LAGRANGIAN FUNCTIONS AND NONDIFFERENTIABLE OPTIMIZATION

A.P. Wierzbicki

December 1978

WP-78-63

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

2361
Laxenburg
Austria

International Institute for Applied Systems Analysis

SUMMARY

Relations between a simple type of convex nondifferentiable optimization problem, $\min_x \max_{i \in I} f_i(x)$, and an equivalent differentiable Lagrangian function approach are investigated in order to provide for a better understanding of quasi-Newton methods introduced recently in nondifferentiable optimization techniques. Conditions for superlinear and quadratic convergence of quasi-Newton methods applied to nondifferentiable optimization problems of this simple type are obtained and their possible implications for problems of more general type are examined. Extensions to nonconvex nondifferentiable problems via augmented Lagrangian functions are discussed.

1. INTRODUCTION

A rapid development and intensive research of nondifferentiable optimization techniques -- see [1], [4] -- 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 [5]. 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 a 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:

$$(1) \qquad \underset{x \in X}{\text{minimize}}\ f(x) \quad ; \quad f(x) = \max_{i \in I} f_i(x)$$

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 and convex functions. It is not 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 $\overline{A} \subset A(\hat{x})$, the matrix composed of $f_{ix}(\hat{x})$ for $i \in \overline{A}$ 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 [6]; however, this condition is rarely satisfied in practical problems. Other second-order conditions 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 assumption of convexity is relaxed at the end of this paper. The assumptions of countability of I and finiteness of A could be actually also 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 modified functions f_i are differentiable. It would seem, therefore, that the considered class of nondifferentiable problems could be extended to cover almost all practically encountered problems. However, still other assumptions are needed for the sake of 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 be fully determined, too:

$$(2) \qquad \partial f(x) = \{g \in R^n : g = \sum_{i \in A} \lambda_i f^*_{ix}(x) \ , \ \lambda_i \geq 0 \ , \ \sum_{i \in A} \lambda_i = 1\}$$

where $f^*_{ix}(x)$ are the gradients of functions f_i at x (written as column vectors, hence the transposition sign *). This assumption is not always satisfied in practical problems of nondifferentiable optimization and can be even considered as contradictory to the very nature of nondifferentiable optimization techniques, where a major problem is precisely an estimation of 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 is this assumption 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. This way, the relations of nondifferentiable techniques for solving problem (1) to known techniques of differentiable optimization can be investigated and some strong (superlinear or even quadratic) convergence properties of a special variant of nondifferentiable quasi-Newton techniques can be concluded. The use of an augmented Lagrangian function in the nonconvex case and the strong properties of this function as specified in [10] result also in second-order necessary and sufficient conditions of optimality for the nondifferentiable problem (1). Under second-order sufficient conditions of optimality, the quasi-Newton techniques can be extended to the nonconvex case.

If $A = A(x)$ and $\partial f(x)$ are not known explicitly, and only some subgradients $g \in \partial f(x)$ can be computed without specifying $f_{ix}(x)$ and the baricentric coordinates λ_i, 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, unless some information on the baricentric coordinates is available.

2. A BASIC LEMMA

A fundamental problem 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 $0 \in G$, or, 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:

$$\underset{(y,g)\in YG}{\text{minimize}} \left(\tfrac{1}{2}||g||^2_{H^{-1}} + \sum_{i\in A} \alpha_i y_i\right) \quad ;$$

(3)

$$YG = \{(y,g) : g = \sum_{i\in A} y_i g_i \ , \ \sum_{i\in A} y_i = 1 \ , \ y_i \geq 0 \ , \ i\in A\} \quad .$$

The following lemma is actually only an extension of the results given in [5], 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

$$\underset{(\bar{x}_0,\bar{x})\in\bar{X}_0}{\text{minimize}} \; (\bar{x}_0 + \tfrac{1}{2}\|\bar{x}\|_H^2) \quad ;$$

(4)

$$\bar{X}_0 = \{(\bar{x}_0,\bar{x})\in R^{n+1} : <g_i,\bar{x}> - \bar{x}_0 - \alpha_i \le 0, i\in A\}$$

where $\|\bar{x}\|_H = <\bar{x},H\bar{x}>$ and 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 $<g_i,\bar{x}> - \bar{x}_0 - \alpha_i \le 0$ and with $\hat{\bar{x}} = -H^{-1}\hat{g}$, $\hat{\bar{x}}_0 = -\|\hat{g}\|_{H-1}^2 - \sum_{i\in A}\alpha_i y_i \le 0$. The following equivalences hold:

$$\hat{\bar{x}} = 0 \iff \hat{g} = 0 \iff \hat{\bar{x}}_0 = -\sum_{i\in A}\alpha_i\hat{y}_i = -\min_{i\in A}\alpha_i \iff$$

$$\hat{\bar{x}}_0 = 0 \text{ if any of } \alpha_i = 0 \quad ;$$

generally, $-\hat{\bar{x}}_0 \ge \|\hat{\bar{x}}\|_H^2 \ge 0$ 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 naturally minimizes $\sum_{i\in A}\alpha_i\tilde{y}_i$ over $\tilde{y}\in\tilde{Y} = \{y : y_i \ge 0, \sum_{i\in A} y_i = 1, \sum_{i\in A} y_i g_i = \hat{g}\}$.

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) has clearly 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 $\alpha_i \neq 0$ such that $\sum_{i \in I} \alpha_i h_i = 0$, is equivalent to the existence of $\alpha_i \neq 0$ such that $\sum_{i \in I} \alpha_i = 0$, $\alpha_j = - \sum_{i \neq j} \alpha_i$, and $\alpha_j g_j = - \sum_{i \neq j} \alpha_i g_i$, which, in turn, is equivalent to the existence of $\lambda_i = - \frac{\alpha_i}{\alpha_j}$, $\sum_{i \neq j} \lambda_i = 1$ and $g_j = \sum_{i \neq j} \lambda_j g_i$. 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 $\hat{g} = \sum_{i \in \bar{A}} \tilde{y}_i g_i$, $\sum_{i \in \bar{A}} \tilde{y}_i = 1$, $\tilde{y}_i \geq 0$, and put $\tilde{y}_i = 0$ for $i \notin \bar{A}$. 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 $\hat{\bar{x}}_{01} + \frac{1}{2} \|\hat{\bar{x}}_1\|_H^2 = \hat{\bar{x}}_{02} + \frac{1}{2} \|\hat{\bar{x}}_2\|_H^2$ implies $\hat{\bar{x}}_{01} = \hat{\bar{x}}_{02}$ if $\hat{\bar{x}}_1 = \hat{\bar{x}}_2$; were $\hat{\bar{x}}_1 \neq \hat{\bar{x}}_2$, so would $\bar{x}_0 = \beta \hat{\bar{x}}_{01} + (1 - \beta) \hat{\bar{x}}_{02}$, $\bar{x} = \beta \hat{\bar{x}}_1 + (1 - \beta) \hat{\bar{x}}_2$ for $\beta \in (0;1)$ yield a smaller value of $\bar{x}_0 + \frac{1}{2} \|\bar{x}\|_H^2$.

Consider now problem (3) and define the Lagrangian function:

$$(5) \quad \mathscr{L}(\bar{x}, \bar{x}_0, y, g) \overset{\text{df}}{=} \frac{1}{2} \|g\|_{H^{-1}}^2 + \sum_{i \in A} \alpha_i y_i + \langle \bar{x}, g - \sum_{i \in A} y_i g_i \rangle + \bar{x}_0 (\sum_{i \in A} y_i - 1)$$

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 a saddle-point of the function (5) under the additional constraint $y_i \geq 0$. Hence, if $\hat{\tilde{X}}$, $\hat{\tilde{X}}_0$ 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:

$$
(6)\qquad
\begin{aligned}
\hat{\hat{X}} \times \hat{\hat{X}}_0 \times \hat{Y} \times \{\hat{g}\} &= \operatorname{Arg} \min_{y \geq 0,\, g \in R^n} \max_{(\bar{x},\bar{x}_0) \in R^{n+1}} \mathscr{L}(\bar{x},\bar{x}_0,y,g) = \\
&= \operatorname{Arg} \max_{(\bar{x},\bar{x}_0) \in R^{n+1}} \min_{y \geq 0,\, g \in R^n} \mathscr{L}(\bar{x},\bar{x}_0,y,g)
\end{aligned}
$$

where Arg min max is the set of points resulting in min max, etc. Compute $\tilde{g}$ minimizing $\mathscr{L}$ for a given $\bar{x}$, $\bar{x}_0$, y. Clearly, $\tilde{g} = \tilde{g}(\bar{x}) = -H\bar{x}$; this implies that $\hat{\bar{x}} = -H^{-1}\hat{g}$ and $\hat{\hat{X}} = \{\hat{\bar{x}}\}$, $\hat{\bar{x}}$ is unique. Moreover, after easy computation

$$
(7)\qquad
\begin{aligned}
\mathscr{L}(\bar{x},\bar{x}_0,y,\tilde{g}(\bar{x})) &= -\tfrac{1}{2}\|\bar{x}\|_H^2 - \bar{x}_0 - \sum_{i \in A} y_i(<g_i,\bar{x}>-\bar{x}_0-\alpha_i) = \\
&\overset{df}{=} -\bar{\mathscr{L}}(y,\bar{x}_0,\bar{x})
\end{aligned}
$$

and, at the saddle-point, $-\frac{1}{2}\|\hat{\bar{x}}\|_H^2 - \hat{\bar{x}}_0 = \frac{1}{2}\|\hat{g}\|_{H^{-1}}^2 + \sum_{i \in A} \alpha_i\hat{y}_i$; since $\|\hat{g}\|_{H^{-1}}^2 = \|\hat{\bar{x}}\|_H^2$, this implies $\hat{\bar{x}}_0 = -\|\hat{\bar{x}}\|_H^2 - \sum_{i \in A} \alpha_i\hat{y}_i$. Hence $\hat{\bar{x}}_0$ is also unique, $\hat{\tilde{X}}_0 = \{\hat{\bar{x}}_0\}$. Obviously, $\hat{\bar{x}} = 0 \Leftrightarrow \hat{g} = 0$ and $\hat{\bar{x}}_0 = 0 \Rightarrow \hat{\bar{x}} = 0$, $\hat{g} = 0$.

The function $\bar{\mathscr{L}}$ 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 $<g_i,\bar{x}_1> - \bar{x}_{01} - \alpha_i < 0$ for all $i \in A$. Moreover, it is well known that

$$\operatorname{Arg} \min_{(\bar{x}_0,\bar{x}) \in R^{n+1}} \max_{y \geq 0} \bar{\mathscr{L}}(y,\bar{x}_0,\bar{x}) = \{\hat{\bar{x}}_0\} \times \{\hat{\bar{x}}\} \times \hat{\hat{Y}} \quad ,$$

where $\hat{\bar{x}}_0$, $\hat{\bar{x}}$ are unique solutions of (4) and $\hat{\hat{Y}}$ is the set of corresponding Lagrange multipliers. But relation (7) implies that:

$$(8) \quad \hat{\hat{X}} \times \hat{\hat{X}}_0 \times \hat{Y} = \operatorname{Arg} \min_{y \geq 0} \max_{(\bar{x},\bar{x}_0) \in R^{n+1}} \mathscr{L}(\bar{x},\bar{x}_0,y,\tilde{g}(\bar{x})) =$$

$$= \operatorname{Arg} \max_{y \geq 0} \min_{(\bar{x},\bar{x}_0) \in R^{n+1}} \bar{\mathscr{L}}(y,\bar{x}_0,\bar{x}) = \{\hat{\bar{x}}\} \times \{\hat{\bar{x}}_0\} \times \hat{\hat{Y}} \quad .$$

Hence $\hat{\hat{Y}} = \hat{Y}$. 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 [12],[14]. 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, $\hat{\bar{x}}_0 = \sum_{i \in A} \alpha_i \hat{y}_i = -\min_{i \in A} \alpha_i \Rightarrow \hat{g} = 0 \Leftrightarrow \hat{\bar{x}} = 0$.

A large part of the above lemma can be found in [5], however, without the full interpretation of $\hat{\bar{x}}_0$, $\hat{\bar{x}}$ as Lagrange multipliers for (3) and without the uniqueness nor Lipschitz-continuity arguments. It is also observed in [5] 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 for 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 [15],[2]; 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 allows also a straight-forward generalization for problems with infinite and uncountable numbers of variables and constraints in Hilbert spaces.

3. QUASI-NEWTON METHODS IN NONDIFFERENTIABLE CONVEX 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:

$$\text{(9)} \qquad \underset{(x_0,x) \in X_0}{\text{minimize}}\; x_0 \quad ; \quad X_0 = \{(x_0,x) \in R^1 \times X : f_i(x) - x_0 \leq 0, i \in I\}$$

with the activity set $A(x)$ defined equivalently by

$$\text{(10)} \qquad A(x) = \{i \in I : f_i(x) - \hat{x}_0(x) = 0\; ; \; \hat{x}_0(x) = \max_{i \in I} f_i(x)\} \quad .$$

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:

$$\text{(11)} \quad L(y,x_0,x) = x_0 + \sum_{i \in I} y_i(f_i(x) - x_0) = x_0(1 - \sum_{i \in I} y_i) + \sum_{i \in I} y_i f_i(x)$$

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 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 such primal-dual methods is probably the most efficient algorithm of nondifferentiable optimization.

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

$$(12) \qquad L^*_x(y,x_0,x) = \sum_{i \in A} y_i f^*_{ix}(x) \underset{(\text{if } A = A(x))}{=} g \in \partial f(x) \quad .$$

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 be also equivalent to minimizing the function:

$$(13) \qquad F(\hat{y},x) = \sum_{i \in A(\hat{x})} \hat{y}_i f_i(x) \quad .$$

However, not only the optimal values $\hat{y}_i$ are not known, but also the activity set $A(x)$ changes, often in arbitrary neighborhood of $\hat{x}$. Also, the strong activity set

$$(14) \qquad SA(\hat{y}) = \{i \in I : \hat{y}_i > 0\}$$

can be made stable in the neighborhood of $(\hat{y},\hat{x})$ only if $y_i > 0$ are admissible for some $i \notin A(x)$. These difficulties are not uniquely related to nondifferentiable problems; they are also known in constraint differentiable problems. A typical way of resolving them -- see, e.g., [14] -- 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

$$(15a) \qquad S \subset SA(y) \quad , \quad A(x) \subset A \quad , \quad S \subset A$$

and that, for (y,x) in some neighborhood of $(\hat{y},\hat{x})$:

$$(15b) \qquad S = SA(\hat{y}) \subset A(\hat{x}) = A \quad .$$

A measure of the distance from (y,x) to $(\hat{y},\hat{x})$ is useful when constructing such approximations. Define

$$(16a) \qquad w = \|(L_x,\tilde{L}_y)\|$$

where

$$(16b) \qquad L_x = \sum_{i\in A} y_i f_{ix}(x) \quad ; \quad \tilde{L}_{y_i} = \begin{cases} 0, & y_i = 0 \\ y_i(f_i(x) - \hat{x}_0(x)), & y_i > 0 \end{cases} .$$

Here $\tilde{L}_{y_i}$ is not precisely the derivative of L in y_i, but measures the violation of Kuhn-Tucher necessary conditions of optimality of $(\hat{y},\hat{x})$; if $A(\hat{x}) \subset A$ and $L_x = 0$, $\tilde{L}_{y_i} = 0$ for $i \in A$, $w = 0$, then clearly, $y = \hat{y}$ and $x = \hat{x}$. Moreover, the following lemma holds:

Lemma 2. Suppose $\hat{x}$ is an optimal solution of problem (9), $\hat{x} \in \operatorname{int} X$, and let $\hat{y}$ be the corresponding vector of Lagrange multipliers, with $\sum_{i\in I} \hat{y}_i = 1$. Suppose that the vectors $h_i = (-1, f_{ix}(\hat{x})) \in R^{n+1}$ are linearly independent for $i \in A(\hat{x})$ (hence, $\hat{y}$ is unique) and let the matrix $\hat{L}_{xx} = \sum_{i\in SA(\hat{y})} \hat{y}_i f_{ixx}(\hat{x})$ be positive definite (hence, $\hat{x}$ is unique). Then there exists a neighborhood $U(\hat{y},\hat{x})$ and a constant $\delta > 0$ such that:

$$(17) \qquad \|(y-\hat{y}, x-\hat{x})\| \leq \delta \cdot w \quad \text{for all } (y,x) \in U(\hat{y},\hat{x})$$

where w is defined by (16a,b) with $S = SA(y)$, $A \supset A(x)$.

For the proof of the lemma see, e.g. [14]. Lemmas 1,2.

Observe that the assumption that the vectors $h_i = (-1, f_{ix}(\hat{x})) \in R^{n+1}$, $i \in I$ are linearly independent is much weaker than the Haar condition. In terms of $f_{ix}(\hat{x})$, the linear independence of h_i is equivalent to the fact that no $f_{jx}(\hat{x})$ can be expressed as a convex combination of other $f_{ix}(\hat{x})$, see the proof of Lemma 1. The Haar condition that each collection of $f_{ix}(\hat{x})$ forms a matrix of maximal rank implies that there should be at least $(n+1)$ of the vectors $f_{ix}(\hat{x})$, $|A(\hat{x})| \geq n+1$ (it is easy to show that otherwise either $\hat{x}$ is not optimal, or the Haar condition is not satisfied), each set of n of the vectors $f_{ix}(\hat{x})$ is linearly independent, and that at least $n+1$ of the baricentric coordinates $\hat{y}_i$ of $0 \in \partial f(\hat{x})$ are positive. In terms of the vectors $h_i = (-1, f_{ix}(\hat{x}))$ the Haar condition has an equivalent formulation that the vectors h_i, $i \in I$, span the entire space R^{n+1}, that is, the cone $\{h_i\}_{i\in I} = \{h \in R^{n+1} : h = \sum_{i\in I} \lambda_i h_i, \lambda_i \geq 0\}$ has nonempty interior in R^{n+1} and, if $\hat{x}$ is optimal, the vector $e_0 =$ $= (-1,0,\ldots,0) \in \text{int cone } \{h_i\}_{i\in I}$. Thus, the last form of the Haar condition has a rather straightforward geometric interpretation; this interpretation is closely related to the necessary condition of optimality $0 \in \partial f(\hat{x})$, which takes the form $e_0 \in \text{cone } \{h_i\}_{i\in I}$ in terms of h_i. The assumption in Lemma 2 that h_i are linearly independent is satisfied very often for $|A(\hat{x})| < n+1$. Generally, the use of vectors h_i, though not quite common in the theory of nondifferentiable optimization, gives a nice insight into the fundamental conditions of optimality.

Consider now an approximation of the subdifferential $\partial f(x)$ by the set G:

$$(18) \qquad G = \{g \in R^n : g = \sum_{i\in A} y_i f^*_{ix}(x) ; \sum_{i\in A} y_i = 1 , y_i \geq 0 , i \in A\}$$

and assume that $0 = \sum_{i\in A} \hat{y}_i f_{ix}(x) \in G$. Although $\partial f(x) \subsetneq G$ if $A(x) \subsetneq A$ and $f_{ix}(x)$ for $i \in A \setminus A(x)$ are not convex combinations of $f_{ix}(x)$, $i \in A(x)$, 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):

$$(19a) \qquad \underset{(y,g)\,\in\, YG}{\text{minimize}} \left(\frac{1}{2}\,\|g\|^2_{H^{-1}} + \sum_{i\in A} x_i y_i\right) ; \quad \alpha_i = \hat{x}_0(x) - f_i(x) \quad ;$$

$$YG = \left\{(y,g) : g = \sum_{i\in A} y_i\, f^*_{ix}(x) ; \sum_{i\in A} y_i = 1 , y_i \geq 0\right\}$$

where H^{-1} is a positive definite matrix, not chosen yet. But, due to Lemma 1, (19a) is equivalent to:

$$(19b)\qquad \underset{(\bar{x}_0,\bar{x})\in\bar{X}_0}{\text{minimize}}\ (\bar{x}_0 + \tfrac{1}{2}||\bar{x}||_H) \quad ;$$

$$\bar{X}_0 = \{(\bar{x}_0,\bar{x}) \in R^{n+1} : f_{ix}(x)\bar{x} - \bar{x}_0 - \hat{x}_0(x) + f_i(x) \leq 0, i \in A\}$$

and the choice of H^{-1} or H is now clear: (19b) is a well-known quadratic approximation problem for the Lagrangian function (11), see, e.g. [12],[14], and the optimal choice of the matrix H is to approximate the Hessian of the Lagrangian function (11) as closely as possible,

$$(19c)\qquad H \approx L_{xx}(\hat{y},\hat{x}_0,\hat{x}) = \hat{L}_{xx} = \sum_{i\in SA(\hat{y})} \hat{y}_i f_{ixx}(\hat{x}) \quad ,$$

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 :

$$(19d)\qquad w = (\tfrac{1}{2}||\sum_{i\in A} y_i f_{ix}(x)||^2_{H^{-1}} + \sum_{i\in A} y_i(\hat{x}_0(x) - f_i(x)))^{\frac{1}{2}} \quad .$$

But this coincides precisely with the minimized function in (19a) and can be interpreted as follows: given a point (y,x) and the set A, w or $(w)^2$ can be determined from (19d). By solving (19a) in y, new $\hat{y}$, $\hat{\bar{x}}$, $\hat{\bar{x}}_0$ and:

$$(19e)\qquad (\hat{w})^2 = \tfrac{1}{2}||\hat{g}||^2_{H^{-1}} + \sum_{i\in A} \alpha_i\hat{y}_i = \tfrac{1}{2}||\hat{\bar{x}}||^2_H + \sum_{i\in A} \alpha_i\hat{y}_i =$$

$$= -\hat{\bar{x}}_0 - \tfrac{1}{2}||\hat{\bar{x}}||^2_H$$

are found. Clearly, $(w)^2 \geq (\hat{w})^2$. But $(\hat{w})^2$ can be interpreted also 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 $\hat{y}$ does not denote the optimal Lagrange multiplier for the original problem, but only for its approximation (19b)) -- see Section 3.4. 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 give also an estimation of the distance $f(x) - f(\hat{x})$ from above, thus being more useful for stopping tests; 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 further sections in more detail, and the above discussion justifies only the alternate use of w, $\hat{w}$ and $-\hat{\bar{x}}_0$ in the proposed algorithms.

3.2. 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), by S^k - the approximation of the set $SA(\hat{y})$. If (y^k, x^k) converge to $(\hat{y}, \hat{x})$ so that w^k converges to zero, then the following formulae for A^k, S^k can be used:

$$\text{(20a)} \qquad A^k = \{i \in I : f_i^k - x_0^k + \frac{1}{\rho}\, y_i^k \geq -\eta_f^k\}$$

$$\text{(20b)} \qquad S^k = \{i \in A^k : y_i^k \geq \eta_y^k\}$$

where $\rho > 0$ is a chosen constant, depending on the scaling of the problem (see Section 5; 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:

$$\text{(20c)} \qquad \eta_f^k = \xi_f (w^{k-1})^{\frac{1}{2}} \; ; \quad \eta_y^k = \min\,(0.01, \xi_y (w^{k-1})^{\frac{1}{2}})$$

where ξ_f, ξ_y are chosen constant; 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 $S^k = SA(\hat{y})$, $A^k = A(\hat{x})$ for sufficiently small w^k even if the Lipschitz constants are not known explicitly. This follows from the following lemma:

Lemma 3. Suppose $\hat{x}$ is a unique solution of problem (9) and $\hat{y}$ the corresponding unique Lagrange multiplier. Let the sets $A(\hat{x})$, $SA(\hat{y})$ be defined by (10), (14) and A^k, S^k by (20c,b) with $\lim\limits_{k\to\infty} \frac{w^k}{\eta_f^k} = \lim\limits_{k\to\infty} \frac{w^k}{\eta_y^k} = 0$, where w^k is defined by (16a,b). Then there exists a number $\tilde{w} > 0$ such that:

$$(21) \qquad A^k = A(\hat{x}) \quad , \quad S^k = SA(\hat{y})$$

$$\text{for all } (y^k,x^k) \in U(\hat{y},\hat{x}) = \{(y^k,x^k) : w^k < \tilde{w}\} \ .$$

For the proof of the lemma see [14], p.3.3.

However, the above results are valid independently of the norm used when defining w^k. If the norm (19d) is used and $-\hat{\bar{x}}_0^{k-1}$ approximates w^k from above, a more useful expression than (20c) can be obtained. Suppose the range of f, denoted by Rf, can be estimated. Then, after some heuristic reasoning, assuming that the initial $|\hat{\bar{x}}_0^0| = Rf$, $\eta_f^0 = 10^{-2}Rf$ and $\eta_y^0 = 10^{-2}$ and expecting the final accuracy to be related to $|\hat{\bar{x}}_0^k|$ of the order $10^{-6}Rf$, the following expressions:

$$(20d) \qquad \eta_f^k = \xi_f|\hat{\bar{x}}_0^{k-1}|^{1/3} \ ; \ \xi_f = 10^{-2}(Rf)^{2/3} \ ; \ \eta_y^k = \xi_y|\hat{\bar{x}}_0^{k-1}|^{1/3} ;$$

$$\xi_y = 10^{-2}(Rf)^{-1/3} ; \ \rho = \frac{1}{Rf}$$

satisfy the assumptions and result in $\eta_f^k = 10^{-4}Rf$, $\eta_y^k = 10^{-4}$ if $|\hat{\bar{x}}_0^{k-1}| = 10^{-6}Rf$. This means that a function f_i such that $f(\hat{x}) - f_i(\hat{x}) < 10^{-4}Rf$ might be still counted to 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. This can be, however, considered as an acceptable risk -- particularly since it will be shown later that the exact estimation of activity (21) does not influence the simple convergence of algorithms and is needed only when establishing superlinear or quadratic convergence.

3.3. 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 (19a,b) to the Lagrangian function (11). Subroutines for a variable metric approximation of the Hessian matrix of this function (discussed in Section 3.5) 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 or suggested $\varepsilon_{ff} = 10^{-6}Rf$, $\gamma \in (0;1)$-desired ratio 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 $\hat{x}_0^0 = Rf$, $y_i^1 = \frac{1}{|I|}$, $i \in I$, $k = 1$.

Step 1. Compute η_f^k, η_y^k from (20d). Compute f^k and f_i^k for $i \in I$ and determine the sets A^k and S^k (20a,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 remaining y_i^k to obtain $\sum_{i \in A^k} y_i^k = 1$. Compute w^k (19d). If $(w^k)^2 < |\hat{x}_0^{k-1}| < \varepsilon_{ff}$, stop. If $k > 1$, update H^k.

Step 2. Solve the problem (19a) to obtain $\hat{y}^k$, $\hat{g}^k$, compute $\hat{\bar{x}}^k$, $\hat{\bar{x}}_0^k$ from Lemma 1 and $\hat{w}^k$ from (19e).

Step 3. Set $(\tau^k = 1)$ $\tilde{x}^k = x^k + \hat{\bar{x}}^k$. If $|\hat{\bar{x}}_0^k| \leq \gamma|\hat{\bar{x}}_0^{k-1}| \leq \gamma^4 Rf$ and $w^k \leq \gamma w^{k-1}$ is not satisfied, compute $\tilde{f}^k = f(\tilde{x}^k)$. If either:

$$(22a) \qquad f^k + m_a \hat{\bar{x}}_0^k \geq \tilde{f}^k \geq f^k + m_b \hat{\bar{x}}_0^k$$

or $|\hat{\bar{x}}_0^k| \leq \gamma|\hat{\bar{x}}_0^{k-1}| \leq \gamma^4 Rf$ and $w^k \leq \gamma w^{k-1}$, set $x^{k+1} = \tilde{x}^k$, $y^{k+1} = \tilde{y}^k$, $k := k+1$, go to step 1.

Step 4. Perform a linear search for τ^k such that:

$$(22b) \qquad f^k + m_a \tau^k \hat{\bar{x}}_0^k \geq f(x^k + \tau^k \hat{\bar{x}}_0^k) \geq f^k + m_b \tau^k \hat{\bar{x}}_0^k$$

(or any other $\tau^{k'}$ resulting in $f(x^k + \tau^{k'} \hat{\bar{x}}_0^k) < f(x^k + \tau^k \hat{\bar{x}}_0^k)$ where τ^k satisfies (22b), see Appendix 1). Set $x^{k+1} = x^k + \tau^k \hat{\bar{x}}^k$, $y^{k+1} = y^k + \tau^k(\hat{y}^k - y^k)$, $k := k+1$, go to step 1.

Comments. Observe that, when computing f^k, all f_i^k for $i \in I$ must be evaluated. It is best to combine this with the determination of sets A^k, S^k, saving f_i^k only for $i \in A^k$. But it is not known whether $\tau^k = 1$ will be accepted when checking condition (22a). Therefore, if $|\hat{\bar{x}}_0^k|$ is already small enough and decreases and the desired convergence ratio γ for w^k is attained, $\tau^k = 1$ is accepted without checking. Actually, w^k is computed only for this purpose -- and for double-checking the stopping test. Other redundant information, as the sets S^k, values $\hat{g}^k$, $\hat{w}^k$, or even the rescaled values y_i^k, could also not be computed if the computation of w^k were deemed unnecessary. But this information is valuable for the analysis of the algorithm and 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 either in [5] or in [12],[14]. It is only necessary to note that $(w^k)^2 < |\hat{\bar{x}}_0^{k-1}|$ will be eventually satisfied, if $|\hat{\bar{x}}_0^k|$ converges to zero (see Section 3.4) and that $w^k \leq \gamma w^{k-1}$ implies convergence if $|\hat{\bar{x}}_0^k|$ is small enough and decreases. Actually, the double-check in step 3 is also redundant, since the linear convergence of $|\hat{\bar{x}}_0^k|$ alone implies convergence of the algorithm in the convex case; but the algorithm is constructed to be applicable also for only locally convex cases.

Theorem 4. 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 $h_i = (-1, f_{ix}(\hat{x})) \in R^{n+1}$ be linearly independent for $i \in A(\hat{x}) = \{i \in I : f_i(\hat{x}) = f(\hat{x})\}$ implying 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 such a neighborhood of point $\hat{x}$ 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 $(\hat{y}, \hat{x})$.

For proof of the theorem combine the results given, for example, in [5], [12].

3.4. Properties of Quadratic Approximations to Lagrange Functions

For the superlinear or quadratic convergence of the above algorithm, two basic properties of the quadratic approximation problems (19a,19b) are of importance.

Lemma 5. Let the assumptions of Theorem 4 and Lemma 3 hold. Then there exist 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 $(19a) \Leftrightarrow (19b)$ have solutions with $\hat{\bar{x}}^k$, $\hat{y}^k = y^k + \hat{\bar{y}}^k$ satisfying the following inequality:

$$(23) \qquad \|\hat{\bar{y}}^k, \hat{\bar{x}}^k\| \leq \beta w^k$$

where w^k is defined as in (16a,b) with any norm, for example, with the norm (19d).

For a general proof of the lemma, see, e.g., [14]; when using the norm (19d) for w^k, the proof becomes quite straightforward.

Lemma 6. Let the assumptions of Theorem 4 and Lemma 3 hold. Suppose the solutions of problems (19a)$\Leftrightarrow$(19b) define $x^{k+1} = x^k + \hat{\bar{x}}^k$, $y^{k+1} = \hat{\bar{y}}^k = y^k + \hat{y}^k$. Then w^{k+1} defined as in (16a,b) with any norm and at the point (y^{k+1}, x^{k+1}) satisfies the following inequality:

$$(24) \qquad w^{k+1} \leq \| (H^k - L^k_{xx}) \hat{\bar{x}}^k \| + o(\hat{\bar{y}}^k, \hat{\bar{x}}^k)$$

where $L^k_{xx} = \sum_{i \in A^k} y^k_i f_{ixx}(x^k)$ and $o(z)$ denotes a function such that $\lim_{\|z\| \to 0} \frac{o(z)}{\|z\|} = 0$.

For a general proof of the lemma, see, e.g., [12],[14]; again, the proof can be simplified when considering the particular norm (19d) for w^{k+1}.

Many further conclusions can be drawn from a more detailed analysis of Lemmas 1,5,6 and the specifical norm (19d) for w. For example, the general relation (24) can be transformed to the form:

$$(25) \qquad (w^{k+1})^2 \leq - \hat{\bar{x}}^k_0 + $$

$$+ \left(\| (H^k - L^k_{xx}) \hat{\bar{x}}^k \|_{(H^k)^{-1}} - \| H^k \hat{\bar{x}}^k \|^2_{(H^k)^{-1}} \right) + o^2 (\hat{\bar{x}}^k, \hat{\bar{y}}^k)$$

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 6 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.

3.5. 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 $\hat{x}$, the sets A^k with the property $A^k = A(\hat{x})$ even if evaluated at (y^k,x^k) in a neighborhood of $(\hat{y},\hat{x})$ were defined. If the following (matrix-valued) function is defined:

(26a) $$\tilde{L}^k_{xx} = \tilde{L}_{xx}(\hat{y}^k,x^k) = \sum_{i\in A^k} \hat{y}^k_i f^k_{ixx}$$

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 6 instead of L^k_{xx}. It is the 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:

(26b) $$\tilde{L}^k_{xx}\, s^j = r^j + o(s^j,\hat{\bar{y}}^j,\ldots,s^k,\hat{\bar{y}}^k)$$

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

(26c) $$s^j = x^j - x^{j-1} \quad (= \hat{\bar{x}}^{j-1}, \text{ if } \tau^{j-1} = 1)$$

(26d) $$r^j = \sum_{i\in A^k} \hat{y}^j_i \,(f^{j*}_{ix} - f^{(j-1)*}_{ix}) \quad .$$

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 (26b) 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:

$$(27a) \qquad H^k s^k = r^k$$

$$(27b) \qquad H^k s^j = r^j + o(s^j, \hat{y}^j, \ldots, s^{k-1}, \hat{y}^{k-1}) \quad , \quad j < k$$

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 (27b). If a rank-one variable metric is used, relations (27a,b) hold independently from step-size coefficients and from the choice of directions; on the other hand, a rank-one variable metric approximation H^k can become ill-definite even if $\tilde{L}^k_{xx}$ are positive definite. However, there are special variants of the rank-one variable metric that guarantee the positive definiteness of H^k [3].

If $N \geq n$ and the data $\{s^j\}^k_{k-N+1}$ span entire R^n, then it can be shown [3] that the relations (26b) and (27a,b) imply together, that

$$(28) \qquad (\tilde{L}^k_{xx} - H^k) s^{k+1} = o(\underset{\sim}{s}^k, \underset{\sim}{\hat{y}}^k) \quad ; \quad \underset{\sim}{s}^k = (s^{k+1}, s^k, \ldots, s^{k-N+1}) \quad ;$$

$$\underset{\sim}{\hat{y}}^k = (\hat{y}^k, \ldots, \hat{y}^{k-N+1}) \quad .$$

If $s^j = \hat{x}^{j-1}$, then the estimate (28) together with (24) from Lemma 6 results in the superlinear convergence of a quadratic approximation method, see next section. Note, however, that the estimate (28) does not imply (although 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 the reason why the quadratic convergence of a quasi-Newton method can be obtained practically only when $H^k = \tilde{L}^k_{xx}$ is explicitly computed.

3.6. Superlinear and Quadratic Convergence of Quadratic Approximation Methods

Lemmas 5, 6 together with the properties of variable metric H^k result in the following theorem:

Theorem 7. Let the assumptions of Theorem 4 and Lemma 3 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 $\|(\tilde{L}^k_{xx} - H^k)\,\hat{\bar{x}}^k\| \leq \xi w^k$, then $w^{k+1} < \gamma w^k$ and $|\hat{\bar{x}}_0^{k+1}| < \gamma|\hat{\bar{x}}_0^k|$ and the algorithm from Section 3.3 converges with the desired convergence rate. If

$$\lim_{k\to\infty} \frac{\|(\tilde{L}^k_{xx} - H^k)\,\hat{\bar{x}}^k\|}{w^k} = 0 \quad ,$$

then the algorithm converges superlinearly, $\lim\limits_{k\to\infty} \frac{w^{k+1}}{w^k} = 0$. If $\tilde{L}^k_{xx} = H^k$ and the second-order derivatives $f_{ixx}(\cdot)$, $i \in A(\hat{x})$, are Lipschitz-continuous, then the algorithm converges quadratically, $\limsup\limits_{k\to\infty} \frac{w^{k+1}}{(w^k)^2} = \bar{a} < +\infty$.

The proof of the theorem is quite standard -- see, for example, the proof of Theorem 1 in [14] -- and is omitted here.

It is worth to note that practical experience with quadratic approximation methods shows that they are the most efficient among a wide class of various algorithms for constrained differentiable optimization [12]. A similar performance might be expected for the algorithm from Section 3.3, 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 [7] describing an algorithm of a similar nature -- though different in many details and in the theoretical justification -- for the same class of problems; the results of numerical tests given in [7] confirm the expectation of a high practical efficiency of the algorithm 3.3.

4. POSSIBLE EXTENSIONS AND RELATED RESULTS

4.1. Nonconvex Nondifferentiable Optimization with Explicitly Given Subdifferentials

If the functions f_i in problem (1) are not even locally convex, then the Lagrangian function (11) for the equivalent problem (9) might have no saddle points, although the problems (1),(9) can

in column form), then $\hat{x}$ is a locally unique solution of problems (1),(9), and $(\hat{y},\rho,\hat{x}_0,\hat{x})$ with $\hat{x}_0 = f(\hat{x})$ is a saddle-point of the function (29).

Similarly, the following necessary condition can be obtained:

Second-order necessary condition of optimality for nondifferentiable nonconvex problem (1) with explicitly given subdifferentials. If $\hat{x} \in \text{int } X$ is a solution to problem (1) where the functions f_i are twice differentiable, and if the vectors $\hat{h}_i = (-1,f_{ix}(\hat{x})$ are linearly independent for $i \in A(\hat{x}) = \{i \in I: f_i(\hat{x}) = f(\hat{x})\}$, then there exist $\hat{y}_i \geq 0$, $\sum_{i \in A(\hat{x})} \hat{y}_i = 1$, such that $\sum_{i \in A(\hat{x})} \hat{y}_i f_{ix}(\hat{x}) = 0$. Moreover, there exists a $\bar{\rho} \geq 0$ such that for $\rho \geq \bar{\rho}$ the following matrix Λ^A_{xx} is positive semidefinite

$$(31) \qquad \Lambda^A_{xx} = \sum_{i \in A(\hat{x})} \hat{y}_i f_{ixx}(\hat{x}) + \rho \sum_{i \in A(\hat{x})} f^*_{ix}(\hat{x}) f_{ix}(\hat{x}) \quad .$$

The derivation of these conditions from the known properties of augmented Lagrangian functions [11], see also [14], is rather technical, and is omitted here. If the functions f_i are convex, then $\rho = 0$ can be used in both (30),(31). If $\rho > 0$ and $S(\hat{y}) \neq A(\hat{x})$, then the augmented Lagrangian function (29) is not twice differentiable at $(\hat{y},\rho,\hat{x}_0,\hat{x})$, but the matrices Λ^S_{xx} and Λ^A_{xx} give a lower and an upper approximation to the Hessian Λ_{xx} -- see [14]. Following the results given in [14], a quadratic approximation algorithm extending the algorithm from Section 3.3 to the even locally nonconvex case can be derived. The algorithm uses the sets S^k (defined redundantly in algorithm 3.3) in order to determine convexifying terms for the quadratic approximation problem (19b), which takes now the form:

$$(32a) \qquad \underset{(\bar{x}^k_0,\bar{x}^k) \in \bar{X}^k_0}{\text{minimize}} \; (\bar{x}^k_0 + \frac{1}{2}\langle \bar{x}^k, H^k \bar{x}^k \rangle + \rho \sum_{i \in S^k} (\frac{1}{2}(f^k_{ix}\bar{x}^k - \bar{x}^k_0)^2 - \alpha^k_i (f^k_{ix}\bar{x}^k - \bar{x}^k_0)))$$

or, equivalently:

$$\text{(32b)} \quad \underset{(\bar{x}_0^k,\bar{x}^k)\in\bar{X}_0^k}{\text{minimize}} \quad (\bar{x}_0^k(1+\rho\sum_{i\in S^k}\alpha_i^k) + \frac{1}{2}\rho|S^k|(x_0^k)^2 + \frac{1}{2}\langle \bar{x}^k,(H^k+\rho F^{k*}F^k)\bar{x}^k\rangle + $$

$$+\rho\sum_{i\in S^k}(\alpha_i^k + \bar{x}_0^k)f_{ix}^k\bar{x}^k)$$

where F^k is a matrix composed of 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:

$$\text{(32c)} \qquad \bar{X}_0^k = \{(\bar{x}_0^k,\bar{x}^k) \in R^{n+1} : f_{ix}^k\bar{x}^k - \bar{x}_0^k - \alpha_i^k \le 0 \text{ , } i \in A^k\} \quad .$$

It is interesting to note that, if $S^k = A^k$ and all constraints are active at a solution of (32), then the problem is fully equivalent to a dual problem as in Lemma 1; otherwise, the dual problem for (32) is more complicated, but its investigation 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 entire 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 be proved also for the nonconvex case by a modification of results given in [14].

4.2. Nondifferentiable Optimization with Implicitly Given Subdifferentials

For a more general class of problems of nondifferentiable optimization, where $\partial f(x)$ 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, etc., a large number of algorithms has been proposed. (See e.g. [8].) This is motivated by the fact that such problems arise quite often, for example, in large scale algorithms of optimization, as well as in many other cases. However, in most cases some additional knowledge related to baricentric coordinates, etc., is implied by the specific nature of the problem, and the assumption of the lack of such knowledge 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 by Lemarechal in [5], together with convergence proofs. However, it was not specified in [5] what the matrix H^k should approximate; it was only required that H^k is uniformly positive definite, which is sufficient for simple convergence. The results of previous sections of this paper make it clear, that H^k should either approximate (in a sense described in Section 3.5) 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 show also that such an approximation is actually impossible, if no additional knowledge on baricentric coordinates is assumed. The use of subsequent $g_k \in \partial f(x^k)$ gives no second-order information, if $g_k = \sum_{i\in A(x^k)} \lambda_i^k f_{ix}^k$ and 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, $j = 0,\ldots,k$, gives more information for, at least if $\hat{g}^k$ converges to zero, then some corresponding baricentric coordinates should converge to $\hat{y}_i$; but subsequent $\hat{g}^k$ give an average information related to many previous x^j, $j = 0,\ldots,k$, and it is difficult to extract from them the current information related to x^k, 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 only implicitly given subdifferentials; but they show that some stronger assumptions either related to a particular choice of subgradients, or to the basic nature of the problem, are really necessary. For example, if the Haar condition is satisfied, then even a linear approximation algorithm could be superlinearly convergent. In any case, the problem of obtaining superlinearly convergent algorithms for nondifferentiable optimization with implicitly given subdifferentials requires further study.

4.3. Other Extensions and Research Directions

Some of the results of this paper, as, for example, Lemma 1, can be generalized for problems with infinitely and uncountably many constraints. The continuous minmax problem,

$$\underset{x \in X}{\text{minimize}} \ \max_{z \in Z} f(x,z)$$

can be attacked by this approach, and, in the convex case, should not present extreme difficulties; the nonconvex case is then, however, essentially more complex, since only a partial generalization of the augmented Lagrangian theory to infinite-dimensional spaces is now available [13].

REFERENCES

[1] Balinski, M.L. and P. Wolfe, eds., Nondifferentiable Optimization, *Mathematical Programming Study 3*, North-Holland Publ. Co., Amsterdam, 1975.

[2] Hohenbalken, B. von, Least Distance Methods for the Scheme of Polytopes, *Mathematical Programming*, 15(1978), 1-11.

[3] Kreglewski, T. and A.P. Wierzbicki, *Further Properties and Modifications of the Rank-One Variable Metric Method*, International Conference on Mathematical Programming, Zakopane, 1977.

[4] Lemarechal, C., *Nondifferentiable Optimization: Subgradient and ε-subgradient Methods*. Lecture Notes: Numerical Methods in Optimization and Operations Research, Springer Verlag, 1975, 191-199.

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

[6] Madsen, K. and H. Schjaer-Jacobsen, Linearly Constrained Minimax Optimization, *Mathematical Programming*, 14(1977), 208-223.

[7] Madsen, K. and J. Hald, *A Two-stage Algorithm for Minimax Optimization*, Institute for Numerical Analysis, Technical University of Denmark, Report No. NI-78-11, 1978.

[8] Mifflin, R., *An Algorithm for Constrained Optimization With Semismooth Functions*, RR-77-3, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1977.

[9] Nurminski, E.A., The Quasigradient Method for Solving Nonlinear Programming Problems, *Cybernetics*, 9, 1(1973), 145-150, Plenum Publ. Co., New York - London.

[10] Rockafellar, R.T., Augmented Lagrange Multiplier Functions and Duality in Nonconvex Programming, *SIAM Journal for Control and Optimization*, 12(1974), 497-513.

[11] Rockafellar, R.T., Lagrange Multipliers in Optimization, *SIAM-AMS Proc.*, 9(1976), 145-168.

[12] 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.

[13] 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 of Control and Optimization*, 15(1977), 25-56.

[14] Wierzbicki, A.P., *A Quadratic Approximation Method Based on Augmented Lagrangian Functions for Nonconvex Nonlinear Programming Problems*, WP-78- , International Institute for Applied Systems Analysis, Laxenburg, Austria, 1978.

[15] Wolfe, P., Finding a Nearest Point in a Polytope, *Mathematical Programming*, 11(1976), 128-149.

APPENDIX 1

An Efficient Line-Search 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 the previous run of the line-search algorithm). The sequence $\{\tau_i\}$ ends with a value $\tau_g = \tau_i$ satisfying two conditions:

(a) $f_{\tau_i} \leq f(x^k) + m_a \tau_i \hat{x}_0^k$

(b) $f_{\tau_i} \geq f(x^k) + m_b \tau_i \hat{x}_0^k$

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 $f_{\tau_i}$. If $f_{\tau_i} < f_f$, set $\tau_f := \tau_i$, $f_f := f_{\tau_i}$. If
      $f_{\tau_i}$ satisfies (a) and (b), stop.

(ii)  If $f_{\tau_i}$ 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 $f_{\tau_i}$ 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 r-times increased or decreased. $\omega^{i+1} = 2$ means that both a lower bound τ_{min} and an upper bound τ_{max} for τ_f are already found and they should be tightened by computing τ_{i+1} as a geometrical mean of them. The last value of τ_g of τ_i, satisfying (a) and (b), gives often useful information. If some external bounds limit the value of τ_i, the algorithm must be accordingly modified.