IIASA COLLABORATIVE PROCEEDINGS SERIES

CP-82-S8

PROGRESS IN NONDIFFERENTIABLE OPTIMIZATION

IIASA COLLABORATIVE PROCEEDINGS SERIES

CP–81–S1 LARGE-SCALE LINEAR PROGRAMMING
Proceedings of an IIASA Workshop
G.B. Dantzig, M.A.H. Dempster, and M.J. Kallio, *Editors*

CP–81–S2 THE SHINKANSEN PROGRAM: TRANSPORTATION, RAILWAY, ENVIRONMENTAL, REGIONAL, AND NATIONAL DEVELOPMENT ISSUES
A. Straszak, *Editor*

CP–82–S1 HUMAN SETTLEMENT SYSTEMS: SPATIAL PATTERNS AND TRENDS
Selected Papers from an IIASA Conference
T. Kawashima and P. Korcelli, *Editors*

CP–82–S2 RISK: A SEMINAR SERIES
H. Kunreuther, *Editor*

CP–82–S3 THE OPERATION OF MULTIPLE RESERVOIR SYSTEMS
Proceedings of an International Workshop, Jodłowy Dwor, Poland
Z. Kaczmarek and J. Kindler, *Editors*

CP–82–S4 NONPOINT NITRATE POLLUTION OF MUNICIPAL WATER SUPPLY SOURCES: ISSUES OF ANALYSIS AND CONTROL
Proceedings of an IIASA Task Force Meeting
K.-H. Zwirnmann, *Editor*

CP–82–S5 MODELING AGRICULTURAL–ENVIRONMENTAL PROCESSES IN CROP PRODUCTION
Proceedings of an IIASA Task Force Meeting
G. Golubev and I. Shvytov, *Editors*

CP–82–S6 LIQUEFIED ENERGY GASES FACILITY SITING: INTERNATIONAL COMPARISONS
H. Kunreuther, J. Linnerooth, and R. Starnes, *Editors*

CP–82–S7 ENVIRONMENTAL ASPECTS IN GLOBAL MODELING
Proceedings of the 7th IIASA Symposium on Global Modeling
G. Bruckmann, *Editor*

CP–82–S8 PROGRESS IN NONDIFFERENTIABLE OPTIMIZATION
E.A. Nurminski, *Editor*

PROGRESS IN NONDIFFERENTIABLE OPTIMIZATION

E.A. Nurminski
Editor

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
Laxenburg, Austria
1982

International Standard Book Number 3-7045-0050-X

Collaborative papers in this *Special* series sometimes report work done at the International Institute for Applied Systems Analysis and sometimes work done elsewhere. They are reviewed at IIASA, but receive only limited external review, and are issued after limited editorial attention. The views or opinions they express do not necessarily represent those of the Institute, its National Member Organizations, or other organizations supporting the work.

PREFACE

The System and Decision Sciences Area plays a dual role at the International Institute for Applied Systems Analysis (IIASA), both providing methodological assistance to other research groups and carrying out fundamental research on new methods and models for use in applied systems analysis. The Optimization Task of the System and Decision Sciences (SDS) Area contributes in both of these fields by helping to solve optimization problems arising in applied areas and also by providing an international forum for large-scale and dynamic linear programming and for nondifferentiable optimization. In the second of these capacities, SDS sponsors annual task-force meetings on various aspects of optimization, bringing together research workers from both East and West to discuss advances in methodology and implementation.

This volume grew out of the second meeting on nondifferentiable optimization, a field whose most important applications lie in treating problems of decision-making under uncertainty. Many important advances were made between the first meeting in 1977 and the second in 1978--new results were obtained in the theory of optimality conditions, and there was more understanding of the relationships between various classes of nondifferentiable functions. All of these new developments were discussed

at the meeting, the reports presented by the participants covering the theory of generalized differentiability, optimality conditions, and the numerical testing and applications of algorithms.

After the meeting the participants prepared extended versions of their contributions; these revised papers form the core of this volume, which also contains a bibliography of over 300 references to published work on nondifferentiable optimization, prepared by the Editor.

It is hoped that this volume will be of use to those already working in nondifferentiable optimization and will stimulate the interest of those currently unfamiliar with this new and rapidly expanding field.

CONTENTS

INTRODUCTION

E. Nurminski (Editor)
International Institute for Applied Systems Analysis,
Laxenburg, Austria

IIASA's interest in nondifferentiable optimization (NDO) is based on the great practical value of NDO techniques. This new field of mathematical programming provides specialists in applied areas with tools for solving non-traditional problems arising in their work and with new approaches and ideas for treating traditional problems. Nondifferentiable optimization is concerned with the new type of optimal decision problems which have objectives and constraints resulting from the behavior of different complex subsystems, the solutions of auxiliary extremum problems, and so on. A common feature of these problems is that the objectives and constraints inevitably have poor analytical properties.

Good analytical properties are essential both for performing comprehensive theoretical analysis and for producing efficient computational methods which are acceptable in practice. The most important of these analytical properties are the existence and continuity of derivatives of various orders.

Unfortunately, derivatives are very sensitive to manipulation--many standard operations and representations used in economics or operations research destroy the property of differentiability.

As an example, consider the piecewise method of representing the response function for different ranges of variables. This type of approximation often has a discontinuity in the first-order or higher-order derivatives at the boundary between consecutive intervals.

The incorporation of uncertainty in environmental parameters or systems characteristics by means of the minimax principle provides another example. In this case, the resulting criteria and constraints will almost certainly have discontinuous derivatives, regardless of how well-behaved the initial equations may have been.

Many of the procedures used in multiobjective optimization create an auxiliary nondifferentiable problem in the search for a compromise solution, and game-theoretical approaches, equilibrium formulations and decomposition are also important sources of nondifferentiable problems.

The absence of derivatives leads to many theoretical difficulties and numerous practical failures in solving certain problems in operations research and systems analysis. The lack of continuous derivatives makes it very difficult to predict with a good degree of accuracy the effect of small changes in control variables--and this hinders the performance of many numerical algorithms.

These, then, were the main motivations for the study of nondifferentiable functions--that is, functions for which derivatives do not exist in the traditional sense of the word. The problem was approached from many angles: these included the study of generalized differentiability and properties of generalized derivatives, the analysis of extremum problems and optimality conditions, and the development of computer algorithms.

The first IIASA meeting on nondifferentiable optimization was concerned mainly with the development of algorithms. It summarized past developments in both East and West, outlined the fields of application, provided test examples, and gave a comprehensive bibliography compiled by participants and other contributors throughout the world.

Since then many important results have been obtained in the general theory of differentiability, and these and other results are discussed in the eight papers contained in this volume.

In the first paper, Yu.M. Ermoliev considers the fundamental connection between nondifferentiable and stochastic optimization, and the various questions raised by this relationship.

J.L. Goffin's paper is concerned with acceleration in the relaxation method for linear inequalities. This is closely related to acceleration in subgradient optimization procedures and it is shown that there are certain features in the performance of these methods which are not explained by current theory. Experiments conducted by the author show that subgradient optimization techniques perform better than existing theory would predict on the basis of the worst-case estimates. Using arguments taken from the theory of successive over-relaxation, rates of convergence are shown to improve in selected experiments.

The paper by C. Lemaréchal is devoted to numerical experiments in which various types of algorithms are applied to a number of test problems. The algorithms considered range from those which have a good performance in smooth cases to the robust ellipsoid algorithm.

The paper contributed by R. Mifflin describes an algorithm for the minimization of certain semismooth functions defined by the author. These functions are quite general and are likely to cover most practical applications. The algorithm combines the idea of the cutting plane method with a quadratic term which the author suggests as a second-order approximation of the Lagrangian associated with the optimal multipliers of the subproblems.

A number of algorithms for convex optimization are based on the idea of ε-subgradients, which have certain remarkable theoretical properties. The paper by E. Nurminski discusses general aspects of the use of ε-subgradients in nondifferentiable convex optimization.

The latest results in the theory of optimality conditions and differentiability were presented at the meeting by

R.T. Rockafellar. His paper on this subject deals with the refinement of the properties of generalized gradients for functions which satisfy regularity requirements in addition to the Lipschitz condition. Additional properties of this type make it possible to establish useful connections between the directional differentiability of nondifferentiable functions and the monotonicity properties of subdifferential mappings.

The links between nondifferentiable optimization and structured decision-making problems are considered in the paper by A. Ruszczyński. A two-stage decision problem is shown to give rise to nondifferentiable problems with specific types of nondifferentiability for which simple subgradient-type algorithms are proposed. An important feature of this approach is that it also allows random factors to be included in the formulation of the problem, and this makes it more realistic in terms of applications.

A. Wierzbicki discusses the theoretical and computational possibilities connected with the use of augmented Lagrangian functions in the last paper of this volume.

METHODS OF NONDIFFERENTIABLE AND STOCHASTIC OPTIMIZATION AND THEIR APPLICATIONS

Yu. M. Ermoliev
International Institute for Applied Systems Analysis,
Laxenburg, Austria

1. INTRODUCTION

Optimization methods are of great practical importance in systems analysis. They allow us to find the best behaviour of a system, determine the optimal structure and compute the optimal parameters of the control system, etc. The development of nondifferentiable and stochastic optimization allows us to state, and effectively solve, new complex optimization problems which are impossible to solve by classical optimization methods.

The term nondifferentiable optimization (NDO) was introduced by Balinski and Wolfe [1975] for extremal problems with an objective function and constraints that are continuous but have no continuous derivatives. This term is now also used for problems with discontinuous functions, although in these cases it might be better to use the terms nonsmooth optimization (NSO) or, in particular, discontinuous optimization (DCO).

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

Nondifferentiable and stochastic optimization are natural developments of classic optimization methods. Some important

classes of nondifferentiable and stochastic optimization problems are well-known and have been investigated long ago: problems of Chebyshev approximations, game theory and mathematical statistics. It should also be noted that, from the conventional viewpoint, there is no major difference between functions with continuous gradients which change rapidly and functions with discontinuous gradients. Each of the above mentioned classes was investigated by its own "homemade" methods. General approaches (extremum conditions, numerical methods) were developed at the beginning of the 1960's. The main purpose of this article is to review briefly some important applications and nondescent procedures of nondifferentiable and stochastic optimization. Clearly, the interests of the author have influenced the content of this article.

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

2. OPTIMIZATION OF LARGE-SCALE SYSTEMS

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

$$\sum_{j=1}^{n} a_{0j}x_j = \min \tag{1}$$

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

$$x_j \geq 0 \quad , \quad j = \overline{1,n} \tag{3}$$

Here Y is a given discrete set. For example, the use of duality theory for solving discrete programming problems [Balinski and Wolfe 1975; Lasdon 1970] necessitates the minimization of

nondifferentiable functions of the kind

$$f(x) = \max_{y \in Y} \left(\sum_{j=1}^{n} a_j(y) x_j - b(y) \right) , \qquad (4)$$

where Y is some discrete set. This problem reduces to problems of the kind (1) - (3).

Clearly in this case the total number of constraints may be equal to 100^{100}. However, these constraints have a form which does not impose heavy demands on the computer core and one can try to find their solution with the finite methods of linear programming. But the number of vertices of the feasible polyhedral set for such problems is so large that the application of the conventional simplex method, or its variants, yields very small steps at each iteration and consequently very slow convergence. Moreover, the known finite methods are not robust against computational errors. The use of nondifferentiable optimization has made it possible to develop easily implementable iterative decomposition schemes of the gradient type. These approaches do not use the basic solution of the linear programming problem which enables one to start the computational process from any point, and leads to computational stability. Nondifferentiable decomposition techniques (see, for instance, Shor [1967] and Ermoliev and Ermolieva [1973]) are based on the following ideas.

Let the linear programming problem have the form

$$\langle c,x \rangle + \langle d,y \rangle = \min$$

$$Ax + Dy \geq b$$

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

We assume that for fixed x it is easy to find a solution $y(x)$ with respect to y. For example, the matrix D may have a block diagonal structure, with x being the connecting variable. The main difficulty here is to find the value x^* for the optimal

solution $(x^*, y(x^*))$. The search for x^* is equivalent to the minimization of the nonsmooth function

$$f(x) = (c,x) + \min_{\substack{Dy \geq b - Ax \\ y \geq 0}} (d,y) = (c,x) + (d,y(x)) \quad . \qquad (5)$$

Another approach is to consider the dual problem:

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

$$uD \leq d \quad ,$$

$$uA \leq c \quad ,$$

$$u \geq 0 \quad .$$

Let us examine the Lagrangian function

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

subject to constraints

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

In this case the search for x^* is equivalent to the minimization of the nonsmooth function

$$f(x) = (c,x) + \max_{\substack{UD \leq d \\ u \geq 0}} (u, b - Ax) \text{ for } x \geq 0 \quad . \qquad (6)$$

The well-known Dantzig-Wolfe decomposition is also based on this principle. A subproblem of minimization with respect to variables u, subject to

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

is solved easily because the matrix D is assumed to have a special structure.

A parametric decomposition method [Ermoliev and Ermolieva 1973] reduces linear programming problems, which may not have block diagonal structure, to nondifferentiable optimization problems by introducing additional parameters. In this case, there is the possibility of splitting the linear programming problem into arbitrary parts, in particular, of singling out subproblems which correspond to blocks of nonzero elements in the constraint matrix.

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

$$y_3 = \min \tag{7}$$

$$\begin{aligned} &\boxed{a_{11}y_1 + a_{12}y_2} + \boxed{a_{13}y_3} \le b_1 \ , \\ &a_{21}y_1 + \boxed{a_{22}y_2} + \boxed{a_{23}y_3} \le b_2 \ , \end{aligned} \tag{8}$$

where

$$b_1 \ge 0 \ , \quad b_2 \ge 0 \ , \quad y_j \ge 0 \ , \quad j = 1,2,3 \ .$$

Let it be necessary to cut this problem, for example, into three parts as shown in constraints (8).

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

$$y_3 = \min$$

$$\begin{aligned} &a_{11}y_1 + a_{22}y_2 \le x_{11} \ , \quad a_{13}y_3 \le x_{12} \ , \\ &a_{21}y_1 \le x_{21} \ , \quad a_{23}y_3 \le x_{23} \ , \\ &a_{22}y_2 \le x_{22} \ . \end{aligned} \tag{9}$$

This problem produces three subproblems with the desired structure. If the minimum value y_3 is denoted as $f(x)$ then it is easy to show that solving the problem (7)-(8) is equivalent to solving (9) for the value of x which minimizes the nondifferentiable function $f(x)$ under the constraints:

$$x_{11} + x_{12} \leq b_1 \quad ,$$

$$x_{21} + x_{22} + x_{23} \leq b_2 \quad ,$$

$$x_{ij} \geq 0 \quad , \quad i = 1,2; \quad j = 1,2,3 \tag{10}$$

3. MINIMAX PROBLEMS, PROBLEMS OF GAME THEORY

The problem (4) is the simplest minimax problem. More general deterministic minimax problems are formulated as follows [Danskin 1967; Demyanov and Malozemov 1974].

For a given function

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

it is necessary to minimize

$$f(x) = \max_{y \in Y} g(x,y) = g(x,y(x)) \tag{11}$$

for $x \in X$. Independently of the smoothness of $g(x,y)$ the function $f(x)$, as a rule, has no continuous derivatives. A particular class of minimax problems thus arises in approximation theory e.g., in problems of the best Chebyshev approximation, in approximation by splines, and in mathematical statistics.

The solution of a system of inequalities

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

for $g(x,y) = d_y(x)$, $y \in Y = \{1,2,\ldots,m\}$ can also be found by minimization of the function (11). The solution of the general

problem of nonlinear programming,

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

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

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

In game theory more complex problems arise in the minimization of the function

$$f(x) = g(x,y(x)) \tag{12}$$

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

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

Independently of the smoothness of the functions $g(x,y)$, $h(x,y)$ the function $f(x)$ in any given case may have no continuous derivatives and may be discontinuous at any point. For $h(x,y) = x \cdot y$, $g(x,y) = x + y$, $Y = |-1,1|$, we obtain

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

The function $h(x,y(x)) = xy(x) = |x|$ is continuous, but does not have continuous derivatives at the point $x = 0$. The function $f(x) = x + y(x)$ may be regarded as discontinuous at $x = 0$.

4. OPTIMIZATION OF PROBABILISTIC SYSTEMS

Taking the influence of uncertain random factors into account, even in the simplest extremal problems, leads to

complex extremal problems with nonsmooth functions. For example, for deterministic ω a set of solutions to the inequality

$$\omega x \leq 1 \quad ,$$

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

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

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

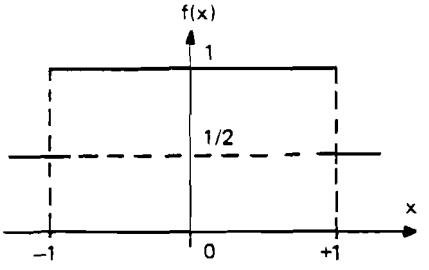

Figure 1.

A quite general stochastic programming (stochastic optimization) problem can be formulated as follows [Ermoliev and Nekrilova 1967, Ermoliev 1976]:

$$\min \{F^0(x) \mid F^i(x) \leq 0, i = \overline{1,m}, x \in X\} \quad , \tag{13}$$

where

$$F^\nu(x) = Ef^\nu(x,\omega) = \int f^\nu(x,\omega) P(d\omega) \quad , \quad \nu = \overline{0,m} \quad . \tag{14}$$

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

$$P\{g^{i}(x,\omega) \leq 0\} \geq p_{i} \quad , \quad i = \overline{1,m}$$

become constraints of the type (13) - (14) if we assume that

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

The problem (13) - (14) is more difficult than the conventional nonlinear programming problem. The main difficulty, besides the nondifferentiability, is connected with the condition (14). As a rule, it is practically impossible to compute the precise values of the integrals (14) and therefore one cannot calculate the precise values of the function $F^{\nu}(x)$. For example, it is only rarely possible for special kinds of distributions and functions $g^{i}(x,\omega)$ to find the expression $P\{g^{i}(x,\omega) \leq 0\}$ as a function of x. Usually only values of the random quantities $f^{i}(x,\omega)$ are available - not values of $F^{\nu}(x)$. To determine whether the point x satisfies the constraints

$$F^{i}(x) = Ef^{i}(x,\omega) \leq 0 \quad , \quad i = \overline{1,m}$$

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

5. ON EXTREMUM CONDITIONS

The difference between nondifferentiable and stochastic optimization problems on the one hand, and the classic problem of deterministic optimization on the other, is apparent in the optimality conditions. If $f(x)$ is a convex differentiable function then the necessary and sufficient conditions for the

minimum have the form:

$$f_x(x) = 0 \quad , \tag{15}$$

where

$$f_x(x) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right) \quad .$$

In the nondifferentiable case this condition transforms into the requirement (Figure 2)

$$0 \in \{\hat{f}_x(x)\} \tag{16}$$

where

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

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

$$f(y) - f(x) \geq \langle \hat{f}_x(x), y-x \rangle \quad , \quad \forall y \quad . \tag{17}$$

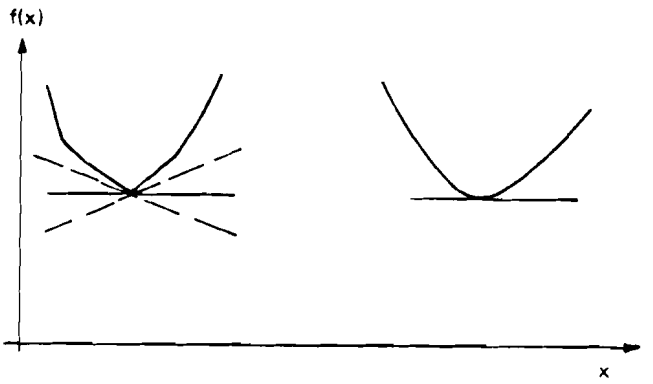

Figure 2.

It should be noted that the notation $\hat{f}_x(x)$ for a subgradient used here is convenient in cases where a function depends on several groups of variables and the subgradient is to be taken with respect to one of them.

The complexity of nondifferentiable optimization problems results from the impossibility of using (16) in practice to discover whether a specific point x is a point corresponding to the minimum of f(x).

This discussion requires one to test whether the 0-vector belongs to the set $\{\hat{f}_x(x)\}$, which usually has no constructive description. A further complication arises from checking the conditions (15) and (16) in stochastic optimization problems. Generally speaking, even checking the conditions (15) in the stochastic case leads to a verification of the statistical hypothesis that for fixed x the mathematical expectation of the random vector $f_x(x,\omega)$ is 0, that is,

$$Ef_x(x,\omega) = 0 \quad .$$

In such cases, the development of direct numerical procedures for finding optimal solutions becomes extremely important.

6. DETERMINISTIC METHODS OF NONDIFFERENTIABLE OPTIMIZATION

There are two different classes of nondifferentiable optimization methods: the nondescent methods which started their development in the early 60's at the Institute of Cybernetics in Kiev [Shor 1964; Ermoliev 1966] and the descent methods which appeared in the western scientific literature in the 70's (see Balinski and Wolfe [1975] for a bibliography).

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

Let us attempt to generalize the known gradient methods of the kind

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

where x^s is an approximate solution at the s-th iteration, and ρ_s are step-size multipliers, for convex functions $f(x)$ with a discontinuous gradient. Difficulties arise connected with the choice of step multipliers ρ_s in the similar procedure

$$x^{s+1} = x^s - \rho_s \hat{f}_x(x^s), \quad s = 0,1,\ldots \tag{18}$$

or more generally

$$x^{s+1} = \Pi_X(x^s - \rho_s \hat{f}_x(x^s)), \quad s = 0,1,\ldots \tag{19}$$

where $\hat{f}_x(x^s)$ is a subgradient of $f(x)$ at $x = x^s$ and $\Pi_X(\cdot)$ is a projection operator on set X.

In practice, it is difficult to review the whole set of subgradients and choose the one which lies in the opposite direction to the domain of smaller values of the objective function (see Figure 3). Usually one can get only one subgradient, and therefore there is no guarantee that a step according to procedure (18) will lead into the domain of smaller values of $f(x)$. The nondescent procedure (18) was proposed by N.Z. Shor [1964] and called the method of generalized gradients.

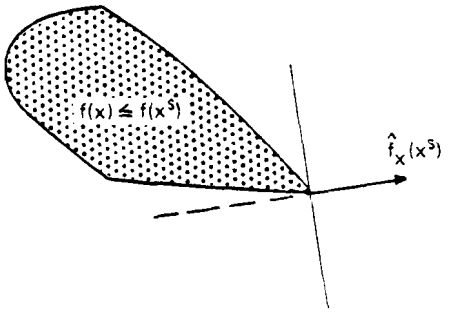

Figure 3.

It allows the use of any subgradient in the subdifferential. General conditions for its convergence were first independently obtained by Ermoliev [1966] and by Polyak [1967] , where ρ_s satisfies the conditions

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

These conditions are very natural as (18), (19) are non-descent processes, i.e., the value of the objective function does not necessarily decrease from iteration to iteration even for arbitrarily small ρ_s.

The influence and close relationship of I. I. Eremin's research on solutions of systems of inequalities and nonsmooth penalty functions on this area of work should be noted [Eremin 1965].

More recently the method (18) has been developed further (see Shor [1976] for a review) and rates of convergence have been studied.

E. A. Nurminski [1973] studied the convergence of methods of type (18) for the functions satisfying the following condition:

$$f(y) - f(x) \geq \langle f_x(x), y-x \rangle + o(\|y - x\|) \quad .$$

Moreover, he proposed a new technique for proving convergence based on the *reductio ad absurdum* argument; he then adapted this technique for studying the convergence of nondescent methods for nonconvex, nonsmooth optimization.

As has already been said, the algorithms constructed on the basis of (18) are simple and require relatively little storage. Thus let us consider an application of the method (19) to the

development of iterative schemes of decomposition. For the function (5) one of the generalized gradients at point x^s is

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

where u^s are dual variables corresponding to $y(x^s)$. Therefore, the iterative scheme of decomposition according to the procedure (19) has the form

$$x^{s+1} = \max \{0, x^s - \rho_s (c - u^s A)\}, \quad s = 0,1,\ldots , \tag{20}$$

For the problem (7) - (8), if y^s is an approximate solution of the subproblem (9) for $x = x^s = \{x^s_{ij}\}$ and u^s are dual variables corresponding to y^s, then

$$x^{s+1} = \pi_x(x^s - \rho_s u^s), \quad s = 0,1,\ldots , \tag{21}$$

where $\pi_x(\cdot)$ is the projection operator on the set (10). There is a very simple algorithm for obtaining $\pi_x(\cdot)$ on sets of type (10).

For the minimax problem (12) in the case when $g(x,y)$ for each $y \in Y$ is a convex function with respect to x, the subgradient is defined as $\hat{f}_x(x) = \hat{g}_x(x,y)|_{y = y(x)} = g_x(x,y(x))$. If $g(x,y)$ is continuously differentiable with respect to x then

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

It should be noted that the above-mentioned nondifferentiable procedures of decomposition make it relatively simple to take into account the special structure of a given problem. For example,

consider the linear problems of optimal control: to find a control $x = (x(0),\ldots,x(N-1))$ and a trajectory $z = (z(0),\ldots,z(N))$, satisfying the state equations:

$$z(k+1) = A(k)z(k) + B(k)x(k) + a(k) \quad ,$$

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

the constraints

$$G(k)z(k) + D(k)x(k) \geq b(k) \quad ,$$

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

and minimize the objective function

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

where $x(k) \in R^n$, $z(k) \in R^r$. The difficulty of this problem is connected with the state constraints. If matrix $G(k) \equiv 0$, we can solve this problem with the help of Pontzjagin's principle.

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

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

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

and constraints

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

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

which minimize

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

We have the following analog of the iterative scheme of decomposition considered above:

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

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

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

$$+ \langle d(k) - p(k+1)B(k) - \lambda(x)D(k), x^s(k) \rangle]$$

under constraints

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

$$p(N) = -c(N) \quad , \lambda(k) \geq 0 \quad , \; k = N-1, \ldots, 0 \quad .$$

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

Some original work carried out individually by Wolfe and Lemarechal (see Balinski and Wolfe [1975]) on descent methods is, on the one hand, a generalization of the ε-steepest descent algorithms studied by Demyanov [1974] and, on the other hand, formally similar to algorithms of conjugate gradients, coinciding with them in the differentiable case.

Since it is impossible to obtain the whole set $\{\hat{f}_x(x^s)\}$ at the point x^s, Wolfe and Lemarechal tried to construct it approximately at each iteration. The further development of subgradient schemes resulted in the creation of ε-subgradients, which were

introduced by Rockafellar [1974]. The early results in this field are due to Rockafellar [1970], Bertsecas and Mitter [1973], Lemarechal [1975], and Nurminski and Zhelikhovski [1977]. Recent research revealed properties of ε-subgradient mappings such as Lipschitz continuity which make ε-subgradient methods attractive both in theoretical and practical respects.

7. STOCHASTIC METHODS

Two classes of deterministic methods have been discussed: nondescent methods and descent methods. The first class is easy to use on the computer but does not result in a monotonic decrease in the objective function. The second class gives monotonic descent but has a complex logic. Both classes have a common shortcoming: they require the exact computation of a subgradient (in a differentiable case this would be the gradient). Often, however, there are problems in which the computation of subgradients is practically impossible. Random directions of search is a simple alternative method of constructing nondifferentiable optimization stochastic descent procedures that do not require the exact computation of a subgradient and which are easy to use on the computer.

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

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

Stochastic nondescent methods of random search (stochastic optimization) are of prime importance in the solution of the most difficult problem arising in stochastic programming, in which it is

impossible to compute either subgradients or exact values of objective and constraint functions. The presence of random components in the search directions of nondescent procedures allows one to overcome local minima, points of discontinuity, etc. A quite general scheme of nondescent methods in stochastic optimization was studied in Ermoliev and Nekrilova [1967], and Ermoliev [1976] under the name stochastic quasigradient (SQG) methods. SQG methods generalize the well-known stochastic approximation methods for optimization of the expectation of random functions to problems involving general constraints with differentiable and nondifferentiable functions. For deterministic nonlinear optimization problems these methods can be regarded as methods of random search. Consider the problem

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

We assume here that $F^\nu(x)$, $\nu = \overline{0,m}$ are convex functions, and X is a convex set. Let $\hat{F}^\nu_x$ denote a subgradient of the function $F^\nu(x)$:

$$F^\nu(z) - F^\nu(x) \geq (\hat{F}^\nu_x(x), z - x) \quad .$$

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

$$E\{\xi^\nu(s) \mid x^0, \ldots, x^s\} = \hat{F}^\nu_x(x^s) + a^\nu(s) \quad ,$$

$$E\{\zeta_\nu(s) \mid x^0, \ldots, x^s\} = F^\nu(x^s) + b_\nu(s) \quad ,$$

where $a^\nu(s)$ is a vector, $b_\nu(s)$ is a number depending upon x^0, $x^1, \ldots, x^s, \ldots$, where usually $a^\nu(s) \to 0, b_\nu(s) \to 0$ (in any sense) for $s \to \infty$. Thus in these methods, instead of exact values of

$\hat{F}_x^\nu(x^s), F^\nu(x^s)$, quantities $\xi^\nu(s), \zeta_\nu(s)$ are used. For further understanding, it is important to see that the random values $\zeta_\nu(s)$ and vectors $\xi^\nu(s)$ are easily calculated. For example, if

$$F^\nu(x) = Ef^\nu(x,\omega)$$

then $\zeta_\nu(s) = f^\nu(x^s,\omega^s)$ where the ω^s result from mutually independent draws of ω. We have

$$E\{\zeta_\nu(s) \mid x^s\} = E\{f^\nu(x^s,\omega) \mid x^s\} = f^\nu(x^s) \quad .$$

If functions $f^\nu(x,\omega)$ are differentiable with respect to x and

$$\xi^\nu(s) = f_x^\nu(x^s,\omega^s),$$

then under reasonable assumptions we will have

$$E\{\xi^\nu(s) \mid x^s\} = Ef_x^\nu(x^s,\omega^s) = (Ef_x^\nu(x^s,\omega^s) \mid x^s\})_x = F_x'(x^s).$$

It should be stressed that SQG methods are applicable not only to stochastic programming problems, but also to NDO deterministic problems, without having to compute values of subgradients. For example, for the deterministic minimax problem (12) consider the vector

$$\xi^0(s) = (3/2)\frac{g(x^s+\Delta_s h^s, y(x^s)) - g(x^s, y(x^s))}{\Delta_s} h^s, \qquad (22)$$

where $\Delta_s > 0, h^s$ is the result of independent random draws of the random vector $h = (h_1,\ldots,h_n)$ whose components are independently and uniformly distributed over $[-1,1]$. (22) satisfies the condition

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

where $\hat{f}_x^0(x^s)$ is a subgradient of the function (12) and $|a^0(s)| \le \text{const}\cdot\Delta_s$, if $g(x,y)$ has uniformly limited second

derivatives with respect to $x \in X$. It is remarkable that independent of the dimensionality of the problem, the vectors (22) can be found by calculating the function $g(x,y)$ at two points only. This is particularly important for extremal problems of large dimensionality.

Let us now discuss briefly one particular type of SQG method - the stochastic quasigradient projection method.

Let it be required to minimize the convex function $F^0(x)$, $x \in X$, where X is a convex set.

The method is defined by the relations:

$$x^{s+1} = \pi_x(x^s - \rho_s \xi^0(s)), \quad s = 0,1,\ldots, \tag{23}$$

where $\pi_x(\cdot)$ is a projection operation on X, ρ_s are step multipliers. The method (23) has been proposed in Ermoliev and Nekrilova [1967]. The characteristic requirements under which the sequence $\{x^s\}$ converges with probability 1 to the solution, are: if $\|x^k\| \le B$, $k = \overline{0,s}$, then $E\{\|\xi^0(s)\|^2 | x^0,\ldots,x^s\} \le C_B$ where B, C_B are constants; ρ_s are step multipliers which may depend upon $x^0, x^1, \ldots, x^s$ and

$$\rho_s \ge 0, \quad \sum_{s=0}^{\infty} \rho_s = \infty \text{ with probability 1}, \tag{24}$$

$$\sum_{s=0}^{\infty} E\{\rho_s^2 + \rho_s \|a^0(s)\|\} < \infty \quad . \tag{25}$$

In the special case when the ρ_s are deterministic and independent of $(x^0,\ldots x^s)$ then, under (24), (25) we obtain from method (23) using the random direction (22) that

$$\rho_s \ge 0, \quad \sum_{s=0}^{\infty} \rho_s = \infty, \quad \sum_{s=0}^{\infty} \rho_s \Delta_s < \infty, \quad \sum_{s=0}^{\infty} \rho_s^2 < \infty \quad . \tag{26}$$

The important application of SQG methods is concerned with the optimization of probabilistic systems by simulation. To obtain

the desired parameters of the probabilistic system, one can often use a Monte-Carlo simulation technique. Let

$$f^0(x,\omega^s)$$

be the outcome of the s-th simulation, where $f^0(x,\omega^s)$ is the random quantity, which depends on an unknown vector of parameters, and noise observations ω^s. Suppose that it is desired to choose the vector $x = (x_1,\ldots,x_n)$, which minimizes the expected value of the performance function defined by the expression

$$F^0(x) = Ef^0(x,\omega) = \int f^0(x,\omega)P(d\omega).$$

The main difficulty of this problem is that the distribution $P(d\omega)$ is unknown and we cannot get the precise value of the function $F^0(x)$ (it is theoretically impossible). A statistical estimate $\xi^0(x)$ of the gradient for a differentiable function $F^0(x)$ could be calculated analogously to (22):

$$\xi^0(s) = (3/2)\frac{f^0(x^s + \Delta_s h^s,\omega^{s1}) - f^0(x^s,\omega^{s0})}{\Delta_s} h^s,$$

where $f^0(x^s + \Delta_s h^s,\omega^{s1}), f^0(x^s,\omega^{s0})$ are outcomes of simulations for $x = x^s + \Delta_s h^s$ and $x = x^s$.

If the second derivatives of the function $F^0(x)$ are bounded for $x \in X$, then

$$E(\xi_j^0(s)\,|\,x^s) = (3/2)\sum_{i=1}^{n} F^0_{x_i}(x^s)Eh_i^s h_j^s + O(\Delta_s) = F^0_{x_j}(x^s) + O(\Delta_s),$$

since $Eh_i h_j = 0$, if $i \neq j$ and $Eh_j^2 = 2/3$. The step-size Δ_s of the finite difference approximation could be chosen according to conditions (26).

8. CONCLUSION

Some applied NDO and STO problems have been briefly discussed in this work. Deterministic, stochastic, descent and nondescent methods were considered. Each one requires some definite information about objective and constraint functions. Deterministic descent methods use the exact values of these functions and their subgradients; stochastic descent methods use only the exact values of functions; deterministic nondescent methods require only the exact values of subgradients; stochastic nondescent methods use neither the values of functions nor the exact values of their subgradients. Obviously, every method reveals its advantages in a specific class of extremum problems: for instance, complex stochastic programming problems are soluble only by stochastic nondescent methods.

REFERENCES

Balinski, M.L., and P. Wolfe (Eds.) 1975. Nondifferentiable Optimization. Mathematical Programming Study 3. Amsterdam: North Holland Publishing Co.

Bertsecas, D.P., and S.K. Mitter. 1973. A descent numerical method for optimization problems with nondifferentiable cost functionals. SIAM Journal on Control 11:637-652.

Danskin, J.M. 1967. The Theory of Min-Max. Berlin: Springer-Verlag.

Demyanov, V.F., and V.N. Malozemov. 1974. Introduction to Mini-Max. New York: John Wiley and Sons.

Eremin, I. 1965. A generalization of the Mozkin-Agmon relaxation method. Uspekhi Mathematischeski Nauk 20:183-187.

Ermoliev, Yu.M. 1966. Methods of solution of nonlinear extremal problems. Kibernetika 2(4).

Ermoliev, Yu.M. 1970. On the stochastic penalty functions method. Kibernetika 1.

Ermoliev, Yu.M. 1976. Methods of Stochastic Programming. Moscow: Nauka (in Russian).

Ermoliev, Yu.M., and L.G. Ermolieva. 1973. A method of parametric decomposition. Kybernetika 9(2).

Ermoliev, Yu.M., and Z. Nekrilova. 1967. Stochastic Gradient Methods and Their Application in Optimal Decision Theory. Notes, Seminar on the Theory of Optimal Solution. Academy of Sciences of the U.S.S.R., Kiev.

Lasdon, L.S. 1970. Optimization Theory for Large Systems. London: Macmillan.

Lemarechal, C. 1975. Nondifferentiable Optimization: Subgradient and ε-Subgradient Methods. Lecture Notes on Optimization and Operations Research.

Nurminski, E.A. 1973. The quasigradient method for the solving of nonlinear programming problems. Cybernetics 9(1):145-150. New York/London: Plenum Publishing Co.

Nurminski, E.A., and A.A. Zhelikhovski. 1977. ε-Quasigradient method for solving nonsmooth extremal problems. Cybernetics 13(1):109-114. New York/London: Plenum Publishing Co.

Polyak, B.T. 1967. A general method for solving extremal problems. Doklady Academii Nauk SSR 174:33-36.

Rockafellar, R.T. 1970. Conjugate convex functions in optimal control and the calculus of variations. Journal of Mathematical Analysis and Applications 32:174-222.

Rockafellar, R.T. 1974. The multiplier method of Hesten and Powell applied to convex programming. Journal of Optimization Theory and Applications 12:9.

Shor, N.A. 1964. On a Structure of the Algorithms for Numerical Solution of the Optimal Planning and Designing. Dissertation, Institute of Cybernetics, Kiev.

Shor, N.Z. 1967. Application of the generalized gradient method in block-programming. Kibernetika 3(3).

Shor, N.Z. 1976. Generalizations of gradient methods for non-smooth functions and their applications to mathematical programming. Economics and Mathematical Methods 12(2):337-356.

ACCELERATION IN THE RELAXATION METHOD FOR LINEAR INEQUALITIES AND SUBGRADIENT OPTIMIZATION*

J.L. Goffin
McGill University
Montreal, Canada

1. INTRODUCTION

Subgradient optimization has been shown in many experiments to be an effective solution technique to the problem of maximizing piecewise linear concave functions defined through the use of the Dantzig-Wolfe decomposition principle applied to some combinatorial problems. This effectiveness showed up in some problems of rather respectable size [10, 12, 15], even though it can be shown to perform arbitrarily badly in two-dimensional problems.

A convergence theory developed by Shor [21, 22] and the author [7] quantifies the rates of convergence of subgradient optimization, in function of condition numbers which measure the good behaviour of the function to optimize. When applied to a quadratic function, subgradient optimization can achieve the rate of convergence of the steepest ascent method provided that the function be not too well conditioned, that the condition number (or the eigenvalue ratio of the matrix defining the

*This research was supported in part by the D.G.E.S. (Quebec) and the N.R.C. of Canada under grant A4152.

quadratic function) be known, and that an overestimate of the distance between the initial point and the solution point be known.

When applied to piecewise linear problems, subgradient optimization always performed significantly better than the theory of [22] and [7] indicated, the gap in performance between experience and theory being wide enough to justify further study.

One key fact to notice is that subgradient optimization is closely related to the relaxation method for solving systems of inequalities, to the extent that the rates of convergence that have been proved for both methods are identical [8,9], under some assumptions (namely that the function not be too well conditioned, that the relaxation parameter be equal to one and that every non-zero extreme point of every subdifferential of the function have the same norm).

A second key fact to notice is that the relaxation method for solving systems of linear inequalities is related to the Kacmarz projection method, the successive overrelaxation method and the Southwell method for solving systems of linear equalities. It has been observed, and proved for some cases, that the S.O.R. method can be accelerated by using values of the relaxation parameter greater than one (if this parameter is equal to one, then the S.O.R. method is known as the Gauss-Seidel, or Nekrasov method). This indicates that acceleration is conceivable for the relaxation method for linear inequalities, and thus also for subgradient optimization.

A study of the acceleration of the relaxation method for a system of two inequalities done in [6] and [9] shows that the same value of the relaxation parameter which accelerates the most the S.O.R. technique in a related linear system of equations also accelerates the convergence of the relaxation method for inequalities, but that this accelerated rate is still bad if the system is badly conditioned (even if the rate is improved by an "order of magnitude"). Experiments suggest that if the relaxation parameter goes above the optimal value given by S.O.R. theory, the convergence accelerates still more, for reasons

different from the ones used in the S.O.R. theory. And thus acceleration, even if it could be proved satisfactorily, is only part of the explanation of the effectiveness of subgradient optimization.

The other part of the explanation will have to relate the combinatorial structure of the original problem studied to the condition number of the function defined through the Dantzig-Wolfe decomposition principle. This paper will attempt to show ("indicate" or "allude to" might be more appropriate) by using experiments and some theory that:

1. The relaxation method for linear inequalities, and subgradient optimization can be accelerated (and in some cases faster than the S.O.R. theory would predict).
2. For some classes of combinatorial problems, uniform lower bounds (dependent upon the dimension of the problem) on the condition number of the associated functions probably exist.
3. The "harder" the combinatorial problem is, the better the condition number of the associated function is.

Given the rather incomplete nature of the theory of acceleration in the S.O.R. method, it should be clear that a general proof of these three points is probably unattainable.

2. SUBGRADIENT OPTIMIZATION AND THE RELAXATION METHOD

Let $f(x)$ be a finite concave function defined on R^n and $\partial f(x)$ be its subdifferential at x, i.e.,

$$\partial f(x) = \{u \in R^n : f(y) \leq f(x) + (u, y-x), \ \forall\, y \in R^n\}.$$

For every x, $\partial f(x)$ is a convex, compact set, and if f is differentiable then $\partial f(x) = \{\nabla f(x)\}$, the gradient of f.

In this paper $f(x)$ will always be a <u>piecewise linear function</u>:

$$f(x) = \min_{i \in I} \{(a^i, x) + b^i\}$$

where I is a finite set, $a^i \in R^n, b^i \in R$.

Letting $I(x) = \{i \in I : f(x) = \langle a^i, x\rangle + b^i\}$, then

$$\partial f(x) = \{u \in R^n : u = \sum_{i \in I(x)} \lambda_i a^i, \quad \sum_{i \in I(x)} \lambda_i = 1, \lambda_i \geq 0, \ \forall\, i \in I(x)\}$$

is the convex hull of the set $\{a^i : i \in I(x)\}$.

It will be assumed that the optimal value $f^* = \max\{f(x) : x \in R^n\}$ is reached on the optimal set $P = \{x \in R^n : f(x) = f^*\}$. It is well known that $P = \{x \in R^n : 0 \in \partial f(x)\}$. The projection of x on P will be denoted by $x^*(x)$, and the distance from x to P by $d(x) = |x - x^*(x)|$, where $|....|$ means Euclidian norm.

The term subgradient optimization will be used if f^* is not known, and the term relaxation method will be used if f^* is known. The reason for this is that if f^* is known, the optimal set can be defined by a system of linear inequalities:

$$\langle a^i, x\rangle + b^i - f^* \geq 0, \quad i \in I$$

A description of both subgradient optimization and of the relaxation method follows:

1. Choose $x^0 \in R^n$.
2. Compute a subgradient of f at $x^q : u^q \in \partial f(x^q)$ (or it could be restricted to $u^q \in \{a^i : i \in I(x^q)\}$. If $u^q = 0$, an optimal point has been found (also if $f(x^q) = f^*$ in the case of the relaxation method).
3. The next point x^{q+1} of the sequence will be obtained by moving from x^q in the direction of u^q by a certain step size. Go back to 2 with q+1 replacing q.

The selection of the step size depends on whether f^* is known or not:

$$x^{q+1} = x^q + \lambda_q \frac{u^q}{\|u^q\|} \quad \text{subgradient optimization } (f^* \text{ unknown}) \qquad (2.1)$$

$$x^{q+1} = x^q + \sigma_q \frac{f^* - f(x^q)}{\|u^q\|^2} u^q \quad \text{relaxation method } (f^* \text{ known}). \qquad (2.2)$$

The sequences $\{\lambda_q\}$ and $\{\sigma_q\}$ that will be studied are given by:

$$\lambda_q = \lambda_o \rho^q \quad , \quad \rho \in (0,1) \quad , \quad \lambda_o > 0$$

and

$$\sigma_q = \sigma \quad , \quad \sigma \in (0.2].$$

In both cases a convergence theory is available: in [22], [7] for subgradient optimization and in [1], [6] and [9] for the relaxation method. Rates of convergence depend upon condition numbers which are defined below. For subgradient optimization one defines:

$$\mu(x) = \min_{u \in \partial f(x)} \frac{\langle u, x^*(x)-x \rangle}{|u| \, |x^*(x)-x|} \qquad \text{for every } x \notin P$$

and

$$\mu = \inf_{x \notin P} \mu(x)$$

and for the relaxation method:

$$\tilde{\mu}(x) = \min_{u \in \partial f(x)} \frac{f^*-f(x)}{|u| \, |x^*(x)-x|} \qquad \text{for every } x \notin P$$

and

$$\tilde{\mu} = \inf_{x \notin P} \tilde{\mu}(x) \quad .$$

The concavity of f implies that $\tilde{\mu}(x) \leq \mu(x)$ and thus $\tilde{\mu} \leq \mu$.

The convergence theorems are as follows:

Theorem 2.1 For subgradient optimization (where $\lambda_q = \lambda_0 \rho^q$), let

$$C = \max\left\{\frac{1}{\rho},\ \frac{\mu - \sqrt{(\mu^2-(1-\rho^2))}}{1-\rho^2}\right\}, \quad D = \frac{\mu + \sqrt{(\mu^2-(1-\rho^2))}}{1-\rho^2}$$

$$z(\mu) = \begin{cases} \sqrt{(1-\mu^2)} & \text{if } \mu \leq \frac{1}{2}\sqrt{2}, \\ \frac{1}{2\mu} & \text{if } \mu > \frac{1}{2}\sqrt{2}. \end{cases}$$

Then

(A) $\rho \geq z(\mu)$ and $d(x^0) \in [\lambda_0 C, \lambda_0 D]$ implies that for all q: $d(x^q) \leq d(x^0)\rho^q$,

(B) $\rho \geq z(\mu)$ and $d(x^0) < \lambda_0 C$ implies that for all q: $d(x^q) \leq \lambda_0 C \rho^q$,

(C) $\rho < z(\mu)$ or $d(x^0) > \rho_0 D$ may lead to the convergence of $\{x^q\}$ to a non-optimal point [7].

Theorem 2.2 For the relaxation method:

$$d(x^{q+1}) \leq \sqrt{1-\sigma(2-\sigma)\tilde{\mu}^2}\ \ d(x^q)$$

and furthermore if dim P = n, there exists a $\sigma^* \in [1,2)$ such that if $\sigma \in (\sigma^*, 2]$, then convergence is finite [9].

It should be clear that in general μ and $\tilde{\mu}$ are different, but also there is a neighbourhood of P such that $\mu(x) = \tilde{\mu}(x)$. One could define condition numbers close to the optimal set (say μ_c and $\tilde{\mu}_c$: note that they are equal): the number $\tilde{\mu}_c$ could be used in Theorem 2.2 instead of $\tilde{\mu}$, provided that q is large enough, so that the iterates are close to the optimal set; a similar statement cannot be made easily about Theorem 2.1, because the proof technique is quite different (and more global in nature).

Thus the rates of convergence of both methods involve condition numbers which are related to one another.

In one case of interest [8], where all the extreme points of all the subdifferentials of f have the same norm (unless the norm is zero) then $\mu = \tilde{\mu} = \mu_c = \tilde{\mu}_c$, and thus in that case the rates of convergence (as given by the available theories) of the relaxation method with $\sigma = 1$, and of subgradient optimization with $\rho = z(\mu)$ (and $\mu \leq \sqrt{2}/2$) are identical. In this case it also follows that condition numbers far from the optimal set can be no worse than close to the optimal set. The condition that all norms be equal is met by the relaxation method of Agmon under its maximal distance implementation (and all norms are one): it simply means that all a^i have been normalized.

This whole theory, though correct, is incomplete, in the sense that it fails to account for the interaction between successive iterates: the theory only considers the worst step-to-step behaviour of the sequences, putting bounds on this, but ignoring the fact that it is impossible for all successive iterates to behave according to this worst bound. This reasoning is very similar in nature to the one used in proving the convergence, and the acceleration, in the S.O.R. technique for solving systems of linear equalities: the linear operators that relate an iterate to the next one all have a spectral radius of one, but the product of n operators (which defines a cycle of n iterations in our terminology, or of <u>one</u> iteration according to the S.O.R. terminology) has a spectral radius less than one.

In this paper we will attempt to extend the ideas of the acceleration of the S.O.R. method to the relaxation method for inequalities and to subgradient optimization.

3. EXAMPLES AND EXPERIMENTATION

Example 1: Systems of linear equalities

$$(a^i, x) + b^i = 0 \quad , \quad x \in R^n, \quad i \in I \tag{3.1}$$

(where usually I contains n elements).

Solving this system is equivalent to maximizing

$$f(x) = \min_{i\in I} \min \{(a^i,x) + b^i, \quad -(a^i,x)-b^i\} \tag{3.2}$$

where $f^* = 0$ if and only if the system has a solution.

Example 2: Dual of a transportation problem

$$f(x) = \sum_{j=1}^{n} s_j x_j + \sum_{k=1}^{m} d_k \min_{j=1,\ldots,n} (c_{jk}-x_j) \tag{3.3}$$

where $\sum_{j=1}^{n} s_j = \sum_{k=1}^{m} d_k$.

Example 3: Dual of an assignment problem

$$\text{in example 2, let } m = n, \quad s_j = 1, \quad d_k = 1 \ \forall j,k. \tag{3.4}$$

We experimented on TR 48, the dual of a transportation problem with n = m = 48, the data of which can be found in [17]. This problem has the peculiarity that the optimal solution is "unique" (i.e., up to an additive constant), and thus we were able to compute $\mu(x^q)$ for the successive iterates. We observed that for all experiments performed inf $\mu(x^q) \cong .0021$, and thus μ and $\tilde{\mu}$ are less than .0021, from which it follows that $z(\mu) = \sqrt{1-\mu^2} \geq .999998$. Actual experimentation with the relaxation method with $\sigma = 1$ showed an asymptotic rate of convergence of .999986 (quite disastrous as it indicates that in order to improve the accuracy of the optimum by one digit, one should go through something like 150,000 iterations). This pessimistic result should be tempered by the observation that the methods (subgradient optimization and relaxation) behaved very well in the opening game (the early iterations) before slowing down.

Experiments were performed in order to measure the actual convergence rates of the relaxation method as a function of σ. The theory says that the rate of convergence is $\sqrt{1-\sigma(2-\sigma)\tilde{\mu}^2}$, i.e., it worsens when σ goes from one to two.

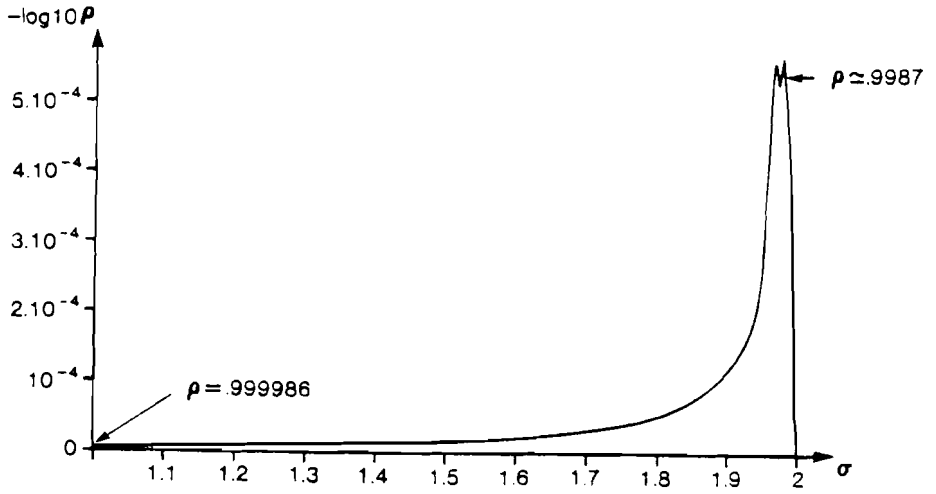

Figure 1. Rate of convergence of the relaxation method applied to TR 48 as a function of σ.

The results, shown in Figure 1, clearly indicate that acceleration takes place, and that a best rate of convergence of about .9985 occurs for $\sigma \in (1.96, 1.98)$. This is an improvement by a factor of one hundred over what happened when $\sigma = 1$.

Closer examination of the convergence showed that:

1. Convergence was very good in the first few hundred iterations: good opening game.
2. It worsened considerably after this (to a rate of the order of .9998): weak middle game.
3. It improved again, starting about when all the subgradients used define linear pieces which go through the optimum point, thus showing that acceleration is an asymptotic behaviour; it also shows that the $d(x^q)$ encountered are all in the range .16 to .26: good end game.

It seemed logical then to experiment with subgradient optimization; we tried with $\lambda_0 = .5000$, $\sigma = .999$, $x^0 = 0$ (where $d(x^0) \cong 1978.$), for 30,000 iterations. Even though we had some worries

about the middle game it converged at a rate given by ρ (on that the theory is clear: if it converges, it must be at a rate ρ). It is not clear at this point whether this is luck or an indication of an underlying theory of acceleration of subgradient optimization (my own feeling is that getting over the middle game is luck, but that acceleration in the final game is probably provable).

A study of the acceleration phenomenon in the relaxation method for inequalities will be attempted in the next sections on a few very particular examples.

4. SYSTEMS OF LINEAR EQUALITIES

Let a system of linear equalities be given by:

$$(a^i, x) + b^i = 0, \quad i \in I = \{1,2,\ldots,n\} \tag{4.1}$$

where the a^i are linearly independent column vectors. Another notation is of course $Ax + b = 0$, where

$$A = \begin{pmatrix} a^{1T} \\ \vdots \\ a^{nT} \end{pmatrix} \qquad \text{and} \qquad b = \begin{pmatrix} b^1 \\ \vdots \\ b^n \end{pmatrix}.$$

Let D be a diagonal matrix, with $D_{ii} = |a^i|^2$, then (4.1) can be written as

$$\tilde{A}x + \tilde{b} = 0, \quad \tilde{A} = D^{-1/2}b \quad . \tag{4.2}$$

If one defines η and $\tilde{\eta}$ by $x = \tilde{A}^T\tilde{\eta} = A^T\eta$, then equivalent systems may be defined by

$$\tilde{\Gamma}\tilde{\eta} + \tilde{b} = 0 \text{ where } \tilde{\Gamma} = \tilde{A}\tilde{A}^T \tag{4.3}$$

$$\Gamma\eta + b = 0 \text{ where } \Gamma = AA^T \quad . \tag{4.4}$$

If x^* denotes the (unique) solution to (4.1) and (4.2) and $\tilde{\eta}^*$ and η^* the (unique) solutions to (4.3) and (4.4), clearly

$$x^* = \tilde{A}^T \tilde{\eta}^* = A^T \eta^* \quad , \quad \tilde{\eta}^* = D^{1/2} \eta^* \quad .$$

The matrix Γ is the Grammian of the vectors a^i, while $\tilde{\Gamma}$ is the Grammian constructed on the vectors of unit norm $\tilde{a}^i = a^i / |a^i|$. Clearly $\tilde{\Gamma}$ is simply the matrix of the cosines of the angles between the vectors a^i, $i = 1,\ldots,n$. Of course Γ and $\tilde{\Gamma}$ are both positive definite symmetric matrices.

The S.O.R. technique (or extrapolated Gauss-Seidel) can be written as follows:

1. x^0 arbitrary ($q=0$)
2. define i by $i = q \pmod n + 1$
   then
3. $\tilde{\eta}^{q+1} - \tilde{\eta}^* = (I - \sigma E_i \tilde{\Gamma})(\tilde{\eta}^q - \tilde{\eta}^*)$
   $q \leftarrow q+1$ and go to 2 .

E_i is a matrix with a one in position (i,i) and zeros elsewhere; note also that $\tilde{\eta}^{q+1}$ is computable in function of $\tilde{\eta}^q$ ($\tilde{\eta}^*$ is not known but $\tilde{\Gamma}\tilde{\eta}^* = -\tilde{b}$ is).
If we let $\tilde{\delta}(\tilde{\eta}) = \tilde{\Gamma}\tilde{\eta} + \tilde{b}$, the residue, then also

$$\tilde{\delta}(\tilde{\eta}^{q+1}) = (I - \sigma \tilde{\Gamma} E_i)\tilde{\delta}(\tilde{\eta}^q)$$

$$= (I - \sigma E_i \tilde{\Gamma})^T \tilde{\delta}(\tilde{\eta}^q) \quad \text{as } \tilde{\Gamma} E_i = (E_i \tilde{\Gamma})^T \quad .$$

In the S.O.R. theory one defines a sequence by $\tilde{\mathcal{H}}^0 = \tilde{\eta}^0, \ldots, \tilde{\mathcal{H}}^k = \tilde{\eta}^{kn}$; the "iteration" from $\tilde{\mathcal{H}}^k$ to $\tilde{\mathcal{H}}^{k+1}$ will be characterized here as a <u>cycle</u> of n iterations. The reason for this is of course that the "iteration" from $\tilde{\mathcal{H}}^k$ to $\tilde{\mathcal{H}}^{k+1}$ is given by a linear operator, and thus the convergence can be studied, and rates of convergence can be identified with the spectral radius of this operator. If we define $\tilde{L}$ and $\tilde{U}$ by $\tilde{\Gamma} = I - \tilde{L} - \tilde{U}$, where $-\tilde{L}(-\tilde{U})$ are respectively, the strictly lower

(upper) part of $\tilde{\Gamma}$, then

$$\tilde{H}^{k+1} - \tilde{\eta}^* = (I-\sigma\tilde{L})^{-1}((1-\sigma)I + \sigma\tilde{U})(\tilde{H}^k - \tilde{\eta}^*)$$

the classical formula; clearly $(I-\sigma\tilde{L})^{-1}((1-\sigma)I - \sigma\tilde{U})$ is equal to $(I-\sigma E_n\tilde{\Gamma}) \cdot (I-\sigma E_{n-1}\tilde{\Gamma}) \ldots (I-\sigma E_1\tilde{\Gamma})$. If we let $R = (I-\sigma\tilde{L})^{-1}((1-\sigma)I + \sigma\tilde{U})$ then also $\tilde{\rho}(\tilde{\eta}^{kn+n}) = \tilde{\Gamma}R\tilde{\Gamma}^{-1} \approx (\tilde{\eta}^{kn})$.

Now let us define the Kacmarz (extrapolated) method applied to (4.1) or (4.2):

1. x^0 arbitrary $(q = 0)$
2. i defined by $i = q(\text{mod } n) + 1$
3. $x^{q+1} - x^* = (I-\sigma\tilde{a}^i\,\tilde{a}^{iT})(x^q - x^*)$
   $q \leftarrow q+1$ and go to 2.

Again, as $\tilde{a}^{iT}x^* = -\tilde{b}^i$, x^{q+1} is computable in function of x^q. If we let $\tilde{r}(x) = \tilde{A}x + \tilde{b}$, and $r(x) = Ax + b$, the residues, then it can be checked that:

$$\tilde{r}(x^{q+1}) = (I-\sigma\tilde{\Gamma}E_i)\tilde{r}(x^q) \quad .$$

Observe that:

$$\tilde{A}^T E_i \tilde{\Gamma}\tilde{A}^{-1T} = \tilde{A}^T E_i \tilde{A} = \tilde{a}^i\tilde{a}^{iT},$$

and thus

$$(I-\sigma\tilde{a}^i\tilde{a}^{iT}) = \tilde{A}^T(I-\sigma E_i\tilde{\Gamma})\tilde{A}^{-1T} \quad .$$

Also,

$$\tilde{r}(x^{kn+n}) = \tilde{\Gamma}R\tilde{\Gamma}^{-1}\tilde{r}(x^{kn})$$

$$x^{kn+n} = \tilde{A}^T R\tilde{A}^{-1T}x^{kn} \quad .$$

We have thus shown the following "theorem".

Theorem 4.1

If we use the Kacmarz method applied to (4.1) (4.2) or the Gauss-Seidel method applied to (4.3) and (4.4), with the same

relaxation parameter σ and starting points x^0 (for 4.1 and 4.2), η^0 (for 4.4), $\tilde{\eta}^0$ (for 4.3) satisfying

$$x^0 = \tilde{A}^T\tilde{\eta}^0 = A^T\eta^0 \quad ,$$

then at every iteration, it is also true that

$$x^q = \tilde{A}^T\tilde{\eta}_q = A^T\eta^q \quad .$$

This fact was noticed by Kahan [14]. And thus the rates of convergence of both methods are identical, and given by the spectral radius of R, (which is known to be less than one if $\sigma \in (0,2)$).

As to every positive definite symmetric matrix Γ one can associate (real) Cholesky factors by $\Gamma = AA^T$, it is also clear that the S.O.R. technique applied to $\Gamma\eta + b = 0$ is equivalent to the extrapolated Kacmarz method applied to $Ax + b = 0$.

All of this can be found in Nicolaides [18].

The following theorem characterizes the convergence of the various methods in terms of the asymptotic behaviour of the directions pointing from the iterates to the optimal point.

Theorem 4.2

Assume that the largest modulus eigenvalue (Λ) of R is real, positive and the unique root of modulus Λ of the characteristic polynomial, with unit (right) eigenvector e_Λ, then the sequence generated by the S.O.R. method applied to (4.3) will, unless $\tilde{\eta}^0 - \tilde{\eta}$ is perpendicular to the left eigenvector $\bar{e}_\Lambda$ corresponding to Λ, satisfy:

$$\lim_{k\to\infty} \frac{\tilde{\eta}^{kn} - \tilde{\eta}^*}{|\tilde{\eta}^{kn} - \tilde{\eta}^*|} = \delta e_\Lambda \qquad \text{(where } \delta \text{ is either +1 or -1)}$$

$$\lim_{k\to\infty} \frac{|\tilde{\eta}^{kn+n} - \tilde{\eta}^*|}{|\tilde{\eta}^{kn} - \tilde{\eta}^*|} = \Lambda$$

$$\lim_{k\to\infty} \frac{\tilde{\eta}^{kn+\ell}-\tilde{\eta}^*}{|\tilde{\eta}^{kn}-\tilde{\eta}^*|} = (I-\sigma E_\ell \tilde{\Gamma})\ldots(I-\sigma E_1\tilde{\Gamma})\delta e_\Lambda, \quad 1 \le \ell \le n \quad .$$

Proof

Let $R = SJS^{-1}$ where J is the Jordan canonical form associated to R. Then there is only one Jordan canonical box corresponding to Λ, and it is of dimension 1. Let (ℓ,ℓ) be the location of Λ in J.

Classically: $J^q = \Lambda^q E_\ell + o(\Lambda^q)$, where $o(\Lambda^q)$ represents a $(n \times n)$ matrix which goes to zero faster than Λ^q.

Thus

$$R^q = \Lambda^q s_\ell s_\ell^{-T} + o(\Lambda^q)$$

where s_ℓ^{-T} is the ℓ-th row of S^{-1}, and s_ℓ is the ℓ-th column of S which can be shown to be proportional to e_Λ (it could have been chosen equal to e_Λ), while s_ℓ^- is proportional to e_Λ^-.

The theorem follows provided that $(e_\Lambda^-, \tilde{\eta}^o-\tilde{\eta}^*) \neq 0$.

To summarize, the convergence of $\tilde{\eta}^k$ to $\tilde{\eta}^*$ takes place along n one-sided asymptotes, used in a cyclic order; the convergence is geometric, with a crisper definition of this concept than in general (the decrease of the distance over n steps tends as a limit to Λ, and thus the average rate per iteration is $\sqrt[n]{\Lambda}$). Of course an almost identical theorem holds for the Kacmarz method.

If the eigenvalue(s) of largest modulus is complex, then the rate of convergence of the Kacmarz and of the S.O.R. method is still given by the n-th root of the modulus, but the existence of asymptotes depends on discussions on the rationality of the argument (in degrees) of the root of largest modulus. If the eigenvalue of largest modulus were negative, then there would be n <u>two</u>-sided asymptotes.

Now the relaxation method of Agmon defined in (2.2) and applied to (3.1) is identical in spirit to the Kacmarz method (a similar argument could be made to compare Southwell's relaxation

method applied to (4.3) and the S.O.R. method applied to (4.1)); the difference is that the sequence of indices used is not given by repeating the cycle 1,2,3,...,n but by choosing at each iteration the index of the largest residue, in absolute value. The same linear operators are used to go from one iteration to the next, but in an order, not necessarily uniquely defined, determined in a different fashion; the order is not defined a priori, but is dependent on the iteration sequence.

It might be worth pointing out that in the Agmon's relaxation method of (2.2) (and in Southwell's method), the normalization of $a^i, i = 1,\ldots,n$, changes the sequences generated, as the index which gives the maximum residue depends on the normalization.

The reason for this section is given by the observation that if in the relaxation method of Agmon (2.2) applied to (3.1) the indices chosen form cycles repeating the integers 1,2,3,...,n in sequence then the convergence theory of the Kacmarz method and of the S.O.R. method applies exactly. It also indicates why the discussion of 5.2 is needed; if in the relaxation method (2.2) the cycle 1,...,n has repeated itself often enough for the non-dominating eigenvalues to have lost their powers, then this cycle will repeat itself ad infinitum, so that an exact asymptotic theory of (2.2) would be available.

If the relaxation method of (2.2) used the index i(q) at iteration $q = 0,1,2,\ldots$, and if there exists an index q^* and an integer p such that $i(q+p) = i(q)$ for all $q \geq q^*$ (i.e., the indices used are cyclic), then all that has been said above holds, mutatis mutandis; the S.O.R. theory can be applied with a Grammian constructed on the vectors $a^{i(q)}, a^{i(q+1)}, \ldots, a^{i(q+p-1)}$, (in that order), where the Grammian might be positive semidefinite if some indices are repeated within a cycle, or if the a^i used are not linearly independent.

The study of when a cyclic order of indices will appear (in an asymptotic sense) has not been done yet, except for a two-dimensional linear system [9].

5. EXAMPLES OF APPLICATION OF THE THEORY OF ACCELERATION OF THE RELAXATION METHOD

1. A two-dimensional example was studied in [9] quite extensively. Let $a^1 = (\cos\alpha, \sin\alpha)^T$, $a^2 = (\cos\alpha, -\sin\alpha)^T$, $b^1 = b^2 = 0$, and $\alpha\in(0,\pi/4)$; a study of the relaxation method of Agmon (2.2) applied to $(a^i,x) + b^i \geq 0$, $i = 1,2$, $x\in R^n$ was done and it is already quite messy. A short summary follows. Let $\sigma^* = 2/(1+\sin 2\alpha)$, then if $\sigma\in(\sigma^*,2]$ finite convergence occurs for any initial point; if $\sigma\in(1,\sigma^*]$, then infinite convergence occurs along 2 (one-sided) asymptotes for all starting points in an open angle, while finite convergence occurs for all other starting points (except for the boundary of that open angle, where unstable infinite convergence is theoretically possible).

If one deals with the system of equalities $a^ix + b^i = 0$, $i = 1,2$, then if $\sigma\in(1,\sigma^*]$, the results are essentially identical (except for the finite convergence part). If one uses the Kacmarz method for this system of equalities, and if $\sigma\in(1,\sigma^*)$, identical results can be shown. In the relaxation method applied to both equalities or inequalities (for $\sigma\in(1,\sigma^*]$) one uses the constraints one after the other (maybe after a few iterations in the case of equalities; in the case of inequalities, this applies only if convergence is not finite, and if so, the cyclic order starts at the first iteration). And thus for $\sigma\in(1,\sigma^*]$, the convergence theory is identical to that given by the S.O.R. technique. We should mention that σ^* is the optimal value of the relaxation parameter, as predicted by the S.O.R. theory.

If $\sigma\in(\sigma^*,2)$, then the S.O.R. theory shows a rate of convergence of $\sigma-1$ which is the modulus of the two complex eigenvalues of the operator R, as defined in section 4. If $\sigma\in(\sigma^*,2)$, then the case of the relaxation method of Agmon applied to the equalities has not been studied in a satisfactory manner: the indices of the equalities are not used in any easily recognizable cycles, and thus the whole theory of linear operators does not seem to help. Experiments have been performed (and also with subgradient optimization), but will be reported elsewhere. Improvements of rates of convergence over the S.O.R. theory seem typical.

2. An assignment problem will be studied in some detail. It has been chosen, because we think that it is the hardest one from the point of view of the relaxation method (note again that, when we talk of relaxation method, it implies that we make the unrealistic assumption that f^* is known).

This problem is defined by a cost matrix which is zero: $c_{ij} = 0$, $i,j = 1,\ldots,n$; we will denote this assignment problem by A_n^H.

Clearly:

$$f(x) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \min_j (-x_j) \tag{5.1}$$

$$= \sum_{i=1}^{n} x_i - n \max_j (x_j) \quad . \tag{5.2}$$

Note that these two expressions are not defined in terms of the same set of linear pieces:

In the first case

$$f(x) = \min_{i \in I} \{(a^i, x)\}$$

where $a^i \in R^n$ is any vector satisfying $(a^i, e) = 0$, (a^i, e^k) is an integer (+ or -) less than one, $k = 1,\ldots,n$, (e is a vector of ones, e^k is a vector with one in position k and zeros elsewhere).

In the second case

$$f(x) = \min_{i=1,\ldots,n} (v^i, x) \ ,$$

where $v^i = e - ne^i$, $i = 1,\ldots,n$.

The implications of this distinction are not earth-shaking, but they may lead to distinct sequences; this is related to the question: when there is more than one subgradient, which one is chosen? This depends on the exact way subproblems are formulated and solved.

The optimal set is $P = \{\alpha e : \alpha \text{ real}\}$; notice that $(\partial f(x), e) = 0$, $\forall x \in R^n$ and thus with either the relaxation method or subgradient optimization $(x^0, e) = (x^q, e) = (x^\infty, e)$, where x^∞, the point to which the sequence converges (if it does converge) is given by:

$$x^\infty = \frac{(x^0, e)}{n} e.$$

To simplify the notation, we will assume that $(x^0, e) = 0$, and thus "the" optimal point is $x^* = 0$. The whole sequence takes place in the (n-1)-dimensional subspace $S = \{x \in R^n : (e, x) = 0\}$; and thus the problem (5.1) or (5.2) is effectively a problem in (n-1) dimensions.

If $x \in S$ then $f(x) = -n \max_{i=1,\ldots,n} x_i$; and thus

$\partial f(x) = \{v^i\}$ iff $x_i > x_j \ \forall j \neq i, \quad j = 1,\ldots,n$.

It is possible to check that:

1. The extreme points of $\partial f(0)$ are v^i, $i = 1,\ldots,n$
2. $(v^i, v^j) = -n$ if $i \neq j$
   $|v^i|^2 = n(n-1)$
3. The solution of the n-1 equations $(v^i, x) = 0$, $i = 1,2, j-1, j+1,\ldots,n$ such that $x \in S$ is given by v^j.
4. $\mu = \tilde{\mu} = \frac{1}{n-1}$

Proof

$$\mu = \tilde{\mu} = \inf_{x \notin P} \inf_{u \in \partial f(x)} \frac{(u, x^*(x) - x)}{|u| |x^*(x) - x|} .$$

The cosine function is quasiconcave (on the domain where it is positive), and thus the inner infimum is attained at extreme points of $\partial f(x)$. But for every x, every extreme point of $\partial f(x)$ has norm n(n-1).

Thus
$$\mu = \tilde{\mu} = \inf_{x \notin P} \frac{f^* - f(x)}{\sqrt{n(n-1)}\,|x^*(x) - x|}$$

$$= \inf_{x \notin P} \left\{ \frac{n \max_i x_i}{\sqrt{n(n-1)}} \;:\; \sum_{i=1}^{n} x_i = 0,\ \sum_{i=1}^{n} x_i^2 = 1 \right\}$$

$$= \frac{1}{n-1}$$

(the inf max is attained for $(x_i = \frac{1}{\sqrt{n(n-1)}} \quad x_j = -\sqrt{\frac{n-1}{n}}$, $j \neq i)$, $i = 1,\ldots,n)$ Q.E.D.

5. The minimum μ or $\tilde{\mu}$ is attained for

$$x = \lambda v^j + Ke, \quad \forall \lambda > 0, \quad \forall K, \quad j = 1,\ldots,n \ .$$

6. $\sqrt{\frac{n}{n-1}}\, d(x,P) \leq f^* - f(x) \leq \sqrt{n(n-1)}\, d(x,P)$

the bound on the right is tight if $x = -\lambda v^j + Ke$, $\forall \lambda > 0$, $\forall K$, $j = 1,\ldots,n$ and on the left if $x = \lambda v^j + Ke$, $\forall \lambda > 0$, $\forall K$, $j = 1,\ldots,n$.

Thus the rate of convergence, as proved in [7] for the relaxation method is $\sqrt{1-\sigma(2-\sigma)\frac{1}{(n-1)^2}}$, and the sustainable rate of convergence [7], [21] for subgradient optimization is $z(\mu) = \sqrt{1-\left(\frac{1}{n-1}\right)^2}$. This is not very good. A better theory certainly exists.

We will try to show that improvements in the theory can probably be made by using the comparison with the S.O.R. technique, provided that all subgradients (except one) are used

in a cyclical order. What we will do here is to exhibit a sequence generated by the relaxation method (2.2) applied to A_n^H which is cyclical, and then indicate that this sequence is stable.

Let the initial point be:

$$x_1^o = \alpha \quad , \quad x_2^o = \alpha(\rho - \frac{\sigma}{n-1}) \quad , \quad x_3^o = \alpha(\rho^2 - \frac{\sigma}{n-1}\frac{1-\rho^2}{1-\rho})\ldots$$

$$x_{n-1}^o = \alpha(\rho^{n-2} - \frac{\alpha}{n-1}\frac{1-\rho^{n-2}}{1-\rho}) = -(\sigma-1)\frac{\alpha}{\rho}, \quad x_n^o = -\frac{\sigma}{n-1}\alpha\frac{1}{1-\rho} \quad .$$

Note that $x_1^o > x_2^o > \ldots > x_{n-1}^o > x_n^o$,

and that $\sum_{k=1}^{n} x_k^o = (x^o, e) = 0$. (We assume $\rho \neq 1$).

At the first iteration, the subgradient $v^1 = e - ne^1$ is chosen, and the second iterate will be:

$$x_1^1 = -(\sigma - 1)\alpha \ , \quad x_2^1 = \rho\alpha \ , \quad x_3^1 = \rho\alpha(\rho - \frac{\sigma}{n-1} \ldots$$

$$x_{n-1}^1 = \rho\alpha(\rho^{n-3} - \frac{\sigma}{n-1}\frac{1-\rho^{n-3}}{1-\rho}) \ , \quad x_n^1 = -\frac{\sigma}{n-1}\rho\alpha\frac{1}{1-\rho} \quad .$$

where

$$-(\sigma-1)\alpha = \rho\alpha\left[\rho^{n-2} - \frac{\sigma}{n-1}\frac{1-\rho^{n-2}}{1-\rho}\right] \quad . \tag{5.3}$$

x^1 is simply ρx^o with a shift of one down of all component indices, except the n-th, (which remains last), and the 1-st which becomes the (n-1)-st. Condition (5.3) simply expresses the fact that the formula is consistent and that the sequence will repeat itself ((5.3) is an equation between ρ and σ).

Clearly,

$$x^{(k+1)} = \sigma P x^{(k)} \tag{5.4}$$

where $p_{nn} = 1$

$p_{i,i-1} = 1$ for $i = 2,\ldots,n-1$

$p_{1,n-1} = 1$

$p_{i,j} = 0$ otherwise,

and also

$$x^{(k+n-1)} = \rho^{n-1}P^{n-1}x^{(k)} = \rho^{n-1}x^{(k)}.$$

The subgradients are used in the cyclical sequence

$$v^1, v^2, \ldots, v^{n-1}, v^1, v^2, \ldots, v^{n-1}, \ldots$$

It can be checked that:

1. $d(x^q, P) = \rho^q \alpha \sqrt{\frac{n\sigma(2-\sigma)}{n-1}} \quad \frac{1}{1-\rho^2}$

2. $\mu(x^q) = \tilde{\mu}(x^q) = \sqrt{\frac{1-\rho^2}{\sigma(2-\sigma)}}$

3. $x^{k+\ell(n-1)} = -\rho^{\ell(n-1)} \left(\sum_{i=1}^{n-1} \tilde{\eta}_i^k v^i \right) \Big/ \sqrt{n(n-1)}$

$$k = 0,\ldots,n-2$$

$$\text{where } \tilde{\eta}_i^k = \frac{1}{n} x_i^k - \frac{1}{n} x_n^k > 0 \qquad \forall i = 1,\ldots,n-1 \qquad (5.5)$$

4. The vector $(\tilde{\eta}_i^0,\ i=1,\ldots,n-1)$ is an eigenvector of the S.O.R. operator $R = (I-\sigma\tilde{L})^{-1}((1-\sigma)I+\sigma\tilde{U})$ where $\tilde{\Gamma}$ is the Grammian constructed on $(v^1,\ldots,v^{n-1})/\sqrt{n(n-1)}$.

The operator R of the S.O.R. method has a characteristic equation (in variable ω) which is:

$$\det[(\omega + \sigma - 1)I - \sigma\tilde{U} - \sigma\omega\tilde{L}] = 0$$

$$\text{or } \det \begin{vmatrix} \omega+\sigma-1 & -\sigma/(n-1)\ldots \\ -\sigma\omega/(n-1) & \omega+\sigma-1 \quad \ldots \\ \vdots & \vdots \end{vmatrix} = 0$$

which can be computed as

$$\frac{\omega(\omega + \sigma - 1 + \sigma/(n-1))^{n-1}}{\omega - 1} = \frac{(\omega + \omega\sigma/(n-1) + \sigma - 1)^{n-1}}{\omega - 1} \quad . \tag{5.6}$$

This polynomial can be related to (5.3) which is obtained by letting $\rho^{n-1}=\omega$ and taking the (n-1)-th (real, positive) root of (5.6):

$$\begin{aligned} 0 &= \rho^{n-1} - \frac{\sigma}{n-1}(\rho^{n-2} + \ldots + \rho) + \sigma - 1 \\ &= \frac{\rho(\rho^{n-1} + (\sigma-1) + \sigma/(n-1)) - (\rho^{n-1}(1+\sigma/n-1) + \sigma - 1)}{\rho - 1} \quad . \end{aligned} \tag{5.7}$$

Every root ρ of (5.7) gives rise to a root of (5.6) given by $\omega=\rho^{n-1}$. If (5.7) has (n-1) distinct roots ρ_i, such that $\omega_i=\rho_i^{n-1}$ are distinct, then ω_i, i=1,...,n-1 are the (n-1) roots of (5.6).

It is not too surprising that (5.3) is identical to (5.7). In order that the sequence of points given by (5.2), (5.3), (5.4) be observable in practice, the stability of that sequence must be proved. Stability follows if two conditions are met:

Condition 5.1 The largest modulus root of (5.6), say ω^+, must be real, positive, unique and larger in modulus than all other roots of (5.6) (the same statement holds mutatis mutandis for $\rho^+ = (\omega^+)^{1/(n-1)}$ if all roots of (5.7) are distinct); the vector $\tilde{\eta}^0$ given by (5.5) is the eigenvector of the operator R corresponding to ω^+.

Condition 5.2 For a neighbourhood of $\tilde{\eta}^0$ (given by 5.5), one needs that the same cyclic use of subgradients as given by (5.2), (5.3) and (5.4) be preserved for all iterations. A sketch of the long and uneventful proof follows (if one assumes all eigenvalues to be distinct so that a full set of eigenvectors exists): let $\bar{\eta}^0 = \tilde{\eta}^0 + \varepsilon^0$, where $\tilde{\eta}^0$ is given by (5.5) and ε^0 is a linear combination of the eigenvectors of R (excluding $\tilde{\eta}^0$); the equations expressing that at iteration $k+\ell(n-1)$ (where $0 \leq k \leq n-1$), the subgradient to be used is v^k, are linear inequalities in terms of

variables which are the components (possibly complex) of ε° in terms of the nondominating eigenvectors of R and are such that the coefficients of the variables are proportional to the ℓ-th power of the corresponding eigenvalues (multiplied by constants which depend only on k and R), while the constant term is proportional $(\omega^{+})^{\ell}$ (times constant which depend on k and R). Every one of these inequalities for every k=0,...,n-1, ℓ=0,...,∞, each of which contains zero, contains a neighbourhood of zero. Thus, if $\varepsilon_{\circ}$ belongs to some neighbourhood of zero, the indices are used in the cyclical order 1,2,...,n-1,1,2,...,n-1 ... .

A proof of this second condition for stability would become interesting if some useful characterization of these neighbourhoods of attraction into a given cyclical order could be given.

We will now move to experimentation on A^{H}_{11}. The number 11 was chosen because the routine (Jenkins-Traub) we had available to us broke down for n rather small on the polynomials given by (5.6) or (5.7).

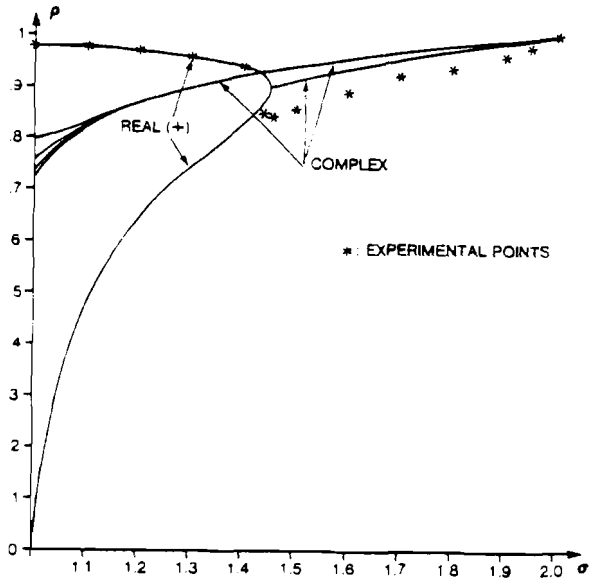

Figure 2. Roots of (5.7) and experimental rates of convergence of the relaxation method applied to A^{H}_{11}.

For n=11, Figure 2 shows the modulus of the roots of (5.7) for $\sigma \in [1,2]$. It was checked that the 10-th power of these roots are the roots of (5.6) because they are distinct (except at isolated points, where, by continuity, it does not matter), and in fact (5.6) was solved and it confirmed this (note: the routine we used broke down on (5.6) for values of σ above 1.7). The roots of (5.7) are shown here because they represent what has been called the rate of convergence of the relaxation method (2.2), while the roots of (5.6) represent the rate of convergence of cycles (n-1 iterations) of the S.O.R. method; comparisons with experiments reported in other works can be made.

It thus follows that we have shown that for $\sigma \in [1, 1.43]$ the rate of convergence of the relaxation of (2.2) applied to A^H_{11}, as function of σ, and for a set of initial points of dimension 11, is given by the graph of the largest root.

A similar graph could be drawn for any value of n. Experiments and the theory of polynomials show that (for $\sigma \in (1,2)$):

1. (5.7) has either two, one double, or no positive real roots.
2. (5.7) has one negative root if n is even.
3. The complex roots of (5.7) get to be extremely close in modulus but not identical (except, of course, for the conjugate pairs), as σ increases from 1 to 2, while the negative root (for n even) is always less in modulus than the complex roots, but comes very close in modulus when σ goes from 1 to 2.
4. The two complex roots which originate from the positive real roots, when σ increases, remain, in modulus, significantly less than the other complex roots.
5. The largest modulus root of (5.7) has a modulus greater than or equal to $(\sigma-1)^{1/(n-1)}$ (this is classical in the S.O.R. theory).

It should be noted that if $\sigma < 1$, then R is a positive matrix, and thus the existence of a dominating positive real root follows from the Perron (Froebinius) theorem.

Experiments to verify all of this were made, and the observed rate of convergence as a function of σ is plotted (Figure 2). For $\sigma \in [1, 1.43]$ the experiments confirm the theory quite exactly (if $\sigma \in (0,1)$ the same would be true). What experiments also indicate is that, after a few initial iterations, n-1 subgradients are used in a cycle (of order n-1); of course <u>any</u> (n-1) subgradients could be used in <u>any</u> cycle of order n-1. In this problem, cycling seems to be an attracting behaviour for all starting points (if $\sigma \leq 1.43$), in that cycling occurs after some iterations.

Note that the rate of convergence obtained for $\sigma=1.43$ is about .93, while the rate given by previous theories (for $\sigma=1$) is $\sqrt{1-\tilde{\mu}^2} = \sqrt{\frac{99}{10}} \simeq .995$; and thus the theory of acceleration is a significant improvement.

We should mention that this example is very particular: the matrix R depends, in general, on the set of subgradients used in a cycle, and on the ordering of the cycle; but in this case, because of the fact that (v^i, v^j)=Constant if $i \neq j$, R does not change. It would imply that, in general, a whole family of matrices R (and of polynomials) would need to be studied; and of course that the idea of <u>one</u> rate of convergence disappears (there might be many). Figure 2 also clearly shows that if $\sigma \in (1.43, 2)$, the rate of convergence is quite different from the modulus of the eigenvalues of R (in fact, the concept of rate of convergence in this case has not been exactly defined).

The whole set of n subgradients is used in a noncyclical way, with no recognizable pattern. It also means that the theory of linear operators does not seem to provide any insights about convergence theory (of course convergence takes place at a rate of at most $\sqrt{1-\sigma(2-\sigma)\mu^2}$, but this is a rate which ignores the interactions between successive iterates). It also shows that in this case, when noncycling behaviour occurs, an extra acceleration (beyond the acceleration given by the S.O.R. theory) takes place. At this point it is quite unclear how this could be studied theoretically.

3. Almost all that was said in the previous section can be interpreted in terms of the solution of the system of linear equalities $(v^i,x)=0$, $i=1,\ldots,n-1$ where $x\in S=\{x\in R^n: \sum_{j=1}^{n} x_j=0\}$. It shows that for values of σ below the optimal relaxation parameter (1.43), choosing indices according to the maximum residue gives rise to the same rate of convergence as a cyclic choice of indices.

Before ending this section, we would like to emphasize that we have dealt with only a few examples, and that not all that has been observed can be extended verbatim. A general theory might not be unattainable but it seems to be exceedingly complex.

6. SUBGRADIENT OPTIMIZATION

In most optimization problems the optimal value of f^* is not known, and thus the relaxation method (2.2) is not implementable. Part of the interest in it comes from the fact that the two theories on rates of convergence (subgradient optimization and relaxation method) appear to be related. Bounds on the rates of convergence of the two methods were shown to be identical, under some conditions, in [8]; those bounds did not account for any acceleration properties.

No theory has yet been developed which would show the acceleration of subgradient optimization. So back to experimentation on the example of section (5.2) (A_{11}^H).

Note that in the sequence of iterates given by (5.3), it is easy to see that $|x^{(k+1)}-x^{(k)}| = \frac{\sigma}{\sqrt{n(n-1)}}\sigma\rho^k$, and thus that if one started subgradient optimization from the point $x^{(0)}$ with $\lambda_k=\frac{\sigma}{\sqrt{n(n-1)}}\alpha$ and σ and ρ related by (5.3), then subgradient optimization could theoretically generate the same sequence of points (5.4) as the relaxation method. This did not happen in practice, because rounding errors made the sequence unstable.

What was observed in practice is that the sustainable rate of convergence of subgradient optimization was around .92, even

though the sequence of iterates used all n subgradients in a pattern with no discernable regularity. With subgradient optimization rates of convergence between .92 and .9 were sometimes achieved, but it required that λ_0 be chosen very carefully, as convergence failed for high values of λ_0 as well as for low values of λ_0. This example (as well as TR 48) clearly indicates a relationship between the rates of convergence of subgradient optimization and of the relaxation method; and thus that some form of acceleration occurs in subgradient optimization.

7. CONCLUSIONS, CONJECTURES AND AREAS OF FUTURE RESEARCH

Subgradient optimization has worked well in practice on quite a few problems generated from combinatorial problems. We think that a complete explanation of this practical effectiveness will require four parts.

Part 1

The theory of acceleration of the classical S.O.R. theory extends, sometimes, to the relaxation method for inequalities and also to subgradient optimization. This is essentially what has been done in this paper, through a mixture of theory, examples and experiments. A general theory would need to add a few layers of complexity to the S.O.R. theory (in cases where it is not very complete, i.e., where matrices are positive definite), and, if the behaviour is noncyclical, then the S.O.R. theory does not seem to help.

Part 2

For some classes of combinatorial problems there are universal bounds on μ and $\tilde{\mu}$. For instance, we conjecture that for any assignment problem, $\mu=\tilde{\mu}=\frac{1}{n-1}$.

This conjecture is a bit misleading: the values of the condition numbers close to the optimal set (μ_c and $\tilde{\mu}_c$) could be much above $1/(n-1)$. In this paper we showed that μ, $\tilde{\mu}$, μ_c or $\tilde{\mu}_c$ are not very accurate indicators of the rate of convergence; so we will also conjecture that the assignment problem A_n^H (for which $\mu_c = \tilde{\mu}_c = 1/(n-1)$) will give the lowest rate of convergence

for both the relaxation method and subgradient optimization, if the acceleration phenomenon is taken into account.

Part 3

To every problem one could associate a rate of convergence achievable by subgradient optimization. And thus we could make probabilistic statements for the rates of convergence of subgradient optimization if the data of a problem varies according to certain probability distributions.

In fact every experiment with subgradient optimization could be viewed as a simulation experiment designed to estimate the probability distribution of the convergence rates. One should point out, that the experiments reported in the published literature are a very biased sample as it is a practice (sometimes unfortunate) not to report failed experiments.

For instance, problem A^H_{48} was studied as a few assignment problems of this size have been studied by Held, Wolfe and Crowder [12]. In [7], it was shown that for A 48 (an assignment problem studied also in [12]) the best rate of convergence was .85. For A^H_{48}, the roots of (5.7) were studied, even though the routines for solving polynomials broke down. We found, by using Rouché's theorem (of the theory of complex variables) that the minimum value of ρ, subject to the conditions that it be the strictly dominating real positive root of (5.7), was around .9926 for a value of σ around 1.693. (Rouché's theorem was used to show that all roots of (5.7), except .9926, are in a complex circle $|z| \leq .9926-\varepsilon)$. Comparing .9926 to .85 seems to indicate that "random" assignment problems will give rise to rates of convergence which are reasonably good, on the average.

Other classes of combinatorial problems should be studied; in fact an assignment problem was chosen here, because it is reasonably easy to study, and we should at least say that it would be quite absurd to solve assignment problems by using subgradient optimization.

Part 4

We will also conjecture that the hardest problems from the point of view of subgradient optimization not only have low probability, but also are the easiest from a combinatorial point of view.

For instance, A_n^H is trivial, and it is hard from the point of view of subgradient estimation. It should be possible to relate μ_c and $\tilde{\mu}_c$ to the nature of the solution to the primal assignment problem.

It also indicates why subgradient optimization works quite well within a branch and bound framework [10]: subgradient optimization seems to work better when the combinatorial nature of the problem is hard.

REFERENCES

1. Agmon, Shmuel. 1954. The relaxation method for linear inequalities. Canadian J. Math., Vol. 6, pp. 382-392.

2. Allen, D.N. de G. 1954. Relaxation Methods in Engineering and Science, McGraw-Hill, New York.

3. Camerini, P.M., L. Fratta and F. Maffioli. 1975. On improving relaxation methods by modified gradient techniques. In Nondifferentiable Optimization, Mathematical Programming Study 3, edited by M.L. Balinski and P. Wolfe, North-Holland, Amsterdam, pp. 26-34.

4. Eremin, I.I. 1962. An iterative method for Chebyshev approximations of incompatible systems of linear inequalities. Dokl. Akad. Nauk SSSR, Vol. 143, pp. 1254-1256. English translation: Soviet Math. Dokl., Vol. 30, pp. 570-572.

5. Eremin, I.I. 1966. On systems of inequalities with convex functions on the left sides. Izv. Akad. Nauk SSSR Ser. Mat., Vol. 30, pp. 265-278. English translation: Amer. Math. Soc. Transl., Vol. 88, pp. 67-83.

6. Goffin, J.L. 1971. On the finite convergence of the relaxation method for solving systems of inequalities. Operations Research Center report ORC 71-36, Univ. of California at Berkeley.

7. Goffin, J.L. 1977. On convergence rates of subgradient optimization methods. Mathematical Programming, Vol. 13, pp. 329-347.

8. Goffin, J.L. 1978. Nondifferentiable optimization and the relaxation method in nonsmooth optimization. In Nonsmooth Optimization, edited by C. Lemaréchal and R. Mifflin, Pergamon Press, Oxford, pp. 31-50.

9. Goffin, J.L. 1977. The relaxation method for solving systems of inequalities. Working paper 77-54, Faculty of Management, McGill University, Montreal, Quebec.

10. Held, M. and R.M. Karp. 1971. The traveling-salesman problem and minimum spanning trees: part II. Mathematical Programming, Vol. 1, pp. 6-25.

11. Held, M., R.M. Karp and P. Wolfe. 1972. Large-scale optimization and the relaxation method. In Proceedings of the 25th National ACM Meeting, held at Boston, Mass., August 1972.

12. Held, M., P. Wolfe and H. Crowder. 1974. Validation of subgradient optimization. Mathematical Programming, Vol. 6, pp. 62-88.

13. Herman, G.T., A. Lent and S.W. Rowland. 1973. ART: mathematics and applications. A report on the mathematical foundations and on the applicability to real data of the algebraic reconstruction techniques. J. Theor. Biol., Vol. 42, pp. 1-32.

14. Kahan, W. 1958. Gauss-Seidel methods of solving large systems of linear equations. Doctoral thesis, University of Toronto.

15. Kennington, J. and M. Shalaby. 1977. An effective subgradient procedure for minimal cost multicommodity flow problems. Management Science, Vol. 23, No. 9.

16. Motzkin, Th. and I.J. Schoenberg. 1954. The relaxation method for linear inequalities. Canadian J. Math., Vol. 6, pp. 393-404.

17. Lemaréchal C. and R. Mifflin (Editors). 1978. Nonsmooth Optimization. Proceedings of an IIASA Workshop, 28 March to 8 April, 1977. Pergamon Press, Oxford.

18. Nicolaides, R.A. 1974. On a geometrical aspect of S.O.R. and the theory of consistent ordering for positive definite matrices. Numer. Math., Vol. 23, pp. 99-104.

19. Oettli, W. 1972. An iterative method, having linear rate of convergence, for solving a pair of dual linear programs. Mathematical Programming, Vol. 3, pp. 302-311.

20. Polyak, B.T. 1969. Minimization of nonsmooth functionals. Zh. Vychisl. Mat. Mat. Fiz (USSR Computational Mathematics and Mathematical Physics), Vol. 9, pp. 509-521.

21. Shor, N.Z. 1968. The rate of convergence of the generalized gradient descent method. Kibernetica, Vol. 4 (3), 98-99. English translation: Cybernetics, Vol. 4 (3), 79-80.

22. Shor, N.Z. 1976. Generalized gradient methods for non-smooth functions and their applications to mathematical programming problems. Ekonomika i Matematicheskie Metody, Vol. 12 (2), 337-356 (in Russian).

23. Southwell, R.V. 1940. Relaxation Methods in Engineering Science. Oxford University Press, Oxford.

NUMERICAL EXPERIMENTS IN NONSMOOTH OPTIMIZATION

Claude Lemaréchal
INRIA, France

1. INTRODUCTION

The aim of this paper is to show the behavior of various methods for nonsmooth optimization, applied to various problems. In Section 2 we present the methods, in Section 3 the problems and the results. In Section 4 we draw some conclusions.

Notation. The function to be minimized (without constraints) is $f(x)$, $x \in R^N$ and we denote by $(.,.)$ (resp. $|\cdot|$) the scalar product (resp. the norm) in R^N. The sequence of iterates generated by an algorithm is $x_1, \ldots, x_n$ and $g_1, \ldots, g_n$ denote the subgradients that it uses; objective values and subgradients are computed simultaneously at each execution of a subprogram, characteristic of each problem, but the same for each algorithm. When no confusion is possible, we will denote by x the current iterate x_n, and by x_+ the forthcoming iterate x_{n+1}; the same notation g and g_+ will hold for the subgradients; i will index iterations performed before the current one. The direction moved from x_n will be d or d_n, and the stepsize a or a_n.

2. THE METHODS

The first two methods are heuristic, in that they are supposed to be applicable only to differentiable objective functions.

2.1 BFGS. In this method the direction is $d = -Hg$ and the next iterate is

$$x_+ = x - aHg \quad ,$$

where H is the $N \times N$ quasi-Newton matrix computed by the Broyden-Fletcher-Goldfarb-Shanno formula (Powell 1975). One has $H_1 = I$, the identity matrix and setting $y = g_+ - g$

$$H_+ = H - (d\tilde{y}H + Hy\tilde{d})/(d,y) + [a + (y,Hy)/(d,y)]d\tilde{d}/(d,y) \quad ,$$

where $\tilde{y}$ denotes the transpose.

The stepsize a is computed as follows:

• One starts from an initial guess $a > 0$. We take the one suggested by Fletcher (Wolfe 1975): $a = 2(f_n - f_{n-1})/(g, d)$, i.e., the stepsize that would minimize f along d, if f were quadratic, and if its corresponding decrease were equal to the decrease obtained at the previous iteration.

• One tests if the stepsize meets a certain stopping criterion, namely, if

$$\begin{cases} (g_+, d) \geq m_1(g, d) \quad , \text{ and} \\ f_+ \leq f + m_2 a(g, d) \quad , \end{cases} \tag{1}$$

where $0 < m_2 < m_1 < 1$ are preassigned coefficients (in fact $m_1 = 0.7$, $m_2 = 0.1$).

• If a is not convenient, one performs a series of adjustments until an a is found that satisfies (1). Noting that, at a point $x = x_n + ad_n$, $(g(x_n + ad_n)d_n)$ is the derivative with respect to a of the one dimensional function $f(x_n + ad_n)$, (1) means that the objective has sufficiently decreased, and the derivative sufficiently increased, relative to the initial slope $(g_n, d_n) = (g, d)$.

The way the adjustments are made is rather standard and not significant enough to be described here. Suffice it to say that it is a safeguarded cubic interpolation.

These characteristics make BFGS equivalent to the subroutine VA13A of the Harwell Library. In nonsmooth optimization, such a method is heuristic. In particular d_n may be uphill. Therefore, the only possible stopping criterion is the case of failure: the algorithm is run until any positive a satisfying (1) cannot be found. This event means that x_n is (very close to) a kink. However, it is likely to occur only when x_n is close to a minimum, thanks to the fact that the line search is by no means an attempt to minimize f along d_n. This intuitive statement is supported by numerical validation, which makes quasi-Newton methods, well-known and easy to program, reasonable "faute de mieux" for nonsmooth optimization.

2.2 SHOR. The second method is Shor's (1971) dilation of the space along the difference of two successive gradients. Here d is again of the form -Hg and

$$x_+ = x + a\, d/|d| \quad .$$

The matrix H is a product of orthogonal affinities along $g_1 - g_{i-1}$. We refer to (Shor and Shabashova 1972) for the complete statement of the algorithm.

In this method there is no line-search, the stepsize is computed off-line as a geometric sequence:

$$a_n = a_1 q^{n-1} \quad , \quad q \in]0, 1[\quad ,$$

where the parameters a_1 and q have to be tuned. Unfortunately we do not know any method to do it properly, so, in the present study, we had to do it empirically by running the algorithm with several values of a_1 and q--for each problem--and taking the combination that gives apparently the best results. Of course, this is possible only when the optimal solution is known. Therefore, we cannot really speak of an implementable algorithm.

However, as with 2.1, numerical evidence will demonstrate excellent behavior of this method.

The next methods are all based on the use of ε-subgradients.

2.3 EPSDES. In this method, which could be called method of ε-descent (Lemaréchal 1974), the direction is computed by orthogonal projection of the origin onto a set of subgradients, and the line-search is an approximate one-dimensional minimization.

If $G = \{g_1, \ldots, g_k\}$ is a finite set in R^N, we denote by NrG the unique solution of

$$\begin{cases} \min \frac{1}{2}|s|^2 \ , \\ s = \sum_{i=1}^{k} \lambda_i g_i \ , \\ \lambda_i \geq 0 \ , \quad i = 1, \ldots, k \ , \\ \sum_{i=1}^{k} \lambda_i = 1 \ . \end{cases}$$

In EPSDES, a number $\varepsilon > 0$ is managed along the iterations. It is normally kept fixed, and it is divided by 10 when it is recognized that x_n (approximately) minimizes f within ε.

An iteration consists of finding a direction of ε-descent d_n, i.e., such that there exists $a > 0$ satisfying

$$f(x_n + ad_n) \leq f(x_n) - \varepsilon \ ,$$

and then x_n is updated to x_{n+1} such that a decrease of ε is obtained.

Constructing such a direction is itself a subalgorithm, made of a series of line-searches along trial directions $d_n^1, \ldots, d_n^k$ until the proper decrease is obtained. At x_n, having generated $g_n^1, \ldots, g_n^k$ (one starts with $g_n^1 = g_n$, subgradient

at x_n) one computes

$$s_n^k = \mathrm{Nr}\ \{g_n^1, \ldots, g_n^k\} \text{ and } d_n^k = -s_n^k \quad .$$

Then a line-search is performed along d_n^k

- if a decrease by ε is obtained, then the iteration is finished,
- if not, a new g_n^{k+1} is obtained, that is approximately a subgradient at the minimum of f along d_n^k; g_n^{k+1} is added to G and a new direction d_n^{k+1} is computed.

Of course, the line-search is possible only if $s_n^k \neq 0$. Hence, a convergence parameter $\eta > 0$ is used and the test

$$|s_n^k| \leq \eta \quad ,$$

is checked. If it is met, one has the approximate optimality condition

$$f(y) \geq f(x_n) - \varepsilon - \eta|y-x_n| \quad , \quad \forall_y \quad R^N \quad , \qquad (2)$$

(this is true only is f is convex; in the nonconvex case, the method is heuristic).

We refer to (Lemaréchal 1974) for an accurate statement of the algorithm. We consider it rather special. It strongly relies upon convexity and appears as extremely heavy. Only a coarse experimental version has been programmed, and numerical results will show that, although it is very reliable, the convergence is usually very slow.

2.4 CHAINE. This method is essentially that of (Lemaréchal 1975), but with a line-search based on (Wolfe 1975). This line-search is rather fundamental because it is a direct extension of 2.1, and is identical for 2.4, 2.5, and 2.6. Therefore we expose it first. The only difference with (Wolfe 1975) is the test for null step.

At x_n are given the direction d_n and the two coefficients $m_1 = 0.2$, $m_2 = 0.1$. Also a tolerance $\varepsilon' > 0$ is given and one has on hand a number $p < 0$ that estimates $f'(x_n; d_n)$ (normally $p = f'(x_n, d_n) = (d_n, g_n)$). The aim of the line-search is to produce a point $y = x_n + ad_n$ and a subgradient g at this point such that:

$$(d_n, g) \geq m_1 p \quad , \tag{3}$$

and either

$$f(y) \leq f(x_n) + am_2 p \quad , \tag{4}$$

or

$$|f(x_n) - f(y) + a(g, d_n)| \leq \varepsilon' \quad . \tag{5}$$

In cases (3), (4) a normal descent step is made from x_n to $x_{n+1} = y = x_n + ad_n$. It is a serious step (compare with (1)).

If x_n is (very close to) a kink, it may happen that $m_2 p < f'(x_n, d_n)$ and then (4) is impossible to obtain. Then a (sufficiently small) stepsize is found satisfying (3), (5); x_{n+1} is taken as x_n, only the new gradient is used for the forthcoming iteration. It is a null step.

As for the direction, it is $d = -NrG$, where $G = \{g_1, \ldots, g_n\}$ and is reinitialized when a certain test is met, namely

$$(g_n, x_n - x_1) > \varepsilon \quad .$$

The convergence parameter ε is managed throughout the algorithm. When the direction is computed, the parameters for the line-search are $p = -|d_n|^2$ and $\varepsilon' = 0.1\,\varepsilon$. The set G is reintialized on two more occasions:

- when $|d| \leq \eta$ (see (2)). Then ε is also divided by 10,
- when the number of subgradients to be stored exceeds a

preassigned limit; this prevents the algorithm from needing an infinite amount of storage.

Note that this algorithm is apparently similar to 2.3. However, there are two characteristics that make it fundamentally different:

- the test for descent,
- the test for reinitializing G,

(in 2.3, G is reinitialized every time a descent is made).

2.5 DYNEPS. This algorithm was first presented in (Lemaréchal 1976). See also (Lemaréchal 1979). Its line-search is identical to that of 2.4, and a convergence parameter ε is also managed. The direction is computed in a slightly different way than 2.3 and 2.4.

Call y_i the point at which the subgradient g_i has been computed (observe that $y_i = x_i$ if the step a_{i-1} was a serious one; for each null step, $y_i = x_{i-1} + a_{i-1}d_{i-1} \neq x_i$; for $i = 1$ $y_1 = x_1$). Then at each iteration define the numbers α_i^n, $i = 1, \ldots, n$ as follows:

$$\alpha_1^1 = 0 \quad \text{and, for } n = 1, \ldots,$$

$$\alpha_{n+1}^{n+1} = \begin{cases} 0 & \text{if } a_n \text{ is a serious step} \\ |f(x_n) - f(y_{n+1}) + a_n(g_{n+1}, d_n)| & \text{if it is a null step.} \end{cases}$$

For each serious iteration n, update α_i^n to

$$\alpha_i^{n+1} = |\alpha_i^n + f(x_{n+1}) - f(x_n) - (g_{n+1}, x_{n+1} - x_n)| \quad ,$$

$$i = 1, \ldots, n \quad ,$$

(note that this formula would leave α_i^n unchanged if it were applied at a null iteration). These formulae allow to define α_i^n recursively at each iteration. There is at least one i for which $\alpha_i^n = 0$ (it is the index of the last serious iteration performed before the present one) and $\alpha_n^n \leq \varepsilon'$ (see (5)).

If the absolute values were neglected, one would have

$$\alpha_i^n = f(x_n) - f(y_i) - (g_i, x_n - y_i) \quad ,$$

which is the error made at x_n when f is linearized at x_i. In the convex case, these absolute values do not play any part.

Then the direction finding problem is now $s_n = NrG(\varepsilon)$ where

$$G(\varepsilon) = \{g = \sum_{i=1}^{n} \lambda_i g_i \ / \ \sum \lambda_i = 1, \ \lambda_i \geq 0 \quad , \quad \sum \lambda_i \alpha_i^n \leq \varepsilon\} \quad .$$

More precisely, the algorithm is as follows:

x_1, $g_1 \in \partial f(x_1)$ are given, together with the tolerances $\underline{\varepsilon} > 0$, $\eta > 0$, and the coefficients $m_1 = 0.2$, $m_2 = 0.1$. Set $\overline{\varepsilon} = +\infty$, $n = 1$, $\alpha_1 = 0$. Choose some ε in $[\underline{\varepsilon}, \overline{\varepsilon}]$.

Step 1. Solve

$$\begin{cases} \min \left| \sum_{i=1}^{n} \lambda_i g_i \right|^2 \quad , \\ \sum \lambda_i = 1 \quad , \quad \lambda_i \geq 0 \quad , \\ \sum \lambda_i \alpha_i \leq \varepsilon \quad , \end{cases}$$

and let $s = \sum \lambda_i g_i$ be the solution. Also call $u \geq 0$ the multiplier of the last constraint.

Step 2. If $|s| > \eta$ then set $d_n = -s$ and go to Step 3. Otherwise if $\varepsilon \leq \underline{\varepsilon}$ then stop. Otherwise take $\varepsilon = \max(\underline{\varepsilon}, 0.1\varepsilon)$ and go to Step 1.

Step 3. Apply the line-search of § 2.4 with $p = -[|d_n|^2 + u\varepsilon]$ and $\varepsilon' = 0.1\varepsilon$.

Obtain $y = x_n + a_n d_n$ and $g_{n+1} \in \partial f(y)$.

In cases (3), (4) go to Step 4. In case (3), (5) go to Step 5.

Step 4. (serious iteration). Set $x_{n+1} = y$, $\alpha_{n+1} = 0$. Change the α_i's $i = 1, \ldots, n$; choose a new $\varepsilon \in [\underline{\varepsilon}, \bar{\varepsilon}]$. Set $n = n + 1$ and go to Step 1.

Step 5. (null iteration). Set $x_{n+1} = x_n$; compute α_{n+1}, set $n = n + 1$ and go to Step 1.

This algorithm is not totally defined because the choice of ε at Step 4 is somewhat arbitrary. We have studied several possibilities for this choice:

ε is not changed in Step 4 (thus the only occasion at which ε is changed is in Step 2, when it is divided by 10), (6)

$$\varepsilon_+ = a|s|^2 \quad , \tag{7}$$

$$\varepsilon_+ = a[|s|^2 + u\varepsilon] \quad , \tag{8}$$

$\varepsilon_+ = \Delta f$ where $\Delta f > 0$ is the decrease of f obtained during the last serious iteration , (9)

$\varepsilon_+ = k[f(x_n) - \min f]$ where k is a fixed number in $]0, 1]$ (this last rule supposed that the optimal cost min f is known; we have tested $k = 0.1$, $k = 0.5$ and $k = 1$) . (10)

2.6 BFEPS. Finally we have tested a rough version of the algorithm presented in (Lemaréchal 1978). As in 2.1, a quasi-Newton matrix H is updated at each serious iteration, by the BFGS formula. Then $d = -Hs$ where $s = \sum \lambda_i g_i$ is the solution of

$$\begin{cases} \min \frac{1}{2} \left(\sum \lambda_i g_i, H \sum \lambda_i g_i\right) + \sum \lambda_i \alpha_i \\ \sum \lambda_i = 1 \quad , \quad \lambda_i \geq 0 \quad . \end{cases}$$

As for the line-search, it is the same as in 2.4 except that, instead of (5), the test for null step is $a \leq \frac{1}{10}$.

3. THE TESTS

We now present application of these algorithms to the test problems of (Lemaréchal and Mifflin eds. 1979). Since each algorithm is applied to each test-problem, the study is comparative. However, the results should not be considered as reliable enough to allow for accurate comparative conclusions. The reason is that the algorithms tested here are experimental, only CHAINE being a polished product. Thus our presentation is more an illustration of the behavior of various methods, than a normative analysis of their respective performances. Only very large differences (in speed of convergence for example) are conclusive.

The speed of a method is characterized by two numbers:

- number of line-searchers (i.e. number of times a direction is computed),
- number of computations of function-gradient.

We think that the second one is probably the most significant, since nonsmooth optimization seems normally devoted to problems in which function-gradient are expensive to compute.

We now review the problems and give the results.

3.1 MAXQUAD. In this problem, $x \in R^{10}$ and f is the maximum of five convex quadratics:

$$f(x) = \max_{k=1}^{5} (A_k x, x) - (b_k, x) \quad .$$

See Lemaréchal and Mifflin eds. 1979) for the definition of A_k and b_k.

Table 1 displays the results for methods 2.1 to 2.6. Each line of the table corresponds to an iteration (line-search) and gives, for each method, the cumulative number of times function-gradient have computed, and the current value of the objective function.

The parameters of 2.2 are $a_1 = 10$, $q = 0.95$. For method 2.5, the rule for ε at Step 4 is (9). Other rules for the same method are exhibited in Table 2, which reads as Table 1.

Iter.	BFGS		SHOR		TPSDES		SHANNO		DYNEPS		BFEPS	
1	1	5337	1	5337	1	5337	1	5337	1	5337	1	5337
2	2	162	2	1228	3	162	2	162	2	162	2	162
3	5	58	3	105	6	87	5	88	3	51	5	58
4	7	38	4	67	9	62	18	61	5	14	7	38
5	9	33	5	30	13	3.85	20	56	8	4.7	9	33
6	11	32	6	9.5	17	3.85	22	7.4	17	2.2	11	32
7	14	31	7	4.3	19	3.85	25	3.16	19	1.26	14	31
8	15	31	8	4.3	22	3.85	51	2.69	21	-.077	15	31
9	16	31	9	4.3	27	3.85	54	2.66	27	-.739	16	31
10	17	30	10	4.3	29	3.85	57	2.61	32	-.747	17	30
11	18	29	11	2.9	35	3.85	61	.251	35	-.759	18	30
12	19	29	12	1.6	40	3.85	67	.364	39	-.826	19	29
13	20	29	13	1.6	44	3.85	70	.317	42	-.8279	20	29
14	22	29	14	.46	49	3.85	72	.227	49	-.8330	22	29
15	24	26	15	.09	54	3.85	74	-.316	68	-.8376	24	26
16	25	23	16	-.46	59	3.85	75	-.355	69	-.8384	25	23
17	27	11	17	-.46	64	3.85	81	-.506	71	-.8389	27	11
18	32	4.39	18	-.46	70	3.85	82	-.684	75	-.8410	32	4.4
19	34	-.13	19	-.53	72	3.85	85	-.697	78	-.8412	34	-.16
20	36	-.352	20	-.63	76	3.85	87	-.775	81	-.8413	36	-.36
									93	-.8413		
e		e		e		e		e	94	-.8413		e
t		t		t		t		t	96	-.8414		t
c		c		c		c		c				c
			25	-.751								
			30	-.8036								
			35	-.8300								
			40	-.8401								
	60	-.8302	45	-.8404							55	-.8268
	78	-.8336	50	-.8410			91	-.7946			67	-.8339
	88	-.8341	55	-.8413			93	-.7974			89	-.8343
	97	-.8343	60	-.8413	136	-.435	95	-.7989			99	-.8346
	105	-.8377	65	-.8414	162	-.651	108	-.8265			105	-.8397
	113	-.8405			1000	-.8010	113	-.8389			116	-.8408
	127	-.8412			1005	-.8391	129	-.8412				
	157	-.8412			1016	-.8409	148	-.8413				
					1100	-.8411	166	-.8414				
					1105	-.8413	186					
					1169	-.8414						

Table 1.- MAXQUAD

Comparative tests

iter.	(6)		(7)		(8)		(10) k=0.1		(10) k=0.5		(10) k=1.	
1	1	5337	1	5337	1	5337	1	5337	1	5337	1	5337
2	2	162	2	162	2	162	2	162	2	162	2	162
3	5	58	5	58	5	58	5	87	6	53	5	58
4	7	37	7	38	7	37	6	5.8	8	12	7	30
5	10	32	9	16	9	16	14	4.5	10	7.2	8	17
6	12	32	11	11	11	13	15	3.2	14	4.8	9	11
7	13	32	14	2.7	17	2.3	16	1.0	16	3.2	10	4.7
8	14	32	16	1.7	21	1.5	18	0.29	19	0.14	13	2.4
9	15	32	22	.79	24	1.2	22	-.16	22	-.46	15	2.1
10	16	32	31	-.39	27	0.6	28	-.50	25	-.68	18	0.
11	17	32	33	-.52	30	-.03	31	-.52	36	-.773	20	-.35
12	19	7.3	48	-.66	32	-.09	34	-.56	43	-.822	21	-.48
13	22	3.1	52	-.700	35	-.26	35	-.59	47	-.823	26	-.62
14	24	2.8	53	-.750	37	-.549	36	-.63	51	-.837	27	-.78
15	26	1.4	56	-.768	40	-.552	38	-.801	53	-.838	30	-.79
16	28	1.4	58	-.8261	45	-.649	46	-.808	58	-.8397	32	-.821
17	32	.87	60	-.8261	51	-.8203	47	-.817	59	-.8405	34	-.8291
18	33	.87	61	-.8273	55	-.8204	48	-.819	77	-.8407	36	-.8314
19	34	.61	69	-.8371	64	-.8285	51	-.825	84	-.8411	48	-.8401
20	36	.61	72	-.8391	66	-.8352	52	-.826	86	-.8413	50	-.8401
2	37	.56	74	-.8407	68	-.8366	54	-.828	88	-.8413	53	-.8405
22	38	.56	95	-.8412	70	-.8408	58	-.831	91	-.8413	73	-.8405
23	39	.54	96	-.8412	72	-.8408	71	-.833	94	-.8413	78	-.8405
24	43	.54	97	-.8413	73	-.8408	73	-.834	97	-.8414	84	-.8413
25	46	-.44	105	-.8413	74	-.8409	77	-.8391			92	-.8413
26	47	-.44	107	-.8414	78	-.8411	79	-.8399			94	-.8413
27	58	-.50			82	-.8413	81	-.8402			97	-.8413
28	61	-.58			85	-.8413	85	-.8402				
29	63	-.72					87	-.8412				
30	66	-.72					92	-.8412				
							99	-.8413				
e		e					101	-.8414				
t		t										
c		c										
50	113	-.8399										
55	123	-.8408										
60	133	-.8409										
65	142	-.8409										

Table 2.- MAXQUAD Various choices of ε in DYNEPS

3.2 SHELL. An example of exact penalty is given with the second problem of Colville, where all the constraints are eliminated by an ℓ_1 penalty. See (Lemaréchal and Mifflin eds. 1979) for the precise definition of the problem.

Method 2.2 uses the parameters $a_1 = 10$, $q = 0.97$.

It turns out that, for the present problem, most methods fail to converge to the solution (this bad behavior is apparently due to nonconvexity). Therefore, it is meaningless to compare speeds of convergence, so we rather compare robustness by recording, for each method, the best objective value obtained when the method stops. In order to have more illustrative results, we have made two series of experiments, where, in the second series, function-gradient are computed in double precision.

Table 3 shows, for each experiment, the number of line-searches, the number of computations of function-gradient, and the final objective value. Except for method 2.2, the behavior is very bad and double precision does not substantially improve the situation.

It would be frustrating to limit this illustration of ℓ_1 penalty with such negative results, so Table 4 (which reads as Table 1) shows the same kind of experiments on the first problem of Colville (Shell primal).

3.3 EQUIL. This is a set of 3 examples of computation of economic equilibria. In terms of nonsmooth optimization they can be written

$$\begin{cases} \min \ \max_{k=1}^{N} \ f_k(x) \\ \sum x_j = 1 \ , \quad x_j \geq 0 \ , \end{cases} \tag{11}$$

where N = 5, 8 and 10 respectively (and, as usual, $x \in R^N$). See (Lemaréchal and Mifflin eds. 1979) for the definition of f_k. The functions f_k are defined only for $x > 0$ and tend to $+\infty$ if a coordinate of x tends to 0. Therefore we extend each f_k by

		nb. iter.	nb. obj.	final obj.val.	nb. iter.	nb. obj.	final obj.val.
B F G S		43	137	65.7	38	104	84.3
S H O R		199	199	32.6	365	365	32.4
EPSDES		980	3562	32.7	747	3001	32.5
CHAINE		244	979	39.6	193	787	38.2
DYNEPS	(6)	63	224	36.5	92	329	33.9
	(7)	111	463	34.1	69	306	34.4
	(8)	56	246	34.1	105	376	33.6
	(9)	85	337	33.8	73	271	37.2
	0.1	56	187	36.7	103	404	33.2
	0.5	86	307	34.7	42	161	34.4
	1	78	251	35.1	113	370	34.1
BFEPS		38	112	49.7	9	34	1648
		SIMPLE precision			DOUBLE precision		

Table 3.- SHELL DUAL Comparative tests

iter	BFGS		SHOR		EPSDES		CHAINE		DYNEPS		BFEPS	
1	1	20	1	20	1	20	1	20	1	20	1	20
2	2	8	2	20	3	8	2	8	2	8	2	8
3	4	4	3	20	5	-6	4	-17	4	4	4	4
4	5	-2	4	20	7	-18	8	-26	11	-26	5	-2
5	7	-3	5	20	9	-24	11	-26	15	-27.7	7	-3
6	13	-14	6	20	11	-24	13	-27.6	16	-29.0	13	-11
7	14	-26.6	7	20	13	-25	14	-27.9	17	-30.24	14	-24
8	16	-28.7	8	20	15	-29.4	15	-27.9	21	-31.03	17	-26.4
9	20	-29.4	9	20	17	-29.4	17	-28.4	28	-32.00	18	-30.28
10	23	-31.20	10	0	19	-29.4	20	-30.84	31	-32.20	21	-31.38
11	25	-31.41	11	-12	21	-29.4	23	-31.52	33	-32.22	23	-31.53
12	28	-31.92	12	-24	23	-29.4	26	-31.60	36	-32.25	26	-32.18
13	34	-32.27	13	-24	26	-31.14	27	-31.60	41	-32.26	30	-32.24
14	37	-32.28	14	-24	29	-31.14	28	-31.72	46	-32.26	32	-32.24
15	39	-32.29	15	-28.3	31	-31.14	35	-31.80	50	-32.27	34	-32.312
16	40	-32.318	16	-30.26	36	-31.14	38	-32.05	56	-32.319	36	-32.333
17	53	-32.320	17	-30.89	40	-31.14	39	-32.30	66	-32.329	40	-32.345
18	54	-32.324	18	-31.18	44	-31.14	40	-32.30	67	-32.331		
19	56	-32.338	19	-31.18	48	-31.14	42	-32.30	69	-32.332		
20	58	-32.344	20	-31.27	52	-31.14	45	-32.318	71	-32.338		
	61	-32.347							73	-32.346		
e	63	-32.347		etc.		etc.		etc.	75	-32.346		
t	64	-32.347							76	-32.346		
c	70	-32.348	25	-32.17					85	-32.347		
			30	-32.297					86	-32.348		
			35	-32.322			75	-32.341				
			40	-32.336			79	-32.341				
			45	-32.344			100	-32.347				
			50	-32.344			105	-32.347				
			55	-32.348	180	-32.27	110	-32.348				
					190	-32.307						
					200	-32.320						
					210	-32.324						
					220	-32.343						
					230	-32.343						
					240	-32.347						
					248	-32.348						

Table 4.- SHELL PRIMAL

Comparative tests

$+\infty$ outside the feasible domain. Furthermore, we do not really consider f_k but rather its restriction to the manifold $\sum x_j = 1$ (the gradient of such a restricted function is the projection of the original gradient onto the subspace $\sum g_j = 0$--see (Lemaréchal and Mifflin eds. 1979)). As a result the minimax problem is unconstrained and methods 2.1-2.6 are applied as they stand.

For method 2.2, the paramenters are $a_1 = \frac{1}{N}$ and $q = 0.95$.

In addition to 2.5, we have tested a variant, which we call ECONEW, similar to DYNEPS except that the set $G(\varepsilon)$ is larger: to the vectors $g_1, \ldots, g_n$ that are subgradients of f at $y_1, \ldots, y_n$, one appends the vectors

$$g_{n+k} = \nabla f_k(x_n) \quad , \quad \text{with the coefficients}$$

$$\alpha_{n+k} = f(x_n) - f_k(x_n) \quad .$$

This makes d_n similar to the direction of the Newton method, when applied to the minimax problem (11). It so happens that (11) is a Haar problem so that the Newton method has a super-linear rate of convergence (and so should have ECONEW).

Table 5 shows the result, for the 3 problems. For each method it displays the number of line-searches, of computations of function-gradient, and the final value of the max-function f. Figure 1 illustrates the relative behavior of DYNEPS and ECONEW on the third problem (N = 10), with ε in Step 4 given by (6). It exhibits the better asymptotic behavior of ECONEW.

3.4 TR48. This is the dual of a transportation problem in R^{48} and the function to be minimized is

$$f(x) = \sum_{i=1}^{48} s_i x_i + d_i \max_{j=1}^{48} (a_{ij} - x_j) \quad .$$

It is a piecewise linear function, with a very large number of linear pieces. The optimum value is -638565 but, working in single precision, one should be satisfied with objective values around -638500.

		5 X 3			8 X 5			10 X 5		
		nb. iter.	nb. obj.	final obj.val.	nb. iter.	nb. obj.	final obj.val.	nb. iter.	nb. obj.	final obj.val.
BFEPS		14	83	.7	10	52	.6	10	41	3.
DYNEPS & ECONEW	CHAINE	28	72	.000	44	95	.000	89	192	.000
	EPSDES	47	238	.000	82	395	.000	143	518	.000
	SHOR	34	34	.007	40	40	.006	51	51	.06
	BFGS	16	63	.66	16	68	.4	33	103	.9
	(6)	11	42	.019	17	43	.008	25	61	.09
		11	37	.001	14	44	.001	12	28	.011
	(7)	14	40	.008	16	40	.007	26	69	.08
		12	31	.011	16	39	.006	22	52	.004
	(8)	12	39	.008	17	43	.003	28	60	.08
		13	34	.011	17	43	.004	19	54	.001
	(9)	13	39	.009	18	52	.007	26	56	.04
		13	37	.001	17	43	.006	23	42	.04
	0.1	16	58	.000	19	54	.003	30	74	.082
		16	41	.000	21	55	.015	25	68	.082
	0.5	13	52	.007	16	41	.008	27	66	.09
		11	36	.02	18	59	.001	18	47	.05
	1.	14	42	.002	17	42	.008	26	64	.09
		14	37	.03	19	46	.000	23	59	.000

Table 5.- EQUIL Comparative tests

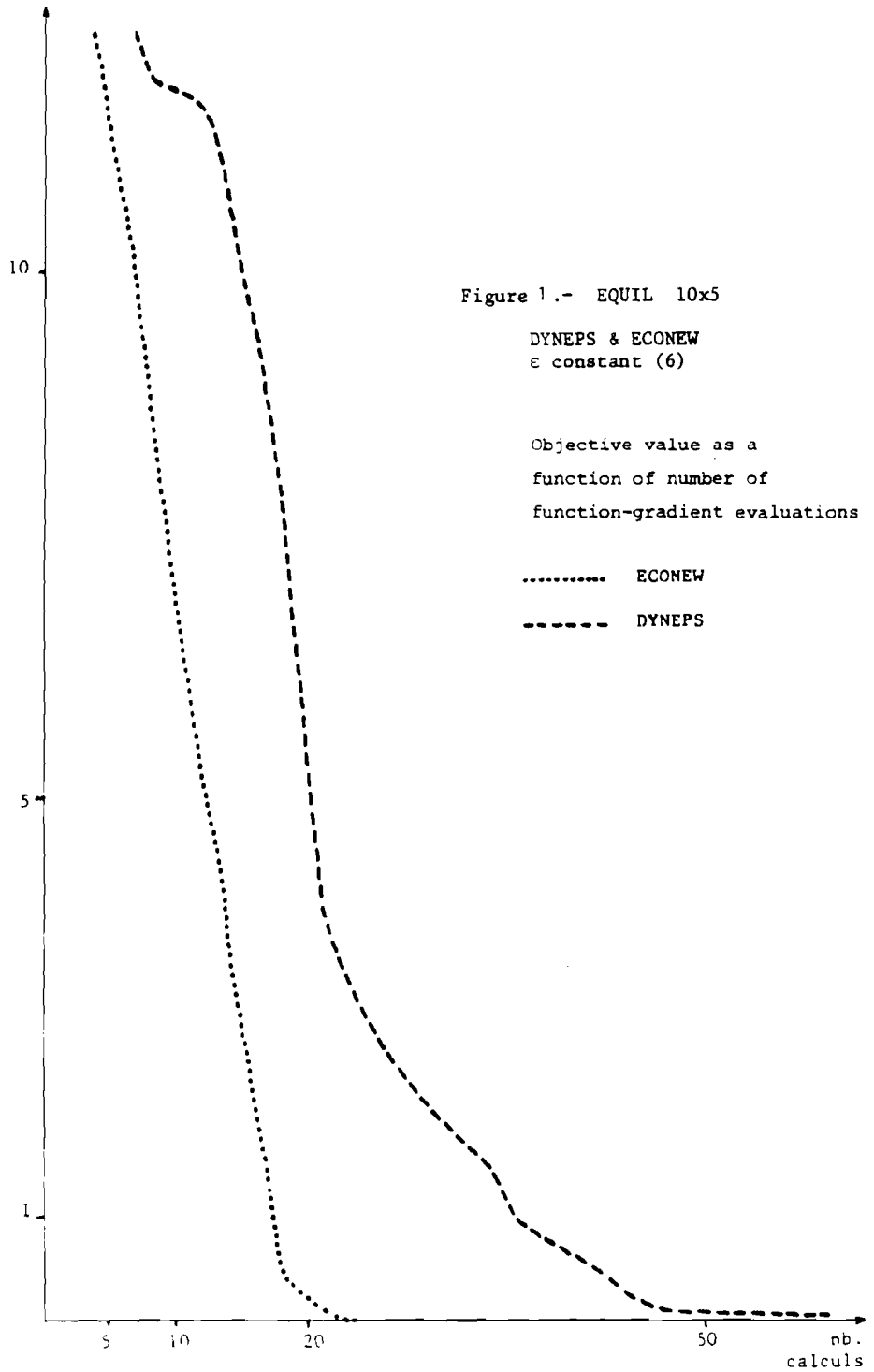

Figure 1.- EQUIL 10x5

DYNEPS & ECONEW
ε constant (6)

Objective value as a function of number of function-gradient evaluations

..... ECONEW

- - - - - DYNEPS

For method 2.2, the parameters are $a_1 = 1000$, $q = 0.985$. Table 6 (which reads as Table 3) shows the results, and demonstrates the excellent behavior of method 2.2. Method 2.3 was so slow that we could not consider that it converged in a reasonable amount of CPU time.

3.5 HILBERT. Finally, for curiosity, we have tested these algorithms on an ill-conditioned quadratic function:

$$f(x) = \frac{1}{2}(x, Ax) - (b, x) \quad , \quad x \in R^{50} \quad ,$$

where A is the Hilbert matrix $a_{ij} = \frac{1}{i+j-1}$, and b is such that the solution is $(1, \ldots, 1)$.

From properties of Hilbert matrices (reasonable norm but small coercivity constant) it is easy to identify the optimal cost, but impossible to obtain the optimal solution up to a reasonable precision. Therefore we measure the quality of a solution x, not by its cost f(x), but by its deviation from the optimum: $\sum_1^{50} (x_i - 1)^2$.

Since the objective function is differentiable, we have tested--in addition to algorithms of §2--two versions of the conjugate gradient method, namely the subroutines VA08A and VA14A of the Harwell library.

Table 7 shows, for each method, the number of line searches, of computations of function-gradient, and the final deviation from the optimal solution. A note (*) indicates that the method has stopped through roundoff errors in the line-search; otherwise the stopping criterion has normally worked.

It has been practically impossible to implement method 2.2 because, in this example, we could not really know which criterion to use for adjusting parameters a_1 and q.

The results are generally modest (the contrary would have been a big surprise) but it is astonishing to see that the most precise algorithm for estimating the optimal solution is 2.3, whereas this algorithm is devised to estimate the optimal cost.

		nb. iter.	nb. obj.	final obj.val.
	B F G S	289	414	-638 346
	S H O R	285	285	-638 563
	EPSDES	∞	∞	
	CHAINE	415	629	-638 057
	(6)	244	368	-638 203
	(7)	332	627	-638 328
D Y N E P S	(8)	325	643	-638 536
	(9)	400	748	-638 308
	0.1	209	398	-638 308
	0.5	207	422	-638 521
	1.	139	227	-638 475
	BFEPS	167	257	-638 330

Table 6.- TR48 Comparative tests

		nb. iter.	nb. obj.	final obj.val.
	B F G S	8	16	.098 (1)
	VA08A	5	15	.100 (1)
	VA14A	8	20	.099 (1)
	S H O R	100	100	.139
	EPSDES	70	276	.004
	CHAINE	16	35	.013
DYNEPS	(6)	15	38	.010 (1)
	(7)	15	35	.054 (1)
	(8)	15	31	.018
	(9)	13	32	.078 (1)
	0.1	20	38	.013
	0.5	11	29	.091 (1)
	1.	14	31	.034 (1)
	BFEPS	11	21	.018

Table 7.- HILBERT Comparative tests

4. CONCLUSION

The aim of this paper was first to demonstrate the validity of some recently proposed algorithms for nonsmooth optimization. We think that the variety of experiments (although they are purely academic) shows that these algorithms do behave consistently, even if their convergence is not always very fast. Failures have been recorded only in one instance (Shell dual) but some improvements of ε-subgradient methods are under study to better cope with nonconvexity, and more satisfactory results have already been obtained.

We have also exhibited the fact that it can be good practice to use a quasi-Newton method in nonsmooth optimization The convergence is rather rapid, and often a reasonably good approximation of the optimum is found; this, in our opinion, is essentially due to the fact that inaccurate line-searches are made. Of course, there is no theoretical possibility to prove convergence to the right point (in fact counterexamples exist) neither are there any means to assess the results.

In terms of rapidity of convergence, Shor's dilatation of the space along the difference of two successive gradients is an excellent method. However, it must be recalled that the question of stepsize is not yet solved, and this prevents the method from being really implementable.

Finally it is rather amusing to compare the results on problems 3.4 and 3.5. Since 3.4 is piecewise linear, one should expect better results with methods 2.4 to 2.5 (which are based on piecewise linear approximations). On the other hand, problem 3.5 being quadratic, it is with methods 2.1 and 2.2 (which are based on quadratic approximations) that better results should be expected. Table 6 and 7 show that it is the contrary that happens, and this raises the question: is there a well-defined frontier between quadratic and piecewise linear functions, or more generally, between smooth and nonsmooth functions?

APPENDIX: Experiments with the ellipsoid algorithm.

Because an exceptional attention has been recently given to Shor's method of dilation of the space along the gradient (the so-called ellipsoid algorithm, popularized by Khachyian) we have also programmed this method, as described in:

N.Z. Shor : Cut-off method with space extension in convex programming problems, *Cybernetics* 1 (1977) 94-96.

In this method, a parameter R must be given, which estimates the distance from the initial iterate x_1 to the optimal solution x^*. Because x^* is known in each of the present examples, we just set $R = |x_1 - x^*|$.

Table 8 shows the results by this method, with the 7 test-problems described in Section 3.

Table 8. Experiments with the ellipsoid algorithm.

	Number of calculations of function-gradient	Final value of objective function
MAXQUAD	1383	-.8414 (1)
SHELL DUAL	4399	32.35 (2)
EQUIL 5 X 3	160	0.03
EQUIL 8 X 5	398	0.03
EQUIL 10 X 5	714	0.09
TR 48	331	-631000. (3)
HILBERT	241	0.003 (4)

(1) Value after 500 calculations : -.82

(2) Value after 1000 calulations : 83

(3) Starts diverging after that iteration

(4) Final value of $|x_n - x^*|^2$.

REFERENCES

Lemaréchal, C. (1974) An Algorithm for Minimizing Convex Functions in Rosenfeld ed. Proceedings of the IFIP Congress, Stockholm, 553-556. North Holland Publishing Company.

Lemaréchal, C. (1975) An extension of "Davidon" methods to nondifferentiable problems in Balinski and Wolfe eds. Nondifferentiable Optimization, Mathematical Programming Study 3:95-109. North Holland Publishing Company.

Lemaréchal, C. (1976) Combining Kelley's and Conjugate Gradient Methods in Prekopa, ed. Abstracts of the IX Symposium on Mathematical Programming, Budapest.

Lemaréchal, C. (1979) Bundle methods in nondifferentiable optimization in Lemaréchal and Mifflin eds. Nonsmooth Optimization 79-102. Pergamon Press.

Lemaréchal, C. (1978) Nonsmooth Optimization and Descent Methods. RR-78-4. Laxenburg, Austria: International Institute for Applied Systems Analysis.

Powell, M.J.D. (1975) Some Global Convergence Properties of a Variable Metric Algorithm for Minimization Without Exact Line-Searches. AERE Report Css15. Harwell.

Shor, N.Z. (1971) A minimization method using the operation of dilatation of the space in the direction of the difference of two successive gradients. Cybernetics 3:450-459.

Shor, N.Z. and L.P. Shabashova (1972) Solution of minimax problems by the method of generalized gradient descent with dilation of the space. Cybernetics 1:88-94.

Wolfe, P. (1975) A method of conjugate subgradients for minimizing nondifferentiable functions in Balinski and Wolfe eds. Nondifferentiable Optimization, Mathematical Programming Study 3:145-173. North Holland Publishing Company.

A set of nonsmooth optimization test-problems in Lemaréchal and Mifflin eds. Nonsmooth Optimization 151-165. Pergamon Press. (1979)

CONVERGENCE OF A MODIFICATION OF LEMARÉCHAL'S ALGORITHM FOR NONSMOOTH OPTIMIZATION*

Robert Mifflin
Department of Pure and Applied Mathematics
Washington State University
Pullman, Washington
USA

1. INTRODUCTION

We consider the problem of minimizing a locally Lipschitz continuous function f on R^n. We give a modification of an algorithm due to Lemarechal [2] and show convergence to a stationary point of f if f also satisfies a weak "semismoothness" [3,4] hypothesis that is most likely satisfied by continuous functions arising in practical problems. The method combines a generalized cutting plane idea with quadratic approximation of a Lagrangian. Even for the case of a convex f, as considered in [2], this version differs from the original method, because of its rules for line search termination and the associated updating of the search direction finding subproblem. More specifically, our version does not require a user-specified uniform lower bound on the line search stepsizes.

A point $\bar{x}\in R^n$ is stationary if $0\in\partial f(\bar{x})$ where ∂f is the generalized gradient [1] of f, i.e., $\partial f(x)$ is the convex hull of all limits of sequences of the form $\{\nabla f(x_k) : \{x_k\} \to x$ and f is

*This material is based upon work supported by the National Science Foundation under Grant No. MCS 78-06716.

differentiable at each x_k}. Important properties of the mapping $\partial f(\cdot)$ are uppersemicontinuity and local boundedness. If f is convex (concave), ∂f equals the subdifferential (superdifferential) of f or if f is continuously differentiable (C^1) ∂f equals the (ordinary) gradient $\{\nabla f\}$. It is possible to determine ∂f or at least to give one element of $\partial f(x)$ at each x for many other functions f such as those that are pieced together from C^1 functions (for example, via maximization and/or minimization operations occurring in decomposition, relaxation, duality and/or exact penalty approaches to solving optimization problems).

In order to implement the algorithm we suppose that we have a subroutine that can evaluate a function $g(x)\in\partial f(x)$ for each $x\in R^n$. Of course, we are especially interested in the case where g is discontinuous at stationary points of f. Associated with f (and g) let $\alpha : R^n\times R^n \to R_+$ be a nonnegative-valued function satisfying

$$\alpha(x,y) \to 0 \text{ if } x \to \bar{x} \text{ and } y \to \bar{x}, \tag{1a}$$

$$\alpha(z,y) - \alpha(x,y) \to 0 \text{ if } x \to \bar{x}, \quad z \to \bar{x} \text{ and } y \to \bar{y}, \tag{1b}$$

and

$$\bar{g}\in\partial f(\bar{x}) \text{ if } x \to \bar{x}, \quad y \to \bar{y}, \quad g(y) \to \bar{g} \text{ and } \alpha(x,y) \to 0. \tag{1c}$$

$\alpha(x,y)$ is intended to be an indication of how much $g(y)\in\partial f(y)$ deviates from being a generalized gradient at x. If f is convex we take

$$\alpha(x,y) = f(x) - [f(y) + \langle g(y), x-y\rangle],$$

which measures the deviation from linearity of f between y and x. For a nonconvex f a possibility is to take $\alpha(x,y)$ to be some "distance" between x and y, in which case property (1c) follows from the uppersemicontinuity of ∂f.

2. THE ALGORITHM

Given a positive integer k, n-vectors x_k, $g(x_k)$, y_i and $g(y_i)$ for i=1,2,...,k and a positive definite n×n matrix A_k solve for $(d,v) = (d_k,v_k) \in R^{n+1}$ the k^{th} quadratic programming subproblem:

$$\text{minimize } 1/2\langle d, A_k d\rangle + v$$

$$\text{subject to } \langle g(x_k), d\rangle \leq v$$

$$-\alpha(x_k, y_i) + \langle g(y_i), d\rangle \leq v \text{ for } i = 1,2,\ldots,k.$$

If $v_k = 0$ stop.

Otherwise ($v_k < 0$ and $d_k \neq 0$) perform a line search from x_k along d_k to find (if possible) two (possibly the same) stepsizes $t_L \geq 0$ and $t_R > 0$ and two corresponding points, $x_L = x_k + t_L d_k$ and $y_R = x_k + t_R d_k$, such that

$$f(x_L) \leq f(x_k) + m_L t_L v_k \tag{2a}$$

and

$$-\alpha(x_L, y_R) + \langle g(y_R), d_k\rangle \geq m_R v_k, \tag{2b}$$

where m_L and m_R are fixed parameters satisfying $0 < m_L < m_R < 1$. If the line search is successful repeat the above procedure with the $k+1^{st}$ subproblem defined by setting

$$x_{k+1} = x_L \text{ and } y_{k+1} = y_R,$$

replacing in the subproblem constraints

$$g(x_k) \text{ by } g(x_L) \text{ and } \alpha(x_k, y_i) \text{ by } \alpha(x_L, y_i)$$

$$\text{for } i = 1,2,\ldots,k,$$

appending the constraint

$$-\alpha(x_{k+1}, y_{k+1}) + \langle g(y_{k+1}), d\rangle \leq v,$$

and replacing A_k by a positive definite matrix A_{k+1}.

3. REMARKS ON THE ALGORITHM

For the starting point x_1 we may set $y_1=x_1$. If this is the case the $\alpha(x_1,y_1) = 0$ and $g(x_1)=g(y_1)$ so that the two starting constraints are the same. In general it may be the case that $x_k=y_i$ for some $i\in\{1,2,\ldots,k\}$. In this case the first constraint may be disregarded during subproblem solution, because it is included among the other constraints.

Elsewhere we will report on how to extend the numerically stable constrained least squares algorithm in [5] for solving more general quadratic programming problems. This will result in a reliable method for solving the dual of the subproblem given here.

The scalar v_k can be interpreted as an approximation to the directional derivative of f at x_k in the direction d_k. By convex quadratic programming duality theory, as in [2],

$$v_k = -\langle d_k, A_k d_k\rangle - \sum_{i=1}^{k} \lambda_{ik}\alpha(x_k,y_i) \tag{3a}$$

and

$$A_k d_k = -\lambda_{0k} g(x_k) - \sum_{i=1}^{k} \lambda_{ik} g(y_i) \tag{3b}$$

where $\lambda_{ik} \geq 0$ for $i=0,1,\ldots,k$ are dual variables (multipliers) associated with the k^{th} subproblem such that

$$\sum_{i=0}^{k} \lambda_{ij} = 1. \tag{3c}$$

By the first subproblem constraint, the Cauchy-Schwarz inequality, (3a), the nonnegativity of λ_{ik} and α and the positive definiteness of A_k we have

$$-|g(x_k)||d_k| \leq \langle g(x_k), d_k\rangle \leq v_k \leq -\langle d_k, A_k d_k\rangle \leq 0. \tag{4}$$

Therefore if $v_k \neq 0$ then $v_k < 0$ and $d_k \neq 0$ and, hence, the line search may be initiated.

Let $\gamma_k > 0$ be the smallest eigenvalue of A_k. Then from (4)

$$-|g(x_k)|^2/\gamma_k \leq v_k \leq -\gamma_k|d_k|^2 \leq 0 \tag{5}$$

and we have the following:

Lemma 1. If $v_k = 0$ then x_k is stationary.

Proof. If $v_k = 0$ then, from (5), $d_k = 0$ and, by (3),

$$\sum_{i=1}^{k} \lambda_{ik}\alpha(x_k,y_i) = 0 \tag{6a}$$

and

$$\lambda_{0k}g(x_k) + \sum_{i=1}^{k} \lambda_{ik}g(y_i) = 0. \tag{6b}$$

The nonnegativity of λ_{ik} and α and (6a) imply that $\alpha(x_k,y_i) = 0$ for each i such that $\lambda_{ik} > 0$. Thus, by property (1c), $g(y_i)$ is an element of the convex set $\partial f(x_k)$ for each i such that $\lambda_{ik} > 0$. Now, stationarity of x_k follows from (6b) and (3c).

Relative to v_k and d_k the line search termination criterion (2a) guarantees sufficient function value decrease, while (2b) along with the definition of α provides sufficient (approximate) directional derivative increase. More specifically, because $m_R < 1$, (2b) causes (d_k,v_k) to be infeasible in subproblem k+1. If f is weakly uppersemismooth (see Appendix) then, because of (1a) and the parameter inequality $m_L < m_R$, a simple search procedure, such as in [3], can be designed to find t_L and t_R or to generate an increasing sequence $\{t_\ell\}$ such that $\{f(x_k+t_\ell d_k)\} \to -\infty$. In order to deal effectively with regions on which f is smooth it is recommended that $m_L < 1/2$.

Note that due to nondifferentiability of f it is possible that $t_L=0$, so $x_{k+1}=x_k$ and the first constraint is not changed.

But in this case, since $t_R > 0$, $y_{k+1} \neq x_{k+1}$, so the appended constraint is clearly different from the first constraint and its inclusion improves our approximate knowledge of $\partial f(x_k)$. In fact, this difference holds in general when $t_L \neq t_R$ and this is what detects and deals with discontinuities of g.

We do not discuss the important question of updating A_k here. We conjecture that A_k should be chosen to converge to the Hessian of some Lagrangian associated with the limiting optimal multipliers of the subproblems. This is the subject of ongoing research where we are developing tests to identify iteration indices k where we may simultaneously make a variable metric update of A_k and reduce or aggregate the constraint bundle.

4. CONVERGENCE

In this section we establish three lemmas that prove part (a) of the following convergence theorem. Part (b) follows from part (a), because a stationary point for a semiconvex function (see Appendix) is a minimizing point [4] and because every accumulation point of $\{x_k\}$ has the same f-value due to the monotonicity of $\{f(x_k)\}$.

Theorem. Suppose $\{x_k\}$, $\{y_k\}$ and $\{A_k\}$ are uniformly bounded with $\{A_k\}$ uniformly positive definite. Then

(a) at least one of the accumulation points of $\{x_k\}$ is stationary and

(b) if f is semiconvex on R^n, every accumulation point of $\{x_k\}$ minimizes f.

Remark. If $\{x : f(x) \leq f(x_1)\}$ is bounded then $\{x_k\}$ is bounded and $\{y_k\}$ can be made bounded by choosing an additional parameter $\beta > 0$ and imposing the additional line search requirement that

$$|y_R - x_L| = (t_R - t_L)|d_k| \leq \beta. \tag{7}$$

For weakly uppersemismooth functions it is possible to simultaneously satisfy (2) and (7) after a finite number of line search steps.

Consider the following assumption that is trivially satisfied if the matrix sequence $\{A_k\}$ is uniformly positive definite:

$$\text{If } \{d_k\} \text{ has no zero accumulation point then } \{\gamma_k\} \text{ has no zero accumulation point.} \tag{8}$$

Lemma 2. Suppose (8) holds and $\{x_k\}$ and $\{y_k\}$ are uniformly bounded. Then $\{d_k\}$ has at least one zero accumulation point.

Proof. Suppose for purposes of a proof by contradiction that there exists a positive number δ such that

$$|d_k| \geq \delta > 0 \text{ for all } k.$$

Then, by (8), there exists a positive number γ such that

$$\gamma_k \geq \gamma > 0 \text{ for all } k$$

and, by (5)

$$-|g(x_k)|^2/\gamma \leq v_k \leq -\gamma\delta|d_k| \leq -\gamma\delta^2 < 0. \tag{9}$$

Thus, since $\{x_k\}$ is assumed bounded and $\partial f(\cdot)$ is locally bounded, $\{g(x_k)\}$ and, hence, $\{v_k\}$ and $\{d_k\}$ are bounded. Let $\bar{v}$ and $\bar{d}$ be accumulation points of $\{v_k\}$ and $\{d_k\}$, respectively. Then, by (9)

$$\bar{v} \leq -\gamma\delta^2 < 0. \tag{10}$$

By (2a) and (9),

$$f(x_L) - f(x_k) \leq m_L t_L v_k \leq -m_L t_L \gamma\delta|d_k|$$

or, since $x_{k+1}=x_L=x_k+t_L d_k$,

$$f(x_{k+1}) - f(x_k) \leq -m_L\gamma\delta|x_{k+1}-x_k|. \tag{11}$$

For any $p > k+1$, (11) and the triangle inequality imply

$$f(x_p) - f(x_{k+1}) = \sum_{j=k+1}^{p-1} f(x_{j+1}) - f(x_j) \leq -m_L \gamma \delta$$

$$\sum_{j=k+1}^{p-1} |x_{j+1} - x_j| \geq -m_L \gamma \delta |x_p - x_{k+1}|$$

As f is continuous and $\{x_k\}$ is assumed bounded, the monotone nonincreasing sequence $\{f(x_k)\}$ is bounded from below and, hence,

$$\{|x_p - x_{k+1}|\} \to 0 \quad .$$

These facts together with property (1b) and the assumed boundedness of $\{y_k\}$ imply that

$$\{\alpha(x_{k+1}, y_{k+1}) - \alpha(x_p, y_{k+1})\} \to 0 \quad . \tag{12}$$

Also, for any $p \geq k+1$ we have, by the p^{th} subproblem feasibility, that

$$-\alpha(x_p, y_{k+1}) + \langle g(y_{k+1}), d_p \rangle \leq v_p$$

and, by (2b) with $x_L = x_{k+1}$ and $y_R = y_{k+1}$, that

$$-\alpha(x_{k+1}, y_{k+1}) + \langle g(y_{k+1}), d_k \rangle \geq m_R v_k \quad .$$

Subtracting the latter from the former inequality gives

$$\alpha(x_{k+1}, y_{k+1}) - \alpha(x_p, y_{k+1}) + \langle g(y_{k+1}), d_p - d_k \rangle \leq v_p - m_R v_k \quad . \tag{13}$$

Now choose p and k in K, as infinite set of integers where $\{d_k\}_{k \in K} \to \bar{d}$ and $\{v_k\}_{k \in K} \to \bar{v}$, so that from (12), (13) and the boundedness of $\{g(y_{k+1})\}$ we have

$$0 \leq \bar{v} - m_R \bar{v} = (1-m_R)\bar{v} \quad .$$

Since $m_R < 1$, this implies that $\bar{v} \geq 0$, which contradicts (10) and completes the proof.

Lemma 3. Suppose that K is such that $\{d_k\}_{k\in K} \to 0$, $\{A_k d_k\}_{k\in K} \to 0$ and $\{\langle g(x_k), d_k\rangle\}_{k\in K} \to 0$. Then $\{v_k\}_{k\in K} \to 0$,

$$\{\sum_{i=1}^{k} \lambda_{ik} a(x_k, y_i)\}_{k\in K} \to 0,$$

and

$$\{\lambda_{0k} g(x_k) + \sum_{i=1}^{k} \lambda_{ik} g(y_i)\}_{k\in K} \to 0$$

where the $\lambda_{ik} \geq 0$ satisfy (3).

Proof. The conclusions follow from the hypotheses (4), (3a) and (3b).

Lemma 4. In addition to the hypotheses of Lemma 3, suppose that $\{x_k\}_{k\in K}$ and $\{y_k\}_{k\in K}$ are uniformly bounded and let $\bar{x}$ be any accumulation $\{x_k\}_{k\in K}$.
Then $\bar{x}$ is stationary.

Proof. As in the proof of Thm. 5.2 in [3], depending on the local boundedness and uppersemicontinuity of ∂f and the properties of convex combinations, Lemma 3 implies the existence of a positive integer $m \leq n+1$, an infinite subset $J \subset K$ and convergent subsequences

$$\{(x_k, g(x_k))\}_{k\in J} \to (\bar{x}, g^0) \in R^n \times \partial f(\bar{x}), \quad \{(y_k^i, g(y_k^i))\}_{k\in J}$$

$$\to (y^i, g^i) \in R^n \times \partial f(y^i)$$

for $i=1,2,\ldots,m$ and $\{\mu_k^i\}_{k\in J} \to \mu^i \geq 0$ for $i=0,1,\ldots,m$

such that

$$\sum_{i=0}^{m} \mu^i = 1 \quad ,$$

$$\sum_{i=0}^{m} \mu^i g^i = 0 \quad ,$$

and for $i=1,2,\ldots,m$

$$\{\alpha(x_k, y_k^i)\}_{k\in J} \to 0 \text{ if } \mu^i > 0 \quad .$$

Now, stationarity of $\bar{x}$ follows from property (1c) as in the proof of Lemma 1.

5. EXTENSION TO CONSTRAINED PROBLEMS

Finally, we remark that using ideas in [3], the algorithm can be extended to become a feasible point method for dealing with constrained optimization problems involving semismooth functions.

6. APPENDIX

A function $f : R^n \to R$ is <u>weakly uppersemismooth</u> [3] at $x\in R^n$ if

(a) f is Lipschitz continuous on a ball about x.

(b) for each $d\in R^n$ and for any sequences $\{t_k\}\subset R_+$ and $\{g_k\}\subset R^n$ such that $\{t_k\} \downarrow 0$ and $g_k \in \partial f(x+t_k d)$ it follows that $\liminf_{k\to\infty} \langle g_k, d\rangle \geq \limsup_{t\downarrow 0} [f(x+td) - f(x)]/t$

It can be shown that the right-hand side of the above inequality is in fact equal to

$$f'(x;d) = \lim_{t\downarrow 0} [f(x+td) - f(x)]/t \quad ,$$

the <u>directional derivative</u> of f at x in the direction d.

The class of weakly uppersemismooth functions strictly contains the class of semismooth [4] functions. This latter class is closed under composition and contains convex, concave, C^1 and many other locally Lipschitz functions such as ones that result from piecing together C^1 functions.

A function $f : R^n \to R$ is <u>semiconvex</u> [4] at $x \in R^n$ if

(a) f is Lipschitz continuous on a ball about x;
and for each $d \in R^n$, $f'(x;d)$ exists and satisfies

(b) $f'(x;d) = \max\{\langle g,d\rangle : g \in \partial f(x)\}$

(c) $f'(x;d) \geq 0$ implies $f(x+d) \geq f(x)$.

An example of a nondifferentiable nonconvex function that is both semismooth and semiconvex is $\log(1+|x|)$ for $x \in R$.

REFERENCES

1. F.H. Clarke. 1975. Generalized Gradients and Applications. Trans. Amer. Math. Soc. 205, pp. 247-262.

2. C. Lemarechal. 1978. Nonsmooth Optimization and Descent Methods. RR-78-4. International Institute for Applied Systems Analysis, Laxenburg, Austria.

3. R. Mifflin. 1977. An Algorithm for Constrained Optimization with Semismooth Functions. Mathematics of Operations Research 2, pp. 191-207.

4. R. Mifflin. 1977. Semismooth and Semiconvex Functions in Constrained Optimization. SIAM J. Control and Optimization 15, pp. 959-972.

5. R. Mifflin. 1979. A Stable Method for Solving Certain Constrained Least Squares Problems. Mathematical Programming 16, pp. 141-158.

SUBGRADIENT METHOD FOR MINIMIZING WEAKLY CONVEX FUNCTIONS AND ε-SUBGRADIENT METHODS OF CONVEX OPTIMIZATION

E.A. Nurminski
International Institute for Applied Systems Analysis
Laxenburg, Austria

1. INTRODUCTION

Optimization methods are very important in systems analysis. Not all systems analysis problems are optimization problems, of course, but in any systems problem optimization methods are important and useful tools. The power of these methods and their ability to handle different problems makes it possible to analyze and construct very complicated systems. Economic planning, for instance, would be greatly limited without the use of linear programming (LP) techniques.

However, linear programming is not the only method of optimization. Problems including factors such as uncertainty, only partial knowledge of the system, and conflicting goals require more sophisticated methods for their solution--methods such as nondifferentiable optimization.

This paper considers the common situation which arises when the outcomes of particular decisions cannot be estimated without solving a difficult auxiliary problem. The solution of this auxiliary problem can be very time-consuming and may limit the analysis of different decisions in the original problem. This paper develops methods of optimal decision making which avoid the direct comparison of decisions and which use only information which is readily accessible from a computational point of view.

2. THE PROBLEM

This paper deals with the finite-dimensional unconditional extremum problem

$$\min_{x \in E^n} f(x) \tag{1}$$

where the objective function has no continuous derivatives with respect to the variable $x = (x_1, \ldots, x_n)$. Various methods were discussed and suggested in relevant literature to solve problem (1) with many types of non-differentiable objective functions. The bibliography published in [1] gives a fairly good notion of these works. It should be emphasized that the nondifferentiability of the objective function in problem (1) is, as a rule, due to complexity of the function's structure. A representative example is minimax problems where the objective function $f(x)$ is a result of maximization of some function $g(x,y)$ with respect to variables y:

$$f(x) = \max_{y \in Y} g(x,y) \tag{2}$$

In this case even a simple computation of the value of f at some fixed point may be quite a time-consuming task which requires, strictly speaking, an infinite number of operations. With this in mind, it seems to be interesting from the standpoint of theory and practice to investigate the feasibility of solution of problem (1) with an approximate computation of the function $f(x)$ and of its subgradients (if the latter are determined for a given type of nondifferentiability). To the best of our understanding, ε - subgradients of functions of the form (2), introduced by R.T. Rockafellar [2], are quite convenient for constructing numerical methods, and so we offer here some results generalizing efforts in this direction [3-5].

3. WEAKLY CONVEX FUNCTIONS

The discussion of a class of the non-differentiable functions broader than the convex functions enables us to gain substantially in generality at the expense of a minor increase in complexity. Properties of the class which will be treated are described by the following definition [6]:

Definition The continuous function $f(x)$ is called the weakly convex function if for each x there exists at least one vector g such that

$$f(y) \geq f(x) + (g, y - x) + r(x,y) \qquad (3)$$

for all y, and the residual term $r(x,y)$ satisfies the condition of uniform smallness with respect to $\|x - y\|$ in each compact subset of E^n, i.e., in any compact set $K \subset E^n$ for any $\varepsilon > 0$ there exists $\delta_K > 0$ such that for $\|x - y\| \leq \delta_K$, $x, y \in K$

$$|r(x,y)| \|x-y\|^{-1} \leq \varepsilon$$

Notice that no constraints are imposed on the sign of the residual term $r(x,y)$. Furthermore, strengthening (3) it is possible to add to $r(x,y)$ any expression of the form $\phi(\|x - y\|)$, where

$$\phi(t) \leq 0 \quad , \quad \phi(t)t^{-1} \to 0 \qquad \text{for } t \to +0$$

The term weakly convex functions is suggested by analogy to the strongly convex functions studied by B.T. Polyak [7].

We will call the vector g, satisfying (3), the subgradient of the function $f(x)$ and will denote a set of subgradients at the point x by $\partial f(x)$.

We will now describe some simple properties of weakly convex functions and of their subgradients.

Lemma 1. With respect to x, $\partial f(x)$ is a convex, closed, bounded and upper semicontinuous multivalued mapping.

The proof of these properties presents no special problems.

Lemma 2. Let $f(x,\alpha)$ be continuous with respect to α and weakly convex with respect to x for each α belonging to the compact topological space A. That is,

$$f(y,\alpha) - f(x,\alpha) \geq (g_\alpha, y - x) + r_\alpha(x,y) \tag{4}$$

for all y, and here $r_\alpha(x,y)$ satisfies the condition of uniform smallness uniformally with respect to $\alpha \in A$. Then

$$f(x) = \max_{\alpha \in A} f(x,\alpha) \tag{5}$$

is a weakly convex function.

The proof is rather simple.

Let

$$A(x) = \{\alpha : f(x,\alpha) = f(x)\}$$

Then, considering (4) for $\alpha \in A(x)$, we obtain

$$f(y) - f(x) \geq f(y,\alpha) - f(x,\alpha) \geq$$

$$\geq (g_\alpha, y - x) + r_\alpha(x,y) \geq$$

$$\geq (g_\alpha, y - x) + \bar{r}(x,y)$$

where

$$-\bar{r}(x,y) = \sup_{\alpha \in A} |r_\alpha(x,y)|$$

It is easily seen that $\bar{r}(x,y)$ satisfies necessary conditions of uniform smallness and the lemma is proved.

The proof of Lemma 2 helps in understanding the procedure of calculation of subgradients of the weakly convex functions. Specifically, for functions of the form (5) the vector $g_\alpha \in G_\alpha(x), \alpha \in A(x)$ is the subgradient of the function $f(x)$ at the point x. It follows from Lemma 1 that an arbitrary vector

$$g \in \mathrm{co}\{g_\alpha, \alpha \in A(x)\} = G(x)$$

is also the subgradient.

The finding of even one element of the set $G(x)$ may be a non-trivial problem and, ignoring efforts spent to calculate for the fixed α the subgradient $g_\alpha \in \partial f_\alpha(x)$, it can be said that problems of computing $f(x)$ and of its subgradient $g \in \partial f(x)$ are equal in complexity.

In establishing necessary extremum conditions for weakly convex functions of great importance is the existence of directional derivatives and a formula for their computation in terms of subgradients.

Lemma 3. The weakly convex function $f(x)$ is directionally differentiable and

$$\frac{\partial f(x)}{\partial e} = \lim_{h \to +0} \frac{f(x+he) - f(x)}{h} = \max_{g \in \partial f(x)} (g,e)$$

Proof. Let

$$\phi(h) = f(x + he) - f(x)$$

It is easily seen that $\phi(h)$ as a function of h is weakly convex. Denote the set of subgradients of $\phi(h)$ by $\partial\phi(h)$. Assume the contrary of what the lemma asserts:

$$\underline{a} = \lim_{h \to +0} \frac{\phi(h)}{h} < \overline{\lim_{h \to +0}} \frac{\phi(h)}{h} = \bar{a}$$

and let $\{\tau_k\} \to \tau$ and $\{\sigma_k\} \to \sigma$ be sequences of values of h such that

$$\lim_{k\to\infty} \frac{\phi(\tau_k)}{\tau_k} = \overline{a} \quad \uparrow$$

$$\lim_{k\to\infty} \frac{\phi(\sigma_k)}{\sigma_k} = \underline{a} \quad \uparrow$$

Furthermore, we have:

$$\phi(\tau_k) \leq g_k^{\tau}\, \tau_k + o(\tau_k) \tag{6}$$

where

$$g_k \in G^{\phi}(\tau_k), \quad o(\tau_k)\tau_k^{-1} \to 0 \qquad \text{for} \quad k\to\infty, \ \tau_k \to +0$$

Without loss of generality it may be assumed that

$$\lim_{k\to\infty} g_k^{\tau} = g^{\tau}$$

Dividing (6) by τ_k and passing to the limit for $k \to \infty$ we obtain

$$g^{\tau} \geq \overline{a}$$

By virtue of Lemma 1 $g^{\tau} \in \partial\phi(0)$, therefore

$$\phi(\sigma_k) \geq g^{\tau}\sigma_k - o(\sigma_k) \tag{7}$$

Dividing (7) by σ_k and passing to the limit when $k\to\infty$ we have a contradiction that proves the differentiability in any direction. By virtue of the weak convexity of f it is easy to obtain

$$\frac{\partial f(x)}{\partial e} \geq \max_{g \in \partial f} (g,e)$$

Now let

$$x^k = x + t_k e, \; t_k \to +0, \; k \to \infty$$

and

$$g_k \in \partial f(x^k), \; g_k \to g \in (x)$$

Then

$$f(x) - f(x^k) \geq (g_k, \; x - x^k) + r(x^k, x)$$

The division of the above inequality by t_k and the pass to the limit when $k \to \infty$ yield:

$$\frac{\partial f(x)}{\partial e} \leq (\tilde{g}, e) \leq \max_{g \in \partial f} (g, e)$$

and thus the proof is completed.

Lemma 3 implies that the necessary condition for the point be extremal is

$$0 \in \partial f(x^*) \tag{8}$$

however, unlike the case with the convex function, this condition is insufficient.

Local properties of the weakly convex functions do not differ from those of the convex functions but their global properties are radically dissimilar. Specifically, the weakly convex functions lack the salient feature of subgradients that enables us to prove the convergence of subgradient method, i.e., the positivity of scalar product of an arbitrary subgradient at some point x in the direction from the extremum point x^*:

$$(g, \; x - x^*) \geq 0 \tag{9}$$

for an arbitrary $g \in \partial f$

This and the fact that a shift in the direction of the antigradient does not assure a decrease in value of a function

being optimized both for the weakly convex and convex functions, complicate tangibly the proof of the subgradient method convergence.

Difficulties that present themselves in proving the convergence of non-relaxation algorithms are of common knowledge. However, in a number of cases they pay, opening new possibilities. In the following chapter we will describe certain criteria of convergence of iterative algorithms which made it possible to prove convergence of a number of algorithms whose behaviour is substantially non-monotonic.

4. CONVERGENCE OF ITERATIVE METHODS OF NON-LINEAR PROGRAMMING

General conditions of convergence of iterative procedures received attention of a lot of researchers. The most fundamental results appear to belong to W.I. Zangwill who suggested necessary and sufficient conditions of convergence of iterative methods of the mathematical programming [7]. However, the convergence theorems derived by W.I. Zangwill do not exhaust investigations conducted in this field, and many authors formulated other conditions that characterize convergence of iterative procedures. In spite of the fact that the later approaches are less general and universal they proved to be more helpful in investigations of specific algorithms. Take [7-9] as an example. It should be emphasized that in the majority of cases these works deal with convergence of algorithms whose objective function decreases monotonically as a process goes and, therefore, they are not applicable, in principle, to the case in hand. These and other reasons served as the starting point in the elaboration of conditions of convergence of iterative procedures with weakened properties of a monotonous variation of the objective function in the progress of the solution of an extremum problem. The approach set forth below is based on author's paper [12].

We will consider an algorithm of the mathematical programming as a certain rule of construction of a sequence $\{x^s\}$ of points of

an n - dimensional Euclidean space E^n. Conditions of convergence of this sequence will be formulated in terms of properties of this sequence and of a certain subset X^* of the space E^n which we will call the solution set. The algorithm will be thought of as the convergent algorithm if each limit point of a sequence generated by it belongs to the set X^*.

The basic convergence theorem is formulated as follows:

Theorem 1. Let the sequence $\{x^s\}$ and the set X^* be such that

A1) If $x^{s_k} \to x^* \in X^*$ then

$$\|x^{s_k+1} - x^{s_k}\| \to 0$$

A2) There exists a compact set K such that

$$x^s \in K$$

A3) If $x^{s_k} \to x' \notin X^*$, then there exists $\varepsilon_0 > 0$ such that for all $\varepsilon \leq \varepsilon_0$ and any k's there exists a point x^{t_k}, $t_k > s_k$ such that

$$\|x^{t_k} - x^{s_k}\| > \varepsilon$$

We will assume

$$t_k = \min_{t > s_k} \{t: \|x^t - x^{s_k}\| > \varepsilon\}$$

A4) There exists a continuous function W(x) such that

$$\overline{\lim_{k \to \infty}} W(x^{t_k}) < \lim_{k \to \infty} W(x^{s_k}) = W(x')$$

for arbitrary sequences $\{s_k\}$, $\{t_k\}$ satisfying condition A3.

A5) The function W assumes on X^* an everywhere incomplete set of values.

Then all limit points of the sequence $\{x^s\}$ belong to the set X^*.

This theorem is proved in [12]. A version of conditions given there varies to some extent from that given above; however, proofs of both theorems are practically similar. An assertion weaker than Theorem 1 is also of interest.

Theorem 2. Under the conditions of Theorem 1 A1-A4 there exists a limit point of the sequence $\{x^s\}$ which belongs to the set X^*. The proof of this theorem employs the same arguments as those of the proof of Theorem 1.

5. MINIMIZATION OF WEAKLY CONVEX FUNCTIONS

In this section we shall study convergence of the recurrent procedure

$$x^{s+1} = x^s - \rho_s g^s, \quad s = 0,1,... \tag{10}$$

for finding the unconditional minimum of the weakly convex function f. In the above relation $\rho_s > 0$ are step multipliers, $g^s \in \partial f(x^s)$ is the subgradient of the objective function f at the point x^s. Requirements placed upon the sequence of step multipliers will be stipulated in what follows.

To prove convergence of procedure (10) requires an auxiliary geometrical lemma. In a simplified form such lemma was first proved in [6].

Lemma 4. Let D be a convex compact set which does not contain a zero and let $\{y^n\}$ be an arbitrary set of vectors from D. By means of a sequence of numbers σ_n such that

$$0 \le \sigma_n \le 1, \ \sigma_n \to 0, \ \Sigma \sigma_n = \infty$$

let us form a sequence of vectors $\{z^n\}$ as follows:

$$z^0 = y^0$$

$$z^{n+1} = z^n + \sigma_n(y^{n+1} - z^n), \quad n = 0,1,\ldots$$

Denote by $\{n_k\}$ a sequence of indexes such that

$$(z^{n_k}, y^{n_k+1}) \geq \gamma > 0 \tag{11}$$

Then for some $\gamma > 0$ such a sequence exists and

$$\sum_{s=n_k}^{n_{k+1}-1} \sigma_s \leq C < \infty$$

Proof. It is obvious that $\{z^n\} \subset D$. Since $0 \in D$, then constants δ and Δ exist such that

$$0 < \delta \leq \|z^n\| \leq \Delta < \infty$$

Let us consider now the changes in the length of vectors z^n :

$$\|z^{n+1}\|^2 = \|z^n + \sigma_n(y^{n+1} - z^n)\|^2 = \|z^n\|^2 +$$

$$+ \sigma_n^2\|y^{n+1} - z^n\|^2 + 2\sigma_n(z^n, y^{n+1} - z^n) \leq \|z^n\|^2 +$$

$$+ 4\Delta^2\sigma_n^2 + 2\sigma_n((z^n, y^{n+1}) - \|z^n\|^2) \leq \|z^n\|^2 +$$

$$+ 4\Delta^2\sigma_n^2 + 2\sigma_n((z^n, y^{n+1}) - \delta^2)$$

If for all n

$$(z^n, y^{n+1}) \leq \frac{1}{2}\delta^2$$

then

$$\|z^{n+1}\|^2 \leq \|z^n\|^2 + 4\Delta^2\sigma_n^2 - \delta^2\sigma_n$$

Since $\sigma_n \to 0$, then for sufficiently large n

$$\|z^{n+1}\|^2 \leq \|z^n\|^2 - \frac{\delta^2}{2}\sigma_n$$

Summing the above inequality with respect to n from N to N + M - 1 we obtain

$$0 \leq \|z^{N+M}\|^2 \leq \|z^N\|^2 - \tag{12}$$

$$-\frac{\delta^2}{2}\sum_{n=N}^{N+M-1}\sigma_n \leq \Delta^2 - \frac{\delta^2}{2}\sum_{n=N}^{N+M-1}\sigma_n$$

The pass to the limit when $M \to \infty$ leads to a contradiction to the supposition (11). It follows that there exists a sequence $\{n_k\}$ such that

$$(z^{n_k}, y^{n_k+1}) \geq \gamma = \frac{1}{2}\delta^2 > 0$$

Further, from (12) it follows for sufficiently large k that

$$0 \leq \Delta^2 - \frac{\delta^2}{2}\sum_{s=n_k}^{n_{k+1}-1}\sigma_s$$

Hence

$$\sum_{s=n_k}^{n_{k+1}-1}\sigma_s \leq 2\,\frac{\Delta^2}{\delta^2}$$

which completes the proof.

The main result which will be proved here later is the proposition about convergence of procedure (10). At first the

solution set will be defined using the necessary extremum conditions:

$$X^* = \{x^* : 0 \in G(x^*)\}$$

The following theorem is valid:

Theorem 3. Let

$$\rho_s, \varepsilon s \to +0, \ \Sigma \rho_s = \infty$$

and the sequence $\{x^s\}$ defined by (10) be bounded. Then all limit points of this sequence belong to the set X^*.

Proof. In proving this theorem we shall employ the general conditions of convergence described in Section 3.

The objective function $f(x)$ is chosen as $W(x)$ and it is demonstrated that conditions A1-A4 will be also satisfied. For simplicity, we will assume that condition A5 is satisfied.

It is obvious, that the satisfaction of conditions A1,A2 follows directly from the assumptions of the proof.

Let $\{x^{n_k}\}$ be a convergent subsequence and

$$\lim_{k \to \infty} x^{n_k} = x' \in X^*$$

In this case $0 \bar{\in} G(x')$ and by virtue of $G(x)$ being upper semi-continuous it is possible to choose so small $\delta > 0$ that

$$0 \bar{\in} co\{G(x), \ \|x - x'\| \leq \delta\}$$

This is also true for the ε - subgradients. It is always possible to choose so small $\varepsilon, \delta > 0$ that

$$0 \bar{\in} co\{G_\gamma(x), \ \|x - x'\| \leq \delta, \ \gamma < \varepsilon\} = \tilde{G}_{\varepsilon,\delta}(x')$$

Then, if condition A3 is not satisfied, for k's large enough

$$g^s \in \tilde{G}_{\varepsilon,\delta}(x'), \quad s \geq n_k$$

and by virtue of separation theorems there exists a vector e such that

$$(g^s,e) \leq -C < 0$$

Therewith

$$(x^{s+1},e) = (x^s - \rho_s g^s,e) =$$

$$= (x^s,e) - \rho_s(g^s,e) \geq (x^s,e) + C\,\rho_s$$

The above inequality implies because of our assumptions an unlimited growth of the inner product (x^s,e). This implication obviously contradicts the assumption and, therefore, proves that condition A3 is satisfied.

Let for some small $\varepsilon > 0$

$$m_k = \min_{m \geq n_k} \{ m : \|x^m - x^{n_k}\| > \varepsilon \}$$

Requirements placed on ε will be refined later.

We meet the dominant difficulty at the following step of the proof; an estimation of a decrease in the objective function when passing from the point x^{n_k}. As the directions $-g^s$ are, generally speaking, not the directions of decrease in the function f(x) the problem of estimation of the function decrease is fairly difficult and rather unwieldy in view of the large number of computation.

Let us fix a sufficiently large k and examine a difference

$$f(x^m) - f(x^{n_k}) \leq (g^m, x^m - x^{n_k}) + \quad - r(x^{n_k}, x^m), \quad m > n_k$$

Estimate with greater precision the addend on the right side of this inequality.

$$(g^m, x^m - x^{n_k}) = -(g^m, \sum_{s=n_k}^{m-1} \rho_s g^s) =$$

$$= - \sum_{s=n_k}^{m-1} \rho_s (g^m, (\sum_{s=n_k}^{m-1} \rho_s)^{-1} \sum_{s=n_k}^{m-1} \rho_s g^s) =$$

$$= - \sum_{s=n_k}^{m-1} \rho_s (g^m, z_k^{m-1})$$

Vectors z_k^m can be obtained by means of the recurrent formula:

$$z_k^{s+1} = z_k^s + \sigma_s^{(k)} (g^{s+1} - z_k^s), \quad s = n_k, \; n_k + 1, \; n_k + 2, \ldots,$$

with the initial condition

$$z_k^{n_k} = g^{n_k}$$

and coefficients $\sigma_s^{(k)}$ equal to

$$\sigma_s^{(k)} = \rho_s (\sum_{m=n_k}^{s} \rho_m)^{-1}$$

It is easily seen that $0 \leq \sigma_s^{(k)} \leq 1$

$$\sum_{s > n_k} \sigma_s^{(k)} = \infty, \quad \sigma_s^{(k)} \to 0 \qquad \text{for } s \to \infty$$

Then by virtue of Lemma 4 there exists a sequence $\{s_i^k,\ i = 1,2,\ldots,\}$ of indexes such that

$$(g^{s_i^k},\ z^{s_i^k-1}) \geq \gamma > 0$$

and here

$$f(x^{s_i^k}) - f(x^{n_k}) \leq -\gamma \sum_{s=n_k}^{s_i^k-1} \rho_s +$$

$$+ \qquad r(x^{n_k},\ x^{s_i^k})$$

Choose from the sequence $\{s_i^k,\ i = 1,\ldots,\}$ a maximum index whose value does not exceed the index m_k and denote it by ν_i^k :

$$\nu_1^k = s_i^k \leq m_k < s_{i+1}^k$$

From the inequality (Lemma 4)

$$\sum_{s=s_i^k}^{s_{i+1}^k-1} \sigma_s^{(k)} \leq C$$

it follows that for sufficiently large k's

$$1 \geq \prod_{s=S^k}^{s_{i+1}^k-1} (1 - \sigma_s^{(k)}) \geq p > 0$$

which implies that

$$\left(\sum_{s=n_k}^{\nu_i^k-1} \rho_s\right) \left(\sum_{s=n_k}^{s_{i+1}^k-1} \rho_s\right)^{-1} \geq p$$

The above inequality may be put in another form:

$$\sum_{s=\nu_1^k}^{m_k-1} \rho_s \leq q \sum_{s=n_k}^{m_k-1} \rho_s$$

where $q = 1 - p < 1$

Summing up it is possible to say that we have constructed as a result the point $x^{\nu_1^k}$ such that

$$f(x^{\nu_1^k}) - f(x^{n_k}) \leq -\gamma \sum_{s=n_k}^{\nu_1^k-1} \rho_s - r(x^{n_k}, x^{\nu_1^k}) \quad (13)$$

and therewith

$$\sum_{s=\nu_1^k}^{m_k-1} \rho_s \leq q \sum_{s=n_k}^{m_k-1} \rho_s \quad (14)$$

If in a similar reasoning the point $x^{\nu_1^k}$ is considered as the initial one, than it is possible to show the existence of a point $x^{\nu_2^k}$ such that

$$f(x^{\nu_2^k}) - f(x^{\nu_1^k}) \leq -\gamma \sum_{s=\nu_1^k}^{\nu_2^k-1} \rho_s - r(x^{\nu_1^k}, x^{\nu_2^k})$$

and

$$\sum_{s=\nu_2^k}^{m_k-1} \rho_s \leq q \sum_{s=\nu_1^k}^{m_k-1} \rho_s \leq q^2 \sum_{s=n_k}^{m_k-1} \rho_s$$

Let us fix an arbitrary small $\tau > 0$ and repeat this process a required number of times in order to construct a sequence of points $\{x^{\nu_i^k},\ i = 1,2,\ldots,M\}$ such that for each i inequalities similar to (13)-(14) be satisfied:

$$f(x^{\nu_{i+1}^k}) - f(x^{\nu_i^k}) \le \gamma \sum_{s=\nu_i^k}^{\nu_{i+1}^k - 1} \rho_s - r(x^{\nu_i^k}, x^{\nu_{i+1}^k}) , \qquad (15)$$

$$\sum_{s=\nu_i^k}^{m_k - 1} \rho_s \le q^i \sum_{s=n_k}^{m_k - 1} \rho_s$$

and $q^m \le \tau$. It obviously suffices to repeat the above reasonings no more than $M = [\log_q \tau] + 1$ times. Summing (15) with respect to i from zero to $M - 1$ we obtain (assuming $\nu_0^k = n_k$ and denoting $\nu_M^k = t_k$):

$$f(x^{t_k}) - f(x^{n_k}) \le -\gamma \sum_{s=n_k}^{t_k - 1} \rho_s - \sum_{i=0}^{M-1} r(x^{\nu_i^k}, x^{\nu_{i+1}^k})$$

Addend in the right part of the inequality is evaluated as follows:

$$\left| \sum_{i=0}^{M-1} r(x^{\nu_i^k}, x^{\nu_{i+1}^k}) \right| \le \sum_{i=0}^{M-1} |r(x^{\nu_i^k}, x^{\nu_{i+1}^k}) \le$$

$$\leq M \sup |r(x,y)| = M\, \bar{r}_\varepsilon(x^{n_k})$$

$$\|x - x^{n_k}\| \leq \varepsilon$$

$$\|y - x^{n_k}\| \leq \varepsilon$$

For the k's that are large enough $\|x^{n_k} - x'\| \leq \varepsilon$ therefore

$$M\, \bar{r}_\varepsilon(x^{n_k}) \leq M \sup_{\substack{\|x-x'\| \leq 2\varepsilon \\ \|y-x'\| \leq 2\varepsilon}} |r(x,y)| \leq \varepsilon\, \delta(\varepsilon)$$

where $\delta(\varepsilon) \to 0$ for $\varepsilon \to 0$

Finally we obtain:

$$f(x^{m_k}) - f(x^{n_k}) \leq f(x^{t_k}) - f(x^{n_k}) +$$

$$+ |f(x^{m_k}) - f(x^{t_k})| \leq -\gamma \sum_{s=n_k}^{t_k-1} \rho_s + \varepsilon\delta(\varepsilon) +$$

$$+ C\,\|x^{m_k} - x^{t_k}\| \leq -\gamma \sum_{s=n_k}^{m_k-1} \rho_s + \gamma\tau \sum_{s=n_k}^{m_k-1} \rho_s$$

$$+ \varepsilon\delta(\varepsilon) + C' \sum_{s=t_k}^{m_k-1} \rho_s \leq -(\gamma - \gamma\tau) \sum_{s=n_k}^{m_k-1} \rho_s$$

$$+ \varepsilon\delta(\varepsilon) + C'\tau \sum_{s=n_k}^{m_k-1} \rho_s \leq -(\gamma - \gamma\tau - C'\tau) \sum_{s=n_k}^{m_k-1} \rho_s +$$

$$+ \varepsilon\,\delta(\varepsilon) \quad .$$

where τ may be assumed to be so small that

$$\gamma - \gamma\tau - C'\tau > \frac{1}{2}\gamma$$

In doing so we obtain:

$$f(x^{m_k}) - f(x^{n_k}) \leq -\frac{\gamma}{2} \sum_{s=n_k}^{m_k-1} \rho_s + \varepsilon\,\delta(\varepsilon) \qquad (16)$$

Furthermore,

$$\varepsilon < \|x^{m_k} - x^{n_k}\| \leq C \sum_{s=n_k}^{m_k-1} \rho_s$$

Substituting this estimate into (16) we obtain:

$$f(x^{m_k}) - f(x^{n_k}) \leq -\frac{\gamma\varepsilon}{2C} + \varepsilon\,\delta(\varepsilon)$$

It may be always assumed that

$$\delta(\varepsilon) \leq \frac{\gamma}{4C}$$

hence

$$f(x^{m_k}) - f(x^{n_k}) \leq -\frac{\gamma\varepsilon}{4C}$$

Passing to the limit when $k \to \infty$ we obtain:

$$\overline{\lim_{k\to\infty}}\; W(x^{m_k}) < \lim_{k\to\infty}\; w(x^{n_k})$$

which is what it was required to prove.

As a result the convergence of algorithm (10) is a consequence of the satisfaction of conditions A1-A5 of Theorem 1.

6. CONVEX CASE

For the problem of convex minimization some results can be obtained describing the behavior of process (10) in the case when $\varepsilon_s = \varepsilon = \text{const}$.

Theorem 4. Let the objective function $f(x)$ be convex

$$\rho_s \to +0 \ , \ \Sigma \rho_s = \infty \ , \ \varepsilon_s = \varepsilon > 0$$

Then, if the sequence $\{x^s\}$ is bounded, there exists only one convergent subsecuence $\{x^{s_k}\}$ such that

$$\lim_{k\to\infty} x^{s_k} = \bar{x}$$

and

$$f(\bar{x}) \leq \min_{x\in E^n} f(x) + \varepsilon$$

Proof. The proof will be based on the same formalism as in Theorem 3. Let

$$X^* = \{x^* : f(x^*) = \min f(x), \ x \in E^n\}$$

and

$$X^*_\varepsilon = \{x^* : f(x^*) \leq \min_{x\in E^n} f(x) + \varepsilon\}$$

Denote

$$W(x) = \min_{x^*\in X^*} \|x - x^*\|^2$$

In our case the role of a set of solutions will be played by X_ε^* . Let us verify whether conditions A1-A4 from Section 2 can be satisfied. It is obvious, that on no account condition A5 can be satisfied in this case and, therefore, it is possible to prove only a weakened convergence of process (10) in the spirit of Theorem 2.

Conditions A1, A2 are obviously satisfied in assumptions of this theorem: verify whether condition A3 is satisfied. Let there be some subsequence:

$$\lim_{k\to\infty} x^{n_k} = x' \bar{\in} X_\varepsilon^*$$

that is,

$$f(x') > \min_{x\in E^n} f(x) + \varepsilon$$

Assume the contrary to condition A3, that is,

$$\lim_{s\to\infty} x^s = x'$$

Then for an arbitrary $\delta > 0$ for a sufficiently large k

$$\|x^s - x'\| \leq \delta$$

for $s > n_k$. Choose $\delta > 0$ in such a way that the set

$$U_{4\delta}(x') = \{x : \|x-x'\| \leq 4\delta\}$$

does not intersect with the set X^*_ε : $U_{4\delta}(x') \cap X_\varepsilon^* = \phi$. Then in suppositions of the proof for an arbitary $x^* \in X^*$ and $s > n_k$:

$$\|x^{s+1} - x^*\|^2 = \|x^s - \rho_s g^s - x^*\| =$$

$$= \|x^s - x^*\|^2 + \rho_s^2 \|g^s\|^2 - 2\rho_s (g^s, x^s - x^*) \leq \tag{17}$$

$$\leq \|x^s - x^*\|^2 + C\rho_s^2 - 2\rho_s(g^s, x^s - x^*)$$

since

$$x^s \in U_{4\delta}(x')$$

then

$$\varepsilon < f(x^s) - f(x^*) \leq (g^s, x^s - x^*) + \varepsilon$$

whence we have for $s > n_k$

$$(g^s, x^s - x^*) \geq \gamma > 0$$

Substituting the above inequality into (17) we obtain

$$W(x^{s+1}) \leq W(x^s) + C\rho_s^2 - 2\gamma\rho_s$$

or for sufficiently large k

$$W(x^{s+1}) \leq W(x^s) - \gamma\rho_s \tag{18}$$

Summing (18) with respect to s from n_k to m-1 we obtain:

$$W(x^m) \leq W(x^{n_k}) - \gamma \sum_{s=n_k}^{m-1} \rho_s \tag{19}$$

Passing in the above inequality to the limit when $m \to \infty$ we have a contradiction to the boundedness of the continuous function $W(x)$ on $U_{4\delta}(x')$. The obtained contradiction proves the fact that condition A3 is satisfied. Let

$$m_k = \min_{m > n_k} m : \|x^m - x^{n_k}\| > \delta$$

For k's that are large enough

$$U_\delta(x^{n_k}) \subset U_{2\delta}(x') \subset U_{4\delta}(x')$$

therefore the estimate of (19) is also valid for $m = m_k$

$$W(x^{m_k}) \leq W(x^{n_k}) - \gamma \sum_{s=n_k}^{m_k-1} \rho_s$$

However,

$$\delta < \|x^{m_k} - x^{n_k}\| \leq C \sum_{s=n_k}^{m_k-1} \rho_s$$

By means of the above estimate we finally obtain:

$$W(x^{m_k}) \leq W(x^{n_k}) - \frac{\gamma\delta}{C}$$

and passing to the limit when $k \to \infty$

$$\overline{\lim_{k\to\infty}} W(x^{m_k}) < \lim_{k\to\infty} W(x^{n_k})$$

which, by virtue of Theorem 2, proves our proposition.

In all probability the assertion of this theorem cannot be strengthened unless additional hypotheses concerning the choice of vectors g^s from appropriate sets $G_\varepsilon(x^s)$ of ε - subgradients are involved.

It is also of interest to estimate a deviation of the limit points of the sequence $\{x^s\}$ from the set of solutions X_ε^* . If we denote

$$d = \sup_{x_\varepsilon^* \in X_\varepsilon^*} \inf_{x^* \in X^*} \|x^* - x_\varepsilon^*\|$$

then from geometrical considerations it is easily shown that all

limit points of the sequence $\{x^s\}$ occur in the set

$$x^* + d_\varepsilon S$$

where S is a unit ball and the addition is meant in Minkovsky's sense.

7. APPENDIX AND GENERALIZATIONS

An essential feature that distinguishes the result of Theorem 3 as compared to that obtained earlier in [13] is, as applied to minimax problems of the type

$$\min_x \max_y f(x,y) \tag{20}$$

the possibility to rid oneself of the check of exactness of the solution of an auxiliary problem of finding the internal maximum:

$$\varphi(x) = \max_y f(x,y)$$

This enables us to justify the application of Arrow-Gurwitz' method

$$x^{s+1} = x^s - \rho_s f'_x(x^s,y^s) \tag{21}$$

$$y^{s+1} = y^s + \delta_s f'_y(x^s,y^s) \tag{22}$$

in the solution of problem (20) on the basis of broader assumptions than common assumptions of strict convexity-concavity or similar ones. Under some of them concerning the relation between step multipliers it proves to be possible to consider iterative relation (2) as the ε - subgradient method of minimization of the

function $\phi(x)$. Convergence of method (21)-(22) is here an implication of Theorem 3. Results obtained in this field are described in more detail in [14]. Of great practical interest is also the development of methods for regulating step multipliers in procedure (10). Basically, Theorem 3 asserts that the ε - subgradient methods converge under the same assumptions as the subgradient methods. In all probability, the ideas that underlie the subgradient methods are applicable to the ε - subgradient methods when their step multipliers are regulated, and furthermore, the computational effect is also the same.

A non-formal requirement here consists in giving up the exact computation of the objective function as stated earlier in the introduction to this paper. For instance, the generalization on the case of ε - subgradient method of step regulation [11] presents no difficulties.

REFERENCES

[1] Balinski, M.L. and P. Wolfe, eds., Nondifferentiable Optimization, *Mathematical Programming, Study 3*, North-Holland Publishing Co., Amsterdam, 1975.

[2] Rockafellar, R.T. *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

[3] Bertsecas, D.P. and S.K. Mitter, A Descent Numerical Method for Optimization Problems with Nondifferentiable Cost Functionals, *SIAM Journal Control*, Vol. 11, 4 (1973).

[4] Lemarechal, C., Nondifferentiable Optimization; Subgradient and ε - Subgradient Methods, Lecture Notes: Numerical Methods in Optimization and Operations Research, Springer Verlag, August 1975, 191-199.

[5] Rockafellar, R.T., The Multiplier Method of Hesten and Powell Applied to Convex Programming, *JOTA*, Vol. 12, 6 (1974).

[6] Nurminski, E.A., The Quasigradient Method for the Solving of Nonlinear Programming Problems, *Cybernetics*, Vol. 9, 1, (Jan-Feb 1973), 145-150, Plenum Publishing Corporation, N.Y., London.

[7] Zangwill, W.I., Convergence Conditions for Nonlinear Programming Algorithms, *Management Science*, Vol. 16, 1 (1969), 1-13.

[8] Wolf, P., *Convergence Theory in Nonlinear Programming*, North-Holland Publishing Company, 1970, 1-36.

[9] Meyer, G.G.L., A Systematic Approach to the Synthesis of Algorithms, *Numerical Mathematics*, Vol. 24, 4 (1975), 277-290.

[10] Rheinboldt, W.C., A Unified Convergence Theory for a Class of Iterative Processes, *SIAM Journal Numerical Analysis*, Vol. 5, 1 (1968),

[11] Nurminski, E.A. and A.A. Zhelikhovski, Investigation of One Regulating Step, *Cybernetics*, Vol. 10, 6 (Nov-Dec 1974), 1027-1031, Plenum Publishing Corporation, New York.

[12] Nurminski, E.A., Convergence Conditions for Nonlinear Programming Algorithms, *Kybernetika*, 6 (1972), 79-81 (in Russian).

[13] Nurminski, E.A. and A.A. Zhelikhovski, ε - Quasigradient Method for Solving Nonsmooth External Problems, *Cybernetics*, Vol. 13, 1 (1977), 109-114, Plenum Publishing Corporation, N.Y., London.

[14] Nurminski, E.A. and P.I. Verchenko, Convergence of Algorithms for Finding Saddle Points, *Cybernetics*, Vol. 13, 3, 430-434, Plenum Publishing Corporation, N.Y., London.

FAVORABLE CLASSES OF LIPSCHITZ-CONTINUOUS FUNCTIONS IN SUBGRADIENT OPTIMIZATION

R. Tyrrell Rockafellar
Department of Mathematics
University of Washington
Seattle, Washington
USA

1. INTRODUCTION

A function $f : R^n \to R$ is said to be *locally Lipschitzian* if for each $x \in R^n$ there is a neighborhood X of x such that, for some $\lambda \geq 0$,

$$|f(x'') - f(x')| \leq \lambda |x'' - x'| \quad \text{for all } x' \in X,\ x'' \in X \tag{1.1}$$

Examples include continuously differentiable functions, convex functions, concave functions, saddle functions and any linear combination or pointwise maximum of a finite collection of such functions.

Clarke (1975 and 1980), has shown that when f is locally Lipschitzian, the generalized directional derivative

$$f^\circ(x;v) = \limsup_{\substack{x' \to x \\ t \searrow 0}} \frac{f(x' + tv) - f(x')}{t} \tag{1.2}$$

is for each x a finite, sublinear (i.e., convex and positively homogeneous) function of v. From this it follows by classical convex analysis that the set

$$\partial f(x) = \{y \in R^n \mid y \cdot v \leq f^\circ(x;v) \text{ for all } v \in R^n\} \tag{1.3}$$

is nonempty, convex, compact, and satisfies

$$f^\circ(x;v) = \max \{y \cdot v \mid y \in \partial f(x)\} \qquad \text{for all } v \in R^n. \qquad (1.4)$$

The elements of $\partial f(x)$ are what Clarke called "generalized gradients" of f at x, but we shall call them *subgradients*. As Clarke has shown, they are the usual subgradients of convex analysis when f is convex or concave (or for that matter when f is a saddle function). When f is continuously differentiable, $\partial f(x)$ reduces to the singleton $\{\nabla f(x)\}$.

In subgradient optimization, interest centers on methods for minimizing f that are based on being able to generate for each x at least one (but not necessarily every) $y \in \partial f(x)$, or perhaps just an approximation of such a vector y. One of the main hopes is that by generating a number of subgradients at various points in some neighborhood of x, the behavior of f around x can roughly be assessed. In the case of a convex function f this is not just wishful thinking, and a number of algorithms, especially those of bundle type (e.g., Lemarechal 1975 and Wolfe 1975) rely on such an approach. In the nonconvex case, however, there is the possibility, without further assumptions on f than local Lipschitz continuity, that the multifunction $\partial f : x \to \partial f(x)$ may be rather bizarrely disassociated from f. An example given at the end of this section has f locally Lipschitzian, yet such that there exist many other locally Lipschitzian functions g, not merely differing from f by an additive constant, for which $\partial g(x) = \partial f(x)$ for all x. Subgradients alone cannot discriminate between the properties of these different functions and therefore cannot be effective in determining their local minima.

Besides the need for conditions that imply a close connection between the behavior of f and the nature of ∂f, it is essential to ensure that ∂f has adequate continuity properties for the construction of "approximate" subgradients and in order to prove the convergence of various algorithms involving subgradients. The key seems to lie in postulating the existence of the ordinary directional derivatives

$$f'(x;v) = \lim_{t \downarrow 0} \frac{f(x + tv) - f(x)}{t} \tag{1.5}$$

and some sort of relationship between them and ∂f. Mifflin (1977a and 1977b), most notably has worked in this direction.

In the present article we study the relationship between f' and ∂f for several special classes of locally Lipschitzian functions that suggest themselves as particularly amenable to computation. First we give some new results about continuity properties of f' when f belongs to the rather large class of functions that are "subdifferentially regular". Next we pass to functions f that are *lower*-C^k for some k, $1 \leq k \leq \infty$, in the following sense: for each point $\bar{x} \in R^n$ there is for some open neighborhood X of $\bar{x}$ a representation

$$f(x) = \max_{s \in S} F(x,s) \qquad \text{for all } x \in X, \tag{1.6}$$

where S is a compact topological space and $F : X \times S \to R$ is a function which has partial derivatives up to order k with respect to x and which along with all these derivatives is continuous not just in x, but jointly in $(x,s) \in X \times S$. We review the strong results obtained by Spingarn (forthcoming) for lower-C^1 functions, which greatly illuminate the properties treated by Mifflin (1977b), and we go on to show that for $k \geq 2$ the classes of lower-C^k functions all coincide and have a simple characterization.

Before proceeding with this, let us review some of the existence properties of f' and continuity properties of ∂f that are possessed by any locally Lipschitzian function. This will be useful partly for background but also to provide contrast between such properties, which are not adequate for purposes of subgradient optimization, and the refinements of them that will be featured later.

Local Lipschitz continuity of a function $f : R^n \to R$ implies by a classical theorem of Rademacher (see Stein 1970) that for almost every $x \in R^n$, ∇f is differentiable at x, and moreover that the gradient mapping ∇f, on the set where it exists, is locally bounded.

Given any $x \in R^n$, a point where f may or not happen to be differentiable, there will in particular be in every neighborhood of x a dense set of points x' where $\nabla f(x')$ exists, and for any sequence of such points converging to x, the corresponding sequence of gradients will be bounded and have cluster points, each of which is, of course, the limit of some convergent subsequence. Clarke demonstrated in Clarke (1975) that $\partial f(x)$ is the convex hull of all such possible limits:

$$\partial f(x) = \text{co}\{\lim \nabla f(x') \mid x' \to x, \text{ f differentiable at } x'\}. \quad (1.7)$$

Two immediate consequences (also derivable straight from properties of $f^\circ(x;v)$ without use of Rademacher's theorem) are first that ∂f is *locally bounded*: for every x one has that

$$\bigcup_{x' \in X} \partial f(x') \text{ is bounded for some neighborhood X of x,} \quad (1.8)$$

and second that ∂f is *upper semicontinuous* in the strong sense:

$$\begin{aligned} &\text{for any } \varepsilon > 0 \text{ there is a } \delta > 0 \text{ such that} \\ &\partial f(x') \subset \partial f(x) + \varepsilon B \text{ whenever } |x'-x| \leq \delta \quad , \end{aligned} \quad (1.9)$$

where

$$B = \text{closed unit Euclidean ball} = \{x \mid |x| \leq 1\} \quad . \quad (1.10)$$

The case where $\partial f(x)$ consists of a single vector y is the one where f is *strictly* differentiable at x with $\nabla f(x) = y$, which by definition means

$$\lim_{\substack{x' \to x \\ t \downarrow 0}} \frac{f(x' + tv) - f(x')}{t} = y \cdot v \quad \text{for all } v \in R^n \quad . \quad (1.11)$$

This is pointed out in Clarke (1975). From (1.7) it is clear that this property occurs if and only if x belongs to the domain of ∇f, and ∇f is continuous at x relative to its domain.

We conclude this introduction with an illustration of the abysmal extent to which ∂f could in general, without assumptions beyond local Lipschitz continuity, fail to agree with ∇f on the domain of ∇f and thereby lose contact with the local properties of f.

Counterexample

There is a Lipschitzian function $f : R^n \to R$ such that

$$\partial f(x) = [-1,1]^n \qquad \text{for all } x \in R^n \ . \tag{1.12}$$

To construct f, start with a measurable subset A of R such that for every nonempty open interval $I \subset R$, both $\text{mes}[A \cap I] > 0$ and $\text{mes}[A \setminus I] > 0$. (Such sets do exist and are described in most texts on Lebesgue measure.) Define $h : R \to R$ by

$$h(t) = \int_0^t \theta(\tau)\, d(\tau) \ , \qquad \text{where } \theta(t) = \begin{cases} 1 & \text{if } t \in A \ , \\ -1 & \text{if } t \in A \ . \end{cases}$$

Since $\|\theta\|_\infty = 1$, h is Lipschitzian on R with Lipschitz constant $\lambda = 1$. Hence $h'(t)$ exists for almost every t, and $|h'(t)| \leq 1$. In fact $h' = \theta$ almost everywhere, from which it follows by the choice of A that the sets $\{t \mid h'(t) = 1\}$ and $\{t \mid h'(t) = -1\}$ are both dense in R. Now let

$$f(x) = \sum_{i=1}^{m} h(x_i) \qquad \text{for } x = (x_1, \ldots, x_n) \ .$$

Then f is Lipschitzian on R^n with gradient

$$\nabla f(x) = (h'(x_1), \ldots, h'(x_n))$$

existing if and only if $h'(x_i)$ exists for $i = 1, \ldots, n$. Therefore $\nabla f(x) \in [-1,1]^n$ whenever $\nabla f(x)$ exists, and for each of the corner points e of $[-1,1]^n$ the set $\{x \mid \nabla f(x) = e\}$ is dense in R^n. Formula (1.7) implies then that (1.12) holds.

Note that every translate $g(x) = f(x - a)$ has $\partial g \equiv \partial f$, because ∂f is constant, and yet $g - f$ may be far from constant.

2. SUBDIFFERENTIALLY REGULAR FUNCTIONS

A locally Lipschitzian function $f : R^n \to R$ is *subdifferentially regular* if for every $x \in R^n$ and $v \in R^n$ the ordinary directional derivative (1.5) exists and coincides with the generalized one in (1.2):

$$f'(x;v) = f^\circ(x;v) \qquad \text{for all } x,v.$$

Then in particular $f'(x;v)$ is a finite, subadditive function of v; this property in itself has been termed the *quasidifferentiability* of f at x by Pshenichnyi (1971).

THEOREM 1. (Clarke 1975). *If f is convex or lower-C^k on R^n for some $k \geq 1$, then f is not only locally Lipschitzian but subdifferentially regular.*

Clarke did not study lower-C^k functions as such but proved in Clarke (1975) a general theorem about the subgradients of "max functions" represented as in (1.6) with $F(x,s)$ not necessarily differentiable in x. His theorem says in the case of lower-C^k functions that

$$\partial f(x) = \operatorname{co}\{\nabla_x F(x,s) \mid s \in I(x)\} \tag{2.1}$$

where

$$I(x) = \arg \max_{s \in S} F(x,s) \ . \tag{2.2}$$

It follows from this, (1.4), and the definition of subdifferential regularity, that

$$f'(x;v) = \max\{\nabla_x F(x,s) \cdot v \mid s \in I(x)\} \tag{2.3}$$

for lower-C^1 functions, a well known fact proved earlier by Danskin (1967).

The reader should bear in mind, however, that Theorem 1 says considerably more in the case of lower-C^k functions than just this.

By asserting the equality of f' and f°, it implies powerful things about the semicontinuity of f' and strict differentiability of f. We underline this with the new result which follows.

THEOREM 2. *For a function* $f : R^n \to R$, *the following are equivalent:*

(a) f *is locally Lipschitzian and subdifferentially regular;*

(b) $f'(x;v)$ *exists finitely for all* x,v, *and is upper semicontinuous in* x.

Proof.

(a) ⇒ (b). This is the easy implication; since $f' = f^\circ$ under subdifferential regularity, we need only apply (1.4) and (1.9).

(b) ⇒ (a). For any x' and v the function $Q(t) = f(x' + tv)$ has both left and right derivatives at every t by virtue of (b):

$$Q'_+(t) = f'(x' + tv\,;\, v), \qquad Q'_-(t) = -f'(x' + tv\,;\, -v) \tag{2.4}$$

Moreover, the upper semicontinuity in (b) implies that for any fixed x and v there is a convex neighborhood X of x and a constant $\lambda \geq 0$ such that

$$f'(x' + tv\,;\, v) \leq \lambda \text{ and } -f'(x' + tv\,;\, v) \geq -\lambda \text{ when } x' + tv \in X \tag{2.5}$$

Since Q has right and left derivatives everywhere and these are locally bounded, it is the integral of these derivatives (cf. Saks (1937)):

$$Q(t_1) - Q(t_0) = \int_{t_0}^{t_1} Q'_+(\tau)\,d\tau = \int_{t_0}^{t_1} Q'_-(\tau)\,d\tau$$

From this and (2.5) it follows that

$$|f(x' + tv) - f(x')| \leq \lambda t \text{ when } x' \in X,\ x' + tv \in X .$$

Thus the local Lipschitz property (1.1) holds as long as $x'' - x'$ is some multiple of a fixed v. To complete the argument, consider not just one v but a basis $v_1, \ldots, v_n$ for R^n.

Each $x \in R^n$ has convex neighborhoods X_i and constants $\lambda_i \geq 0$ such that

$$|f(x' + tv_i) - f(x')| \leq \lambda_i t \quad \text{when } x' \in X_i,\ x' + tv_i \in X_i \tag{2.6}$$

Then there is a still smaller neighborhood X of x and a constant $\alpha \geq 0$ such that for $x' \in X$ and $x'' \in X$ one has

$$x'' = x' + t_1 v_1 + \ldots + t_n v_n$$

with x' and $x' + t_1 v_1 \in X_1$, $x' + t_1 v_1$ and $(x' + tv_1) + tv_2 \in X_2$, and so forth, and

$$|t_1| + \ldots + |t_n| \leq \alpha |x'' - x'| \quad .$$

Then by (2.6)

$$|f(x'') - f(x')| \leq |f(x' + t_1 v_1) - f(x')| + |f(x' + tv_1 + tv_2) - f(x' + tv_1)| + \ldots$$

$$\leq \lambda_1 t_1 + \lambda_2 t_2 + \ldots + \lambda_n t_n$$

$$\leq (\lambda_1 + \lambda_2 + \ldots + \lambda_n)\, \alpha |x'' - x'| \quad .$$

In other words, f satisfies the Lipschitz condition (1.1) with $\lambda = (\lambda_1 + \ldots + \lambda_n)\alpha$. Thus f is locally Lipschitzian.

We argue next that $f'(x; v) \leq f^\circ(x; v)$ for all x, v by (1.2), and therefore via (1.7) that

$$f^\circ(x; v) = \limsup_{x' \to x} f'(x'; v) \quad . \tag{2.7}$$

The "lim sup" in (2.7) is just $f'(x'; v)$ under (b), so we conclude that $f'(x; v) = f^\circ(x; v)$. Thus (b) does imply (a), and the proof of Theorem 2 is complete. □

COROLLARY 1. *Suppose* f *is locally Lipschitzian and subdifferentially regular on* R^n *and let* D *be the set of all points where* f *happens to be differentiable. Then at each* $x \in D$, f *is in fact strictly differentiable. Furthermore, the gradient mapping is continuous relative to* D.

COROLLARY 2. *If* f *is locally Lipschitzian and subdifferentially regular on* R^n, *then* ∂f *is actually single-valued at almost every* $x \in R^n$.

These corollaries are immediate from the facts about differentiability of f that were cited in §1 in connection with formula (1.7). The properties they assert have long been known for convex functions but have not heretofore been pointed out as properties of all lower-C^k functions. They hold for such functions by virtue of Theorem 1.

COROLLARY 3. *Suppose* f *is locally Lipschitzian and subdifferentially regular on* R^n. *If* g *is another locally Lipschitzian function on* R^n *such that* $\partial g = \partial f$, *then* $g = f + \text{const}$.

Proof. By Corollary 2, ∂g is single-valued almost everywhere. Recalling that g is strictly differentiable wherever ∂g is single-valued, we see that at almost every $x \in R^n$ the function $h = g - f$ is strictly differentiable with $\nabla h(x) = \nabla g(x) - \nabla f(x) = 0$. Since h is locally Lipschitzian, the fact that $\nabla h(x) = 0$ for almost all x implies h is a constant function. □

COROLLARY 4. *Suppose* f *is locally Lipschitzian and subdifferentially regular on* R^n. *Then for every continuously differentiable mapping* $\xi: R \to R^n$, *the function* $Q(t) = f(\xi(t))$ *has right and left derivatives* $Q'_+(t)$ *and* $Q'_-(t)$ *everywhere, and these satisfy*

$$Q'_+(t) = \limsup_{\tau \to t} Q'_+(\tau) = \limsup_{\tau \to t} Q'_-(\tau) \quad ,$$
$$Q'_-(t) = \liminf_{\tau \to t} Q'_+(\tau) = \liminf_{\tau \to t} Q'_-(\tau) \quad . \tag{2.8}$$

Proof. The function Q is itself locally Lipschitzian and subdifferentiably regular (cf. Clarke 1980). Apply Theorem 2 to Q, noting that $Q'_+(t) = Q'(t;1) = Q^\circ(t;1)$ and $Q'_-(t) = -Q'(t;-1) = -Q^\circ(t;-1)$, and hence also $\partial Q(t) = [Q'_-(t), Q'_+(t)]$. The reason $Q'_+(\tau)$ and $Q'_-(\tau)$ can appear interchangeably in (2.8) is that by specialization of (1.7) to Q, as well as the characterizations of Q'_+ and Q'_- just mentioned, one has

$$Q'_+(\tau) = \limsup_{\tau' \to \tau} Q'(\tau') \quad , \quad Q'_-(\tau) = \liminf_{\tau' \to \tau} Q'(\tau') \quad ,$$

where the limits in this case are over the values τ' where $Q'(\tau')$ exists. □

3. LOWER-C^1 FUNCTIONS AND SUBMONOTONICITY

The multifunction $\partial f : R^n \rightrightarrows R^n$ is said to be *monotone* if

$$(x' - x'') \cdot (y' - y'') \geq 0 \text{ whenever } y' \in \partial f(x'), \; y'' \in \partial f(x'') \quad (3.1)$$

This is an important property of long standing in nonlinear analysis, and we shall deal with it in §4. In this section our aim is to review results of Spingarn (forthcoming) on two generalizations of monotonicity and their connection with subdifferentially regular functions and lower-C^1 functions. The generalized properties are as follows: ∂f is *submonotone* if

$$\liminf_{\substack{x' \to x \\ y' \in \partial f(x')}} \frac{(x' - x) \cdot (y' - y)}{|x' - x|} \geq 0 \quad , \quad \forall x, \; \forall y \in \partial f(x) \quad , \tag{3.2}$$

and it is *strictly submonotone* if

$$\liminf_{\substack{x' \to x \\ x'' \to x \\ y' \in \partial f(x') \\ y'' \in \partial f(x'')}} \frac{(x'' - x') \cdot (y'' - y')}{|x'' - x'|} \geq 0 \quad , \quad \forall x \quad . \tag{3.3}$$

To state the results, we adopt Spingarn's notation:

$$\partial f(x)_v = \{y \in \partial f(x) \mid (y' - y) \cdot v \leq 0, \forall y' \in \partial f(x)\} \qquad (3.4)$$

Thus $\partial f(x)_v$ is a certain face of the compact convex set $\partial f(x)$, the one consisting of all the points y at which v is a normal vector. Let us also recall the notion of *semismoothness* of f introduced by Mifflin (1977): this means that

$$\text{whenever } x^j \to x,\ v^j \to v,\ t_j \downarrow 0,\ y^j \to y, \text{ with } y^j \in \partial f(x^j + t_j v^j), \text{ then one has } y \cdot v = f'(x;v) \quad . \qquad (3.5)$$

THEOREM 3 (Spingarn (forthcoming)). *The following properties of a locally Lipschitzian function* $f:R^n \to R$ *are equivalent:*

(a) f *is both subdifferentially regular and semismooth;*

(b) ∂f *is submonotone;*

(c) ∂f *is directionally upper semicontinuous in the sense that for every* $x \in R^n$, $v \in R^n$ *and* $\varepsilon > 0$, *there is a* $\delta > 0$ *such that*

$$\partial f(x + tv') \subset \partial f(x)_v + \varepsilon B \quad \text{when } |v' - v| < \delta \text{ and } 0 < t < \delta . \qquad (3.6)$$

THEOREM 4 (Spingarn (forthcoming)). *The following properties of a locally Lipschitzian function* $f:R^n \to R$ *are equivalent:*

(a) f *is lower* C^1;

(b) ∂f *is strictly submonotone:*

(c) ∂f *is strictly directionally upper semicontinuous in the sense that for every* $x \in R^n$, $v \in R^n$ *and* $\varepsilon > 0$, *there is a* $\delta > 0$ *that*

$$(y''-y')\cdot v' \geq -\varepsilon \text{ when } |x'-x| < \delta,\ |v'-v| < \delta,\ 0 < t < \delta,\ y' \in \partial f(x') \text{ and } y'' \in \partial f(x'+tv'). \tag{3.7}$$

Spingarn has further given a number of valuable counter-examples in his forthcoming paper. These demonstrate that

$$\partial f \text{ submonotone} \not\Rightarrow \partial f \text{ strictly submonotone}, \tag{3.8}$$

$$f \text{ subdifferentially regular} \not\Rightarrow f \text{ lower-}C^1, \tag{3.9}$$

$$f \text{ quasidifferentiable and semismooth} \not\Rightarrow f \text{ subdifferentially regular.} \tag{3.10}$$

Comparing Theorems 3 and 4, we see that lower-C^1 functions have distinctly sharper properties than the ones of quasidifferentiability and semismoothness on which Mifflin, for instance, based his minimization algorithm (1977a). In perhaps the majority of applications of subgradient optimization the functions are actually lower-C^1, or even lower-C^∞. This suggests the possibility of developing improved algorithms which take advantage of the sharper properties. With this goal in mind, we explore in the next section what additional characteristics are enjoyed by lower-C^k functions for $k > 1$.

4. LOWER-C^2 FUNCTIONS AND HYPOMONOTONICITY

The properties of lower-C^k functions for $k \geq 2$ turn out, rather surprisingly, to be in close correspondence with properties of convex functions It is crucial, therefore, that we first take a look at the latter. We will have an opportunity at the same time to verify that convex functions are special examples of lower-C^∞ functions. The reader may have thought of this as obvious, because a convex function can be represented as a maximum of affine (linear-plus-a-constant) functions, which certainly are C^∞. The catch is, however, that a representation must be constructed in terms of affine functions which depend *continuously* on a parameter s ranging over a *compact* set, if the definition of lower-C^∞ is to be satisfied.

We make use now of the concept of monotonicity of ∂f defined at the beginning of §3.

THEOREM 5. *For a locally Lipschitzian function* $f:R^n \to R$, *the following properties are equivalent:*

(a) f *is convex ;*

(b) ∂f *is montone ;*

(c) for each $\bar{x} \in R^n$ *there is a neighborhood* X *of* $\bar{x}$ *and a representation of* f *as in* (1.6) *with* S *a compact topological space,* F(x,s) *affine in* x *and continuous in* s .

Proof. (a) $\Rightarrow$ (c). In terms of the conjugate f^* of the convex function f, we have the formula

$$f(x) = \max_{y \in R^n} \{y \cdot x - f^*(y)\} \text{ for all } x , \tag{4.1}$$

where the maximum is attained at y if and only if $y \in \partial f(x)$ (see Rockafellar 1970, §23). Any $\bar{x}$ has a compact neighborhood X on which ∂f is bounded. The set

$$S = \{(y,\beta) \in R^{n+1} \mid \exists x \in X \text{ with } y \in \partial f(x), \beta = y \cdot x - f(x)\}$$

is then compact, and we have as a special case of (4.1)

$$f(x) = \max_{(y,\beta) \in S} \{y \cdot x - \beta\} .$$

This is a representation of the desired type with $s = (y,\beta)$, $F(x,s) = y \cdot x - \beta$.

(c) $\Rightarrow$ (a). The representations in (c) imply certainly that f is convex relative to some neighborhood of each point. Thus for any fixed x and v the function $Q(t) = f(x + tv)$ has left and right derivatives Q'_- and Q'_+ which are nondecreasing in some neighborhood of each t. These derivatives are then nondecreasing relative to $t \in (-\infty, \infty)$, and it follows from this that

Q is a convex function on $(-\infty,\infty)$ (cf. Rockafellar 1970, §24) Since this is true for every x and v, we are able to conclude that f itself is convex.

(a) $\Rightarrow$ (b). This is well-known (cf. Rockafellar 1970, §24).

(b) $\Rightarrow$ (a). A direct argument could be given, but we may as well take advantage of Theorem 3. Monotonicity of ∂f trivially implies submonotonicity, so we know from Theorem 3 that f is subdifferentially regular. Fixing any x and v, we have by the monotonicity of ∂f that

$$((x + t''v) - (x + t'v)) \cdot (y'' - y') \geq 0 \quad \text{when}$$

$$t' < t,\ y' \in \partial f(x + t'v),\ y'' \in \partial f(x + t''v) .$$

This implies

$$\sup_{y' \in \partial f(x + t'v)} y' \cdot v \leq \inf_{y'' \in \partial f(x + t''v)} y'' \cdot v = -\sup_{y'' \in \partial f(x + t''v)} [-y'' \cdot v],$$

or equivalently (by 1.4) and subdifferential regularity)

$$f'(x + t'v; v) \leq -f'(x + t''v; -v) \quad \text{when } t' \leq t'' . \tag{4.2}$$

Since also

$$-f'(x'; -v) \leq f'(x'; v) \text{ for all } x', v,$$

by the sublinearity of $f'(x'; \cdot)$, (4.2) tells us that the function $Q(t) = f(x + tv)$ has left and right derivatives which are everywhere nondecreasing in $t \in (-\infty, \infty)$. Again as in the argument that (c) implies (a), we conclude from this fact that f is convex on R^n. □

COROLLARY 5. *Every convex function* $f: R^n \to R$ *is in particular lower-*C^∞.

Proof. In the representation in (c) we must have $F(x,s) = a(s)\cdot x - \alpha(s)$ for certain $a(s) \in R^n$ and $\alpha(s) \in R$ that depend continuously on x. This is the only way that $F(x,s)$ can be affine in x and continuous in s. Then, of course, $F(x,s)$ has partial derivatives of all orders with respect to x, and these are all continuous in (x,s). □

Let us now define two notions parallel to Spingarn's submonotonicity and strict submonotonicity: ∂f is *hypomonotone* if

$$\liminf_{\substack{x' \to x \\ y' \in \partial f(x')}} \frac{(x'-x)\cdot(y'-y)}{|x'-x|^2} > -\infty \quad \text{for all } x \text{ and } y \in \partial f(x) \tag{4.3}$$

and *strictly hypomonotone* if

$$\liminf_{\substack{x' \to x \\ x'' \to x \\ y' \in \partial f(x') \\ y'' \in \partial f(x'')}} \frac{(x''-x')\cdot(y''-y')}{|x''-x'|^2} > -\infty \quad \text{for all } x\,. \tag{4.4}$$

Clearly hypomonotone implies submonotone, and strictly hypomonotone implies strictly submonotone. We have little to say here about hypomonotonicity itself, but the importance of strict hypomonotonicity is demonstrated by the following result.

THEOREM 6. *For a locally Lipschitzian function f on R^n, the following properties are equivalent:*

(a) f is lower-C^2 ;

(b) ∂f is strictly hypomonotone ;

(c) For every $\bar{x} \in R^n$ there is a convex neighborhood X of $\bar{x}$ on which f has a representation

$$f = g - h \text{ on } X \text{ with } g \text{ convex, } h \text{ quadratic convex.} \tag{4.5}$$

(d) For every $\bar{x} \in R^n$ there is a neighborhood X of $\bar{x}$ and a representation of f as in (1.6) with S a compact topological space, $F(x,s)$ quadratic in x and continuous in s.

Proof.

(a) $\Rightarrow$ (c). Choose any $\bar{x}$ and consider on some neighborhood X of $\bar{x}$ a representation (1.6) of f as in the definition of f being lower-C^2: $F(x,s)$ has second partial derivatives in x, and these are continuous with respect to (x,s). Shrink X if necessary so that it becomes a compact convex neighborhood of $\bar{x}$. The Hessian matrix $\nabla_x^2 F(x,s)$ depends continuously on (x,s) ranging over a compact set $X \times S$, so we have

$$\min_{\substack{(x,s) \in X \times S \\ |v| = 1}} v \cdot \nabla_x^2 F(x,s) v > -\infty .$$

Denote this minimum by $-\rho$ and let

$$G(x,s) = F(x,s) + (\rho/2)|x|^2 . \tag{4.6}$$

Then

$$v \cdot \nabla_x^2 G(x,s) v = v \cdot [\nabla^2 F(x,s) + \rho I] v \geq 0 \tag{4.7}$$

for all $(x,s) \in X \times S$ when $|v| = 1$ and hence also in fact for all $v \in R^n$, because both sides of (4.7) are homogeneous of degree 2 with respect to v. Thus $\nabla_x^2 G(x,s)$ is a positive semidefinite matrix for each $(x,s) \in X \times S$, and $G(x,s)$ is therefore a convex function of $x \in X$ for each $s \in S$. The function

$$g(x) = \max_{s \in S} G(x,s)$$

is accordingly convex, and we have from (4.6) and (1.6) that (4.5) holds for this and $h(x) = (\rho/2)|x|^2$.

(c) $\Rightarrow$ (d). Given a representation as in (c), we can translate it into one as in (d) simply by plugging in a representation of g of the type described in Theorem 5(c).

(d) $\Rightarrow$ (a). Any representation of type (d) is a special case of the kind of representation in the definition of f being lower-C^2 (in fact lower-C^∞); if a quadratic function of x depends

continuously on s, so must all its coefficients in any expansion as a polynomial of degree 2.

(c) $\Rightarrow$ (b). Starting from (4.5) we argue that $\partial f(x) = \partial g(x) - \partial h(x)$ (cf. Clarke 1980, §3, and Rockafellar 1979, p.345), where ∂g happens to be monotone (Theorem 5) and ∂h is actually a linear transformation: $y \in \partial f(x)$ if and only if $y = Ax$, where A is symmetric and positive semidefinite. For $y' \in \partial f(x')$, $y'' \in \partial f(x'')$, we have $y' + Ax' \in \partial g(x')$ and $y'' + Ax'' \in \partial g(x'')$, so from the monotonicity of ∂g it follows that

$$\begin{aligned} 0 &\leq (x' - x'') \cdot ([y' + Ax'] - [y'' + Ax'']) \\ &= (x' - x'') \cdot (y' - y'') + (x' - x'') \cdot A(x' - x'') \quad . \end{aligned} \tag{4.8}$$

Choosing $\rho > 0$ large enough that

$$v \cdot Av \leq \rho |v|^2 \qquad \text{for all } v \in R^n$$

we obtain from (4.8) that

$$\begin{aligned} (x'' - x') \cdot (y'' - y') \geq \rho |x'' - x'|^2 \quad \text{when} \quad & x' \in X, \ x'' \in X, \\ & y' \in \partial f(x'), \\ & y'' \in \partial f(x'') . \end{aligned} \tag{4.9}$$

Certainly (4.4) holds then for $x = \bar{x}$, and since $\bar{x}$ was an arbitrary point of R^n we conclude that ∂f is hypomonotone.

(b) $\Rightarrow$ (c). We are assuming (4.4), so for any $\bar{x}$ we know we can find a convex neighborhood X of $\bar{x}$ and a $\rho > 0$ such that (4.9) holds. Let $g(x) = f(x) + (\rho/2)|x|^2$, so that $\partial g = \partial f + \rho I$ (cf. Clarke 1980, §3, and Rockafellar 1979, p. 345). Then by (4.9), ∂g is monotone on X, and it follows that g is convex on X (cf. Theorem 5; the argument in Theorem 5 is in terms of functions on all of R^n, but it is easily relativized to convex subsets of R^n). Thus (4.5) holds for this g and $h(x) = (\rho/2)|x|^2$. □

COROLLARY 6. *If a function* $f:R^n \to R$ *is lower-*C^2*, it is actually lower-*C^∞*. Thus for* $2 \leq k \leq \infty$ *the classes of lower-*C^k *functions all coincide.*

Proof. As noted in the proof that (d) $\Rightarrow$ (a), any representation of the kind in (d) actually fits the definition of f being lower-C^∞.

COROLLARY 7. *Let* $f:R^n \to R$ *be lower-*C^2*. Then at almost every* $x \in R^n$, f *is twice-differentiable in the sense that there is a quadratic function* q *for which one has*

$$f(x') = q(x') + o(|x' - x|^2) \quad .$$

Proof. This is a classical property of convex functions (cf. Alexandroff 1939), and it carries over to general lower-C^2 functions via the representation in (c).

Counterexample

Since the lower-C^k functions are all the same for $k \geq 2$, it might be wondered if the lower-C^1 functions are really any different either. But here is an example of a lower-C^1 function that is not lower-C^2. Let $f(x) = -|x|^{3/2}$ on R. Then f is of class C^1, hence in particular a lower-C^2, and there would exist by characterization (d) in Theorem 6 numbers α, β, γ, such that

$$f(x) \geq \alpha + \beta x + \gamma x^2 \quad \text{for all } x \text{ near } 0, \text{ with equality when } x = 0 .$$

Then $\alpha = f(0) = 0$ and $-|x|^{3/2} \geq \beta x + \gamma x^2$, from which it follows on dividing by $|x|$ and taking the limits $x \downarrow 0$ and $x \uparrow 0$ that $\beta = 0$. Thus γ would have to be such that $-|x|^{3/2} \geq \gamma |x|^2$ for all x sufficiently near 0, and this is impossible. Therefore f is not lower-C^2.

REFERENCES

Alexandroff, A.D. 1939. Almost everywhere existence of the second differential of a convex function and some properties of convex surfaces connected with it. Leningrad State Univ. Ann., Math. Ser. 6:3-35 (Russian).

Clarke, F.H. 1975. Generalized gradients and applications. Trans. Amer. Math. Soc. 205:247-262.

Clarke, F.H. 1980. Generalized gradients of Lipschitz continuous functionals. Advances in Math.

Danskin, J.M. 1967. The Theory of Max-Min and its Applications to Weapons Allocations Problems. Springer-Verlag. New York.

Lemarechal, C. 1975. An extension of Davidon methods to non-differentiable problems. Math. Programming Study 3: Non-differentiable Optimization, 95-109. North Holland.

Mifflin, R. 1977a An algorithm for constrained optimization with semismooth functions. Math. of Op. Research 2: 191-207.

Mifflin, R. 1977b Semismooth and semiconvex functions in constrained optimization. SIAM J. Control Opt. 15: 959-972.

Pshenichnyi, B.N. 1971. Necessary Conditions for an Extremum. Marcel Dekker. New York.

Rockafellar, R.T. 1970. Convex Analysis. Princeton Univ. Press.

Rockafellar, R.T. 1977. Directionally Lipschitzian functions and subdifferential calculus. Proc. London Math. Soc. 39: 331-355.

Saks, S. 1937. Theory of the Integral. Hafner. New York.

Spingarn, J.E. (forthcoming) Submonotone subdifferentials of Lipschitz functions. Trans.Amer.Math.Soc.

Stein, E.M. 1970. Singular Integrals and Differentiability Properties of Functions. Princeton Univ. Press.

Wolfe, P. 1975. A method of conjugate gradients for minimizing nondifferentiable functions. Math. Programming Study 3: Nondifferentiable Optimization, 145-173. North Holland.

NONDIFFERENTIABLE FUNCTIONS IN HIERARCHICAL CONTROL PROBLEMS

Andrzej Ruszczyński
Institute of Automatic Control
Technical University of Warsaw
Poland

1. INTRODUCTION

The aim of this paper is to highlight the importance of the theory of nondifferentiable optimization and of equations with nondifferentiable operators in a modern branch of control theory - the theory of *hierarchical control systems*.

The hierarchical approach usually applied to large-scale decision problems is based on the partial decentralization of the decision-making process. The typical decision problem involves an object and a decision maker. In the classical (centralized) approach the decision maker observes the object and chooses the decision variables according to certain preferences. Each such operation is connected with the solution of an optimization problem. In practice, however, we often have to deal with large-scale objects which are systems composed of several interconnected subsystems. In this situation the centralized approach, though mathematically correct, has several drawbacks due to the need for large-scale information processing, data transmission, etc. Therefore we aim at organizing the decision-making process in such a way that the decision variables related to a definite subsystem are chosen on the basis of information relevant for this subsystem. Thus, a decision maker is associated with each

subsystem and takes part in the decision-making process; each of them observes his own subsystem and chooses his decision variables by solving an optimization subproblem related to this subsystem. It is clear that a simple decentralization of this type cannot be considered effective if there are interactions between the subsystems, and for this reason a supreme decision maker (called the *coordinator*) is introduced. The aim of the coordinator is to influence the lesser decision makers so as to make their decisions consistent with the global objective. This is achieved through the introduction (by the coordinator) of certain parameters (*coordination variables*) into the lower-level problems and the modification of the set of preferences (i.e., objectives and/or constraints) used by the other decision makers. It should be stressed that the coordinator has no direct access to the system and that he may not override the decisions of the other decision makers. It is also assumed that he receives aggregated information about the effect of his decisions. The decision-making structure outlined above, which comprises the coordinator and several lower-order decision makers, is called the hierarchical control system.

At this point it should be said that the above approach does not necessarily lead to strictly optimal decisions, and this is the main reason why it is difficult to convince a mathematician of the value of hierarchical control systems. The advantages of the hierarchical approach lie beyond the mathematics and are connected with adaptability, information privacy, data transmission and other preferences which cannot be formalized and which, obviously, will not be discussed in this paper (see, however, refs. 2,4, and 7). However, once a definite hierarchical structure has been chosen, interesting and well-defined mathematical problems arise. These problems are connected with the coordinator's task, which is to optimize the performance of the whole system by the appropriate choice of values for the coordination variables. From the outline given above it will be evident to experts in optimization that the function *coordination variables → performance* will be a nonsmooth function, even if the infimal problems involve differentiable functions. This is the main reason for including this paper in a volume concerned with nondifferentiable optimization.

In the following analysis we shall assume that the subsystems being controlled are described by the following equations:

$$y_i = f_i(c_i,u_i,z_i) \quad , \qquad i \in \overline{(1,N)} \tag{1.1}$$

where i is an index representing a particular subsystem, y_i denotes the output influencing other subsystems, c_i denotes the local decision variables (control), u_i denotes the input originating from other subsystems, and z_i denotes disturbances. We assume that $y_i \in Y_i \subset \underset{\sim}{Y}_i$, $c_i \in C_i \subset \underset{\sim}{C}_i$, $u_i \in U_i \subset \underset{\sim}{U}_i$, $z_i \in Z_i$, f_i: $C_i \times U_i \times Z_i \to Y_i$, and $\underset{\sim}{Y}_i$, $\underset{\sim}{C}_i$, $\underset{\sim}{U}_i$ are finite dimensional spaces. The interconnections between the subsystems are described by the following equations:

$$u_i = \sum_{j=1}^{N} H_{ij}\, y_j \tag{1.2}$$

where the H_{ij} are linear operators (usually given by 0-1 matrices). For brevity we use the following notation: $y = (y_1,\ldots,y_N)$, $Y = \underset{i=1}{\overset{N}{X}} Y_i \subset \underset{\sim}{Y} = \underset{i=1}{\overset{N}{X}} \underset{\sim}{Y}_i$; $c = (c_1,\ldots,c_N)$, $C = \underset{i=1}{\overset{N}{X}} C_i \subset \underset{\sim}{C} = \underset{i=1}{\overset{N}{X}} \underset{\sim}{C}_i$; $u = (u_1,\ldots,u_N)$, $U = \underset{i=1}{\overset{N}{X}} U_i \subset \underset{\sim}{U} = \underset{i=1}{\overset{N}{X}} \underset{\sim}{U}_i$; $z = (z_1,\ldots,z_N)$, $Z = \underset{i=1}{\overset{N}{X}} Z_i$. Using this notation we can write equations (1.1), (1.2) in the compact form

$$y = f(c,u,z) \quad , \tag{1.3}$$

$$u = Hy \ (u_i = H_i y) \quad , \tag{1.4}$$

where f: $C \times U \times Z \to Y$, H: $\underset{\sim}{Y} \to \underset{\sim}{U}$, H_i: $\underset{\sim}{Y} \to \underset{\sim}{U}_i$. It is assumed that for any $c \in C$ and $z \in Z$ the set of equations (1.3), (1.4) defines unique interactions $u(c,z) \in U$, $u(c,z) = (u_1(c,z),\ldots,u_N(c,z))$.

We assume that there is a performance index $q(c,z)$ defined for the system,

$$q(c,z) = \sum_{i=1}^{N} q_i(c_i,u_i,z_i) \tag{1.5}$$

where $u_i = u_i(c,z)$. Also, the controls c and inputs u must satisfy the constraints

$$g_i(c_i,u_i,z_i) \leq 0, \quad i \in \overline{1,N} \tag{1.6}$$

where $g_i : C_i \times U_i \times Z_i \to R^{J_i}$. Here J_i represents the dimensionality of this vector constraint.

In the next two parts of the paper we shall describe two examples of the hierarchical approach to the problem of finding the value $\hat{c}$ which minimizes the function (1.5) subject to (1.3), (1.4), and (1.6). These examples represent the two main ideas on which hierarchical systems are based - the primal approach, which involves a two-stage minimization, and the dual approach, which uses coordination variables (prices) to represent interactions.

However, before proceeding to these problems we should explain the notation used in the following sections. If $\phi : R^n \to R^1$ then $\nabla\phi(x)$ and $\nabla^2\phi(x)$ denote the gradient and the hessian, respectively. If $\phi : R^n \to R^m$, $\phi(x) = (\phi_1(x),\ldots,\phi_m(x))$ then $\nabla\phi(x)$ denotes the matrix with columns $\nabla\phi_i(x)$ $(i = 1,\ldots,m)$, and $\phi_x(x)$ denotes the derivative, considered to be a linear operator. For a linear operator (matrix) A we use A^* to denote the operator (matrix) conjugate to A. Finally, a function $\phi : R^n \to R^m$ is described as convex if all functions ϕ_i $(i = 1,\ldots,m)$ are convex.

2. THE PRIMAL APPROACH

2.1 The decentralized control system

In this section we assume that the disturbance $z = (z_1,\ldots,z_N)$ which influences the system is a random variable. We also assume that it is possible to measure z. The optimal decision rule is then simple: observe z and choose the value of c which minimizes q(c,z) (defined by (1.5)) subject to the constraints (1.3), (1.4), and (1.6). However, this approach may not be satisfactory for some particular reason and so we are going to develop a decision rule $z \to c$, which can be split into N rules of the form $z_i \to c_i$. To achieve this we shall adopt the *primal approach* derived from large-scale optimization theory (see, for example, ref.5), and analyzed in the context of

control in refs. 2-4. Let us assume that the desired values of interactions $y = (y_1,\ldots,y_N)$, $u = Hy = (u_1,\ldots,u_N)$ are fixed. Under this assumption we formulate N independent problems for *local decision makers* as follows: (DM_i) *observe* z_i *and choose the value* $\hat{c}_i$ *which solves the problem*

$$\min q_i(c_i,u_i,z_i)$$

subject to

$$f_i(c_i,u_i,z_i) - y_i = 0, \tag{2.1}$$

and

$$g_i(c_i,u_i,z_i) \leq 0. \tag{2.2}$$

Note that u_i and y_i are fixed parameters for DM_i. Let $C_i(u_i,y_i,z_i)$ denote the feasible set for DM_i, defined by (2.1) and (2.2), and let $\hat{c}_i(u_i,y_i,z_i)$ be the solution of DM_i.

Let $C(u,y,z) = \sum_{i=1}^{N} C_i(u_i,y_i,z_i)$, $\hat{c}(u,y,z) = (\hat{c}_1(u_1,y_1,z_1),\ldots,\hat{c}_N(u_N,y_N,z_N))$. We define the set

$$Y_0 = \{y\in Y:\ P\{C(Hy,y,z) \neq \emptyset\} = 1\}, \tag{2.3}$$

where P denotes the probability. We assume that $Y_0 \neq \emptyset$, and that for any $y\in Y_0$, $u = Hy$ and for almost all z with respect to measure P, the local problems DM_i have solutions. Thus, for any $y\in Y_0$, the set of problems DM_i ($i\in\overline{1,N}$) defines a mapping $z \to \hat{c}(Hy,y,z)$, composed of local mappings $z_i \to \hat{c}_i(H_iy,y_i,z_i)$. Thus, any $y\in Y_0$ defines a *decentralized control system* of the type we were trying to achieve. It remains only to make the performance of this control system as good as possible by appropriate adjustment of the desired values of interactions y (and $u = Hy$). This is the *coordination task* and we shall focus our attention on this problem.

Let $\varphi : Y_0 \times Z \to R^1$ be defined by

$$\varphi(y,z) = q(\hat{c}(Hy,y,z),z)) \tag{2.4}$$

and let

$$\Phi(y) = E\{\phi(y,z)\} \tag{2.5}$$

where E denotes the mathematical expectation. The coordination task is to solve the following optimization problem:

$$\min_{y \in Y_0} \Phi(y), \tag{2.6}$$

which, when written explicitly, takes the form

$$\min_{y \in Y_0} E\left\{ \sum_{i=1}^{N} \min_{c_i \in C_i(H_i y, Y_i, z_i)} q_i(c_i, H_i y, Y_i, z_i) \right\}. \tag{2.7}$$

This is a large-scale *two-stage stochastic programming problem* [1] (see also the paper by Yu. M. Ermoliev in this volume). The solution of problems of this type presents certain serious difficulties. Firstly, in order to obtain the value of Φ at a given point y it is necessary to compute the integral $\Phi(y) = \int \phi(y,z)\, P(dz)$. The analytical calculation of this integral is impossible (except in some trivial cases). In addition, the numerical computation of the integral appears to be difficult and time-consuming. Secondly, even in very simple cases, the function $y \to \phi(y,z)$ is nonsmooth because its values are obtained from the optimization problems dependent on the parameter y. Thus the function Φ is generally nondifferentiable. Thirdly, the feasible set Y_0 is defined indirectly and so it is difficult to verify whether a given y belongs to Y_0.

With a view to the above-mentioned difficulties we have developed a special algorithm which is able to tackle problem (2.7). This algorithm is based on the *stochastic subgradient method* due to Yu. M. Ermoliev and others.

2.2 Properties of the coordination problem

In this section we shall investigate the essential characteristics of problem (2.6). We shall make the following assumptions.

1. There exists an open bounded set $\Omega \subset Y_0$ such that the solution of the problem (2.6) belongs to Ω.

2. The set $\hat{c}(H\Omega,\Omega,Z)$ is bounded.

3. The functions $(c,u) \to f(c,u,z)$, $(c,u) \to g(c,u,z)$, $(c,u) \to q(c,u,z)$ are continuously differentiable for almost all z and all $y \in \Omega$, $u = Hy$, $c = \hat{c}(u,y,z)$, and the derivatives are uniformly bounded.

4. The functions $z \to (f(c,u,z), g(c,u,z), q(c,u,z))$ are integrable.

5. For almost all z and all $y \in \Omega$ the solutions of the problems DM_i satisfy necessary optimality conditions. It is possible to choose Lagrange multipliers uniformly bounded on $\Omega \times Z$.

6. One of the following conditions is satisfied:
   (a) The functions $(c_i,u_i) \to f_i(c_i,u_i,z_i)$ are linear and the functions $(c_i,u_i) \to g_i(c_i,u_i,z_i)$, $(c_i,u_i) \to q_i(c_i,u_i,z_i)$ are convex for almost all z.
   (b) The function $y \to \hat{c}(Hy,y,z)$ is uniformly Lipschitz continuous on Ω for $z \in Z$.

For brevity we use the expressions "almost all" and "integrable" for "P-almost all" and "P-integrable" and we write "Z" instead of "$Z \setminus Z_0$, where Z_0 is a set of null measure".

We shall now prove that the function φ is *weakly convex* and we shall develop formulae for the calculation of its *subgradients* (see ref. 8 and the paper by E.A. Nurminski in this volume for definitions of a weakly convex function and a subgradient).

Lemma 1. The function $y \to \varphi(y,z)$ is weakly convex on Ω for almost all z.

Proof. Let us define the Lagrange function for DM_i:

$$L_i(c_i,u_i,y_i,z_i,\lambda_i,\mu_i) = q_i(c_i,u_i,z_i) + \langle\lambda_i, f_i(c_i,u_i,z_i)-y_i\rangle + \langle\mu_i, g_i(c_i,u_i,z_i)\rangle.$$

We denote by $\hat{\lambda}_i(u_i,y_i,z_i)$, $\hat{\mu}_i(u_i,y_i,z_i)$, any multipliers corresponding to the solution of DM_i. Let

$$L(c,u,y,z,\lambda,\mu) = \sum_{i=1}^{N} L_i(c_i,u_i,y_i,z_i,\lambda_i,\mu_i)$$

be the global Lagrange function and let $\hat{\lambda}(u,y,z) = (\hat{\lambda}_1(u_1,y_1,z_1), \ldots, \hat{\lambda}_N(u_N,y_N,z_N))$ and $\hat{\mu}(u,y,z) = (\hat{\mu}_1(u_1,y_1,z_1), \ldots, \hat{\mu}_N(u_N,y_N,z_N))$.

Then for any $y \in \Omega$, $u = Hy$, and for almost all z the following equality holds:

$$\phi(y,z) = L(\hat{c}(u,y,z),\ u,y,z,\hat{\lambda}(u,y,z),\ \hat{\mu}(u,y,z)).$$

Let $\bar{y} \in \Omega$ and $\bar{u} = H\bar{y}$. For brevity we use the notation:

$$c = \hat{c}(u,y,z),\ \bar{c} = \hat{c}(\bar{u},\bar{y},z),\ \lambda = \hat{\lambda}(u,y,z),\ \bar{\lambda} = \hat{\lambda}(\bar{u},\bar{y},z),$$

$$\mu = \hat{\mu}(u,y,z),\ \bar{\mu} = \hat{\mu}(\bar{u},\bar{y},z).$$

We then have

$$\phi(\bar{y},z) - \phi(y,z) = L(\bar{c},\bar{u},\bar{y},z,\bar{\lambda},\ \bar{\mu}) - L(c,u,y,z,\lambda,\mu) \geq$$

$$\geq L(\bar{c},\bar{u},\bar{y},z,\lambda,\mu) - L(c,u,y,z,\lambda,\mu) \qquad (2.8)$$

By virtue of condition 3 the function $(c,u,y) \to L(c,u,y,z,\lambda,\mu)$ is continuously differentiable for almost all z. Thus

$$\begin{aligned}\phi(\bar{y},z) - \phi(y,z) \geq\ & \langle \nabla_c L(c,u,y,z,\lambda,\mu),\ \bar{c} - c\rangle + \\ & + \langle \nabla_u L(c,u,y,z,\lambda,\mu),\ \bar{u} - u\rangle + \\ & + \langle \nabla_y L(c,u,y,z,\lambda,\mu),\ \bar{y} - y\rangle + \\ & + r_L(c,u,y,\bar{c},\bar{u},\bar{y},z,\lambda,\mu)\end{aligned} \qquad (2.9)$$

where the residual term r_L is small with respect to $\|(\bar{c},\bar{u},\bar{y}) - (c,u,y)\|$. Note that it follows from the necessary optimality conditions for DM_i that $\nabla_c L(c,u,y,z,\lambda,\mu) = 0$. Next

$$\langle \nabla_u L(c,u,y,z,\lambda,\mu),\ \bar{u} - u\rangle = \langle H^* \nabla_u L(c,u,y,z,\lambda,\mu),\ \bar{y} - y\rangle.$$

If condition 6a holds then the function $(c,u,y) \to L(c,u,y,z,\lambda,\mu)$ is convex and $r_L \geq 0$. On the other hand, if condition 6b holds then r_L is small with respect to $\|\bar{y} - y\|$. In both cases, for almost all z

$$\phi(\bar{y},z) - \phi(y,z) \geq \langle \xi, \bar{y} - y \rangle + r_z(\bar{y},y), \tag{2.10}$$

where

$$\xi = H^*\nabla_u L(c,u,y,z,\lambda,\mu) + \nabla_y L(c,u,y,z,\lambda,\mu), \tag{2.11}$$

and the residual term $r_z(\bar{y},y)$ satisfies the condition of uniform smallness with respect to $\|\bar{y} - y\|$ on compact subsets of Ω. Finally, the function $y \to \phi(y,z)$ is continuous for almost all z. Consequently, taking into account (2.10) and (2.11), it is weakly convex, and the lemma has been proved.

Corollary. For any $y \in \Omega$ and for almost all z the vector defined by (2.11) is a subgradient of the function $y \to \phi(y,z)$ at the point y.

Lemma 2. The function Φ is weakly convex on Ω. For any $y \in \Omega$ the vector

$$d = E\{H^*\nabla_u L(c,u,y,z,\lambda,\mu) + \nabla_y L(c,u,y,z,\lambda,\mu)\} \tag{2.12}$$

with $c = \hat{c}(Hy,y,z)$, $u = Hy$, $\lambda = \hat{\lambda}(Hy,y,z)$, $\mu = \hat{\mu}(Hy,y,z)$ is a subgradient of Φ at the point y, i.e., $d \in \partial\Phi(y)$.

Proof. Observe that by virtue of assumptions 3 and 4 the function $z \to \phi(y,z)$ is integrable. Hence the function Φ is well-defined and continuous on Ω.

Now let us consider the inequalities (2.8), (2.9), and (2.10). In practice, by virtue of condition 5 we can always assume that the functions $z \to \hat{\lambda}(u,y,z)$, $z \to \hat{\mu}(u,y,z)$, $z \to \hat{c}(u,y,z)$ are measurable (this is true if we choose solutions and multipliers according to a certain ordering rule, e.g., solutions and multipliers of minimum norm, if they are not unique). Thus the functions $z \to \nabla_c L(\hat{c}(u,y,z),u,y,z, \hat{\lambda}(u,y,z), \hat{\mu}(u,y,z))$, $z \to \nabla_u L(\hat{c}(u,y,z), u,y,z, \hat{\lambda}(u,y,z), \hat{\mu}(u,y,z))$, and $z \to \nabla_y L(\hat{c}(u,y,z), u,y,z, \hat{\lambda}(u,y,z), \hat{\mu}(u,y,z))$ are integrable. The integrability of these gradients follows from assumptions 3 and 4 and the finite dimensionality of the c-,u- and y-spaces. Consequently the vector d is well-defined by (2.12). If condition 6a holds, then $r_z(\bar{y},y) = 0$ and the proposition of the lemma follows immediately from (2.10). If condition 6b holds, then it follows from the above considerations that the function $z \to r_L(\hat{c}(Hy,y,z), Hy,y,\hat{c}(H\bar{y},\bar{y},z), H\bar{y},\bar{y},z, \hat{\lambda}(Hy,y,z), \hat{\mu}(Hy,y,z))$ is integrable, and

the function $(y,\bar{y}) \to r_L(\hat{c}(Hy,y,z),\ Hy,y,\ \hat{c}(H\bar{y},\bar{y},z),\ H\bar{y},\bar{y},\ \hat{\lambda}(Hy,y,z),\ \hat{\mu}(Hy,y,z))$ is uniformly small with respect to $\|\bar{y}-y\|$ for $z\in Z$. Hence the function $r_z(\bar{y},y)$ in (2.10) is integrable and $E\{r_z(\bar{y},y)\}$ is small with respect to $\|\bar{y}-y\|$. The proposition of the lemma follows from (2.10).

2.3 The algorithm

In this section we shall assume that the set Ω satisfying condition 1 is known and that there is a point $y^0\in\Omega$ such that $\Phi(y) \geq \Phi(y^0) + \gamma$ for all $y\in Y_0\setminus\Omega$ and for some $\gamma > 0$.

Let π_i ($i\in\overline{1,N}$) be binary random variables associated with the problems DM_i such that $p_i = P\{\pi_i = 1\} > 0$. Let $\pi = (\pi_1,\ldots,\pi_N)$.

We shall now develop a two-level stochastic algorithm for the solution of problem (2.6). The algorithm operates in conjunction with two random number generators, which produce sequences $\{\pi^k\}$ and $\{z^k\}$ from a series of mutually independent observations over π and z. Based on these results, the algorithm constructs a random sequence $\{y^k\}$, which is supposed to converge to the solution of problem (2.6). The k-th iteration of the algorithm consists of the following operations:

1. Draw a value $\pi^k = (\pi_1^k,\ldots,\pi_N^k)$, independent of the previous draws π^j, z^j $(j < k)$.
2. For a given y^k determine the parameters $y_i^k, u_i^k = H_i y^k$ for the problems DM_i.
3. Draw a value $z^k = (z_1^k,\ldots,z_N^k)$, independent of all π^j $(j \leq k)$, z^j $(j < k)$.
4. Solve those problems DM_i for which $\pi_i^k = 1$, and define

$$c_i^k = \hat{c}_i(u_i^k,\ y_i^k,\ z_i^k),$$

$$\lambda_i^k = \hat{\lambda}_i(u_i^k,\ y_i^k,\ z_i^k),$$

$$\mu_i^k = \hat{\mu}_i(u_i^k,\ y_i^k,\ z_i^k),$$

$$\alpha_i^k = \nabla_{u_i} q_i(c_i^k,\ u_i^k,\ z_i^k) + \nabla_{u_i} f_i(c_i^k,\ u_i^k,\ z_i^k)\,\lambda_i^k + \nabla_{u_i} g_i(c_i^k,\ u_i^k,\ z_i^k)\,\mu_i^k.$$

5. Compute the vectors

$$\alpha^k = (\pi_1^k \alpha_1^k / p_1, \ldots, \pi_N^k \alpha_N^k / p_N),$$

$$\beta^k = (\pi_1^k \lambda_1^k / p_1, \ldots, \pi_N^k \lambda_N^k / p_N),$$

$$\xi^k = H^* \alpha^k - \beta^k \quad .$$

6. Compute

$$y^{k+1} = \begin{cases} y^k - \rho_k \xi^k, & \text{if } y^k - \rho_k \xi^k \in \Omega, \\ y^0, & \text{otherwise,} \end{cases} \tag{2.13}$$

where $\rho_k \geq 0$, and ρ_k depends only on $(y^0, \ldots, y^k)$.

Theorem 1. Let the probability that the following conditions are satisfied be 1:

$$\rho_{k+1}/\rho_k \to 1, \quad \sum_{k=0}^{\infty} \rho_k = \infty \quad . \tag{2.14a}$$

Additionally, let

$$\sum_{k=0}^{\infty} E\rho_k^2 < \infty \quad . \tag{2.14b}$$

Then all accumulation points of the sequence $\{y^k\}$ are stationary (i.e., belong to the set $Y^* = \{y \in \Omega : 0 \in \partial\Phi(y)\}$) with probability 1, and the sequence $\{\Phi(y^k)\}$ converges.

Proof. Observe that

$$E\{\xi^k \mid y^0, \ldots, y^k, z^k\} = H^* \nabla_u L(c^k, u^k, y^k, z^k, \lambda^k, \mu^k) + \nabla_y L(c^k, u^k, y^k, z^k, \lambda^k, \mu^k)$$

with $c^k = \hat{c}(Hy^k, y^k, z^k)$, $u^k = Hy^k$, $\lambda^k = \hat{\lambda}(Hy^k, y^k, z^k)$, $\mu^k = \hat{\mu}(Hy^k, y^k, z^k)$. Thus, by virtue of Lemma 2,

$$E\{\xi^k \mid y^0, \ldots, y^k\} = E\{E\{\xi^k \mid y^0, \ldots, y^k, z^k\} \mid y^0, \ldots, y^k\} \in \partial\Phi(y^k). \tag{2.15}$$

Convergence of the sequence $\{y^k\}$ satisfying (2.13)-(2.15) to the set Y^* follows from results obtained by Nurminski and Ermoliev (see ref. 8 and ref. 1, Chapter IV, Theorem 5).

The above theorem is of great practical importance in computation. Observe that for a given sequence of draws $\{z^k\}$, $\{\pi^k\}$ the algorithm generates a path $\{y^k(\omega)\}$ for the stochastic process $\{y^k\}$ (ω denotes the event corresponding to one run of the algorithm). It follows from Theorem 1 that almost all of the paths taken by this process converge to the set Y^*. Therefore, for any practical purposes, one can consider only a single trajectory of the process $\{y^k\}$. It is worth noting that neither the values nor the subgradients of Φ are necessary to generate the path $\{y^k(\omega)\}$.

Finally, let us discuss briefly the role played by the binary variable π^k. If $\pi^k = (1,\ldots,1)$ for all $k \geq 0$, then we have to solve all lower-level problems at each iteration of the algorithm. This can be costly if we are dealing with a large problem involving many subproblems DM_i. The random variable π has therefore been introduced in order to overcome this difficulty. At the k-th iteration of the algorithm we solve only those problems DM_i for which $\pi_i^k = 1$. In particular, if π^k always has only one nonzero component then only one subproblem is solved at each iteration of the algorithm.

2.4 The penalty method

It was assumed in the previous section that the set Ω which satisfies condition 1 is known. This assumption seems to be rather restrictive and we shall try to avoid the need to invoke it. A serious difficulty arises at this point; we cannot guarantee that all points y^k belong to the set Y_0. In order to overcome this difficulty we shall introduce artificial variables into the problems DM_i and formulate modified lower-level problems as follows:

($\widetilde{DM}_i$) *observe* z_i *and choose* c_i, v_i, w_i, σ_i *which solve the problem*

$$\min\ [q^i(c_i,u_i,z_i) + R\sigma_i]$$

subject to

$$f_i(c_i,u_i,z_i) - y_i + v_i = 0\ , \tag{2.16}$$

$$g_i(c_i,u_i,z_i) + w_i \le 0, \tag{2.17}$$

$$(v_i,w_i,\sigma_i)\in\Gamma_i = \{(v_i,w_i,\sigma_i):$$

$$\|v_i\|^2 + \|w_i\|^2 \le \sigma_i^2,\ \sigma_i \ge 0\}. \tag{2.18}$$

The variables v_i,w_i are the artificial variables, and the constant $R > 0$ is a penalty coefficient.

Let $\tilde{c}_{Ri}(u_i,y_i,z_i)$, $\tilde{v}_{Ri}(u_i,y_i,z_i)$, $\tilde{w}_{Ri}(u_i,y_i,z_i)$, and $\tilde{\sigma}_{Ri}(u_i,y_i,z_i)$ be the solutions of DM_i. As before, let $\tilde{c}_R(u,y,z) = (\tilde{c}_{R1},\ldots,\tilde{c}_{RN})$, $\tilde{v}_R(u,y,z) = (\tilde{v}_{R1},\ldots,\tilde{v}_{RN})$, $\tilde{w}(u,y,z) = (\tilde{w}_{R1},\ldots,\tilde{w}_{RN})$, $\tilde{\sigma}(u,y,z) = (\tilde{\sigma}_{R1},\ldots,\tilde{\sigma}_{RN})$. Let

$$\tilde{\phi}_R(y,z) = q(\tilde{c}_R(Hy,y,z),z) + R\,\|\tilde{\sigma}_R(Hy,y,z)\|_1,$$

where

$$\|\sigma\|_1 = \sum_{i=1}^{\infty} |\sigma_i|,$$

and let

$$\tilde{\Phi}_R(y) = E\tilde{\phi}_R(y,z).$$

Let us consider the problem

$$\min\ \tilde{\Phi}_R(y). \tag{2.19}$$

We assume that we know an open bounded set $\tilde{\Omega}$ such that for R sufficiently large the solution of (2.19) belongs to $\tilde{\Omega}$. We also assume that conditions 2-6 from Section 2.2 are satisfied for DM_i with Ω replaced by $\tilde{\Omega}$ and $\hat{c}$ replaced by $\tilde{c}_R$. Note that we no longer assume that $\tilde{\Omega}\subset Y_0$.

Using a method similar to that employed in Section 2.2, it may be shown that the function $\tilde{\phi}_R$ is weakly convex on $\tilde{\Omega}$. Moreover, we can solve the problem (2.19) using a modified version of the algorithm described in Section 2.3; at each iteration of the algorithm we solve the problems $\tilde{DM}_i$ instead of DM_i, and we use multipliers corresponding to (2.16) and (2.17) instead of multipliers corresponding to (2.1) and (2.2). This algorithm will generate a sequence $\{y^k\}$ which converges with probability one to the set of stationary points of problem (2.19). However, even for a very large penalty coefficient R the solutions of (2.19) are not necessarily the solutions of (2.6), although the exact penalty approach is used. This is due to the stochastic nature of the problem; the operation of averaging may smooth out the penalty function. Thus we have to study the relations between the solutions of (2.6) and the solutions of (2.19). Let $\hat{y}\in\Omega$ be a solution of (2.6) and let $\tilde{y}_R\in\tilde{\Omega}$ be a solution of (2.19) for a sufficiently large value of R.

Lemma 3. Let the functions $(c_i,z_i) \to q_i(c_i,u_i,z_i)$ be uniformly bounded from below for $u\in H\tilde{\Omega}$. Then for sufficiently large values of R

$$E \|\tilde{\sigma}_R(H\tilde{y}_R,\tilde{y}_R,z)\|_1 \leq \text{const}/R. \tag{2.20}$$

Proof. Note that $\tilde{\phi}_R(y) \leq \phi(y)$ for $y\in\tilde{\Omega}\cap Y_0$. Thus

$$\tilde{\phi}_R(\tilde{y}_R) \leq \tilde{\phi}_R(\hat{y}) \leq \phi(\hat{y}) \tag{2.21}$$

On the other hand

$$\tilde{\phi}_R(\tilde{y}_R) = R\,E\|\tilde{\sigma}_R(H\tilde{y}_R,\tilde{y}_R,z)\|_1 + E\,q(\tilde{c}_R(H\tilde{y}_R,\tilde{y}_R,z),z) \geq$$

$$\geq R\,E\|\tilde{\sigma}_R(H\tilde{y}_R,\tilde{y}_R,z)\|_1 + \text{const}. \tag{2.22}$$

The inequality (2.20) follows from (2.21) and (2.22), and the lemma has been proved.

Corollary. For any $\varepsilon > 0$

$$P\{\|\tilde{\sigma}_R(H\tilde{y}_R,\tilde{y}_R,z)\|_1 \geq \varepsilon\} \leq \text{const}/\varepsilon R.$$

Now let us consider the sequence $\{y^s\}$ such that $y^s = \tilde{y}_{R_s}$, $R_s \to \infty$.

Theorem 2. Any accumulation point of the sequence $\{y^s\}$ is a solution of problem (2.6).

Proof. Let $u^s = Hy^s$, $c^s(z) = \tilde{c}_{R_s}(u^s,y^s,z) = (c_1^s(z),\ldots,c_N^s(z))$, $\sigma^s(z) = \tilde{\sigma}_{R_s}(u^s,y^s,z) = (\sigma_1^s(z),\ldots,\sigma_N^s(z))$.

Let y^* be a limit point of the sequence $\{y^s\}$. Obviously, we can always assume that $y^s \to y^*$. We shall prove that for almost all z the sequence $\{\sigma^s(z)\}$ converges. It is easy to verify that the point $c_i^s(z_i)$ is a solution of the problem

$$\min\,[q_i(c_i,u_i^s,z_i) + R_s\,T_i(c_i,u_i^s,y_i^s,z_i)],$$

where

$$T_i(c_i,u_i,y_i,z_i) = (\|f_i(c_i,u_i,z_i) - y_i\|^2 + \|g_i(c_i,u_i,z_i)_+\|^2)^{1/2}$$

and $T_i(c_i^s(z_i),u_i^s,y_i^s,z_i) = \sigma_i^s(z_i)$. Thus, for any $k \geq 0$

$$q_i(c_i^k(z_i),u_i^k,z_i) + R_k\sigma_i^k(z_i) \leq q_i(c_i^s(z_i),u_i^k,z_i) +$$

$$+ R_kT_i(c_i^s(z_i),u_i^k,y_i^k,z_i) \leq q_i(c_i^s(z_i),u_i^k,z_i) +$$

$$+ R_k\sigma_i^s(z_i) + R_k(T_i(c_i^s(z_i),u_i^k,y_i^k,z_i) - T_i(c_i^s(z_i),u_i^s,y_i^s,z_i)).$$

Hence

$$\sigma_i^k(z_i) \leq \sigma_i^s(z_i) + \alpha_i^{s,k}(z_i),$$

where $\alpha_i^{s,k}(z_i) \to 0$ as $s,k \to \infty$. This property follows from the convergence of the sequences $\{u_i^s(z_i)\}$, $\{y_i^s(z_i)\}$ and the continuity of

the function $(u_i,y_i) \to T_i(c_i,u_i,y_i,z_i)$. It follows from the last inequality that the sequence $\{\sigma^s(z)\}$ converges with probability one. Consequently, from Lemma 3, $\lim \sigma^s(z) = 0$ for almost all z. Let $c^*(z)$ be any accumulation point of the sequence $\{c^s(z)\}$. Then

$$P\{f(c^*(z),Hy^*,z) - y^* = 0\} = 1 ,$$

$$P\{g(c^*(z),Hy^*,z) \leq 0\} = 1 .$$

Thus $y^* \in Y_0$. On the other hand

$$\Phi(\hat{y}) \geq \tilde{\Phi}_{R_s}(y^s) \geq E\, q(c^s(z)) .$$

Consequently $\Phi(\hat{y}) = \Phi(y^*)$, and the theorem has been proved.

The above theorem shows that use of the penalty method with a sufficiently large penalty coefficient yields good approximations to the solution of problem (2.6). Thus we have overcome the last of the three difficulties mentioned at the end of Section 2.1.

3. THE DUAL APPROACH

3.1 The interaction balance method with feedback

In this section we shall describe a variant of the *dual method* for large-scale optimization problems. For simplicity, we assume that z in (1.3) is constant and that the performance index (1.5) and constraints (1.6) do not depend on z. Thus we have to solve the following optimization problem:

$$\min \sum_{i=1}^{N} q_i(c_i,u_i)$$

subject to

$$u - H\, f(c,u,z) = 0, \tag{3.1}$$

$$g_i(c_i,u_i) \leq 0, \quad i \in \overline{\{1,N\}}. \tag{3.2}$$

If the exact value of the parameter z is known and the functions involved are sufficiently regular, then the solution of the above

problem can be found by well-known large-scale optimization techniques (see, for example, ref. 5). However, in some control and large-scale optimization problems we encounter situations in which these techniques are not applicable. In control problems we do not usually know the exact values of the parameters z; we can only observe the behavior of the controlled system. In some large-scale optimization problems we cannot handle the constraints (3.1), even though we know z, because the functions $(c,u) \to f(c,u,z)$ are non-differentiable. In both cases we have to content ourselves with an approximate solution which must, however, satisfy the constraints (3.1) and (3.2). In this section we develop a hierarchical method which generates suboptimal solutions of this type. The method is based on the *interaction balance principle* suggested in ref. 2 and analyzed in refs. 6,9, and 10.

Let us introduce a coordination variable (price) p such that $p = (p_1,\ldots,p_N)$, $p_i \in U_i$ and let us specify the problems to be solved by the local decision-making units associated with the subsystems as follows:

$$\min L_i(c_i,u_i,p) \tag{3.3}$$

subject to

$$g_i(c_i,u_i) \leq 0 ,$$

where

$$L_i(c_i,u_i,p) = q_i(c_i,u_i) + \langle p_i,u_i\rangle - \sum_{j=1}^{N} \langle p_j,H_{ji}\phi_i(c_i,u_i)\rangle \quad . \tag{3.4}$$

We assume that the functions ϕ_i are continuously differentiable approximations of the functions $(c_i,u_i) \to f_i(c_i,u_i,z_i)$. Let $\bar{c}(p) = (\bar{c}_1(p),\ldots,\bar{c}_N(p))$ and $\bar{u}(p) = (\bar{u}_1(p),\ldots,\bar{u}_N(p))$ be the solutions of problem (3.3). When the control $\bar{c}(p)$ is applied to the real system described by (3.1) it produces interactions $u(\bar{c}(p),z)$. The *coordination task* is to find price $\tilde{p}(z)$ such that

$$\bar{u}(\tilde{p}(z)) - u(\bar{c}(\tilde{p}(z)),z) = 0 . \tag{3.5}$$

This principle is an extension of the dual method of optimization to the class of problems considered in this section. Indeed, observe that if $\phi_i(c_i,u_i) \equiv f_i(c_i,u_i,z)$ then the function

$$L(c,u,p) = \sum_{i=1}^{N} L_i(c_i,u_i,p) = \sum_{i=1}^{N} q_i(c_i,u_i) + \langle p, u - Hf(c,u,z)\rangle$$

is the Lagrange function for the original problem. Note also that condition (3.5) is equivalent to the condition $\bar{u}(p) = Hf(\bar{c}(p),\bar{u}(p),z)$, which is the optimality condition in the dual method of optimization [5,7]. However, if $\phi_i(c_i,u_i) \neq f_i(c_i,u_i,z_i)$ then the solution given by (3.5) is not necessarily optimal, although it still satisfies constraints (3.1) and (3.2). Some bounds on the loss of optimality have been derived in ref. 6, but it is still difficult to convince a mathematician of the validity of the above approach since it is motivated by preferences which cannot be formalized. However, once the coordination problem (3.5) has been formulated, we are faced with a well-defined mathematical problem: find a method for the solution of equation (3.5). This is not a trivial matter since the functions $p \to \bar{c}(p)$, $p \to \bar{u}(p)$ and $c \to u(c,z)$ are generally nondifferentiable.

In the next section we shall investigate in detail the properties of the function $p \to (\bar{c}(p),\bar{u}(p))$. In Section 3.3, we shall construct an iterative algorithm based on the properties of this function which can be used to solve equation (3.5). However, before proceeding to these problems we shall introduce some useful notation and make a number of assumptions.

1. For any $c \in C$ the equation $u = H\phi(c,u)$ defines a unique solution $u_0(c)$. The linear operator $I - H\phi_u(c,u)$ is nonsingular for $c \in C$.

We introduce the following notation: $x = (c,u)$, $x \in X$, $Q(x) = \sum_{i=1}^{N} q_i(c_i,u_i)$, $F(x) = u - H\phi(c,u)$, $D(x) = u - u_0(c)$, $D_*(x,z) = u - u(c,z)$. Using this notation we can write the *lower-level problems* in the form

$$\min\ [L(x,p) = Q(x) + \langle p, F(x)\rangle] \tag{3.6}$$

subject to

$$g(x) \leq 0 .$$

The *coordination problem* (3.5) takes the form

$$D_*(\bar{x}(p),z) = 0 \ , \tag{3.7}$$

where $\bar{x}(p) = (\bar{c}(p), \bar{u}(p))$ is the solution of (3.6).

Let p^0 be the solution of the equation $F(\bar{x}(p)) = 0$, and let $x^0 = \bar{x}(p^0)$. This solution may easily be obtained by the classical dual method. Let Ω_p, Ω_x be neighborhoods of p^0, x^0 which satisfy the relation $\bar{x}(\Omega_p) \subset \Omega_x$. We then make the following assumptions:

2. The functions Q,F, and g are twice differentiable on Ω_x and the function g is convex.
3. There exists a constant $\nu > 0$ such that $\nabla^2_{xx} L(x,p) \geq \nu I$ for $x \in \Omega_x$, $p \in \Omega_p$ (i.e., the operator $\nabla^2_{xx}(L(x,p) - \nu I)$ is positive semidefinite).

3.2 Properties of lower-level solutions

In order to investigate the properties of the function $p \to \bar{x}(p)$, we have to study the effects of the scalar inequality constraints $g_{ij}(x) \leq 0$ $(i \in \overline{1,N}, j \in \overline{1,J_i})$ involved in problem (3.6). Let

$$I_0(p) = \{(i,j): g_{ij}(\bar{x}_i(p)) = 0\},$$

and

$$\Gamma(p,x) = (g_{ij}(x))_{(i,j) \in I_0(p)}.$$

We assume that for $p \in \Omega_p, x = \bar{x}(p)$, the linear operator $\Gamma_x(p,x)$ is nonsingular. Under this assumption there exist unique Lagrange multipliers $\mu(p) \in R^J$ which correspond to the constraints in problem (3.6). We define the operator $W(p) = \nabla^2_{xx}(L(\bar{x}(p),p) + \langle \mu(p), g(\bar{x}(p)) \rangle)$. Obviously, from assumptions 2 and 3, $W(p) \geq \nu I$ for $p \in \Omega_p$.

Let $p^1, p^2 \in \Omega_p$ and let $p(t) = (1-t)p^1 + t\,p^2$. We shall investigate the behavior of the function $t \to \bar{x}(p(t))$ for $t \in [0,1]$. Since the set $I_0(p(t))$ may change as t ranges between 0 and 1, this function may be nondifferentiable.

Lemma 4. If the set $I_0(p(t))$ is constant in a neighborhood of the point $t_0 \in (0,1)$ then the function $t \to \bar{x}(p(t))$ is differentiable at t_0, and

$$\frac{d\,\bar{x}(p(t))}{dt} = -\,B(p(t_0))F_x^*(\bar{x}(p(t_0)))\quad (p^2-p^1), \tag{3.8}$$

where

$$B(p) = \begin{cases} W^{-1} & \text{, if } I_0(p) = 0 \\ W^{-1} - W^{-1}\,\Gamma_x^*\,(\Gamma_x W^{-1}\,\Gamma_x^*)^{-1}\,\Gamma_x W^{-1} & \text{, otherwise.} \end{cases} \tag{3.9}$$

For brevity we have written $W(p)$ as W and $\Gamma_x(p,\bar{x}(p))$ as Γ_x. Moreover, for any $p \in \Omega_p$, $B(p) = B^*(p)$ and

$$N(B(p)) = R(\Gamma_x^*(p,\bar{x}(p))) , \tag{3.10}$$

$$0 \leq B(p) \leq W(p)^{-1} . \tag{3.11}$$

($N(A)$ and $R(A)$ denote the null space and the range of the operator A, respectively).

Proof. The propositions of the lemma follow immediately from the necessary optimality conditions for problem (3.6) and the assumption that the operator $\Gamma_x(p,\bar{x}(p))$ is nonsingular.

Observe that the function $t \to B(p(t))$ is in general discontinuous and so the function $t \to d\bar{x}(p(t))/dt$ is also discontinuous. However, if the set $I_0(p(t))$ changes a countable number of times as t ranges between 0 and 1, then the functions $t \to \bar{x}(p(t))$, $t \to D(\bar{x}(p(t)))$ are absolutely continuous and

$$D(\bar{x}(p^2)) - D(\bar{x}(p^1)) = -\int_0^1 D_x(\bar{x}(p(t)))B(p(t))F_x^*(\bar{x}(p(t)))\,(p^2-p^1)\,dt \tag{3.12}$$

This condition does not seem to be very restrictive in practice and we shall impose it on our problem.

Since we may assume that $D_*(x,z) \approx D(x)$, equation (3.12) gives a good approximation to the difference $D_*(\bar{x}(p^2),z) - D_*(\bar{x}(p^1),z)$. Thus it seems reasonable to construct an algorithm for the solution of (3.7) based on equation (3.12), in a way similar to that

used in ref. 12. However, this leads to a number of problems. If p is close to p^0 then $D_x(\bar{x}(p))$ and $F_x(\bar{x}(p))$ are close to $D_x(x^0)$ and $F_x(x^0)$, respectively, but no statement of this type can be made about the operator $B(p)$. Moreover, it seems unreasonable to compute $B(p)$, since this involves many time-consuming operations. In the next section we shall develop an algorithm for the solution of (3.7) which does not require the computation of $B(p)$; this is due to the properties (3.10) and (3.11) of the operator $B(p)$.

3.3 The algorithm

Let $\varepsilon > 0$ and let A be a selfconjugate positively defined operator such that $A : \underset{\sim}{X} \to \underset{\sim}{X}$, $m_A I \leq A \leq M_A I$. We define operators $D_0 = D_x(x^0)$, $F_0 = F_x(x^0)$, and $E = (D_0 A F_0^*)^{-1}$. Let us consider the algorithm

$$p^{k+1} = p^k + E\, D_*(\bar{x}(p^k), z). \tag{3.13}$$

It is easy to verify that the matrix E is well-defined, since it follows from assumption 1 that $N(D_0) = N(F_0)$. We shall prove that the mapping $p^k \to p^{k+1}$ has a fixed point in a neighborhood of p^0 and that the algorithm (3.13) converges to this fixed point. To do this we must make the following important assumption.

4. There exists $\delta > 0$ such that for $p \in \Omega_p$ and $v \in R(F_0^*)$ the inequality $d(v, R(\Gamma_x^*(p, \bar{x}(p)))) \geq \delta \| v \|$ is satisfied, where $d(v,M)$ denotes the distance between the point v and the set M.

We now introduce a new norm $\| \cdot \|_0$ in $\underset{\sim}{U}$, derived from the scalar product

$$\langle v^1, v^2 \rangle_0 = \langle F_0^* v^1,\ A F_0^* v^2 \rangle \quad .$$

Since $N(F_0^*) = \{0\}$ then the norm $\| \cdot \|_0$ is well defined.

Lemma 5. Let $V(p) = p + E D(\bar{x}(p))$. Let $\| W(p) \| \leq M_w$ for $p \in \Omega_p$. Finally, let the inequality

$$\| D_x(\bar{x}(p^1)) B(p^2) F_x^*(\bar{x}(p^1)) - D_x(\bar{x}(p^2)) B(p^2) F_x^*(\bar{x}(p^2)) \|_0 \leq L \| p^1 - p^2 \|_0 \tag{3.14}$$

hold for any $p^1, p^2 \in \Omega_p$. Then for any $p^1, p^2 \in \Omega_p$

$$\| V(p^2) - V(p^1) \|_0 \leq \alpha(\varepsilon) \| p^2 - p^1 \|_0 +$$

$$+ \frac{1}{2} \varepsilon L \|E\|_0 (\|p^1 - p^0\|_0 + \|p^2 - p^0\|_0) \|p^2 - p^1\|_0 , \quad (3.15)$$

where

$$\alpha(\varepsilon) = (1 - \frac{2\varepsilon\delta^2}{M_A M_W} + \frac{\varepsilon^2}{\nu^2 m_A^2})^{1/2}. \quad (3.16)$$

Proof. Let $p(t) = (1-t)p^1 + t\,p^2$.

Let $S = R(F_0^*)$ and let P_s denote the projection on S. We define linear operations $A_s: S \to S$, $B_s(p): S \to S$, $F_s: S \to \underset{\sim}{U}$, $D_s: S \to \underset{\sim}{U}$ as follows: $A_s = P_s A|_S$, $B_s(p) = P_s B(p)|_S$, $F_s = F_0|_S$, $D_s = D_0|_S$. Next we define

$$R = \int_0^1 B_s(p(t))dt .$$

It follows from (3.10) that for any $x \in S$

$$\langle x, B_s(p(t))x\rangle = \langle x, B(p(t))x\rangle = \langle s, W(p(t))^{-1}s\rangle ,$$

where $s = x - \Gamma_x^*(\Gamma_x W^{-1} \Gamma_x^*)^{-1} \Gamma_x W^{-1} x$. For brevity we have written $W(p(t))$ as W, and $\Gamma_x(p(t), \bar{x}(p(t)))$ as Γ_x. Hence, from (3.11) and assumption 4, we have

$$\langle x, B_s(p(t))x\rangle \geq \frac{1}{M_W} \|s\|^2 \geq \frac{\delta^2}{M_N} \|x\|^2 .$$

Thus

$$\langle x, Rx\rangle \geq \frac{\delta^2}{M_N} \|x\|^2 \quad (3.17)$$

for $x \in S$.

Let $\bar{v} = (I - \varepsilon E D_0 R F_0^*)(p^2 - p^1)$. Since $N(D_0) = N(F_0) \perp S$, then $E = (D_s A_s F_s^*)^{-1} = (F_s^*)^{-1} A_s^{-1} D_s^{-1}$. Thus

$$\|\bar{v}\|_0^2 = \langle p^2-p^1,\ p^2-p^1\rangle_0 - 2\varepsilon \langle p^2-p^1,\ (F_s^*)^{-1}A_s^{-1}R\ F_0^*(p^2-p^1)\rangle_0 +$$

$$+ \varepsilon^2\langle (F_s^*)^{-1}A_s^{-1}\ R\ F_0^*(p^2-p^1),\ (F_s^*)^{-1}A_s^{-1}\ R\ F_0^*\ (p^2-p^1)\rangle_0\ . \qquad (3.18)$$

We shall now estimate the components of (3.18). It follows from the definition of $\langle\cdot,\cdot\rangle_0$ and from (3.17) that

$$\langle p^2-p^1,\ (F_s^*)^{-1}\ A_s^{-1}\ R\ F_0^*(p^2-p^1)\rangle_0 \geq \frac{\delta^2}{M_w}\ \|F_0^*\ (p^2-p^1)\|^2\ .$$

Next, we obtain from (3.11) the inequality

$$\langle (F_s^*)^{-1}\ A_s^{-1}\ R\ F_0^*\ (p^2-p^1),\ (F_s^*)^{-1}\ A_s^{-1}\ R\ F_0^*(p^2-p^1)\rangle_0 =$$

$$= \langle A_s^{-1}\ R\ F_0^*(p^2-p^1),\ R\ F_0^*(p^2-p^1)\rangle \leq \frac{1}{\nu^2 m_A}\ \|F_0^*\ (p^2-p^1)\|^2\ .$$

Finally, it follows from the definition of the norm $\|\cdot\|_0$ that

$$\frac{1}{M_A}\ \|p^2-p^1\|_0^2 \leq \|F_0^*\ (p^2-p^1)\|^2 \leq \frac{1}{m_A}\ \|p^2-p^1\|_0^2\ .$$

Using the above three estimates, we obtain from (3.18) the inequality

$$\|\bar{v}\|_0 \leq \alpha(\varepsilon)\ \|p^2-p^1\|_0 \qquad (3.19)$$

where $\alpha(\varepsilon)$ is defined by (3.16).

It is now easy to prove the inequality (3.15). It follows from (3.12) that

$$V(p^2) - V(p^1) = \bar{v} + \varepsilon E \int_0^1 (D_0B(p(t))F_0^* -$$

$$- D_x(\bar{x}(p(t))B(p(t)))F_x^*(\bar{x}(p(t)))) (p^2-p^1)dt$$

Next, by virtue of (3.14),

$$\|V(p^2) - V(p^1)\|_0 \leq \|\bar{v}\|_0 + \varepsilon L\ \|E\|_0\ \|p^2-p^1\| \int_0^1 \|p(t)-p^0\|_0\, dt \leq$$

$$\leq \|v\|_0 + \frac{1}{2}\varepsilon L\ \|E\|_0\ (\|p^1-p^0\|_0 + \|p^2-p^2\|_0)\ \|p^2-p^1\|_0. \qquad (3.20)$$

The inequality (3.15) follows from (3.19) and (3.20), and the lemma has been proved.

Now let us define the operation

$$\tilde{D}(x,z) = D_*(x,z) - D(x) = u_0(c) - u(c,z) .$$

Theorem 3. Let the assumptions of Lemma 5 be satisfied. Let also

$$\|\tilde{D}(\bar{x}(p^1),z) - \tilde{D}(\bar{x}(p^2),z)\|_0 \le \tilde{L}\,\|p^1-p^2\|_0 \tag{3.21}$$

for $p^1,p^2\in\Omega_p$. Next, let

$$\|\,E\,D_*(x^0,z)\,\|_0 \le \eta . \tag{3.22}$$

We define the following constants: $\xi = (1-\alpha(\varepsilon))/\varepsilon$, $b = \|E\|_0$, $h = bL\eta$, $r = (\xi - b\tilde{L} - [(\xi - b\tilde{L})^2 - 2h]^{1/2})/bL$. Let the following conditions be satisfied:

$$\sqrt{2h} + b\tilde{L} \le \xi \tag{3.23}$$

$$K(p^0;r) = \{p\in\underset{\sim}{U}: \|p-p^0\|_0 \le r\}\subset\Omega_p . \tag{3.24}$$

Then equation (3.7) has a unique solution $\tilde{p}(z)$ in $K(p^0;r)$ and the sequence $\{p^k\}$ generated by (3.13) converges to $\tilde{p}(z)$.

Proof. We define the operation $V_*(p) = p + \varepsilon\,E\,D_*(\bar{x}(p),z)$. It follows from Lemma 5 and (3.21) that

$$\|V_*(p^2) - V_*(p^1)\|_0 \le (\alpha(\varepsilon) + \varepsilon b\tilde{L})\,\|p^2-p^1\|_0 +$$

$$+ \tfrac{1}{2}\varepsilon bL\,(\|p^1-p^0\|_0 + \|p^2-p^0\|_0)\,\|p^2-p^1\|_0 \tag{3.25}$$

for any $p^1,p^2\in\Omega_p$. We define the function $\psi\colon R^1\to R^1$,

$$\psi(t) = \tfrac{1}{2}\varepsilon\,bL\,t^2 + (\alpha(\varepsilon) + \varepsilon b\tilde{L})\,t + \varepsilon\eta .$$

It follows from (3.25) that for $p^1,p^2\in\Omega_p$ and $t \ge \frac{1}{2}(\|p^1-p^0\|_0 +$ $+ \|p^2-p^0\|_0)$ the following inequality holds:

$$\|V_*(p^2) - V_*(p^1)\|_0 \leq (\alpha(\varepsilon) + \varepsilon b\tilde{L} + \varepsilon bL\, t)\|p^2 - p^1\|_0 \leq \psi'(t)\|p^2 - p^1\|_0 . \tag{3.26}$$

Next,

$$\|V_*(p^0) - p^0\|_0 \leq \varepsilon n = \psi(0) . \tag{3.27}$$

In addition to the sequence $\{p^k\}$ we shall consider the sequence $\{t_k\}$ defined by the formula $t_{k+1} = \psi(t_k)$, $t_0 = 0$. It is evident that $t_k \uparrow r$, since r is the smaller of the two roots of the equation $\psi(t) = t$. We are going to prove that $\{t_k\}$ is the majorizing sequence for the sequence $\{p^k\}$, i.e., that $\|p^{k+1} - p^k\|_0 \leq t_{k+1} - t_k$. It follows from (3.27) that

$$\|p^1 - p^0\|_0 \leq \|V_*(p^0) - p^0\|_0 \leq t_1 - t_0 .$$

Let us suppose that

$$\|p^j - p^{j-1}\|_0 \leq t_j - t_{j-1}$$

for $1 \leq j \leq k$. Let $p(\tau) = (1-\tau)p^{k-1} + \tau p^k$, and let $0 = \tau_0 \leq \tau_1 \ldots \leq \tau_N = 1$. It follows from (3.26) that

$$\|p^{k+1} - p^k\|_0 = \|V_*(p^k) - V_*(p^{k-1})\|_0 \leq \sum_{i=0}^{N-1} \|V_*(p(\tau_i)) - V_*(p(\tau_{i+1}))\|_0 \leq$$

$$\leq \sum_{i=1}^{N} \psi'(\rho_i)\|p(\tau_i) - p(\tau_{i+1})\|_0 ,$$

where

$$\rho_i = \frac{1}{2}(\tau_i + \tau_{i+1})(t_k - t_{k-1}) + t_{k-1} \geq$$

$$\geq \frac{1}{2}(\|p(\tau_i) - p^0\|_0 + \|p(\tau_{i+1}) - p^0\|_0) .$$

Approaching the limit as $N \to \infty$, we find that $\sup |\tau_{i+1} - \tau_i| \to 0$ and we obtain the inequality

$$\|p^{k+1} - p^k\|_0 \leq \int_0^1 \psi'(t_{k-1} + \tau(t_k - t_{k-1}))(t_k - t_{k-1})d\tau =$$

$$= \int_{t_{k-1}}^{t_k} \psi'(t)dt = \psi(t_k) - \psi(t_{k-1}) = t_{k+1} - t_k \quad .$$

Thus, we have proved by induction that $\|p^{k+1}-p^k\|_0 \leq t_{k+1}-t_k$ for all $k \geq 0$, and hence that the sequence $\{p^k\}$ converges to the solution $\tilde{p}\in K(p^0;r)$ of equation (3.7). In order to prove that this solution is unique, let us assume that $V_*(\bar{p}) = \bar{p}$ for some $\bar{p}\in K(p^0;r)$. Let $\bar{t}_0 = \|\bar{p} - p^0\|_0$ and $\bar{t}_{k+1} = \psi(t_k)$. Adopting a method similar to that used above, it is possible to prove that $\|\bar{p} - p^k\| \leq \bar{t}_k - t_k$. But $\lim \bar{t}_k = \lim t_k = r$, and hence $\bar{p} = \tilde{p}$. Finally, let us note that in a similar fashion it can be proved that the solution $\tilde{p}$ is unique in $K(p^0;\bar{r})$, where $\bar{r} = (\xi - b\tilde{L} + [(\xi - b\tilde{L})^2 - 2h]^{1/2})/b\tilde{L}$ is the greater of the roots of the equation $\psi(t) = t$. Moreover, if the algorithm (3.13) is initialized with $\bar{p}^0\in$ int $K(p^0;r)$ it generates a sequence $\{\bar{p}^k\}$ which converges to $\tilde{p}$. Thus the theorem has been proved.

To sum up, we have solved equation (3.7) which involves non-differentiable functions $p \to \bar{x}(p)$ and $x \to D_*(x,z)$, using a Newton-like algorithm (3.13). The algorithm exploited a number of special features of the problem under consideration, the most important of which were: $D_*(x,z) = D(x)$ (conditions (3.21) and (3.23)), $N(D_0) = N(G_0)$ (assumption 1) and $B > 0$ on S (Lemma 4 and (3.17)). It is worth noting that if we replace (3.23) by a strict inequality then the operation V_* has contraction mapping properties in $K(p^0;r)$. This feature is important for control problems in which z is a time-varying parameter, because it is then possible to trace the moving solution $\tilde{p}(z(t))$ of the nonstationary equation $D_*(\bar{x}(p),z(t)) = 0$ (see ref. 11).

REFERENCES

1. Yu.M. Ermoliev, *Methods of Stochastic Programming* (in Russian), Nauka, Moscow, 1976.

2. W. Findeisen, *Multilevel Control Systems* (in Polish), PWN, Warsaw, 1974 (German translation: Hierarchische Steuerungssysteme, Verlag Technic, Berlin, 1977).

3. W. Findeisen et al., On-line hierarchical control for steady-state systems, *IEEE Trans. Automat. Contr. (Special Issue on Decentralized Control and Large Scale Systems)*, Vol. AC-23 (1978), 189-209.

4. W. Findeisen, F.N. Bailey, M. Brdyś, K. Malinowski, P. Tatjewski, and A. Woźniak, *Control and Coordination in Hierarchical Systems*, Volume 9 in the IIASA International Series on Applied Systems Analysis, John Wiley, Chichester, 1980.

5. L.S. Lasdon, *Optimization Theory for Large-Scale Systems*, Macmillan, New York, 1970.

6. L. Malinowski and A. Ruszczyński, Application of interaction balance method to real process coordination, *Control and Cybernetics*, 4, 2, (1975).

7. M.D. Mesarovic et al., *Theory of Hierarchical, Multilevel Systems*, Academic Press, New York, 1970.

8. E.A. Nurminski, The quasigradient method for the solution of nonlinear programming problems, *Cybernetics*, 9, 1 (1974), 145-150.

9. A. Ruszczyński, An algorithm for the interaction balance method with feedback (in Polish), *Archiwum Automatyki i Telemechaniki*, 21, 1 (1976), 137-150.

10. A. Ruszczyński, Convergence conditions for the interaction balance algorithm based on an approximate mathematical model, *Control and Cybernetics*, 5, 4 (1976), 29-43.

11. A. Ruszczyński, Coordination of nonstationary systems, *IEEE Trans. Automat. Contr.*, Vol. AC-24 (1979), 51-62.

12. A.I. Zinchenko, On the approximate solution of functional equations with nondifferentiable operators (in Russian), *Matem. Fizika*, 14, (1973), 55-58.

LAGRANGIAN FUNCTIONS AND NONDIFFERENTIABLE OPTIMIZATION

A.P. Wierzbicki
International Institute for Applied Systems Analysis
Laxenburg, Austria

1. INTRODUCTION

Rapid development and intensive research into nondifferentiable optimization techniques (Balinski and Wolfe, 1975) has resulted recently in algorithms that are closely related or even equivalent in the differentiable case to known and effective techniques of differentiable optimization. A very interesting quasi-Newton technique for nondifferentiable optimization was proposed and partly investigated in Lemaréchal (1978). To understand fully possible weak and strong points of quasi-Newton methods in nondifferentiable optimization, a more exhaustive study of various relations between nondifferentiable and differentiable problems is needed. Because of the large variety of nondifferentiable problems, this goal cannot be achieved in a short paper. However, some theoretical insight can be obtained by analyzing the most simple type of nondifferentiable problems:

$$\underset{x \in X}{\text{minimize}}\ f(x) \quad ; \quad f(x) = \max_{i \in I} f_i(x) \tag{1}$$

where X is a convex set with nonempty interior in R^n (possibly $X = R^n$), I is a countable set of indexes (possibly finite). It is assumed that f is bounded from below on X and that $\max_{i \in I} f_i(x)$

for each $x \in X$ is attained at a finite subset $A(x) = A \subset I$; $f_i: R^n \to R^1$ are twice-differentiable functions. It is not necessarily assumed that the Haar condition is satisfied, that is, if $\hat{x} \in \text{Arg} \min_{x \in X} \max_{i \in I} f_i(x)$, then for any subset $\bar{A} \subset A(\hat{x})$, the matrix composed of $f_{ix}(\hat{x})$ for $i \in \bar{A}$ has its maximal rank. If this condition is satisfied, then $\hat{x}$ is uniquely determined by $f_{ix}(\hat{x})$, $i \in A(\hat{x})$ only, and some efficient algorithms for solving the problem (1) are known (Madsen and Schjaer-Jacobsen, 1977); however, this condition is rarely satisfied in practical problems. Other conditions of second-order type resulting in the uniqueness of $\hat{x}$ are further assumed to hold, together with conditions implying the uniqueness of baricentric coordinates in subdifferential sets.

The functions f_i might be assumed convex or not; this problem is discussed in detail later. The assumptions of countability of I and finiteness of A could actually be relaxed, although this generalization is beyond the scope of this paper. If the functions f_i are not differentiable, it is often possible to reformulate the problem (1) by enlarging the set I in such a way that the modified functions f_i are differentiable. It would seem, therefore, that the class of nondifferentiable problems considered could be extended to cover almost all problems encountered in practice. However, still more assumptions are needed for a theoretical investigation: that the activity set $A(x) = A$ can be determined explicitly for each $x \in X$ and that the subdifferential $\partial f(x)$ of f at x can also be fully determined:

$$\partial f(x) = \{g \in R^n : g = \sum_{i \in A} \lambda_i f^*_{ix}(x) \ , \ \lambda_i \geq 0, \ \sum_{i \in A} \lambda_i = 1\} \qquad (2)$$

where $f^*_{ix}(x)$ are the gradients of functions f_i at x (written as column vectors, nence the transposition sign *). This assumption is not always satisfied in practical problems of nondifferentiable optimization and can even be considered as contradictory to the very nature of nondifferentiable optimization techniques, where one of the main problems is to estimate the subdifferential $\partial f(x)$ without knowing its full description. On the other hand,

in order to obtain a better theoretical insight, it is useful to proceed in two stages: first, investigate the implications that $A = A(x)$ and $\partial f(x)$ are known explicitly, then try to relax this assumption and check for which theoretical properties this assumption is crucial.

Under the assumption that $A = A(x)$ and $\partial f(x)$ are known explicitly, problem (1) is equivalent to a constrained differentiable optimization problem which can be studied by introducing a normal or an augmented Lagrangian function, depending on convexity assumptions. In this way, the relation of nondifferentiable techniques for solving problem (1) to known techniques of differentiable optimization can be investigated, sufficient conditions of optimality can be studied and some strong (superlinear or even quadratic) convergence properties of a special variant of nondifferentiable quasi-Newton techniques can be deduced. These strong results substantiate the introduction of a special class of *nondifferentiable optimization problems with explicitly known subdifferentials*, which are in fact equivalent to differentiable problems and can be solved efficiently by appropriate techniques--provided the number of elements in the activity set, $|A(x)|$, is not too large to make the explicit definition of the subdifferential computationally cumbersome.

If $A = A(x)$ and $\partial f(x)$ are not known explicitly, or their explicit definition is computationally cumbersome, then only some subgradients $g \in \partial f(x)$ can be computed without specifying $f_{ix}(x)$ and the baricentric coordinates λ_i. This constitutes the opposite class of *nondifferentiable optimization problems with implicit subdifferentials*. In this class, it is very difficult to construct a quasi-Newton method that would converge superlinearly since the subgradient g cannot be used to obtain a sufficiently accurate approximation of a Hessian matrix that would result in strong convergence properties; in fact, such a construction seems to be impossible. However, many algorithms with sublinear or linear convergence are known for the case of implicit subdifferentials and these algorithms are in fact more practically efficient for problems in which the explicit definition of the subdifferential is computationally cumbersome.

2. A BASIC LEMMA

One of the fundamental problems in nondifferentiable optimization is as follows. Given the set $\partial f(x)$ or an approximation G thereof, expressed by convex combinations of a set of vectors $g_i \in R^n$ for $i \in A$ and by some accuracy parameters $\alpha_i \geq 0$, $i \in A$ (where $A = A(x)$, $g_i = f'_{ix}(x)$, $\alpha_i = 0$ if the set $\partial f(x) = G$ is given explicitly), check whether $0 \in G$, and if not, find the vector $\hat{g} \in G$ of minimal norm, subject to accuracy corrections if $\alpha_i > 0$. When using quasi-Newton methods of nondifferentiable optimization, the norm in which $\hat{g}$ is minimized must be chosen according to some other properties of the problem. Therefore, denote $\|g\|^2_{H^{-1}} = \langle g , H^{-1}g \rangle$, where $H^{-1}: R^n \to R^n$ is a given positive definite matrix. The basic problem can be stated as follows:

$$\text{minimize } \left(\frac{1}{2}\|g\|^2_{H^{-1}} + \sum_{i \in A} \alpha_i y_i\right) \quad ; \tag{3}$$

$$YG = \{(y,g) : g = \sum_{i \in A} y_i g_i \ , \ \sum_{i \in A} y_i = 1 \ , \ y_i \geq 0 \ , \ i \in A\} \ .$$

The following lemma is actually only an extension of the results given in Lemaréchal (1978), but, since it is fundamental for the investigation of relations between nondifferentiable optimization methods and equivalent Lagrangian function approaches, it is presented here in detail.

Lemma 1. The problem (3) is equivalent to the following dual problem

$$\underset{(\bar{x}_0, \bar{x}) \in \bar{X}_0}{\text{minimize}} \ \left(\bar{x}_0 + \frac{1}{2}\|\bar{x}\|^2_H\right) \quad ; \tag{4}$$

$$\bar{X}_0 = \{(\bar{x}_0, \bar{x}) \in R^{n+1} : \langle g_i, \bar{x} \rangle - \bar{x}_0 - \alpha_i \leq 0 \quad , \quad i \in A\}$$

where $\|\bar{x}\|^2_H = \langle \bar{x} , H\bar{x} \rangle$. The equivalence is to be understood in the sense that, if $\hat{g}$, $\hat{y}$ are solutions to the problem (3) with

Lagrange multipliers $\hat{\bar{x}}$ for the constraint $g - \sum_{i\in A} y_i g_i = 0$ and $\hat{\bar{x}}_0$ for the constraint $\sum_{i\in A} y_i - 1 = 0$, then $\hat{\bar{x}}_0$, $\hat{\bar{x}}$ are solutions of problem (4) with Lagrange multipliers $\hat{y}_i$ for constraints $\langle g_i, \bar{x}\rangle - \bar{x}_0 - \alpha_i \le 0$ with $\hat{\bar{x}} = -H^{-1}\hat{g}$, $\hat{\bar{x}}_0 = -\|\hat{g}\|^2_{H^{-1}} - \sum_{i\in A}\alpha_i y_i \le 0$. The following equivalences also hold:

$$\hat{\bar{x}} = 0 \Leftrightarrow \hat{g} = 0 \Leftrightarrow \hat{\bar{x}}_0 = -\sum_{i\in A}\alpha_i \hat{y}_i = -\min_{i\in A}\alpha_i \Leftrightarrow$$

$$\hat{\bar{x}}_0 = 0 \text{ if any of } \alpha_i = 0 \ ;$$

generally, $-\hat{\bar{x}}_0 \ge \|\hat{x}\|^2_H \ge 0$ and $-\hat{\bar{x}}_0 \ge \sum_{i\in A}\alpha_i\hat{y}_i \ge 0$. Moreover, the solutions $\hat{\bar{x}}_0$, $\hat{\bar{x}}$, $\hat{g}$ of the problems (3), (4) are unique, whereas $\hat{y}$ is unique if vectors $h_i = (-1, g_i) \in R^{n+1}$ are linearly independent for $i \in A$, and, generally, $\hat{y} \in \hat{Y}$, where $\hat{Y}$ is a compact convex set. Even if $\hat{y}$ is not unique, it minimizes $\sum_{i\in A}\alpha_i\tilde{y}_i$ over $\tilde{y} \in y = \{y : y_i \ge 0, \sum_{i\in A} y_i = 1, \sum_{i\in A} y_i g_i = \hat{g}\}$. If $\hat{y}$ is unique, then, for any positive definite H^{-1}, the solutions $\hat{\bar{x}}_0$, $\hat{\bar{x}}$, $\hat{g}$, $\hat{y}$ depend Lipschitz-continuously on the data g_i, α_i.

Proof. Both problems are convex. Consider first the question of the uniqueness of their solution. Problem (3) clearly has a unique solution $\hat{g}$ in g and, if h_i are linearly independent for $i \in A$, a unique solution $\hat{y}$ in y. Observe that the linear dependence of the vectors $h_i = (-1, g_i) \in R^{n+1}$, that is, the existence of $\alpha_i \ne 0$ such that $\sum_{i\in I}\alpha_i h_i = 0$, is equivalent to the existence of $\alpha_i \ne 0$ such that $\sum_{i\in I}\alpha_i = 0$, $\alpha_j = -\sum_{i\ne j}\alpha_i$, and $\alpha_j g_j = -\sum_{i\ne j}\alpha_i g_i$, which, in turn, is equivalent to the existence of $\lambda_i = -(\alpha_i/\alpha_j)$, $\sum_{i\ne j}\lambda_i = 1$, and $g_j = \sum_{i\ne j}\lambda_j g_i$. If h_i are linearly independent, such a situation cannot occur and an arbitrary g_j, $j \in I$, cannot be a convex combination of other g_i; this implies the uniqueness of baricentric coordinates $\hat{y}_i$. If h_i are linearly dependent, choose a minimal subset $\bar{A} \subset A$ such that $g = \sum_{i\in\bar{A}}\tilde{y}_i g_i$, $\sum_{i\in\bar{A}}\tilde{y}_i = 1$,

$\tilde{y}_i \geq 0$, and put $\tilde{y}_i = 0$ for $i \notin A$. If the choice of $\bar{A}$ is not unique, define such a $\tilde{y}$ for each $\bar{A}$; define a set $\tilde{Y}$ as the convex hull of all such $\tilde{y}$. Since all $y \in \tilde{Y}$ result in the same $\hat{g}$, the set of optimal $\hat{y}$ is defined by $\hat{Y} = \text{Arg min}_{y \in \tilde{Y}} \sum_{i \in A} \alpha_i y_i$; $\hat{Y}$ is a compact and closed set. Problem (4) has a unique solution $(\hat{\bar{x}}_0, \hat{\bar{x}})$, since $\hat{\bar{x}}_{01} + \frac{1}{2}\|\hat{\bar{x}}_1\|^2_H = \hat{\bar{x}}_{02} + \frac{1}{2}\|\hat{\bar{x}}_2\|^2_H$ implies $\hat{\bar{x}}_{01} = \hat{\bar{x}}_{02}$ if $\hat{\bar{x}}_1 = \hat{\bar{x}}_2$; if $\hat{\bar{x}}_1 \neq \hat{\bar{x}}_2$, then $\bar{x}_0 = \beta\hat{\bar{x}}_{01} + (1-\beta)\hat{\bar{x}}_{02}$, $\bar{x} = \beta\hat{\bar{x}}_1 + (1-\beta)\hat{\bar{x}}_2$ for $\beta \in (0;1)$ would yield a smaller value of $\bar{x}_0 + \frac{1}{2}\|\bar{x}\|^2_H$.

Consider now problem (3) and define the Lagrangian function:

$$\mathcal{L}(\bar{x}, \bar{x}_0, y, g) \overset{df}{=} \frac{1}{2}\|g\|^2_{H^{-1}} + \sum_{i \in A} \alpha_i y_i + \langle \bar{x}, g - \sum_{i \in A} y_i g_i \rangle + \bar{x}_0 \left(\sum_{i \in A} y_i - 1 \right) \qquad (5)$$

Since the equality constraints in (3) are affine, each solution $(\hat{y}, \hat{g})$ of (3) together with the corresponding Lagrange multipliers $(\hat{\bar{x}}, \hat{\bar{x}}_0)$ are saddle points of the function (5) under the additional constraint $y_i \geq 0$. Hence, if $\hat{X}$, $\hat{X}_0$ denote sets of possible Lagrange multipliers $\hat{\bar{x}}$, $\hat{\bar{x}}_0$ for $(\hat{y}, \hat{g}) \in \hat{Y} \times \{\hat{g}\}$, then:

$$\hat{X} \times \hat{X}_0 \times \hat{Y} \times \{\hat{g}\} = \text{Arg} \min_{y \geq 0, g \in R^n} \max_{(x, x_0) \in R^{n+1}} \mathcal{L}(\bar{x}, \bar{x}_0, y, g) \qquad (6)$$

$$= \text{Arg} \max_{(\bar{x}, \bar{x}_0) \in R^{n+1}} \min_{y \geq 0, g \in R^n} \mathcal{L}(\bar{x}, \bar{x}_0, y, g)$$

where Arg min max is the set of points resulting in min max, etc. Compute the value g that minimizes $\mathcal{L}$ for a given $\bar{x}$, $\bar{x}_0$, y. Clearly, $\tilde{g} = \tilde{g}(x) = -H\bar{x}$; this implies that $\hat{\bar{x}} = -H^{-1}\hat{g}$ and $\hat{\hat{X}} = \{\hat{\bar{x}}\}$ is unique. Moreover, after easy computation

$$\mathcal{L}(\bar{x},\bar{x}_0,y,\tilde{g}(x)) = -\frac{1}{2}\|\bar{x}\|^2_H - \bar{x}_0 - \sum_{i\in A} y_i(\langle g_i,\bar{x}\rangle\bar{x}_0 - \alpha_i) \tag{7}$$

$$\stackrel{df}{=} -\bar{\mathcal{L}}(y,\bar{x}_0,\bar{x})$$

and at the saddle point, $-\frac{1}{2}\|\hat{\bar{x}}\|^2_H - \hat{\bar{x}}_0 = \frac{1}{2}\|\hat{g}\|^2_{H^{-1}} + \sum_{i\in A}\alpha_i\hat{y}_i$; since $\|\hat{g}\|^2_{H^{-1}} = \|\hat{\bar{x}}\|^2_H$, this implies $\hat{\bar{x}}_0 = -\|\hat{\bar{x}}\|^2_H - \sum_{i\in A}\alpha_i\hat{y}_i$. Hence $\hat{\bar{x}}_0$ is also unique, $\hat{\bar{x}}_0 = \{\hat{\bar{x}}_0\}$. Obviously, $\hat{\bar{x}} = 0 \Rightarrow \hat{g} = 0$ and $\hat{\bar{x}}_0 = 0 \Rightarrow \hat{\bar{x}} = 0$, $\hat{g} = 0$.

The function $\bar{\mathcal{L}}$ in (7) is the Lagrangian function for problem (4). Observe that problem (4) satisfies the Slater condition, since $\bar{x}_1 = 0$, $\bar{x}_{01} > 0$ are admissible for the problem and $\langle g_i, \bar{x}_1\rangle - \bar{x}_{01} - \alpha_i < 0$ for all $i \in A$. Moreover, it is well known that

$$\operatorname{Arg}\min_{(\bar{x}_0,\bar{x})\in R^{n+1}} \max_{y\geq 0} \bar{\mathcal{L}}(y,\bar{x}_0,x) = \{\hat{\bar{x}}_0\} \times \{\hat{\bar{x}}\} \times \hat{Y} \quad ,$$

where $\hat{\bar{x}}_0$, $\hat{\bar{x}}$ are unique solutions of (4) and $\hat{Y}$ is the set of corresponding Lagrange multipliers. But relation (7) implies that:

$$\hat{X} \times \hat{\bar{x}}_0 \times \hat{Y} = \operatorname{Arg}\min_{y\geq 0} \max_{(\bar{x},\bar{x}_0)\in R^{n+1}} \mathcal{L}(\bar{x},\bar{x}_0,y,\tilde{g}(\bar{x})) \tag{8}$$

$$= \operatorname{Arg}\max_{y\geq 0} \min_{(\bar{x},\bar{x}_0)\in R^{n+1}} \bar{\mathcal{L}}(y,\bar{x}_0,\bar{x}) = \{\hat{\bar{x}}\} \times \{\hat{\bar{x}}_0\} \times \hat{Y} \quad .$$

Hence $\hat{\bar{Y}} = \hat{Y}$. If $\hat{Y} = \{\hat{y}\}$, the Lipschitz-continuity of $\hat{\bar{x}}$, $\hat{\bar{x}}_0$, $\hat{y}$, $\hat{g}$ in g_i and α_i results from general properties of solutions of sets of equations and inequalities--(see Szymanowski, 1977 and Wierzbicki, 1978). Moreover, since $\hat{\bar{x}}_0$ is the solution of (4), $\hat{\bar{x}} = 0 \Rightarrow \hat{\bar{x}}_0 \geq -\alpha_i$, $i \in A$, and $\hat{\bar{x}}_0 = -\min_{i\in A}\alpha_i$; if any of $\alpha_i = 0$, then $\hat{\bar{x}} = 0 \Rightarrow \hat{\bar{x}}_0 = 0$. Conversely, $\hat{\bar{x}}_0 = \sum_{i\in A}\alpha_i\hat{y}_i = -\min_{i\in A}\alpha_i \Rightarrow \hat{g} = 0 \Rightarrow \hat{\bar{x}} = 0$.

A large part of the above lemma is given in Lemaréchal (1978), however, without the full interpretation of $\hat{\bar{x}}_0$, $\hat{\bar{x}}$ as Lagrange multipliers for (3) and without the uniqueness or Lipschitz-continuity arguments. It is also observed in Lemaréchal (1978) that problem (3) is easier to solve computationally than (4); in fact, the equation $\hat{g} = \sum_{i \in A} \hat{y}_i g_i$ defines $\hat{g}$ explicitly, and is treated as a constraint in the lemma only in order to provide an interpretation for $\hat{\bar{x}}$. There exist very efficient algorithms for solving (3) in $\hat{y}$ and $\hat{g}$, if $\alpha_i = 0$ (see Wolfe, 1976 and Hohenbalken, 1978); these algorithms can also be adapted to the case when $\alpha_i > 0$. Once $\hat{y}$ and $\hat{g}$ are defined, $\hat{\bar{x}}$ and $\hat{\bar{x}}_0$ are easily computed.

Lemma 1 also allows a straightforward generalization for problems with infinite and innumerable variables and constraints in Hilbert spaces; but this generalization is beyond the scope of this paper.

3. NONDIFFERENTIABLE OPTIMIZATION WITH EXPLICIT SUBDIFFERENTIALS

3.1 Fundamentals

If the activity set $A(x)$ and the subdifferential $\partial f(x)$ are given explicitly at each $x \in X$, then the nondifferentiable problem (1) is equivalent to the following differentiable one:

$$\begin{aligned} &\underset{(x_0,x)\in X_0}{\text{minimize}}\ x_0 \quad ; \\ &X_0 = \{(x_0,x) \in R^1 \times X : f_i(x) - x_0 \leq 0 \ , \ i \in I\} \end{aligned} \tag{9}$$

with the activity set $A(x)$ defined equivalently by

$$\begin{aligned} &A(x) = \{i \in I : f_i(x) - \hat{x}_0(x) = 0 \quad ; \\ &\hat{x}_0(x) = \max_{i\in I} f_i(x)\} \ . \end{aligned} \tag{10}$$

If the functions f_i are convex, then problem (9) is convex and clearly satisfies the Slater condition with any $x_1 \in X$ and $x_{01} > \hat{x}_0(x_1)$. Thus, the normal Lagrange function:

$$L(y,x_0,x) = x_0 + \sum_{i\in I} y_i(f_i(x) - x_0)$$
$$= x_0(1 - \sum_{i\in I} y_i) + \sum_{i\in I} y_i f_i(x) \qquad (11)$$

has a saddle point $(\hat{y},\hat{x}_0,\hat{x})$ at a solution $(\hat{x}_0,\hat{x})$ of problem (9) with a corresponding Lagrange multiplier $\hat{y}$, whereas $\hat{x}$ is a solution of (1) and $\hat{x}_0 = f(\hat{x}) = \min_{x\in X} \max_{i\in I} f_i(x)$ is the minimal value of f. It is assumed further that $\hat{x}$ is a unique solution to (9) and an internal point of X.

If the number $|I|$ of constraints in (9) is large, then a purely dual method for solving (9) by assuming arbitrary $y = \{y_i\}_{i\in I}, y_i \geq 0$ and then minimizing the Lagrangian function (11) is clearly not efficient. But a primal-dual method for solving (9), which consists of determining the activity set A(x) or an approximation A thereof and eliminating inactive constraints by setting $y_i = 0$ for $i \in I\backslash A$, might be quite efficient; it is shown further that one of these primal-dual methods is probably the most efficient algorithm for nondifferentiable optimization, if $|A(\hat{x})|$ is not too large.

Suppose $y_i \geq 0$ for $i \in A$ are chosen in such a way that $\sum_{i\in A} y_i = 1$. Then $L_{x_0}(y,x_0,x) = 0$ and

$$L_x^*(y,x_0,x) = \sum_{i\in A} y_i f_{ix}^*(x) \quad (\text{if } A = A(x)) = g \in \partial f(x) \quad . \qquad (12)$$

Thus, if only $A = A(\hat{x})$ and $\hat{y}_i \geq 0$, $i \in A(\hat{x})$, $\sum_{i\in A(\hat{x})} \hat{y}_i = 1$ such that $\sum_{i\in A(\hat{x})} \hat{y}_i f_i(\hat{x}) = g = 0$ were known, then solving the equivalent problems (1), (9) would also be equivalent to minimizing the function:

$$F(\hat{y},x) = \sum_{i\in A(\hat{x})} \hat{y}_i f_i(x) \quad . \tag{13}$$

However, not only are the optimal values $\hat{y}_i$ not known, but the activity set A(x) also changes, often in arbitrary neighborhoods of $\hat{x}$. Also, the strong activity set

$$S(\hat{y}) = \{i \in I : \hat{y}_i > 0\} \tag{14}$$

might change in any neighborhood of (y,x). These difficulties are not uniquely related to nondifferentiable problems; they are also known in constrained differentiable problems. A typical way of resolving them (see, e.g., Wierzbicki, 1978) is to construct approximations A of A(x) and S of SA(y) such that $y_i = 0$ for $i \notin A$, $y_i > 0$ for $i \in S$, and

$$S \subset S(y) \quad , \quad A(x) \subset A \quad , \quad S \subset A \tag{15a}$$

and that, for (y,x) in some neighborhood of $(\hat{y},\hat{x})$:

$$S = S(\hat{y}) \subset A(\hat{x}) = A \quad . \tag{15b}$$

A detailed way of constructing such an approximation is discussed in Section 4. Here note only that a measure of the distance from $(\hat{y},\hat{x})$ to (y,x) is useful when constructing such approximations. Define

$$w = \| (L_x, \tilde{L}_y) \| \tag{16a}$$

where

$$L_x = \sum_{i\in A} y_i f_{ix}(x) \quad ; \quad \tilde{L}_{y_i} = y_i (f_i(x) - \hat{x}_0(x)) \quad . \tag{16b}$$

Here L_{y_i} is not precisely the derivative of L in y_i, but measures the violation of Kuhn-Tucker necessary conditions for optimality of $(\hat{y},\hat{x})$--if $A(\hat{x}) \subset A$ and $L_x = 0$, $\tilde{L}_{y_i} = 0$ for $i \in A$, $w = 0$, and $\hat{x}$, $\hat{y}$ are unique, then clearly $y = \hat{y}$ and $x = \hat{x}$. A lemma on the estimation of $\|(y - \hat{y}, x - \hat{x})\|$ by w is given in Section 4.

3.2 Quadratic Approximations

Consider now an approximation of the subdifferential $\partial f(x)$ by the set G:

$$G = \{g \in R^n : g = \sum_{i\in A} y_i f^*_{ix}(x) \quad ; \quad \sum_{i\in A} y_i = 1 \ , \ y_i \geq 0 \ , \ i \in A\} \tag{17}$$

and assume that $0 = \sum_{i\in A}\hat{y}_i f_{ix}(x) \in G$. Although G is only an approximation, the relation $0 \in G$ might imply $x = \hat{x}$ provided that $\sum_{i\in A}\hat{y}_i(\hat{x}_0(x) - f_i(x)) = 0$, since then $L_x = 0$, $\tilde{L}_{y_i} = 0$, and $w = 0$. This leads to a problem analogous to (3):

$$\underset{(y,g)\in YG}{\text{minimize}} \ (\tfrac{1}{2}\|g\|^2_{H^{-1}} + \sum_{i\in A}\alpha_i y_i) \quad ; \quad \alpha_i = \hat{x}_0(x) - f_i(x) \quad ; \tag{18a}$$

$$YG = \{(y,g) : g = \sum_{i\in A} y_i f^*_{ix}(x) \quad ; \quad \sum_{i\in A} y_i = 1 \ , \ y_i \geq 0\}$$

where H^{-1} is a positive definite matrix, not chosen as yet. But, due to Lemma 1, (18a) is equivalent to:

$$\underset{(\bar{x}_0,\bar{x})\in\bar{X}_0}{\text{minimize}} \ (\bar{x}_0 + \tfrac{1}{2}\|\bar{x}\|_H) \quad ; \tag{18b}$$

$$\bar{X}_0 = \{(\bar{x}_0,\bar{x}) \in R^{n+1} : f_{ix}(x)\bar{x} - \bar{x}_0 - \bar{x}_0(x) + f_i(x) \leq 0, \quad i \in A\}$$

and the choice of H^{-1} or H is now clear: (18b) is a well-known quadratic approximation problem for the Lagrangian function (11), (see, e.g., Szymanowski, 1977 and Wierzbicki, 1978) and the optimal choice of the matrix H is to approximate the Hessian of the Lagrangian function (11) as closely as possible,

$$H \approx L_{xx}(\hat{y},\hat{x}_0,\hat{x}) = \hat{L}_{xx} = \sum_{i\in SA(\hat{y})} \hat{y}_i f_{ixx}(\hat{x}) \quad , \tag{19a}$$

either by direct computation of second-order derivatives (a Newton-type method) or, for example, by variable metric techniques based on the data $\sum_{i\in S} y_i f_{ix}(x)$ for (y,x) close to $(\hat{y},\hat{x})$ and S close to $SA(\hat{y})$.

Another useful interpretation of problem (19a) results from its relation to the distance w. Observe that the norm used in (16a) might be arbitrary and, after a slight redefinition of $\tilde{L}_{y_i}$, the following specific expression for w can be used:

$$w = (\tfrac{1}{2}\| \sum_{i\in A} y_i f_{ix}(x) \|^2_{H^{-1}} + \sum_{i\in A} y_i(\hat{x}_0(x) - f_i(x)))^{\frac{1}{2}} \quad . \tag{19b}$$

However, this coincides precisely with the minimized function in (18a) and can be interpreted as follows: given a point (y,x) and the set A, w or $(w)^2$ can be determined from (19b). By solving (18a) in y, new $\hat{y}$, $\hat{\bar{x}}$, $\hat{\bar{x}}_0$ and

$$(\hat{w})^2 = \frac{1}{2}\|\hat{g}\|^2_{H^{-1}} + \sum_{i\in A}\alpha_i\hat{y}_i = \frac{1}{2}\|\hat{\bar{x}}\|^2_H + \sum_{i\in A}\alpha_i\hat{y}_i \tag{19c}$$

$$= -\hat{\bar{x}}_0 - \frac{1}{2}\|\hat{\bar{x}}\|^2_H$$

are found. Clearly, $(w)^2 \geq (\hat{w})^2$. On the other hand, $(\hat{w})^2$ can also be interpreted as an upper bound for a new $(w)^2$, obtained after x is changed to $x + \hat{\bar{x}}$ and y is changed to $\hat{y}$ (here y does not denote the optimal Lagrange multiplier for the original problem, but only for its approximation (18b))--see Section 4. Another interpretation of $\hat{\bar{x}}_0$ and $(\hat{w})^2$ is that both approximate the gain $f(x) - f(x + \hat{\bar{x}})$ of the objective function f--$\hat{\bar{x}}_0$ is a linear approximation of this gain and $(\hat{w})^2$ a quadratic one. Clearly, the linear approximation is more optimistic than the quadratic one, but, because of convexity, the linear approximation can also give an estimation of the distance $f(x) - f(\hat{x})$ from above, thus being more useful for some algorithmic purposes; moreover, $-\hat{\bar{x}}_0$ also gives an estimation from above for the new $(w)^2$, obtained after changing x to $x + \hat{\bar{x}}$ and y to $\hat{y}$. All these properties are discussed in more detail in Section 4; the above discussion only justifies the role of quadratic approximations in nondifferentiable optimization.

3.3 Sufficient and Necessary Conditions of Optimality

The basic necessary condition of optimality for nondifferentiable problems

$$0 \in \partial f(\hat{x}) \tag{20}$$

is known to hold under various assumptions related to the definition of the subdifferential $\partial f(\hat{x})$--see Clarke (1975), Mifflin (1977), Nurminski (1973). In particular, if the functions f_i in the problem (1) are continuously differentiable, then the function f is weakly convex--see Nurminski (1973)--and

the necessary condition (20) holds with $\partial f(\hat{x})$ defined by (2). If the functions f_i are convex, then the condition (20) is also sufficient. However, even if the functions f_i are convex, the condition (20a) does not necessarily specify $\hat{x}$ uniquely. To obtain uniqueness, we must either assume strong convexity of f_i, or use the Haar condition. The Haar condition is sufficient for a unique solution of (20) even in the nonconvex case, but the requirements of the Haar condition are rather strong. In order to weaken these requirements, second-order approximations for nondifferentiable problems might be used.

The next three subsections present a more detailed discussion of the above-mentioned conditions. First, a geometric interpretation of the first-order conditions--the condition (20) and the Haar condition--is given, with an indication of possible generalizations of these conditions. Second, a second-order uniqueness condition for convex problems is derived in a natural way from the normal Lagrangian function. Third, second-order sufficient and necessary conditions for the optimality of a solution $\hat{x}$ of the problem (1) in the nonconvex case are derived from augmented Lagrangian functions.

3.4 First-Order Conditions

The condition (20) can be interpreted geometrically if we consider vectors $h_i \in R^{n+1}$ of the form:

$$h_i = (-1, f_{ix}(x)) \quad \text{for } i \in A(x) \quad ;$$
$$\hat{h}_i = (-1, f_{ix}(\hat{x})) \quad \text{for } i \in A(\hat{x}) \tag{21a}$$

We shall also use the following notation:

$$e_0 = (-1, 0) \in R^{n+1} \tag{21b}$$

where 0 denotes the zero element in R^n. Taking into account (2), it immediately follows that the necessary condition (20) can be written equivalently as

$$e_0 \in -K^* \stackrel{df}{=} \{h \in R^{n+1} : h = \sum_{i \in A(\hat{x})} \alpha_i \hat{h}_i \ , \ \alpha_i \geq 0\} \tag{21c}$$

The condition (21c) can also be used in infinite-dimensional spaces. In fact, if $f : E \to R^1$ is a subdifferentiable function defined on a linear topological space E_1 and its subdifferential $\partial f(x) \subset E^*$ is defined, where E^* is the dual space to E, then we can define $-K^* = \{h \in R^1 \times E^* : h = \alpha(-1,g) \ , g \in \partial f(x) \ , \alpha \geq 0\}$, and $e_0 \in -\hat{K}^*$ is equivalent to $0 \in \partial f(\hat{x})$. Note also that the polar cone to $-\hat{K}^*$, defined by $\hat{K} = \{k \in R^1 \times E : \langle h,k \rangle \leq 0$ for all $h \in -\hat{K}^*\}$, is a conical approximation of the epigraph of the function f at $\hat{x}$. The particular sense of this approximation depends, clearly, on the type of definition of the subdifferential we use; however, it is not the goal of this paper to pursue possible generalizations in detail (see Wierzbicki, 1972, for a discussion of similar ideas in nondifferentiable dynamic optimization). If $E = R^n$ and $-\hat{K}^*$ is given by (21c), then $\hat{K}$ is the tangent cone to the epigraph of $f(x) = \max_{i \in I} f_i(x)$ at $\hat{x}$. The geometric interpretation of the first-order necessary condition (21c) is given in Figure 1: any element of the tangent cone $\hat{K}$ must have a nonpositive scalar product with the downward pointing vector e_0.

An interpretation and generalization of the first-order Haar sufficient condition is given in the following lemma.

Lemma 2. The Haar sufficient condition for a point $\hat{x}$ satisfying (20) to be a unique (local in nonconvex case) solution of the problem (1)--usually formulated as the requirement that matrices $F_{\bar{A}}$ have maximal rank for all subsets $\bar{A}$ of $A(\hat{x})$, where $F_{\bar{A}}$ is a matrix composed of vectors $f_{ix}(\hat{x})$ for $i \in \bar{A}$--can be equivalently

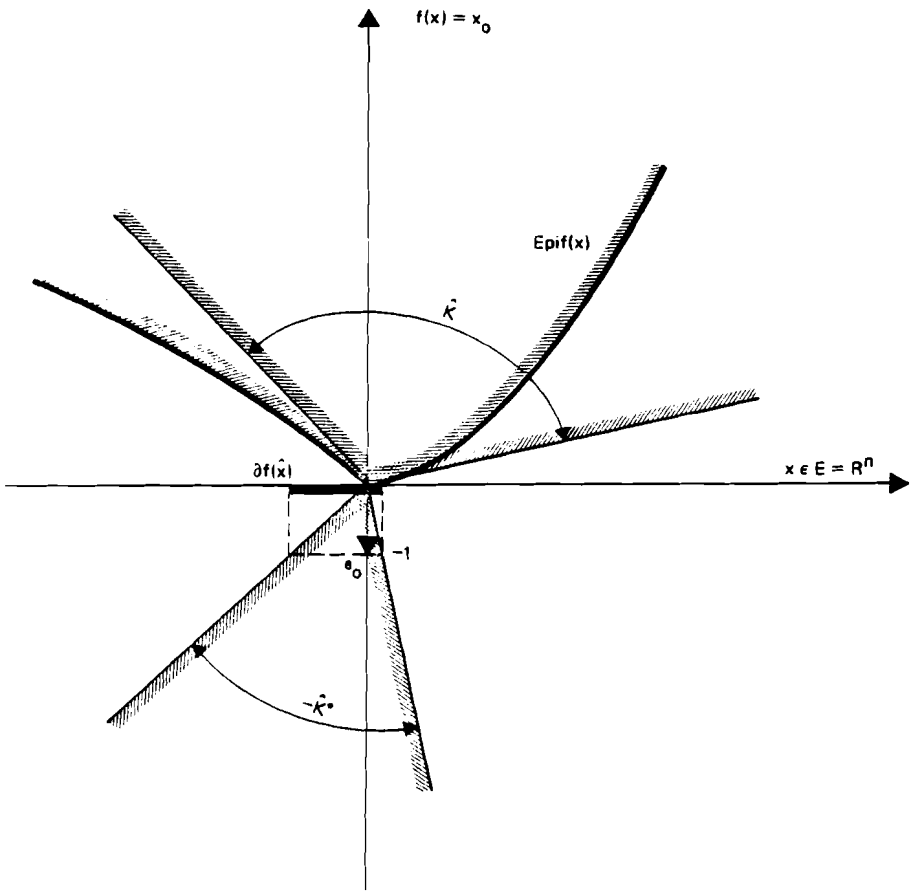

Figure 1. A geometric interpretation of the first-order necessary condition (21c).

stated in the following form:

$$e_0 \in -\hat{\dot{K}}^* \stackrel{df}{=} \{h \in -\hat{K}^* : \langle h,k\rangle < 0 \text{ for all } k \in K,\ k \neq 0\} \quad . \tag{22}$$

Before proving the lemma, observe that $\hat{\dot{K}}^*$ is defined as the quasi-interior of the cone $\hat{K}^*$, polar to the conical approximation $\hat{K}$ of the epigraph of f. This suggests that (22) might be used as a generalization of the Haar condition in a linear topological space. However, this conjecture is not proven here; we limit the lemma to the equivalence of (22) and the Haar condition in R^n. Observe that in such a case the quasi-interior is simply the interior of $\hat{K}^*$.

Proof. Note that the Haar condition implies necessarily that there are at least n + 1 elements of $A(\hat{x})$. Suppose that there are less than this, no more than n. According to (20) there exist nonzero $\hat{\lambda}_i$ such that $\sum_{i\in A(\hat{x})} \hat{\lambda}_i f_{ix}(\hat{x}) = 0$ and thus all the vectors $f_{ix}(\hat{x})$ for $i \in A(\hat{x})$ are always linearly dependent; but if no more than n vectors of n elements each are linearly dependent, we cannot form matrices of full rank from all collections of these vectors. Thus, there must be at least n + 1 vectors $f_{ix}(\hat{x})$ for $i \in A(\hat{x})$, and each collection of, say, n of them must be linearly independent. This implies in turn that at least n + 1 baricentric coordinates λ_i must be positive, since, if $|A(\hat{x})| = n + 1$, then for each j we have $\sum_{i\in A(\hat{x}),i\neq j} \lambda_i f_{ix}(\hat{x}) = -\hat{\lambda}_j f_{jx}(\hat{x}) \neq 0$ and $\hat{\lambda}_j \geq 0$ which implies $\hat{\lambda}_j > 0$. But if at least n + 1 baricentric coordinates are positive, then $e_0 = (-1,0)$ is in the interior of the cone $\hat{K}^* = \text{cone}\,(-1, f_{ix}(\hat{x}))$. Conversely, if e_0 is in the interior of $\hat{K}^*$, then at least n + 1 of vectors $f_{ix}(\hat{x})$ sum up to zero with coefficients greater than zero, and each collection of n or less of these vectors is linearly independent.

The geometrical interpretation of this reformulated Haar condition is also given in Figure 1. However, we see from Lemma 2 that the Haar condition is very restrictive: at least n + 1 functions $f_i(x)$ must be active at $\hat{x}$ and, on the basis of

the implicit function theorem, we can determine $\hat{x}$ uniquely from the set of n equations $f_i(x) = f_j(x)$, for any fixed $j \in A(\hat{x})$ and for n chosen $i \in A(\hat{x})$, $i \neq j$. However, there are often cases when $|A(\hat{x})| \leq n$; then the cone $\hat{K}^*$ does not have an interior, the Haar condition cannot be satisfied, and x must be determined on the basis of our additional information--this time using second-order conditions.

3.5 Second-Order Uniqueness Condition for the Convex Case

If all the functions f_i and thus the function f are convex, then the strong local convexity of the function f at $\hat{x}$ suffices for the uniqueness of $\hat{x}$. However, the function f is not differentiable and we cannot use the classical definition of strong convexity via second-order derivatives. On the other hand, we could use the Lagrangian function (11) for the equivalent problem (9) to characterize the strong convexity of the problem (1).

Lemma 3. If, at a solution $\hat{x}$ of the problem (1) satisfying (20) with $\partial f(\hat{x})$ defined by (2), the following matrix

$$L_{xx}(\hat{y}, \hat{x}_0, \hat{x}) = \sum_{i \in A(\hat{x})} \hat{y}_i f_{ixx}(\hat{x}) \tag{23}$$

is positive definite, where $\hat{y}_i = \hat{\lambda}_i$ are defined (not necessarily uniquely) by the baricentric coordinates of 0 in $\partial f(\hat{x})$, then $\hat{x}$ is a locally unique solution of (1). If, additionally, the functions f_i are convex, then $\hat{x}$ is the globally unique solution. Moreover, if the vectors $\hat{h}_i = (-1, f_{ix}(\hat{x}))$ for all $i \in A(\hat{x})$ are linearly independent, then the baricentric coordinates $\hat{y}_i$ are also defined uniquely.

Proof. It is known that if a Lagrangian function has a local or global minimum in primal variables while the dual variables satisfy the Kuhn-Tucker conditions, then this minimum is also a local or global solution to the primal problem. But the positive

definiteness of $L_{xx}(\hat{y},\hat{x}_0,\hat{x})$ suffices for a local or, in the convex case, global minimum of the Lagrangian function (11) in x. In x_0, the condition $\sum_{i\in A(\hat{x})}\hat{y}_i = 1$ guarantees a nonunique minimum; however, x_0 is only an auxiliary variable. As for the Kuhn-Tucker conditions, they are satisfied by the definition of $A(\hat{x})$, the definition of the baricentric coordinates $\hat{y}_i$ and by the condition (20) since $\hat{y}_i = 0$ for $i \notin A(\hat{x})$, $\hat{y}_i \geq 0$ for $i \in A(\hat{x})$, $\sum_{i\in A(\hat{x})}\hat{y}_i(f_i(\hat{x}) - \hat{x}_0) = \sum_{i\in I}\hat{y}_i(f_i(\hat{x}) - \hat{x}_0) = 0$, and $L_x(\hat{y},\hat{x}_0,\hat{x}) = \sum_{i\in A(\hat{x})}\hat{y}_i f_{ix}(\hat{x}) = 0$, $L_{x_0} = 1 - \sum_{i\in A(\hat{x})}\hat{y}_i = 0$. The uniqueness of baricentric coordinates y_i if h_i are linearly independent has been established in the proof of Lemma 1.

Observe that the requirement that the vector $(-1, f_{ix}(\hat{x})) = \hat{n}_i$ is linearly independent is much less restrictive than the Haar condition; this requirement can also be satisfied if $|A(\hat{x})| < n + 1$. On the other hand, the requirement that $L_{xx}(\hat{y},\hat{x}_0,\hat{x})$ is positive definite is rather restrictive and not really necessary: it would suffice for this Hessian matrix to be positive definite only in the subspace tangent to strongly active constraints. This, however, is related to the more general form of sufficient conditions for nonconvex problems.

3.6 Second-Order Sufficient and Necessary Conditions for the Nonconvex Case

If the functions f_i are not even locally convex, then the normal Lagrangian function (11) for the equivalent problem (9) might have no saddle-points, but only an inflexion point in x at a solution $\hat{x}$ of the problems (1), (9). However, an augmented Lagrangian function might have a saddle-point.

An augmented Lagrangian function for the problem (9) has the form

$$\Lambda(y,\rho,x_0,x) = x_0 + \frac{1}{2\rho}\sum_{i\in I}\left((f_i(x) - x_0 + \frac{y_i}{\rho})_+^2 - \frac{(y_i)^2}{\rho}\right) \tag{24}$$

where the operation $(\cdot)_+$ is defined by $(z)_+ = \max(0,z)$ and $\rho \geq 0$ is a penalty coefficient; it can be shown that with $\rho = 0$ the function (24) reduces to the normal Lagrangian function (11). The sufficient conditions for an augmented Lagrangian function of type (24) to have a saddle-point, resulting in the optimality of $\hat{x}$, were given by Rockafellar (1974, 1976). However, for the purpose of this paper, these results must be slightly modified, since the augmented Lagrangian (24) need not necessarily have a minimum in the auxiliary variable x_0.

Lemma 4. If $\hat{x} = \arg\min_{x \in X} \Lambda(\hat{y},\rho,\hat{x}_0,x)$ with $\hat{x}_0 = \max_{i \in I} f_i(\hat{x})$, $\rho > 0$, and $\hat{y} = \arg\max_{y \in R^{|I|}} \Lambda(y,\rho,\hat{x}_0,\hat{x})$, then $\hat{x}$ is an optimal solution of the problems (1), (9), that is, $\max_{i \in I} f_i(x) \geq \hat{x}_0$ for all $x \in X$.

Proof. Since Λ is a differentiable function of its arguments, the unrestricted maximization in y implies that $(f_i(x) - x_0 + \frac{y_i}{\rho})_+ = \frac{y_i}{\rho}$ for all $i \in I$, which can happen if and only if (see also Wierzbicki and Kurcyusz (1977) for possible generalizations of this equivalence) $\hat{y}_i \geq 0$, $f_i(\hat{x}) = \hat{x}_0$ for $i \in A(\hat{x})$ and $\hat{y}_i = 0$, $f_i(\hat{x}) < \hat{x}_0$ for $i \notin A(\hat{x})$. Thus, y_i are Lagrange multipliers for the problem (9). If we require, additionally, that $\Lambda_{x_0}(\hat{y},\rho,\hat{x}_0,\hat{x}) = 1 + \rho \sum_{i \in I}(f_i(x) - x_0 - \frac{y_i}{\rho})_+ = 1 - \sum_{i \in A(\hat{x})} y_i = 0$, then y_i can also be interpreted as baricentric coordinates; however, we do not need to state in the lemma that $\hat{x}_0$ minimizes Λ. Since we assume that $\hat{x}$ does minimize Λ, we obtain after obvious transformations:

$$\sum_{i \in I}(f_i(x) - x_0 + \frac{y_i}{\rho})_+^2 \geq \sum_{i \in I}(\frac{y_i}{\rho})^2 \quad \text{for all } x \in X \ . \tag{25a}$$

Suppose now that the thesis of the lemma is not true and there exists an $x \in X$ such that $f_i(x) < \hat{x}_0$ for all $i \in I$. However, this would imply that

$$(f_i(x) - \hat{x}_0 + \frac{\hat{y}_i}{\rho})_+ < \frac{\hat{y}_i}{\rho} \quad \text{for all } i \in I \tag{25b}$$

which would contradict (25a). Thus the thesis is true.

Actually, Lemma 4 can be proven under much more general assumptions. Without changing the proof, X might be any set--say, in linear topological space. The countability and finiteness of I is also not essential: I might be generalized to represent, for example, any subset of a Hilbert space, as in Wierzbicki and Kurcyusz (1977).

The way in which Lemma 4 is proven suggests the following equivalent statement of the first-order necessary condition (20) for the optimality of solutions of problems (1), (9):

Lemma 5. If $\hat{x}$ is an optimal solution of problems (1), (9), then there exist $\hat{y}_i \geq 0$, $\hat{y}_i = 0$ for $i \notin A(\hat{x}) = \{i \in I : f_i(\hat{x}) = f(\hat{x}) = \max_{j \in I} f_j(\hat{x})\}$ such that $\hat{x}$ and $\hat{y}$ are stationary points of $\Lambda(y,\rho,\hat{x}_0,x)$ with $\hat{x}_0 = \max_{i \in I} f_i(\hat{x})$ and an arbitrary $\rho > 0$. Moreover, $\hat{y}$ is a global maximal point of Λ. If, additionally, $\sum_{i \in A(\hat{x})} \hat{y}_i = 1$, then $\hat{x}_0$ is a global minimal point of Λ.

Proof. If x is an optimal solution of (1), (9), then it satisfies (20) with $\partial f(\hat{x})$ defined by (2). However, the scaling of $\hat{y}_i$ might be arbitrarily changed in (24), since ρ is an arbitrary positive number; hence we need not require that $\sum_{i \in A(\hat{x})} y_i = 1$. It was shown in the proof of Lemma 4 that $y_i \geq 0$, $f_i(\hat{x}) = \hat{x}_0$ for $i \in A(\hat{x})$ and $\hat{y}_i = 0$, $f_i(\hat{x}) < \hat{x}_0$ for $i \notin A(\hat{x})$ are equivalent to a stationary point of Λ in y; it is also a globally maximal point, since Λ is a concave function of y, see Wierzbicki and Kurcyusz (1977). The condition $0 \in \partial f(\hat{x})$ or $\sum_{i \in A(\hat{x})} \hat{y}_i f_{ix}(\hat{x}) = 0$ is equivalent to the statement that $\hat{x}$ is a stationary point of Λ, since $\Lambda_x(\hat{y},\rho,\hat{x}_0,\hat{x}) = \rho \sum_{i \in I} (f_i(\hat{x}) - \hat{x}_0 + \frac{\hat{y}_i}{\rho})_+ f_{ix}(\hat{x}) = \sum_{i \in A(x)} \hat{y}_i f_{ix}(\hat{x}) = 0$. If $\sum_{i \in A(\hat{x})} \hat{y}_i = 1$, then also $\Lambda_{x_0}(\hat{y},\rho,\hat{x}_0,\hat{x}) = 0$; since Λ is convex in x_0, this implies that $\hat{x}_0$ is a global minimal point.

The conclusions of Lemmas 4 and 5 might be summarized as follows: although an optimal solution x of problems (1), (9) is a stationary point of the augmented Lagrangian function (24), the corresponding multipliers $\hat{y}_i$ maximize this function (without any constraints--the nonnegativeness of $\hat{y}_i$ results from this maximization) and, if $\hat{y}_i$ are interpreted as baricentric coordinates through appropriate scaling, $\hat{x}_0$ also minimizes this function; we are not generally certain that $\hat{x}$ minimizes this function. If it does, this is also a sufficient condition for optimality.

From Lemma 4, a sufficient condition for a local solution of a nonconvex problem of the type (1) can be derived:

Theorem 1. Suppose that at a given $\hat{x} \in \text{int } X$ there exist $\hat{y}_i \geq 0$, $\hat{y}_i = 0$ for $i \notin A(\hat{x})$, such that $\sum_{i \in A(\hat{x})} \hat{y}_i f_{ix}(\hat{x}) = 0$, where $A(\hat{x}) = \{i \in I : f_i(\hat{x}) = f(\hat{x}) = \max_{j \in I} f_j(\hat{x})\}$. If, for some $\rho > 0$, the following matrix is positive definite:

$$\Lambda^S_{xx} = \sum_{i \in A(\hat{x})} \hat{y}_i f_{ixx}(\hat{x}) + \rho \sum_{i \in S(\hat{y})} f^*_{ix}(\hat{x}) f_{ix}(\hat{x}) \tag{26}$$

where $S(\hat{y}) = \{i \in A(\hat{x}) : \hat{y}_i > 0\}$ and $f^*_{ix}(\hat{x})$ denotes the gradient of f_i in column form, then $\hat{x}$ is a local solution of problems (1), (9). With $\hat{x}_0 = \max_{i \in I} f_i(\hat{x})$, if $\sum_{i \in A(\hat{x})} \hat{y}_i = 1$, then $\hat{y}$ and $(\hat{x}_0, \hat{x})$ correspond to a saddle-point of the function (26).

Proof. Consider the function

$$\Lambda^S(y, \rho, x_0, x) = x_0 + \frac{1}{2\rho} \sum_{i \in S(\hat{y})} (f_i(x) - x_0 + \frac{y_i}{\rho})^2 - \frac{1}{2\rho} \sum_{i \in A(\hat{x})} (\frac{y_i}{\rho})^2 \tag{27a}$$

It has been proven in Wierzbicki (1978) that this function is a local quadratic approximation from below of the function (24), that is, there exists a neighborhood $U(\hat{y},\hat{x}_0,\hat{x})$ of $(\hat{y},\hat{x}_0,\hat{x})$ such that

$$\Lambda^S(y,\rho,x_0,x) \leq \Lambda(y,\rho,x_0,x) \quad \text{for all } (y,x_0,x) \in U(\hat{y},\hat{x}_0,\hat{x}) \tag{27b}$$

where the inequality can be replaced by equality if $A(\hat{x}) = S(\hat{y})$.

Here $\hat{y},\hat{x}_0,\hat{x}$ denote a stationary point of the function (24). If we choose $\hat{x}_0 = \max_{i\in I} f_i(\hat{x})$ then, under the assumptions of the theorem, $\hat{y}$ and $\hat{x}$ indeed correspond to a stationary point, while $\hat{y}$ is a global maximum point--see the proof of Lemma 5. However, we do not necessarily require that $\hat{y}_i$ are scaled to $\sum_{i\in A(\hat{x})}\hat{y}_i = 1$, and $\hat{x}_0$ is not necessarily a stationary point; due to the particular form of (24), (27a), the inequality (27b) also holds for an arbitrarily fixed $\hat{x}_0$.

While evaluating the Hessian matrix of Λ^S in x at $\hat{y},\rho,\hat{x}_0,\hat{x}$, we obtain Λ^S_{xx} as given in (26). If it is positive definite, then Λ^S has a unique local minimum in x at $\hat{x}$, with $y,\rho,\hat{x}_0$ fixed (it is easy to check that $\hat{x}$ is also a stationary point of Λ^S). With fixed $\hat{y},\rho,\hat{x}_0$ the inequality (27b) implies that the function Λ also has a unique local minimum in x at $\hat{x}$. Hence, Lemma 4 can be applied with X replaced by a neighborhood of $\hat{x}$, and we conclude that $\hat{x}$ is a locally optimal solution of problems (1), (9). Observe that the uniqueness of the minimum of Λ in x with fixed $y = \hat{y}$ does not necessarily imply the uniqueness of x as a solution of problems (1), (9). If we scale $\hat{y}_i$ to $\sum_{i\in A(\hat{x})}\hat{y}_i = 1$, then Lemma 5 implies that $\hat{y}$ and $(\hat{x}_0,\hat{x})$ represent a saddle-point of Λ. Even in this case, $\hat{y}_i$ might be nonunique if $S(\hat{y}) \neq A(\hat{x})$, or if the vectors $\hat{n}_i = (-1, f_{ix}(\hat{x}))$ for $i \in A(\hat{x})$ are linearly dependent.

Theorem 1 is in fact only a slight modification and adaptation to the particular problem (1) of the results given by Rockafellar (1974, 1976); similarly, his results on the second-order necessary conditions of optimality can be adapted to obtain:

Theorem 2. If $\hat{x}$ is an optimal solution for problems (1), (9) satisfying (20) with $\partial f(\hat{x})$ given by (2), then, for any $\rho > 0$, the following matrix

$$\Lambda_{xx}^{A} = \sum_{i \in A(\hat{x})} \hat{y}_i f_{ixx}(\hat{x}) + \rho \sum_{i \in A(\hat{x})} f_{ix}^{*}(\hat{x}) f_{ix}(\hat{x}) \tag{28}$$

is positive semidefinite.

Proof. Following Rockafellar (1974, 1976) we conclude that for the optimality of $\hat{x}$ it is necessary that the following Hessian matrix

$$\Lambda_{\tilde{x}\tilde{x}}^{A} = \hat{L}_{\tilde{x}\tilde{x}} + \rho \sum_{i \in A(\hat{x})} \hat{h}_i^{*} \hat{h}_i \tag{29a}$$

be positive semidefinite, where $\tilde{x} = (x_0, x)$, $\hat{h}_i = (-1, f_{ix}(\hat{x}))$ and

$$\hat{L}_{\tilde{x}\tilde{x}} = \begin{bmatrix} 0 & 0 \\ 0 & \hat{L}_{xx} \end{bmatrix}, \quad \hat{L}_{xx} = \sum_{i \in A(\hat{x})} \hat{y}_i f_{ixx}(\hat{x}) \tag{29b}$$

Now compute a quadratic form $\langle (\bar{x}_0, \bar{x}), \Lambda_{\tilde{x}\tilde{x}}^{A}(\bar{x}_0, \bar{x}) \rangle$ to obtain

$$\langle(\bar{x}_0,\bar{x}) , \Lambda^{A}_{\tilde{x}\tilde{x}}(\bar{x}_0,\bar{x})\rangle = \langle\bar{x} , \hat{L}_{xx}\bar{x}\rangle + \rho \sum_{i\in A(\hat{x})} (f_{ix}(\hat{x})\bar{x} - \bar{x}_0)^2 \tag{29c}$$

This expression must be nonnegative for all $(\bar{x}_0,\bar{x})$, in particular if $\bar{x}_0 = 0$; but if $\bar{x}_0 = 0$, then

$$\langle(0,\bar{x}) , \Lambda^{A}_{\tilde{x}\tilde{x}}(0,\bar{x})\rangle = \langle\bar{x} , \Lambda_{xx}\bar{x}\rangle \tag{29d}$$

Thus, the positive semidefiniteness of Λ^{A}_{xx} is necessary. Theorem 2 also has the following interpretation. As shown in Wierzbicki (1978), the function:

$$\Lambda^{A}(y,\rho,x_0,x) = x_0 + \tfrac{1}{2}\rho \sum_{i\in A(\hat{x})} (f_i(x) - x_0 + \frac{y_i}{\rho})^2 - \tfrac{1}{2}\rho \sum_{i\in A(\hat{x})} (\frac{y_i}{\rho})^2 \tag{30a}$$

is an upper bound for the function (24) in a neighborhood $U(\hat{y},\hat{x}_0,\hat{x})$ of $(\hat{y},\hat{x}_0,\hat{x})$:

$$\Lambda(y,\rho,x_0,x) \leq \Lambda^{A}(y,\rho,x) \quad \text{for all } (y,x_0,x) \in U(\hat{y},\hat{x}_0,\hat{x}) \tag{30b}$$

Thus, the upper-bound function Λ^{A} must satisfy the second-order necessary condition for a minimum in x at $\hat{x}$ if $\hat{x}$ is optimal.

4. A QUASI-NEWTON METHOD FOR NONDIFFERENTIABLE OPTIMIZATION WITH EXPLICIT SUBDIFFERENTIALS

The possibility of constructing quadratic approximations to Lagrangian functions for the problems (1), (9), in both the convex and nonconvex cases, justifies the analysis of quasi-Newton methods for nondifferentiable optimization. A specific algorithm of this class, based on a corresponding algorithm for differentiable optimization described in Wierzbicki (1978), is presented in this section. For the sake of clarity, only the convex case is described in detail, although subsection 4.7 implies the possibility of extending this algorithm to the nonconvex case.

4.1 Estimation of the Distance from the Optimal Solution

Most algorithms of nonlinear or nondifferentiable optimization produce a sequence $\{x^k,y^k\}$ of approximations of the primal and dual solution $(\hat{x},\hat{y})$. Some measure of the distance of (x^k,y^k) from $(\hat{x},\hat{y})$ is explicitly or implicitly used. Here, we shall use the variable $w^k = \|(L_x^k,\tilde{L}_{y_k}^k\|$ with $L_x^k = \sum_{i\in A^k} y_i^k f_i(x^k)$ and $\tilde{L}_{y_i}^k = y_i^k(f_i(x^k) - \hat{x}_0(x^k))$, $\hat{x}_0(x^k) = \max_{i\in I} f_i(x^k)$--see equations (16a, b)--for this purpose. This is justified by the following lemma.

Lemma 6 (Wierzbicki, 1978). Suppose $\hat{x}$ is an optimal solution of problem (9) with convex functions f_i. Let $\hat{x} \in \text{int } X$, and let $\hat{y}$ be the corresponding vector of Lagrange multipliers, with $\sum_{i\in I}\hat{y}_i = 1$. Suppose that the vectors $\hat{h}_i = (-1,f_{ix}(\hat{x}))$ are linearly independent for $i \in A(\hat{x})$--hence y is unique--and let the matrix $\hat{L}_{xx} = \sum_{i\in S(\hat{y})}\hat{y}_i f_{ixx}(\hat{x})$ be positive definite--hence, $\hat{x}$ is unique. Then there exists a neighborhood $U(\hat{y},\hat{x})$ and a constant $\rho > 0$ such that

$$\|(y^k - \hat{y}\,,\, x^k - \hat{x})\| \leq \delta \cdot w^k \quad \text{for all } (y^k,x^k) \in U(\hat{y},\hat{x}) \tag{31}$$

where w^k is defined as in the preceding paragraph, with an arbitrarily chosen norm, and with $A^k \supset A(x^k)$.

In particular, we can use the norm given in equation (19b) to obtain w^k corresponding to the minimal function in (18a).

4.2 Approximations of Activity Sets

Consider now the situation when (y^k, x^k) are given elements of a sequence $\{y^k, x^k\}_0^\infty$. Denote by the upper index k all values of functions evaluated at (y^k, x^k) with $\hat{x}_0(x^k) = x_0^k$, etc. Denote by A^k the approximation of the set $A(\hat{x})$ as evaluated at (y^k, x^k) and by S^k the approximation of the set $S(\hat{y})$. If (y^k, x^k) converges to $(\hat{y}, \hat{x})$ and w^k converges to zero, then the following formulae for A^k, S^k can be used:

$$A^k = \{i \in I : f_i^k - x_0^k + \frac{1}{\rho} y_i^k \geq -\eta_f^k\} \tag{32a}$$

$$S^k = \{i \in A^k : y_i^k \geq \eta_y^k\} \tag{32b}$$

where $\rho > 0$ is a chosen constant, depending on the scaling of the problem; (clearly $y_i^k \in [0, 1]$ but $f(x^k) = x_0^k$ can have arbitrary scaling), and where $\eta_f^k > 0$, $\eta_y^k > 0$ and η_f^k, η_y^k converge to zero but more slowly than w^k. For example, formulae of the following type may be used:

$$\eta_f^k = \xi_f (w^{k-1})^{\frac{1}{2}} \quad ; \quad \eta_y^k = \min\,(0.01\,,\, \xi_y (w^{k-1})^{\frac{1}{2}}) \tag{32c}$$

where ξ_f, ξ_y are chosen constants; again, the best choice of these constants depends on the scaling of the problem, that is, on Lipschitz constants for functions f_i or on the norms of gradients f_{ix}. But the assumption that η_f^k, η_y^k converge to zero more slowly than w^k implies the desired result $S^k = S(\hat{y})$, $A^k = A(\hat{x})$ for sufficiently small w^k even if the Lipschitz constants are not known explicitly. This follows from the following lemma:

Lemma 7 (Wierzbicki, 1978). Suppose $\hat{x}$ is a unique solution of problem (9) and $\hat{y}$ the corresponding unique Lagrange multiplier. Let the sets $A(\hat{x})$, $S(\hat{y})$ be defined by (10), (14) and A^k, S^k by (32c, b) with $\lim_{k\to\infty}(w^k/\eta_f^k) = \lim_{k\to\infty}(w^k/\eta_y^k) = 0$, where w^k is defined by (16a, b). Then there exists a number $\tilde{w} > 0$ such that

$$A^k = A(\hat{x}) \quad , \quad S^k = S(\hat{y}) \tag{33}$$

for all $(y^k, x^k) \in U(\hat{y}, \hat{x}) = \{(y^k, x^k) : w^k < \tilde{w}\}$.

However, the above results are valid independently of the norm used when defining w^k. If the norm (19b) is used and $-\hat{\hat{x}}_0^{k-1}$ approximates w^k from above, a more useful expression than (32c) can be obtained. Suppose the range of f, denoted by Rf, can be estimated. Then, after some heuristic reasoning, assuming that the initial $|\hat{\hat{x}}_0^0| = Rf$, $\eta_f^0 = 10^{-2}Rf$ and $\eta_y^0 = 10^{-2}$ and expecting the final accuracy to be related to $|\hat{\hat{x}}_0^k|$ of the order $10^{-6}Rf$, the following expressions:

$$\eta_f^k = \xi_f |\hat{\hat{x}}_0^{k-1}|^{\frac{1}{2}} \quad ; \quad \xi_f = 10^{-2}(Rf)^{\frac{1}{2}} \quad ;$$
$$\eta_y^k = \xi_y |\hat{\hat{x}}_0^{k-1}|^{\frac{1}{2}} \quad ; \quad \xi_y = 10^{-2}(Rf)^{-\frac{1}{2}} \quad ; \quad \rho = \frac{1}{Rf} \tag{32d}$$

satisfy the assumptions and result in $\eta_f^k = 10^{-4}Rf$, $\eta_y^k = 10^{-4}$ if $|\hat{\hat{x}}_0^{k-1}| = 10^{-6}Rf$. This means that a function f_i such that $f(\hat{x}) - f_i(\hat{x}) < 10^{-4}Rf$ might still be included in the probably active set A^k and a Lagrange multiplier with $\hat{y} < 10^{-4}$ might be excluded from the strongly active set S^k. However, this can be considered as an acceptable risk--particularly since it will be shown later that the exact estimation of activity (33) does not influence the simple convergence of algorithms and is needed only when establishing superlinear or quadratic convergence.

4.3 A Quadratic Approximation Algorithm for Nondifferentiable Optimization With Explicit Subdifferentials

The algorithm minimizes a function $f(x) = \max_{i \in I} f_i(x)$ for $x \in R^n$, where a minimal point $\hat{x}$ is supposed to exist (a modification for the case $x \in X$ where X is a compact convex set is possible but not described here). The functions f_i are assumed to be convex and twice-differentiable. It is also assumed that the values $f(x^k) = f^k$, $f_i(x^k) = f_i^k$, $f_{ix}(x^k) = f_{ix}^k$ can be computed for i in any subset of I. The algorithm is based on quadratic approximations (18a, b) to the Lagrangian function (11). Subroutines for a variable metric approximation of the Hessian matrix of this function (discussed in Section 4.5) and for a directional search (described, for example, in Appendix 1) are assumed to be available.

Step 0. Choose parameters x^1 - initial guess of the solution, supplied by the user; Rf - estimated range of the function values, supplied by the user; ε_{ff} - final accuracy of function values, supplied by the user (suggested $\varepsilon_{ff} = 10^{-6}$Rf); $\gamma \in (0\ ;\ 1)$ - desired rate of convergence of gradient values (suggested $\gamma = 0.1$); $m_a \in (0\ ;\ 0.5)$, $m_b \in (0.5\ ;\ 1)$ - linear search parameters (suggested $m_a = 0.3$, $m_b = 0.7$); H^1 - initial approximation of the Hessian (suggested $H^1 = I$). Set $\hat{\hat{x}}_0^0 = Rf$, $y_i^1 = \frac{1}{|I|}$, $i \in I$, $k = 1$.

Step 1. Compute η_f^k, η_y^k from (32). Compute f^k and f_i^k for $i \in I$ and determine the sets A^k and S^k (32a, b), saving only f_i^k for $i \in A^k$. Compute f_{ix}^k and $\alpha_i = f^k - f_i^k$ for $i \in A^k$. Set $y_i^k = 0$ for $i \notin A^k$, rescale proportionally the remaining y_i^k to obtain $\sum_{i \in A^k} y_i^k = 1$. Compute w^k (19b). If $(w^k)^2 < |\hat{\hat{x}}_0^{k-1}| < \varepsilon_{ff}$, stop. If $k > 1$, update H^k.

Step 2. Solve the problem (18a) to obtain $\hat{y}^k$, $\hat{g}^k$, compute $\hat{\hat{x}}^k$, $\hat{\hat{x}}_0^k$ from Lemma 1 and $\hat{w}^k$ from (19c).

Step 3. Set $(\tau^k = 1)\tilde{x}^k = x^k + \hat{\hat{x}}^k$. If $|\hat{\hat{x}}_0^k| \leq \gamma |\hat{\hat{x}}_0^{k-1}| \leq \gamma^4 Rf$ and $w^k \leq \gamma w^{k-1}$ is not satisfied, compute $\tilde{f}^k = f(\tilde{x}^k)$. If either

$$f^k + m_a \hat{\hat{x}}_0^k \geq \tilde{f}^k \geq f^k + m_b \hat{\hat{x}}_0^k \tag{34a}$$

or $|\hat{\tilde{x}}_0^k| \leq \gamma|\hat{x}_0^{k-1}| \leq \gamma^4 Rf$ and $w^k \leq \gamma w^{k-1}$, set $x^{k+1} = \tilde{x}^k$, $y^{k+1} = \tilde{y}^k$, $k := k + 1$, and go to Step 1.

Step 4. Perform a linear search for τ^k such that:

$$f^k + m_a\tau^k\hat{x}_0^k \geq f(x^k + \tau^k\hat{x}_0^k) \geq f^k + m_b\tau^k\hat{x}_0^k \qquad (34b)$$

(or any other $\tau^{k'}$ resulting in $f(x^k + \tau^{k'}\hat{x}_0^k) < f(x^k + \tau^k\hat{x}^k)$, where τ^k satisfies (34b), see Appendix 1). Set $x^{k+1} = x^k + \tau^k\hat{x}^k$, $y^{k+1} = y^k + \tau^k(\hat{y}^k - y^k)$, $k := k + 1$, and go to Step 1.

Comments. Observe that all f_i^k for $i \in I$ must be evaluated when computing f^k. It is best to combine this with the determination of sets A^k, S^k, saving only f_i^k for $i \in A^k$. But it is not known whether $\tau^k = 1$ will be accepted when checking condition (34a). Therefore, if $|\hat{x}_0^k|$ is already small enough and decreases and the desired convergence rate γ for w^k is attained, $\tau^k = 1$ is accepted without checking. In fact, w^k is computed only for this purpose--and to double-check the stopping test. Other redundant information, such as the sets S^k, the values $\hat{g}^k$, $\hat{w}^k$, or even the rescaled values y_i^k, need not be computed if the computation of w^k were deemed unnecessary. However, this information is valuable in analyzing the algorithm and in possible debugging.

A full analysis of the simple convergence of the algorithm is omitted here, since the proof of the following theorem can be easily derived from results given in Lemaréchal (1978), Szymanowski (1977), or Wierzbicki (1978). It is only necessary to note that $(w^k)^2 < |\hat{x}_0^{k-1}|$ will eventually be satisfied if $|\hat{x}_0^k|$ converges to zero (see Section 4.4), and that $w^k \leq \gamma w^{k-1}$ implies convergence if $|\hat{x}_0^k|$ is small enough and decreases. The double-check in Step 3 is also redundant, since the linear convergence of $|\hat{x}_0^k|$ alone implies convergence of the algorithm in the convex case; but the algorithm is also designed to be used in cases of only local convexity.

Theorem 3. Suppose $\hat{x}$ is the unique minimizing point of $f(x) = \max_{i \in I} f_i(x)$, where f_i are twice-differentiable functions, and let the vectors $\hat{h}_i = (-1, f_{ix}(\hat{x})) \in R^{n+1}$ be linearly independent for $i \in A(\hat{x}) = \{i \in I : f_i(\hat{x}) = f(\hat{x})\}$. This implies that the corresponding Lagrange multiplier vector $\hat{y}$, $\hat{y}_i \geq 0$ for $i \in A(\hat{x})$, $\hat{y}_i = 0$ for $i \notin A(\hat{x})$ and $\sum_{i \in A(\hat{x})} \hat{y}_i = 1$, is also unique. Let $\hat{L}_{xx} = \sum_{i \in A(\hat{x})} \hat{y}_i f_{ixx}(\hat{x})$ be positive definite. Let $U(\hat{x})$ be a neighborhood of point $\hat{x}$ such that the (not necessarily convex) function f has no generalized subdifferentials containing zero other than at point $x = \hat{x}$; if f is convex, let $U(\hat{x}) = R^n$. Let the matrices H^k be uniformly positive definite. Then, for any $x^1 \in U(\hat{x})$, the sequence (y^k, x^k) generated by the above algorithm with $\varepsilon_{ff} = 0$ converges to the point $(\hat{y}, \hat{x})$.

To prove the theorem, combine the results given, for example, in Lemaréchal (1975) and Wierzbicki (1978).

4.4 Properties of Quadratic Approximations to Lagrange Functions

Two basic properties of quadratic approximation problems (18a, b) are important for the superlinear or quadratic convergence of the above algorithm:

Lemma 8. Let the assumptions of Theorem 3 and Lemma 3 hold. Then there exists a neighborhood $U(\hat{y}, \hat{x})$ of $(\hat{y}, \hat{x})$ and a number $\beta > 0$ such that, for any $(y^k, x^k) \in U(\hat{y}, \hat{x})$, problems (18a) $\rightarrow$ (18b) have solutions with $\hat{\hat{x}}^k$, $\hat{y}^k = y^k + \hat{\hat{y}}^k$ satisfying the following inequality:

$$\| \hat{\hat{y}}^k , \hat{\hat{x}}^k \| \leq \beta w^k \tag{35}$$

where w^k is defined as in (16a, b) with any norm, for example, the norm (19b).

For a general proof of the lemma, see, e.g., Wierzbicki (1978); when using the norm (19b) for w^k the proof becomes quite straightforward.

Lemma 9. Let the assumptions of Theorem 3 and Lemma 3 hold. Suppose the solutions of problems (18a) ↔ (18b) define $x^{k+1} = x^k + \hat{\bar{x}}^k$, $y^{k+1} = y^k + \hat{y}^k$. Then w^{k+1} defined as in (16a, b) with any norm satisfies the following inequality at the point (y^{k+1}, x^{k+1}):

$$w^{k+1} \leq \| (H^k - L^k_{xx})\hat{\bar{x}}^k \| + o(\hat{\bar{y}}^k, \hat{\bar{x}}^k) \tag{36}$$

where $L^k_{xx} = \sum_{i \in A^k} y^k_i f_{ixx}(x^k)$ and $o(z)$ denotes a function with the property that $\lim_{\|z\| \to 0} o(z)/\|z\| = 0$.

For a general proof of the lemma, see, e.g., Szymanowski (1977) and Wierzbicki (1978); again, the proof can be simplified by considering the particular norm (19b) for w^{k+1}.

Many further conclusions can be drawn from a more detailed analysis of Lemmas 1, 8, 9 using the specific norm (19b) for w. For example, the general relation (36) can be transformed to:

$$(w^{k+1})^2 \leq -\hat{\bar{x}}^k + (\| (H^k - L^k_{xx})\hat{\bar{x}}^k \|_{(H^k)^{-1}} - \| H^k \hat{\bar{x}}^k \|^2_{(H^k)^{-1}} + o^2(\hat{\bar{x}}^k, \hat{\bar{y}}^k) \tag{37}$$

which indicates that, for (y^k, x^k) in a neighborhood of $(\hat{y}, \hat{x})$ and for the norm of $(H^k - L^k_{xx})\hat{\bar{x}}^k$ small enough when compared to the norm of $H^k \hat{\bar{x}}^k$, the inequality $(w^{k+1})^2 \leq -\hat{\bar{x}}^k_0$ holds. More generally, Lemma 9 indicates that the norm of $(H^k - L^k_{xx})\hat{\bar{x}}^k$ is responsible for the speed of convergence of quadratic approximation algorithms.

4.5 Properties of Variable Metric Approximations

A variable metric H^k should approximate the Hessian matrix

$$\hat{L}_{xx} = L_{xx}(\hat{y}, \hat{x}_0, \hat{x}) = \sum_{i \in A(\hat{x})} \hat{y}_i f_{ixx}(\hat{x}) \quad .$$

Since A(x) changes in every neighborhood of $\hat{x}$, it is necessary to define sets A^k with the property $A^k = A(\hat{x})$ even if evaluated at (y^k,x^k) in a neighborhood of $(\hat{y},\hat{x})$. If the following (matrix-valued) function is defined:

$$\tilde{L}^k_{xx} = \tilde{L}_{xx}(\hat{y}^k,x^k) = \sum_{i\in A^k} \hat{y}^k_i f^k_{ixx} \tag{38a}$$

then this function is continuous in $(\hat{y}^k,x^k)$ and $\tilde{L}_{xx}(\hat{y},\hat{x}) = \hat{L}_{xx}$; moreover, it can be shown that $\tilde{L}^k_{xx}$ can be used in Lemma 9 instead of L^k_{xx}. It is this matrix $\tilde{L}^k_{xx}$ that can be approximated by a variable metric technique.

A typical variable metric approximation of the $(n \times n)$ matrix $\tilde{L}^k_{xx}$ is based on a set of data $\{s^j, r^j\}^k_{j=k-N+1}$ such that:

$$\tilde{L}^k_{xx} s^j = r^j + o(s^j,\hat{\bar{y}}^j,\ldots,s^k,\hat{\bar{y}}^k) \tag{38b}$$

where $o(\cdot)$ is a function converging to zero faster than the norm of its arguments. The number of data varies; clearly $N \geq n$ is required for a sensible approximation. The data s^j, r^j related to the function $\tilde{L}_{xx}$ can be defined by

$$s^j = x^j - x^{j-1} (= \hat{\bar{x}}^{j-1} \text{ , if } \tau^{j-1} = 1) \tag{38c}$$

$$r^j = \sum_{i\in A^k} \hat{y}^j_i (f^{j*}_{ix} - f^{(j-1)*}_{ix}) \quad . \tag{38d}$$

Observe that $r^j \neq \hat{g}^j - \hat{g}^{j-1} = r^j + \sum_{i\in A^k} \hat{\bar{y}}^j f^{(j-1)*}_{ix}$; if $\hat{g}^j - \hat{g}^{j-1}$ were used instead of r^j, the requirements (38b) could not be satisfied, since the difference between them converges to zero only as fast as $\hat{\bar{y}}^j$. The matrix H^k approximating $\tilde{L}^k_{xx}$ is now constructed in a way that guarantees that:

$$H^k s^k = r^k \tag{39a}$$

$$H^k s^j = r^j + o(s^j, \hat{\tilde{y}}^j, \ldots, s^{k-1}, \hat{\tilde{y}}^{k-1}) \quad , \qquad j < k \qquad (39b)$$

under various additional assumptions. In the most widely used rank-two variable metric procedures, an increasingly accurate directional search resulting in almost conjugate subsequent directions of search is needed to guarantee (39b). If a rank-one variable metric procedure is used, relations (39a, b) are independent of the step-size coefficients and of the choice of directions; on the other hand, a rank-one variable metric approximation H^k may not be positive definite even if $\tilde{L}^k_{xx}$ are positive definite. However, there are special variants of the rank-one variable metric that guarantee that H^k will be positive definite (Kreglewski, 1977).

If $N \geq n$ and the data $\{s^j\}^k_{k-N+1}$ span R^n, then it can be shown (Kreglewski, 1977) that the relations (38b) and (39a, b) imply together that

$$(\tilde{L}^k_{xx} - H^k) s^{k+1} = o(\underset{\sim}{s}^k, \hat{\underset{\sim}{\tilde{y}}}^k) \quad ; \qquad (40)$$

$$\underset{\sim}{s}^k = (s^{k+1}, s^k, \ldots, s^{k-N+1}) \quad ; \quad \hat{\underset{\sim}{\tilde{y}}}^k = (\hat{\tilde{y}}^k, \ldots, \hat{\tilde{y}}^{k-N+1}) \quad .$$

If $s^j = \hat{\tilde{x}}^{j-1}$, then the estimate (40) together with (36) from Lemma 9 results in the superlinear convergence of a quadratic approximation method (see next section). Note, however, that estimate (40) does not imply (although it is implied by) $\lim_{k\to 0} \|\tilde{L}^k_{xx} - H^k\| = 0$; only rather special types of variable metric procedures approximate $\tilde{L}_{xx}$ in the norm. This is why the quadratic convergence of a quasi-Newton method can be obtained in practice only when $H^k = \tilde{L}^k_{xx}$ is computed explicitly.

4.6 Superlinear and Quadratic Convergence of Quadratic Approximation Methods

Lemmas 8 and 9, together with the properties of variable metric H^k, result in the following theorem:

Theorem 4. Let the assumptions of Theorem 3 and Lemma 7 hold. Then, for any desired convergence rate $\gamma \in (0;1)$, there exists a number $\xi = \xi(\gamma) > 0$ and a neighborhood $U(\hat{y},\hat{x})$ of $(\hat{y},\hat{x})$ such that, if $(y^k,x^k) \in U(\hat{y},\hat{x})$ and $\|(\tilde{L}_{xx}^k - H^k)\hat{\hat{x}}^k\| \leq \xi w^k$, then $w^{k+1} < \gamma w^k$ and $|\hat{\hat{x}}_0^{k+1}| < \gamma|\hat{\hat{x}}_0^k|$ and the algorithm from Section 4.3 converges at the desired rate. If

$$\lim_{k\to\infty} \frac{\|(\tilde{L}_{xx}^k - H^k)\hat{\hat{x}}^k\|}{w^k} = 0 \quad ,$$

then the algorithm converges superlinearly, $\lim_{k\to\infty} (w^{k+1}/w^k) = 0$. If $\tilde{L}_{xx}^k = H^k$ and the second-order derivatives $f_{ixx}(\cdot)$, $i \in A(\hat{x})$, are Lipschitz-continuous, then the algorithm converges quadratically, $\lim_{k\to\infty} \sup (w^{k+1}/(w^k)^2) = \bar{a} < +\infty$.

The proof of the theorem is quite standard--see, for example, the proof of Theorem 1 in Wierzbicki (1978)--and is omitted here.

It is worth noting that practical experience with quadratic approximation methods shows that they are the most efficient algorithms for constrained differentiable optimization (Szymanowski, 1977). A similar performance might be expected from the algorithm given in Section 4.3, since it is only an adaptation of quadratic approximation methods to the special class of nondifferentiable problems. Moreover, the author's attention was recently drawn to a paper (Madsen and Schjaer-Jacobsen, 1977) describing an algorithm similar in nature--though different in many details and in the theoretical justification--to that of Section 4.3, for the same class of problems; the results of numerical tests given by Madsen and Schjaer-Jacobsen confirm that the algorithm given in Section 4.3 should be very efficient in practice.

4.7 Nonconvex Nondifferentiable Optimization With Explicitly Given Subdifferentials

Following the results given in Section 3.7, it is possible to derive a quadratic approximation algorithm extending the algorithm from Section 4.3 to even the locally nonconvex case.

The algorithm uses the sets S^k (defined redundantly in algorithm 4.3) in order to determine convexifying terms for the quadratic approximation problem (18b), which now takes the form:

$$\underset{(\bar{x}_0^k,\,\bar{x}^k)\in\bar{X}_0^k}{\text{minimize}} \; (\bar{x}_0^k + \tfrac{1}{2}\langle \bar{x}^k, H^k\bar{x}^k\rangle + \rho \sum_{i\in S^k} (\tfrac{1}{2}(f_{ix}^k \bar{x}^k - \bar{x}_0^k)^2 - \alpha_i^k (f_{ix}^k \bar{x}^k - \bar{x}_0^k))) \tag{41a}$$

or, equivalently:

$$\underset{(\bar{x}_0^k,\,\bar{x}^k)\in\bar{X}_0^k}{\text{minimize}} \; (\bar{x}_0^k(1 + \rho \sum_{i\in S^k} \alpha_i^k) + \tfrac{1}{2}\rho|S^k|(x_0^k)^2 + \tfrac{1}{2}\langle \bar{x}^k, (H^k + \rho F^{k*}F^k)\bar{x}^k\rangle + \rho \sum_{i\in S^k} (\alpha_i^k + \bar{x}_0^k) f_{ix}^k \bar{x}^k) \tag{41b}$$

where F^k is a matrix composed of elements f_{ix}^k for $i \in S^k$, $\alpha_i^k = f^k - f_i^k$, $|S^k|$ is the number of elements in S^k, and

$$\bar{X}_0^k = \{(\bar{x}_0^k, \bar{x}^k) \in R^{n+1} : f_{ix}^k \bar{x}^k - \bar{x}_0^k - \alpha_i^k \leq 0 \;,\; i \in A^k\} \;. \tag{41c}$$

It is interesting to note that, if $S^k = A^k$ and all constraints are active for a solution of (41), the problem is fully equivalent to a dual problem as in Lemma 1; in all other cases the dual problem for (41) is more complicated, but might lead to interesting results. A quadratic approximation algorithm requires a variable metric approximation either of the matrix $H^k \approx \sum_{i\in A^k} y_i^k f_{ixx}^k$, or of the matrix $H^k + \rho F^{k*}F^k$; the latter is positive definite, if the second-order sufficient condition of optimality is satisfied. Under this assumption, the superlinear convergence of the algorithm can also be proved for the nonconvex case by a modification of results given in Wierzbicki (1978).

4.8 Nondifferentiable Optimization With Implicitly Given Subdifferentials

A large number of algorithms has been proposed for the more general class of nondifferentiable optimization problems in which $\partial f(x)$ are not given explicitly and it is possible to compute only function values $f(x)$ and subgradients $g \in \partial f(x)$ without any more specific knowledge of their baricentric coordinates (see, e.g., Mifflin, 1977). This is largely because such problems arise quite often in large-scale optimization algorithms, as well as in many other cases. However, in most such problems some additional knowledge of baricentric coordinates, etc., is implied by the specific nature of the problem; ignoring this information is a simplification resulting in more straightforward, but less effective, algorithms.

The first quasi-Newton algorithm of this type, based in fact on results closely related to Lemma 1, was given with convergence proofs by Lemaréchal (1975). However, Lemaréchal did not specify what the matrix H^k should approximate; it was required only that H^k should be uniformly positive definite, which is sufficient for simple convergence. The results given in previous sections of this paper make it clear that H^k should approximate (in the sense described in Section 3.6) either the Hessian $\sum_{i \in A^k} y_i^k f_{ixx}^k$ or, in the nonconvex case, the augmented Hessian of type (30).

But the results of previous sections also show that such an approximation is actually impossible if no additional knowledge of the baricentric coordinates is assumed. The use of consecutive $g_k \in \partial f(x^k)$ gives no second-order information, if $g_k = \sum_{i \in A(x^k)} \lambda_i^k f_{ix}^k$ where λ_i^k might be arbitrary, not even converging to the optimal baricentric coordinates $\hat{y}_i$ (if they are unique) if x_k converges to $\hat{x}$. The use of the elements $\hat{g}^k$ closest to zero as a convex combination of previous g_j, $j = 0,1,\ldots,k$ gives more information, at least if $\hat{g}^k$ converges to zero, because then some corresponding baricentric coordinates should converge to $\hat{y}_i$; but later $\hat{g}^k$ yield averaged information related to many previous x^j, $j = 0,1,\ldots,k$, and it is difficult to extract from these the current information related to x^k that is necessary for a variable metric approximation.

The above remarks do not prove that it is impossible to construct a superlinearly convergent algorithm for nondifferentiable optimization with subdifferentials given only implicitly; however, they do show that some stronger assumptions, either related to the choice of subgradients or to the basic nature of the problem, are necessary. For example, if the Haar condition is satisfied, then even a linear approximation algorithm could be superlinearly convergent. However, it is clear that the problem of obtaining superlinearly convergent algorithms in cases of nondifferentiable optimization with implicitly given subdifferentials requires further study.

4.9 Other Extensions and Research Directions

Some of the results given in this paper, for example, Lemmas 1 and 2, can be generalized for problems with infinite or innumerable constraints. The continuous minimax problem

$$\underset{x\in X}{\text{minimize}} \max_{z\in Z} f(x,z)$$

can be approached in this way, and, in the convex case, should not present great difficulties; the nonconvex case is, however, essentially more complex, since only a partial generalization of the augmented Lagrangian theory to infinite-dimensional spaces is now available--see Wierzbicki and Kurcyusz (1977).

APPENDIX 1

An Efficient Line-Search Method for Nonsmooth Optimization

It is assumed that, at a given point x^k, a search direction $\hat{x}^k$ and a linear estimation of the difference $f(x^k + \hat{x}^k) - f(x^k) \approx \hat{x}_0^k < 0$ are given. Function values $f_{\tau_i} = f(x^k + \tau_i \hat{x}^k)$ are computed in order to find $f_f = \min_{\tau_i} f_{\tau_i}$ and $\tau_f = \arg\min_{\tau_i} f_{\tau_i}$, where τ_i are elements of a specially generated sequence. The sequence $\{\tau_i\}$ starts with $\tau_0 = 1$ (or, optionally, with the value accepted for τ_f in a previous run of the line-search algorithm). The sequence $\{\tau_i\}$ ends with a value $\tau_g = \tau_i$ which satisfies two conditions:

(a) $$f_{\tau_i} \leq f(x^k) + m_a \tau_i \hat{\bar{x}}_0^k$$

(b) $$f_{\tau_i} \geq f(x^k) + m_b \tau_i \hat{\bar{x}}_0^k$$

where $0 < m_a < m_b < 1$; suggested values for m_a and m_b are $m_a = 0.3$, $m_b = 0.7$. To generate the sequence, an expansion or contraction ratio r is also used; suggested value r = 10.

The algorithm is as follows:

(0) Set $\tau_0 (=1)$, $\omega^0 := 0$, $f_f = f(x^k)$, $\tau_f := 0$, $i := 0$,

(i) Compute f_{τ_i} . If $f_{\tau_i} < f_f$, set $\tau_f := \tau_i$, $f_f := f_{\tau_i}$. If f_{τ_i} satisfies (a) and (b), stop.

(ii) If f_{τ_i} does not satisfy (a), set $\tau_{max} := \tau_i$. If $\omega^i = 0$ or $\omega^i = -1$, set $\omega^{i+1} := -1$. If $\omega^i = +1$, set $\omega^{i+1} = 2$.

(iii) If f_{τ_i} does not satisfy (b), set $\tau_{min} := \tau_i$. If $\omega^i = 0$ or $\omega^i = +1$, set $\omega^{i+1} := +1$. If $\omega^i = -1$, set $\omega^{i+1} = 2$.

(iv) If $|\omega^{i+1}| = 1$, set $\tau_{i+1} := r^{\omega^{i+1}} \tau_i$. If $\omega^{i+1} = 2$, set $\tau_{i+1} := (\tau_{max} \cdot \tau_{min})^{\frac{1}{2}}$. Set $i := i + 1$, go to (i).

Comment: $\omega^{i+1} = \pm 1$ means that τ_{i+1} should be increased or decreased by a factor of r. $\omega^{i+1} = 2$ means that both a lower bound τ_{min} and an upper bound τ_{max} for τ_f have already been found and they should be tightened by computing τ_{i+1} as their geometrical mean. The last value τ_g of τ_i, which satisfies (a) and (b), often gives useful information. If some external bounds limit the value of τ_i, the algorithm must be modified accordingly.

REFERENCES

Balinski, M.L. and P. Wolfe, eds., Nondifferentiable Optimization, *Mathematical Programming Study 3*, North-Holland Publ.Co., Amsterdam, 1975.

Clarke, F.H., Generalized Gradients and Applications, *Transactions of the American Mathematical Society*, 205(1975), 247-262.

Hohenbalken, B. von, Least Distance Methods for the Scheme of Polytopes, *Mathematical Programming*, 15(1978), 1-11.

Kreglewski, T. and A.P. Wierzbicki, *Further Properties and Modifications of the Rank-One Variable Metric Method*, International Conference on Mathematical Programming, Zakopane, 1977.

Lemaréchal, C., *Nondifferentiable Optimization: Subgradient and ε-Subgradient Methods*, Lecture Notes: Numerical Methods in Optimization and Operations Research, Springer Verlag, 1975, 191-199.

Lemaréchal, C., *Nonsmooth Optimization and Descent Methods*, RR-78-4, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1978.

Madsen, K. and H. Schjaer-Jacobsen, Linearly Constrained Minimax Optimization, *Mathematical Programming*, 14(1977), 208-223.

Mifflin, R., *An Algorithm for Constrained Optimization With Semismooth Functions*, RR-77-3, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1977.

Nurminski, E.A., The Quasigradient Method for Solving Nonlinear Programming Problems, *Cybernetics*, 9, 1(1973), 145-150.

Rockafellar, R.T., Augmented Lagrange Multiplier Functions and Duality in Nonconvex Programming, *SIAM Journal on Control and Optimization*, 12(1974), 497-513.

Rockafellar, R.T., Lagrange Multipliers in Optimization, *SIAM-AMS Proceedings*, 9(1976), 145-168.

Szymanowski, J. et al., *Computational Methods of Optimization. Basic Research and Numerical Tests* (in Polish), Research Report, Institute of Automatic Control, Technical University of Warsaw, 1977.

Wierzbicki, A.P., Maximum Principle for Semiconvex Performance Functionals, *SIAM Journal on Control and Optimization*, 10(1972), 444-459.

Wierzbicki, A.P. and St. Kurcyusz, Projection on a Cone, Penalty Functionals and Duality Theory for Problems with Inequality Constraints in Hilbert Space, *SIAM Journal on Control and Optimization*, 15(1977), 25-26.

Wierzbicki, A.P., *A Quadratic Approximation Method Based on Augmented Lagrangian Functions for Nonconvex Nonlinear Programming Problems*, WP-78-61, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1978.

Wolfe, P., Finding a Nearest Point in a Polytope, *Mathematical Programming*, 11(1976), 128-149.

BIBLIOGRAPHY ON NONDIFFERENTIABLE OPTIMIZATION

E. Nurminski

International Institute for Applied Systems Analysis,

Laxenburg, Austria

This is a research bibliography with all the advantages and shortcomings that this implies. The author has used it as a bibliographical data base when writing papers, and it is therefore largely a reflection of his own personal research interests. However, it is hoped that this bibliography will nevertheless be of use to others interested in nondifferentiable optimization.

1. MONOGRAPHS

This section contains monographs related to nondifferentiable optimization. It also contains bibliographies on nondifferentiable optimization which are not parts of books, articles in journals, reports etc.

References

M.A. Aizerman, E.M. Braverman, and L.I. Rozonoer, *Potential Functions Method in Machine Learning Theory (in Russian)*, Nauka, Moscow (1970).

J.-P. Aubin, *Mathematical Methods of Game and Economic Theory*, North-Holland, Amsterdam (1979).

. J.-P. Aubin, *Methodes Explicites de l'Optimisation*, Dunod, Paris (1982).

. J.-P. Aubin, *Lecons d'Analyse Non-Lineaire et de ses Motivations Economiques*, Presses Universitaires de France, Paris (1984).

. A. Auslender, *Optimization: Methodes Numeriques*, Masson, Paris (1976).

. M. Avriel, *Nonlinear Programming, Analysis and Methods*, Prentice Hall, Englewood Cliffs, New Jersey (1976).

. J. Cea, *Optimization: Theorie et Algorithmes*, Dunod, Paris (1971).

. J.M. Danskin, *The Theory of Max-Min*, Springer, New York (1967).

. G.B. Dantzig, *Linear Programming and Extensions*, Princeton University Press, Princeton (1963).

. V.F. Demyanov and A.M. Rubinov, *Approximation Methods in Optimization Problems*, American Elsevier, New York (1970).

. V.F. Demyanov and V.N. Malozemov, *Introduction to Minimax*, John Wiley, New York (1974).

. V.F. Demyanov and L.V. Vasiliev, *Nondifferentiable Optimization (in Russian)*, Nauka, Moskow (1981).

. I. Ekeland and R. Temam, *Analyse Convexe et Problemes Variationnelles*, Dunod, Paris (1974).

. Yu.M. Ermoliev, *Stochastic Programming Methods (in Russian)*, Nauka, Moscow (1976).

. V.V. Fedorov, *Numerical Methods for Maxmin Problems (in Russian)*, Nauka, Moscow (1979).

. E.G. Golstein, *Theory of Convex Programming*, American Mathematical Society, Translations of Mathematical Monographs, Vol. 36, Providence, R.I. (1972).

. A.M. Gupal, *Stochastic Methods for Solving Nonsmooth Extremum Problems (in Russian)*, Naukova Dumka, Kiev (1979).

. J. Gwinner, "Bibliography on Nondifferentiable Optimization and Nonsmooth Analysis," *Journal of Computational and Applied Mathematics* **7**(4) pp. 277-285 (1981).

. V.Ja. Katkovnik, *Linear Estimates and Stochastic Optimization Problems (in Russian)*, Nauka, Moscow (1976).

. P.J. Kelly and M.L. Wiess, *Geometry and Convexity*, John Wiley, New York (1979).

. L.S. Lasdon, *Optimization Theory for Large Systems*, Macmillan, New York (1970).

. P.J. Laurent, *Approximation et Optimization*, Hermann, Paris (1972).

. C. Lemarechal and R. Mifflin, *Nonsmooth Optimization: Proceedings of a IIASA Workshop*, Pergamon Press, Oxford (1978).

. O. Mangasarian, *Nonlinear Programming*, McGraw-Hill, New York (1969).

. A.S. Nemirovski and D.B. Judin, *Complexity and Effectiveness of Optimization Methods (in Russian)*, Nauka, Moscow (1979).

. E.A. Nurminski, *Numerical Methods for Solving Deterministic and Stochastic Minmax Problems (in Russian)*, Naukova Dumka, Kiev (1979).

. E.A. Nurminski, "Bibliography on Nondifferentiable Optimization," WP-82-32, International Institute for Applied Systems Analysis, Laxenburg, Austria (1982).

. J.M. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York (1970).

. B.N. Pshenichniy, *Necessary Conditions for Extremum Problems (in Russian)*, Nauka, Moscow (1969). English translation: Marcel Dekker, New York (1971).

. B.N. Pshenichniy and Yu.M. Danilin, *Numerical Methods for Extremum Problems (in Russian)*, Nauka, Moscow (1975).

. B.N. Pshenichniy, *Convex Analysis and Extremal Problems (in Russian)*, Nauka, Moscow (1980).

. A.W. Robers and D.E. Varberg, *Convex Functions*, Academic Press, New York (1973).

. R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton (1972).

. R.T. Rockafellar, "The Theory of Subgradients and its Applications to Problems of Optimization. Convex and Nonconvex Functions," in *Research and Education in Mathematics, Vol. 1*, ed. K.H. Hoffman and R. Wille,Heldermann, Berlin (1981).

. D.L. Russell, *Calculus of Variations and Control Theory*, Academic Press, New York (1976).

. J.F. Shapiro, *Mathematical Programming: Structures and Algorithms*, John Wiley, New York (1979).

. N.Z. Shor, *Methods for Minimization of Nondifferentiable Functions and their Applications (in Russian)*, Naukova Dumka, Kiev (1979).

. Ja.Z. Tsypkin, *Adaptation and Learning in Automatic Systems (in Russian)*, Nauka, Moscow (1968). English translation: Academic Press, New York (1971).

. J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York (1972).

2. ALGORITHMS

This section deals with algorithms. It also contains reports on applications of nondifferentiable optimization and computational experiments in this field.

References

S. Agmon, "The Relaxation Method for Linear Inequalities," *Canadian Journal of Mathematics* **6** pp. 382-392 (1954).

A. Auslender, "Methodes Numeriques pour la Decomposition et la Minimisation de Fonctions Non-Differentiables," *Numerische Mathematik* **18** pp. 213-223 (1971).

A. Auslender, "Programmation Convexe avec Erreurs: Methodes de Epsilon-Sous-Gradients," *Comptes Rendus de l'Academie des Sciences (Paris) ,Serie A* **284**(2) pp. 109-112 (1977).

A. Auslender, "Penalty Methods for Computing Points that Satisfy Second Order Necessary Conditions," *Mathematical Programming* **17**(2) pp. 229-238 (1979).

A.B. Bakushinski and B.T. Poljak, "On the Solution of Variational Inequalities (in Russian)," *Doklady Akademii Nauk SSSR* **219** pp. 1038-1041 (1974). English translation: Soviet Mathematics Doklady pp. 1705-1710.

J.W. Bandler and C. Charalambous, "Nonlinear Programming Using Minimax Techniques," *Journal of Optimization Theory and Applications* **13** pp. 607-619 (1974).

M.S. Bazaraa and J.J. Goode, "The Traveling Salesman Problem: A Duality Approach," *Mathematical Programming* **13**(2) pp. 221-237 (1977).

M.S. Bazaraa, J.J. Goode, and R.L. Rardin, "A Finite Steepest-Ascent Method for Maximizing Piecewise-Linear Concave Functions," *Journal of Optimization Theory and Applications* **25**(3) pp. 437-442 (1978).

. M.S. Bazaraa and J.J. Goode, "A Survey of Various Tactics for Generating Lagrangian Multipliers in the Context of Lagrangian Duality," *European Journal of Operational Research* **3**(3) pp. 322-338 (1979).

. L.G. Bazhenov, "On the Convergence Conditions of a Method for Minimizing Almost-Differentiable Functions (in Russian)," *Kibernetika* **8**(4) pp. 71-72 (1972). English translation: Cybernetics Vol. 8(4)pp. 607-609.

. D.P. Bertsekas, "Stochastic Optimization Problems with Nondifferentiable Cost Functionals," *Journal of Optimization Theory and Applications* **12** pp. 218-231 (1973).

. D.P. Bertsekas and S.K. Mitter, "A Descent Numerical Method for Optimization Problems with Nondifferentiable Cost Functionals," *SIAM Journal on Control* **11** pp. 637-652 (1973).

. D.P. Bertsekas, "Nondifferentiable Optimization Via Approximation," pp. 1-25 in *Nondifferentiable Optimization*, ed. M.L. Balinski and P. Wolfe,Mathematical Programming Study 3, North-Holland, Amsterdam (1975).

. D.P. Bertsekas, "On the Method of Multipliers for Convex Programming," *IEEE Transactions on Automatic Control* **20** pp. 358-388 (1975).

. D.P. Bertsekas, "Necessary and Sufficient Condition for a Penalty Method to be Exact," *Mathematical Programming* **9**(1) pp. 87-99 (1975).

. D.P. Bertsekas, "A New Algorithm for Solution of Nonlinear Resistive Networks Involving Diodes," *IEEE Transactions on Circuit Theory* **23** pp. 599-608 (1976).

. D.P. Bertsekas, "Minimax Methods Based on Approximation," pp. 463-465 in *Proceedings of the 1976 Johns Hopkins Conference on Information Science and Systems*, Baltimore, Maryland (1976).

D.P. Bertsekas, "Approximation Procedures Based on the Method of Multipliers," *Journal of Optimization Theory and Applications* **23**(4) pp. 487-510 (1977).

B. Birzak and B.N. Pshenichniy, "On Some Problems in the Minimization of Nonsmooth Functions (in Russian)," *Kibernetika* **2**(6) pp. 53-57 (1966). English translation: Cybernetics Vol. 2(6) pp. 43-46.

C.E. Blair and R.G. Jeroslow, "An Exact Penalty Method for Mixed-Integer Programs," *Mathematics of Operations Research* **6**(1) pp. 14-18 (1981).

J. Bracken and J.T. McGill, "A Method for Solving Mathematical Programs with Nonlinear Programs in the Constraints," *Operations Research* **22** pp. 1097-1101 (1974).

R. Brooks and A.M. Geoffrion, "Finding Everett's Lagrange Multipliers by Linear Programming," *Operations Research* **14**(6) pp. 1149-1153 (1966).

A. Butz, "Iterative Saddle Point Techniques," *SIAM Journal on Applied Mathematics* **15** pp. 719-726 (1967).

P.M. Camerini, L. Fratta, and F. Maffioli, "A Heuristically Guided Algorithm for the Traveling Salesman Problem," *Journal of the Institution of Computer Science* **4** pp. 31-35 (1973).

P.M. Camerini, L. Fratta, and F. Maffioli, "On Improving Relaxation Methods by Modified Gradient Techniques," pp. 26-34 in *Nondifferentiable Optimization*, ed. M.L. Balinski and P. Wolfe,Mathematical Programming Study 3, North-Holland, Amsterdam (1975).

J. Cea and R. Glowinski, "Minimisation des Fonctionnelles Non-Differentiables," Report No. 7105, I.N.R.I.A, Le Chesnay, France (1971).

C. Charalambous, "Nonlinear Least p-th Optimization and Nonlinear Programming," *Mathematical Programming* **12**(2) pp. 195-225 (1977).

C. Charalambous, "A Lower Bound for the Controlling Parameters of the Exact Penalty Functions," *Mathematical Programming* 15(3) pp. 278-290 (1978).

C. Charalambous, "On Conditions for Optimality of the Nonlinear L-1 Problem," *Mathematical Programming* 17(2) pp. 123-135 (1979).

J. Chatelon, D. Hearn, and T.J. Lowe, "A Subgradient Algorithm for Certain Minimax and Minisum Problems," *Mathematical Programming* 14(2) pp. 130-145 (1978).

E.W. Cheney and A.A. Goldstein, "Newton's Method for Convex Programming and Chebyshev Approximation," *Numerische Mathematik* 1(1) pp. 253-268 (1959).

A.R. Conn, "Constrained Optimization Using a Nondifferentiable Penalty Function ," *SIAM Journal on Numerical Analysis* 10 pp. 760-784 (1973).

A.R. Conn, "Linear Programming Via a Nondifferentiable Penalty Function," *SIAM Journal on Numerical Analysis* 13(1) pp. 145-154 (1976).

G. Cornuejols, M.L. Fisher, and G.L. Nemhauser, "Location of Bank Accounts to Optimize Float: An Analytic Study of Exact and Approximate Algorithms," *Management Science* 23(8) pp. 789-810 (1977).

H. Crowder, "Computational Improvements for Subgradient Optimization," RC4907, IBM, T. Watson Research Center, New York (1974).

J. Cullum, W.E. Donath, and P. Wolfe, "The Minimization of Certain Nondifferentiable Sums of Eigenvalues of Symmetric Matrices," pp. 35-55 in *Nondifferentiable Optimization*, ed. M.L. Balinski and P. Wolfe,Mathematical Programming Study 3, North-Holland, Amsterdam (1975).

G.B. Dantzig, "General Convex Objective Forms," in *Mathematical Methods in the Social Sciences*, ed. S. Karlin, K.J. Arrow, and P. Suppes,Stanford

University Press, Stanford (1960).

V.F. Demyanov, "On Minimax Problems (in Russian)," *Kibernetika* **2**(6) pp. 58-66 (1966). English translation: Cybernetics Vol. 2(6) pp. 47-53.

V.F. Demyanov, "Algorithms for Some Minimax Problems," *Journal of Computer and System Sciences* **2** pp. 342-380 (1968).

V.F. Demyanov and A.M. Rubinov, "Minimization of Functionals in Normed Spaces," *SIAM Journal on Control* **6** pp. 73-89 (1968).

V.F. Demyanov, "Seeking a Minimax on a Bounded Set (in Russian)," *Doklady Academii Nauk SSSR* **191** pp. 517-521 (1970).

V.F. Demyanov, "A Continuous Method for Solving Minimax Problems (in Russian)," *Zurnal Vycislitel'noi Matematiki i Matematiceskoi Fiziki* **15** pp. 592-598 (1975). English translation: USSR Computational Mathematics and Mathematical Physics Vol. 15 pp. 45-52.

V.F. Demyanov and L.V. Vasiliev, "The Relaxation Method of Generalized Gradient (in Russian)," *Optimizacia* **19(36)** pp. 48-52 (1977).

V.F. Demyanov, L.V. Vasiliev, and S.A. Lisina, "Minimization of a Convex Function by Means of E-Subgradients (in Russian)," pp. 3-22 in *Control of Dynamical Systems*, Leningrad State University, Leningrad (1978).

V.F. Demyanov, "A Multistep Method for Generalized Gradient Descent (in Russian)," *Zurnal Vycislitel'noi Matematiki i Matematiceskoi Fiziki* **18**(5) pp. 1112-1118 (1978).

V.F. Demyanov, "A Modified Generalized Gradient Method in the Presence of Constraints (in Russian)," *Vestnik Leningradskogo Universiteta, Matematika, Mekhanika, Astronomiya* **19**(4) pp. 25-29 (1978). English translation: to appear in Vestnik Leningradskogo Universiteta.

V.F. Demyanov and L.V. Vasiliev, "The Method of (e,m,t)-Generalized Gradient Descent in the Presence of Constraints (in Russian)," *Vestnik Leningradskogo Universiteta, Matematika, Mekhanika, Astronomiya* **12**(1980).

V.F. Demyanov, "Subgradient Method and Saddle Points (in Russian)," *Vestnik Leningradskogo Universiteta, Mathematika, Mekhanika, Astronomiya* **13**(7) pp. 17-23 (1981).

L.C.W. Dixon, "Reflections on Nondifferentiable Optimization, Part I:The Ball-Gradient," *Journal of Optimization Theory and Applications* **32**(2) pp. 123-134 (1980).

L.C.W. Dixon and M. Gaviano, "Reflections on Nondifferentiable Optimization, Part 2: Convergence," *Journal of Optimization Theory and Applications* **32**(3) pp. 259-276 (1980).

J. Elzinga and T.G. Moore, "A Central Cutting Plane Algorithm for the Convex Programming Problem," *Mathematical Programming* **8** pp. 134-145 (1975).

I.I. Eremin, "A Generalization of the Motzkin-Agmon Relaxation Method (in Russian)," *Uspekhi Matematicheski Nauk* **20** pp. 183-187 (1965).

I.I. Eremin, "The Relaxation Method of Solving Systems of Inequalities with Convex Functions on the Left Side (in Russian)," *Doklady Academii Nauk SSSR* **160** pp. 994-996 (1965). English translation: Soviet Mathematics Doklady Vol. 6 pp. 219-221(1965).

I.I. Eremin, "On the Penalty Method in Convex Programming (in Russian)," *Doklady Akademii Nauk SSSR* **173**(4) pp. 63-67 (1967).

I.I. Eremin, "Standard Iteration Nonsmooth Optimization Process for Nonstationary Convex Programming Problems.Part 1 (in Russian)," *Zurnal Vycislitel'noi Matematiki i Matematiceskoi Fiziki* **18**(6) pp. 1430-1442 (1978).

I.I. Eremin, "Standard Iteration Nonsmooth Optimization Process for Nonstationary Convex Programming Problems.Part 2 (in Russian)," *Zurnal Vycislitel'noi Matematiki i Matematiceskoi Fiziki* **19**(1) pp. 112-120 (1979).

I.I. Eremin, "On Nonstationary Processes of Mathemetical Programming (in Russian)," pp. 13-19 in *Methods of Optimization and Pattern Recognition in Planning Problems*, Institute of Mathematics and Mechanics, Sverdlovsk (1980).

Yu.M. Ermoliev, "Methods of Solution of Nonlinear Extremal Problems (in Russian)," *Kibernetika* **2**(4) pp. 1-17 (1966).

Yu.M. Ermoliev and Z.V. Nekrylova, "Some Methods of Stochastic Optimization (in Russian)," *Kibernetika* **2**(6) pp. 96-98 (1966). English translation: Cybernetics Vol. 2(6) pp. 77-78.

Yu.M. Ermoliev and N.Z. Shor, "On the Minimization of Nondifferentiable Functions (in Russian)," *Kibernetika* **3**(1) pp. 101-102 (1967). English translation: Cybernetics Vol. 3(1) pp. 72-80.

Yu.M. Ermoliev and N.Z. Shor, "Method of Random Walk for the Two-Stage Problems of Stochastic Programming and its Generalization (in Russian)," *Kibernetika* **4**(1) pp. 90-92 (1968). English translation: Cybernetics Vol.4(1) pp. 59-60.

Yu.M. Ermoliev, "On the Method of Generalized Stochastic Gradients and Stochastic Quasi-Fejer Sequences (in Russian)," *Kibernetika* **5**(2)(1969). English translation: Cybernetics Vol. 5(2) pp. 208-220.

Yu.M. Ermoliev and L.G. Ermolieva, "The Method of Parametric Decomposition (in Russian)," *Kibernetika* **9**(2) pp. 66-69 (1973). English translation: Cybernetics Vol. 9(2) pp. 262-266.

. Yu.M. Ermoliev and E.A. Nurminski, "Limit Extremal Problems (in Russian)," *Kibernetika* **9**(4) pp. 130-132 (1973). English translation: Cybernetics Vol. 9(4) pp. 691-693.

. Yu. M. Ermoliev, "Stochastic Models and Methods of Optimization (in Russian)," *Kibernetika* 11(4) pp. 109-119 (1975). English translation: Cybernetics Vol. 11(4) pp. 630-641.

. J.P. Evans and F.J. Gould, "Application of the Generalized Lagrange Multiplier Technique to a Production Planning Problem," *Naval Research Logistic Quarterly* **18**(1) pp. 59-74 (1971).

. J.P. Evans, F.J. Gould, and J.W. Tolle, "Exact Penalty Functions in Nonlinear Programming," *Mathematical Programming* **4** pp. 72-97 (1973).

. H. Everett, "Generalized Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources," *Operations Research* **11** pp. 399-417 (1963).

. J.E. Falk, "A Linear Max-Min Problem," *Mathematical Programming* **5**(2) pp. 169-188 (1973).

. A. Feuer, "An Implementable Mathematical Programming Algorithm for Admissible Fundamental Functions," Ph. D. Dissertation, Columbia University, New York (1974).

. M.L. Fisher, W.D. Northup, and J.F. Shapiro, "Using Duality to Solve Discrete Optimization Problems: Theory and Computational Experience," pp. 56-94 in *Nondifferentiable Optimization*, ed. M.L. Balinski and P. Wolfe, Mathematical Programming Study 3, North-Holland, Amsterdam (1975).

. M.L. Fisher, "A Dual Algorithm for the One-Machine Scheduling Problem," *Mathematical Programming* 11(3) pp. 229-251 (1976).

M.L. Fisher, "The Lagrangian Relaxation Method for Solving Integer Programming Problems," *Management Science* **27**(1)(1981).

R. Fletcher, "An Exact Penalty Function for Nonlinear Programming with Inequalities," *Mathematical Programming* **5**(2) pp. 129-150 (1973).

R. Fletcher, "Methods Related to Lagrangian Functions," pp. Chapter 8 in *Numerical Methods for Constrained Optimization*, ed. P.E. Gill and W. Murray,Academic Press, London (1974).

R. Fletcher, "Conjugate Gradient Methods for Indefinite Systems," pp. 73-89 in *Numerical Analysis, Dundee*, ed. G. A. Watson,Springer, Berlin (1975).

R. Fletcher, "Methods for Solving Nonlinear Constrained Optimization Problems," University of Dundee Report NA 16 (1976). Also to be published in the Proceedings of the York State-of-the-Art Conference.

A.A. Gaivoronski, "Nonstationary Stochastic Programming Problems (in Russian)," *Kibernetika* **14**(4) pp. 89-92 (1978). English translation: Cybernetics Vol. 14(4) pp. 575-579.

A.M. Geoffrion, "Primal Resource-Directive Approaches for Optimizing Nonlinear Decomposable Systems," *Operations Research* **18**(3) pp. 375-403 (1970).

A.M. Geoffrion, "Duality in Nonlinear Programming: A Simplified Applications - Oriented Development," *SIAM Review* **13** pp. 1-37 (1971).

A.M. Geoffrion and G. Graves, "Multicommodity Distribution System Design by Benders' Decomposition," *Management Science* **20**(5) pp. 822-844 (1974).

O.V. Glushkova and A.M. Gupal, "Numerical Methods for the Minimization of Maximum Functions without Calculating Gradients (in Russian)," *Kibernetika* **16**(5) pp. 141-143 (1980).

O.V. Glushkova and A.M. Gupal, "On Nonmonotonous Methods for Minimizing Nonsmooth Functions with Gradient Averaging (in Russian)," *Kibernetika* **16**(6) p. 128 (1980).

J.L. Goffin, "On the Finite Convergence of the Relaxation Method for Solving Systems of Inequalities," ORC 71-36, Operations Research Center Report, University of California, Berkeley (1971).

J.L. Goffin, "On Convergence Rates of Subgradient Optimization Methods," *Mathematical Programming* **13**(3) pp. 329-347 (1977).

J.L. Goffin, "The Relaxation Method for Solving Systems of Linear Inequalities," *Mathematics of Operations Research* **5**(3) pp. 388-414 (1980).

A.A. Goldstein, "Optimization with Corners," pp. 215-230 in *Non-Linear Programming, Vol. 2*, Academic Press, New York (1975).

A.A. Goldstein, "Optimization of Lipschitz Continuous Functions," *Mathematical Programming* **13** pp. 14-22 (1977).

E.G. Golstein, "Generalized Gradient Method for Finding Saddle Points (in Russian)," *Ekonomika i Matematicheskie Metody* **8**(4)(1970).

R.C. Grinold, "Steepest Ascent for Large-Scale Linear Programs," *SIAM Review* **14** pp. 447-464 (1972).

L. Grippo and A. La Bella, "Some Abstract Models of Convergent Algorithms in Nondifferentiable Optimization," R.78-19, Universita di Roma, Instituto di Automatica (1978).

M. Grotschel, L. Lovasz, and A. Schrijver, "The Ellipsoid Method and its Consequences in Combinatorial Optimization," Report No. 80151-OR, Institute fur Okonometrie und Operations Research, Rheinische Friedrich-Wilhelms-Universitat, Bonn (1980).

A.M. Gupal, "One Stochastic Programming Problem with Constraints of a Probabilistic Nature (in Russian)," *Kibernetika* **10**(5) pp. 94-100 (1974). English translation: Cybernetics Vol. 10(6) pp. 1019-1026.

A.M. Gupal, "On a Minimization Method for Almost-Differentiable Functions (in Russian)," *Kibernetika* **13**(1) pp. 114-116 (1977).

A.M. Gupal and V.I. Norkin, "A Minimization Algorithm for Discontinuous Functions (in Russian)," *Kibernetika* **13**(2) pp. 73-75 (1977).

A.M. Gupal, "Method for the Minimization of Functions Satisfying the Lipschitz Condition (in Russian)," *Kibernetika* **16**(5) pp. 91-94 (1980).

O.V. Guseva, "Convergence Rate of the Generalized Stochastic Gradient Method (in Russian)," *Kibernetika* **7**(4) pp. 143-145 (1971). English translation: Cybernetics Vol. 7(4) pp. 734-742.

J. Hald and H. Schrajer-Jacobsen, "Linearly Constrained Minimax Optimization without Calculating Derivatives," *Third Symposium on Operations Research (University of Mannheim, Mannheim), Section 1*, pp. 289-301 (1978).

J. Hald and K. Madsen, "A 2-Stage Algorithm for Minmax Optimization," pp. 225-239 in *International Symposium on Systems Optimization and Analysis, Rocquencourt, December 11-13,1978*, Lecture Notes in Control and Information Science, Vol.14, Springer, Berlin (1979).

J. Hald and K. Madsen, "Combined LP and Quasi-Newton Methods for Minimax Optimization," *Mathematical Programming* **20**(1) pp. 49-62 (1981). [Also technical report NI-79-05, Numerisk Institut Danmarks Tekniske Hojskole, Lyngby, Denmark, 1979.]

S.P. Han, "Variable Metric Methods for Minimizing a Class of Nondifferentiable Functions," *Mathematical Programming* **20**(1) pp. 1-13

(1981).

. M. Held and R.M. Karp, "The Traveling Salesman Problem and Minimum Spanning Trees. Part 1," *Operations Research* **18** pp. 1138-1162 (1970).

. M. Held and R.M. Karp, "The Traveling Salesman Problem and Minimum Spanning Trees. Part 2," *Mathematical Programming* **1** pp. 6-25 (1971).

. B. von Hohenbalken, "Least Distance Methods for the Scheme of Polytopes," *Mathematical Programming* **15** pp. 1-11 (1978).

. D.B. Judin and A.S. Nemirovski, "Evaluation of Information Complexity for Mathematical Programming Problems (in Russian)," *Ekonomika i Matematicheskie Metody* **12**(1) pp. 128-142 (1976).

. D.B. Judin and A.S. Nemirovski, "Information Complexity and Effective Methods for Solving Convex Extremum Problems (in Russian)," *Ekonomika i Matematicheskie Metody* **12**(2) pp. 357-369 (1976).

. C.Y. Kao and R.R. Mayer, "Secant Approximation Methods for Convex Optimization," pp. 143-162 in *Mathematical Programming at Oberwolfach*, ed. H. Konig, B. Korte, and K. Ritter,Mathematical Programming Study 14, North-Holland, Amsterdam (1981).

. A.A. Kaplan, "A Convex Programming Method with Internal Regularization (in Russian)," *Doklady Akademii Nauk SSSR* **241**(1) pp. 22-25 (1978). English translation: Soviet Mathematics Doklady Vol. 19(4) pp. 795-799.

. S. Kaplan, "Solution of the Lorie-Savage and Similar Integer Programming Problems by the Generalized Lagrange Multiplier Method," *Operations Research* **14** pp. 1130-1136 (1966).

. N.N. Karpinskaja, "Methods of Penalty Functions and the Foundations of Pyne's Method (in Russian)," *Avtomatika i Telemekhanika* **28** pp. 140-146 (1967). English translation: Automation and Remote Control Vol. 28 pp.

124-129.

J.E. Kelley, "The Cutting Plane Method for Solving Convex Programs," *Journal of the Society for Industrial and Applied Mathematics* **8**(4) pp. 703-712 (1960).

L.G. Khachiyan, "A Polynomial Algorithm in Linear Programming (in Russian)," *Doklady Akademii Nauk SSSR* **244** pp. 1093-1096 (1979).

L.G. Khachiyan, "Polynomial Algorithm in Linear Programming (in Russian)," *Zurnal Vycislitel'noi Matematiki i Matematiceskoi Fiziki* **20**(1) pp. 51-68 (1980). English translation: USSR Computational Mathematics and Mathematical Physics Vol. 20(1) pp.53-72.

K.C. Kiwiel, "A Variable Metric Method of Centers for Nonsmooth Minimization," CP-81-23, International Institute for Applied Systems Analysis, Laxenburg, Austria (1981).

S.K. Korovin and V.I. Utkin, "Method of Piecewise Smooth Penalty Functions (in Russian)," *Avtomatika i Telemekhanika* **37** pp. 94-105 (1976). English translation: Automation and Remote Control Vol.37 pp. 39-48.

B. Korte and R. Schrader, "A Note on Convergence Proofs for Shor-Khachiyan Methods," Report No. 80156 - OR, Institute fur Okonometrie und Operations Research, Rheinische Friedrich-Wilhelms-Universitat, Bonn (1980).

M.K. Kozlov, S.P. Tarasov, and L.G. Khachiyan, "Polynomial Solvability of Convex Quadratic Programming (in Russian)," *Doklady Akademii Nauk SSSR* **248**(5) pp. 1049-1051 (1980).

O.V. Kupatadze, "On the Gradient Method for Minimizing Nonsmooth Functions (in Russian)," in *Optimalnye i Adaptivnye Sistemy*, Trudy 4 Vsesojuzn. Sovesch. po Avt.Upr. (Tbilisi, 1968), Nauka, Moscow (1972).

A.I. Kuzovkin and V.M. Tihomirov, "On the Quantity of Observations Required to Find a Minimum of a Convex Function (in Russian)," *Ekonomika i Matematicheskie Metody* **3**(1) pp. 95-103 (1967).

C. Lemarechal, "An Algorithm for Minimizing Convex Functions," pp. 552-556 in *Information Processing '74*, ed. J.L. Rosenfeld,North-Holland, Amsterdam (1974).

C. Lemarechal, "Note on an Extension of Davidon Methods to Nondifferentiable Functions," *Mathematical Programming* **7**(3) pp. 384-387 (1974).

C. Lemarechal, "Nondifferentiable Optimization; Subgradient and E-Subgradient Methods," pp. 191-199 in *Lecture Notes in Economics and Mathematical Systems, Vol. 117*, ed. W. Oettli,Springer, Berlin (1975).

C. Lemarechal, *Combining Kelley's and Conjugate Gradient Methods*, Abstracts, 9th International Symposium on Mathematical Programming, Budapest (1976).

C. Lemarechal, "A View of Line-Searches," pp. 59-78 in *Optimization and Optimal Control*, ed. A.Auslender, W.Oettli, and J.Stoer,Proceedings of a Conference held at Oberwolfach, March 16-22, 1980, Lecture Notes in Control and Information Science, Vol.30, Springer, Berlin (1981).

A.Yu. Levin, "On an Algorithm for the Minimization of Convex Functions (in Russian)," *Doklady Akademii Nauk SSSR* **160** pp. 1244-1247 (1965). English translation: Soviet Mathematics Doklady Vol.6 pp. 286-290.

E.S. Levitin and B.T. Poljak, "Convergence of Minimizing Sequences in Conditional Extremum Problems (in Russian)," *Doklady Akademii Nauk SSSR* **168** pp. 993-996 (1966). English translation: Soviet Mathematics Doklady Vol.7 pp. 764-767.

E.S. Levitin, "A General Minimization Method for Nonsmooth Extremal Problems (in Russian)," *Zurnal Vycislitel'noi Matematiki i Matematiceskoi Fiziki* **9** pp. 783-806 (1969). English translation: USSR Computational Mathematics and Mathematical Physics Vol. 9 pp. 63-69.

D.G. Luenberger, "Control Problems with Kinks," *IEEE Transactions on Automatic Control* **15** pp. 570-575 (1970).

K. Madsen, "An Algorithm for Minimax Solution of Overdetermined Systems of Nonlinear Equations," *Journal of the Institute of Mathematics and its Applications* **16** pp. 321-328 (1975).

K. Madsen, "Minimax Solution of Nonlinear Equations without Calculating Derivatives," pp. 110-126 in *Nondifferentiable Optimization*, ed. M.L. Balinski and P. Wolfe,Mathematical Programming Study 3, North-Holland, Amsterdam (1975).

K. Madsen and H. Schrajer-Jacobsen, "Linearly Constrained Minimax Optimization," *Mathematical Programming* **14**(2) pp. 208-223 (1978).

O.L. Mangasarian, "Iterative Solution of Linear Programs," Computer Science Technical Report #327, Computer Science Department,University of Winsconsin, Madison (1979).

R.E. Marsten, "The Use of the Boxstep Method in Discrete Optimization," pp. 127-144 in *Nondifferentiable Optimization*, ed. M.L. Balinski and P. Wolfe,Mathematical Programming Study 3, North-Holland, Amsterdam (1975).

R.E. Marsten, W.W. Hogan, and J.W. Blankenship, "The Boxstep Method for Large-Scale Optimization," *Operations Research* **23**(3) pp. 389-405 (1975).

G.D. Maystrovsky, "On a Gradient Method for Searching for Saddle Points (in Russian)," *Ekonomika i Matematicheskie Metody* **12**(5) pp. 917-929 (1976).

R. Mifflin, "Semismooth and Semiconvex Functions in Constrained Optimization," RR-76-21, International Institute for Applied Systems Analysis, Laxenburg, Austria (1976). [Also in SIAM Journal on Control and Optimization Vol.15(6) pp.959-972 (1977).]

R. Mifflin, "An Algorithm for Constrained Optimization with Semismooth Functions," *Mathematics of Operations Research* **2** pp. 191-207 (1977).

R. Mifflin, "A Stable Method for Solving Certain Constrained Least-Squares Problems," *Mathematical Programming* **16**(2) pp. 141-158 (1979).

V.S. Mikhalevich, Yu.M. Ermoliev, V.V. Skurba, and N.Z. Shor, "Complex Systems and Solution of Extremal Problems (in Russian)," *Kibernetika* **3**(5) pp. 29-39 (1967). English translation: Cybernetics Vol. 3(5) pp. 25-34.

V.S. Mikhalevich, I.V. Sergienko, and N.Z. Shor, "A Study of Methods for Solving Optimization Problems and Their Applications (in Russian)," *Kibernetika* **17**(4) pp. 89-113 (1981).

H. Mine and M. Fukushima, "A Minimization Method for the Sum of a Convex Function and a Continuously Differentiable Function," *Journal of Optimization Theory and Applications* **33**(1) pp. 9-24 (1981).

T. Motzkin and I.J. Schoenberg, "The Relaxation Method for Linear Inequalities," *Canadian Journal of Mathematics* **6** pp. 393-404 (1954).

J.A. Muckstadt and S.A. Koenig, "An Application of Lagrangian Relaxation to Scheduling in Power-Generation Systems," *Operations Research* **25**(3) pp. 387-403 (1977).

W. Murray and M.L. Overton, "Steplength Algorithms for a Class of Nondifferentiable Functions," CS-78-679, Stanford University, Stanford (1978).

A.S. Nemirovski, "Effective Iterative Methods of Solving Equations with Monotone Operators (in Russian)," *Ekonomika i Matematicheskie Metody* **17**(2) pp. 344-359 (1981).

V.I. Norkin, "Two Random Search Algorithms for the Minimization of Nondifferentiable Functions (in Russian)," in *Mathematical Methods in Operations Research and Reliability Theory*, ed. Yu.M. Ermoliev and I.N. Kovalenko,Ukrainian Academy of Sciences, Institute of Cybernetics, Kiev (1978).

V.I. Norkin, "Minimization Method for Nondifferentiable Functions with Averaging of Generalized Gradients (in Russian)," *Kibernetika* **16**(6) pp. 88-89 (1980).

E.A. Nurminski, "Quasigradient Method for Solving Nonlinear Programming Problems (in Russian)," *Kibernetika* **9**(1) pp. 122-125 (1973).

E.A. Nurminski, "Minimization of Nondifferentiable Functions Under Noise (in Russian)," *Kibernetika* **10**(4) pp. 59-61 (1974).

E.A. Nurminski and A.A. Zhelikhovski, "Investigation of One Regulating Step in a Quasi-Gradient Method for Minimizing Weakly Convex Functions (in Russian)," *Kibernetika* **10**(6) pp. 101-105 (1974). English translation: Cybernetics Vol. 10(6) pp. 1027-1031.

E.A. Nurminski and A.A. Zhelikhovski, "An E-Quasigradient Method for the Solution of Nonsmooth Extremal Problems (in Russian)," *Kibernetika* **13**(1) pp. 109-113 (1977). English translation: Cybernetics Vol.13.

E.A. Nurminski, "On E-Subgradient Methods of Nondifferentiable Optimization," pp. 187-195 in *International Symposium on Systems Optimization and Analysis, Rocquencourt, December 11-13,1978*, Lecture Notes in Control and Information Science, Vol.14,Springer, Berlin (1979).

E.A. Nurminski, "An Application of Nondifferentiable Optimization in Optimal Control," pp. 137-158 in *Numerical Optimization of Dynamic Systems*, ed. L.C.W. Dixon and G.P. Szego,North-Holland, Amsterdam (1980).

M.W. Padberg and M.R. Rao, "The Russian Method for Linear Inequalities," Technical Report, Graduate School of Business Administration, New York University, New York (1979).

V.M. Panin, "Linearization Method for the Discrete Minimax Problem (in Russian)," *Kibernetika* **16**(3) pp. 86-90 (1980).

G. Papavassilopoulos, "Algorithms for a Class of Nondifferentiable Problems," *Journal of Optimization Theory and Applications* **34**(1) pp. 41-82 (1981).

S.C. Parikh, "Approximate Cutting Planes in Nonlinear Programming," *Mathematical Programming* **11**(2) pp. 194-198 (1976).

T. Pietrzykowski, "An Exact Potential Method for Constrained Maxima," *SIAM Journal on Numerical Analysis* **6** pp. 217-238 (1969).

B.T. Poljak, "A General Method of Solving Extremal Problems (in Russian)," *Doklady Akademii Nauk SSSR* **174** pp. 33-36 (1967). English translation: Soviet Mathematics Doklady Vol. 8.

B.T. Poljak, "Minimization of Nonsmooth Functionals (in Russian)," *Zurnal Vycislitel'noi Matematiki i Matematiceskoi Fiziki* **9**(3) pp. 509-521 (1969). English translation: USSR Computational Mathematics and Mathematical Physics Vol.9(3) pp 14-29 (1969).

B.T. Poljak and Ja.Z. Tsypkin, "Pseudogradient Adaptation and Training (in Russian)," *Avtomatika i Telemekhanika* **34** pp. 45-68 (1973). English translation: Automation and Remote Control Vol. 34 pp. 377-397.

. B.T. Poljak, "Stochastic Regularized Algorithms," in *Supplement to Preprints*, Stochastic Control Symposium, IFAC, Budapest (1974).

. B.T. Poljak, "Convergence and Convergence Rates of Iterative Stochastic Algorithms, I. General Case (in Russian)," *Automatika i Telemekhanika* **37**(12) pp. 83-94 (1976). English translation: Automation and Remote Control Vol. 37 pp. 1858-1868.

. R.A. Poljak, "On the Best Convex Chebyshev Approximation (in Russian)," *Doklady Akademii Nauk SSSR* **200** pp. 538-540 (1971). English translation: Soviet Mathematics Doklady Vol. 12 pp. 1441-1444.

. R.A. Poljak, "Controlling Sequence Methods for Solution of Dual Problems in Convex Programming (in Russian)," pp. 95-111 in *Mathematical Methods for Solution of Economic Problems*, ed. N.P. Fedorenko and E.G. Golstein,Optimal Planning and Control Series, Vol.8, Nauka, Moscow (1979).

. M.J.D. Powell, "A View of Unconstrained Optimization," pp. 117-152 in *Optimization in Action*, ed. L.C.W. Dixon,Academic Press, London (1976).

. B.N. Pshenichniy, "Dual Method in Extremum Problems (in Russian)," *Kibernetika* **1**(3) pp. 89-95 (1965). English translation: Cybernetics Vol. 1(3) pp. 91-99.

. B.N. Pshenichniy, "Convex Programming in a Normalized Space (in Russian)," *Kibernetika* **1**(5) pp. 46-54 (1965). English translation: Cybernetics Vol. 1(5) pp. 46-57

. B.N. Pshenichniy, "On One Method for Solving a Convex Programming Problem (in Russian)," *Kibernetika* **16**(4) pp. 48-55 (1980).

. J.K. Reid, "On the Method of Conjugate Gradients for the Solution of Large Sparse Systems of Linear Equations," pp. Chapter 16 in *Large Sparse Sets of Linear Equations*, ed. J.K. Reid,Academic Press, London (1971).

- S.M. Robinson, "A Subgradient Algorithm for Solving K-Convex Inequalities," pp. 237-245 in *Optimization and Operations Research*, ed. M.A. Wolfe,Proceedings of the Conference held at Oberwolfach, 1975, Lecture Notes on Economics and Mathematical Systems, Vol.117 Springer, Berlin (1976).
- R.T. Rockafellar, "Augmented Lagrangians and Applications of the Proximal Point Algorithm in Convex Programming," *Mathematics of Operations Research* 1(2) pp. 97-116 (1976).
- R.T. Rockafellar, "Monotone Operators and the Proximal Point Algorithm," *SIAM Journal on Control and Optimization* 14 pp. 877-898 (1976).
- S.V. Rzhevsky, "Use of the Convex Function Subdifferential in One Method of Function Minimization (in Russian)," *Kibernetika* **16**(1) pp. 109-111 (1980).
- R. Saigal, "The Fixed Point Approach to Nonlinear Programming," pp. 142-157 in *Point-to-Set Maps in Mathematical Programming*, ed. P. Huard,Mathematical Programming Study 10, North-Holland, Amsterdam (1979).
- C. Sandi, "Subgradient Optimization," *Comb. Optimiz. Lect. Summer Sch. Comb. Optim. Urbino, 1977*, pp. 73-91 (1979).
- M.A. Sepilov, "The Generalized Gradient Method for Convex Programming Problems (in Russian)," *Ekonomika i Matematicheskie Metody* **11**(4) pp. 743-747 (1975).
- M.A. Sepilov, "On the Generalized Gradient Method for Extremal Problems (in Russian)," *Zurnal Vycislitel'noi Matematiki i Matematiceskoi Fiziki* **16**(1) pp. 242-247 (1976).
- J.F. Shapiro, "Generalized Lagrange Multipliers in Integer Programming," *Operations Research* **19**(1) pp. 68-75 (1971).

J.F. Shapiro, "Nondifferentiable Optimization and Large-Scale Linear Programming," pp. 196-209 in *International Symposium on Systems Optimization and Analysis, Rocquencourt, December 11-13,1978*, Lecture Notes in Control and Information Science, Vol.14, Springer, Berlin (1979).

M.B. Shchepakin, "On the Modification of a Class of Algorithms for Mathematical Programming (in Russian)," *Zurnal Vycislitel'noi Matematiki i Matematiceskoi Fiziki* **19**(6) pp. 1387-1395 (1979).

N.Z. Shor, "Application of the Gradient Method for the Solution of Network Transportation Problems (in Russian)," Notes, Scientific Seminar on Theory and Application of Cybernetics and Operations Research, Institute of Cybernetics of the Ukrainian Academy of Sciences, Kiev (1962).

N.Z. Shor, "On the Structure of Algorithms for the Numerical Solution of Problems of Optimal Planning and Design (in Russian)," Dissertation, Kiev (1964).

N.Z. Shor, "Application of Generalized Gradient Descent in Block Programming (in Russian)," *Kibernetika* **3**(3) pp. 53-55 (1967). English translation: Cybernetics Vol. 3(3) pp. 43-45.

N.Z. Shor, "The Rate of Convergence of the Generalized Gradient Descent Method (in Russian)," *Kibernetika* **4**(3) pp. 98-99 (1968). English translation: Cybernetics Vol. 4(3) pp. 79-80.

N.Z. Shor and M.B. Shchepakin, "Algorithms for Solving Two-Stage Stochastic Programming Problems (in Russian)," *Kibernetika* **4**(3)(1968). English translation: Cybernetics Vol. 4(3) pp. 48-50.

N.Z. Shor, "Convergence Rate of the Gradient Descent Method with Dilation of the Space (in Russian)," *Kibernetika* **6**(2) pp. 80-85 (1970). English translation: Cybernetics Vol. 6(2) pp. 102-108.

N.Z. Shor, "Utilization of the Operation of Space Dilation in the Minimization of Convex Functions (in Russian)," *Kibernetika* **6**(1) pp. 6-12 (1970). English translation: Cybernetics Vol. 6(1) pp. 7-15.

N.Z. Shor and P.R. Gamburd, "Certain Questions Concerning the Convergence of the Generalized Gradient Method (in Russian)," *Kibernetika* **7**(6) pp. 82-84 (1971). English translation: Cybernetics Vol. 7(6) pp. 1033-1036.

N.Z. Shor and N.G. Zhurbenko, "A Minimization Method Using Space Dilation in the Direction of the Difference of Two Successive Gradients (in Russian)," *Kibernetika* **7**(3) pp. 51-59 (1971). English translation: Cybernetics Vol. 7(3) pp. 450-459.

N.Z. Shor, "A Class of Almost-Differentiable Functions and a Minimization Method for Functions of this Class (in Russian)," *Kibernetika* **8**(4) pp. 65-70 (1972). English translation: Cybernetics Vol. 8(4) pp. 599-606.

N.Z. Shor and L.P. Shabashova, "Solution of Minimax Problems by the Generalized Gradient Method with Space Dilation (in Russian)," *Kibernetika* **8**(1) pp. 82-94 (1972). English translation: Cybernetics Vol. 8(1) pp. 88-94.

N.Z. Shor, "Convergence of a Gradient Method with Space Dilation in the Direction of the Difference between Two Successive Gradients (in Russian)," *Kibernetika* 11(4) pp. 48-53 (1975). English translation: Cybernetics Vol. 11(4) pp. 564-570.

N.Z. Shor, "Generalized Gradient Methods for Nonsmooth Functions and Their Applications to Mathematical Programming Problems (in Russian)," *Ekonomika i Matematicheskie Metody* **12**(2) pp. 332-356 (1976).

N.Z. Shor, "Cut-Off Method with Space Extension in Convex Programming Problems (in Russian)," *Kibernetika* **13**(1) pp. 94-96 (1977). English transla-

tion: Cybernetics Vol. 13(1) pp. 94-96.

N.Z. Shor, L.A. Galustova, and A.I. Momot, "Application of Mathematical Methods in Optimum Designing of the Unified Gas Supply System with Regard for Dymanics of Its Development (in Russian)," *Kibernetika* **14**(1) pp. 69-74 (1978).

N.Z. Shor and V.I. Gershovitch, "On One Family of Solution Algorithms for Problems of Convex Programming (in Russian)," *Kibernetika* **15**(4) pp. 62-67 (1979).

V.A. Skokov, "Note on Minimization Methods Using Space Dilation (in Russian)," *Kibernetika* **10**(4) pp. 115-117 (1974). English translation: Cybernetics Vol. 10(4) pp. 689-692.

J.J. Strodiot and V.H. Nguyen, "An Exponential Penalty Method for Nondifferentiable Minimax Problems with General Constraints," *Journal of Optimization Theory and Applications* **27**(2)(1979).

Ja.Z. Tsypkin and B.T. Poljak, "Attainable Accuracy of Adaptation Algorithms (in Russian)," *Doklady Academii Nauk SSSR* **218**(3) pp. 532-535 (1974).

S.P. Uryasyev, "On One Regulation of a Step in Limiting Extreme-Value Problems (in Russian)," *Kibernetika* **17**(1) pp. 105-108 (1981).

P. Wolfe, "Convergence Theory in Nonlinear Programming," pp. Chapter 1 in *Integer and Nonlinear Programming*, ed. J. Abadie,North-Holland, Amsterdam (1970).

P. Wolfe, M. Held, and H. Crowder, "Validation of Subgradient Optimization," *Mathematical Programming* **6** pp. 62-88 (1974).

P. Wolfe, "Note on a Method of Conjugate Subgradients for Minimizing Nondifferentiable Functions," *Mathematical Programming* **7**(3) pp. 380-383

(1974).

. P. Wolfe, "Finding the Nearest Point in a Polytope," *Mathematical Programming* 11(2) pp. 128-149 (1976).

. I. Zang, "Discontinuous Optimization by Smoothing," *Mathematics of Operations Research* 6(1) pp. 140-152 (1981).

. N.G. Zhurbenko, E.G. Pinaev, N.Z. Shor, and G.N. Yun, "Choice of Fleet Composition and Allocation of Aircraft to Civil Airline Routes (in Russian)," *Kibernetika* 12(4) pp. 138-141 (1976). English translation: Cybernetics Vol.12(4) pp. 636-641.

. N.G. Zhurbenko, "Study of One Class of Algorithms for Minimization of Nonsmooth Functions and Their Application to Solution of Large-Scale Problems (in Russian)," Dissertation, Kiev (1977).

. S. Zlobec, "A Note on Optimization Methods with Conditioned Gradients," *Zeitschrift fur Angewandte Mathematik und Mechanik* 59 pp. 279-281 (1979).

. S.I. Zukhovitski, R.A. Poljak, and M.E. Primak, "An Algorithm for the Solution of the Problem of Convex Chebyshev Approximation (in Russian)," *Doklady Akademii Nauk SSSR* 151 pp. 27-30 (1963). English translation: Soviet Mathematics Doklady Vol.4 pp. 901-904.

3. GENERALIZED DIFFERENTIABILITY

This section contains papers on general notions of differentiability, optimality conditions in the nondifferentiable case, properties of perturbation functions in parametric programming and the stability of optimum programs in connection with nondifferentiable optimization.

References

. A.D. Alexandrov, "The Existence Almost Everywhere of the Second Differential of a Convex Function and Some Associated Properties of Convex Surfaces (in Russian)," *Ucenye Zaapiski Leningradskogo Gosudarstvennogo Universiteta Seriya Matematika* **37**(6) pp. 3-35 (1939).

. E. Asplund and R.T. Rockafellar, "Gradients of Convex Functions," *Transactions of the American Mathematical Society* **139** pp. 443-467 (1969).

. N.N. Astafiev, "Stability and Marginal Values of Convex Programming Problems (in Russian)," *Sibirskii Matematicheski Zhurnal* **19**(3) pp. 491-503 (1978).

. H. Attouch and R.J.-B. Wets, "Approximation and Convergence in Nonlinear Optimization," WP-80-142, International Institute for Applied Systems Analysis, Laxenburg, Austria (1980).

. J.-P. Aubin and F.H. Clarke, "Multiplicateurs de Lagrange en Optimisation Non-Convexe et Applications," *Comptes Rendus de l'Academie des Sciences (Paris), Serie A* **285** pp. 451-454 (1977).

. J.-P. Aubin, "Gradients Generalises de Clarke," *Annales des Sciences Mathematiques de Quebec* **2** pp. 197-252 (1978).

. J.P. Aubin and F.H. Clarke, "Shadow Prices and Duality for a Class of Optimal Control Problems," *SIAM Journal on Control and Optimization* **17**(5) pp. 567-568 (1979).

. J.-P. Aubin, "Further Properties of Lagrange Multipliers in Nonsmooth Optimization," *Journal of Applied Mathematics and Optimization* **57** pp. 79-90 (1980).

. J.-P. Aubin, "Contingent Derivatives of Set-Valued Maps and Existence of Solutions to Nonlinear Inclusions and Differential Inclusions," pp. 160-232 in

Advances in Mathematics. Supplementary Studies, ed. L. Nachbin,Academic Press (1981).

J.P. Aubin, "Lipschitz Behavior of Solutions to Convex Minimization Problems," WP-81-76, International Institute for Applied Systems Analysis, Laxenburg, Austria (1981).

A. Auslender, "Convex Programming with Errors: Methods of e-Subgradients," in *Survey of Mathematical Programming*, ed. A. Prekopa,Proceedings of 9th International Mathematical Programming Symposium, Budapest, August 23-27, 1976, North-Holland, Amsterdam - Akademiai Kiado, Budapest (1979).

A. Auslender, "Differentiable Stability in Nonconvex and Nondifferentiable Programming," pp. 29-41 in *Point-to-Set Maps in Mathematical Programming*, ed. P. Huard,Mathematical Programming Study 10, North-Holland, Amsterdam (1979).

A. Auslender, "Sur la Differentiabilite de la Fonction d'Appui du Sousdifferential a e-Pres," *Comptes Rendus de l'Academie des Sciences (Paris), Serie A* **292**(3) pp. 221-224 (1981).

M. Avriel and I. Zang, "Generalized Arcwise-Connected Functions and Characterization of Local-Global Minimum Properties," *Journal of Optimization Theory and Applications* **32**(4) pp. 407-426 (1980).

M.S Bazaraa, J.F. Goode, and C.M. Shetty, "Optimality Criteria in Nonlinear Programming Without Differentiability," *Operations Research* **19** pp. 77-86 (1971).

A. Ben-Tal and A. Ben-Israel, "Characterization of Optimality in Convex Programming:Nondifferentiable Case," *Applicable Analysis* **9**(2) pp. 137-156 (1979).

V.I. Berdysev, "Continuity of Multivalued Mappings Connected with the Minimization of Convex Functionals (in Russian)," *Doklady Akademii Nauk SSSR* **243**(3) pp. 561-564 (1978).

V.V. Beresnev, "On Necessary Optimality Conditions for the Systems of Discrete Inclusions (in Russian)," *Kibernetika* **13**(2) pp. 58-64 (1977).

J.M. Borwein, "Fractional Programming without Differentiability," *Mathematical Programming* 11(3) pp. 283-290 (1976).

J.M. Borwein, "Tangent Cones, Starshape and Convexity," *International Journal of Mathematics and Mathematical Sciences* 1(4) pp. 497-498 (1978).

J.M. Borwein, "The Minimum of a Family of Programs," *Third Symposium on Operations Research (University of Mannheim, Mannheim)* 1 pp. 100-102 (1978).

J.M. Borwein, "A Multivalued Approach to the Farkas Lemma," in *Point-to-Set Maps in Mathematical Programming*, ed. P. Huard,Mathematical Programming Study 10, North-Holland, Amsterdam (1979).

J.M. Borwein, "A Note on Perfect Duality and Limiting Lagrangians," *Mathematical Programming* **18**(3) pp. 330-337 (1980).

J. Bracken and J.T. McGill, "Mathematical Programs with Optimization Problems in the Constraints," *Operations Research* **21** pp. 37-44 (1973).

A. Brondsted and R.T. Rockafellar, "On the Subdifferentiability of Convex Functions," *Proceedings of the American Mathematical Society* **16** pp. 605-611 (1965).

A. Brondsted, "On the Subdifferential of the Supremum of Two Convex Functions," *Mathematica Scandinavica* **31** pp. 225-230 (1972).

S.L. Brumelle, "Convex Operators and Supports," *Mathematics of Operations Research* **3**(2) pp. 171-175 (1978).

S. Chandra and M. Chanramohan, "Duality in Mixed Integer Nonconvex and Nondifferentiable Programming," *Zeitschrift fur Angewandte Mathematik und Mechanik* **59**(4) pp. 205-209 (1979).

F.H. Clarke, "Generalized Gradients and Applications," *Transactions of the American Mathematical Society* **205** pp. 247-262 (1975).

F.H. Clarke, "A New Approach to Lagrange Multipliers," *Mathematics of Operations Research* **1**(2) pp. 165-174 (1976).

F.H. Clarke, "Optimal Control and the True Hamiltonian," *SIAM Review* **21**(2) pp. 157-166 (1979).

F.H. Clarke, "Generalized Gradients of Lipschitz Functions," *Advances in Mathematics* **40**(1) pp. 52-67 (1980).

F.H. Clarke, "Nonsmooth Analysis and Optimization," pp. 847-853 in *Proceedings of the International Congress of Mathematics (Helsinki 1978)*, Finnish Academy of Sciences, Helsinki (1980).

J.P. Crouzeix, "Conditions for Convexity of Quasiconvex Functions," *Mathematics of Operations Research* **5**(1) pp. 120-125 (1980).

J.P. Crouzeix, "Some Differentiability Properties of Quasiconvex Functions on R-n," pp. 9-20 in *Optimization and Optimal Control*, ed. A.Auslender, W.Oettli, and J.Stoer,Lecture Notes in Control and Information Science, Vol.30, Springer (1981).

J.M. Danskin, "The Theory of Max-Min with Applications," *SIAM Journal on Applied Mathematics* **14** pp. 641-664 (1966).

V.F. Demyanov and V.N. Malozemov, "The Theory of Nonlinear Minimax Problems (in Russian)," *Uspekhi Matematcheski Nauk* **26** pp. 53-104 (1971).

V.F. Demyanov, "Second Order Directional Derivatives of a Function of the Maximum (in Russian)," *Kibernetika* **9**(5) pp. 67-69 (1973). English transla-

tion: Cybernetics Vol.9(5) pp. 797-800.

V.F. Demyanov and V.K. Somesova, "Conditional Subdifferential of Convex Functions (in Russian)," *Doklady Akademii Nauk SSSR* **242**(4) pp. 753-756 (1978).

V.F. Demyanov and A.M. Rubinov, "On Quasidifferentiable Functionals (in Russian)," *Doklady Academii Nauk SSSR* **250**(1) pp. 21-25 (1980).

V.F. Demyanov, "On the Relation between the Clarke Subdifferential and the Quasidifferential (in Russian)," *Vestnik Leningradskogo Universiteta* **13** pp. 18-24 (1980). English translation: Vestnik Leningrad Universitet Vol.13 pp 183-189.

V.F. Demyanov and A.M. Rubinov, "Some Approaches to Nonsmooth Optimization Problems (in Russian)," *Economika i Matematicheskie Metody* **17**(6) pp. 1153-1174 (1981).

S. Dolecki and S. Rolewicz, "Exact Penalty for Local Minima," *SIAM Journal on Control and Optimization* **17** pp. 596-606 (1979).

A.Y. Dubovitskii and A.A. Milyutin, "Extremum Problems in the Presence of Constraints (in Russian)," *Doklady Akademii Nauk SSSR* **149** pp. 759-761 (1963). English translation: Soviet Mathematics Doklady Vol. 4 pp. 452-455(1963).

A.Y. Dubovitskii and A.A. Milyutin, "Extremum Problems in the Presence of Restrictions (in Russian)," *Zurnal Vycislitel'noi Matematiki i Matematiceskoi Fiziki* **5** pp. 395-453 (1965). English translation: USSR Computational Mathematics and Mathematical Physics Vol.5

Pham Canh Duong and Hoang Tuy, "Stability, Surjectivity and Local Invertability of Nondifferentiable Mappings," *Acta Mathematica Vietnamica* **3**(1) pp. 89-105 (1978).

J.J.M. Evers and H.van Maaren, "Duality Principles in Mathematics and Their Relations to Conjugate Functions," Technical Report, Department of Applied Mathematics, Technische Hogeschool Twente, Enschede, The Netherlands (1981).

W. Fenchel, "On Conjugate Convex Functions," *Canadian Journal of Mathematics* **1** pp. 73-77 (1949).

R. Fletcher and G.A. Watson, "First and Second Order Conditions For a Class of Nondifferentiable Optimization Problems," *Mathematical Programming* **18**(3) pp. 291-307 (1980).

J. Gauvin and J.W. Tolle, "Differential Stability in Nonlinear Programming," *SIAM Journal on Control and Optimization* **15**(2) pp. 294-311 (1977).

J. Gauvin, "The Generalized Gradient of a Marginal Function in Mathematical Programming," *Mathematics of Operations Research* **4**(4) pp. 458-463 (1979).

E.G. Golstein and N.V. Tretyakov, "Modified Lagrangians in Convex Programming and Their Generalizations," pp. 86-97 in *Point-to-Set Maps in Mathematical Programming*, ed. P. Huard,Mathematical Programming Study 10, North-Holland, Amsterdam (1979).

B.D. Graven, "Implicit Function Theorems and Lagrange Multipliers," *Numerical Functional Analysis and Optimization* **6**(2) pp. 473-486 (1980).

R.C. Grinold, "Lagrangian Subgradients," *Management Science* **17** pp. 185-188 (1970).

A.M. Gupal, "On the Properties of Functions Satisfying the Local Lipschitz Condition (in Russian)," *Kibernetika* **14**(4) pp. 140-142 (1978).

J. Gwinner, "Closed Images of Convex Multivalued Mappings in Linear Topological Spaces with Applications," *Journal of Mathematical Analysis and*

Applications **60**(1) pp. 75-86 (1977).

J. Gwinner, "Contribution a la Programmation non Differentiable dans des Espaces Vectoriels Topologiques," *Comptes Rendus de l'Academie des Sciences (Paris), Serie A* **289**(10) p. 523 (1979).

J. Gwinner, "On Optimality Conditions for Infinite Programs," pp. 21-28 in *Optimization and Optimal Control*, ed. A.Auslender, W.Oettli, and J.Stoer,Proceedings of a Conference held at Oberwolfach, March 16-22, 1980, Lecture Notes in Control and Information Science, Vol.30, Springer (1981).

H. Halkin, "The Method of Dubovitskii-Milyutin in Mathematical Programming," pp. 1-12 in *Symposium on Optimization and Stability Problems in Continuum Mechanics, 24 August 1971, Los Angeles*, Springer (1973).

H. Halkin, "Implicit Functions and Optimization Problems without Continuous Differentiability of the Data," *SIAM Journal on Control* **12**(2) pp. 229-236 (1974).

H. Halkin, "Mathematical Programming without Differentiability," pp. 279-288 in *Calculus of Variations and Control Theory*, ed. D.L. Russell,Academic Press, New York (1976).

H. Halkin, "Necessary Conditions for Optimal Control Problems with Differentiable and Nondifferentiable Data," pp. 77-118 in *Mathetatical Control Theory*, ed. W.A. Coppel,Lecture Notes in Mathematics, Vol.68, Springer, Berlin (1978).

W. Heins and S.K. Mitter, "Conjugate Convex Functions, Duality, and Optimal Control Problems. I. Systems Governed by Ordinary Differential Equations," *Information Sciences* **2**(2) pp. 211-243 (1970).

J.-B. Hiriart-Urruty, "Contribution a la Programmation Mathematique Cas Deterministe et Stochastique," These Serie E N 247, Universite de Clermont-Ferrand, Clermont-Ferrand, France (1977).

J.-B. Hiriart-Urruty, "Generalized Gradients of Marginal Value Function," *SIAM Journal on Control and Optimization* **16** pp. 301-316 (1978).

J.-B. Hiriart-Urruty, "Refinements of Necessary Optimality Conditions in Nondifferentiable Programming," *Applied Mathematics and Optimization* **5**(1) pp. 63-82 (1979).

J.-B. Hiriart-Urruty, "Tangent Cones, Generalized Gradients and Mathematical Programming in Banach Spaces," *Mathematics of Operations Research* **4**(1) pp. 79-97 (1979).

J.-B. Hiriart-Urruty, "New Concepts in Nondifferentiable Programming," *Bulletin of the Mathematical Society of France* **60** pp. 5-85 (1979). [Also in Journees d'Analyse Non Convexe, May 1980]

J.-B. Hiriart-Urruty, "Lipschitz r-Continuity of the Approximate Subdifferential of a Convex Function," *Mathematica Scandinavica* **47** pp. 123-134 (1980).

J.-B. Hiriart-Urruty, "Extension of Lipschitz Functions," *Journal of Mathematical Analysis and Applications* **77**(2) pp. 539-554 (1980).

J.-B. Hiriart-Urruty, "E-Subdifferential Calculus," in *Proceedings of the Colloquium "Convex Analysis and Optimization", 28-29 February, 1980,* Imperial College, London (forthcoming).

W.W. Hogan, "Directional Derivatives of Extremal-Value Functions with Applications to the Completely Convex Case," *Operations Research* **21**(1) pp. 188-209 (1973).

W.W. Hogan, "Point-to-Set Maps in Mathematical Programming," *SIAM Review* **15** pp. 591-603 (1973).

A.D. Ioffe and V.L. Levin, "Subdifferentials of Convex Functions (in Russian)," *Transactions of the Moscow Mathematical Society* **26** pp. 1-72 (1972).

A.D. Ioffe, "Nonsmooth Analysis: Differential Calculus of Nondifferentiable Mappings," *Transactions of the American Mathematical Society* **266**(1) pp. 1-56 (1981).

V. Kankova, "Differentiability of the Optimal Function in a Two-Stage Stochastic Nonlinear Programming Problem (in Czech)," *Ekonomiko-Matematickyi Obzor* **14**(3) pp. 322-330 (1978).

D. Klatte, "On the Lower Semicontinuity of Optimal Sets in Convex Programming," pp. 104-109 in *Point-to-Set Maps in Mathematical Programming*, ed. P. Huard,Mathematical Programming Study 10, North-Holland, Amsterdam (1979).

K.O. Kortanek, "Perfect Duality in Semi-Infinite and Generalized Convex Programming," *Operations Research Verfahren* **25** pp. 79-88 (1977).

C. Lemarechal and E.A. Nurminski, "Differentiability of a Support Function of an E-Subgradient," WP-80-101, International Institute for Applied Systems Analysis, Laxenburg, Austria (1980).

C. Lemarechal and E.A. Nurminski, "Sur la Differentiabilite de la Fonction d'Appui du Sous-Differentiel Approche," *Comptes Rendus de l'Academie des Sciences (Paris), Serie A* **290** pp. 855-858 (1980).

A.V. Levitin, "A Generalization of the Property of Strong Convexity and General Theorems on the Gradient Method for Minimization of Functionals (in Russian)," pp. 112-120 in *Mathematical Methods for Solution of Economic*

Problems, ed. N.P. Fedorenko and E.G. Golstein,Nauka, Moskow (1977).

E.S. Levitin and B.S. Darhovski, "Quadratic Optimality Conditions for a Class of Nonsmooth Problems of Mathematical Programming (in Russian)," *Doklady Akademii Nauk SSSR* **244**(2) pp. 270-273 (1979).

E.S. Levitin, "Second-Order Conditions in Nonsmooth Mathematical Programming Problems Applied to the Minimax with Bound Variables (in Russian)," *Doklady Akademii Nauk SSSR* **244**(2) pp. 286-290 (1979).

P.O. Lindberg, "A Generalization of Fenchel Conjugation Giving Generalized Lagrangians and Symmetric Nonconvex Duality," in *Survey of Mathematical Programming*, ed. A. Prekopa,North-Holland, Amsterdam - Akademiai Kiado Proceedings of 9th International Mathematical Programming Symposium, (1979).

P.O. Lindberg, "A Relaxational Framework for Duality," Technical Report TRITA-MAT-1981-12, Department of Mathematics, Royal Institute of Technology, Stockholm (1981).

C. Malivert, J.-P. Penot, and M. Thera, "Minimization d'une Fonction Reguliere sur un Ferme non Regulier et Non-Convexe d'un Espace de Hilbert," *Comptes Rendus de l'Academie des Sciences (Paris), Serie A* **286** pp. 165-168 (1978).

O. Mangasarian, "Uniqueness of Solution in Linear Programming," *Linear Algebra and Applications* **25** pp. 152-161 (1979).

H. Massam and S. Zlobec, "Various Definitions of the Derivative in Mathematical Programming," *Mathematical Programming* **7**(2) pp. 144-161 (1974).

E.J. McShane, "Extension of Range of Functions," *Bulletin of the American Mathematical Society* **40** pp. 837-842 (1930).

R.A. Minch, "Applications of Symmetric Derivatives in Mathematical Programming," *Mathematical Programming* **1** pp. 307-321 (1971).

J.J. Moreau, "Fonctionnelles Sous-Differentiables," *Comptes Rendus de l'Academie des Sciences (Paris), Serie A* **257** pp. 4117-4119 (1963).

J.J. Moreau, "Semi-Continuite de Sous-Gradient d'une Fonctionnelle," *Comptes Rendus de l'Academie des Sciences (Paris), Serie A* **260** pp. 1067-1070 (1965).

E.A. Nurminski, "On the Continuity of E-Subdifferential Mappings (in Russian)," *Kibernetika* **13**(5) pp. 148-149 (1977). English translation: Cybernetics Vol.13(5) pp. 790-791(1978).

W. Oettli, "Symmetric Duality, and a Convergent Subgradient Method for Discrete, Linear, Constrained Approximation Problems with Arbitrary Norms Appearing in the Objective Function and in the Constraints," *Journal of Approximation Theory* **14**(1)(1975).

J.-P. Penot, "Calcul Sous-Differentiel et Optimisation," *Journal of Functional Analysis* **27**(2) pp. 248-276 (1978).

J.-P. Penot, "The Use of Generalized Subdifferential Calculus in Optimization Theory," *Operations Research Verfahren* **31** pp. 495-511 (1979).

J.-P. Penot, "A Characterization of Tangential Regularity," *Nonlinear Analysis, Theory, Methods and Applications* 5(6) pp. 625-643 (1981).

J.-P. Penot, "On the Existence of Multipliers in Mathematical Programming in Banach Spaces," pp. 89-104 in *Optimization and Optimal Control*, ed. A.Auslender, W.Oettli, and J.Stoer,Proceedings of a Conference held at Oberwolfach, March 16-22, 1980 Lecture Notes in Control and Information Science, Vol.30, Springer (1981).

. B.N. Pshenichniy, *Lecons sur les Jeux Differentiels*, Controle Optimal et Jeux Differentiels, Vol.4, I.N.R.I.A., Le Chesnay, France (1971).

. B.N. Pshenichniy, "Necessary Extremum Conditions for Differential Inclusions (in Russian)," *Kibernetika* **12**(6) pp. 60-78 (1976).

. B.N. Pshenichniy, "On Necessary Extremum Conditions for Nonsmooth Functions (in Russian)," *Kibernetika* **13**(6) pp. 92-96 (1977).

. S.M. Robinson and R.R. Meyer, "Lower Semicontinuity of Multivalued Linearization Mappings," *SIAM Journal on Control* **11**(3) pp. 525-533 (1973).

. S.M. Robinson and R.H. Day, "A Sufficient Condition for Continuity of Optimal Sets in Mathematical Programming," *Journal of Mathematical Analysis and Applications* **45**(2) pp. 506-511 (1974).

. S.M. Robinson, "Regularity and Stability of Convex Multivalued Functions," *Mathematics of Operations Research* **1**(2) pp. 130-143 (1976).

. S.M. Robinson, "A Characterization of Stability in Linear Programming," *Operations Research* **25**(3) pp. 435-447 (1977).

. S.M. Robinson, "Generalized Equations and Their Solutions, Part 1: Basic Theory," pp. 128-141 in *Point-to-Set Maps in Mathematical Programming*, ed. P. Huard,Mathematical Programming Study 10, North-Holland, Amsterdam (1979).

. S.M. Robinson, "Some Continuity Properties of Polyhedral Multifunctions," pp. 206-214 in *Mathematical Programming at Oberwolfach*, ed. H. Konig, B. Korte, and K. Ritter,Mathematical Programming Study 14, North-Holland, Amsterdam (1981).

. R.T. Rockafellar, "Conjugate Convex Functions in Optimal Control and the Calculus of Variations," *Journal of Mathematical Analysis and Applications* **32** pp. 174-222 (1970).

R.T. Rockafellar, "Augmented Lagrange Multiplier Functions and Duality in Nonconvex Programming," *SIAM Journal on Control* **12** pp. 268-285 (1974).

R.T. Rockafellar and R.J.-B. Wets, "Measures as Lagrange Multipliers in Multistage Stochastic Programming," *Journal of Mathematical Analysis and Applications* **60**(2) pp. 301-313 (1977).

R.T. Rockafellar, "Higher Derivatives of Conjugate Convex Functions," *International Journal of Applied Analysis* 1(1) pp. 41-43 (1977).

R.T. Rockafellar, "Clarke's Tangent Cones and the Boundaries of Closed Sets in R-n," *Nonlinear Analysis* **3**(1) pp. 145-154 (1978).

R.T. Rockafellar, "Directionally Lipschitzian Functions and Subdifferential Calculus," *Proceedings of the London Mathematical Society* **39**(3) pp. 331-355 (1979).

R.T. Rockafellar and J.E. Spingarn, "The Generic Nature of Optimality Conditions in Nonlinear Programming," *Mathematics of Operations Research* **4**(4) pp. 425-430 (1979).

R.T. Rockafellar, "Generalized Directional Derivatives and Subgradients of Nonconvex Functions," *Canadian Journal of Mathematics* **32** pp. 257-280 (1980).

R.T. Rockafellar, "Favorable Classes of Lipschitz Continuous Functions in Subgradient Optimization," WP-81-1, International Institute for Applied Systems Analysis, Laxenburg, Austria (1981).

R.T. Rockafellar, "Augmented Lagrangians and Marginal Values in Parameterized Optimization Problems," WP-81-19, International Institute for Applied Systems Analysis, Laxenburg, Austria (1981).

S. Rolevicz, "On Conditions Warranting Phi-2 Subdifferentiability," pp. 215-224 in *Mathematical Programming at Oberwolfach*, , ed. H. Konig, B. Korte,

and K. Ritter,Mathematical Programming Study 14, North-Holland, Amsterdam (1981).

G. Salinetti and R.J.-B. Wets, "On the Convergence of Sequences of Convex Sets in Finite Dimensions," *SIAM Review* **21** pp. 18-33 (1979).

M. Schechter, "More on Subgradient Duality," *Journal of Mathematical Analysis and Applications* **71**(1) pp. 251-261 (1979).

C. Singh, "A System of Inequalities and Nondifferentiable Mathematical Programming," *Journal of Optimization Theory and Applications* **27**(2) pp. 291-299 (1979).

J.E. Spingarn, "Submonotone Subdifferentials of Lipschitz Functions," *Transactions of the American Mathematical Society* **264**(1) pp. 77-89 (1981).

J.E. Spingarn, "On Optimality Conditions for Structured Families of Nonlinear Programming Problems," *Mathematical Programming* **22**(1) pp. 82-92 (1982).

S. Tanimoto, "Nondifferentiable Mathematical Programming and Convex-Concave Functions," *Journal of Optimization Theory and Applications* **31**(3) pp. 331-342 (1980).

L. Thibault, "Subdifferentiability of Compactly Lipschitzian Vector-Valued Functions," *Trauvux Sem. Anal. Convexe* **8**(5) pp. 1-54 (1978).

L. Thibault, "Subdifferentiability of Compactly Lipschitzian Mappings," *Operations Research Verfahren* **31** pp. 637-645 (1979).

Hoang Tuy, "Stability Property of a System of Inequalities," *Acta Mathematica Vietnamica* **2**(1) pp. 3-16 (1977).

J.P. Vial, "Strong Convexity of Sets and Functions," Discussion Paper 8018, Center for Operations Research and Econometrics, Universite Catholique de Louvain (1980).

J. Warga, "Derivatives, Containers, Inverse Functions and Controllability," pp. 13-46 in *Calculus of Variations and Control Theory*, ed. D.L. Russell,Academic Press, New York (1976).

J. Warga, "Controllability and a Multiplier Rule for Nondifferentiable Optimization Problems," *SIAM Journal on Control and Optimization* **16** pp. 803-812 (1978).

R.J.-B. Wets, "Stochastic Programs with Fixed Recourse: The Equivalent Deterministic Program," *SIAM Review* **16**(3) pp. 309-339 (1974).

R.J.-B. Wets, "Convergence of Convex Functions, Variational Inequalities and Convex Optimization Problems," pp. 376-403 in *Variational Inequalities and Complementarity Problems*, ed. R.W. Cottle, F. Giannessi, and J-L. Lions,Wiley, Chichester (1980).

A.P. Wierzbicki, "Maximum Principle for Semismooth Performance Functionals," *SIAM Journal on Control* **10**(3) pp. 444-459 (1972).

A.P. Wierzbicki and S. Kurcyusz, "Projection on a Cone: Generalized Penalty Functions and Duality Theory for Problems with Inequality Constraints in Hilbert Space," *SIAM Journal on Control and Optimization* **15** pp. 25-56 (1977).

V.I. Zabotin and Yu.A. Polonsky, "Preconvex Sets, Mappings and Their Applications to Extreme-Value Problems (in Russian)," *Kibernetika* **17**(1) pp. 71-74 (1981).

S. Zlobec and B. Mond, "Duality for Nondifferentiable Programming without Constraint Qualification," *Utilitas Mathematica* **15** pp. 291-302 (1979).

S. Zlobec and A. Ben-Israel, "Perturbed Convex Programs. Continuity of Optimal Solutions and Optimal Values," *Operations Research Verfahren* **31** pp. 737-749 (1979).