Constrained Minimization Methods

Article *in* USSR Computational Mathematics and Mathematical Physics · December 1966

DOI: 10.1016/0041-5553(66)90114-5

2 authors, including:

Boris T. Polyak
Institute of Control Sciences

CONSTRAINED MINIMIZATION METHODS*

E.S. LEVITIN and B.T. POLYAK

Moscow

(*Received* 24 *September* 1965)

Introduction

CONSTRAINED extremum problems form a wide class of problems often encountered in pure and applied mathematics. We only need to recall examples such as linear and non-linear programming, the Lagrange problem in the calculus of variations, optimal control problems, problems of best approximation and variational problems for partial differential equations. Each of these types was at one time considered in isolation. To-day, a general mathematical theory of extremal problems is being created on the basis of functional analysis. For instance, the scheme proposed by A.Ya. Dubovitskii and A.A. Milyutin [1] enables the necessary conditions for a minimum to be obtained in a unified way for all such problems, starting with the duality theorem in linear programming and ending with the Pontryagin maximum principle for optimal control. Very general results have also been obtained regarding the existence and uniqueness of the extremum.

In the present paper we consider another aspect of extremal problems, namely methods of solving them, from the same unified functional-analytic viewpoint. The methods are classified and theorems proved for certain of them. These general theorems are illustrated by some specific examples.

Using this approach we are naturally prevented from considering many familiar methods which in essence use some specific feature of the particular problem (e.g. methods of dynamic programming). This paper consequently makes no pretence at offering a survey of numerical methods of minimization. Our aim is to give as general a statement as possible

* *Zh. vȳchisl. Mat. mat. Fiz.* 6, 5, 787 - 823, 1966.

of certain methods which are in part familiar in computational practice, and to state precisely the conditions for their convergence. We pay hardly any attention to another important aspect of the problem, namely the computational side. But it seems to us that our purely theoretical results on the convergence of methods do in fact throw light on this aspect. Some discussion of this topic will be found in Section 12.

1. Notes on the mathematical theory of extremal problems

We devote this section to the mathematical theory that will be utilized below. A familiarity with the general ideas of functional analysis [2, 3] is assumed.

All functionals and sets are considered below in a real Banach space E.

As usual, we call the functional $f(x)$ convex on Q if $f((x+y)/2) \leqslant \frac{1}{2} f(x) + \frac{1}{2} f(y)$ for all $x, y \in Q$ (or strictly convex if the inequality is strict). Throughout, we shall only consider continuous convex functionals. Hence the convexity condition can also be written as $f(\alpha x + (1-\alpha)y) \leqslant \alpha f(x) + (1-\alpha) f(y)$, $0 \leqslant \alpha \leqslant 1$. Similarly, the set Q is called convex if $z \in Q$ for all $x, y \in Q$, where $z = \frac{1}{2}(x+y)$, and strictly convex if z is an interior point of Q. The linear functional $c \in E^*$ is called a support functional to $f(x)$ at the point x if $f(x+y) \geqslant f(x) + (c, y)$ for all $y \in E$, and a support functional to Q at the point x if $(c, x) \leqslant (c, y)$ for all $y \in Q$. If the zero linear functional is a support to $f(x)$ at the point x^*, then x^* is a minimum of $f(x)$ on E. Here and below, E^* is the conjugate space and (c, x) is the value of the linear functional $c \in E^*$ on the element $x \in E$ (in particular, the scalar product in Hilbert space). A convex functional $f(x)$ has a support functional at any point x. This support functional is unique and is the same as the gradient $f'(x)$ if $f(x)$ is differentiable (the Fréchet derivative is always understood). With $f(x)$ differentiable, the convexity condition is equivalent to $(f'(x) - f'(y), x - y) \geqslant 0$ for any x and y. If the convex functional has a second derivative $f''(x)$, then $f''(x)y, y) \geqslant 0$ for all x and y. Finally, a convex functional is weakly semicontinuous from below.

Most results concerning the existence of an extremum can be obtained from the following.

Theorem 1.1

A functional weakly semicontinuous from below has a minimum on a weakly compact set.

In particular, it follows from this that

Theorem 1.2

A convex functional $f(x)$ has a minimum on any bounded closed convex set Q of reflexive space E.

We quote an elementary theorem on the uniqueness of the minimum.

Theorem 1.3

If, in addition to the conditions of Theorem 1.2, $f(x)$ is strictly convex, or if Q is strictly convex, while $f(x)$ has no zero support functionals on Q, then its minimum is unique.

We now turn to the convergence of minimizing sequences. Since the successive approximations in many minimization methods do not in general satisfy constraints, it is worth introducing the following:

Definition [4]. We call $x^n \in E$, $n = 1, 2, \ldots$, a generalized minimizing sequence (GMS) for $f(x)$ on Q if

a) $\lim\limits_{n\to\infty} f(x^n) = f^* = \inf\limits_{x\in Q} f(x)$, b) $\lim\limits_{n\to\infty} \rho(x^n, Q) = \lim\limits_{n\to\infty} \inf\limits_{x\in Q} \|x^n - x\| = 0.$

Theorem 1.4

In the circumstances of Theorem 1.2, we can extract from any GMS a subsequence weakly convergent to the minimum. In the circumstances of Theorem 1.3, any GMS is weakly convergent to the (unique) minimum.

We shall be interested in cases where strong, as well as weak, convergence of the GMS can be proved. For this, we introduce the following definition [4, 5]. The functional $f(x)$ is called uniformly convex if there exists a function $\delta(\tau)$, $\delta(\tau) > 0$ for $\tau > 0$ (which can be assumed monotonic), such that $f((x + y)/2) \leqslant \frac{1}{2}(f(x) + f(y)) - \delta(\| x - y \|)$ for all x, y. This condition is equivalent to the following: for any $x \in E$ there exists a support functional $c \in E^*$ such that $f(x + y) \geqslant f(x) + (c, y) + \delta(\| y \|)$ for any y. Further, we call a set Q uniformly convex if there exists a function $\delta(\tau)$, $\delta(\tau) > 0$ for $\tau > 0$, such that

$(x + y)/2 + z \in Q$ for any x, $y \in Q$ and any z: $\| z \| \leqslant \delta(\| x - y \|)$. Finally, we call a uniformly convex functional (set) strongly convex if $\delta(\tau) = \gamma\tau^2$, $\gamma > 0$. Such functionals and sets have the following properties. A uniformly convex function $f(x)$ is bounded from below in E, while the set $S = \{x\colon f(x) \leqslant \lambda\}$ is bounded for all λ. If, in addition, $f(x)$ satisfies a Lipschitz condition on S, then S is uniformly convex. Every uniformly convex set different from E is bounded. In finite-dimensional space every bounded strictly convex set is uniformly convex, while every strictly convex functional is also uniformly convex on any bounded set. If $f(x)$ is differentiable, the uniform convexity condition is equivalent to $(f'(x) - f'(y), x - y) \geqslant \delta(\| x - y \|)$. If $f(x)$ is twice differentiable, the condition for strong convexity is equivalent to $f''(x)y, y) \geqslant \gamma \| y \|^2$, $\gamma > 0$, for all x, y.

Theorem 1.5

If Q is a closed convex set, and $f(x)$ a uniformly convex functional, then every GMS is (strongly) convergent to the (unique) minimum x^* of $f(x)$ on Q. Here $f(x) - f(x^*) \geqslant \delta(\| x - x^* \|)$ for all $x \in Q$.

Theorem 1.6

If Q is a closed uniformly convex set, and $f(x)$ a functional weakly semicontinuous from below, with a unique minimum at the boundary point x^* on Q, then every GMS is (strongly) convergent to x^*. In particular, if $f(x)$ is a convex functional, having no zero support functionals on Q, then $f(x) - f(x^*) \geqslant \lambda\delta(\| x - x^* \|)$, $\lambda > 0$, for all $x \in Q$.

As we shall see below, only a slight strengthening of the conditions of Theorems 1.4 - 1.6 is required for convergence of many minimization methods.

In many problems the admissible set Q is specified by $Q = \{x\colon g(x) \leqslant 0\}$, where $g(x)$ is a functional. In connection with the concept of GMS, it is important to know when the condition $g(x^n) \to +0$ implies $\rho(x^n, Q) \to 0$. We call the constraint $g(x) \leqslant 0$ correct if this is true. It turns out that $g(x) \leqslant 0$ is correct in the following cases: (a) if E is finite-dimensional, $g(x)$ is continuous and $S = \{x\colon g(x) \leqslant \varepsilon\}$ is bounded for some $\varepsilon > 0$; (b) if $g(x)$ is convex, Q is bounded and there exists x^0 such that $g(x^0) \geqslant 0$; (c) if $g(x)$ is convex and $\| g'(x) \| \geqslant \varepsilon > 0$ for all x such that $g(x) = 0$ (here $g'(x)$ denotes any support functional to $g(x)$ at the point x; (d) if $g(x)$ is differentiable, Q bounded, $g'(x)$ satisfies a Lipschitz condition, and there exist $\varepsilon > 0$, $\delta > 0$ such that $\| g'(x) \| \geqslant \varepsilon$ for $|g(x)| \leqslant \delta$; (e) if $g(x) \geqslant 0$ for all x (i.e. $Q = \{x\colon g(x) = 0\}$), $g(x)$ is differentiable, $g'(x)$ satisfies a Lipschitz condition, and

$\| g'(x) \|^2 \geqslant \lambda g(x)$, $\lambda > 0$; (f) if $g(x)$ is uniformly convex.

Notice that, if the initial problem is such that the *GMS* convergence cannot be said to be strong, yet at the same time good convergence to the solution is desirable in the sense of closeness in norm as well as with respect to the functional, the artificial device of A.N. Tikhonov regularization can be used to obtain a strongly convergent sequence [6, 4].

In conclusion we dwell briefly on the necessary and sufficient conditions for an extremum. In the case of unconstrained problems, such conditions have been well known ever since the foundations of analysis and the calculus of variations were laid. They were later extended to the case of constraints in the form of equations (the method of Lagrange multipliers). But it is only recently that the necessary conditions for an extremum have been stated for problems of mathematical programming (the Kuhn - Tucker Theorem, [7]) and for optimal control problems (Pontryagin's maximum principle, see [8]). The necessary condition for a minimum was obtained by L.V. Kantorovich [9] in 1940 for the general problem of minimizing a functional on a set of Banach space. Recently, A.Ya. Dubovitskii and A.A. Milyutin [1] have developed a unified technique for obtaining the extremum conditions for a very wide variety of problems.

We shall use the conditions for a minimum of the functional $f(x)$ on a convex set Q of Banach space E in the following form (x^* is the minimum throughout what follows):

(a) if $f(x)$ is differentiable, then

$$(f'(x^*), x - x^*) \geqslant 0 \quad \text{for all} \quad x \in Q, \tag{1.1}$$

i.e. $f'(x^*)$ is a support functional to Q at x^*;

(b) if $f(x)$ is differentiable and E Hilbert space, then

$$\alpha \| f'(x^*) \| = \min_{x \in Q} \| x^* - \alpha f'(x^*) - x \| \quad \text{for all} \quad \alpha > 0, \tag{1.2}$$

i.e. the projection of the vector $x^* - \alpha f'(x^*)$ on Q is the same as x^*;

(c) if $f(x)$ is convex and twice differentiable, then

$$(f'(x^*), x - x^*) + {}^1/_2(f''(x^*)(x - x^*), x - x^*) \geqslant 0, \tag{1.3}$$

or

$$\min_{x\in Q}[f(x^*)+(f'(x^*),x-x^*)+{}^1/_2(f''(x^*)(x-x^*),x-x^*)]=f(x^*).$$

We shall see below that the method of gradient projection (Section 5) can be regarded as an iterational means of satisfying condition (1.2), the conditional gradient method (Section 6) a means of satisfying condition (1.1), and Newton's method (Section 7) condition (1.3).

2. Examples of extremal problems

General minimization methods will be illustrated throughout this paper by examples, mainly from optimal control problems. We describe these problems in the present section and indicate conditions ensuring the functional and set properties participating in the general theorems (smoothness, convexity, uniform convexity etc.).

1. The problem of mathematical programming amounts to minimizing a function $f(x)$ of m variables, $x = (x_1, \ldots, x_m)$ on a set Q specified by the constraints $g_i(x) \leqslant 0$, $i = 1, \ldots, r$. We shall often reduce different problems to a sequence of problems of linear ($f(x)$, $g_i(x)$ linear) or quadratic ($f(x)$ quadratic, $g_i(x)$ linear) programming, for which the final methods of solution are well known [7, 10].

2. The optimal control problem [8] amounts to minimizing the functional

$$f(u)=\int_0^T F(x(t),\ u(t),t)\,dt+\Phi(x(T)), \tag{2.1}$$

where the phase variables $x(t)$ and control $u(t)$ are connected by

$$\frac{dx}{dt}=\varphi(x,u,t), \qquad x(0)=c, \tag{2.2}$$

in the presence of certain constraints. We shall only consider problems in which the time T and the initial state $x(0)$ are fixed. We assume that $x(t) = (x_1(t), \ldots, x_m(t))$, $u(t) = (u_1(t), \ldots, u_r(t))$, $r \leqslant m$. The solution will be sought in the class of functions $u(t) \in L_2{}^r(0, T)$, i.e.

$$\|u\|_{L_2^r}=\left(\int_0^T\|u(t)\|_{E_r}^2dt\right)^{1/2}=\left(\int_0^T\sum_{i=1}^r u_i^2(t)\,dt\right)^{1/2}.$$

This class of functions is fairly wide (in particular, it includes all bounded measurable $u(t)$); on the other hand, the space $L_2{}^r$ is Hilbert, which is extremely handy for various methods of minimization. The notation $f(u)$ (and not $f(x, u)$) in (2.1) emphasizes that $u(t)$ is regarded as the independent variable, while $x(t)$ is found (for the given $u(t)$) as a solution of (2.2).

We note some important particular cases of the functional $f(u)$:
(a) $F \equiv 0$ is the problem of terminal state optimization (e.g. $\Phi(x(T)) = \|x(T) - d\|_m$ (b) $\Phi \equiv 0$, $F(x, u, t) = F(x, t)$ is the problem of best approximation of a given trajectory (e.g. $F(x, t) = \|x(t) - a(t)\|^2_{E_m}$) (c) $\Phi \equiv 0$, $F(x, u, t) = F(u, t)$ is the problem of control power minimization (e.g. $F(u,t) = \|u(t)\|^2_{E_r}$).

An important particular case of system (2.2) is the linear problem in which (2.2) becomes

$$\frac{dx}{dt} = A(t)x + B(t)u, \qquad x(0) = c, \tag{2.3}$$

where $A(t)$, $B(t)$ are $m \times m$ and $m \times r$ matrices, continuously dependent on t.

In the particular case when $A(t) \equiv 0$, $B(t) \equiv I$ (I is the unit matrix), we obtain the classical functional of the calculus of variations.

We shall not quote the proofs of the following properties of the functional $f(u)$, since they are quite laborious and of little interest.

a. Let $\varphi(x, u, t)$ satisfy a Lipschitz condition in x and u for all x, u, t (i.e. for instance, $\|\varphi(x_1, u, t) - \varphi(x_2, u, t)\|_{E_m} \leqslant K\|x_1 - x_2\|_{E_m}$ where K is independent of u and t; other conditions of this type will be understood similarly). Let $\Phi(x)$ and $F(x, u, t)$ satisfy a Lipschitz condition in x on any set bounded in x and let $|F(x, u + \bar{u}, t) - F(x, u, t)| \leqslant M\|u\|_{E_r}\|\bar{u}\|_{E_r} + R\|\bar{u}\|^2_{E_r}$ (implying that $F(x, u, t)$ satisfies a Lipschitz condition in u on any set bounded in u and that $F(x, u, t)$ as a function of u is increasing not faster than quadratically). Then $f(u)$ is a continuous functional, satisfying a Lipschitz condition on any bounded set.

b. Let $\varphi(x, u, t)$ have partial derivatives with respect to x and u, satisfying a Lipschitz condition and bounded for all x, u, t, while

$F(x, u, t)$ has partial derivatives with respect to x and u satisfying a Lipschitz condition on any set bounded in x. Finally, let $\Phi(x)$ be differentiable and its gradient satisfy a Lipschitz condition on any set bounded in x. Then $f(u)$ is differentiable, its gradient satisfies a Lipschitz condition everywhere, and has the form

$$f'(u) = h(t) = F_u - \varphi_u{}^*\psi, \tag{2.4}$$

where $\psi(t)$ is the solution of the system

$$\frac{d\psi}{dt} = -\varphi_x{}^*\psi + F_x, \qquad \psi(T) = -\Phi'(x(T))$$

(here, as throughout what follows, the subscript denotes differentiation with respect to the corresponding argument, and the asterisk the conjugate operator, in this case the transposed matrix).

Notice that the necessary condition $f'(u) = 0$ for a minimum in the unconstrained problem is none other than the classical Euler - Lagrange equation for this problem in the integral form. Notice also that $f'(u)$ is connected with the function $H(x, u, t, \psi)$ of Pontryagin [8] by the simple relationship $h(t) = -H_\mu$.

All further properties of $f(u)$ are quoted for a linear system (2.3) only. Unfortunately, we know of no extension of them to a non-linear system.

c. If $F(x, u, t)$ is continuous with respect to $\{x, u\}$, measurable with respect to t, convex with respect to u, while $\Phi(x)$ is continuous, then $f(u)$ is weakly (and hence strongly) semicontinuous from below.

d. If $F(x, u, t)$ is convex with respect to $\{x, u\}$, and $\Phi(x)$ convex, $f(u)$ must also be convex.

e. If, in addition to property d, $F(x, u, t)$ is uniformly (strongly) convex with respect to u (more precisely, if $F((x_1 + x_2)/2, (u_1 + u_2)/2, t) \leqslant {}^1/_2 F(x_1, u_1, t) + {}^1/_2 F(x_2, u_2, t) - \delta(\|u_1 - u_2\|)$ for any x_1, x_2, u_1, u_2, t, then $f(u)$ is a uniformly (strongly) convex functional.

Properties d and e can be substantially improved [5], Theorem 10, where this is done for the elementary variational problem).

f. If $F(x, u, t)$ is twice continuous differentiable with respect to $\{x, u\}$, and $\Phi(x)$ twice differentiable, then $f(u)$ is twice differentiable and its second derivative $f''(u)$ is

$$(f''(u)\bar{u}, \bar{u}) = \int_0^T [(F_{xx}\bar{x}, \bar{x}) + 2(F_{xu}\bar{x}, \bar{u}) + (F_{uu}\bar{u}, \bar{u})]\,dt + (\Phi''\bar{x}(T), \bar{x}(T)), \tag{2.5}$$

where $d\bar{x}/dt = A\bar{x} + B\bar{u}$, $\bar{x}(0) = 0$.

g. In connection with theorems 1.3, 1.4, 1.6, it is important to know in what cases $f'(u) = 0$. We call system (2.3) non-degenerate if $B^*\psi \not\equiv 0$ in any interval from $[0, T]$ for any non-zero solution of the system $d\psi/dt = -A^*\psi$. This condition means, in particular, that any point $x(T)$ can be reached from the initial point $x(0) = 0$. Let the system be non-degenerate and any of the following conditions be satisfied (for all admissible u): (a) $F \equiv 0, \Phi' \neq 0$; (b) $\Phi \equiv 0, F_x \not\equiv 0$ on any interval $F_u \equiv 0$; в) $\Phi \equiv 0, F_x \equiv 0, F_u \not\equiv 0$ on any interval. Then $f'(u) \neq 0$ for all admissible u.

We now consider the various types of constraint encountered in optimal control problems:

(1) $Q_1 = \{u : u(t) \in M_t$ for almost all $0 \leqslant t \leqslant T\}$; M_t is a set of E_r. For example, $Q_1 = \{u : a_i(t) \leqslant u_i(t) \leqslant b_i(t)\}$;

(2) $Q_2 = \left\{u : \int_0^T G(u(t), t)\,dt \leqslant \lambda\right\}$; the case $G(u,t) =$

or $G(u, t) = (Ru(t), u(t))_{E_r}$, is often encountered, where R is a positive definite matrix;

(3) $Q_3 = \{u : x(t) \in S_t$ for all $0 \leqslant t \leqslant T\}$, $S_t \subset E_m$; particular cases of this constraint are $x(T) = d$ and $q(x(t)) \leqslant 0$ for all $0 \leqslant t \leqslant T$, $q(x)$ is a function of m variables.

Constraints of a rather more general kind are also possible (for instance, $\int_0^T G(x, u, t)\,dt \leqslant \alpha$ or $\{x, u\} \in N_t$ for all $0 \leqslant t \leqslant T$, $N_t \subset E_{m+r}$). We shall not consider them, since it would add to the complexity of the analysis without introducing anything new.

We quote the properties of the sets $Q_1 - Q_3$.

a. If M_t is convex for all t, then Q_1 is convex. If $G(u, t)$ is convex with respect to u for all t, then Q_2 is convex. If S_t is convex for all t, and the system linear, then Q_3 is convex.

b. If M_t is closed for almost all t, then Q_1 is closed. If $G(u, t)$ is continuous with respect to u and $|G(u,t)| \geqslant \alpha\|u\|^2_{E_r} + \beta$ for all t, then Q_2 is closed. If S_t is closed for all t, and the system linear, then Q_3 is closed.

c. If M_t is uniformly bounded in E_r, then Q_1 is bounded. If $G(u,t) \geqslant \gamma\|u\|^2_{E_r}$ for large $\|u\|_{E_r}$ and all t, then Q_2 is bounded.

d. If $M_t \not\equiv E_r$, then Q_1 is certainly not uniformly convex (and does not even contain interior points). If $G(u, t)$ is uniformly (strongly) convex with respect to u for all t (with a function $\delta(\tau)$ independent of t), then Q_2 is uniformly (strongly) convex.

e. The set Q_3 can be specified by means of the functional $Q_3 = \{u : g(u) \leqslant 0\}$. We prove this in three particular cases. Let $Q' = \{u : q_1(x(t)) \leqslant 0$ for all $0 \leqslant i \leqslant T\}$. We can now take $g_1(u) = \max\limits_{0 \leqslant t \leqslant T} q_1(x(T))$.

If $Q'' = \{u : q_2(x(T)) \leqslant 0\}$, we can take $g_2(u) = q_2(x(T))$. Finally if $Q''' = \{u : x(T) = d\}$, then $g_3(u) = \|x(T) - d\|^2_{E_m}$.

We give the conditions under which the constraints $g_i(u) \leqslant 0$, $i = 1$, 2, 3, are correct. Let system (2.3) be non-degenerate. Then $g_3(u)$ is a correct constraint, while $g_1(u)$, $g_2(u)$ are correct provided the $q_i(x)$ are convex and $0 < \varepsilon \leqslant \|q_i'\| \leqslant c$ for $q_i = 0$, $i = 1, 2$, and in addition $q_1(x(0)) < 0$. In particular, the linear conditions $(a(t), x(t))^E \leqslant \lambda(t)$, $0 \leqslant t \leqslant T$, and $(a, x(T)) \leqslant \lambda$ lead to correct constraints.

3. In addition to finite-dimensional and optimal control problems, there is a vast number of extremal problems connected with partial differential equations. We shall not discuss these, since they cannot be stated in as general a way as problems with ordinary differential equations. The methods considered below can nevertheless be readily applied to many concrete problems of this kind (in particular, to those discussed in [11, 44].

3. Classification of the methods

In the general form our problem is to minimize a real function $f(x)$ on some set Q of real Banach space E. Most of the methods for solving this problem seem to fall readily into the following groups.

1. Methods of feasible directions

In these methods we obtain a minimizing sequence of points $x^0, \ldots, x^n, \ldots,$ all of which belong to Q. The functional and constraints are approximated at each point, and the new point obtained by solving an auxiliary problem. This group includes the gradient projection, conditional gradient and Newton methods discussed in Section 5 - 7. There are other methods of this type, which are described for the particular case of mathematical programming in [7].

2. Methods of set approximation

In these the set Q is approximated by a sequence of sets Q_n, for each of which the problem of minimizing $f(x)$ on Q_n is solved. Generally speaking, it is not essential for $Q_n \subset Q$. Examples are the Ritz and cut-off etc. methods, discussed in Sections 8 - 10.

3. Methods of penalty functions

These methods amount in essence to reducing the constrained to an unconstrained extremal problem, by imposing a "penalty" on the initial functional for infringing the constraints. These methods are dealt with in Section 11.

4. Duality methods

We include in this group the methods in which an iterative process is found for selecting the linear functionals ("dual" variables), figuring in the necessary conditions for an extremum. We shall omit these methods here, since it has not yet proved possible to state them in a general enough form. Several methods of this type are described, as applied to finite-dimensional problems, in [12 - 14].

This classification is not conventional and naturally has defects. In particular, some of the methods used for concrete problems need stretching to fit into one particular category. Also, a method can sometimes be classified differently according to the point of view.

Other criteria can be used for classification. For instance, we can group methods according to the highest order of functional derivative employed; e.g. the gradient projection method is of the first order and Newton's of the second. Further, we can call a method k-step if k previous iterations are used to obtain the next. Most of the methods discussed here are one-step (Sections 4 - 7) or null-step (methods of penalty functions (Section 11) or Ritz methods (Section 9), while in the cut-off method all the previous iterations are used to obtain the next. We can also classify methods as stationary or non-stationary, according to whether the method of obtaining the n-th iteration depends on n or not. Finally, we only consider discrete methods (i.e. those in which an iterational sequence x^n is formed), though many have continuous analogues (i.e. we obtain a trajectory $x(t)$, described by a differential equation), The point is that, for the numerical realization of continuous methods, we find ourselves using finite-difference methods to solve the differential equations, i.e. in essence we move over to a discrete method. However, continuous methods can be used e.g. for solving problems on analogue computers.

4. Methods of finding an unconstrained extremum

It seems worth discussing the case of an unconstrained extremum separately, since, firstly, rather more precise theorems can be proved here, and secondly, it is worth stating explicitly the methods whose analogues are used below for finding constrained extrema.

We shall only be concerned with one-step discrete methods, i.e. those in which an iterational sequence of the type

$$x^{n+1} = x^n - \alpha_n P_n(x^n), \qquad \alpha_n \geqslant 0, \tag{4.1}$$

is constructed.

We start with methods of the gradient type. We mean by this phrase that the direction of motion $P_n(x^n)$ is in some sense close to the gradient (see conditions (4.3), (4.4) in Theorem 4.1). Theorems 4.1 - 4.2 are an extension of the results of [15], where a gradient method in the strict sense of the word was investigated in Hilbert, space, i.e.

$$x^{n+1} = x^n - \alpha_n f'(x^n), \qquad \alpha_n \geqslant 0. \tag{4.2}$$

In finite-dimensional space the idea of this method goes back as far as Cauchy.

We shall not give the proofs here, since they follow the same lines as in [15].

Theorem 4.1

Let $f(x)$ be bounded from below on E: $\inf_{x \in E} f(x) = f^* > -\infty$, and differentiable, with $f'(x)$ satisfying a Lipschitz condition with constant M, and let

$$\|P_n(x)\| \leqslant K_1 \|f'(x)\|, \tag{4.3}$$

$$(f'(x), P_n(x)) \geqslant K_2 \|f'(x)\|^2, \qquad K_2 > 0, \tag{4.4}$$

$$0 < \varepsilon_1 \leqslant \alpha_n \leqslant \frac{2K_2}{MK_1} - \varepsilon_2, \qquad \varepsilon_2 > 0. \tag{4.5}$$

for all n. Then, whatever the x^0 in the method (4.1), we have $f(x^{n+1}) \leqslant f(x^n)$, $\lim_{n\to\infty} f'(x^n) = 0$. If, in addition, $f(x)$ is convex and $\{x: f(x) \leqslant f(x^0)\}$ bounded, then $\lim_{n\to\infty} f(x^n) = f^*$.

In particular, if E is a Hilbert space and $P_n(x) = f'(x)$, then conditions (4.3), (4.4) are satisfied for $K_1 = K_2 = 1$, so that Theorem 4.1 gives the convergence conditions for the gradient method (4.2).

To ensure the existence of a minimum and convergence of the sequence x^n, extra restrictions must be imposed on $f(x)$.

Theorem 4.2

In addition to the conditions of Theorem 4.1, let

$$\|f'(x)\|^2 \geqslant \lambda[f(x) - f^*], \quad \lambda > 0. \tag{4.6}$$

Then the sequence (4.1) is convergent to the minimum at the rate of a geometric progression.

The conditions of the theorem only need to be satisfied in a neighbourhood of x^*. Notice also that condition (4.6) does not demand the convexity of $f(x)$ and does not guarantee uniqueness of the minimum. In essence, it merely means that, if the gradient is small at some point, then $f(x)$ is close to its minimum at this point. This condition cannot be extended to the case of a constrained extremum.

If we replace (4.6) by the stricter condition for strong convexity and confine ourselves to method (4.2), a better estimate of the convergence rate can be obtained.

Theorem 4.3

Let E be a Hilbert space, and $f(x)$ twice differentiable, where, for all x, y,

$$m(y, y) \leqslant (f''(x)y, y) \leqslant M(y, y), \qquad m > 0. \tag{4.7}$$

Then, with $\alpha_n \equiv \alpha$, $0 < \alpha < 2/M$ in method (4.2) $\|x^n - x^*\| \leqslant \|x^0 - x^*\| q^n$, $q = \max\{|1 - \alpha M|, |1 - \alpha m|\}$, q is minimal and equal to $(M - m)/(M + m)$ for $\alpha = 2/(M + m)$.

As above, the conditions of the theorem only need to be satisfied in a neighbourhood of x^*.

The gradient method is often suitable in problems where at first sight there appear to be restrictive equations. Let E_1, E_2, E_3 be Hilbert spaces, $x \in E_1$, $y \in E_2$; our problem amounts to minimizing $f(x, y)$ on condition that $P(x, y) = 0$, P is an operator from $E_1 \times E_2$ into E_3. We shall assume that, given any fixed x, the equation $P(x, y) = 0$ has a unique solution $y(x)$. Now, instead of solving the initial conditioned extremum problem, we can seek the unconstrained minimum of the function $\Phi(x) = f(x, y(x))$. There is no need at all to find the function $y(x)$ explicitly here. In fact, the gradient of the functional $\Phi(x)$ is $\Phi'(x) = f_x - [P_y']^{-1} P_x$. Hence, the gradient method has the form

$$x^{n+1} = x^n - \alpha_n [f_x(x^n, y^n) - (P_y'(x^n, y^n))^{-1} P_x(x^n, y^n)], \quad y^{n+1} = y(x^{n+1}). \tag{4.8}$$

In other words, we only need to be able to find $y(x^{n+1})$. This approach is well known, see e.g. [16]. In essence, we use it when, in the optimal control problem (1.1), (1.2) we take only the one independent variable $u(t)$, instead of the two $x(t)$, $u(t)$ connected by equation (1.2).

Another important class of minimization methods includes those of Newton's type. Newton's method for minimizing $f(x)$ is

$$x^{n+1} = x^n - [f''(x^n)]^{-1} f'(x^n). \tag{4.9}$$

This method can be interpreted in two ways. Firstly, as the Newton's method for solving the equation $f'(x) = 0$. Hence all the conditions for convergence of method (4.9) (in particular, Theorem 4.4) can be obtained from the familiar [2, 17] convergence conditions for Newton's method for solving equations. Secondly, if we approximate $f(x)$ by a quadratic functional at the point x^n $f(x^n) + (f'(x^n),\ x - x^n) + {}^1\!/_2(f''(x^n)(x - x^n),\ x - x^n)$, its minimum point will be the same as x^{n+1} in (4.9). We shall use this interpretation when extending Newton's method to the case of constrained problems.

Theorem 4.4

Let condition (4.7) be satisfied, and in addition, let $f''(x)$ satisfy a Lipschitz condition with constant R. Then the sequence x^n in method (4.9) is convergent to the minimum x^* at the rate

$$\|x^n - x^*\| \leqslant \frac{m}{2R} \sum_{k=n}^{\infty} \delta^{2^k}, \text{ where } \delta = \frac{2R}{m} \|x^1 - x^0\|$$

(assuming that $\delta < 1$).

There are methods intermediate between the gradient and Newton's methods, when part of the dependences are approximated linearly, and part quadratically. For instance, when minimizing $\varphi(P(x))$, where $x \in E_1$, is a non-linear operator from E_1 to E_2, $\varphi(y)$ is a functional in E_2, we can approximate $\varphi(y)$ quadratically, and $P(x)$ linearly. We then obtain the method [18]

$$x^{n+1} = x^n - [P'^*(x^n)\varphi''(P(x^n))P'(x^n)]^{-1}P'^*(x^n)\varphi'(P(x^n)). \tag{4.10}$$

In the above problem of minimizing $f(x,\ y)$ under the condition $P(x,y) = 0$, we can linearize $P(x,\ y)$ and approximate $f(x,\ y)$ quadratically. We then get the method

$$x^{n+1} = x^n - [f_{xx} + f_{yy}(P_y{}^{-1}P_x)^*P_y{}^{-1}P_x - f_{xy}P_y{}^{-1}P_x]^{-1}(f_yP_y{}^{-1}f_x + f_x). \tag{4.11}$$

But the convergence conditions for (4.11) are unknown.

We now illustrate the concrete form of the above methods for different examples. The convergence conditions for each follow from the above theorems and the properties of functionals given in Section 2.

1. When minimizing the finite-dimensional function $f(x) = f(x_1, \ldots, x_n)$ the gradient method (4.2) becomes

$$x_i^{n+1} = x_i{}^n - \alpha_n \frac{\partial f(x^n)}{\partial x_i}, \qquad i = 1, \ldots, m. \tag{4.12}$$

The more general methods

$$x_i^{n+1} = x_i^n - \alpha_n^i \frac{\partial f(x^n)}{\partial x_i}, \qquad i = 1, \ldots, m, \tag{4.13}$$

can be regarded as realizations of method (4.1). In Newton's method (4.9), $f''(x^n)$ is the matrix of second derivatives; at each step of this method, therefore, we have to find this matrix and solve a system of linear algebraic equations.

2. In the case of the optimal control problem (2.1), (2.2), the gradient method (4.2) becomes, in accordance with the expression (2.4) for the gradient

$$\begin{gathered} u^{n+1}(t) = u^n(t) - \alpha_n [F_u - \varphi_u{}^*\psi], \\ \frac{d\psi}{dt} = -\varphi_x{}^*\psi + F_x, \qquad \psi(T) = -\Phi'(x(T)), \end{gathered} \tag{4.14}$$

where all the quantities occurring here are evaluated on the trajectory corresponding to $u^n(t)$.

A more general method of the type (4.1) is obtained if the constants $\alpha_n \geqslant 0$ are replaced by the functions $\alpha_n(t) \geqslant 0$ in (4.14).

The method (4.14) seems to have been first applied by Stein [19] to the elementary variational problem; see also [15]. It was proposed for the general optimal control problem in [20 - 23]. Elementary theorems on convergence of this method are only to be found in [23]. It follows from Theorem 4.1 of the present paper that, if the conditions given in Section 2 are satisfied, in which the gradient of $f(u)$ exists and satisfies a Lipschitz condition, while $F(x, u, t) \geqslant \lambda_1$ and $\Phi(x) \geqslant \lambda_2$, then we have $f(u^{n+1}) \leqslant f(u^n)$ and $f'(u^n) \to 0$. in method (4.14). Theorem 4.3 can be applied for the linear problem (2.1), (2.3), and a result obtained on the convergence of u^n to the solution.

We now consider Newton's method for the linear problem (2.1), (2.3). In accordance with expression (2.5) for $f''(u)$, at each step of Newton's method we have to minimize the quadratic functional

$$\begin{aligned} f(\bar{u}) = \int_0^T \Big[& (F_x, \bar{x}) + (F_u, \bar{u}) + \frac{1}{2}(F_{xx}\bar{x}, \bar{x}) + (F_{xu}\bar{x}, \bar{u}) + \\ & + \frac{1}{2}(F_{uu}\bar{u}, \bar{u}) \Big] dt + (\Phi', \bar{x}(T)) + \frac{1}{2}(\Phi''\bar{x}(T), \bar{x}(T)), \end{aligned} \tag{4.15}$$

where $d\bar{x} / dt = Ax + B\bar{u}, \bar{x}(0) = 0$.

Accordingly, Newton's method becomes

$$u^{n+1}(t) = u^n(t) + F_{uu}^{-1}(F_u - 2F_{xu}\bar{x} - B^*\psi), \tag{4.16}$$

where $\bar{x}(t)$, $\psi(t)$ is the solution of the linear boundary-value problem

$$\frac{d\bar{x}}{dt} = A\bar{x} - BF_{uu}^{-1}(F_u - 2F_{xu}\bar{x} - B^*\psi),$$
$$\frac{d\psi}{dt} = -A^*\psi + F_x + F_{xx}\bar{x} + F_{xu}F_{uu}^{-1}(F_u - 2F_{xu}\bar{x} - B^*\psi),$$
$$\bar{x}(0) = 0, \qquad \psi(T) = -\Phi''\bar{x}(T).$$

Finally, the mixed method (4.11) for problem (1.1), (1.2) has the same form as (4.16), if A and B in (4.16) are replaced by φ_x and φ_u respectively.

Newton's method was proposed for the elementary variational problem in [19]; we are not aware of any previous application of it in the literature to the optimal control problem (method (4.16)).

5. Gradient projection method

We turn to methods for solving constrained problems.

The gradient projection method, for minimizing the differentiable functional $f(x)$ on a set Q of Hilbert space E, amounts to forming the sequence

$$x^{n+1} = P_Q(x^n - \alpha_n f'(x^n)), \tag{5.1}$$

where P_Q is the operator of projection on to Q (in other words, $y = P_Q(x)$ is given by $y \in Q$, $\|y - x\| = \inf_{z \in Q} \|x - z\|$). If $Q = E$ (i.e. we have the unconstrained problem), method (5.1) is the same as the gradient method (4.2). A curious point is that the more general method (4.1) does not similarly extend to constrained problems. The necessary condition (1.2) for a minimum is satisfied at a stationary point of method (5.1).

A general theorem on the convergence of this method is

Theorem 5.1

Let Q be a bounded closed convex set of Hilbert space H, $f(x)$ a functional differentiable on Q, where $f'(x)$ satisfies a Lipschitz condition with constant M, and $0 < \varepsilon_1 \leqslant \alpha_n \leqslant 2/(M + 2\varepsilon_2)$, $\varepsilon_2 > 0$. Then sequence (5.1) has the following properties:

(1) $f(x^n)$ is monotonically decreasing and $\lim\limits_{n\to\infty} \|x^{n+1} - x^n\| = 0$;

(2) if $f(x)$ is convex, then

$$\lim_{n\to\infty} f(x^n) = f^* = \inf_{x\in Q} f(x),$$

where $f(x^n) - f^* \leqslant c/n$, and a subsequence of x^n exists, weakly convergent to the minimum x^*;

(3) if $f(x)$ is strictly convex or Q is strictly convex, while $f'(x) \neq 0$ on Q, then x^n is weakly convergent to the (unique) minimum x^*;

(4) if $f(x)$ is uniformly convex or Q is uniformly convex, while $f'(x) \neq 0$ on Q, then x^n is strongly convergent to x^*;

(5) if $f(x)$ is twice differentiable, where $m\|y\|^2 \leqslant (f''(x)y, y) \leqslant M\|y\|^2$, $m > 0$ for all $x \in Q$ and all y, then, with $\alpha_n \equiv \alpha$, $0 < \alpha < 2/M$, sequence (5.1) is convergent to x^* at the rate of a geometric progression: $\|x^n - x^*\| \leqslant C(x^0, \varepsilon)(q + \varepsilon)^n$, where $q = \max\{|1 - \alpha m|, |1 - \alpha M|\}$, $0 \leqslant q < 1$, $\varepsilon > 0$ is arbitrary. q is minimal and equal to $(M - m)/(M + m)$ for $\alpha = 2/(M + m)$.

Proof. We notice first that, since Q is convex and closed, the operator $P_Q(x)$ is uniquely defined for all x. Further,

$$f(x^{n+1}) - f(x^n) = (f'(x^n), x^{n+1} - x^n) + \int_0^1 (f'(x^n + \tau(x^{n+1} - x^n)) -$$

$$- f'(x^n), x^{n+1} - x^n)\,d\tau \leqslant (f'(x^n), x^{n+1} - x^n) + \int_0^1 \|f'(x^n + \tau(x^{n+1} - x^n)) -$$

$$- f'(x^n)\|\|x^{n+1} - x^n\|\,d\tau \leqslant (f'(x^n), x^{n+1} - x^n) + \frac{M}{2}\|x^{n+1} - x^n\|^2 =$$

$$= -\frac{1}{\alpha_n}(x^n - \alpha_n f'(x^n) - x^{n+1}, x^{n+1} - x^n) - \frac{1}{\alpha_n}\|x^{n+1} - x^n\|^2 +$$

$$+\frac{M}{2}\|x^{n+1}-x^n\|^2 \leqslant \left(-\frac{1}{\alpha_n}+\frac{M}{2}\right)\|x^{n+1}-x^n\|^2 \leqslant -\varepsilon_2\|x^{n+1}-x^n\|^2 \leqslant 0$$

(Here, we have used the fact that $(x - P_Q(x),\ y - P_Q(x)) \leqslant 0$ for all $y \in Q$. Thus $f(x^n)$ is monotonically decreasing. Since the boundedness of Q and the Lipschitz condition on $f'(x)$ imply that $f(x)$ is bounded from below on Q, we have $\lim\limits_{n\to\infty} f(x^n)$ exists, so that $\| x^{n+1} - x^n \|^2 \leqslant [f(x^n) - f(x^{n+1})] / \varepsilon_2 \to 0$ as $n \to \infty$.

Further, if $f(x)$ is convex, a minimum x^* must exist, $f(x^*) = f^*$. Since $f(x)$ is convex,

$$0 \leqslant f(x^n) - f^* \leqslant (f'(x^n), x^n - x^*) = (f'(x^n), x^n - x^{n+1}) +$$
$$+\frac{1}{\alpha_n}(x^n - \alpha_n f'(x^n) - x^{n+1}, x^* - x^{n+1}) - \frac{1}{\alpha_n}(x^n - x^{n+1}, x^* - x^{n+1}) \leqslant$$
$$\leqslant \|f'(x^n)\|\,\|x^{n+1}-x^n\| + \frac{1}{\alpha_n}\|x^{n+1}-x^n\|\,\|x^*-x^{n+1}\| \leqslant$$
$$\leqslant \left(\|f'(x^n)\| + \frac{1}{\varepsilon_1}\|x^* - x^{n+1}\|\right)\|x^{n+1}-x^n\|.$$

Since the expression in brackets is bounded, and $\| x^{n+1} - x^n \| \to 0$ by what has been proved, we have $f(x^n) \to f^*$ as $n \to \infty$. We now consider the rate of convergence. Let $\varphi_n = f(x^n) - f^*$. From the last inequality, $\varphi_n \leqslant \lambda \| x^n - x^{n+1} \|$, while we found above that $\varphi_n - \varphi_{n+1} \geqslant \varepsilon_2 \| x^n - x^{n+1} \|^2$. Hence $\varphi_n - \varphi_{n+1} \geqslant \rho\varphi_n^2$. We consider the numerical sequence $p_{n+1} = p_n - \rho p_n^2$. If p_0 is small, we know [24] that $p_n \leqslant c/n$. But, if $\varphi_n \leqslant p_n$ and φ_n is small, then $\varphi_{n+1} \leqslant \varphi_n - \rho\varphi_n^2 \leqslant p_{n+1}$. Since, by what has been proved, $\lim\limits_{n\to\infty}\varphi_n = 0$, then $\varphi_n \leqslant p_n \leqslant c/n$ for all reasonably large n. The last part of the second assertion of the theorem, and also assertions 3 and 4, follow from Theorems 1.4 - 1.6.

Finally

$$\| x^n - x^{n+1} \|^2 = (x^n - \alpha f'(x^n) - x^{n+1}, x^n - x^{n+1}) +$$
$$+ \alpha(f'(x^n), x^n - x^{n+1}) \leqslant \alpha(f'(x^n), x^n - x^{n+1}) =$$
$$= \alpha(f'(x^{n-1}) + A(x^n - x^{n-1}) + r, x^n - x^{n+1}).$$

Here, $A = f''(x^{n-1})$, $r = f'(x^n) - f'(x^{n-1}) - f''(x^{n-1})\ (x^n - x^{n-1})$, $\|r\| = o(\| x^n - x^{n-1} \|)$, I is the unit operator. Thus,

$$\begin{aligned}\| x^n - x^{n+1} \|^2 &\leqslant (x^{n-1} - \alpha f'(x^{n-1}) - x^n, x^{n+1} - x^n) + \\ &+ ((I - \alpha A)(x^n - x^{n-1}), x^{n+1} - x^n) + \alpha (r, x^n - x^{n+1}) \leqslant \\ &\leqslant (\| I - \alpha A \| \| x^n - x^{n-1} \| + \| r \|) \| x^{n+1} - x^n \| \leqslant \\ &\leqslant (\max_{m \leqslant \lambda \leqslant M} |1 - \alpha\lambda| \| x^n - x^{n-1} \| + \| r \|) \| x^{n+1} - x^n \| \leqslant \\ &\leqslant (q \| x^n - x^{n-1} \| + \| r \|) \| x^{n+1} - x^n \|.\end{aligned}$$

Hence

$$\|x^n - x^{n+1}\| \leqslant q\|x^n - x^{n-1}\| + \|r\| = (q + \delta_n)\|x^n - x^{n-1}\|,$$

where $\delta_n = \|r\| / \|x^n - x^{n-1}\| \to 0$ as $n \to \infty$ (since $\|x^n - x^{n-1}\| \to 0$), which proves the last assertion of the theorem.

Notes. 1. The assumption in (4) and (5) of Theorem 5.1 that Q is bounded is superfluous.

2. In (5), it is sufficient to require that the functional be strongly convex and uniquely differentiable. We have confined ourselves to the case of a twice differentiable functional since this simplifies the proof and enables a better estimate of the convergence rate to be given.

3. In the case of unconstrained minimization, method (5.1) is the same as (4.2), and (5) of Theorem 5.1 becomes similar to Theorem 4.3.

4. The estimate of the convergence rate in (2) cannot be improved.

5. While the present paper was in the press, [25] appeared, in which the same method is considered and similar results obtained; but no convergence rate estimate was given.

We now see what aspect the gradient projection method takes on in various concrete problems. We shall not write down its convergence conditions in each case, since they can be obtained from Theorem 5.1 and the properties of functionals and sets described in Section 2. Notice that many of the methods described below have been used in computational practice, though their convergence has not been proved.

1. The finite-dimensional problem min $f(x)$, $x = (x_1, \ldots, x_m)$, under the constraints $a_i \leqslant x_i \leqslant b_i$, $i = 1, \ldots, m$. Method (5.1) becomes

$$x_i^{n+1} = \begin{cases} a_i, & \text{if} \quad x_i^n - \alpha_n \dfrac{\partial f}{\partial x_i} \leqslant a_i; \\ x_i^n - \alpha_n \dfrac{\partial f}{\partial x_i}, & \text{if} \quad a_i < x_i^n - \alpha_n \dfrac{\partial f}{\partial x_i} < b_i; \\ b_i, & \text{if} \quad x_i^n - \alpha_n \dfrac{\partial f}{\partial x_i} \geqslant b_i. \end{cases} \tag{5.2}$$

2. The problem of non-linear programming with linear restrictions: $\min f(x)$, $x = (x_1, \ldots, x_m)$, under the constraints $Ax \leqslant b$, $b \in E_r$, A is an $m \times r$ matrix. Here, at each step of method (5.1), we have to solve the problem of quadratic programming

$$\min_{Ax \leqslant b} \| x^n - \alpha_n f'(x^n) - x \|^2.$$

Notice that Rosen's gradient projection method [26] differs from the present (e.g. the points x^n and x^{n+1} do not need to lie on one face of the polyhedron, as in Rosen's method).

3. The problem of $\min f(x)$ under the constraints $Ax = b$, where $x \in E_1$, $b \in E_2$, E_1, E_2 are Hilbert spaces, A is a non-degenerate (i.e. $AE_1 = E_2$) bounded linear operator from E_1 onto E_2. Method (5.1) now becomes

$$x^{n+1} = x^n - \alpha_n f'(x^n) - A^*(AA^*)^{-1}A(x^n - \alpha_n f'(x^n)). \tag{5.3}$$

Incidentally, (5.3) can also be regarded as a gradient method for an unconstrained extremum, if we regard the subspace $Ax = 0$ as an independent Hilbert space. If, in particular, the constraints are specified by means of a finite number of linear functionals $(c_i, x) = \beta_i$, $i = 1, \ldots, m$, where $c_1, \ldots, c_m$ are linearly independent, then method (5.3) becomes

$$x^{n+1} = x^n - \alpha_n \left[f'(x^n) + \sum_{i=1}^{m} \lambda_i c_i \right], \qquad (c_i, x^0) = \beta_i, \qquad i = 1, \ldots, m, \tag{5.4}$$

where the λ_i satisfy the system of linear algebraic equations

$$\sum_{i=1}^{m} \lambda_i (c_i, c_j) = -(c_j, f'(x^n)), \quad j = 1, \ldots, m.$$

In particular, in the linear optimal control problem (2.1), (2.3) with fixed values of $x(T) = d$ (we can naturally assume $\Phi \equiv 0$ here), the condition $x(T) = d$ is equivalent to specifying the values of m linear functionals of u. Hence method (5.1) becomes

$$u^{n+1}(t) = u^n(t) - \alpha_n(F_u - B^*\psi), \tag{5.5}$$

where $\psi(t)$ is the solution of the linear boundary-value problem

$$\frac{d\psi}{dt} = -A^*\psi + F_x, \qquad \frac{d\bar{x}}{dt} = A\bar{x} + B(F_u - B^*\psi), \qquad \bar{x}(0) = \bar{x}(T) = 0,$$

and the condition $x^0(T) = d$ has to be satisfied for $u^0(t)$.

4. For the problem min $f(x)$ under the constraint $g(x) \leqslant 0$, where f, g are convex differentiable functionals in Hilbert space, method (5.1) becomes

$$x^{n+1} = \begin{cases} x^n - \alpha_n f'(x^n), & \text{if} \quad g(x^n - \alpha_n f'(x^n)) \leqslant 0; \\ \bar{x}, & \text{if} \quad g(x^n - \alpha_n f'(x^n)) > 0; \end{cases} \tag{5.6}$$

where $\bar{x}$ is found from the conditions

$$x^n - \alpha f'(x^n) - \bar{x} = \lambda g'(\bar{x}), \qquad \lambda \geqslant 0, \qquad g(\bar{x}) = 0.$$

For instance, if the restriction $g(x) \leqslant 0$ is linear, $(c,\ x) \leqslant \beta$, then $\bar{x} = x^n - \alpha^n f'(x^n) - \lambda c$, where λ is obtained from the condition $(c,\ \bar{x}) = \beta$. If the constraint $g(x) \leqslant 0$ has the form $\|x\| \leqslant \beta$, then $\bar{x} = \beta[x^n - \alpha_n f'(x^n)] : \|x^n - \alpha_n f'(x^n)\|$. Method (5.6) can be used for the optimal control problem in which one of the following constraints is present:

$$(c, x(T)) \leqslant \beta, \qquad \int_0^T (c(t), u(t))\,dt \leqslant \beta, \qquad \int_0^T \|u(t)\|^2_{E_r}\,dt \leqslant \beta,$$

or, more generally,

$$g(x(T)) \leqslant 0, \qquad \int_0^T g(u(t))\,dt \leqslant 0.$$

In the case of several differentiable constraints $g_i(x) \leqslant 0$ we have to solve, at each step of method (5.1), a finite-dimensional problem of mathematical programming (quadratic programming if the $g_i(x)$ are linear).

5. The optimal control problem (2.1), (2.2) under the constraint $u(t) \in M$, for almost all $0 \leqslant t \leqslant T$, where M_t is a convex closed set of E^r for all t, can be solved by method (5.1) in the form

$$u^{n+1}(t) = u^n(t) - P_{M_t}(u^n(t) - \alpha_n f'(u^n)), \tag{5.7}$$

where $f'(u)$ is given by (2.4), while P_{M_t} is the operator of projection onto M_t in E_r (for each t). In particular, if the constraints are $a_i(t) \leqslant u_i(t) \leqslant b_i(t)$, $i = 1, \ldots, r$, then

$$[P_{M_t}(u(t))]_i = \begin{cases} a_i(t), & \text{if} \quad u_i(t) \leqslant a_i(t); \\ u_i(t), & \text{if} \quad a_i(t) < u_i(t) < b_i(t); \\ b_i(t), & \text{if} \quad u_i(t) \geqslant b_i(t). \end{cases} \tag{5.8}$$

6. Let Q_1, Q_2 be closed convex sets in Hilbert space, Q_1 being bounded. Then there exist $x^* \in Q_1$, $y^* \in Q_2$, such that

$$\| x^* - y^* \| = \rho(Q_1, Q_2) = \inf_{x \in Q_1,\, y \in Q_2} \| x - y \|,$$

i.e. $x^* = P_{Q_1}(y^*)$, $y^* = P_{Q_2}(x^*)$. If Q_1 or Q_2 is strictly convex, while $Q_1 \cap Q_2 = \varphi$, then x^*, y^* are unique. The problem of finding x^*, y^* is equivalent to the problem of minimizing

$$f(x) = \rho^2(x, Q_2) = \| x - P_{Q_2}(x) \|^2 = \inf_{y \in Q_2} \| x - y \|^2 \text{ на } Q_1.$$

We consider the gradient projection method for this last problem. It can be shown that $f(x)$ is differentiable, $f'(x) = 2(x - P_{Q_2}(x))$, $f'(x)$ satisfies a Lipschitz condition with constant 2. Applying method 5.1 with $\alpha_n \equiv ½$, we get

$$x^{n+1} = P_{Q_1}(P_{Q_2}(x^n)). \tag{5.9}$$

Thus the method amounts to successive projection onto Q_2 and Q_1. Since $f(x)$ is convex, we find from (2) of Theorem 5.1 that, for (5.9),

$$\lim_{n \to \infty} \| x^n - P_{Q_2}(x^n) \| = \| x^* - y^* \|.$$

If Q_1 or Q_2 is strictly convex, while $Q_1 \cap Q_2 = \varphi$, we can apply (3) of Theorem 5.1, so that $x^n \xrightarrow{st} x^*$, $P_{Q_2}(x^n) \xrightarrow{st} y^*$. Finally, if Q_1 or Q_2 is uniformly convex, and $Q_1 \cap Q_2 = \varphi$, the strong convergence of method (5.9) will follow from (4) of Theorem 5.1.

6. Conditional gradient method

The essence of this method for minimizing a functional $f(x)$ on a set Q is as follows: we linearize $f(x)$ at the point x^n at each step, solve the auxiliary linear problem of minimizing $(f'(x^n), x)$ on Q, and determine the minimum $\bar{x}^n$ in this problem from the direction of the movement to obtain the next approximation. Thus,

$$(f'(x^n), \bar{x}^n) \leqslant (f'(x^n), x) \quad \text{for all} \quad x \in Q,\ \bar{x}^n \in Q, \tag{6.1}$$

$$x^{n+1} = x^n + \alpha_n(\bar{x}^n - x^n),\ 0 \leqslant \alpha_n \leqslant 1. \tag{6.2}$$

Theorem 6.1

Let Q be a bounded closed convex set in reflexive space E, $f(x)$ be differentiable on Q, where $f'(x)$ satisfies a Lipschitz condition with constant M, and let $\alpha_n = \min\{1, \gamma_n(f'(x^n), x^n - \bar{x}^n) / \| x^n - \bar{x}^n \|^2\}$, where $\bar{x}^n$ is defined in (6.1), and $0 < \varepsilon_1 \leqslant \gamma_n \leqslant (2 - \varepsilon_2)/M$, $\varepsilon_2 > 0$. Then the sequence (6.2) has the following properties:

(1) $f(x^n)$ is monotonically decreasing and $\lim\limits_{n\to\infty} (f'(x^n), x^n - \bar{x}^n) = 0$;

(2) if $f(x)$ is convex, then $\lim\limits_{n\to\infty} f(x^n) = f^* = \inf\limits_{x\in Q} f(x)$, where $f(x^n) - f^* \leqslant c/n$, and a subsequence x^n exists, weakly convergent to the minimum x^*;

(3) if $f(x)$ is strictly convex or Q strictly convex, while $f'(x) \neq 0$ on Q, then x^n is weakly convergent to the (unique) minimum x^*;

(4) if $f(x)$ is uniformly convex or Q uniformly convex, while $f'(x) \neq 0$ on Q, then x^n is strongly convergent to x^*;

(5) if $f(x)$ is convex and $\|f'(x)\| \geqslant \varepsilon > 0$ on Q, while Q is strongly convex, then x^n is convergent to x^* at the rate of a geometric progression.

Proof. We observe first that, by Theorems 1.2, 1.3, the point $\bar{x}^n$ in (6.1) exists (and is unique, if Q is strictly convex).

Just as in the proof of Theorem 5.1, we get

$$f(x^{n+1}) - f(x^n) \leqslant (f'(x^n), x^{n+1} - x^n) + \frac{M}{2}\|x^{n+1} - x^n\|^2 =$$

$$= -\alpha_n(f'(x^n), x^n - \bar{x}^n) + \frac{M\alpha_n^2}{2}\|x^n - \bar{x}^n\|^2.$$

If $1 \leqslant \gamma_n(f'(x^n), x^n - \bar{x}^n) / \|x^n - \bar{x}^n\|^2$, then $\alpha_n = 1$ and

$$f(x^{n+1}) - f(x^n) \leqslant -(f'(x^n), x^n - \bar{x}^n) + \frac{M}{2}\|x^n - \bar{x}^n\|^2 =$$

$$= (f'(x^n), x^n - \bar{x}^n)\left(\frac{M}{2}\frac{\|x^n - \bar{x}^n\|^2}{(f'(x^n), x^n - \bar{x}^n)} - 1\right) \leqslant$$
$$\leqslant (f'(x^n), x^n - \bar{x}^n)\left(\frac{M\gamma_n}{2} - 1\right) \leqslant -\frac{\varepsilon_2}{2}(f'(x^n), x^n - \bar{x}^n).$$

If $1 \geqslant \gamma_n (f'(x^n), x^n - \bar{x}^n) / \|x^n - \bar{x}^n\|^2$, then

$$\alpha_n = \gamma_n \frac{(f'(x^n), x^n - \bar{x}^n)}{\|x^n - \bar{x}^n\|^2}, \qquad f(x^{n+1}) - f(x^n) \leqslant -\frac{(f'(x^n), x^n - \bar{x}^n)^2}{\|x^n - \bar{x}^n\|^2} +$$
$$+\frac{M\gamma_n^2}{2}\frac{(f'(x^n), x^n - \bar{x}^n)^2}{\|x^n - \bar{x}^n\|^2} \leqslant -\frac{\varepsilon_1}{2}\frac{(f'(x^n), x^n - \bar{x}^n)^2}{\|x^n - \bar{x}^n\|^2} \leqslant 0.$$

In both cases, $f(x^{n+1}) \leqslant f(x^n)$, i.e. $f(x^n)$ is monotonically decreasing. Since $f(x)$ is bounded from below on Q (by virtue of the Lipschitz condition on $f'(x)$ and the boundedness of Q), there exists $\lim_{n\to\infty} f(x^n)$, i.e. $\delta_n = f(x^n) - f(x^{n+1}) \to 0$ as $n \to \infty$. But $(f'(x^n), x^n - \bar{x}^n) \leqslant 2\delta_n/\varepsilon_2$ in the first case and $(f'(x^n), x^n - \bar{x}^n) \leqslant (2\delta_n R^2/\varepsilon_2)^{1/2}$ in the second $(R = \sup_{x,y\in Q}\|x - y\|)$. Hence, in both cases, $\lim_{n\to\infty} (f'(x^n), x^n - \bar{x}^n) = 0$. Further, if $f(x)$ is convex, there exists x^*, $f(x^*) = f^*$ and $0 \leqslant f(x^n) - f^* \leqslant (f'(x^n), x^n - x^*) = (f'(x^n, x^n - \bar{x}^n) + (f'(x^n), \bar{x}^n - x^*) \leqslant (f'(x^n), x^n - \bar{x}^n)$. It now follows that $\lim_{n\to\infty} (f(x^n) - f^*) = 0$. We estimate the rate of convergence. Let $f(x^n) - f^* = \varphi_n$, when $\delta_n = \varphi_n - \varphi_{n+1}$. As was shown above, $(f'(x^n), x^n - \bar{x}^n) \leqslant \max\{2(\varphi_n - \varphi_{n+1})/\varepsilon_2, K(\varphi_n - \varphi_{n+1})^{1/2}\} \leqslant K(\varphi_n - \varphi_{n+1})^{1/2}$ for large n. On the other hand (see above), $\varphi_n \leqslant (f'(x^n), x^n - \bar{x}^n)$. Combining these inequalities, we get $\varphi_n^2 \leqslant K^2(\varphi_n - \varphi_{n+1})$, i.e. $\varphi_{n+1} \leqslant \varphi_n - \varphi_n^2/K^2$. Hence we obtain, as in the proof of Theorem 5.1, $\varphi_n \leqslant c/n$. The final assertion of (2) of Theorem 6.1, and also (3) and (4), follow from the general Theorems 1.4 - 1.6 on minimizing sequences.

We turn to the proof of (5). Using the strong convexity of Q, we get

$$-(f'(x^n), x^n - \bar{x}^n) = 2\left(f'(x^n), \frac{x^n + \bar{x}^n}{2} - \lambda\frac{f'(x^n)}{\|f'(x^n)\|}\|\bar{x}^n - x^n\|^2\right) +$$
$$+ 2\left(f'(x^n), \lambda\frac{f'(x^n)}{\|f'(x^n)\|}\|x^n - \bar{x}^n\|^2 - x^n\right) \geqslant 2(f'(x^n), \bar{x}^n) +$$
$$+ 2\lambda\|f'(x^n)\|\,\|\bar{x}^n - x^n\|^2 - 2(f'(x^n), x^n) \geqslant 2(f'(x^n), \bar{x}^n - x^n) +$$
$$+ 2\lambda\varepsilon\|x^n - \bar{x}^n\|^2.$$

Hence $(f'(x^n), x^n - \bar{x}^n) \geqslant 2\lambda\varepsilon \| \bar{x}^n - x^n \|^2$.

Let $\alpha_n = 1$. By what has been proved, $\varphi_n - \varphi_{n+1} \geqslant (\varepsilon_2/2)(f'(x^n), x^n - \bar{x}^n) \geqslant (\varepsilon_2/2)\varphi_n$, i.e. $\varphi_{n+1} \leqslant (1 - \varepsilon_2/2)\varphi_n$.

If $\alpha_n < 1$, then

$$\varphi_n - \varphi_{n+1} \geqslant \frac{\varepsilon_2}{2} \frac{(f'(x^n), x^n - \bar{x}^n)}{\| x^n - \bar{x}^n \|^2} \geqslant \lambda\varepsilon\varepsilon_2 (f'(x^n), x^n - \bar{x}^n) \geqslant \lambda\varepsilon\varepsilon_2\varphi_n,$$

i.e. $\varphi_{n+1} \leqslant (1 - \lambda\varepsilon\varepsilon_2)\varphi_n$.

In both cases, therefore, $\varphi_{n+1} \leqslant q\varphi_n, q = \max\{(1 - \varepsilon_2/2), (1 - \lambda\varepsilon\varepsilon_2)\} <$. Hence, $\varphi_n \leqslant \varphi_0 q^n$. But (Theorem 1.6) $\varphi_n \geqslant \gamma \| x^n - x^* \|^2$, so that $\|x^n - x^*\| \leqslant (\varphi_0/\gamma)^{1/2} q^{n/2}$, which completes the proof.

Notes. 1. The asymptotic formulae $\varphi_n = O(1/n)$ in (2) of Theorem 6.1 cannot be improved without extra assumptions about Q (even if we demand strong convexity of the functional).

2. Method (6.2) was discussed in the general form in [27], but none of the above convergence rate estimates were given. In addition, the coefficient α_n is defined in [27] as in the method of steepest descent, and not as in Theorem 6.1; some unnecessary restrictions are also imposed on $f(x)$ and Q.

3. A whole group of methods, intermediate between the gradient projection and the conditional gradient methods, is considered in [28] (without proof of convergence).

We now examine the aspect of the conditional gradient method for various concrete problems, without dwelling on the convergence conditions, which are readily obtained from the above results.

1. The problem of non-linear programming with linear constraints: $\min f(x)$, $Ax \leqslant b$, $x \in E_m$, $b \in E_r$, where A is an $m \times r$ matrix. To determine $\bar{x}^n$ at each step we have to solve the auxiliary problem of linear programming: $\min (f'(x^n), x)$, $Ax \leqslant b$. Method (6.2) was proposed in this form in [29], and statements (1) and (2) of Theorem 6.1 proved for this case.

2. For the restriction $\| x \| \leqslant \rho$ in Hilbert space $\| \bar{x}^n \| = -\rho f'(x^n) / \| f'(x^n) \|$, so that method (6.2) becomes

$$x^{n+1} = x^n - \alpha_n \left(\rho \frac{f'(x^n)}{\| f'(x^n) \|} + x^n \right). \tag{6.3}$$

It is easily shown that, as $\rho \to \infty$, the method approximates to the ordinary gradient method (4.2) for an unconstrained extremum. The method can be applied in this form both in the finite-dimensional case (constraint $\left(\left(\sum_{i=1}^{m} x_i^2 \right)^{1/2} \leqslant \rho \right)$, and in optimal control problems (constraint

$$\left(\left(\int_0^T \sum_{i=1}^{r} u_i^2(t)\, dt \right)^{1/2} \leqslant \rho. \right)$$

3. For optimal control problems (2.1), (2.2) with constraints of the type $u(t) \in M_t$ for almost all $0 \leqslant t \leqslant T$, M_t is a bounded closed convex set of E^r for all t, method (6.2) becomes

$$u^{n+1}(t) = u^n(t) + \alpha_n (\bar{u}^n(t) - u^n(t)), \tag{6.4}$$

where $\bar{u}^n(t)$ at each instant t realizes on M_t a minimum with respect to u of the linear function $(h(t), u(t))$, where $h(t)$ is given by (2.4). The method was considered in [30, 31] for problems of this type.

Notice that, when system (2.2) and $F(x, u, t)$ are linear in u, the method is the same as the method of [32] (see also [22], Section 6.6), so that a proof of the convergence of the latter method can be obtained in this case.

7. Newton's method

The idea of this method lies in approximating the functional $f(x)$ to be minimized at each step by a quadratic functional (first terms of a Taylor series) and taking the minimum point of this functional on Q as the next approximation. Thus, x^{n+1} is given by the condition

$$\begin{gathered} x^{n+1} \in Q, \qquad f_n(x^{n+1}) \leqslant f_n(x) \qquad \text{for all} \quad x \in Q, \\ f_n(x) = (f'(x^n), x - x^n) + 1/2(f''(x^n)(x - x^n), x - x^n). \end{gathered} \tag{7.1}$$

The modified Newton method [2] can be extended in exactly the same way to the case of constraints, but we shall not dwell on this.

Theorem 7.1

Let Q be a closed convex set in Hilbert space E, $f(x)$ a functional twice differentiable on Q, where $f''(x)$ satisfies the conditions

$$m \| y \|^2 \leqslant (f''(x)\, y, y) \leqslant M \| y \|^2, \qquad m > 0, \tag{7.2}$$

$$\| f''(x) - f''(z) \| \leqslant R \| x - z \| \tag{7.3}$$

for all x, $z \in Q$ and all y. Further, let $\delta = (2R/m) \| x^1 - x^0 \| < 1$, where x^1 is defined in terms of x^0 from (7.1). Then the sequence (7.1) is convergent to the minimum x^*, where

$$\| x^n - x^* \| \leqslant \frac{m}{2R} \sum_{k=n}^{\infty} \delta^{2^k}.$$

Proof. Since $f_n(x)$ has a minimum on Q at the point x^{n+1}, $(f_n'(x^{n+1}),$ x will also have a minimum on Q at this point, see (1.1). Hence

$$(f_n'(x^{n+1}), x^{n+1} - x^n) = (f'(x^n) + f''(x^n)(x^{n+1} - x^n), x^{n+1} - x^n) \leqslant 0.$$

Consequently,

$$f_n(x^{n+1}) = (f'(x^n), x^{n+1} - x^n) + \frac{1}{2}(f''(x^n)(x^{n+1} - x^n), x^{n+1} - x^n) \leqslant$$
$$\leqslant -\frac{1}{2}(f''(x^n)(x^{n+1} - x^n), x^{n+1} - x^n) \leqslant -\frac{m}{2} \| x^{n+1} - x^n \|^2.$$

Thus,

$$\| x^{n+1} - x^n \|^2 \leqslant -\frac{2}{m} f_n(x^{n+1}). \tag{7.4}$$

On the other hand, $f_n(x^{n+1}) = (f'(x^n), x^{n+1} - x^n) + {}^1\!/_2 (f''(x^n)(x^{n+1} - x^n),$ $x^{n+1} - x^n) \geqslant (f'(x^n), x^{n+1} - x^n)$. Further, by (7.3), $f'(x^n) = f'(x^{n-1}) +$ $f''(x^{n-1})(x^n - x^{n-1}) + r$, where $\|r\| \leqslant R\|x^n - x^{n-1}\|^2$. Hence

$$\begin{aligned} -f_n(x^{n+1}) &\leqslant -(f'(x^n), x^{n+1} - x^n) = \\ &= -(f'(x^{n-1}) + f''(x^{n-1})(x^n - x^{n-1}) + r, x^{n+1} - x^n) = \\ &= -(f'_{n-1}(x^n), x^{n+1} - x^n) + (r, x^{n+1} - x^n) \leqslant (r, x^{n+1} - x^n) \leqslant \\ &\leqslant \| r \| \, \|x^{n+1} - x^n\| \leqslant R \| x^n - x^{n-1} \|^2 \|x^{n+1} - x^n\|. \end{aligned}$$

Combining this with inequality (7.4), we get $\|x^{n+1}-x^n\| \leqslant (2R/m)$ $\|x^n-x^{n-1}\|^2$. Hence it follows by induction that $\|x^{n+1}-x^n\| \leqslant$ $(m/2R)\delta^{2^n}$. Hence

$$\| x^k - x^n \| \leqslant \sum_{i=n}^{k-1} \| x^{i+1} - x^i \| \leqslant \frac{m}{2R} \sum_{i=n}^{k-1} \delta^{2^i},$$

i.e. $\|x^k - x^n\| \to 0$ as $k, n \to \infty$. The sequence x^n is therefore fundamental, so that there exists $x^* = \lim_{n\to\infty} x^n$, where $x^* \in Q$ and

$$\| x^n - x^* \| \leqslant \frac{m}{2R} \sum_{i=n}^{\infty} \delta^{2^i}.$$

It is easily shown that the sufficient condition (1.3) for a minimum is satisfied at x^*, which completes the proof of the theorem.

Notes. 1. The condition $\delta = (2R/m)\|x^1 - x^0\| < 1$, characterizing the closeness of x^0 to the minimum is more conveniently checked than any other condition, including $\|x^0 - x^*\|$.

2. If $Q = E$, (7.1) becomes the ordinary Newton method (4.9), and Theorem 7.1 is the same as Theorem 4.4.

3. Newton's method seems not to have been used previously for problems with constraints.

We give some examples of applying Newton's method.

1. For the problem of non-linear programming with linear constraints min $f(x)$, $Ax \leqslant b$, $x \in E^m$, $b \in E^r$, where A is an $m \times r$ matrix, method (7.1) amounts to a sequence of problems of quadratic programming

$$\min_{Ax\leqslant b} \left[(f'(x^n), x - x^n) + \frac{1}{2}(f''(x^n)(x - x^n), x - x^n) \right].$$

2. For the linear optimal control problem (2.1), (2.3) with a fixed value of $x(T) = d$ (so that we can assume $\Phi \equiv 0$), method (7.1) is the same as Newton's method for the unconstrained extremum problem (4.16) except that the boundary conditions become $\bar{x}(0) = \bar{x}(T) = 0$.

3. In the case of the linear optimal control problem (2.1), (2.3) with constraints of the type $u(t) \in M_t$ for almost all $0 \leqslant t \leqslant T$, method (7.1) reduces to a sequence of problems on minimizing a quadratic functional

under the same constraint. The method of solving this latter problem is given in [33].

4. In the case of the non-linear optimal control problem, Newton's method is extremely laborious. However, we can use here a method which is an analogue of (4.11) for the case when constraints are present. This method is therefore the same as the Newton method described above for linear systems, provided we take at each step, as the linear system (2.3), the linearized equation (2.2), i.e. the equation $d\bar{x}/dt = \varphi_x\bar{x} + \varphi_u\bar{u}$, $\bar{x}(0) = 0$.

8. Methods of set approximation

The essence of this group of methods is to replace the initial problem of minimizing $f(x)$ on a set Q by a sequence of problems on minimizing $f(x)$ on sets Q_n approximating to Q. The set sequence Q_n can be chosen in advance (as in the Ritz method), or the next Q_n can be determined from the results of the previous problems (as in the cut-off method). As distinct from all the previous methods discussed, the successive approximations obtained in set approximation methods do not necessarily belong to the initial set Q.

Theorem 8.1

Let Q be any set in Banach space E, and $f(x)$ a functional satisfying a Lipschitz condition on $\bigcup_{n=1}^{\infty} Q_n$ and bounded from below on Q:

$$\inf_{x\in Q} f(x) = f^* > -\infty.$$

Let Q_n be a sequence of sets, and the sequence $x^n \in Q_n$ such that

$$f(x^n) \leqslant \inf_{x\in Q_n} f(x) + \varepsilon_n, \quad \lim_{n\to\infty} \varepsilon_n = 0,$$

where

$$\lim_{n\to\infty} \inf_{y\in Q_n} \| x - y \| = 0 \quad \text{for all} \quad x \in Q, \tag{8.1}$$

$$\lim_{n\to\infty} \inf_{x\in Q} \| x^n - x \| = 0. \tag{8.2}$$

Then $\lim_{n\to\infty} f(x^n) = f^*$.

Proof. Let $y^k \in Q$ be a minimizing sequence for $f(x)$ on Q, i.e. $\lim_{k\to\infty} f(y^k) = f^*$ $(f(y^k) \geqslant f^*)$. By (8.1), given any $y^k \in Q$, there exists a sequence $z^n \in Q_n$ such that $\lim_{n\to\infty} z^n = y^k$. By the Lipschitz condition on $f(x)$, $\lim_{n\to\infty} f(z^n) = f(y^k)$. Hence

$$f(y^k) = \lim_{n\to\infty} f(z^n) \geqslant \overline{\lim_{n\to\infty}} (\inf_{x\in Q_n} f(x)) \geqslant \overline{\lim_{n\to\infty}} (f(x^n) - \varepsilon_n) = \overline{\lim_{n\to\infty}} f(x^n).$$

Passing to the limit as $k \to \infty$, we get $\overline{\lim_{n\to\infty}} f(x^n) \leqslant f^*$. Further, given any $\varepsilon > 0$ there exists for sufficiently large n, $\bar{x}^n \in Q$ such that $\| x^n - \bar{x}^n \| \leqslant \varepsilon$. Hence, by the Lipschitz condition of $f(x)$, $|f(x^n) - f(\bar{x}^n) | \leqslant M\varepsilon$. But $f(\bar{x}^n) \geqslant f^*$, since $\bar{x}^n \in Q$. Hence $\underline{\lim}_{n\to\infty} f(x^n) \geqslant f(\bar{x}^n) - M\varepsilon \geqslant f^* - M\varepsilon$. Since ε is arbitrary, we now have

$$\underline{\lim}_{n\to\infty} f(x^n) \geqslant f^*, \text{ i.e. } \overline{\lim_{n\to\infty}} f(x^n) = \underline{\lim}_{n\to\infty} f(x^n) = f^*.$$

Notes. 1. If the sequence Q_n is such that $Q_n \subset Q$ for all n, (8.2) is obviously satisfied, while if $Q \subset Q_n$ for all n, (8.1) is obviously satisfied.

2. Notice that the theorem does not demand that the sequence Q_n be a good approximation to Q everywhere; it is sufficient that it approximate Q in a neighbourhood of the minimum. In particular, it can happen that $Q_n \supset Q_{n+1} \supset Q$, but $\bigcap_{n=1}^{\infty} Q_n \neq Q$ (as in fact happens in the cut-off method).

3. The sequence x^n obtained in Theorem 8.1 is a GMS (Section 1) so that Theorems 1.4 - 1.6 are applicable.

Some methods of obtaining the set systems Q_n are discussed in the next two sections.

9. Ritz's method

The idea of this method, which is well known for problems without constraints [45], is to approximate the admissible set by finite-dimensional sets.

Let $a^1, \ldots, a^n, \ldots$ be a complete system in Q, i.e. for any $x \in Q$,

$$\lim_{n\to\infty} \inf_{\lambda_1\ldots\lambda_n} \left\| x - \sum_{i=1}^{n} \lambda_i a^i \right\| = 0.$$

We shall call it a basis. We denote by L_n the subspace stretched over $a^1, \ldots, a^n$. We take $Q_n = L_n \cap Q$ and solve the minimization problem for $f(x)$ on Q_n. This problem is finite-dimensional, since it amounts to minimizing $\varphi(\lambda, \ldots, \lambda_n) = f\left(\sum_{i=1}^{n} \lambda_i a^i\right)$ under the constraint $\sum_{i=1}^{n} \lambda_i a^i \in Q$.

Theorem 9.1

Let Q be any set in Banach space E; $a^1, \ldots, a^n$ be a complete system in Q; $f(x)$ a functional semicontinuous from below on Q. Let the sequence x^n be such that $x^n \in Q_n$, $f(x^n) \leqslant \inf_{x\in Q_n} f(x) + \varepsilon_n$, $\lim_{n\to\infty} \varepsilon_n = 0$. Then $\lim_{n\to\infty} f(x^n) = f^* = \inf_{x\in Q} f^n(x)$.

Proof. Let y^k be a minimizing sequence. Since $f(x)$ is semicontinuous from below and the basis is complete, for any y^k we can find $z^k \in Q_n$, $f(z^k) - f(y^k) \leqslant f(y^k) - f^* + \delta_k$, $\lim_{k\to\infty} \delta_k = 0$. The sequence z^k will now also be minimizing, i.e. $\lim_{k\to\infty} f(z^k) = f^*$. But, by definition of x^n, $f(x^n) \leqslant f(z^k) + \varepsilon_n$. Hence $\lim_{n\to\infty} f(x^n) = f^*$.

The Ritz method is worth using when the constraint $\sum_{i=1}^{n} \lambda_i a^i \in Q$ in E_n is not too complicated. We give some examples.

1. For the optimal control problem (2.1), (2.2) under the constraint $u(t) \in M$ for all $0 \leqslant t \leqslant T$ step functions are best taken as the basis. In fact, we introduce the functions $a_{ij}^{(n)}(t)$, $i = 1, \ldots, r$, $j = 1, \ldots, n$

$$a_{ij}^{(n)}(t) = \begin{cases} 1, & \text{if } \quad \frac{T}{n}(j-1) \leqslant t < \frac{T}{n} j, \\ 0, & \text{otherwise} \end{cases}$$

and seek $u(t)$ as a combination of these functions

$$u_i(t) = \sum_{j=1}^{n} \lambda_{ij} a_{ij}^{(n)}(t), \qquad i = 1, \ldots, r \tag{9.1}$$

The constraint $u(t) \in M$ now generates the constraints $\lambda_j \in M$, $j = 1, \ldots, n$, where λ_j is the vector with components $\lambda_{1j}, \ldots, \lambda_{rj}$. In particular, if the initial constraint has the form $|u_j(t)| \leqslant 1$, the constraint on λ_{ij} is unusually simple: $|\lambda_{ij}| \leqslant 1$ for all i and j. The auxiliary problem thus reduces to minimizing a function of rn variables λ_{ij} under simple constraints on these variables. We emphasize that, to evaluate the function for any values of the λ_{ij}, we have to introduce the control $u(t)$, (9.1), integrate system (2.2) with this control then evaluate the functional (2.1). This is not equivalent to the method of finite-difference approximation commonly employed for equations and functionals.

2. If there are no constraints, or they have the form

$$\int_0^T \sum_{i=1}^{r} u_i^2(t)\,dt \leqslant \rho \quad \text{or} \quad \int_0^T u_i^2(t)\,dt \leqslant \rho_i, \quad i = 1, \ldots, r,$$

in the optimal control problem, other bases can be chosen. For instance, we can use a system of trigonometric polynomials or Legendre polynomials.

3. If the optimal control problem (2.1) and (2.3) is linear, we can also apply the Ritz method when there are also constraints on the phase coordinates. For, if there are constraints $g_k(x(t_j)) \leqslant 0$, $k = 1, \ldots, s$, specified at a finite number of points t_j (in particular, the condition $x(T) = d$), then, since $x_i(t_j)$ is a linear functional of u, each such constraint generates a constraint on a linear combination of coefficients λ. In particular, if all the g_k are linear, we obtain the problem of nonlinear programming for λ with linear constraints.

4. If the constraints in problem (2.1), (2.3) are $g(x(t)) \leqslant 0$ for all $0 \leqslant t \leqslant T$, $g(x)$ is a convex function, then (2.1), (2.3) can be approximated by problems with step functions u and constraints $g(x(t_j)) \leqslant 0$, $t_j = Tj/n$, $j = 0, \ldots, n-1$.

Let their solution be $u^n(t)$, $x^n(t)$. Since $q(x)$, $x(t)$ are continuous, $g(x(t_j)) \leqslant 0$ implies $\lim_{n\to\infty} g(x^n(t)) \leqslant 0$ for all t, i.e. $g(u^n) = \max_{0\leqslant t\leqslant T} g(x^n(t)) \to +0$ as $n \to \infty$.

If $g(u) \leqslant 0$ is a correct constraint (see Section 2 for the condition for this), then $u^n(t)$ is a GMS (Section 1) and Theorems 1.4 - 1.6 are applicable. In particular, for the classical variational problem (i.e. for system (2.2) of the type $dx/dt = u$, $r = m$), $g(x(t_j)) \leqslant 0$ implies $g(x(t)) \leqslant 0$ for all t, ($x(t)$ corresponding to piecewise constant $u(t)$, piecewise linearly, so that the constraints $g(x(t)) \leqslant 0$ only need to be checked at the corners of the step-line). Thus we here obtain simply a minimizing sequence, and not a GMS.

5. If an entire system of constraints is given, for each of which the Ritz method (with the same basis) is applicable and reasonably simple, the method will not be too difficult for the problem as a whole. In particular, the Ritz method is easily applied for the linear optimal control problem in which there are constraints of the types $u(t) \in V$ and

$\int_0^T \sum_{i=1}^r u_i^2(t)\,dt \leqslant \rho$ and constraints on the phase coordinates. Herein lies

the important difference from the methods previously considered, where an ability to solve the auxiliary problem (projection, finding the minimum of a linear functional, etc.) on each set separately does not necessarily enable us to solve it on their intersection.

10. Cut-off methods

In these methods the problem of minimizing $f(x)$ on Q is solved successively on the sets $Q_n \supset Q_{n+1} \supset Q$. After finding the minimum point x^n of $f(x)$ on Q_n, we add a new constraint, "cutting-off" the point

$x_n \overline{\in} Q$, which in fact determines the new set Q_{n+1}.

Two convergence theorems are given for these methods below. In the first, the admissible set is specified by means of the constraint $g(x, s) \leqslant 0$ for all $s \in S$, s is a parameter; in the second, by means of the single constraint $g(x) \leqslant 0$. This difference is only apparent. For, introducing the functional $g(x) = \sup_{s \in S} g(x, s)$, we can reduce the constraint $g(x, s) \leqslant 0$ to $g(x) \leqslant 0$. On the other hand, if $g(x)$ is a convex functional, the constraint $g(x) \leqslant 0$ is equivalent to the system of (linear) constraints $g(s) + (g'(s), x - s) \leqslant 0$ for all $s \in E$, where $g'(s)$ is a linear support functional to $g(x)$ at the point s. The real difference between Theorems 10.1 and 10.2 is that the former refers only to the finite - dimensional case; though far weaker conditions are imposed on the functional in it than in Theorem 10.2.

We take the problem of minimizing the functional $f(x)$ under the constraints $x \in Q_0$ and $g(x, s) \leqslant 0$ for all $s \in S$, S is a set. Let $Q = Q_0 \cap \{x : g(x, s) \leqslant 0 \text{ for all } s \in S\}$. The following method of solution is possible. Suppose we have obtained the set Q_n (we start with $n = 0$). We find the minimum of $f(x)$ on Q_n; let this be at the point x^n. We then solve the problem: maximize $g(x^n, s)$ with respect to $s \in S$, let the solution be s^n. We now obtain Q_{n+1} by adding the new constraint $g(x, s^n) \leqslant 0$. Thus

$$x^n \in Q_n, \qquad f(x^n) = \min_{x \in Q_n} f(x), \qquad g(x^n, s^n) = \max_{s \in S} g(x^n, s), \qquad (10.1)$$
$$Q_{n+1} = Q_n \cap \{x: g(x, s^n) \leqslant 0\}.$$

Theorem 10.1

Let Q_0 be compact, $f(x)$ continuous on Q_0, $g(x, s)$ satisfy a Lipschitz condition with respect to x with constant M for all $s \in S$, and method (10.1) be applicable. Then, for the sequence (10.1),

$$\lim_{n \to \infty} f(x^n) = f^* = \min_{x \in Q} f(x),$$

and a subsequence x^{n_i} exists convergent to the solution.

Proof. Since $Q_n \supset Q_{n+1} \supset Q$, we have $f(x^n) \leqslant f(x^{n+1}) \leqslant f^*$. We show that $\underline{\lim}_{n \to \infty} g(x^n, s^n) \leqslant 0$. Let there exist $\varepsilon > 0$ such that $g(x^n, s^n) \geqslant \varepsilon$ for all reasonably large n. Now, $\varepsilon \leqslant g(x^n, s^n) = g(x^n, s^n) - g(x^k, s^n) + g(x^k, s^n) \leqslant M\|x^n - x^k\| + g(x^k, s^n)$. But since $x^k \in Q_{n+1}$ for $k > n$, we have $g(x^k, s^n) \leqslant 0$, i.e. $\|x^n - x^k\| \geqslant \varepsilon / M$ for all reasonably large n and all $k > n$. But this contradicts the compactness of Q_0. Hence $\overline{\lim}_{n \to \infty} g(x^n, s^n) \leqslant 0$. We select a subsequence x^{n_i} for which $\lim_{i \to \infty} g(x^{n_i}, s^{n_i}) \leqslant 0$ and for which there exists $x^* = \lim_{i \to \infty} x^{n_i}$ (this is possible because Q_0 is compact). Now, $g(x^*, s) = g(x^*, s) - g(x^{n_i}, s) + g(x^{n_i}, s) \leqslant M\|x^* - x^{n_i}\| + g(x^{n_i}, s^{n_i})$ Hence $g(x^*, s) \leqslant 0$ for all $s \in S$. Thus, $x^* \in Q$. Since $f(x^{n_i}) \leqslant f^*$, we also have $f(x^*) \leqslant f^*$; on the other hand, $f(x^*) \geqslant f^*$ be definition of f^*. Hence x^* is the solution of our problem.

We now consider the cut-off method (10.2)

$$f(x^n) = \min_{x \in Q_n} f(x), \qquad Q_{n+1} = Q_n \cap \{x : g(x^n) + (g'(x^n), x - x^n) \leqslant 0\},$$

for the problem of minimizing $f(x)$ on Q specified by the condition $Q = \{x : g(x) \leqslant 0\}$, where, as usual, $g'(x^n)$ is any linear support functional to $g(x)$ at the point x^n, and $Q_0 \supset Q$ is a set.

Theorem 10.2

Let $f(x)$ be uniformly convex on $Q_0 \subset E$, where E is reflexive, Q_0 convex and closed, and $g(x)$ a continuous convex functional satisfying a Lipschitz condition on Q_0. We now have, for method (10.2):

1) $\lim\limits_{m\to\infty} f(x^n) \leqslant f^* = \min\limits_{x\in Q} f(x), \quad \overline{\lim\limits_{n\to\infty}}\, g(x^n) \leqslant 0;$

2) if $g(x)$ is a correct constraint, and $f(x)$ satisfies a Lipschitz condition on Q_0, then $\lim\limits_{n\to\infty} x^n = x^*$, $x^* \in Q$, $f(x^*) = f^*$ and $\lim\limits_{n\to\infty} f(x^n) = f^*$;

3) if, in addition, $f(x)$ is a weakly convex functional and there exists $\bar{x}$: $g(\bar{x}) < 0$, $\bar{x}$ is an interior point of Q_0, then the convergence rate satisfies $f(x^n) - f^* \leqslant c_1 / n, \quad x^n - x^* \leqslant c_2 / \sqrt{n}.$

Proof. Since $Q_n \supset Q_{n+1} \supset Q$, we have $f(x^n) \leqslant f(x^{n+1}) \leqslant f^*$, and there exists $f^* \geqslant \bar{f} = \lim\limits_{n\to\infty} f(x^n)$. Since $f(x)$ is uniformly convex, while x^n is the minimum of $f(x)$ on Q_n, we have $f(x) - f(x^n) \geqslant \delta(\|x^n - x\|)$ for all $x \in Q_n$; in particular, $f(x^{n+1}) - f(x^n) \geqslant \delta(\|x^{n+1} - x^n\|)$. Hence it follows that $\lim\limits_{n\to\infty} \|x^{n+1} - x^n\| = 0$. Further, since $x^{n+1} \in Q_{n+1}$, we have $0 \geqslant g(x^n) + (g'(x^n), \quad x^{n+1} - x^n) \geqslant g(x^n) - \|g'(x^n)\| \, \|x^{n+1} - x^n\|$, i.e. $g(x^n) \leqslant M\|x^{n+1} - x^n\|$ (since $\|g'(x^n)\|$ is bounded by virtue of the Lipschitz condition on $g(x)$). Hence $\overline{\lim\limits_{n\to\infty}}\, g(x^n) = 0$.

We turn to the proof of (2) of the theorem. If $g(x^n) \leqslant 0$, then $x^n \in Q$, so that $f(x^n) \geqslant f^*$, i.e. in this case x^n is the solution. Let $g(x^n) > 0$ for all n; now, by what has been proved, $g(x^n) \to +0$ as $n \to \infty$. If $g(x)$ is a correct constraint, it follows from this that $\lim\limits_{n\to\infty} \rho(x^n, Q) = 0$. Theorem 8.1 is now applicable, whence $\lim\limits_{n\to\infty} f(x^n) = f^*$. But it follows from this, by Theorem 1.5, that x^n converges to the solution x^*.

We now estimate the convergence rate for strongly convex functionals. For these, the above condition $f(x^{n+1}) - f(x^n) \geqslant \delta(\|x^{n+1} - x^n\|)$ becomes $f(x^{n+1}) - f(x^n) \geqslant \gamma\|x^{n+1} - x^n\|^2$, $\gamma > 0$. Further, let $g(x) > 0$ for all n in which case there exists $0 < \lambda_n < 1$ such that $g(\bar{x} + \lambda_n(x^n - \bar{x})) = 0$. Using the convexity of $g(x)$, we get $0 = g(\bar{x} + \lambda_n(x^n - \bar{x})) \leqslant (1 - \lambda_n)g(\bar{x}) + \lambda_n g(x^n)$, i.e. $g(x^n) \geqslant (1 - \lambda_n / \lambda_n)[-g(\bar{x})] \geqslant (1 - \lambda_n)[-g(\bar{x})]$. But $\|x^n - (\bar{x} + \lambda_n(\bar{x}^n - \bar{x}))\| \geqslant \rho(x^n, Q)$, so that $1 - \lambda_n \geqslant \rho(x^n, Q) / \|x^n - \bar{x}\| \geqslant c\rho(x^n, Q)$ (because $\{x^n\}$ is bounded). Thus, $g(x^n) \geqslant c\rho(x^n, Q)[-g(\bar{x})] = K\rho(x^n, Q)$, $K > 0$. As was shown above, $g(x^n) \leqslant M\|x^{n+1} - x^n\|$. Hence $\rho(x^n, Q) \leqslant (M / K)\|x^{n+1} - x^n\|$, and since $f^* - f(x^n) \leqslant L\rho(x^n, Q)$, then $f^* - f(x^n) \leqslant (LM / K)\|x^{n+1} - x^n\|$. Let $f^* - f(x^n) = \varphi_n$. We have obtained $\varphi_n \leqslant (LM / K)\|x^{n+1} - x^n\|$, while above we obtained $f(x^{n+1}) - f(x^n) = \varphi_n - \varphi_{n+1} \geqslant \gamma\|x^{n+1} - x^n\|^2$. Combining these, we get

$$\varphi_{n+1} \leqslant \varphi_n - \frac{\gamma K^2}{L^2 M^2} \varphi_n^2.$$

Using the same technique as in the proof of (2) of Theorem 5.1, we obtain the required estimate.

Note. It is not clear whether the convergence rate estimate obtained above can be improved.

We now consider the aspect of the cut-off method in various concrete problems.

1. The general problem of convex programming (minimization of $f(x)$ under the constraints $g_i(x) \leqslant 0$, $i = 1, \ldots, r$, $f(x)$, $g_i(x)$ are convex functions in E_m) can easily be seen to reduce to the problem of minimizing the linear function x_{m+1} under the convex constraints $f(x) - x_{m+1} \leqslant 0$, $g_i(x) \leqslant 0$, $i = 1, \ldots, r$, in space E_{m+1}. Further, the system of convex constraints of the type $g_i(x) \leqslant 0$, $i = 1, \ldots, r$, reduces to a single convex constraint $g(x) \leqslant 0$, where $g(x) = \max_{1 \leqslant i \leqslant r} g_i(x)$. Thus, instead of the initial problem, we can consider the minimization problem for the linear function (c, x) under the convex constraint $g(x) \leqslant 0$. Let Q_0 be a polyhedron (if the constraint $x \in Q_0$ is absent, it has to be artificially introduced, e.g. in the form $|x_i| \leqslant M$ for all i, M is a large number). Now, $Q_{n+1} = Q_n \cap \{x : g(x^n) + (g'(x^n), x - x^n) \leqslant 0\}$ is also a polyhedron. Method (10.2) thus reduces here to solving a sequence of linear

programming problems. The convergence of the method can be proved by Theorem 10.1. For, the constraint $g(x) \leqslant 0$ is equivalent to $g(x, s) = g(s) + (g'(s), x - s) \leqslant 0$ for all s. Method (10.1) is now easily seen to be the same as method (10.2).

This "method of cutting planes" was proposed by Kelley for this problem [34], and he proved its convergence.

2. The problem of best approximation amounts to minimizing

$$f(x) = \max_{s \in S} \left| \varphi(s) - \sum_{i=1}^{m} x_i \varphi_i(s) \right|,$$

where S is a compactum, $\varphi(s)$, $\varphi_i(s)$ are continuous functions on it, and the φ_i, ..., φ_m are linearly independent. In addition, certain auxiliary constraints can be imposed on the coefficients x_i: $x \in Q_0$. This problem amounts to minimizing x_{m+1} under the constraints

$$\left| \varphi(s) - \sum_{i=1}^{m} x_i \varphi_i(s) \right| \leqslant x_{m+1} \quad \text{for all} \quad s \in S, \quad x \in Q_0.$$

Method (10.1) is applicable to this problem, where it takes the following form: x_1^k, ..., x_m^k, x_{m+1}^k is the solution of the problem: min x_{m+1} under the constraints

$$\left| \varphi(s_j) - \sum_{i=1}^{m} x_i \varphi_i(s_j) \right| \leqslant x_{m+1}, \quad j = 1, \dots, k, \quad x \in Q_0, \qquad (10.3)$$

s_{k+1} is given by the condition

$$\left| \varphi(s_{k+1}) - \sum_{i=1} x_i^k \varphi_i(s_{k+1}) \right| = \max_{s \in S} \left| \varphi(s) - \sum_{i=1}^{m} x_i^k \varphi_i(s) \right|.$$

The method thus reduces to solving a sequence of problems of best approximation on a discrete system of points s_1, ..., s_k (which is equivalent to the problem of linear programming if Q_0 is a polyhedron) and to determination of the maximum modulus of the deviation for the next approximation. Methods of this kind are familiar in the theory of best approximation (see e.g. [35], which contains various results on their convergence and also references to other work).

3. We consider the linear optimal control problem (2.1), (2.3) under constraints on the phase coordinates of the type $q(x(t)) \leqslant 0$ for all

$0 \leqslant t \leqslant T$, where $q(x)$ is a convex function. We investigate the method whose n-th step consists in minimizing $f(u)$ under the constraints $q(x^i(t_i)) + (q'(x^i(t_i)), x(t_i) - x^i(t_i)) \leqslant 0$, $i = 1, \ldots, n-1$, where $x^i(t)$ is the solution of the problem at the i-th step, and t_i the point introduced at this step. We introduce the next, n-th point from the condition

$$q(x^n(t_n)) = \max_{0 \leqslant t \leqslant T} q(x^n(t)). \tag{10.4}$$

This method is a variant on method (10.2). For, the constraint $q(x(t)) \leqslant 0$ is equivalent to $g(u) \leqslant 0$, where

$$g(u) = \max_{0 \leqslant t \leqslant T} q(x(t))$$

is a convex functional (as usual, $x(t)$ is the solution of (2.3) corresponding to the control $u(t)$). Further, if (10.4) is satisfied, for one of the support functionals to $g(u)$ at the point u^n the constraint $g(u^n) + (g'(u^n), u - u^n) \leqslant 0$ becomes $q(x^n(t^n)) + (q'(x^n(t_n)), x(t_n) - x^n(t_n)) \leqslant 0$, i.e. is the same as the above. We can therefore apply Theorem 10.2 to prove the convergence. All the assertions of the theorem will certainly be satisfied if $Q_0 = E$, $f(u)$ is a strongly convex functional, and $g(u)$ a correct constraint (the sufficient conditions for this are given in Section 2).

We shall not dwell on methods of solving the auxiliary problem. We merely remark that it consists in minimizing $f(u)$ under a finite number of linear constraints of the type $(c^i, u) \leqslant \alpha_i$, $i = 1, \ldots, n-1$. It can therefore be solved like any analogous finite - dimensional problem of mathematical programming.

11. Method of penalty functions

In all the methods discussed so far, the constrained extremum problem has been reduced to a sequence of simpler constrained extremum problems. The initial problem may also be reduced to a sequence of unconstrained extremum problems. In the method of penalty functions the unconstrained problem is obtained by introducing a penalty on infringement of the constraints. This idea is common in computational practice [20, 22, 36 - 42]; it seems to have been first clearly stated by Courant.

We give some examples of penalty functions. We shall assume that our set is given by means of one constraint $g(x) \leqslant 0$, and consider a sequence of penalty functions of the type $\psi_n(x) = \varphi_n(g(x))$. The following are the most popular examples of such functions.

1. $$\psi_n(x) = K_n g(x)_+, \tag{11.1}$$

where $K_n \geqslant 0, \lim\limits_{n\to\infty} K_n = \infty$; here and below, $g(x)_+ = \max\{0, g(x)\}$.

Obviously, $\psi_n(x) \geqslant 0$ for all x; $\lim\limits_{n\to\infty} \psi_n(x) = \infty$, if $g(x) > 0$; the $\psi_n(x)$ are convex if $g(x)$ is convex.

2. $$\psi_n(x) = (g(x)_+ + 1)^n - 1. \tag{11.2}$$

It is easily shown that $\psi_n(x) \geqslant 0$, $\psi_n(x) = 0$, if $g(x) \leqslant 0$; $\lim\limits_{n\to\infty} \psi_n(x) = \infty$ for $g(x) > 0$; the $\psi_n(x)$ are convex if $g(x)$ is convex.

3. $$\psi_n(x) = [(g(x) + 1)_+]^n. \tag{11.3}$$

Here, $\psi_n(x) \geqslant 0$ for all x, $\lim\limits_{n\to\infty} \psi_n(x) = \infty$ for $g(x) > 0$, as distinct from the previous examples, generally speaking, $\psi_n(x) \neq 0$ for $g(x) \leqslant 0$ (for instance, $\psi_n(x) \equiv 1$ for $g(x) = 0$), but $\lim\limits_{n\to\infty} \psi_n(x) = 0$ for $g(x) < 0$. In addition, the $\psi_n(x)$ are convex if $g(x)$ is convex.

4.

$$\psi_n(x) = \begin{cases} +\infty, & \text{if} \quad g(x) \geqslant 0, \\ -\dfrac{\alpha_n}{g(x)}, & \text{if} \quad g(x) < 0, \end{cases} \tag{11.4}$$

where $\alpha_n \geqslant 0, \lim\limits_{n\to\infty} \alpha_n = 0$. As before, $\psi_n(x) \geqslant 0$ for x, $\lim\limits_{n\to\infty} \psi_n(x) = 0$ for $g(x) < 0$ and the $\psi_n(x)$ are convex for $g(x) < 0$ if $g(x)$ is convex.

The functions (11.4) are not continuous on E. However, we can introduce more complicated functions of the same kind, which are continuous. For example,

$$\psi_n(x) = \begin{cases} -\alpha_n / g(x), & \text{if} \quad g(x) < 0 \text{ and } -\alpha_n/g(x) \leqslant K_n, \\ K_n \left(\dfrac{1 + g(x)}{1 - \varepsilon_n / K_n} \right)^n & \quad \text{otherwise} \end{cases}$$

where $\alpha_n \geqslant 0$, $\lim_{n\to\infty} \alpha_n = 0$, $K_n \geqslant 0$, $\lim_{n\to\infty} K_n = \infty$.

Of course there are other examples of penalty functions. In particular, they can be obtained from (11.1) - (11.4) if the constraint $g(x) \leqslant 0$ is replaced by the equivalent constraint $\varphi(g(x)) \leqslant 0$, where $\varphi(t)$ is a continuous function, and $\varphi(t) > 0$ for $t > 0$ and $\varphi(t) < 0$ for $t < 0$.

We now consider the method of penalty functions itself for the problem of minimizing $f(x)$ under the constraint $x \in Q$, where $Q = \{x : g(x) \leqslant 0\}$. The method consists in solving (up to ε_n, where $\lim_{n\to} \varepsilon_n = 0$) a sequence of problems on the unconstrained minimum of $f(x) + \psi_n(x)$, $\psi_n(x) = \varphi_n(g(x))$, i.e. in determining the sequence x^n:

$$f(x^n) + \psi_n(x^n) \leqslant d_n + \varepsilon_n, \qquad d_n = \inf_{x\in E} [f(x) + \psi_n(x)]. \tag{11.5}$$

Theorem 11.1

Let $f(x)$ and $g(x)$ be continuous, Q be non-empty, $d_n \geqslant d > -\infty$, $\psi_n(x) \geqslant 0$ for all x, n, $\lim_{n\to\infty} \psi_n(x) = 0$ on the set R everywhere dense in Q, and $\lim_{n\to\infty} \varphi_n(t) = \infty$ for $t > 0$. Now, in method (11.5), $f(x^n) \leqslant f^* = \inf_{x\in Q} f(x)$, $\overline{\lim_{n\to\infty}} g(x^n) \leqslant 0$. If the constraint $g(x)$ is correct, and $f(x)$ satisfies a Lipschitz condition in some neighbourhood Q, then $\lim_{n\to\infty} f(x^n) = f^*$, $\lim_{n\to\infty} \rho(x^n, Q) = 0$, i.e. x^n is a GMS.

Proof. Let y^m be a minimizing sequence for $f(x)$ on Q, i.e. $y^m \in Q$, $\lim_{m\to\infty} f(y^m) = f^*$. Given any y^m, we can find $z^m \in R$ such that $|f(y^m) - f(z^m)| \leqslant f(y^m) - f^* + \delta_m$, $\lim_{m\to\infty} \delta_m = 0$. This is possible, since R is dense in Q and $f(x)$ continuous. Obviously, z^m is now also a minimizing sequence. By definition of x^n, we have $f(x^n) + \psi_n(x^n) \leqslant f(z^m) + \psi_n(z^m) + \varepsilon_n$. We take $\varepsilon > 0$, and choose m so that $f(z^m) \leqslant f^* + \varepsilon/3$; since $z^m \in R$, we can find an N so that $\psi_n(z^m) \leqslant \varepsilon/3$ and $\varepsilon_n \leqslant \varepsilon/3$ for all $n \geqslant N$. Now, $f(x^n) + \psi_n(x^n) \leqslant f^* + \varepsilon$. Since ε is arbitrary, this means that $\overline{\lim_{n\to\infty}} (f(x^n) + \psi_n(x^n)) \leqslant f^*$, and since $\psi_n(x^n) \geqslant 0$ we also have $\overline{\lim_{n\to\infty}} f(x^n) \leqslant f^*$. On the other hand, $\overline{\lim_{n\to\infty}} \psi_n(x^n) \leqslant \overline{\lim_{n\to\infty}} (f^* - f(x^n)) \leqslant f^* - d$. Hence it follows (since $\lim_{n\to\infty} \psi_n(x) = \lim_{n\to\infty} \varphi_n(g(x)) = \infty$ for $g(x) > 0$), that $\overline{\lim_{n\to\infty}} g(x^n) \leqslant 0$. Now let $g(x)$ be a correct constraint. In this case, $\overline{\lim_{n\to\infty}} g(x) \leqslant 0$ implies $\lim_{n\to\infty} \rho(x^n, Q) = 0$. Finally, it follows from the Lipschitz condition on $f(x)$ that $f(x^n) \geqslant f^* + L\rho(x^n, Q)$, i.e. $\underline{\lim_{n\to\infty}} f(x^n) \geqslant f^*$. Finally, therefore, $\lim_{n\to\infty} f(x^n) = f^*$.

Note. Penalty functions of the type (11.1) and (11.2) satisfy the conditions of the theorem everywhere, but those of type (11.3) and (11.4) only when $R = \{x : g(x) < 0\}$ is dense in $Q = \{x : g(x) \leqslant 0\}$. The latter is true, in particular, if $g(x)$ is convex and R non-empty. Functions of the type (11.3), (11.4) are certainly continuous if R is empty.

We turn to application of the method of penalty functions to specific problems.

1. In the problem of mathematical programming the constraints $g_i(x) \leqslant 0$, $i = 1, \ldots, r$ can be reduced to the single constraint $g(x) \leqslant 0$ by many methods, e.g.

$$g(x) = \max_{1 \leqslant i \leqslant r} \{g_i(x)\} \quad \text{or} \quad g(x) = \sum_{i=1}^{r} \alpha_i g_i(x)_+, \qquad \alpha_i > 0.$$

It is specially convenient to take

$$g(x) = \sum_{i=1}^{r} \alpha_i [g_i(x)_+]^2, \qquad \alpha_i > 0,$$

since $g(x)$ is now differentiable if the $g_i(x)$ are differentiable. After this, penalty functions of the type (11.1) - (11.4) can be applied.

As already mentioned, if Q is bounded, the constraint $g(x) \leqslant 0$ is correct in the finite - dimensional case and Theorem 11.1 is applicable. The method of penalty functions in the form (11.1) was considered for problems of mathematical programming in [40, 22], in form (11.4) in [38, 39], in a rather more general form of the same type in [41], and in a form similar to (11.3) in [43].

2. If the constraints are in the form of equations $g_i(x) = 0$, $i = 1, \ldots, r$, they can be replaced by a single constraint of the inequality type, e.g.

$$g(x) \leqslant 0, \qquad g(x) = \sum_{i=1}^{r} \alpha_i g_i^2(x), \qquad \alpha_i > 0,$$

after which the method of penalty functions can be employed. Here, it is not possible to use penalty functions of the type (11.3) or (11.4), since the set $R = \{x : g(x) < 0\}$ is empty in this case. Penalty functions of the type (11.1) were investigated for this problem in [42].

3. If the set Q cannot be specified by means of functionals, it can always be specified in the form $g(x) \leqslant 0$. In particular, for any set. $Q = \{x : g(x) \leqslant 0\}$, where $g(x) = \rho(x, Q) = \inf_{y \in Q} \| x - y \|$. This constraint is moreover correct and $g(x) = 0$ for $x \in Q$. If Q is bounded, closed, has an interior point (for simplicity the point 0) and is convex, then it can be specified in the form $Q = \{x : g(x) \leqslant 1\}$ by means of the Minkovskii function $g(x) = \inf_{\lambda > 0,\, x/\lambda \in Q} \{\lambda\}$. This constraint is also correct and convex.

4. We now turn to optimal control problems (2.1), (2.2). We take first constraints on the control specified in the form

$$Q = \{u \in L_2^r : \; u(t) \in M \text{ for almost all } 0 \leqslant t \leqslant T\},$$

M is a convex closed set in E_r having interior points. Let $\varphi_n(u)$ be a system of functions on E_r satisfying the conditions: (1) $\varphi_n(u) \geqslant 0$ for all u; (2) if $u \in M^0$, i.e. u is an interior point of M, then $\lim_{n \to \infty} \varphi_n(u) = 0$, where $\varphi_n(u)$ is monotonically decreasing; (3) if $u \overline{\in} M$, then $\varphi_n(u) \geqslant K_n \rho^2(u, M)$, where $K_n \geqslant 0$, $\lim_{n \to \infty} K_n = \infty$, K_n is independent of u, $\rho(u, M)$ is the distance from u to M (in E_r). Notice that, if $\varphi_n(u) = 0$ for $u \in M$, we no longer require M to be convex or have interior points. The following are some examples of functions $\varphi_n(u)$ satisfying all the conditions. If the constraints have the form

$$\sum_{i=1}^{r} u_i^2(t) \leqslant 1,$$

we can take

$$\varphi_n(u) = K_n \left(\sum_{i=1}^{r} u_i^2 - 1 \right) \qquad \text{or} \qquad \varphi_n(u) = \left(\sum_{i=1}^{r} u_i^2 \right)^n.$$

If the constraints are $|u_i(t)| \leqslant 1,\ i = 1, \ldots, r,$ we can take

$$\varphi_n(u) = K_n(\max_{1\leqslant i\leqslant r}\{|u_i|\} - 1)_+ \quad \text{or} \quad \varphi_n(u) = \sum_{i=1}^{r} |u_i|^n.$$

We introduce

$$\psi_n(u) = \int_0^T \varphi_n(u(t))\,dt.$$

These penalty functions satisfy all the conditions of theorem 11.1. For. $\psi_n(u) \geqslant 0,\ \lim_{n\to\infty}\psi_n(u) = 0$ on $R = \{u\colon\ u(t) \in M^0\}$, and $R = q$, while finally,

$$\psi_n(u) = \int_0^1 \varphi_n(u(t))\,dt \geqslant \int_S \varphi_n(u(t))\,dt \geqslant K_n \int_S \rho^2(u(t), M)\,dt =$$

$$= K_n \int_0^T \|u(t) - \bar{u}(t)\|^2_{E_r}\,dt = K_n\rho^2(u,\ Q),$$

where $S = \{t \in [0, T]\colon\ u(t) \bar{\in} M\}$, $\bar{u}(t)$ is the projection of $u(t)$ on to M. Hence, if the remaining conditions of Theorem 11.1 are satisfied (i.e. the Lipschitz condition on $f(u)$ and the boundedness from below of $f(u) + \psi_n(u)$), then in method (11.5), amounting to consecutive minimization of $f(u) + \psi_n(u)$, without constraints, we have the convergences $f(u^n) \to f^*,\ \rho(u^n, Q) \to 0$.

The method of penalty functions can also be used if there are constraints on the phase coordinates. For instance, with the constraint $g(x(t)) \leqslant 0$ for all $0 \leqslant t \leqslant T$, we can introduce the penalty function

$$\psi_n(u) = K_n[\max_{0\leqslant t\leqslant T} g(x(t))]_+ \quad \text{or} \quad \psi_n(u) = K_n \int_0^T [g(x(t))_+]^2\,dt$$ (the latter is specially suitable, being differentiable if $g(x)$ is differentiable); with the constraint $x(T) = d$ the penalty function can be $\psi_n(u) = K_n\| x(T) - d \|^2_{E_m}$. The conditions of Theorem 11.1 are easily shown to hold here (the criteria for correctness of the constraints are given in Section 2).

If there are several simultaneous constraints, each of them (or part of them) can be replaced independently by penalty functions. But difficulties arise here when verifying if the system of constraints is correct. Notice that the same difficulties arise in this situation in the cut-off method (see Section 10).

The idea of applying penalty functions to optimal control problems occurs in [20, 22, 36, 37]. But only [36] contains a proof of convergence, for one particular case.

12. The computational aspect

We obviously cannot consider the computational aspect of the full range of extremum problems in the present section: this is a very important topic which has not been adequately investigated; the discussion of it would have to be largely based on the results of numerical experiments. We shall merely dwell on the computational aspect that can be illuminated by the convergence theorems proved above.

A few words first about unconstrained extremum problems (Section 4). It is clear from Theorem 4.1 that methods of the gradient type are applicable to an extraordinarily wide class of functions and virtually any initial approximations. For Newton's method on the other hand (Theorem 4.4), very strict conditions are required for the functional being minimized and for the initial approximation. Further, Newton's method involves much more complicated computations at each step compared with the gradient method (namely, we have to find the second derivative $f''(x)$ and transform it). Nevertheless, Newton's method has one advantage: a high rate of convergence, which is sometimes decisive. For, the conditions of Theorem 4.3 are usually satisfied close to the minimum, and the gradient method (with optimal choice of α) is convergent at the rate of a geometric progression with ratio $q = (M - m)/(M + m) = (p - 1)/(p + 1)$, where $p = M/m \geqslant 1$. But in many practical problems, $p = M/m$ is extremely large (e.g. in poorly stated problems of linear algebra, "well organized" functions of several variables, and functions $F(x, u, t)$ weakly dependent on u in optimal control problems). In these cases $q \approx 1 - 2/p$, i.e. q is close to unity and the convergence of the gradient method is slow. In Newton's method, however, we have the quadratic convergence $\| x^n - x^* \| \leqslant C\delta^{2^n}$, where δ is independent of p, but depends on how close the functional $f(x)$ is to quadratic in the neighbourhood of x^*. Hence the convergence can be extremely fast, even for large p. Accordingly, the following combination of methods can be useful. We first use the gradient method, which does not require a good initial approximation and provides fairly fast convergence at the start. After the convergence of the gradient method has slowed down, it usually becomes possible to apply Newton's method, since a good initial approximation has now been obtained. The extra complexity of the computations at each step is compensated by the substantial increase in the rate of convergence.

We now turn to problems under constraints. In the methods of Sections 5 - 10 the initial problem reduces to a sequence of auxiliary extremal problems on minimizing a linear (Section 5) or quadratic (Section 7) functional on Q, on projecting on to Q (Section 6), or on minimizing $f(x)$ on sets Q_n (Sections 8 - 10). Naturally, a given method is only worthwhile if the auxiliary problems can be solved fairly easily. Actually, all the examples in this paper were chosen so that the auxiliary problem could be solved either explicitly or by some finite method (we count as finite, methods which reduce to solving a Cauchy problem for a non-linear ordinary differential equation or to solving a boundary value problem for a linear equation).

As regards convergence rates, in the gradient projection method (Section 5) we have convergence at the rate of a geometric progression for a strongly convex functional and any convex set (Theorem 5.1), whereas in the conditional gradient method (Section 6) convergence of the same type can be proved for a convex functional having no absolute minimum on Q, and a strongly convex set (Theorem 6.1). On the other hand, the auxiliary problems in these methods can be solved fairly easily in roughly the same cases. It follows from this that, if the functional is only convex, or if it is strongly convex, but $p = M/m$ is large, the conditional gradient method is worth using in the case of a strongly convex set. If the set is not strongly convex, this method can give extremely slow convergence. In particular, in the optimal control problem (2.1), (2.3) with $F(x, u, t) = F(x, t)$ and constraints of the type

$\int_0^1 \sum_{i=1}^r u_i^2(t)\,dt \leqslant \rho$ the conditional gradient method is suitable. If $F(x, u, t)$ is strongly convex with respect to $\{x, u\}$, and the constraint is $u(t) \in M$ for all $0 \leqslant t \leqslant T$, the gradient projection method is preferable. As regards the gradient projection and Newton methods, they are in roughly the same relationship as their analogues in the case of an unconstrained extremum.

The question of the convergence rate for set approximation methods is fairly complex. It is important to notice, however, that these methods are still suitable for optimal control problems when the methods of Section 5 - 7 lead to excessively difficult auxiliary problems, namely, when there are constraints on the phase coordinates (see the examples in Sections 9 - 10). In addition, in the cut-off methods (Section 10), $f(x^n) \leqslant f^*$, which enables us to give a lower bound for f^*. When combined with one of the methods of Sections 5 - 7 or 9, in which we always have $f(x^n) \geqslant f^*$, this method can yield an error estimate for the approximate solution.

We now turn to the method of penalty functions (Section 11). This method is above all extremely general. It can be applied to virtually any extremum problem (convergence conditions apart). In addition, the auxiliary problems involved always appear simple. This is by no means the case, however. It turns out that the solution of the unconstrained extremum problem for the functional $\Phi_n(x) = f(x) + \psi_n(x)$ becomes more and more difficult as n increases (provided the minimum does not lie inside Q). We shall illustrate this with an example of minimizing a strongly convex functional $f(x)$ under the one linear constraint $(c,x) = 0$. We introduce the penalty function $\psi_n(x) = K_n(c,x)^2$, $K_n \geqslant 0$, $\lim_{n\to\infty} K_n = \infty$.

We shall solve the unconstrained minimum problem for $\Phi_n(x) = f(x) + \psi_n(x)$ by the gradient method and estimate the rate of convergence. Let x_1 be such that $(c, x_1) = 0$, $\| x_1 \| = 1$, and x_2 such that $(c, x_2) = \lambda > 0$, $\| x_2 \| = 1$. Then, if $(c, x) = 0$, $(\Phi_n''(x)x_1, x_1) = (f''(x)x_1, x_1) \leqslant M$, and $(\Phi_n''(x)x_2, x_2) = (f''(x)x_2, x_2) + K_n\lambda^2 \geqslant m + K_n\lambda^2$. Hence $p_n = M_n/m_n \geqslant (m + K_n\lambda^2) / M$, i.e. $p_n \to \infty$ as $n \to \infty$. But the convergence rate is determined by the number $q_n \approx 1 - 2 / p_n$, so that $q_n \to 1$ as $n \to \infty$. In other words, the rate of convergence of the gradient method in the auxiliary problem is slower, the greater n. The method of penalty functions appears to be suitable only when other methods fail, or in order to obtain an initial approximation for another method. Notice, however, that, by employing penalty functions of the type (11.4), an upper bound can be obtained for f^*, and a lower bound by employing functions of the type (11.1) or (11.2).

Translated by D.E. Brown

REFERENCES

1. DUBOVITSKII, A.Ya. and MILYUTIN, A.A. *Zh. vȳchisl. Mat. mat. Fiz.* 5, 3, 395 - 453, 1965.

2. KANTOROVICH, L.V. and AKILOV, G.P. *Functional analysis in normed spaces* (Funktsional'nyi analiz v normirovannykh prostranstvakh), Fizmatgiz, Moscow, 1959.

3. VAINBERG, M.M. *Variational methods for investigating non-linear*

operators (Variatsionnye metody issledovaniya nelineinykh operatorov), Gost., Moscow, 1956.

4. LEVITIN, E.S. and POLYAK, B.T. *Dokl. Akad. Nauk SSSR*, **168**, 5, 997 - 1000, 1966.

5. POLYAK, B.T. *Dokl. Akad. Nauk SSSR*, **166**, 2, 287 - 290, 1966.

6. TIKHONOV, A.N. *Dokl. Akad. Nauk SSSR*, **162**, 4, 763 - 765, 1965.

7. ZOUTENDIJK, G. *Methods of feasible directions*, Elsevier, 1960.

8. PONTRYAGIN, L.S., BOLTYANSKII, V.G., GAMKRELIDZE, R.V. and MISHCHENKO, E.F. *Mathematical theory of optimal processes* (Matematicheskaya teoriya optimal'nykh protsessov), Fizmatgiz, Moscow, 1961.

9. KANTOROVICH, L.V. *Dokl. Akad. Nauk SSSR*, **28**, 3, 212 - 215, 1940.

10. KYUNTSI, G.P. and KRELLE, V. *Non-linear programming* (Nelineinoe programmirovanie), Soviet Radio, Moscow, 1965.

11. EGOROV, Yu.V. *Zh. vȳchisl. Mat. mat. Fiz.* **3**, 5, 887 - 904, 1963.

12. ARROW, K.J. *et al.* *Studies in linear and non-linear programming*, Stamford, 1958.

13. PSHENICHNYI, B.N. *Zh. vȳchisl. Mat. mat. Fiz.* **5**, 1, 98 - 106, 1965.

14. VULF, F. New methods of non-linear programming, Symp. Application of mathematics to economics, Vol. 3, 312 - 334, Mysl', Moscow, 1965.

15. POLYAK, B.T. *Zh. vȳchisl. Mat. mat. Fiz.* **3**, 4, 643 - 654, 1963.

16. BROWN, R.R. The design of optimum systems, Proc. V ann. computer appl. symp., 1958, Chicago, 1959, 86 - 93.

17. POLYAK, B.T. *Zh. vȳchisl. Mat. mat. Fiz.* **4**, 6, 995 - 1005, 1964.

18. SOKOLOV, S.N. and SILIN, I.N. *Minimization of functionals by the linearization methods* (Nakhozhdenie minimumov funktsionalov metodom linearizatsii), Preprint, OIYaI, D810, 1961.

19. STEIN, M.L. On methods for obtaining solutions of fixed endpoint

problems in the calculus of variations, *J. Res. Nat. Bur. Standards*, 50. 3, 277 - 297, 1953.

20. SHATROVSKII, L.I. *Zh. vȳchisl. Mat. mat. Fiz.* **2**, 3, 488 - 490, 1962.

21. BRYSON, A.E. and DENHAM, W.F. A steepest ascent method for solving optimum programming problems, *J. appl. Mech.* **29**, 2, 247 - 257, 1962.

22. KELLEY, J.E. The gradient method. In symp. *Optimization methods with applications to space flight mechanics*, 244 - 306, Moscow, "Nauka", 1965.

23. AOKI, M. On a successive approximation technique in solving some control optimization problems, *J. Math. anal. Appl.* **5**, 3, 418 - 431, 1962.

24. POLYA, G. and SZEGÖ, G. *Aufgaben und Lehrsatze aus der Analysis*, Springer, Berlin, 1925.

25. GOLDSTEIN, A.A. Convex programming in Hilbert space, *Bull. Am. Math. Soc.* **70**, 5, 709 - 710, 1964.

26. ROSEN, J.B. The gradient projection method for non-linear programming I., *J. soc. ind. appl. Math.* **8**, 2, 181 - 217, 1960.

27. DEMYANOV, V.F. and RUBINOV, A.M. *Vest. LGU*, No. 19, 5 - 18, 1964.

28. DEM'YANOV, V.F. *Kibernetika*, No. 6, 65 - 74, 1965.

29. FRANK, M. and WOLFE, Ph. Algorithm for quadratic programming, *Nav. Res. Log. Quart.* **3**, 1 - 2, 95 - 110, 1956.

30. DEM'YANOV, V.F. *Avtomatika Telemekh.* **25**, 1, 3 - 12, 1964.

31. PSHENICHYI, B.N. *Zh. vȳchisl. Mat. mat. Fiz.* **5**, 2, 236 - 241, 1965.

32. KRYLOV, I.A. and CHERNOUS'KO, F.L. *Zh. vȳchisl. Mat. mat. Fiz.* **2**, 6, 1132 - 1139, 1962.

33. LETOV, A.M. *Avtomatika Telemekh.* **23**, 11, 1405 - 1413, 1962.

34. KELLEY, J.E. The cutting plane method for solving convex programs, *J. soc. Ind. appl. Math.* **8**, 4, 703 - 712, 1960.

35. CHENEY, E.W. and GOLDSTEIN, A.A. Newton's method for convex

programming and Tchebysheff approximation, *Num. Math.* 1, 5, 253 - 259, 1959.

36. KRASOVSKII, N.N. *Prikl. Mat. Mekh.* **24**, 2, 271 - 276, 1960.

37. OSTROVSKII, G.M. *Avtomatika Telemekh.* **23**, 10, 1284 - 1289, 1962.

38. CARROLL, Ch.W. The created response surface technique for optimizing non-linear restrained systems, *Operat. Res.* 9, 2, 169 - 183, 1961.

39. FIACCO, A.V. and McCORMICK, G.P. The sequential unconstrained minimization technique for non-linear programming, *Man. Sci.* 10, 2, 360 - 366; 4, 601 - 618, 1964.

40. PIETRZYKOWSKI, T. Applications of the steepest ascent method to concave programming, *Inform. Process 1962*, Amsterdam, 185 - 189, 1963.

41. POMENTALE, T. A new method for solving conditioned maxima problems. *J. math. Analyt. Appl.* 10, 1, 216 - 220, 1965.

42. BUTLER, T. and MARTIN, A.V. On a method of Courant for minimizing functionals, *J. Math. Phys.* **41**, 3, 291 - 299, 1962.

43. GOLDSTEIN, A.A. and KRIPKE, B.R. Mathematical programming by minimizing differentiable functions, *Num. Math.* 6, 1, 47 - 48, 1964.

44. BUTKOVSKII, A.G. *Optimal control theory for systems with distributed parameters* (Teoriya optimal'nogo upravleniya sistemami s raspredelennymi parametrami), "Nauka" Moscow, 1965.

45. MIKHLIN, S.G. *Numerical realization of variational methods* (Chislennaya realizatsiya variatsionnykh metodov), "Nauka", Moscow, 1966.