NUMERICAL METHODS IN EXTREMAL PROBLEMS

B. N. Pshenichny
and
Yu. M. Danilin

Б. Н. Пшеничный

Ю. М. Данилин

ЧИСЛЕННЫЕ МЕТОДЫ В ЭКСТРЕМАЛЬНЫХ ЗАДАЧАХ

ИЗДАТЕЛЬСТВО «НАУКА»
ГЛАВНАЯ РЕДАКЦИЯ
ФИЗИКО-МАТЕМАТИЧЕСКОЙ ЛИТЕРАТУРЫ
МОСКВА


ИЗДАТЕЛЬСТВО
МИР

NUMERICAL METHODS IN EXTREMAL PROBLEMS

B. N. Pshenichny and Yu. M. Danilin

Translated from the Russian
by V. Zhitomirsky, D. Sc. (Eng.)

MIR
PUBLISHERS
MOSCOW

First published 1978

Revised from the 1975
Russian edition

На английском языке

CONTENTS

PREFACE

Computational methods of solving extremal problems developed very intensively in recent years.

The lists of the literature on these subjects contain at present hundreds of items. This interest in the development of computational methods is not casual. It reflects the important role played by the finding of extrema in diverse applied problems. The problem of an effective minimization of a function with different constraints on the variables is the subject matter of this book.

It should be stressed from the very beginning that recent years have brought changes in the requirements to be met by new computational algorithms. Some ten or fifteen years ago any new algorithm for solving a minimization problem was noticed with interest, but now only the construction of a new algorithm is insufficient. It is now necessary to show in what respect it is better than the existing ones. Thus there arises the problem of comparing the effectiveness of different algorithms. Unfortunately this problem has no simple solution. This is due to the necessity of choosing a criterion of effectiveness and the criteria may be diverse. For instance, we can take as a criterion of effectiveness the accuracy of the result obtained, the time required for computing, the necessary storing capacity of the computer, etc. Also it is often necessary to use rather contradictory criteria in estimating an algorithm.

In selecting algorithms to be included in this book, the authors based their choice on the criterion of accuracy of the result and the rate of convergence of the iterative process. However, even with this limiting condition it is not possible to order all the algorithms in one and only one way and tell which of them is better or worse than another. The reason is that the estimate of the rate of convergence is not made for a particular problem, rather it is applied to a class of problems. Therefore an algorithm which is poor as applied to a broad class of problems can prove effective on a narrower one. This makes it necessary for the calculator to keep a large reserve of algorithms and to apply them depending on the problem to be solved.

It is important to know what ensures a fast rate of convergence of the algorithm. In practice, even the calculating of the first derivative of a function quite often involves certain difficulties; these become insurmountable when trying to calculate the second derivative.

Therefore special stress is laid on the description of the algorithms that require the finding only of the first derivative or only of the value of the function.

In describing the computational methods we consider only the finite dimensional case. This is due to two reasons. First, in using a computer for calculations, the problem is to be approximated anyway by a finite dimensional one. Secondly, most of the known algorithms are comparatively simply generalized for the minimization of functionals without essential changes. This approach made it possible to make the book easily understood by a broad circle of readers, since in order to grasp most of the results described only a knowledge of the principles of mathematical analysis and linear algebra is required.

To avoid the necessity of frequent cross-referencing, not many references are given in the text. Short bibliographic notes follow some of the chapters. The authors did not attempt to comprise all the literature on the questions treated, this being simply impossible because of its vastness. This is why the list of literature given at the end of the book includes only papers and monographs directly used in writing this book.

It should be noted that the authors have not discussed the methods of solving a broad and important class of noncorrect extremal problems, which are treated in the works of A. N. Tikhonov and his followers. The authors have but slightly touched the solving of optimal control problems. These problems have been studied from various points of view and the methods for their solution are given in N. N. Moiseev's monograph *Numerical Methods in the Theory of Optimal Systems.*

The algorithms set forth below are iterative in character. This means that we can construct a finite or infinite sequence of points x_k, $k = 0, 1 \ldots$ which is said to converge to the solving of a minimization problem.

The points of the sequence are related by the equation

$$x_{k+1} = x_k + \alpha_k p_k$$

where p_k is the vector of shift from point x_k and α_k is a step along the direction of p_k. Therefore the description of any of the algorithms given below consists in establishing the method of choosing the vector p_k and the length of the step α_k. It should be noted that the method of choosing the vector p_k determines the general rate of convergence of the process and the method of choosing α_k has an important influence on the amount of calculations at each iteration. Therefore the authors' aim was to give in all cases of choosing α_k a method, such that the required value of α_k could be found after a finite number of iterations without affecting the general rate of convergence.

Let us briefly review the estimates of the rate of convergence, which are in most cases used in this book.

We say that a sequence $\{x_k\}$ converges to point x_* at a *linear rate* or at the *rate of geometrical progression* (with the ratio q) if from a certain k the inequality $\| x_{k+1} - x_* \| \leqslant q \| x_k - x_* \|$ where $0 < q < 1$, is satisfied. If the inequality $\| x_{k+1} - x_* \| \leqslant q_k \| x_k - x_* \|$ is satisfied, where $q_k \to 0$ with $k \to \infty$, we say that the *rate of convergence of the sequence* $\{x_k\}$ *is superlinear*, or *faster than the rate of convergence of any geometric progression*. If $q_k \leqslant C \| x_k - x_* \| \to 0$, then $\| x_{k+1} - x_* \| \leqslant C \| x_k - x_* \|^2$. This estimate is a characteristic of the *quadratic rate of convergence*.

The above estimates will occur in this book also in several other equivalent forms.

Some remarks on the notations used.

As mentioned before, the subject is treated for the case of an n-dimensional vector space which will be denoted by E^n. The vectors will be denoted by lower-case letters x, y, z, etc. and their components by using superscripts so that x^i is the i-th component of vector x. The subscripts denote the elements of a sequence. Matrices are denoted by capital letters A, B, C etc. An asterisk as upper index denotes transposition, i.e. A^* is the transposed matrix A. As a rule vector x means a column-vector so that x^* denotes a row-vector. The scalar product of two vectors is denoted by (x, y), i.e.

$$(x, y) = \sum_{i=1}^{n} x^i y^i.$$

The norm of the vector is understood to be its Euclidean norm, unless otherwise specified:

$$\| x \| = \sqrt{(x, x)}.$$

In conclusion, the authors express their sincere gratitude to G. E. Lybarskaya, L. A. Sobolenko, E. I. Boguslavskaya and V. M. Panin for the invaluable assistance in preparing this book.

Chapter I (except Sec. 5 and partly Sec. 2) and Chap. III (except Sec. 9 and partly Sec. 3) have been written by B. N. Pshenichny.

Chapter II, the third and the fourth subsections of Sec. 2 and the fifth and sixth subsections of Sec. 3, and Sec. 9 of Chap. III have been written by Yu. M. Danilin.

CHAPTER I

INTRODUCTION TO THE THEORY OF MATHEMATICAL PROGRAMMING

This chapter describes some facts from the theory of convex sets and the necessary conditions of the extrema; these facts are necessary for understanding the matter set forth in subsequent chapters.

1. CONVEX SETS

In this section we consider the basic properties of convex sets in an n-dimensional Euclidean space.

Definition. Separation Theorem

Definition 1.1. *A set of points X in E^n is called convex if together with any $x_1, x_2 \in X$ it contains also all points of the form:*

$$x = \lambda x_1 + (1 - \lambda) x_2, \quad 0 \leqslant \lambda \leqslant 1.$$

In geometrical terms this means that if the end points of a segment belong to a convex set X then the whole segment belongs to the set too.

Lemma 1.1. *The following statements hold:*

(1) *The intersection of any number of convex sets is convex.*

(2) *If $x_i \in X$, $i = 1, \ldots, m$, then with any λ_i, $i = 1, \ldots, m$ such that $\sum_{i=1}^{m} \lambda_i = 1$, $\lambda_i \geqslant 0$, a point $x = \sum_{i=1}^{m} \lambda_i x_i$ belongs to X.*

The following theorem and its corollaries are the basic tools using which it is possible to obtain results characterising various properties of convex sets.

Theorem 1.1. *Let X be a convex set, and $\overline{X}$ its closure. If point x_0 does not belong to $\overline{X}$, then there exist a vector $a \in E^n$, $a \neq 0$, and a number $\varepsilon > 0$ such that for all $x \in X$*

$$(a, x) \leqslant (a, x_0) - \varepsilon.$$

Proof. $\overline{X}$ is a closed set, by definition. Let us show that it is convex. Indeed, if $x \in \overline{X}$, then there is a sequence $\{x_k\}$, $k = 1, \ldots$, such that $x_k \in X$, $x_k \to x$. Now let $x, y \in \overline{X}$, $0 \leqslant \lambda \leqslant 1$. Let us prove that $\lambda x + (1 - \lambda)y \in \overline{X}$. Since X is a convex set, it follows from $x_k, y_k \in X$, $x_k \to x$, $y_k \to y$ that

$$\lambda x_k + (1 - \lambda)\, y_k \in X,$$
$$\lambda x_k + (1 - \lambda)\, y_k \to \lambda x + (1 - \lambda)\, y.$$

This means that $\lambda x + (1 - \lambda)\, y \in \overline{X}$, i.e. $\overline{X}$ is convex.
Let us take a point $y \in \overline{X}$ whose distance from x_0 is the least, i.e.

$$\| x - x_0 \| \geqslant \| y - x_0 \|, \quad x \in \overline{X}.$$

Since $\overline{X}$ is convex for all $x \in X$ and $0 \leqslant \lambda \leqslant 1$, we have

$$\lambda x + (1 - \lambda)\, y = y + \lambda\,(x - y) \in X.$$

Therefore

$$\| \lambda x + (1 - \lambda)\, y - x_0 \|^2 = \| y - x_0 + \lambda\,(x - y) \|^2$$
$$= (y - x_0 + \lambda\,(x - y),\ y - x_0 + \lambda\,(x - y))$$
$$= (y - x_0,\ y - x_0) + 2\lambda\,(y - x_0,\ x - y) + \lambda^2\,(x - y,\ x - y)$$
$$= \| y - x_0 \|^2 + 2\lambda\,(y - x_0,\ x - y) + \lambda^2 \| x - y \|^2 \geqslant \| y - x_0 \|^2.$$

The last inequality holds for any λ, varying between zero and unity. Simplifying it we obtain

$$2\,(y - x_0,\ x - y) + \lambda \| x - y \|^2 \geqslant 0;$$

hence with $\lambda = 0$

$$(y - x_0,\ x - y) \geqslant 0.$$

Let $a = x_0 - y$. The last inequality can then be written in the form $(a, x) \leqslant (a, y)$. But

$$(a, y) = (a, x_0) - (a, x_0 - y) = (a, x_0) - \| a \|^2.$$

Setting $\varepsilon = \| a \|^2$, we finally obtain

$$(a, x) \leqslant (a, x_0) - \varepsilon.$$

This inequality holds for any $x \in \overline{X}$. Besides $\varepsilon > 0$ as $x_0 \overline{\in} \overline{X}$ and consequently $y \neq x_0$. Therefore

$$\varepsilon = \| a \|^2 = \| x_0 - y \|^2 > 0.$$

Q.E.D.

Remark. In proving theorem 1.1. we have proved at the same time that the closure of a convex set is convex too. As a simple exercise the reader can prove that the set of interior points of a convex set is convex too.

Corollary 1.1. *Let X be a convex set and x_0 the frontier point of X. Then there is a vector $a \neq 0$ such that*

$$(a, x) \leqslant (a, x_0), \quad x \in X.$$

Corollary 1.2. *If X and Y are convex sets that do not intersect, then there is a vector $a \neq 0$ such that*

$$(a, x) \leqslant (a, y), \quad x \in X, y \in Y.$$

Corollary 1.3. *If X and Y are closed convex sets which do not intersect and one of them is bounded, then there exist a vector $a \neq 0$ and a number $\varepsilon > 0$ such that*

$$(a, x) \leqslant (a, y) - \varepsilon, \quad x \in X, y \in Y.$$

Convex Cones

Definition 1.2. *A set K is called a convex cone if the set is convex and together with every point $x \in K$ it contains all points λx with $\lambda > 0$.*

It is clear that if $x, y \in K$ then $x + y \in K$. In fact, since K is a convex set, point $\frac{1}{2}x + \frac{1}{2}y$ belongs to K. But

$$x + y = 2\left(\frac{1}{2}x + \frac{1}{2}y\right),$$

whence $x + y \in K$ by the definition of a cone. The most important properties of cones are formulated in terms which establish the relation between the original cone and the cone that is its conjugate or dual.

Definition 1.3. *Let K be a convex cone. The set of all vectors $y \in E^n$ satisfying for any $x \in K$ the inequality $(x, y) \geqslant 0$ is called a conjugate cone and denoted by K^*.*

An elementary check shows that K^* is also a convex cone.

Lemma 1.2. *K^* is a closed convex cone.*

Lemma 1.3. *Let K be a convex cone. Then $x_0 \in \overline{K}$ if and only if $(x_0, y) \geqslant 0$ for all $y \in K^*$. If K is closed, then*

$$(K^*)^* = K.$$

Proof. It is evident that if $x_0 \in \overline{K}$, then $(x_0, y) \geqslant 0$ for all $y \in K^*$. Suppose it is false. Let $(x_0, y) \geqslant 0$ for any $y \in K^*$, but $x_0 \notin \overline{K}$.

Since K is a closed convex set and using theorem 1.1, we can assert that there is a vector a such that

$$(a, x_0) \leqslant (a, x) - \varepsilon, \quad x \in \overline{K}.$$

Now a closed cone K always contains point 0. Therefore in particular

$$(a, x_0) \leqslant -\varepsilon. \tag{1.1}$$

On the other hand

$$(a, x) \geqslant 0, \quad x \in \overline{K}. \tag{1.2}$$

Indeed, if for a certain $x_1 \in K$ $(a, x_1) < 0$, then since $\lambda x_1 \in K$ with $\lambda > 0$

$$(a, x_0) \leqslant \lambda (a, x_1) - \varepsilon$$

and the last inequality must be valid for any λ; this is impossible if $(a, x_1) < 0$. Thus (1.2) is valid and consequently $a \in K^*$. Then $(a, x_0) \geqslant 0$ and this contradicts (1.1). This proves the first part of the lemma.

Let us now prove its second part. If $x \in K$, then $(x, y) \geqslant 0$ for all $y \in K^*$, by definition, and therefore $x \in (K^*)^*$, $K \subset (K^*)^*$. Conversely, by definition, $x \in (K^*)^*$ if and only if $(x, y) \geqslant 0$ with any $y \in K^*$. However, it was proved above that in this case $x \in K$, i.e. $(K^*)^* \subset K$. Thus $(K^*)^* = K$. Q.E.D.

Polyhedral cones are an important class of cones encountered in the theory of linear programming.

Definition 1.4. *A cone K is called polyhedral if there exists a finite set of n-dimensional vectors a_i, $i = 1, \ldots, m$ such that with $x \in K$ the expansion*

$$x = \sum_{i=1}^{m} \lambda_i a_i, \quad \lambda_i \geqslant 0, \quad i = 1, \ldots, m \tag{1.3}$$

is valid and conversely (1.3) *implies that $x \in K$.*

Thus a polyhedral cone K is a set of points which can be represented in the form (1.3). A given point $x \in K$ in the form (1.3), speaking generally, is represented not uniquely.

Lemma 1.4. *Let $x \in K$, K being a polyhedral cone. Then there is such an expansion of x in vectors a_i with nonnegative coefficients λ_i, that the number of indices i for which $\lambda_i > 0$ does not exceed n, the number of dimensions of the space; the vectors a_i corresponding to nonzero λ_i are linearly independent.*

Proof. Let $x \in K$, i.e. $x = \sum_{i=1}^{m} \lambda_i a_i$, and $\mathcal{J}$ be the set of those indices i such that $\lambda_i > 0$. Suppose that the number of elements in $\mathcal{J}$ is greater than n, or does not exceed n, but the vectors a_i, $i \in \mathcal{J}$, are

linearly dependent. Since more than n linearly independent vectors cannot exist in an n-dimensional space, there are coefficients α_i, not all zero, such that $\sum_{i\in\mathcal{J}} \alpha_i a_i = 0$. Besides, by definition of $\mathcal{J}$, $\lambda_i = 0$ if $i \notin \mathcal{J}$ and so

$$x = \sum_{i\in\mathcal{J}} \lambda_i a_i, \quad \lambda_i > 0, \quad i \in \mathcal{J}.$$

Subtracting from this relation the preceding one multiplied by ε, we obtain

$$x = \sum_{i\in\mathcal{J}} (\lambda_i - \varepsilon\alpha_i)\, a_i.$$

Without loss of generality we can take that $\alpha_i > 0$, for some $i \in \mathcal{J}$. Setting $\varepsilon_0 = \min_{i\in\mathcal{J},\, \alpha_i>0} \frac{\lambda_i}{\alpha_i}$ and $\bar{\lambda}_i = \lambda_i - \varepsilon_0\alpha_i$, we have

$$x = \sum_{i\in\mathcal{J}} \bar{\lambda}_i a_i$$

where $\bar{\lambda}_i \geqslant 0$ and for one i at least $\bar{\lambda}_i = 0$.

Thus we have obtained an expansion of x in vectors a_i with non-negative coefficients; however the number of strictly positive coefficients has been diminished.

This process can now be applied further until the number of non-zero coefficients becomes less than n or equal to n and vectors a_i for which $\lambda_i > 0$ become linearly independent. Since we have a process of diminishing a whole number, this process obviously cannot be continued infinitely and after a certain number of steps we shall get an expansion which satisfies the conditions of our lemma.

Lemma 1.5. *A polyhedral cone is closed.*

Lemma 1.6. *Let the cone K be defined by a system of linear inequalities*

$$(a_i,\ x) \geqslant 0, \quad i = 1, \ldots, m$$

where $a_i \in E^n$. Then the conjugate cone K^* is a polyhedral cone and consists of points y, which can be presented in the form

$$y = \sum_{i=1}^{m} \lambda_i a_i, \quad \lambda_i \geqslant 0, \quad i = 1, \ldots, m.$$

Proof. Let us consider the cone

$$\widetilde{K} = \{y : y = \sum_{i=1}^{m} \lambda_i a_i, \quad \lambda_i \geqslant 0, \quad i = 1, \ldots, m\}.$$