha_j (x_j^s - \sum_{i=1}^{m} \varepsilon_{ij}^s), \beta_j (\sum_{i=1}^{n} \varepsilon_{ij}^s - x_j) \right\}$$

would show a rather raped tendency of decreasing. This is illustrated schematically in Figure 1 for the test problem of scholl location with data for Turin city (see [64]), n = 23, $p_{ij} = (e^{-c_{ij}}/\sum_j e^{-c_{ij}})$ and where c_{ij} is the distance between demand point i and potential location j.

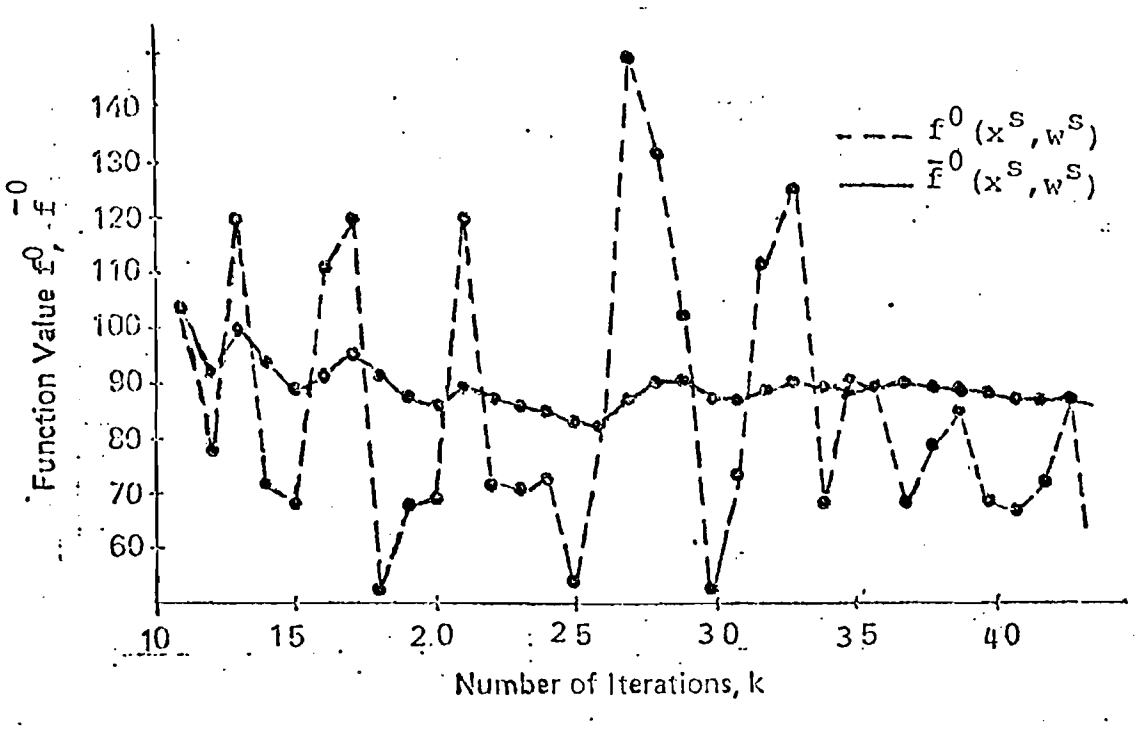

Figure 1. The behavior of the sequences $f^0(x^s,w^s)$ and $\tilde{f}^0(x^s,w^s)$ as a function of the iteration number.

Figure 1 also shows the more regular behavior of the quantities

$$\bar{f}^0(x^s,w^s) = \frac{1}{s}\sum_{k=1}^{s} f^0(x^k,w^k) \quad .$$

After a number of iterations the function $\bar{f}^0(x^s,w^s)$ will achieve a certain level and then will remain almost on the same level. The nonimprovement in the behavior of $\bar{f}^0(x^s,w^s)$ causes switching to a new step regulation $\rho_s = c_1$ (or c_1/s) etc.

For $\alpha_j = \beta_j$ and deterministic demands a_i the results of the computations were generally in a good agreement with well-known solution of such a problem, based on the entropy approach (see [64]).

$$x_j^* \approx \sum_{i=1}^{m} a_i \left(e^{-c_{ij}} / \sum_j e^{-c_{ij}} \right)$$

In many cases the convergence is improved if during some iterations the directions (see [25])

$$\frac{1}{k_s} \sum_{k=s-k_s}^{s} \xi^0(k)$$

are used instead of $\xi^0(s)$.

Formal investigations of the asymptotic rate of convergence of SQG-type procedures were attempted by Poljak (see [42]). A systematic study of asymptotic behavior (asymptotic rate, limiting distributions, etc.) and more general procedures were undertaken in [24]. Note that for the above mentioned step-size selection it is important to have asymptotic behavior of the stochastic procedures with permanent step multiplier (see [23]).

REFERENCES

[1] Ermoliev, Yu.M., and Z.V. Nekrylova. 1967. The Method Stochastic Subgradients and Its Applications. Notes, Seminar on the Theory of Optimal Solution. Academy of Sciences of the U.S.S.R., Kiev.

[2] Ermoliev, Yu.M., and N.Z. Shor. 1968. Method of random walk for two-stage problem and its generalization. Kibernetika, 1.

[3] Ermoliev, Yu.M. 1969. On the stochastic quasi-gradient method and stochastic quasi-Feyer sequences. Kibernetika, 2.

[4] Ermoliev, Yu.M. 1971. General problem of stochastic programming. Kibernetika, 3.

[5] Ermoliev, Yu.M. 1976. Stochastic Programming Methods. Moscow: Nauka.

[6] Gupal, A.M., and L.G. Bajenov. 1972. Stochastic linearization. Kibernetika, 1.

[7] Ermoliev, Yu.M., and A.I. Jastremskiy. 1979. Stochastic Models and Methods in Economic Planning. Moscow: Nauka.

[8] Gupal, A.M. 1974. On the stochastic programming problem with constraints. Kibernetika, 6.

[9] Gupal, A.M. 1977. Method of almost-differentiable function minimization. Kibernetika, 1.

[10] Gupal, A.M. 1979. Stochastic Methods of Nonsmooth Optimization. Kiev: Naukova Dumka.

[11] Nurminski,E.A. 1973. Convergence conditions of algorithms of stochastic programming. Kibernetika, 3.

[12] Nurminski,E.A. 1973. Quasigradient method for solving problems on nonlinear programming. Kibernetika, 1.

[13] Nurminski,E.A. 1979. Numerical Methods for solving Deterministic and Stochastic Minimax Problems. Kiev: Naukova Dumka.

[14] Ermoliev, Yu.M., and E.A. Nurminski. 1973. Limit extremum problems. Kibernetika., 1.

[15] Gupal, A.M. 1974. Optimization method for problems with time-varying functions. Kibernetika, 2.

[16] Nurminski, E.A. 1977. The problem of nonstationary optimization. Kibernetika, 2.

[17] Vertchenko, P.I. 1977. Limit Extremum Problems of Stochastic Optimization. Abstract of dissertation, Press of the Institute of Cybernetics, Kiev.

[18] Gaivoronskiy, A.A. 1977. Methods of Stochastic Nonstationary Optimization, Collection Operations Research and Systems Reliability. Press of the Institute of Cybernetics, Kiev.

[19] Gaivoronskiy, A.A. 1978. Nonstationary stochastic programming problems. Kibernetika, 4.

[20] Gaivoronskiy, A.A., and Yu.M. Ermoliev. 1979. Stochastic optimization and simultaneous parameter estimation. Izvestia Academii Nauk SSSR, Technischeskaj Kibernetika, 4.

[21] Nurminski, E.A., and P.I. Verchenko. 1977. On a convergence of saddle-point algorithms, Kibernetika 3.

[22] Golodnikov, A.N. 1979. Finding of Optimal Distribution Function in Stochastic Programming Problems. Abstract of dissertation, Institute of Cybernetics press, Kiev.

[23] Ermoliev, Yu.M., and Yu.M. Kaniovskiy. 1979. Asymptotic behavior of stochastic programming methods with permanent step-size multiplier USSR, Computational Mathematics and Mathematical Physics, 2.

[24] Kaniovskiy, Yu.M., P.S. Knopov, and Z.V. Nekrylova. 1980. Limiting Theorems of Stochastic Programming Processes. Kiev: Naukova Dumka.

[25] Bajenov, L.G., and A.M. Gupal. 1972. Stochastic analog of conjugate gradients method. Kibernetika, 1.

[26] Gupal, A.M. 1978. Stochastic method of feasible directions of nondifferentiable optimization. Kibernetika, 2.

[27] Ermoliev, Yu.M. 1975. Stochastic models and methods of optimization. Kibernetika, 4.

[28] Ermoliev, Yu.M., and E.A. Nurminski. 1973. Extremum Problems in Statistics and Numerical Methods of Stochastic Programming. Collection: Some Problems of Systems Control and Modeling. Press of the Institute of Mathematics, Ukrainian Academy of Sciences, Kiev.

[29] Ermoliev, Yu.M. 1972. The stochastic problem of optimal control. Kibernetika, 1.

[30] Ermoliev, Yu.M., and A.M. Gupal. 1978. The linearization method in nondifferentiable optimization. Kibernetika, 1.

[31] Ermoliev, Yu.M. 1976. The Stochastic Quasigradient Methods and their Application to the Stochastic Programming Problems with Non-Smooth Functions. Proceedings of the IX Mathematical Programming Symposium. Amsterdam: North Holland.

[32] Gupal, A.M., and V.P. Norkin. 1977. Method of discontinuous optimization. Kibernetika, 2.

[33] Ermoliev, Yu.M., and E.A. Nurminski. 1980. Stochastic quasigradient algorithms for minimax problems. Edited by M. Dempster. Proceedings of the International Conference on Stochastic Programming. London: Academic Press.

[34] Ermoliev, Yu.M., V.P. Gulenko, and T.I. Tsarenko. 1978. Finite-Difference Method in Optimal Control. Kiev: Naukova Dumka.

[35] Girko, V.L. 1976. Stochastic Matrices. Kiev: Naukova Duma.

[36] Shor, N.Z. 1962. Application of the Gradient Method for the Solution of Network Transportation Problems. Notes, Scientific seminar on theory and applications of cybernetics and operations research. Kiev: Academy of Sciences USSR.

[37] Ermoliev, Yu.M. 1966. Methods of solution of nonlinear extremal problems. Kibernetika, 4.

[38] Poljak, B.T. 1967. A general method for solving extremal problems. Soviet Mathematic Doklady, 8.

[39] Wolfe, P., and M.L. Balinski (eds.) 1975. Nondifferentiable Optimization. Mathematical Programming Study 3, North-Holland Publishing Co.

[40] Shor, N.Z. 1979. The Methods of Nondifferentiable Optimization and their Applications. Kiev: Naukova Dumka.

[41] Bajenov, L.G. 1972. Convergence conditions of almost-differentiable function minimization. Kibernetika, 4.

[42] Poljak, B.T. 1976. Convergence and rate of convergence of iterative stochastic algorithms. Automatic and Remote Control, 12.

[43] Katkovnik, V. Ya. 1976. Linear Estimation and Stochastic Optimization Problems. Moscow: Nauka.

[44] Rastrigin, L.A. 1974. Extremal Control Systems. Moscow: Nauka.

[45] Fjedorov, V.V. 1979. Numerical Methods of Maxmin Problems. Moscow: Nauka.

[46] Poljak, B.T. 1978. Nonlinear programming methods in the presence of noise. Mathematical Programming, 14.

[47] Wasan, M.T. 1969. Stochastic Approximations. Cambridge Transactions in Math. and Math. Phys. 58. Cambridge: Cambridge University Press.

[48] Fabian, V. 1965. Stochastic Approximation of Constrained Minima. In: Transactions of the 4th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, 1965, (Prague, 1967).

[49] Kushner, H.J. 1974. Stochastic approximation algorithms for constrained optimization problems. Annals of Statistics 2(4).

[50] Kushner, H.J., and T. Gavin. 1974. Stochastic approximation type methods for constrained systems: algorithms and numerical results. IEEE Transactions Automatic Control, 19.

[51] Kushner, H.J., and E. Sanvincente. 1975. Stochastic approximation of constrained systems with system and constraint noise. Automatic, 11.

[52] Marti, K. 1976. On approximate solutions of stochastic dominance and stochastic penalty methods. Proceedings of the IX. Mathematical Programming Symposium. Amsterdam: North-Holland.

[53] Hiriart-Urruty, J.-B. 1977. Contributions à la Programmation Mathematique: Cas Deterministe et Stochastique. Thèse, D.Sc. Mathematiques, Université de Clermont-Ferrand 11.

[54] Arrow, K.J., L. Hurwicz, and H. Uzawa, eds. 1958. Studies in Linear and Non-linear Programming. Stanford, Calif.: Stanford University Press.

[55] Dupaĉ, V. 1965. A dynamic stochastic approximation method. Annals of Mathematical Statistics, 6.

[56] Dupaĉ, V. 1976. The continuous dynamic Robbins-Monroe procedure. Kybernetika 12, N6.

[57] Uosaki, K. 1974. Some generalizations of dynamic stochastic approximation procedures. The Annals of Statistics, 2, N5.

[58] Fujita, S., and T. Fukao. 1972. Convergence conditions of dynamic stochastic approximation method for non-linear stochastic discrete-time dynamic systems IEEE Transactions on Automatic Control, AC-17, N5.

[59] Ermoliev, Yu.M., and G. Leonardi. 1980. Some Proposals for Stochastic Facility Location Models, WP-80- Laxenburg, Austria: International Institute for Applied Systems Analysis.

[60] Rockafellar, R.T., and R. Wets. 1976. Stochastic convex programming: singular multipliers and extended duality. Pacific J. Math. 62.

[61] Ermoliev, Yu.M. 1980. Some Problems of Linkage Systems. WP-80-102. Laxenburg, Austria: International Institute for Applied Systems Analysis.

[62] Roberts, P.D. 1977. Multilevel approaches to the combined problem of systems optimization and parameter identification. International Journal of Systems Science, 8(3).

[63] Tsypkin, Ya.Z. 1971. Adaptation and Learning in Automatic Systems. New York: Academic Press.

[64] Ermoliev, Yu.M., G. Leonardi, and J. Wira. 1980. Stochastic Quasi-Gradient Methods Applied to a Facility Location Problem. WP-80- , Laxenburg, Austria: International Institute for Applied Systems Analysis.

[65] Nemizovsky, A.S., and D.B. Judin. 1979. Complexity of Problems and Efficiency of Optimization Methods. Moscow: Nauka.

[66] Kushner, H.J., and D.S. Clark. 1978. Stochastic approximation methods for constrained and unconstrained Systems, Applied Mathematical Sciences, 26.

[67] Luce, R.D. and H. Raiffa. 1957. Games and Decisions. New York: John Wiley & Sons.

[68] Michalevitch, M.V. 1979. On optimization nondifferentiable utility functions. Collection: Methods nondifferentiable and stochastic optimization. Press of the Institute of Cybernetics Ukrainian Academy of Sciences, Kiev.

[69] Zangwill, W.I. 1969. Convergence conditions for nonlinear programming. Mang.Sci. vol.16, N1.

[70] Motzhin, T.S. and J.J. Schanberg. 1954. The relaxation method for linear inequalities. Canad.J.Math., 6,3.