ALGORITHM MODELS FOR NONDIFFERENTIABLE OPTIMIZATION

by

E. Polak and D.Q. Mayne

Memorandum No. UCB/ERL M82/34

10 May 1982

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

Algorithm Models for Nondifferentiable Optimization

E. Polak and D. Q. Mayne

Abstract

It is shown that a number of seemingly unrelated nondifferentiable optimization algorithms are special cases of two simple algorithm models: one for constrained and one for unconstrained optimization. In both of these models, the direction finding procedures use parametrized families of maps which are locally uniformly u.s.c. with respect to the generalized gradients of the functions defining the problem. The selection of the parameter is determined by a rule which is analogous to the one used in methods of feasible directions.

Research sponsored by the National Science Foundation grants ECS-79-13148 and CEE-8105790, the Joint Services Electronics Program Contract F49620-79-C-0178, and the UK Science Research Council.

1. Introduction

A formal extension of a differentiable optimization algorithm to the nondifferentiable case consists of replacing gradient vectors $\nabla f(x)$, used by the algorithm in solving a differentiable problem, by the vectors $h(x) = \text{argmin}\{\|h\| \mid h \in \partial f(x)\}$, when applied to a nondifferentiable problem, with $\partial f(x)$ denoting the generalized gradient of $f(x)$, see [C1]. Such formal extensions cannot be shown to converge to stationary points. The reason for this is that while gradients are usually locally uniformly continuous, generalized gradients usually are not even locally uniformly upper-semi-continuous (u. s. c.).

An examination of the nondifferentiable optimization literature, see e. g. [B2, C3, G1, G2, L2-L4, M1, M3, P1-P8], shows that in order to overcome this lack of local uniform upper-semi-continuity, the search direction procedures of nondifferentiable optimization algorithms invariably replace gradients not by generalized gradients, but by better behaved supersets which are obtained in a variety of ways. These supersets reflect the local behavior of the functions in question. When only local Lipschitz continuity is assumed, the supersets consist of bundles of generalized gradients which are generated by exploring a neighborhood about the current iterate, see e. g. [B2, G1, P3]. When the problem functions are convex, subgradient bundles are used as supersets, see e. g. [L2-L4]. When the problem functions are semi-smooth, a special line exploration method can be used to eliminate the need for acquiring a bundle of generalized gradients, see e. g. [M1, M3, P3, P4]. When the problem functions are in some sense piece-wise differentiable and allow one to determine whether one is at a differentiable point or not, the

need for constructing generalized gradient bundles disappears altogether since much simpler supersets can generally be used, as we see from [C3, G2, M3, P1, P5, P6].

In [P7], we find a theory dealing with the extension of differentiable optimization algorithms to the nondifferentiable case. This theory requires the use of bundles of generalized gradients, computed in an ε ball about the current iterate, with the value of $\varepsilon > 0$ controlled by a mechanism analogous to the one used in the Polak method of feasible directions [P9] and in phase I - phase II methods such as those in [P2]. The theory in [P7] does not contribute to the understanding or the construction of algorithms, such as those in [C3, C2, N3, P1, P5, P6], that do not use generalized gradient bundles, and it leads to implementable algorithms only when all the problem functions are semi-smooth.

It has generally been thought that the cumbersome algorithms, which fit within the framework established in [P7], have nothing to do with the highly specialized algorithms in [C3, C2, P5, P6], which exploit the properties of such functions as $f(x) = \max\{\phi(x,t) \mid t \in T\}$, with T a closed interval, or $f(x) = \max$ eigenvalue $(Q(x))$ with $Q(x)$ a differentiable, complex valued Hermitian matrix. It is shown in this paper that this impression is wrong by showing that both classes of nondifferentiable optimization algorithms can be seen as special cases of two simple algorithm models: one for constrained and one for unconstrained optimization. These algorithm models make use of generalized gradient supersets which are locally uniformly u. s. c. with respect to the generalized gradients of the problem functions (a global version of this concept was first used in [P8]). In particular, it is shown in this paper that both the

generalized gradient bundles used in [P7] and the supersets used in the algorithms in [C3, G2, M3, P1, P5, P6] have this local relative u. s. c. property. The algorithms in [C3, G2, M3, P1, P5, P6] solve problems involving functions of the form $f(x) = \phi(g(x))$, where $g: R^n \to R^m$ is continuously differentiable and $\phi: R^m \to R^1$ is locally Lipschitz. It is shown that the supersets used by these algorithms are the generalized gradients of perturbation functions. Some rules for the construction of appropriate perturbation functions are given.

It is to be hoped that as a result of the work reported in this paper, both the exposition of nondifferentiable optimization algorithms and the invention of new ones will be considerably simplified.

2. Unconstrained Optimization

In this section we shall consider algorithm models for solving problems of the form:

$$\min_{x \in \mathbb{R}^n} f(x) \tag{2.1}$$

where $f: \mathbf{R}^n \to \mathbf{R}^1$ is locally Lipschitz continuous. Extensions of our results to normed spaces are quite straightforward and hence will be left to the interested reader.

We recall that a locally Lipschitz function $f(\cdot)$ is differentiable almost everywhere, and that one can define for it a generalized gradient $\partial f(x)$ [C1], by

$$\partial f(x) = \text{co}\{\lim_{i \to \infty} \nabla f(x+v_i)\} \tag{2.2}$$

where the $v_i \to 0$ as $i \to \infty$ are such that $\nabla f(x+v_i)$ exists and co denotes the convex hull of the set in question. It is shown in [C1] that the

map $\partial f(\cdot)$ is u.s.c. in the sense of Berge [B1] and bounded on bounded sets.

When $f(\cdot)$ is only locally Lipschitz, the ordinary directional derivative

$$df(x,h) \triangleq \lim_{\lambda \searrow 0} \frac{f(x+\lambda h)-f(x)}{\lambda} \tag{2.2b}$$

may not exist. Instead, see [C1], one defines the <u>generalized</u> directional derivative of f at x, in the direction h by

$$d_o f(x,h) \triangleq \overline{\lim_{\substack{\lambda \searrow 0 \\ y \to 0}}} \frac{f(x+y+\lambda h)-f(x+y)}{\lambda} . \tag{2.2c}$$

It was shown in [C1] that

$$d_o f(x,h) = \max_{\xi \in \partial f(x)} \langle \xi, h \rangle . \tag{2.2d}$$

As we recall, given an $x_i \in \mathbb{R}^n$ the Armijo gradient method [A1,P9], for differentiable optimization in $\mathbb{R}^n$, first computes the <u>steepest descent</u> direction

$$h(x_i) \triangleq \arg \min_{h \in \mathbb{R}^n} \{\tfrac{1}{2} \|h\|^2 + df(x_i,h)\} \tag{2.3a}$$
$$= -\nabla f(x_i)$$

next, with $\alpha,\beta \in (0,1)$, computes the step size

$$\lambda_i \triangleq \max\{\beta^k | k \in \mathbb{N}_+ , f(x_i+\beta^k h(x_i))-f(x_i) \leq -\beta^k \alpha \|h(x_i)\|^2\} \tag{2.3b}$$

where $\mathbb{N}_+ = \{0,1,2,....\}$; then updates according to

$$x_{i+1} = x_i + \lambda_i h(x_i) . \tag{2.3c}$$

The simplest idea for extending this method (as well as others) to the nondifferentiable case, consists of replacing (2.3a) by

$$h(x_i) \triangleq \arg \min_{h \in \mathbb{R}^n} \{\tfrac{1}{2} \|h\|^2 + d_0 f(x_i, h)\}$$

$$= -\arg \min\{\tfrac{1}{2} \|h\|^2 | h \in \partial f(x_i)\} \tag{2.3d}$$

while leaving (2.3b), (2.3c) unaltered.

Unfortunately, because $\partial f(\cdot)$ is <u>not</u> locally <u>uniformly</u> u.s.c., such extensions fail to be convergent. Consequently, many unconstrained optimization algorithms compute a search direction $h(x_i)$ at x_i by solving an auxiliary problem of the form (2.3d), but with $d_0 f(x_i, h)$ replaced by a kind of ε-generalized derivative $d_\varepsilon f(x_i, h)$, with $\varepsilon \geq 0$, defined by

$$d_\varepsilon f(x,h) \triangleq \max_{\xi \in G_\varepsilon f(x)} (\xi, h) \tag{2.3e}$$

where for every $\varepsilon \geq 0$, and $x \in \mathbb{R}^n$, $\partial f(x) \subset G_\varepsilon f(x)$, and the sets $G_\varepsilon f(x)$ are compact, convex and locally uniformly u.s.c., in a sense to be made clear later, with respect to $\partial f(\cdot)$, thus making up for the lack of local uniform u.s.c. in $\partial f(\cdot)$. We note that because $\partial f(x) \subset G_\varepsilon f(x)$ for all $\varepsilon \geq 0$, we always have $d_0 f(x,h) \leq d_\varepsilon f(x,h)$. When this substitution is made (2.3d) becomes

$$h(x_i) \triangleq \arg \min\{\tfrac{1}{2} \|h\|^2 + d_\varepsilon f(x_i, h)\}$$

$$-\arg \min\{\tfrac{1}{2} \|h\|^2 | h \in G_\varepsilon f(x_i)\}, \tag{2.3f}$$

In addition, a mechanism must be introduced for driving ε to zero. The Polak method of feasible directions [P9] provides an idea for this purpose.

The commonly utilized properties of the maps $G_\varepsilon f(x)$ can be summarized as follows.

Definition 2.1: We shall say that $\{G_\varepsilon f(\cdot)\}_{\varepsilon \geq 0}$, $G_\varepsilon f : \mathbb{R}^n \to 2^{\mathbb{R}^n}$, is a family of convergent direction finding (c.d.f.) maps for the locally Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}^1$ if

(i) for all $x \in \mathbb{R}^n$, $\partial f(x) = G_0 f(x)$;

(ii) for all $x \in \mathbb{R}^n$, $\varepsilon < \varepsilon' \Rightarrow G_\varepsilon f(x) \subset G_{\varepsilon'} f(x)$;

(iii) for any $\varepsilon \geq 0$, $G_\varepsilon(x)$ convex, and it is u.s.c. in (ε, x), in the sense of Berge [B1] at $(0, \hat{x})$, for all $\hat{x} \in \mathbb{R}^n$, and bounded on bounded sets;

(iv) given any $\hat{x} \in \mathbb{R}^n$, $\hat{\varepsilon} > 0$ and $\hat{\delta} > 0$, there exists a $\hat{\rho} > 0$ such that for any $x', x'' \in B(\hat{x}, \hat{\rho}) \triangleq \{x \mid \|x - \hat{x}\| \leq \hat{\rho}\}$ and any $\eta' \in \partial f(x')$ there exists an $\eta'' \in G_\varepsilon f(x'')$ such that $\|\eta'' - \eta'\| \leq \hat{\delta}$. ¤

We note that property (iv) above was referred to in [P8] as "upper-semi-continuity of $G_\varepsilon f(\cdot)$ with respect to $\partial f(\cdot)$" and was found very useful in establishing optimality conditions for minimizing sequences.

The simplest known example of a family of c.d.f. maps (see [P7]) for a function $f(\cdot)$ are the maps $\partial_\varepsilon f(x)$ defined by

$$\partial_\varepsilon f(x) \triangleq \operatorname*{co}_{x' \in B(x,\varepsilon)} \{\partial f(x')\} \tag{2.4}$$

It is obvious by inspection that they satisfy the properties (i)-(iv) in Definition 2.1.

Let $\nu \in (0,1)$, $\varepsilon_0 > 0$, $\delta > 0$ be given. We define the set E by

$$E \triangleq \{0\} \cup \{\varepsilon_0, \nu\varepsilon_0, \nu^2\varepsilon_0, \ldots\} \tag{2.5}$$

and, given a family of c.d.f. maps $\{G_\varepsilon f(\cdot)\}$ for $f(\cdot)$, we define the maps

$h_\varepsilon : \mathbb{R}^n \times \mathbb{R}^1 \to \mathbb{R}^n$ and $\varepsilon : \mathbb{R}^n \to E$ as follows:

$$h_\varepsilon(x) \triangleq -\arg\min\{\tfrac{1}{2}\|v\|^2 \mid v \in G_\varepsilon f(x)\} , \tag{2.6}$$

$$\varepsilon(x) \triangleq \max\{\varepsilon \in E \mid \|h_\varepsilon(x)\|^2 \geq \delta\varepsilon\} . \tag{2.7}$$

The map $\varepsilon(\cdot)$ has the following important property which is crucial to the success of our algorithms.

Proposition 2.1: For every $\hat{x} \in \mathbb{R}^n$ such that $0 \notin \partial f(\hat{x})$, there exists a $\hat{\rho} > 0$ such that $\varepsilon(x) \geq \nu\varepsilon(\hat{x}) > 0$ for all $x \in B(\hat{x},\hat{\rho})$.

Proof: Since $G_\varepsilon f(x)$ is u.s.c. in (ε,x), at $(0,\hat{x})$ for any $\hat{x} \in \mathbb{R}^n$, it follows that $\|h_\varepsilon(x)\|$ is l.s.c. in (ε,x) at $(0,\hat{x})$. Hence, since $0 \notin G_0 f(\hat{x})$, $\varepsilon(\hat{x}) > 0$, and $[\|h_{\varepsilon(\hat{x})}(\hat{x})\|^2 - \delta\varepsilon(\hat{x})] \geq 0$, so that $[\|h_{\nu\varepsilon(\hat{x})}(\hat{x})\|^2 - \delta\nu\varepsilon(\hat{x})] > 0$. It now follows directly by l.s.c. of $[\|h_{\nu\varepsilon(\hat{x})}(x)\|^2 - \delta\nu\varepsilon(\hat{x})]$ in x that there exists a $\hat{\rho} > 0$ such that $\varepsilon(x) \geq \nu\varepsilon(\hat{x})$ for all $x \in B(\hat{x},\hat{\rho})$. ¤

We now proceed to state an algorithm model.

Algorithm Model 2.1.

Parameters: $\delta > 0$, ε_0 (for $\varepsilon(x)$); $\alpha,\beta \in (0,1)$ (for Armijo step size rule).
A family $\{G_\varepsilon(\cdot)\}_{\varepsilon \geq 0}$ of c.d.f. maps.

Data: $x_0 \in \mathbb{R}^n$.

Step 0: Set $i = 0$.

Step 1: Compute $\varepsilon(x_i)$ and $h_i \triangleq h_{\varepsilon(x_i)}$.
If $\varepsilon(x_i) = 0$, stop.

Step 2: Compute the largest $\lambda_i = \beta^{k_i}$, $k_i \in \mathbb{N}_+$ such that

$$f(x_i+\lambda_i h_i) - f(x_i) \leq -\lambda_i \alpha\delta\varepsilon(x_i) . \tag{2.8}$$

Step 3: Set $x_{i+1} = x_i + \lambda_i h_i$, set $i = i+1$ and go to Step 1. ¤

Theorem 2.1: Let $\{x_i\}$ be a sequence constructed by Algorithm Model 2.1. a) If $\{x_i\}$ is finite, with last element x_k, then $0 \in \partial f(x_k)$. b) If $\{x_i\}$ is infinite, then for any accumulation point $\hat{x}$ of $\{x_i\}$, $0 \in \partial f(\hat{x})$ holds.

Proof: a) Since $\varepsilon(x_k) = 0$ if and only if $0 \in \partial f(x_k)$, this part of the theorem is clearly true.

b) Suppose $\{x_i\}$ is infinite and that $x_i \overset{K}{\to} \hat{x}$, with $K \subset \{0,1,2,\ldots\}$ infinite and $0 \notin \partial f(\hat{x})$. Then by Proposition 2.1 there exists an i_0 such that $\varepsilon(x_i) \geq \nu\varepsilon(\hat{x}) > 0$ for all $i \in K$, $i \geq i_0$. Since the sets $G_{\varepsilon_0} f(x_i)$ are bounded on bounded sets and $G_{\varepsilon(x_i)} f(x_i) \subset G_{\varepsilon_0} f(x_i)$ by (ii) of Definition 2.1, it follows that there exists a $b \in (0,\infty)$ such that $\nu\varepsilon(\hat{x})\delta \leq \|h_i\|^2 \leq b$ for all $i \in K$, $i \geq i_0$. Next, by the mean value theorem of Lebourg [L1], for $\lambda \geq 0$,

$$f(x_i+\lambda h_i) - f(x_i) = \lambda \langle h_i, \xi_{i\lambda} \rangle \tag{2.9}$$

with $\xi_{i\lambda} \in \partial f(x_i+s\lambda h_i)$ and $s \in (0,1)$. Referring to (iv) in Definition 2.1, let $\hat{\delta} = (1-\alpha)[\nu\varepsilon(\hat{x})\delta_2]^{1/2}$. Then there exists a $\hat{\rho} > 0$ such that for all $x', x'' \in B(\hat{x},\hat{\rho})$, given any $\eta' \in \partial f(x')$, there exists an $\eta'' \in G_{\nu\varepsilon(\hat{x})} f(x'')$ such that $\|\eta''-\eta'\| \leq (1-\alpha)[\nu\varepsilon(\hat{x})\delta]^{1/2}$. Now let $\hat{\lambda} = \beta^{\hat{k}} \leq \hat{\rho}/2b$, so that if $x_i \in B(\hat{x},\hat{\rho}/2)$, then $(x_i+s\hat{\lambda}h_i) \in B(\hat{x},\hat{\rho})$ for all $s \in (0,1)$. Then there exists an $i_1 \geq i_0$, such that for all $i \in K$, $i \geq i_1$,

$$\begin{aligned} f(x_i+\beta^{\hat{k}}h_i) - f(x_i) &= \beta^{\hat{k}} \langle h_i, \xi_{i\lambda} \rangle \\ &= \beta^{\hat{k}}[\langle h_i, \bar{\xi}_{i\lambda} \rangle + \langle h_i, \xi_{i\lambda}-\bar{\xi}_{i\lambda} \rangle] \end{aligned} \tag{2.10}$$

with $\bar{\xi}_{i\lambda} \in G_{\nu\varepsilon(\hat{x})} f(x_i) \subset G_{\varepsilon(x_i)} f(x_i)$ such that $\|\xi_{i\lambda}-\bar{\xi}_{i\lambda}\| \leq (1-\alpha)(\nu\varepsilon(\hat{x})\delta)^{1/2} \leq (1-\alpha)(\varepsilon(x_i)\delta)^{1/2} \leq (1-\alpha)\|h_i\|$. Since $\langle h_i, \bar{\xi}_{i\lambda} \rangle \leq -\|h_i\|^2$ by construction of h_i, (2.10) now yields that

$$f(x_i+\beta^{\hat{k}}h_i) - f(x_i) \leq \beta^{\hat{k}}[-\|h_i\|^2 + \|h_i\|(1-\alpha)(\varepsilon(x_i)\delta)^{1/2}]$$

$$\leq -\beta^{\hat{k}}\alpha\|h_i\|^2$$

$$\leq -\beta^{\hat{k}}\alpha\delta\varepsilon(x_i)\,. \tag{2.11}$$

Hence for all $i \in K$, $i \geq i_1$, we must have $\lambda_i \geq \beta^{\hat{k}}$ and therefore for all $i \in K$, $i \geq i_1$

$$f(x_{i+1}) - f(x_i) \leq -\beta^{\hat{k}}\alpha\delta\nu\varepsilon(\hat{x})\,. \tag{2.12}$$

Since $\{f(x_i)\}_{i=0}^{\infty}$ is a monotonic decreasing sequence by construction, (2.12) implies that $f(x_i) \to -\infty$ as $i \to \infty$. But by continuity of $f(\cdot)$, and the monotonicity of $\{f(x_i)\}_{i=0}^{\infty}$, we must have that $f(x_i) \to f(\hat{x})$ as $i \to \infty$, and hence we have a contradiction, which completes our proof. ∎

As we have pointed out earlier, the maps $G_\varepsilon f(x) \triangleq \partial_\varepsilon f(x)$ defined in (2.4) are c.d.f. maps. Unfortunately, (see [M1,P7]), only when $f(\cdot)$ is convex or semi-smooth do we know how to construct an adequate approximation to arg $\min\{\|h\| \,|\, h \in \partial_\varepsilon f(x)\}$; consequently implementable algorithms based on $\partial_\varepsilon f(x)$ have been proposed only for these cases.

We now turn to a special class of locally Lipschitz functions $f(\cdot)$ for which it is easy to determine whether any given point x is a point of differentiability or not. For such functions, it is possible to construct much nicer c.d.f. maps than $\partial_\varepsilon f(x)$. An examination of the literature shows that this construction involves the use of the generalized gradients of locally Lipschitz perturbation functions $\hat{f}_v(\cdot)$.

The class of functions we are about to consider have the form

$$f(x) = \phi(g(x)) \tag{2.13}$$

where $g : \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable and $\phi : \mathbb{R}^m \to \mathbb{R}^1$ is locally Lipschitz. We note that by the chain rule [C1]

$$\partial f(x) \subset G\hat{f}_0(x) \triangleq \{\xi | \xi = \frac{\partial g(x)^T}{\partial x} y,$$

$$y \in \partial\phi(z),\ z = g(x)\} \ . \tag{2.14}$$

We now introduce the family of perturbation functions $\{\hat{f}_v(\cdot)\}_{v\in\mathbb{R}^m}$, $\hat{f}_v : \mathbb{R}^n \to \mathbb{R}^1$ defined by

$$\hat{f}_v(x) \triangleq \phi(g(x)+v) \tag{2.15}$$

We note that $\hat{f}_v(\cdot)$ is locally Lipschitz and that by the chain rule [C1],

$$\partial\hat{f}_v(x) \subset G\hat{f}_v(x) \triangleq \{\xi | \xi = \frac{\partial g(x)^T}{\partial x} y,$$

$$y \in \partial\phi(z),\ z = g(x) +v\} \tag{2.16}$$

We will show that there are a number of functions $f(\cdot)$, of the form (2.13), for which, given $x,\varepsilon \geq 0$, it is possible to define vectors $v_\varepsilon(x)$ such that $\|v_\varepsilon(x)\| \leq K\varepsilon$ (with K fixed) and $G_\varepsilon f(x) \triangleq \hat{G}f_{v_\varepsilon(x)}(x)$ are c.d.f. maps. Clearly, we will need the following hypothesis.

Assumption 2.1: For all $x \in \mathbb{R}^n$, $G\hat{f}_0(x) = \partial f(x)$.

The various known rules for constructing the vectors $v_\varepsilon(x)$ can be traced as being derived from those for the function

$$f(x) \triangleq \max_{j\in\underline{m}} g^j(x) \tag{2.17}$$

where $\underline{m} \triangleq \{1,2,\ldots,m\}$. For any $x \in \mathbb{R}^n$ and $\varepsilon \geq 0$, let

$$I_\varepsilon(x) \triangleq \{j \in \underline{m} | f(x) - g^j(x) \leq \varepsilon\} . \tag{2.17a}$$

Then

$$\partial f(x) = \operatorname*{co}_{j\in I_0(x)} \{\nabla g^j(x)\} \tag{2.17b}$$

and Assumption 2.1 holds. Since for any $v \in \mathbb{R}^m$, $\hat{f}_v(x) = \max_{j\in m} (g^j(x)+v^j)$, if we define $v^j(x) \triangleq (f(x)-g^j(x))$ for all $j \in I_\varepsilon(x)$ and set $\overline{v}^j(x) = 0$ otherwise, we find that

$$G\hat{f}_{v_\varepsilon(x)}(x) = \operatorname*{co}_{j\in I_\varepsilon(x)} \{\nabla g^j(x)\} \tag{2.17c}$$

and that for all $v \in \mathbb{R}^m$ such that $\|v\|_\infty \leq \varepsilon$, $G\hat{f}_v(x) \subset G\hat{f}_{v_\varepsilon(x)}(x)$. We now consider the class of functions of the form (2.13) for which a similar fact holds. We shall give some additional examples later.

Proposition 2.2: Let $f : \mathbb{R}^n \to \mathbb{R}^1$ be of the form (2.13) and let $\|\cdot\|$ be some norm on $\mathbb{R}^m$. Suppose that for all $x \in \mathbb{R}^n$ and $\varepsilon > 0$ there exists a $v_\varepsilon(x) \in \mathbb{R}^m$ such that $\|v_\varepsilon(x)\| \leq \varepsilon$ and for all $v \in \mathbb{R}^m$ satisfying $\|v\| \leq \varepsilon$ we have

$$G\hat{f}_v(x) \subset G\hat{f}_{v_\varepsilon(x)}(x)\,. \tag{2.18}$$

Then $G_\varepsilon f(x) \triangleq G\hat{f}_{v_\varepsilon(x)}(x)$ defines a family of c.d.f. maps.

Proof: We refer to Definition 2.1. Because of Assumption 2.1, it is clear that $\partial f(x) = G\hat{f}_0(x) = G_0 f(x)$ for all $x \in \mathbb{R}^n$. (ii) By construction, see (2.18), it follows that if $\varepsilon' > \varepsilon \geq 0$ then for any $x \in \mathbb{R}^n$, $G_\varepsilon f(x) \subset G_{\varepsilon'} f(x)$. (iii) For any $\varepsilon \geq 0$, $G_\varepsilon f(x)$ is obviously convex, and bounded on bounded sets. We now show that $G_\varepsilon f(x)$ is u. s. c. in (ε,x) at $(0,\hat{x})$ for any $\hat{x} \in \mathbb{R}^n$. Let $\hat{x} \in \mathbb{R}^n$ and $\hat{\delta} > 0$ be given and let $\hat{z} \triangleq g(\hat{x})$. Since $\partial\phi(\cdot)$ is u. s. c., there exists a $\rho > 0$ such that $\partial\phi(z) \subset N_\delta(\partial\phi(\hat{z}))$ for all $z \in \mathbb{R}^m$ such that $\|z-\hat{z}\| \leq \rho$, with $N_\delta(\partial\phi(\hat{z}))$ a

$\hat{\delta}$-neighborhood of $\partial\phi(\hat{z})$. Since $g(\cdot)$ is continuous and $\|v_\varepsilon(x)\| \leq \varepsilon$ for any $x \in \mathbb{R}^n$, it follows that there exists a $\hat{\rho} > 0$ such that with $\hat{\varepsilon} = \frac{1}{2}\hat{\rho} > 0$

$$\|g(x) + v_\varepsilon(x) - g(\hat{x})\| \leq \hat{\rho} \tag{2.19}$$

for all $x \in B(\hat{x},\hat{\rho})$ and $\varepsilon \in [0,\hat{\varepsilon}]$. Consequently, $G_\varepsilon f(x)$ is u.s.c. in (ε,x) at $(0,\hat{x})$. We now show that property (iv) of Definition 2.1 holds. Let $\hat{x} \in \mathbb{R}^n$, $\hat{\varepsilon} > 0$ and $\hat{\delta} > 0$ be given. Then, since $g(\cdot)$ is continuous, there exists a $\rho^* > 0$ such that for any $x',x'' \in B(\hat{x},\rho^*)$, $\|g(x')-g(x'')\| \leq \hat{\varepsilon}$. Let $\eta' \in \partial f(x') = G_0 f(x')$, then $\eta' = \frac{\partial g(x')^T}{\partial x} y'$ for some $y' \in \partial\phi(g(x'))$. Now, by definition of $v_{\hat{\varepsilon}}(\cdot)$

$$\hat{G}_v(x'') \subset G_{\hat{\varepsilon}} f(x'') \text{ for all } \|v\| \leq \hat{\varepsilon} . \tag{2.20}$$

Letting $v^* = g(x') - g(x'')$, we find that

$$\begin{aligned}\hat{G}_{v^*}(x'') &= \{\eta \mid \eta = \frac{\partial g(x'')^T}{\partial x} y,\ y \in \partial\phi(g(x'') + [g(x')-g(x'')])\} \\ &= \{\eta \mid \eta = \frac{\partial g(x'')^T}{\partial x} y,\ y \in \partial\phi(g(x'))\} .\end{aligned} \tag{2.21}$$

Since $\|v^*\| \leq \hat{\varepsilon}$, for $y' \in \partial\phi(g(x'))$ as above, we must have $\frac{\partial g(x'')^T}{\partial x} y' \in \hat{G}_{v^*}(x'')$ $\subset G_{\hat{\varepsilon}} f(x'')$. Now $\eta'' \triangleq \frac{\partial g(x'')^T}{\partial x} y' \in G_\varepsilon f(x'')$ and

$$\|\eta'-\eta''\| \leq \|[\frac{\partial g(x')^T}{\partial x} - \frac{\partial g(x'')^T}{\partial x}]\|\|y'\| . \tag{2.21a}$$

Since $\partial\phi(\cdot)$ is bounded on bounded sets and $\frac{\partial g(\cdot)}{\partial x}$ is uniformly continuous on $B(\hat{x},\rho^*)$, it follows from (2.21a) that there is a $\hat{\rho} \in (0,\rho^*]$ such that for any $x',x'' \in B(\hat{x},\hat{\rho})$, given an $\eta' \in \partial f(x')$ there exists a $\eta'' \in G_{\hat{\varepsilon}} f(x'')$ such that $\|\eta'-\eta''\| \leq \hat{\delta}$. This completes our proof. ¤

Apart from the function f(x) defined in (2.17) which satisfies the assumptions of Proposition 2.2, we can cite the following two interesting examples which also fall within the framework of Proposition 2.2.

Consider the function

$$f(x) = \sum_{j\in\underline{m}} |g^j(x)| \tag{2.22}$$

where the $g^j : \mathbb{R}^n \to \mathbb{R}^1$ are continuously differentiable. For any $v \in \mathbb{R}^m$, $\hat{f}_v(x) \triangleq \sum_{j\in\underline{m}} |g^j(x)+v^j|$ and hence, given any $\hat{x} \in \mathbb{R}^n$ and $\varepsilon > 0$, if we define $v^j_\varepsilon(\hat{x}) = -g^j(\hat{x})$ if $|g^j(\hat{x})| \leq \varepsilon$ and set $v^j_\varepsilon(x) = 0$ otherwise, we find that $G\hat{f}_v(\hat{x}) \subset G\hat{f}_{v_\varepsilon(\hat{x})}(x)$ for all $v \in \mathbb{R}^m$ such that $\|v\|_\infty \leq \varepsilon$. This is clear from the fact that, with $\hat{z} = g(\hat{x})$,

$$G\hat{f}_v(\hat{x}) = \sum_{j\notin J(\hat{z}+v)} \nabla g^j(\hat{x}) + \sum_{j\in J(\hat{z}+v)} co\{\nabla g^j(\hat{x}), -\nabla g^j(\hat{x})\} \tag{2.23}$$

where $J(\hat{z}+v) \triangleq \{j \in \underline{m} | |\hat{z}^j+v| = 0\}$.

Finally consider the function

$$f(x) = \max_{\omega\in\Omega} \zeta(x,\omega) \tag{2.24}$$

where $\zeta : \mathbb{R}^n \times \mathbb{R}^1 \to \mathbb{R}^1$ is continuously differentiable and $\Omega \subset \mathbb{R}^1$ is a compact interval. In this case the function $g(x)(\cdot) \triangleq \zeta(x,\cdot)$ assumes values not in $\mathbb{R}^m$, but in $L_\infty(\Omega)$. For any $v \in L_\infty(\Omega)$, we define

$$\hat{f}_v(x) = \max_{\omega\in\Omega}[\zeta(x,\omega) + v(\omega)] \tag{2.25}$$

and obtain that

$$G\hat{f}_v(x) = \operatorname*{co}_{\omega\in\hat{\Omega}_v(x)} \{\nabla_x \zeta(x,\omega)\}, \tag{2.25a}$$

where $\hat{\Omega}_v(x) \triangleq \{\omega \in \Omega | \hat{f}_v(x) = \zeta(x,\omega) + v(\omega)\}$. Clearly, if we set $v_\varepsilon(x)(\omega) = \hat{f}(x) - \zeta(x,\omega)$ for all $\omega \in \tilde{\Omega}_\varepsilon(x) \triangleq \{\omega \in \Omega | f(x) - \zeta(x,\omega) \leq \varepsilon\}$, and

$v_\varepsilon(x)(\omega) = 0$ for all other $\omega \in \Omega$, we find that $G\hat{f}_v(x) \subset G\hat{f}_{v_\varepsilon(x)}(x)$ for all $v \in L_\infty(\Omega)$ such that $\|v\|_\infty \leq \varepsilon$. This results in

$$G_\varepsilon f(x) = \operatorname*{co}_{\omega \in \tilde{\Omega}_\varepsilon(x)} \{\nabla_x \zeta(x,\omega)\} \tag{2.26}$$

The above set may have an infinite number of elements and hence is not a convenient set to use for finding descent directions. Referring to [G2], we find that when $\tilde{\Omega}_0(x)$ is a finite set for all $x \in \mathbb{R}^n$, it is possible to use the much smaller set

$$G_\varepsilon f(x) \triangleq \operatorname*{co}_{\omega \in \Omega_\varepsilon(x)} \{\nabla_x \zeta(x,\omega)\} \tag{2.26a}$$

where $\Omega_\varepsilon(x) \triangleq \{\omega \in \tilde{\Omega}_\varepsilon(x) | \omega$ is a local maximizer of $\zeta(x,\cdot)\}$. It is easy to see that $G_\varepsilon f(x)$ corresponds to the perturbation function $\hat{f}_{v_\varepsilon(x)}(\cdot)$, with $v_\varepsilon(x)(\omega) = f(x) - \zeta(x,\omega)$ for all $\omega \in \Omega_\varepsilon(x)$ and is arbitrary otherwise up to the requirement that $\hat{\Omega}_{v_\varepsilon(x)}(x) = \Omega_\varepsilon(x)$. Quite clearly, $G_\varepsilon f(x)$, as defined in (2.26a) does not satisfy the assumptions of Proposition 2.2. However, showing that the maps $G_\varepsilon f(\cdot)$, defined by (2.26a), are c.d.f. maps is a great deal simpler than the original proof of convergence in [G2], as we now show.

Proposition 2.3: Consider the function $f(\cdot)$ defined by (2.24) with $\Omega = [\omega_0, \omega_f]$. Suppose that for every $x \in \mathbb{R}^n$ $\zeta_\omega(x,\bar{\omega}) = 0$ for at most a finite number of $\bar{\omega} \in \Omega$ and that $\zeta_\omega(x,\omega) \neq 0$ for $\omega \in \{\omega_0, \omega_f\}$. Then the maps $\{G_\varepsilon f(\cdot)\}_{\varepsilon \geq 0}$ defined by (2.26a) form a family of c.d.f. maps. ¤

The proof of Proposition 2.3 requires the following three facts which we establish first.

Fact 2.1: For any $\varepsilon > 0, x \in \mathbb{R}^n$, let

$$\overset{v}{\Omega}_\varepsilon(x) = \{\omega \in \tilde{\Omega}_\varepsilon(x) | \zeta_\omega(x,\omega) = 0\} \cup \{\omega_0, \omega_f\} \tag{2.27}$$

Then $\overset{v}{\Omega}_\varepsilon(\cdot)$ is u.s.c.

Proof: Suppose $x_i \to \hat{x}$ as $i \to \infty$ and $\omega_i \in \overset{v}{\Omega}_\varepsilon(x_i)$ are such that $\omega_i \to \hat{\omega}$ as $i \to \infty$. Then (i) $f(x_i) - \zeta(x_i,\omega_i) \leq \varepsilon$ for all i and hence, by continuity of $f(\cdot)$ and $\zeta(\cdot,\cdot), \hat{\omega} \in \Omega_\varepsilon(x)$, and (ii) either by continuity of $\zeta_\omega(\cdot,\cdot)$ $\zeta_\omega(\hat{x},\hat{\omega}) = 0$ or $\hat{\omega} \in \{\omega_0,\omega_f\}$. Thus $\hat{\omega} \in \overset{v}{\Omega}_\varepsilon(\hat{x})$, which completes the proof. ¤

Fact 2.2: Given $\hat{x} \in \mathbb{R}^n$, $\hat{\varepsilon} > 0$, there exists a $\hat{\rho} > 0$ such that for any $x',x'' \in B(\hat{x},\hat{\rho})$, $\Omega_0(x') \subset \tilde{\Omega}_{\hat{\varepsilon}}(x'')$.

Proof: Since Ω is compact, there exists a $\hat{\rho} > 0$ such that for any $x',x'' \in B(\hat{x},\hat{\rho})$ if $\omega' \in \Omega$ is such that

$$f(x') - \zeta(x',\omega') = 0 \tag{2.28a}$$

then

$$f(x'') - \zeta(x'',\omega') \leq \varepsilon, \tag{2.28b}$$

i.e. $\Omega_0(x') \subset \tilde{\Omega}_\varepsilon(x'')$. ¤

The following result is obvious.

Fact 2.3: Let $\mu_\varepsilon(x) \triangleq \text{meas}(\tilde{\Omega}_\varepsilon(x))$. Then (i) $\varepsilon' < \varepsilon'' \Rightarrow \mu_{\varepsilon'}(x) \leq \mu_{\varepsilon''}(x)$, and (ii) $\mu_\varepsilon(\cdot)$ is continuous. ¤

Proof of Proposition 2.3: Referring to Definition 2.1 we find that (i) $\partial f(x) = G_0(x)$ and (ii) that $\varepsilon < \varepsilon' \Rightarrow G_\varepsilon(x) \Rightarrow G_{\varepsilon'}(x)$, by construction in (2.26a). (iii) Clearly, $G_\varepsilon(x)$ is always convex. Next, let $\hat{x} \in \mathbb{R}^n$ be arbitrary. Since by assumption the set $\overset{v}{\Omega}_\varepsilon(\hat{x})$ is finite for all $\varepsilon \geq 0$, there exists an $\hat{\varepsilon} > 0$ such that $\Omega_\varepsilon(\hat{x}) = \overset{v}{\Omega}_\varepsilon(\hat{x})$ for all $\varepsilon \in [0,\hat{\varepsilon}]$. It now follows from the u.s.c. of $\overset{v}{\Omega}_\varepsilon(\cdot)$ that $\Omega_\varepsilon(\cdot)$ is u.s.c. at $\hat{x}$ for all $\varepsilon \in [0,\hat{\varepsilon}]$ and hence, from the continuity of $\nabla_x\zeta(\cdot,\cdot)$, it follows

that $G_\varepsilon(x)$ is u.s.c. at $(0,\hat{x})$. We now turn to property (iv) in Definition 2.1. Let $\hat{x} \in \mathbb{R}^n$, $\hat{\varepsilon} > 0$ and $\hat{\delta} > 0$ be given. Then (a) there exists a $\rho_1 > 0$ such that for all $x',x'' \in B(\hat{x},\rho_1)$ and $\omega',\omega'' \in [\omega_0,\omega_f]$ satisfying $|\omega'-\omega''| \leq \rho_1$, we have

$$\|\nabla_x \zeta(x',\omega') - \nabla \zeta_x(x'',\omega'')\| \leq \hat{\delta} . \tag{2.29}$$

(b) Since $\Omega_0(\hat{x})$ is a discrete set, there exists an $\varepsilon_1 \in (0,\hat{\varepsilon}]$ such that $\mu_{\varepsilon_1}(\hat{x}) \leq \hat{\delta}/2$. Hence, by continuity of $\mu_{\varepsilon_1}(\cdot)$, there exists a $\rho_2 \in (0,\rho_1]$ such that $\mu_{\varepsilon_1}(x) \leq \rho_1$ for all $x \in B(\hat{x},\rho_2)$.

(c) By Fact 2.2, there exists a $\hat{\rho} \in (0,\rho_2]$ such that $\Omega_0(x') = \tilde{\Omega}_0(x')$ $\subset \tilde{\Omega}_{\varepsilon_1}(x'')$ for all $x',x'' \in B(\hat{x},\hat{\rho})$.

(d) Now consider any $x',x'' \in B(\hat{x},\hat{\rho})$. If $\eta' \in \partial f(x')$, then

$$\eta' = \sum_{\omega' \in \Omega_0(x')} \mu^{\omega'} \nabla_x \zeta(x',\omega') \tag{2.30}$$

where $\mu^{\omega'} \geq 0$ and $\sum_{\omega' \in \Omega_0(x')} \mu^{\omega'} = 1$. By (c) $\Omega_0(x') \subset \tilde{\Omega}_{\varepsilon_1}(x'') \subset \tilde{\Omega}_{\hat{\varepsilon}}(x'')$ and $\mu_{\varepsilon_1}(x'') \leq \rho_1$. Since every disjoint interval of $\tilde{\Omega}_{\varepsilon_1}(x'')$ must contain at least one $\omega'' \in \Omega_{\varepsilon_1}(x'')$, it follows that for every $\omega' \in \Omega_0(x')$ there exists an $\omega_{\omega'} \in \Omega_{\varepsilon_1}(x'') \subset \Omega_{\hat{\varepsilon}}(x'')$ such that $|\omega'-\omega_{\omega'}| \leq \rho_1$. Hence the vector $\eta'' \in G_{\hat{\varepsilon}}(x'')$ defined by

$$\eta'' = \sum_{\omega' \in \Omega_0(x')} \mu^{\omega'} \nabla_x \zeta(x'',\omega_{\omega'}) \tag{2.31}$$

satisfies

$$\begin{aligned} \|\eta''-\eta'\| &\leq \sum_{\omega' \in \Omega_0(x')} \mu^{\omega'} \|\nabla_x \zeta(x'',\omega_{\omega'}) - \nabla_x \zeta(x',\omega')\| \\ &\leq \hat{\delta} \end{aligned} \tag{2.32}$$

which completes our proof. ¤

Next we turn to problems involving eigenvalues of Hermitian matrices, see e.g. [C3,P6]. We shall consider only one case. Let $Q : \mathbb{R}^n \to C^m \times C^m$ be a continuously differentiable, complex matrix valued function such that $Q(x)$ is Hermitian for all x. For any mxm Hermitian matrix M, we denote its eigenvalues as $\sigma^1[M] \geq \sigma^2[M] \ldots \geq \sigma^m[M]$ and we consider the case where

$$f(x) \triangleq \sigma^1[Q(x)] \tag{2.33}$$

Thus for any Hermitian matrix $V \in C^m \times C^m$, we define the perturbation function $\hat{f}_V(\cdot)$ by

$$\hat{f}_V(x) = \sigma^1[Q(x) + V] \tag{2.34}$$

We proceed by analogy with the example in (2.17) in defining a "maximal" perturbation matrix $V_\varepsilon(x)$. Clearly, there exists a matrix of complex orthonormal left eigenvectors $U(x)$ such that $U^*(x)\, U(x) = I$ and

$$Q(x) = U^*(x)\, \Sigma(x)\, U(x) \tag{2.35}$$

where $\Sigma(x) \triangleq \text{diag}(\sigma^1[Q(x)],\ldots,\sigma^m[Q(x)])$. Given $\varepsilon > 0$, we define $V_\varepsilon(x)$ by

$$V_\varepsilon(x) \triangleq U^*(x)\, \Lambda_\varepsilon(x)\, U(x) \tag{2.36}$$

where $\Lambda_\varepsilon(x) \triangleq \text{diag}(\lambda^i_\varepsilon(x))$, with $\lambda^i_\varepsilon(x) = \sigma^1[Q(x)] - \sigma^i[Q(x)]$ for all $i \in I_\varepsilon[Q(x)]$ and $\lambda^i_\varepsilon(x) = 0$ otherwise, where $I_\varepsilon[Q(x)] \triangleq \{i \in \underline{m} \mid \sigma^1[Q(x)] - \sigma^i[Q(x)] \leq \varepsilon\}$. This choice of $V_\varepsilon(x)$ clearly "maximizes" the set

$$\partial \hat{f}_V(x) \triangleq \text{co}\{y \mid y^i = \langle U_V(x)z, \frac{dQ(x)}{dx^i} U_V(x)z\rangle$$
$$i = 1,2,\ldots,n,\ \|z\| = 1\} \tag{2.37}$$

where, given that $I_0[Q(x)+V] = \{1,2,\ldots,k_V(x)\}$ $U_V(x)$ consists of the first $k_V(x)$ columns of $U(x)$ (see [P6] for a proof that $\partial f(x) = \partial \hat{f}_0(x)$). We

claim that $G_\varepsilon f(x) \triangleq G\hat{f}_{V_\varepsilon(x)}(x)$, as defined by (2.36) and (2.37) is a c.d.f. map, but we omit a proof, which can be constructed by referring to [P6]. We note that in [P6] a somewhat larger set $G_\varepsilon f(x)$ was used so as to avoid computational difficulties caused by the need to distinguish between eigenvalues that are very close to being equal. We find that in [P6], for any $\varepsilon \geq 0$, $x \in \mathbb{R}^n$

$$k_\varepsilon(x) \triangleq \max_{k \in \underline{m}} \{k+1 \mid \sigma^i[Q(x)] - \sigma^{i+1}[Q(x)] \leq \varepsilon \text{ for all } i \leq k\} \tag{2.38a}$$

which leads to the definition

$$G_\varepsilon f(x) \triangleq \text{co}\{y \mid y^i = \langle U_\varepsilon(x)z, \frac{\partial Q(x)}{\partial x^i} U_\varepsilon(x)z\rangle ,$$

$$i = 1,2,\ldots,n, \ \|z\| = 1\} \tag{2.38b}$$

where $U_\varepsilon(x)$ consists of the first $k_\varepsilon(x)$ columns of $U(x)$. Since $k_\varepsilon(x) \geq k_{V_\varepsilon(x)}(x)$, it is clear that (2.38b) results in a larger set than (2.37) for $V = V_\varepsilon(x)$; however the general properties relevant to convergence of the two sets are the same.

This concludes our demonstration that a large number of semingly unrelated algorithms for various unconstrained nondifferentiable optimization problems can be seen as manifestations of a single relatively simple principle.

Before proceeding to constrainted optimization problems, it remains to point out that when $f(\cdot)$ is locally Lipschitz function from a Banach space $\mathcal{X}$ into $\mathbb{R}$, the computation of a descent direction according to

$$h(x) = \arg \min_{h \in} \{\tfrac{1}{2}\|h\|^2 + d_\varepsilon f(x,h)\} \tag{2.39a}$$

may not be a tractable problem. In that case, (2.39a) may be replaced by

$$h(x) \in \arg \min_{\|h\| \leq 1} d_\varepsilon f(x,h)$$

$$= \arg \min_{\|h\| \leq 1} \max_{\zeta \in G_\varepsilon f(x)} (\xi,h) \tag{2.39b}$$

where the action of a $\xi \in \mathcal{X}'$, the dual of $\mathcal{X}$, on an $h \in \mathcal{X}$ is denoted by (ξ,h). All the proofs in this section remain valid when (2.39a) is replaced by (2.39b).

3. Constrained Optimization

In this section we restrict ourselves to problems of the form

$$\min\{f^0(x) | f^j(x) \leq 0,\ j \in \underline{m}\}, \tag{3.1}$$

where $\underline{m} \triangleq \{1,2,\ldots,m\}$ and $f^j : \mathbb{R}^n \to \mathbb{R}^1$, $j \in \{0\} \cup \underline{m}$, are locally Lipschitz continuous functions.

We shall assume that we have for all the functions f^j, $j \in \{0\} \cup \underline{m}$, families of convergent direction finding maps $\{G_\varepsilon f^j(\cdot)\}_{\varepsilon \geq 0}$ (see Definition 2.1). We define $\psi : \mathbb{R}^n \to \mathbb{R}^1$ and $\psi(\cdot)_+$, as follows

$$\psi(x) \triangleq \max_{j \in \underline{m}} f^j(x), \tag{3.2}$$

$$\psi(x)_+ = \max\{0,\psi(x)\}. \tag{3.3}$$

We are about to state a phase I-phase II algorithm of a form quite similar to the ones treated in [P2,P7]. First, for any $x \in \mathbb{R}^n$ and $\varepsilon \geq 0$, we define the ε-most violated constraint index set

$$I_\varepsilon(x) \triangleq \{j \in \underline{m} | f^j(x) \geq \psi(x)_+ - \varepsilon\} \tag{3.4a}$$

and we set

$$J_\varepsilon(x) \triangleq \{0\} \cup I_\varepsilon(x)\ . \tag{3.4b}$$

Next, for $\varepsilon \geq 0$ given, we define the phase-I ε-search direction at x by

$$h_{\psi\varepsilon}(x) \triangleq \arg\min_h\{\tfrac{1}{2}\|h\|^2 + \max_{j\in I_\varepsilon(x)} d_\varepsilon f^j(x,h)\}$$

$$= -\arg\min_h \{\tfrac{1}{2}\|h\|^2 | h \in \operatorname{co}_{j\in I_\varepsilon(x)}\{G_\varepsilon f^j(x)\}\}, \tag{3.5}$$

where co denotes the convex hull, and we define the phase-II ε-search direction at x by

$$h_{f\varepsilon}(x) \triangleq \arg\min_h \{\tfrac{1}{2}\|h\|^2 + \max_{j\in J_\varepsilon(x)} d_\varepsilon f^j(x,h)\}$$

$$= -\arg\min_h \{\tfrac{1}{2}\|h\|^2 | h \in \operatorname{co}_{j\in J_\varepsilon(x)}\{G_\varepsilon f^j(x)\}\}. \tag{3.6}$$

Finally, we define the cross-over function

$$\Gamma(x) \triangleq e^{-\gamma\psi(x)_+} \tag{3.7}$$

where $\gamma > 0$ is a parameter. It will become clear shortly that when $\psi(x) > 0$, for appropriate values of ε, $h_{\psi_\varepsilon}(x)$ is a descent direction for $\psi(x)$, while when $\psi(x) \leq 0$, $h_{f\varepsilon}(x)$ is a <u>feasible</u> descent direction for $f(\cdot)$. The cross-over from one to the other is incorporated in the search direction

$$h_\varepsilon(x) \triangleq \Gamma(x)h_{f\varepsilon}(x) + (1-\Gamma(x))h_{\psi_\varepsilon}(x). \tag{3.8}$$

As we have already seen in the preceding section, we need a mechanism for driving ε to zero. To this end, (c.f. (3.13) in [P7]) we define

$$\theta_\varepsilon(x) \triangleq \max\{\|\Gamma(x)\ h_{f\varepsilon}(x)\|^2,\ \|(1-\Gamma(x))h_{\psi\varepsilon}(x)\|^2\} \tag{3.9}$$

and, with E as in (2.5), and $\delta > 0$,

$$\varepsilon(x) \triangleq \max\{\varepsilon \in E | \theta_\varepsilon(x) \geq \delta\varepsilon\}\ . \tag{3.10}$$

Algorithm Model 3.1.

Parameters: $\delta > 0$, ε_0 (for $\varepsilon(x)$); $\alpha,\beta \in (0,1)$ (for Armijo step size rule). Families $\{G_\varepsilon f^j(\cdot)\}_{\varepsilon \geq 0}$, $j \in \{0\} \cup \underline{m}$ of c.d.f. maps.

Data: $x_0 \in \mathbb{R}^n$.

Step 0: Set $i = 0$.

Step 1: Compute $\varepsilon(x_i)$ and $h_i \triangleq h_{\varepsilon(x_i)}(x_i)$.

If $\varepsilon(x_i) = 0$, stop.

Step 2: If $\psi(x_i)_+ > 0$, compute the largest $\lambda_i = \beta^{k_i}$, $k_i \in \mathbb{N}_+$ such that

$$\psi(x_i+\lambda_i h_i) - \psi(x_i) \leq -\lambda_i \alpha\delta\varepsilon(x_i). \tag{3.11}$$

If $\psi(x_i)_+ = 0$, compute the largest $\lambda_i = \beta^{k_i}$, $k_i \in \mathbb{N}_+$ such that

$$f^0(x_i+\lambda_i h_i) - f^0(x_i) \leq -\lambda_i \alpha\delta\varepsilon(x_i) \tag{3.12a}$$

and

$$\psi(x_i+\lambda_i h_i) \leq 0. \tag{3.12b}$$

Step 3: Set $x_{i+1} = x_i + \lambda_i h_i$, set $i = i+1$ and go to Step 1. ¤

To ensure that the above algorithm does not jam at an infeasible point, we must introduce the following commonly used hypothesis:

Assumption 3.1: For every $x \in \mathbb{R}^n$ such that $\psi(x) \geq 0$, $h_{\psi 0}(x) \neq 0$. ¤

To establish the convergence properties of Algorithm Model 3.1 we shall need the following results.

Lemma 3.1: For every $\varepsilon \geq 0$ and any $x \in \mathbb{R}^n$

$$\|h_\varepsilon(x)\|^2 \geq \theta_\varepsilon(x). \tag{3.13}$$

For a proof of this lemma, see Lemma 3.1 in [P7].

Lemma 3.2: The function $\theta_\varepsilon(x)$ defined in (3.9) has the following properties: a) For any $x \in \mathbb{R}^n$, if $\varepsilon'' > \varepsilon' \geq 0$, then $\theta_{\varepsilon''}(x) \leq \theta_{\varepsilon'}(x)$. b) $\theta_\varepsilon(x)$ is l.s.c. in (ε,x) at any $(0,\hat{x})$.

Proof: a) Since $\varepsilon'' > \varepsilon'$ implies that $I_{\varepsilon''}(x) \supset I_{\varepsilon'}(x)$ and that $G_{\varepsilon''}f^j(x) \supset G_{\varepsilon'}f^j(x)$, this part is obvious.
b) By definition of c.d.f. maps, $G_\varepsilon f^j(x)$ is u.s.c. in (ε,x) at any $(0,\hat{x})$, $j = 0,1,\ldots,m$. Let $\hat{x} \in \mathbb{R}^n$ be arbitrary. Then, given $\hat{\varepsilon} > 0$, there exists a $\hat{\rho} > 0$ such that $I_\varepsilon(x) \subset I_{\hat{\varepsilon}}(x) \subset I_{\hat{\varepsilon}}(x)$ for all $(\varepsilon,x) \in [0,\hat{\varepsilon}] \times B(\hat{x},\hat{\rho})$. Because of this and the u.s.c. of the $G_\varepsilon f^j(x)$ at $(0,\hat{x})$, there exist $\varepsilon^* \in (0,\hat{\varepsilon}]$ and $\rho^* \in (0,\hat{\rho}]$ such that

$$\operatorname*{co}_{j\in I_\varepsilon(x)}\{G_\varepsilon f^j(x)\} \subset \operatorname*{co}_{j\in I_{\hat{\varepsilon}}(\hat{x})}\{G_\varepsilon f^j(x)\} \subset N_\delta\Big(\operatorname*{co}_{j\in I_{\hat{\varepsilon}}(\hat{x})}\{G_\varepsilon f^j(\hat{x})\}\Big) \qquad (3.14)$$

where $N_\delta(\cdot)$ denotes a δ neighborhood of the set in parentheses. Consequently, $\|h_{\psi\varepsilon}(x)\|$ is l.s.c. in (ε,x) at any $(0,\hat{x})$. Similarly, it can be shown that $\|h_{f\varepsilon}(x)\|$ is l.s.c. in (ε,x) at any $(0,\hat{x})$. Since $\Gamma(\cdot)$ is continuous, it follows that $\theta_\varepsilon(x)$ is l.s.c. in (ε,x) at any $(0,\hat{x})$, which completes our proof. ¤

The following result can be established in essentially the same way as Proposition 2.1 and hence a proof will be omitted.

Corollary 3.1: For every $\hat{x} \in \mathbb{R}^n$ such that $\theta_0(\hat{x}) > 0$, there exists a $\hat{\rho} > 0$ such that $\varepsilon(x) \geq \nu\varepsilon(\hat{x}) > 0$ for all $x \in B(\hat{x},\hat{\rho})$. ¤

Theorem 3.1: Suppose that Assumption 3.1 holds and that $\{x_i\}$ is a sequence constructed by Algorithm Model 3.1. a) If $\{x_i\}$ is finite, with last element x_k, then $\psi(x_k) \leq 0$ and

$$0 \in \underset{j\in J_0(x_k)}{\text{co}} \{\partial f^j(x_k)\} \tag{3.15a}$$

b) If $\{x_i\}$ is infinite, then for any accumulation point $\hat{x}$ of $\{x_i\}$, we have $\psi(\hat{x}) \leq 0$ and

$$\theta \in \underset{j\in J_0(x)}{\text{co}} \{\partial f^j(\hat{x})\} \tag{3.15b}$$

Proof: a) Suppose that $\{x_i\}_{i=1}^k$ is finite. Then, by construction, $\varepsilon(x_k) = 0$ and hence, by corollary 3.1, $\theta_0(x_k) = 0$, so that

$$\Gamma(x_k)\, h_{f0}(x_k) = (1-\Gamma(x_k))h_{\psi 0}(x_k) = 0 \tag{3.16}$$

Suppose now that $\Gamma(x_k) < 1$, i.e. $\psi(x_k) > 0$, then (3.16) implies that $h_{\psi 0}(x_k) = 0$. But this contradicts Assumption 3.1 and hence we must have $\psi(x_k) \leq 0$. Since $\Gamma(x_k) = 1$, $h_{f0}(x_k) = 0$ and hence, since $\partial f^j(x_k) = G_0 f^j(x_k), j = 0,1,..,m$, we find that (3.15a) must hold.

b) Suppose that $\{x_i\}_{i=0}^{\infty}$ has an accumulation point $\hat{x}$, i.e., that $x_i \overset{K}{\to} \hat{x}$, with $K \subset \mathbf{N}_+$ infinite, that $\psi(\hat{x}) \leq 0$ and (3.15b) fails to hold. We consider two cases.

Case 1: $\psi(x_i) > 0$ for all $i \in \mathbf{N}_+$. Then, by (3.11) $\{\psi(x_i)\}_{i=0}^{\infty}$ is monotone decreasing, and $\psi(x_i) \overset{K}{\to} \psi(\hat{x})$ by continuity of $\psi(\cdot)$. Hence $\psi(x_i) \searrow \psi(\hat{x})$ as $i \to \infty$. We shall now show that this leads to a contradiction and in the process also show that this part of the Algorithm Model 3.1 is well defined.

Since $\psi(x_i) \geq 0$ for all i, we must have $\psi(\hat{x}) \geq 0$ and hence $h_{\psi 0}(\hat{x}) \neq 0$ by Assumption 3.1. Consequently, $\theta_0(\hat{x}) > 0$ either because $\Gamma(\hat{x}) > 0$ or because (3.15b) fails, i.e., because $\Gamma(\hat{x}) = 0$ and $h_{f0}(\hat{x}) \neq 0$. Thus, by Corollary 3.1, there exists an i_0 such that $\varepsilon(x_i) \geq \nu\varepsilon(\hat{x}) > 0$ for all

$i \in K$, $i \geq i_0$. Since the sets $G_{\varepsilon_0} f^j(x_i)$ are bounded on bounded sets and $G_{\varepsilon(x_i)} f^j(x_i) \subset G_{\varepsilon_0} f^j(x_i)$ by (ii) of Definition 2.1, it follows, via Lemma 3.1, that there exists a $b \in (0,\infty)$ such that

$$\nu\varepsilon(\hat{x})\delta \leq \varepsilon(x_i) \leq \|h_i\|^2 \leq b \tag{3.17}$$

For all $i \in K$, $i \geq i_0$. Now, by the mean value theorem of Lebourg [L1], for $j \in \underline{m}$

$$f^j(x_i+\lambda h_i) - \psi(x_i) = [f^j(x_i) - \psi(x_i)] + \lambda \langle \xi^j_{\lambda_i}, h_i \rangle, \tag{3.18}$$

where $\xi^j_{\lambda_i} \in \partial f^j(x_i+s\lambda h_i)$. Now, for any $\xi \in G_{\varepsilon(x_i)} f^j(x_i)$, $j \in I_{\varepsilon(x_i)}(x_i)$, we have by construction that

$$\begin{aligned} \langle h_i, \xi \rangle &= \Gamma(x_i) \ \langle h_{f\varepsilon(x_i)}(x_i), \xi \rangle \\ &\quad + [1-\Gamma(x_i)] \langle h_{\psi\varepsilon(x_i)}(x_i), \ \xi \rangle \geq \Gamma(x_i)\|h_{f\varepsilon(x_i)}(x_i)\|^2 \\ &\quad + [1-\Gamma(x_i)] \ \|h_{\psi\varepsilon(x_i)} \ (x_i)\|^2 \geq \|h_i\|^2 \end{aligned} \tag{3.19}$$

Hence, proceeding as in the proof of Theorem 2.1, we conclude that there is a $\hat{\lambda}_1 = \beta^{\hat{k}_1}$, $\hat{k}_1 \in \mathbb{N}_+$, such that

$$\begin{aligned} f^j(x_i+\beta^{\hat{k}_1}h_i) - \psi(x_i) &\leq -\beta^{\hat{k}_1}\alpha\|h_i\|^2 \\ &\leq -\beta^{\hat{k}_1}\alpha\delta\varepsilon(x_i) \\ &\leq \beta^{\hat{k}_1}\alpha\delta\nu\varepsilon(\hat{x}) \end{aligned} \tag{3.20a}$$

for all $j \in I_{\varepsilon(x_i)}(x_i)$, and where we have made use of the fact that $f^j(x_i) - \psi(x_i) \leq 0$. Next, since $f^j(x_i) - \psi(x_i) < -\varepsilon(x_i) \leq -\nu\varepsilon(x)$ for all $j \notin I_{\varepsilon(x_i)}(x_i)$, and since the h_i are bounded for $i \in K$, it follows

by uniform continuity of f^j, ψ on bounded sets, that there exists a $0 < \hat{\lambda} = \beta^{\hat{k}} \leq \hat{\lambda}_1$ such that

$$f^j(x_i+\hat{\lambda}h_i) - \psi(x_i) \leq -\hat{\lambda}\alpha\delta\varepsilon(x_i) \tag{3.20b}$$

for all $j \notin I_{\varepsilon(x_i)}(x_i)$, $i \in K, i \geq i_0$. Combining (3.20a) and (3.20b) we conclude that $\lambda_i \geq \hat{\lambda}$ for all $i \in K$, $i \geq i_0$ and hence that

$$\psi(x_{i+1}) - \psi(x_i) \leq -\hat{\lambda}\alpha\delta\nu\varepsilon(\hat{x}) \tag{3.21}$$

for all $i \geq i_0$, $i \in K$. But this contradicts the fact that $\psi(x_i) \searrow \psi(\hat{x})$ and hence we are done.

Case 2: There exists an i_0 such that for all $i \geq i_0$, $\psi(x_i) \leq 0$. If $x_i \overset{K}{\to} \hat{x}$ and $\psi(\hat{x}) < -\varepsilon(\hat{x})$, then the theorem follows directly from Theorem 2.1. Hence we only need to consider the case where $\psi(\hat{x}) \geq -\varepsilon(\hat{x})$. Now, for this case, we conclude from Case 1 above that there is an i_0 and a $\hat{\lambda}_1 = \beta^{\hat{k}_1}$ such that for all $i \geq i_0$, $i \in K$

$$\psi(x_i+\beta^{\hat{k}_1}h_i) \leq 0\ , \tag{3.22}$$

and from the proof of Theorem 2.1 that there exists an $i_1 \geq i_0$ and a $\hat{\lambda} = \beta^{\hat{k}} \leq \hat{\lambda}_1$ such that (3.12a) is satisfied for all $i \geq i_1$, $i \in K$, with $\lambda_i = \hat{\lambda}$. Hence we must have that $\lambda_i \geq \hat{\lambda}$ and for all $i \geq i_1$, $i \in K$

$$f^0(x_{i+1}) - f^0(x_i) \leq -\hat{\lambda}\alpha\delta\nu\varepsilon(\hat{x}). \tag{3.23}$$

But $f^0(x_i) \searrow f^0(\hat{x})$ since $x_i \overset{K}{\to} \hat{x}$ and $f^0(\cdot)$ is continuous, which is contradicted by (3.23) and hence the proof is complete. ¤

Since for continuously differentiable functions $f^j(\cdot)$ we may set $G_\varepsilon f^j(x) = \nabla f^j$, it should now be clear that any combination of differentiable functions and functions such as those defined in (2.21),

(2.29) and (2.36), (2.37) may appear in the constraints.

Finally, it remains to point out that when the substitute formula (2,39) is used for problems in Banach spaces, (3.5) and (3.6) become replaced by

$$h_{\psi_\varepsilon}(x) \in \arg \min_{\|h\|\leq 1} \max_{j\in I_\varepsilon(x)} d_\varepsilon f^j(x,h)$$

$$= \arg \min_{\|h\|\leq 1} \max_{\substack{\xi\in co\{G_\varepsilon f^j(x)\} \\ j\in I_\varepsilon(x)}} (\xi,h) \tag{3.24}$$

and

$$h_{f_\varepsilon}(x) \in \arg \min_{\|h\|\leq 1} \max_{j\in J_\varepsilon(x)} d_\varepsilon f^j(x,h)$$

$$= \arg \min_{\|h\|\leq 1} \max_{\substack{\xi\in co\{G_\varepsilon f^j(x)\} \\ j\in J_\varepsilon(x)}} (\xi,h) \tag{3.24}$$

respectively. Again, the arg min max may be set valued.

4. Conclusion

We have shown that a rather large number of nondifferentiable optimization algorithms can be presented and analyzed in a unified way. We have also shown that for an important class of optimization problems, defined by composite functions, efficient nondifferentiable optimization algorithms can be constructed by using the generalized gradients of perturbation functions. Furthermore we have established rules for the construction of these perturbation functions.

5. References

[A1] Armijo, L., "Minimization of functions having Lipschitz continuous first partial derivatives", Pacific Journal of Mathematics, Vol. 16, pp. 1-3, 1966.

[B1] Berge, C., Topological Spaces, Macmillan Co., N.Y., 1963.

[B2] Bertsekas, D.P., and Mitter, S.K., "A Descent numerical method for nondifferentiable cost functionals", SIAM J. Control, Vol. 11, pp. 637-652, 1973.

[C1] Clarke, F., "Generalized gradients and Applications,", Trans. Amer. Math. Soc., Vol. 205, 1975.

[C2] Clarke, F., "A new approach to Lagrange multipliers", Math. of Oper. Research, Vol. 1, pp. 165-174, 1976.

[C3] Cullum, J., Donath, W. E., and Wolfe, P., "The Minimization of certain nondifferentiable sums of eigenvalues of symmetric matrices", in Nondifferentiable Optimization, Math. Programming Study No. 3, M.L. Balinski and P. Wolfe, eds., North Holland, Amsterdam, pp. 35-55, 1975.

[G1] Goldstein, A.A., "Optimization of Lipschitz continuous functions", Math. Programming Vol. 13, pp. 14-22, 1977.

[G2] Gonzaga, C., E. Polak and R. Trahan, "An improved algorithm for optimization problems with functional inequality constraints", IEEE Transactions on Automatic Control, Vol. AC-25, No. 1, pp. 49-54, 1979.

[L1] Lebourg, C., "Valeur Moyenne Pour Gradient Generalise", C.R. Acad. Sci., Paris, Vol. 281, 1975.

[L2] Lemarechal, C., "Nondifferentiable optimization, subgradient and ε-subgradient methods", Lecture Notes, No. 117, Optimization and

Operations Research, Springer Verlag, New York 1976.

[L3] Lemarechal, C., "Minimization of nondifferentiable functions with constraints", Proc. 12th Allerton Conference on Circuit Theory, University of Illinois, Urbana, pp. 16-24, 1974.

[L4] Lemarechal, C., "Extensions diverses des methods de gradient et applications", These Docteur d'Etat, University of Paris, 1980.

[M1] Mifflin, R., "An Algorithm for constrained optimization with semi-smooth functions", Math. of Oper. Res., Vol. 2, No. 2, 1977.

[M2] Mifflin, R., Semi-smooth and semi-convex functions in constrained optimization", SIAM J. Control and Optimization, Vol. 15, No. 6, pp. 959-973, 1977.

[M3] Mayne, D.Q. and E. Polak, "A quadratically convergent algorithm for solving infinite dimensional inequalities", University of California Electronics Research Laboratory, Memo. No. UCB/ERL M/80/11, 1980. J. of Appl. Math. and Optimization, in press.

[P1] Polak, E. and D.Q. Mayne, "An Algorithm for optimization problems with functional inequality constraints", IEEE Transactions on Automatic Control, Vol. AC-21, No. 2, pp. 181-194, 1976.

[P2] Polak, E., R. Trahan, and D.Q. Mayne, "Combined phase-I -- phase-II methods of feasible directions", Math. Programming, Vol. 17, No. 1, pp. 32-61, 1979.

[P3] Polak, E., and A. Sangiovanni Vincentelli, "Theoretical and computational aspects of the optimal design centering, tolerancing and tuning problem", IEEE Trans. Vol. CAS-26, No. 9, pp. 795-813, 1979.

[P4] Polak, E., and D.Q. Mayne, "On the solution of singular value inequalities over a continuum of frequencies", IEEE Trans. on

Automatic Control, Vol. AC-26, No. 3, pp. 690-695, 1981.

[P5] Polak, E., and A. Tits, "A recursive quadratic programming algorithm for semi-infinite optimization problems", University of California, Berkeley, Electronics Research Laboratory, Memo No. UCB/ERL M80/50, 1980, J. Applied Math. and Optimization, in press.

[P6] Polak, E., and Wardi, Y.Y., "A nondifferentiable optimization algorithm for the design of control systems subject to singular value inequalities over a frequency range", Automatica, Vol. 18, No. 3, pp. 267-283, 1982.

[P7] Polak, E., D.Q. Mayne, and Y.Y. Wardi, "On the extension of constrained optimization algorithms from differentiable to non-differentiable problems," University of California, Berkeley, Electronics Research Laboratory Memo No. UCB/ERL M81/78, April 14, 1981, SIAM J. Control and Optimization, in press.

[P8] Polak, E., and Y.Y. Wardi, "A study of minimizing sequences," University of California, Berkeley, Electronics Research Laboratory Memo No. UCB/ERL M82/22, March 22, 1982.

[P9] Polak, E. "Computational Methods in optimization: a unified approach," Academic Press, N.Y., 1081.