METHODS OF NONDIFFERENTIABLE AND STOCHASTIC OPTIMIZATION AND THEIR APPLICATIONS

Yu.M. Ermoliev

December 1978 WP-78-62

Working Papers are internal publications intended for circulation within the Institute only. Opinions or views contained herein are solely those of the author(s).

2361
Laxenburg
Austria

International Institute for Applied Systems Analysis

INTRODUCTION

Optimization methods are of a great practical importance in systems analysis. They allow us to find the best behavior of a system, determine the optimal structure and compute the optimal parameters of the control system etc. The development of nondifferentiable optimization, differentiable and nondifferentiable stochastic optimization allows us to state and effectively solve new complex optimization problems which were impossible to solve by classical optimization methods, for instance optimization problems with numbers of variables in the order of 100^{100}.

The term nondifferentiable optimization (NDO) was introduced by Balinski and Wolfe [1] for extremum problems with an objective function and constraints that are continuous but have no continuous derivatives. Now this term is used also for problems with discontinuous functions though it might be better to use for them the terms nonsmooth optimization (NSO) or, in particular, discontinuous optimization (DCO).

The term stochastic optimization (STO) is used for stochastic extremum problems or for stochastic methods that solve deterministic or stochastic extremum problems.

Nondifferentiable and stochastic optimization are natural developments of classic optimization methods. The interest in nondifferentiable optimization and stochastic optimization is based on two reasons: first, as has been mentioned above a wide range of new applied problems cannot be solved by the classic methods; secondly, the possibility of reducing known difficult problems to nondifferentiable or stochastic optimization problems that permit obtaining their solutions.

For example, from the conventional viewpoint, there is no principal difference between functions with continuous gradients which change rapidly and functions with discontinuous gradients.

Some important classes of nondifferentiable and stochastic optimization problems are well-known and have been investigated long ago: problems of Chebyshev approximations, game theory and mathematical statistics. However, each of these classes was investigated by its own "homemade" methods. General approaches (extremum conditions, numerical methods) were developed at the beginnning of the 1960's. The main purpose of this article is to review briefly some important applications of nondifferentiable and stochastic optimization and to characterize principal directions of research. Clearly, the interests of the author have influenced the content of this article.

1. APPLICATIONS OF NDO & STO

Let us consider some applied problems which require nondifferentiable optimization and stochastic optimization methods.

Optimization of Large-Scale Systems

Many applied problems lead to complex extremum problems with a great number of variables and constraints. For example, there are linear programming problems with a number of variables or constraints in the order of 100^{100}. Formally such problems have one of the following forms:

$$\sum_{j=1}^{n} a_{0j} x_j = \min \qquad (1)$$

$$\sum_{j=1}^{n} a_{ij}(\theta) x_j \geq b_i(\theta) , \qquad \theta \in \Theta, \quad i = \overline{1,m} , \qquad (2)$$

$$x_j \geq 0 , \qquad j = \overline{1,n} \qquad (3)$$

or

$$\sum_{\theta \in \Theta} d_0(\theta) x(\theta) = \min \qquad (4)$$

$$\sum_{\theta \in \Theta} d_i(\theta) x(\theta) \geq \beta_i , \qquad i = \overline{1,m} \qquad (5)$$

$$x(\theta) \geq 0 , \quad \theta \in \Theta . \qquad (6)$$

Here Θ is a given discrete set, for example

$$a_{ij}(\theta) = \sum_{\ell=1}^{r} d_{ij}^{\ell}\theta_{\ell} + \alpha_{ij} \quad ,$$

$$b_{i}(\theta) = \sum_{\ell=1}^{r} \;{}_{i}\,\theta_{\ell} + \beta_{i} \quad ,$$

$$\Theta = \{\theta = (\theta_1,\ldots,\theta_r) : \sum_{\ell=1}^{r} \gamma_{\ell}\theta_{\ell} \leq \gamma \, , \theta_{\ell} = \pm 1 \, , \ell = \overline{1,r}\} \quad .$$

Clearly that for this case the total number of constraints is equal to $2^{r} \cdot m$.

On the other hand these constraints have a form which does not impose heavy demands on the computer core and one can try to find their solution with the known methods of linear programming [2]. However, the number of vertices of the feasible polyhedral set for such problems is so large that the application of the conventional Simplex method or its variants yield very small steps at each iteration and consequently very slow convergence. Moreover the known finite methods are not robust computational errors. The reduction of these problems to problems of nondifferentiable or stochastic optimization made it possible to develop easily implemented iterative decomposition schemes of the gradient type. These approaches do not use the basic solution of the linear programming problem which enables to start the computational process from any point and leads to computational stability. Furthermore, these methods converge faster in practice.

Consider the problem (1) - (3). It can be reduced to the nondifferentiable optimization problem

$$f^{0}(x) = \sum_{j=1}^{n} a_{0j}\, x_{j} = \min \qquad (7)$$

$$f^{i}(x) = \max_{\theta\in\Theta} \left(\sum_{j=1}^{n} a_{ij}(\theta)\, x_{j} - b_{j}(\Theta) \right) \geq 0, \quad i = \overline{1,m} \qquad (8)$$

$$x_{j} \geq 0 \quad , \qquad j = \overline{1,n} \quad , \qquad (9)$$

which has only m constraints.

We consider now some schemes of decomposition which are described in [3]. Let the linear programming problem have the form

$$(c,x) + (d,y) = \min$$

$$Ax + Dy \geq b$$

$$x \geq 0 \quad , \quad y \geq 0 \quad .$$

We assume that for fixed X it is easy to find its solution y(x) with respect to y. For example the matrix D may have a block diagonal structure, with x being the connecting variables. The main difficulty here is to find the value x* of the optimal solution (x*,y(x*)). The search for x* is equivalent to the minimization of the nonsmooth function

$$f(x) = (c,x) + \min_{\substack{Dy\geq b-Ax \\ y\geq 0}} (d,y) = (c,x) + (d,y(x)) \quad . \qquad (10)$$

Another approach is to consider the dual problem:

$$(u,b) = \max$$

$$uD \leq d \quad ,$$

$$uA \leq 0 \quad ,$$

$$u \geq 0 \quad .$$

Let us examine the Lagrangean function

$$(u,b) + (c - uA,x) = (c,x) + (u,b - Ax)$$

where

$$uD \leq d \quad , \quad u \geq 0 \quad , \quad x \geq 0 \quad .$$

In this case the search of x* is equivalent to the minimization of the nonsmooth function (the well-known Dantzig-Wolfe-decomposition is based on this principle)

$$f(x) \;=\; (c,x) + \max_{\substack{uD\le d,\\ u\ge 0}} (u,b-Ax) \quad \text{for } x \ge 0 \quad . \tag{11}$$

A subproblem of minimization with respect to variables u, subject to

$$uD \le d \quad , \quad u \ge 0$$

is solved easily because the matrix D has a special structure by assumption.

A parametric decomposition method [4] reduces linear programming problems which do not have block diagonal structure to nondifferentiable optimization problems by introducing additional parameters. In this case there is the possibility to split the linear programming problem into arbitrary parts, in particular to single out subproblems corresponding to blocks of nonzero elements in the constraint matrix. An analogous idea was also used in [5,6].

Let us analyse the general idea of the method using the concrete example

$$y_3 \;=\; \min \tag{12}$$

$$\begin{aligned} a_{11}y_1 + a_{12}y_2 + \boxed{a_{13}y_3} &\le b_1 \quad , \\ a_{21}y_1 + \boxed{a_{22}y_2} + \boxed{a_{23}y_3} &\le b_2 \quad , \end{aligned} \tag{13}$$

where

$$b_1 \ge 0 \quad , \quad b_2 \ge 0 \quad , \quad y_1 \ge 0 \quad , \quad i = 1,2,3 \quad .$$

Let it be necessary to cut this problem, for example, into three parts as it is shown in constraints (13).

Consider the following subproblem: for the given variable $x = (x_{11},x_{12},x_{21},x_{22},x_{23}) \ge 0$ find $y_1 \ge 0$, $y_2 \ge 0$, $y_3 \ge 0$ for which

$$\begin{aligned} & y_3 = \min \\ & a_{11}y_1 + a_{22}y_2 \le x_{11} \quad , \quad a_{13}y_3 \le x_{12} \\ & \qquad a_{21}y_1 \le x_{21} \quad , \quad a_{23}y_3 \le x_{23} \\ & \qquad a_{22}y_2 \le x_{22} \end{aligned} \tag{14}$$

This problem comes to the three subproblems with the desirable structure. If the minimal value of y_3 is denoted as f(x) then it is easy to show that solving the problem (12) - (13) is equivalent to solving (14) for such x which minimizes the nondifferentiable function f(x) under the constraints:

$$\begin{aligned} & x_{11} + x_{12} \le b_1 \\ & x_{21} + x_{22} + x_{23} \le b_2 \end{aligned} \tag{15}$$

$$x_y \ge 0 \quad , \quad i = 1,2; \quad j = 1,2,3$$

Similar methods are conveniently applied in the linkage of submodels.

Discrete Programming, Minimax Problems, Problems of Game Theory

The use of duality theory for solving discrete programming problems [1,2] of large dimension necessitates the minimization of nondifferentiable functions of the kind

$$f(x) = \max_{y \in Y} \left(\sum_{j=1}^{n} a_j(y) x_j - b(y) \right)$$

where Y is some discrete set. This problem reduces to problems of the kind (1) - (3) (if we use methods of classical optimization):

$$x_{n+1} = \min$$

$$\sum_{j=1}^{n} a_j(y)\, x_j - b(y) \leq x_{n+1} \quad , \quad y \in Y \quad ,$$

$$x_j \geq 0 \quad , \quad j = \overline{1,n}$$

However, solution of this problem by linear programming methods is out of question and therefore NDO should be used for minimization of the associated function (16) below.

More general deterministic minimax problems are formulated in the following manner [7,8]: For a given function

$$g(x,y), \quad x \in X \subsetneq R^n \quad , \quad y \in Y \subseteq R^m$$

it is necessary to minimize

$$f(x) = \max_{y \in Y} g(x,y) = g(x,y(x)) \tag{16}$$

for $x \in X$. Independently of the smoothness of $g(x,y)$ the function $f(x)$ as a rule has no continuous derivatives. A particular class of the minimax problems arises in approximation theory, e.g. in problems of the best Chebyshev approximation of the function $r(y)$ by linear combinations of the functions $\delta_j(y)$:

$$g(x,y) = \left| r(y) - \sum_{j=1}^{n} x_j \delta_j(y) \right| \quad .$$

Similar problems arise in mathematical statistics, in game theory with zero sum games, in filtration theory, identification, approximation by splines etc.

A solution of systems of inequalities

$$d_i(x) \leq 0 \quad , \quad i = \overline{1,m}$$

for $g(x,y) = d_y(x)$, $y \in Y = \{1,2,\ldots,m\}$ can be reduced to minimization of the function (16). This idea was used in the work

[9] for computing economic equilibria through nonsmooth optimization. A solution of the general problem of nonlinear programming

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

can also be reduced to this problem, if it is assumed that

$$g(x,y) = f^0(x) + \sum_{i=1}^{m} y_i f^i(x), y \in Y = \{y : y = (y_1,\ldots,y_m), y_i \geq 0, i = \overline{1,m}\}.$$

In game theory, and in the theory multiobjective optimization, more complex problems arise in the minimization of the function

$$f(x) = g(x,y(x)) \tag{17}$$

for $x \in X$ where $y(x)$ is such that

$$h(x,y(x)) = \max_{y \in Y} h(x,y) \quad .$$

Independently of the smoothness of the functions $g(x,y)$, $h(x,y)$ the function $f(x)$ in the given case will have no continuous derivatives and will be discontinuous in general. For $h(x,y) = x \cdot y$, $g(x,y) = x + y$, $Y = [1,1]$, we obtain

$$y(x) = \begin{cases} 1, & \text{if } x > 0 \quad , \\ -1, & \text{if } x < 0 \quad . \end{cases}$$

The function $h(x,y(x)) = xy(x) = |x|$ is continuous but does not have continuous derivatives at the point $x = 0$. Function $f(x) = x + y(x)$ is discontinuous. That is why the value of such models in applications depends on the development of numerical methods for discontinuous optimization.

Optimization of Probabilistic Systems

Taking into account the influence of uncertain random factors even in the simplest extremum problems leads to complex extremum problems with nonsmooth functions. For example for deterministic ω a set of solutions of the inequality

$$\omega x \leq 1$$

where ω, x are scalars, defines a semi-axis. If ω is a random variable it is natural to consider the function

$$f(x) = P\{\omega x \leq 1\}$$

and to find x which maximizes $f(x)$. If $\omega = \pm 1$ with probability 0.5 then $f(x)$ is a discontinuous function (see Figure 1).

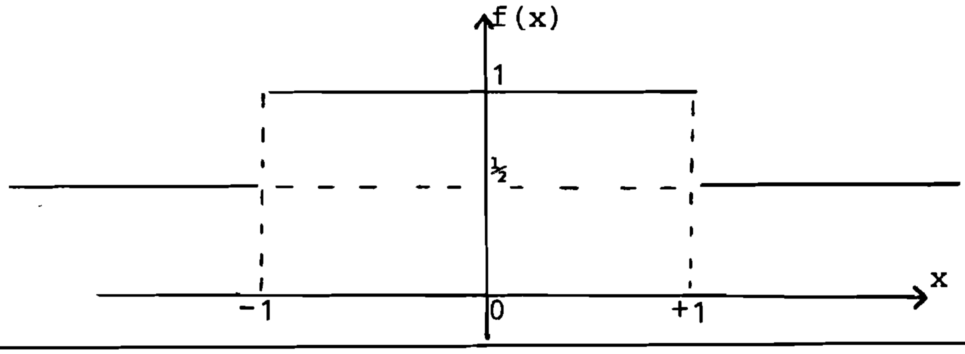

Figure 1

Since many complex systems are under the influence of the uncertain random factors, nonsmooth optimization becomes even more important.

Health Care Systems: Patients may be sick for random time intervals, the diagnosis, the results of medical treatments are partly random, epidemies are similar to random processes, accidents are random as well, and so on.

Communication and Computer Networks: Unreliability of facilities and channels, random character of the load etc.

Food and Agriculture: Harvests are strongly dependent upon weather fluctuations which are essentially random, technological progress, demands, supply of resources, forecasting investment for the development of new ideas, for new kinds of products etc.

A rather general problem of the stochastic programming can be formulated [10] as follows

$$\min \{F^0(x):F^i(x) \leq 0, i=1,m, x\in X\} \tag{18}$$

where

$$F^\nu(x) = Ef^\nu(x,\omega) = \int f^\nu(x,\omega)P(d\omega) \quad , \quad \nu = \overline{0,m} \tag{19}$$

Here $f^\nu(x,\omega)$, $\nu = \overline{0,m}$ are random functions, and ω is a random factor which we shall consider as an element of the probability space (Ω,A,P). For example conditions like

$$P\{g^i(x,\omega) \leq p\} \geq p_i \quad , \quad i = \overline{1,m}$$

become constraints of the type (18) - (19) if we assume that

$$f^i(x,\omega) = \begin{cases} p_i - 1, & \text{if } g^i(x,\omega) \leq 0 \quad , \\ p_i \quad , & \text{if } g^i(x,\omega) > 0 \quad . \end{cases}$$

The problem is more difficult than the conventional nonlinear programming problem.

It has been noted above that taking into account random parameters even in simple linear programming problems leads to nondifferentiable optimization problems. The main difficulty of the problem (18) - (19), besides the nondifferentiability, is connected with the condition (19). The examples considered below show that as a rule it is practically impossible to compute the precise values of the integrals (19) and therefore one can not calculate the precise values of the functions $F^\nu(x)$.

Usually only values of the random quantities $f^\nu(x,\omega)$ are available instead of $F^\nu(x)$. To determine whether the point x satisfies the constraints

$$F^i(x) \quad = \quad \underline{F}\, f^i(x,\omega) \ \le 0 \quad , \qquad i = \overline{1,m}$$

is then a complicated problem of verifying the statistical hypothesis that the mathematical expectation of the random quantities $f^i(x,\omega)$ is nonpositive.

Other applications

Many applied problems reduce to problems of optimal control with discontinuous trajectories (in state space), for example in impulse control, and in the control of systems with varying structure. In inventory control theory a trajectory of the system is discontinuous at the instances of deliveries, and (Fig. 2) here the value of the discontinuity can serve as control variable.

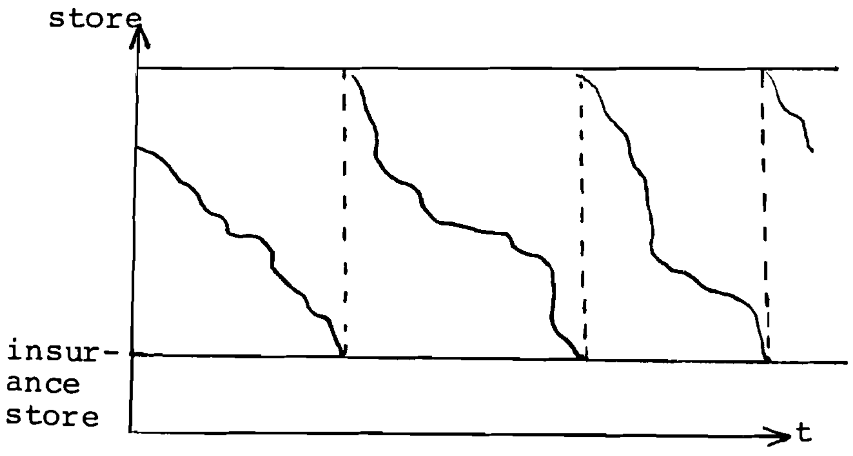

Figure 2

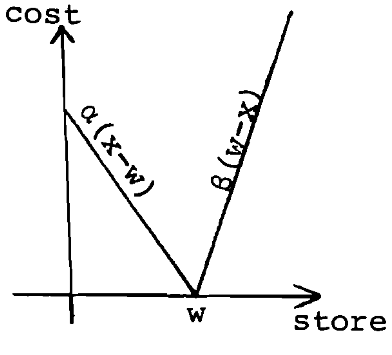

Figure 3

In static inventory problems the cost function has a graph as shown in Figure 3, where ω is demand, d, β are the store expenditures and losses respectively.

Very important applications which lead to nondifferentiable and stochastic optimization problems are the problems of long term planning. In these problems a typical cost function versus the output is given in Figure 4.

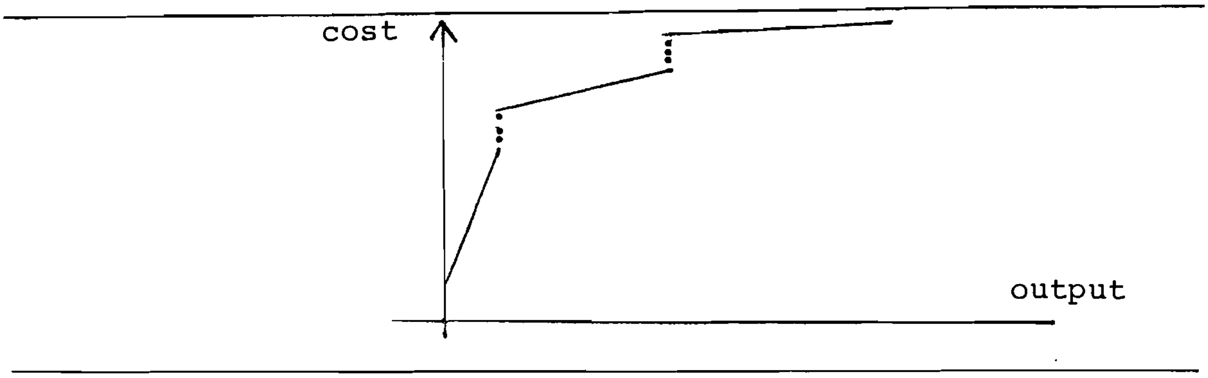

Figure 4

The steps of this function correspond to additional reconstruction investments for larger-scale plants.

Let us consider a model of long-term planning for optimal composition of an agriculture machinery park [10]. Let $b_i(k)$ be a quantity of work of the ith kind (harvesting, planting etc.) at the kth period, $x_{ij}(k)$ is the number of machines of the jth type for the ith kind of work; $W_{ij}(k)$ is a shift in the performance of the machines. It is required to minimize

$$\sum_{i,j,k} C_{ij}(k)\, x_{ij}(k) + \sum_j \max_k \sum_{i,j} x_{ij}(k)\, \lambda_k \quad ,$$

$$\sum_j W_{ij}(k)\, x_{ij}(k) \geq b_i(k) \quad , \qquad x_{ij}(k) \geq 0$$

where $C_{ij}(k)$ are shift expenses, λ_k are annual depreciations.

If we take into account that $b_i(k)$ are usually random values we obtain a stochastic minimax problem.

2. ON EXTREMUM CONDITIONS

The peculiarity of nondifferentiable and stochastic optimization problems in comparison with the classic problem of deterministic optimization becomes apparent already in optimality conditions. If f(x) is a convex differentiable function then the necessary and sufficient conditions of the minimum have the form:

$$f_x(x) = 0 \quad , \tag{20}$$

where

$$f_x = \frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \quad .$$

In the nondifferentiable case this condition transforms into requirement (Figure 5)

$$0 \in \{\hat{f}_x(x)\} \tag{21}$$

where

$$\{\hat{f}_x(x)\} = \partial f(x)$$

is a set (the subdifferential) of generalized gradients (the subgradients). These vectors $\hat{f}_x(x)$ satisfy the inequality

$$f(y) - f(x) \geq (\hat{f}_x(x), y-x) \quad , \quad \forall y \quad . \tag{22}$$

It should be noted that the notation $\hat{f}_x(x)$ 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, problems of two-stage stochastic programming etc. which are considered below.)

The complexity of nondifferentiable optimization problems results from the impossibility of practical usage of (21) for the answer to the question whether a specific point x may be a point of the minimum of f(x).

This discussion requires testing whether the 0-vector belongs to the set $\{\hat{f}_x(x)\}$ which usually has no constructive description. A further complication is checking the conditions (20), (21) by statistical methods. For example verifying the statistical hypothesis that for fixed x the mathematical expectation of the random vector $f_x(x,\omega)$ is 0, that is, whether

$$\underline{F}\, f_x(x,\omega) \quad = \quad 0 \quad .$$

Deterministic Methods of Nondifferentiable Optimization

There are two different classes of nondifferentiable optimization methods: the non-descent methods which started their development in the early 60's at the Institute of Cybernetics in Kiev [11,12] and the descent ones which appeared in the '70's in the western scientific literature (see [1] for a bibliography).

Let us discuss briefly the basic ideas of these two approaches.

An attempt to generalize the known gradient methods of the kind

$$x^{s+1} \quad = \quad x^s - \rho_s f_x(x^s) \quad , \qquad s = 0,1,\ldots$$

where x^s is an approximate solution at the s-th iteration, and ρ_s are step-size multipliers, for functions f(x) with a discontinuous gradient requires definition of an analogue of the gradient at points where the usual gradient does not exist. For almost differentiable functions the definition is made by limit transfer. A generalized gradient (almost gradient) of the almost differentiable function f(x) at point x is a vector $\hat{f}_x(x)$ belonging to the convex hull of the limit points of all sequences $\{f_x(x^s)\}$ where $\{x^s\}$ is a sequence of points at which the gradients $f_x(x^s)$ exist and whose limit point is x.

If f(x) is a convex function we get a set of vectors $\hat{f}_x(x)$ which satisfy (22).

Let us note that a convex function has a gradient almost everywhere. There are classes of problems however, in which every point with rational coordinates has no gradient and therefore, in any computational process at each iteration, we have to deal with a point of nondifferentiability.

Principal difficulties are connected with the choice of step multipliers ρ_s even for convex functions. It is impossible in practice to review the whole set of subgradients and to choose that one in the opposite direction to which leads the domain of smaller values of the objective function. Usually one can get only one of the subgradients and therefore there is no guarantee that a step according to the procedure

$$x^{s+1} = x^s - \rho_s \hat{f}_x(x^s) \quad , \qquad s = 0,1,\ldots \tag{23}$$

or to the more general one

$$x^{s+1} = \Pi_X(x^s - \rho_s \hat{f}_x(x^s)) \quad , \ s = 0,1,\ldots \tag{24}$$

(where $\Pi_X(\cdot)$ is a projection operator on the set X), will lead into the domain of the smaller values of f(x) (Figure 6).

Figure 6

To avoid this problem procedure (23) was proposed in 1962 by N.Z. Shor [11] and called the method of generalized gradients. It allows the use of any subgradient in the subdifferential. General conditions for its convergence have first been obtained by Y.M. Ermoliev [12] and independently by B.T. Polyak [13], where the ρ_s should satisfy the conditions

$$\rho_s \nrightarrow 0 \quad , \qquad \sum_{s=0}^{\infty} \rho_s = \infty \quad .$$

These conditions are very natural as (23) is a nondescent process i.e. the value of the objective function does not necessarily decrease from iteration to iteration even for arbitrarily small ρ_s.

The influence and close relations of research by I.I. Eremin on solutions of systems of inequalities and on nonsmooth penalty functions [14] to this area of work should be noted.

Since then the method (23) has been further developed (see review [16]) and rates of convergence have been studied.

E.A. Nurminski [16] studied the convergence of methods of the type (23) for the functions satisfying the following condition

$$f(y) - f(x) \geq (\hat{f}_x(x), y-x) + 0(||y-x||) \quad . \qquad (25)$$

Moreover he proposed a new proof technique for convergence based on the arguments ad absurdo, i.e. he adapted this technique for studying the convergence of nondescent methods of non-convex, non-smooth optimization.

As has already been said the algorithms constructed on the basis of (23) are simple and require relatively little storage. Thus let us consider an application of the method (24) to the development of iterative schemes of decomposition. For the function (10) one of the generalized gradients at point x^s is

$$\hat{f}_x(x^s) = c - u^s A$$

where u^s are dual variables corresponding to $y(x^s)$. Therefore the iterative scheme of decomposition according to the procedure (24) has the form

$$x^{s+1} = \max \{0, x^s - \rho_s(c - u^s A)\}, \quad s=0,\ldots \tag{26}$$

The same may be obtained by considering the function (14): if y^s is an approximate solution of the subproblem (15) for $x = x^s = \{x^s_{ij}\}$ and u^s are dual variables corresponding to y^s, then

$$x^{s+1} = \pi_x(x^s - \rho_s u^s), \quad s = 0,1,\ldots \tag{27}$$

where $\pi_x(\cdot)$ is the projection operator on the set (15). A very simple algorithm for the solution to this problem exists.

For the minimax problem (17) in the case when $g(x,y)$ for each $y \in Y$ is a convex function with respect to x, the subgradient is defined as $\hat{f}_x(x) = \hat{g}_x(x,y)|_{y=y(x)} = g_x(x,y(x))$ If $g(x,y)$ is continuously differentiable with respect to x then

$$\hat{f}_x(x) = g_x(x,y)_{y = y(x)} = g_x(x,y(x)).$$

If we use this formula for function (11), we obtain the following iterative method of decomposition:

$$x^{s+1} = \max \{0, x^s - \rho_s(c - u^s A)\}, \quad s = 0,1\ldots$$

where u^s is a solution of the subproblem :

$$(u, \ell - Ax^s) = \max, \quad u D \leq d, \quad u \geq 0$$

The iterative methods of decomposition based on the non-differentiable approach are effective techniques for the solution of different complex optimization problems. For example, for linear problems of optimal control we can use the method considered above. Consider the following problem: to find a control $x = (x(0),\ldots,x(N-1))$ and a trajectory $z = (z(0),\ldots,z(n))$, satisfying the state equations:

$$z(k+1) = A(k) + B(k) + a(k)$$

$$z(0) = z^0 \ , \ K = 0,1\ldots, N-1 \ ,$$

the constraints

$$G(k)z(k) + D(k)x(k) \leq \ell(k) \ , \ k = 0,1\ldots, N-1$$

$$u(k) \geq 0$$

and minimize the objective function

$$(c(N) \ , \ z(N) + \sum_{k=0}^{N-1} [c(k),z(k)) + (d(k),x(k))] \ ,$$

where $x(k) \in R^n$, $z(k) \in R^r$ The difficulty of this problem is connected with the state constraints. If matrice $G(t)$ 0, we can solve this problem with the help of the Pontzjagin's principle.

The dual problem [34] is to find dual control $\lambda = (\lambda(N-1),\ldots, \lambda(0))$ and dual trajectory $p = (p(N),\ldots p(0))$, subject to state equations

$$p(k) = p(k+1) \ A(k) - \lambda(k)G(k) + c(k)$$

$$p(N) = c(N), \ k = N-1,\ldots,0$$

and constraints

$$p(k+1) \ B(k) + \lambda(k)D(k) \leq d(k)$$

$$\lambda(k) \geq 0$$

which minimize

$$(P(0),z^0) \sum_{k=0}^{N-1} [(p(k+1),a(k)) + (\lambda(k),b(k))] \ .$$

We have the following analog of the iterative scheme of decomposition considered above (for finding the optimal control):

$$x^{s+1}(k) = \max \{0, x^s(k) - \rho_s [p^s(k+1)B(k) - \lambda^s(k)D(k) - d(k)]\}$$

where $\lambda^s(k)$ (k=N-1,...0) , $p^s(k)$, k = N-1,...,0 is a solution of the subproblem : minimize the linear function:

$$(p(0),z^0) + \sum_{k=N-1}^{0} [(p(k+1),a(k)) + (\lambda(k),b(k) +$$

$$+ d(k) - p(k+1)B(x) - \lambda(x)D(x),x^s{}_{(k)})]$$

under constraints

$$p(k) = p(k+1)A(k) - \lambda(k)G(k) + c(k)$$

$$p(N) = c(N) \ , \ k = N-1,\ldots,0 \ ,$$

$$\lambda(k) \geq 0 \ , \ k = N-1,\ldots, 0 \ ,$$

We may use the well-known Pontzjagin's principle for solving this problem. Its solution is reduced to the solution of N simple static linear programming problems.

Original work by Wolfe and Lemarechal (see [1]) on descent methods are, on one hand, a generalization of algorithms of ε-steepest descent studied by V.F. Demyanov [8] and on the other hand they are formally similar to algorithms of conjugate gradients and coincide with them in the differentiable case.

The set $\{\hat{f}_x(x^s)\}$ is required to implement the descent process. Since at the point x^s it is impossible to get the whole set $\{\hat{f}_x(x^s)\}$ an attempt can be made to construct it approximately. In Wolfe and Lemarechal's works, the following idea is used for this purpose. If at the point x^s the movement in the direction opposite to the subgradient $\hat{f}_x(x^s)$ leads to the decrease of the objective function by not less than $\varepsilon > 0$ (this is essential for convergence) the movement to x^{s+1} is made in this direction. If not, as trial step to a point z^{s1} is made in this direction, the subgradient $\hat{f}_x(z^{s1})$ is calculated and one returns to the point $z^{s0} = x^s$. The convex hull of $\hat{f}_x(z^{s0})$ and $\hat{f}_x(z^{s1})$ in a certain sense approximates $\{\hat{f}_x(x^s)\}$ from which one finds the element of the hull which has the least norm. If it is near zero, it should be excepted, according to the optimality criterion (21) that x^s is near optimal. Let the norm of this element be distinct from zero. If the direction from this point leads to a decrease of the objective function by not less than ε the move from x^s to x^{s+1} is made in this direction. If this is not true, only a trial step is made to a point z^{s2} , $\hat{f}_x(z^{s2})$ is calculated, then one returns to x^{s0}. The convex hull of the vectors $\hat{f}_x(z^{s0})$, $\hat{f}_x(z^{s1})$, $\hat{f}_x(z^{s2})$ is considered and so on.

The further development of subgradient schemes resulted in the creation of ε-subgradient processes. This technique, instead of subgradients, uses ε-subgradients introduced by Rockafellar [17]. The early results in this direction belong to Rockafellar [18], D. Bertsecas [19], C. Lemarechal [20], Nurminski and Zhelikhovski [21]. The recent research unveiled such properties of ε-subgradient mappings as Lipschitz continuity which make ε-subgradient methods attractive both in theoretical and practical respects.

Stochastic methods of NDO

Two classes of deterministic methods were discussed: non-descent and descent ones. The first class of the methods is easy to use on the computer but it does not result in a monotonic decreasing of the objective function. The second class obtains monotonic descent but has a complex logic and is rather difficult for computer implementation. Both classes have a common short coming, they require the exact computation of a subgradient (in a differentiable case - the gradient). Often however, there are problems in which the computation of subgradients is practically impossible. Random directions of search is a simple alternative method to construct nondifferentiable optimization descent procedures that do not require an exact computation of a subgradient and which are easy to use on the computer.

There are various ideas on how to construct methods of random search in deterministic problems which only require the exact values of objective and constraint functions. One of the simplest methods is as follows : from the point x^s, the direction of the descent is chosen at random and the motion in this direction is made with a certain step. The length of this step may be chosen in various ways, in particular such that:

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

Such methods are easy to implement on the computer and they can be made to have a good asymptotic behaviour. As shown in [22], they can have a geometric rate of convergence which is rare for the deterministic methods considered above.

Nondescent methods of random search are of prime importance in the solution of the most difficult problem arising in stochastic programming. In these extremum problems it is impossible to compute either subgradients or exact values of objective and constraint functions. The presence of random components in the search directions of nondescent procedures permits overcoming local minima, points of discontinuity, etc. Let us analyse first, in detail, the above mentioned difficulties of stochastic programming problems by way of concrete examples and then consider the general ideas for descent SQMs.

<u>The stochastic programming problem</u>

The problem (18)-(19) represents a general stochastic programming problem. It is a model of optimization of a stochastic system in which the decision (planned values of the system parameters x) is considered independent of the random factors. Such a situation is typical for planning the development of systems which will work in a random enviroment for a long time. There are other classes of stochastic systems in which the decisions are based on the actual knowledge of the random parameters of the system and thus the decision x becomes a random vector. Such situations usually occur in real-time control and short-term planning. In practice this problem can sometimes (via a decision rule) be reduced to the problem (18)-(19).

The main difficulty of problem (18)-(19), as has been noted, is that the functions $F^v(x)$, $v = \overline{o,m}$ often have no continuous derivatives. Another important difficulty is connected with condition (19). Let us consider some examples.

1. The two-stage problem

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 is resistant to random variations of the initial data. For this purpose the notion of correction is introduced and the losses connected with this correction are considered. An optimal long-term plan should minimize the total expenditures for the realization of the plan and for its possible correction.

The simplest two-stage stochastic programming problem may be formulated in the following way:

The decision z consists of two separate parts:

$$z = (x,y) = (x_1,...,x_n,y_1,...,y_m)$$

where with every z a certain loss is associated:

$$(c,x) + d,y)$$

Every decision variable should satisfy constraints:

$$Ax + Dy = \ell, \quad x \geq 0, \; y \geq 0 \; .$$

All coefficients $w = (d,\ell,A,D)$ are random variables and a decision is chosen in two stages.

Stage 1. The long-term decision x is made.

Stage 2. The random parameters $w = (d,\ell,A,D)$ are observed and a corrective solution y is derived from the known w:

$$\min \{(d,y) : Dy = B - Ax , y \geq 0\}$$

The problem is to find such vector x that the function

$$F^0(x) = Ef^0(x,w) = (c,x) + \tag{28}$$

$$+ E \min_{\substack{Dy=\ell-Ax \\ y \geq 0}} (d,y) = (c,x) + E(d,y(x,w))$$

has a minimum value.

It is evident that $F^0(x)$ is a convex, but in the general case nonsmooth function since the operation of the minimization is present under the integral sign. The value of the function

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

can be calculated without difficulty. To calculate $F^0(x)$ it is necessary to find the distribution of the $(d,y(x,w))$ as a function of x and then to calculate the corresponding integral (28) which is possible only in rare cases.

The problem (28) is strongly connected with large scale linear programming problems. For instance, if w has a discrete distribution: $w \in \{2,2,\ldots,N\}$ and $w = k$ with probability p_k and

$$p_k \geq 0, \sum_{k=1}^{N} p_k = 1$$

then the initial problem becomes the following:

$$(c,x) + (d(1),y(1) + (d(2),y(2)+\ldots+(d(N),y(N) = \min \qquad (31)$$

$$\begin{array}{llll} A(1)x + D(1)y(1) & & & = \ell(1) \\ A(2)x + & D(2)y(2) & & = \ell(2) \qquad (32) \\ A(N)x + & & D(N)y(N) & = \ell(N) \end{array}$$

$$x \geq 0,\ y(1) \geq 0,\ y(2) \geq 0,\ldots,y(N) \geq 0 \qquad (33)$$

where $y(k)$ is the correction of the plan if $w = k$. The number N may be very large. If only the coordinates of the vector $\ell = (\ell_1,\ldots,\ell_m)$ are random and each of them has two independent outcomes then $N = 2^m$.

2. The stochastic minimax problems

The objective function of the simplest stochastic minimax problem looks as follows

$$F^0(x) = E \max_{1 \leq i \leq M} \Big[\sum_{j=1}^{N} a_{ij}(w)x_j - \ell_1(w)\Big] \qquad (29)$$

or more generally

$$F^0(x) = E \max_{y \in Y} g(x,y,w) \qquad (30)$$

It should be noted that the two-stage problem (28) and the stochastic minimax problems generalize the problems (10),(16),(17).

A very important particular class of stochastic minimax problems arises in inventory control problems (a stochastic model of optimal structure of an agricultural machinery park is also stochastic minimax problem). Thus the expected expenditures in planning the stock $x_1,\ldots,x_n$ of nonhomogeneous products equal

$$F^0(x) = E \max \{\alpha(\sum_{j=1}^{n} \gamma_j x_j - w), \beta(w - \sum_{j=1}^{n} \gamma_j x_j)\} ,$$

where w represents demand, α,β are storage expenditures and losses and γ_j are the coefficients of substitution.

For problems (29),(30) it is again easy to calculate

$$f^0(x,w) = \max_{1 \le i \le M} [\sum_{j=1}^{n} a_{ij}(w) x_j - \ell_i(w)]$$

but $F^0(x)$ remains difficult. It is a convex but often a nonsmooth function.

3. The stochastic problem of optimal control

The same difficulties are inherent in stochastic problems of the theory of optimal control. Taking into account the dynamics of a complex system leads to the following very general problem: find $x = (x(0),x(2),\ldots,x(N-1))$ which minimizes

$$F^0(x) = E\ \Phi(z(0),\ldots,z(N),x(0),\ldots,x(N-1),w) \tag{34}$$

under the constraints

$$z(k+1) = g(z(k),x(k),w,k), z(0) = z^0, \tag{35}$$

$$x(k) \in X(k), k = 0,1\ldots,N-1 \tag{36}$$

In particular, one might have

$$F^0(x) = E \max_k \|z(k) - z^*(k)\| \tag{37}$$

Thus the solution of even the simplest stochastic programming problem which we considered above requires the development of numerical methods of optimization without using exact functional values. The stochastic quasigradient methods [10,23] allow to solve successfully the above mentioned problems with the rather arbitrary but in practice useful measures P(dw).

The general idea of stochastic quasigradient methods

Consider the problem

$$\min \ F^0(x) \ : \ f^i(x) \leq 0, \ i = 1,m, \ x \in X\}$$

We assume here that $F^v(x), v=0,\bar{M}$ are convex functions, i.e. where $\hat{F}_x^v$ is a subgradient and the set

$$F^v(z) - F^v(x) \geq (\hat{F}_x^v(x), z-x)$$

is convex.

In stochastic quasigradient (SQG) methods the sequence of approximations $x^0, x^1 \ldots, x^s \ldots$, is constructed with the help of random vectors $\xi^v(s)$ and random quantities $\Theta_v(s)$ which are stochastic estimates of the values of subgradients $\hat{F}_x^v(x^s)$ and of the function $\hat{F}^v(x^s)$:

$$E(\xi^v(s)(x^0, \ldots, x^s) = \hat{F}_x^v(x^s) + a^v(s)$$

$$E(\Theta_v(s)(x^0, \ldots, x^s) = F^v(x^s) + \ell_v(s) \ ,$$

where $a^v(s)$ is a vector, $\ell_v(s)$ is a number depending upon x^0, $x^1, \ldots, x^s, \ldots$, where usually $a^v(s) \to 0, \ell_v(s) \to 0$ for $s \to \infty$. Thus in these methods instead of exact values of $\hat{F}_x^v(x^s), F^v(x^s)$ $\xi^v(s), \theta, (s)$ are used. For further understanding it is important to see that the random values $\theta_v(s)$ and vectors $\xi^v(s)$ are easily calculated. For example, if

$$F^v(x) = Ef^v(x,w)$$

then $\theta_v(s) = f^v(x^s, w^s)$ where the w^s result from mutually independent draws of w.

We have

$$E(\theta_V(s)\ /x^s) = E(f^V(x^s,w)/x^s) = f^V(x^s)$$

For a two-stage problem

$$\xi^0(s) = c - u(x^s,w^s)\ A(w^s) \tag{38}$$

where $u(x^s,w^s)$ are dual variables corresponding to the second-stage optimal plan $y(x^s,w^s)$. It can be shown [10] that

$$E(\xi^o(s)/x^s) = \hat{F}_x^0(x^s)$$

where $\hat{F}_x^0(x^s)$ is a subgradient of the function (28). For the objective function (29) of the stochastic minimix problem the vector $\xi^0(s) = (\xi_1^0(s),\dots,\xi_n^0(s))$ is calculated by the formula

$$\xi_j^0(s) = a_{i_s j}(w^s) \tag{39}$$

where i_s is defined by the relation

$$\sum_{j=1}^{n} a_{i_s j}(w^s)x_j^s - \ell_{i_s}(w^s) = \max_i \ [\sum_{j=1}^{n} a_{ij}(w^s)x_j - \ell_i(w^s)]$$

It may be shown [10] that

$$M(\xi^0(s)/x^s) = \hat{F}_x^0(x^s)$$

where $\hat{F}_x^0(x^s)$ is a subgradient of the function (29). It should be noted that stochastic quasigradient methods are also applicable to NDD deterministic problems, without requiring values of subgradients. For example, for the deterministic minimax problem (17) the vector

$$\xi^o(s) = \frac{g(x^s+\Delta_s h^s, y(x^s)) - g(x^s,y(x^s))}{\Delta_s}\ h^s \tag{40}$$

where $\Delta_s > 0, h^s$ is the result of independent random draws of the random vector $h = (h, \ldots, h_n)$ whose components are independently and uniformly distributed over $[- \quad, 1]$. (40) satisfies the condition

$$E(\xi^0(s)/ x^s = \hat{f}^0_x(x^s) + a^0(s)$$

where $\hat{f}^0_x(x^s)$ is a subgradient of the function (17) and $\|a^0(s)\| \leq \Delta_s$ const, if $g(x,y)$ has uniformly limited second derivatives with respect to $x \in x$, $y \in Y$. For the objective function of the stochastic minimax problem (30) the vector $\xi^0(s)$ have the same formula (also see [20]) :

$$\xi^0(s) = \frac{g(x^s + \Delta_s h^s, y(x^s, w^s), w^s) - g(x^s, y(x^s, w^s), w^s)}{\Delta_s} h^s$$

It is remarkable that independent of the dimensionality of the problem the vectors (40),(41) it can be found by calculating the functions $g(x,y)$, $g(x,y,w)$ at two points only. This is particularly important for extremum problems of large dimensionality. Let us consider a number of SQG methods in which $\theta_v(s), \xi^v(s)$ are used instead of $F^v(x^s)$, $\hat{F}^v_x(x^s)$.

THE SQG METHODS

1. The stochastic quasigradient projection method

Let is be required to minimize the convex function in $x \in X$, where X is a convex set.

The method is defined by the relations:

$$x^{s+1} = \pi_x(x^s - \rho_s \xi^0(s)), \quad s = 0, 1, \ldots, \tag{42}$$

where $\pi_x(\cdot)$ is a projection operation on X, ρ_s one step multipliers.

If $\xi^0(s) = \hat{F}^0_x(x^s)$, we obtain the well-known method of generalized gradients (28). If

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

where the function $F^0(x)$ has uniformly limited second derivatives, it can be shown that for

$$\xi^0(s) = \sum_{j=1}^{n} \frac{f^0(x^s+\Delta_s \ell^j \quad w^{s^j}) - f(x^s,w^{s^j})}{\Delta s} \; e^j \qquad (43)$$

we have

$$E\ \xi^0(s)(x^s) = F^0_x(x^s) + a^0(s)$$

where $||a^0(s)|| \leq \text{const}\ \Delta_s$, $w^{so}_{,..,}w^{si}_{,..,}w^{sn}$ result from independent draws over w. Then the method (41)-(42) corresponds to the stochastic approximation method [24,25]. The method (42) has been proposed in [23]. The characteristic requirements, under which the sequence $\{x^s\}$ converges with probability to the solution, are: if $\|x^k\| \leq B$, $k = \overline{0,s}$, then $E(||\xi^0(s)||^2/x^0,\dots,x^s) \leq C_B$, where B,C_B are constants; ρ_s are step multipliers which may depend upon x^0,x^1,x^s :

$$\rho_s \geq 0, \sum_{s=o}^{\infty} \rho_s = \infty \quad \text{with probability 1,} \qquad (44)$$

$$\sum_{s=0}^{\infty} E(\rho_s^2+\rho_s||\ a^0(s)\ ||) < \infty \qquad (45)$$

Particularly if ρ_s are deterministic and independent of $(x^0,\dots x^s)$ then, under (44),(45) we obtain for the method (41) using the random direction (40)-(41) that

$$\rho_s \geq 0\ ,\ \Sigma\,\rho_s = \infty,\ \Sigma\,\rho_s\,\Delta_s < \infty \quad \Sigma\,\rho_s \ < \infty\ ,$$

The methods, which we shall consider below, converge under conditions approximately analogous to those mentioned above.

From (38) and (41) we obtain the following method of solving two-stage problems:

(i) For given x^s observe the random realisations of d,A Bℓ, which we note as:

$$d(s)A(s), B(s), \ell(s)$$

(ii) Solve the problem

$$(d^s, y) = \min$$

$$B(s)y \leq \ell_s - A_s X^s ,$$

$$y \geq 0$$

and calculate dual variables u^s.

(iii) get

$$\xi^0(s) = c(½) + u(s)\ A(s) \text{ and change } X^s :$$

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

go to (i).

It is worth noticing that this method can be regarded as a stochastic version of the iterative procedure of decomposition (28). It is simply implemented on the computer and it permits to solve extremely large-scale problems of the kind (32)-(33).

2. The stochastic linearization method.

Let the function $F^0(x)$ have continuous derivatives. If $F_x^0(x^s)$ is known then standard linearization is defined by the relations:

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

$$(F_x^0(x^s),\ \bar{x}^s) = \min_{x \in X} (F_x^0(x^s), x)$$

In the case where $F_x^0(x^s)$ is unknown, the stochastic variant of this method has been studied in [10,26] and is defined by the relations

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

$$(v^0(s),\bar{x}^s) = \min_{x \in X} (v^0(s),x) \tag{46}$$

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

where δ_s satisfy the conditions of the kind (44)-(45). It is worth noting that if instead of $v^0(s)$ the vectors $\xi^0(s)$ are used that, as simple examples show, the method does not converge. If $\xi_s = 1/s+1$ then

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

In this method on every iteration the subproblem is to be solved in the region X. For this problem the well-known methods of nonlinear programming or linear programming can be applied and will not require great computational efforts especially as an initial approximation of $\bar{x}^s$ the point $\bar{x}^{s+1}$ is chosen.

Consider now the general problem of minimization of the function

$$F^0(x) \tag{47}$$

under conditions

$$F^i(x) \leq 0,\ i = \overline{i,m} \tag{48}$$

$$x \in X \tag{49}$$

where $F^V(x)$ are convex functions, X a convex set.

3. The penalty functions stochastic method

Constraints (48) can be taken into account by means of penalty functions and instead of the general problem we can consider the problem of minimizing the function

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

on the set X.

Since it is practically impossible to calculate $F^i(x)$ in problems of the stochastic programming i.e. it is impossible to find min $(0,F^i(x))$, [27] defined the relations

$$x^{s+1} = \pi_X(x^s-\rho_s(\xi^0(s) + c \sum_{i \geqq 1} \min(0,z_i(s))\xi^i(s)), s=0,1..., \quad (5\dot{0})$$

$$z_i(s+1) = z_i(s) + \delta_s(\theta_i(s)-z_i(s)), \; s=0,1..., \quad (51)$$

4. Besides the above mentioned methods there are many others (see [29]). In particular Gupal &28] has studied the method characterized by the relations:

$$x^{s+1} = \pi_X(x^s-\rho_s\zeta^s), \; s=0,1...,$$

$$\zeta^s = \begin{cases} \zeta^0(s), \text{ if } zi_s(s) = \max\limits_{i=1 \leq m} z_i(s) \leq 0 \\ \zeta^i s(s), \text{ if } z_{i_s}(s) > 0, \end{cases}$$

where the values $z_i(s)$ are defined by the relations (51).

5. Non-convex functions

In [16] the convergence of the stochastic quasigradient methods for the functions $F^v(x)$ satisfying the condition (25) was studied. We also note the investigation of the minimization of almost-everywhere differentiable functions and discontinuous functions [28,29,30,35]. In this paper the simple and easily implemented methods for the problems (17) and others, appearing in the theory at multicriteria optimization were developed. In these papers the convergence of the following methods have been studied.

$$x^{s+1} = x^s - \rho_s \sum_{j=i} \frac{f(\bar{x}^s+\Delta_s e^j) - f(\bar{x}^s)}{\Delta_s} e^j \tag{52}$$

where e^j are unit vectors of the point $\bar{x}^s$ is randomly chosen in a neighborhood of the point x^s with radius $r_s \to 0, s\to\infty$.

The procedures, as (52), are based on the general ideas of solving limit extremum problems, which have begun to be developed in [33].

6. The limit extremum problems.

Briefly, the essence of this theory is the following. Let it be required to minimize the function f(x) without continuous derivatives. A sequence is considered to be composed of "good" functions $f^s(x)$, e.g., smooth ones which converge at f(x) for $s \to \infty$ and the procedures of the following form:

$$x^{s+1} = x^s - \rho_s f_x^s(x^s), \quad s=0,1\ldots, \tag{53}$$

Under rather general conditions it is possible to show that

$$f^s(x^s) \to \min f(x)$$

Often approximate functions have the form of mathematical expectations

$$f^s(x) = \int f(x-\theta)\ P_s(\theta) = Ef(x-\theta^s) \tag{54}$$

where the measure $P_s(dw)$ for $s \to \infty$ centers at the point x. Hence instead of the procedure (53) the realization of which requires exact value of the gradient of the mathematical expectation (54) the stochastic quasigradient methods are used which employ the vectors ξ^s satisfying the condition

$$E(\xi^s/x^0,\dots,x^s) = \hat{f}_x^s(x^s) + a^s,$$

where $\hat{f}_x^s(x)$ -subgradient of the function $f^s(x)$. For example, if for function (54) we consider random vector (stochastic quasigradient) ξ^s type (43) then we obtain the method (52); if we consider random vector type (41), we obtain the following method

$$x^{s+1} = x^s - \rho_s \frac{f(\tilde{x}^s+\Delta_s h^s) - f(\tilde{x}^s)}{\Delta_s} h^s, \; s=0,1\dots,$$

where $x^s = (\tilde{x}^s,\dots,\tilde{x}_n^s)$ is a random point P_s distributed in a neighborhood of the point x^s. If f(x) satisfies the Lipschitz local condition, then distributions P_s can be uniformly in an n-dimensional cube with the side r_s, e.g. $\tilde{x}_j^s$ are independent random values uniformly distributed on intervals $[x_j^s - \frac{r_s}{2}, x_j^s + \frac{r_s}{2}]$. Similar distributions are applicable when f(x) is semicontinuous function. Then the function

$$F^s(x) = Ef(x-h^s-\tau^s)$$

where h^s, τ^s are random vectors with independent components uniformly distributed on $[-\frac{r_s}{2}, \frac{r_s}{2}]$ is smooth, $F^s(x) \to f(x)$ is uniform in any bounded domain.

These approaches seem to be very important in nonsmooth and particularly discontinuous optimization. Thus in [35] it was shown that general scheme of linearization method may be used for the optimization of a wide range of nonconvex nonsmooth functions. Let us examine a problem of minimization of a function f(x) under constraints $x \in X$, where f(x) satisfies the Lipschitz local condition, X is a convex compact in R^n. The following method is considered

$$x^{s+1} = x^s + \rho_s(\bar{x}^s - x^s),\quad 0 \leq \rho_s \leq 1,\quad s=0,1\ldots,$$

$$(v(s),\bar{x}^s) = \min_{x \in X} \ (v(s),x)'$$

$$v(s+1) = v(s) + \delta_s(\theta^s - v(s)) ,$$

where $x^0 \in X$; δ_s satisfy the conditions of the kind (44)-(45);

$$\theta^s = \frac{1}{r_s} \sum_{j=1} [f(\tilde{x}_1^s,\ldots,x_j^s + \frac{r_s}{2},\ldots,\tilde{x}_n^s) - f(\tilde{x}_1^s,\ldots,x_j^s - \frac{r_s}{2},\ldots,\tilde{x}_n^s)]e^j,$$

x_j^s are independent random values uniformly distributed on intervals $[x_j^s - \frac{r_s}{2},\ x_j^s + \frac{r_s}{2}]$.

Some applied NDO, STO problems were briefly discussed in this work. There are many applications of STO numerical methods in mathematical statistics, complex systems, identification, reliability, inventory control, production allocation [10]. The deterministic, stochastic, descent and nondescent methods were considered. Each one requires some definite information about objective and constraint functions. Deterministic descent methods use the exact values of these functions and their subgradients, stochastic descent methods use only the exact values of functions; deterministic nondescent methods require only exact values of subgradients; stochastic nondescent methods do not use values of functions and exact values of their subgradients. Obviously, every method reveals its advantages in a specific class of extremum problems, for instance, complex stochastic programming problems are solvable only by stochastic nondescent methods.