NOTES ON THE MATHEMATICAL FOUNDATIONS OF NONDIFFERENTIABLE OPTIMIZATION IN ENGINEERING DESIGN

by

E. Polak

Memorandum No. UCB/ERL M84/15

2 February 1984

NOTES ON THE MATHEMATICAL FOUNDATIONS OF NONDIFFERENTIABLE OPTIMIZATION IN ENGINEERING DESIGN

by

E. Polak

Memorandum No. UCB/ERL M84/15

2 February 1984

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

NOTES ON THE MATHEMATICAL FOUNDATIONS OF NONDIFFERENTIABLE OPTIMIZATION IN ENGINEERING DESIGN

by

E. POLAK

Department of Electrical Engineering and Computer Sciences
and the Electronics Research Laboratory
University of California, Berkeley, California 94720

0. INTRODUCTION.

In most engineering fields, the design process begins with the identification of one or more structural or system configurations which can satisfy the overall objectives. Once a configuration is chosen, parameters for the components or other elements must be determined. The most widespread computer-aided design systems, whether in automatic control, electronics, or structures, assist in the parameter determination phase by means of simulation or response evaluation programs. Such programs are executed to evaluate an initial design. Next, some procedure is followed for adjustment of selected design parameters or the system configuration in order to achieve an optimum final design.

Unfortunately, humans are rather inept at solving heuristically the multiparameter adjustment problems that frequently arise in engineering design. As a result, engineers are turning more and more frequently to optimization for final design parameter adjustment. Referring to [Pol 1] we find that quite commonly engineering design problems lead to optimization problems with a finite number of design parameters and an infinite number of nonlinear inequality constraints. Such optimization problems are often referred to as *semi-infinite*. They form a special class of nondifferentiable optimization problems. Because they have a great deal of structure, it is possible to devise reasonably efficient algorithms for their solution.

These notes collect in one volume the mathematical results in continuity, differentiability, convexity, properties of max functions, nonsmooth analysis, and optimality conditions which are essential to the understanding of nondifferentiable and semi-infinite optimization. In addition, they present an axiomatic structure which should enable the reader to grasp the essential features of first order algorithms for semi-infinite optimization. Although specific problems are dealt with only superficially in these notes, the reader will find that the papers dealing with specific problems will be accessible to him or her, as a result of familiarity with these notes. The only major topic in nonsmooth analysis which is omitted from these notes is that of semi-smooth functions. The reason for this is that while semi-infinite optimization problems frequently involve semi-smooth functions, they have considerable structure which eliminates the need for the use of the brute force techniques associated with semi-smooth optimization algorithms. For details, the reader is referred to Polak- Mayne-Wardi [Pol 4].

1. Continuity.

We summarize the various concepts of continuity which play a role in optimization theory. Since in the context of optimization algorithms one generally deals with sequences rather than with neighborhoods, we shall give sequential alternatives whenever possible. A good reference on the topics in this section is [Ber 1].

Definition 1.1: A function $f:\mathbb{R}^n \to \mathbb{R}^m$ is said to be *continuous at* $\hat{x} \in \mathbb{R}^n$ if for every $\delta > 0$ there exists a $\hat{\rho} > 0$ such that

$$\|f(x) - f(\hat{x})\| \le \delta \quad \forall\, x \in B(\hat{x}, \hat{\rho}) \tag{1.1}$$

where

$$B(\hat{x}, \hat{\rho}) \triangleq \{x \in \mathbb{R}^n \mid \|x - \hat{x}\| \le \hat{\rho}\} \tag{1.2}$$

$f(\cdot)$ is said to be *continuous* if $f(\cdot)$ is continuous at all $\hat{x} \in \mathbb{R}^n$. ▪

Exercise 1.1 Prove the following result:

Proposition 1.1: $f:\mathbb{R}^n \to \mathbb{R}^m$ is continuous at $\hat{x}$ if and only if for *any* sequence $\{x_i\}_{i=0}^{\infty}$ in $\mathbb{R}^n$, $x_i \to \hat{x}$ *as* $i \to \infty \Rightarrow f(x_i) \to f(\hat{x})$ as $i \to \infty$. ▪

Definition 1.2: A function $f:\mathbb{R}^n \to \mathbb{R}$ is said to be *upper semi-continuous* at $\hat{x}$ (u.s.c.) if for every $\delta > 0$ there exists a $\hat{\rho} > 0$ such that

$$f(x) - f(\hat{x}) \le \delta \quad \forall\, x \in B(\hat{x}, \hat{\rho}) \tag{1.3}$$

$f(\cdot)$ is said to be u.s.c. if it is u.s.c. at all $x \in \mathbb{R}^n$. ▪

Exercise 1.2: Prove the following result:

Proposition 1.2: $f:\mathbb{R}^n \to \mathbb{R}$ is u.s.c. at $\hat{x}$ if for any sequence $\{x_i\}_{i=0}^{\infty}$ in $\mathbb{R}^n$, $x_i \to \hat{x}$ as $i \to \infty \Rightarrow$

$$\overline{\lim} f(x_i) \leq f(\hat{x}) \tag{1.4}$$

■

Definition 1.3: A function $f:\mathbb{R}^n \to \mathbb{R}$ is said to be *lower semi-continuous* if $-f(\cdot)$ is u.s.c. ■

Exercise 1.3: Show that if $f:\mathbb{R}^n \to \mathbb{R}$ is l.s.c. at $\hat{x}$ if and only if for *any* sequence $\{x_i\}_{i=0}^{\infty} \in \mathbb{R}^n$, $x_i \to \hat{x} \Rightarrow \underline{\lim} f(x_i) \geq f(\hat{x})$. ■

The simplest way to think of $\overline{\lim}$ and $\underline{\lim}$ is in terms of cluster points.

Definition 1.4: Let $\{x_i\}_{i=0}^{\infty}$ be a sequence in $\mathbb{R}^n$. Then $\hat{x}$ is said to be a *cluster point* (or *accumulation point*) of $\{x_i\}_{i=0}^{\infty}$ if for any $\delta>0$, $k \geq 0$ there exists an integer $l \geq k$ such that

$$\|x_l - \hat{x}\| \leq \delta \tag{1.6}$$

i.e., $\hat{x}$ is an accumulation point of $\{x_i\}_{i=0}^{\infty}$ if there is a subsequence $\{x_i\}_{i \in K}$, $K \subset \{0,1,2,\ldots\}$ such that $x_i \xrightarrow{K} \hat{x}$ as $i \to \infty$. ■

Fact 1.1: Let S be the set of accumulation points of a bounded sequence $\{y_i\}_{i=0}^{\infty}$ Then S is compact and

$$\overline{\lim} y_i = \max\{y \mid y \in S\} \tag{1.7a}$$
$$\underline{\lim} y_i = \min\{y \mid y \in S\} \tag{1.7b}$$

■

Exercise 1.4: Show that $f:\mathbb{R}^n \to \mathbb{R}$ is u.s.c. if and only if $\forall\, b \in \mathbb{R}$, $\{x \in \mathbb{R}^n \mid f(x) < b\}$ is open. Also, $f:\mathbb{R}^n \to \mathbb{R}$ is l.s.c. if and only if $\forall\, b \in \mathbb{R}$, $\{x \in \mathbb{R}^n \mid f(x) > b\}$ is open. ■

Next we turn to point-to-set functions. For example, let $\varphi:\mathbb{R}^n \text{x} \mathbb{R}^m \to \mathbb{R}$ be a continuous function. We can define the point-to-set valued function

$$F(x) \triangleq \{y \in \mathbb{R}^m \mid \varphi(x,y) \leq 0\} \tag{1.8}$$

which maps $\mathbb{R}^m$ into $2^{\mathbb{R}^m}$. As another example, consider

$$M(x) \triangleq arg \max_{y \in Y} \varphi(x,y) \tag{1.9}$$

where $Y \subset \mathbb{R}^m$ is compact, which also maps $\mathbb{R}^n$ into $2^{\mathbb{R}^m}$.

The most important concept for point-to-set maps is that of upper semi-continuity, though some use can also be made of lower semi-continuity. Note that the definitions, below, have nothing to do with the ones that we gave for functions from $\mathbb{R}^n$ into $\mathbb{R}$.

Definition 1.5: A function (map) $f:\mathbb{R}^n \to 2^{\mathbb{R}^m}$ is said to be *upper-semi-continuous* (u.s.c.) at $\hat{x}$ if

a) $f(\hat{x})$ is compact and

b) for every open set G such that $f(\hat{x}) \subset G$ there exists a $\hat{\rho}>0$ such that $f(x) \subset G$ for all $x \in B(\hat{x},\hat{\rho})$ (See Fig. (1.1)).

A function $f:\mathbb{R}^n \to 2^{\mathbb{R}^m}$ is u.s.c. if it is u.s.c. at every $x \in \mathbb{R}^n$. ■

Definition 1.6: A function $f:\mathbb{R}^n \to 2^{\mathbb{R}^m}$ is said to be *lower-semi-continuous* (l.s.c.)

at $\hat{x}$ if for every open set G such that $f(\hat{x})\cap G\neq\emptyset$ there exists a $\hat{\rho}>0$ such that $f(x)\cap G\neq\emptyset$ for all $x\in B(\hat{x},\hat{\rho})$ where $\emptyset$ denotes the empty set. (See Fig. (1.2)).

A function $f:\mathbb{R}^n\to 2^{\mathbb{R}^m}$ is l.s.c. if it is l.s.c. at every $x\in\mathbb{R}^n$. ▪

Definition 1.6: A function $f:\mathbb{R}^n\to 2^{\mathbb{R}^m}$ is said to be continuous if it is both u.s.c. and l.s.c. ▪

Note that when $f:\mathbb{R}^n\to\mathbb{R}$ is either u.s.c. or l.s.c. in the sense of set valued maps, it is continuous in the ordinary sense.

Exercise 1.5: Prove the following result:

Proposition 1.3: Suppose that $f:\mathbb{R}^n\to 2^{\mathbb{R}^m}$ is l.s.c. at $\hat{x}$ and $f(\hat{x})$ is compact. Then for any $\hat{\delta}>0$ there exists a $\hat{\rho}>0$ such that

$$f(x)\cap B(y,\hat{\delta})\neq\emptyset \quad \forall\, x\in B(\hat{x},\hat{\rho}),\ \forall\, y\in f(\hat{x}) \tag{1.10}$$

▪

Upper and lower semi-continuity can also be given a sequential interpretation in terms of limit points and cluster (accumulation) points.

Definition 1.7: Consider a sequence of sets $\{A_i\}_{i=0}^{\infty}$ in $\mathbb{R}^n$.

a) The point $\hat{x}$ is said to be a *limit point* of $\{A_i\}_{i=0}^{\infty}$ if $d(\hat{x},A_i)\to 0$ as $i\to\infty$, where

$$d(\hat{x},A_i)\triangleq inf\{\|x-\hat{x}\|\ |x\in A_i\} \tag{1.11}$$

i.e., if there exist $x_i\in A_i$ such that $x_i\to\hat{x}$ as $i\to\infty$.

b) The point $\hat{x}$ is a *cluster (accumulation) point* of $\{A_i\}_{i=0}^{\infty}$ if 0 is an accumulation point of $\{d(\hat{x},A_i)\}_{i=0}^{\infty}$, i.e., if there exist $x_i\in A_i$ and a subset $K\subset\{0,1,2,...\}$ such that $x_i\xrightarrow{K}\hat{x}$ as $i\to\infty$.

c) We denote the *set* of *limit points* of $\{A_i\}$ by $\underline{Lim}A_i$ and the *set* of *cluster* points of $\{A_i\}$ by $\overline{Lim}A_i$. ▪

Exercise 1.6: Prove the following result:

Proposition 1.4:

a) A function $f: \mathbb{R}^n \to 2^{\mathbb{R}^m}$, such that $f(x)$ is compact for all $x \in \mathbb{R}^n$ and bounded on bounded sets, is u.s.c. at $\hat{x}$ if and only if for any sequence $x_i \to \hat{x}$ as $i \to \infty$ $\overline{Lim} f(x_i) \subset f(\hat{x})$.

b) A function $f: \mathbb{R}^n \to 2^{\mathbb{R}^m}$ is l.s.c. at $\hat{x}$ if and only if for any sequence $x_i \to \hat{x}$ as $i \to \infty$, $\underline{Lim} f(x_i) \supset f(\hat{x})$. ▪

Exercise 1.7: Suppose that $\varphi(x,\cdot)$ has compact level sets for each $x \in \mathbb{R}^n$. Show that $F(\cdot)$ as defined in (1.8) is u.s.c. ▪

This conludes our excursion into the world of continuity concepts.

2. Derivatives.

We shall now present the specific concepts of differentiation that we need in optimization.

Definition 2.1: Let $f : I\!R^n \to I\!R^m$. We say that $Df : I\!R^n \times I\!R^n \to I\!R^m$ is a *differential* for $f(\cdot)$ at $\hat{x} \in I\!R^n$ if

a) $Df(\hat{x};\cdot)$ is linear.

b)

$$\lim_{||h||\to 0} \frac{||f(\hat{x}+h)-f(\hat{x})-Df(\hat{x};h)||}{||h||} = 0 \tag{2.1}$$

When $f : I\!R^n \to I\!R^m$ has a differential at all $x \in I\!R^n$, we say that $f(\cdot)$ is *differentiable*. ▪

Since $Df(\hat{x};\cdot)$ is a linear map from $I\!R^n$ into $I\!R^m$, there exists a $m \times n$ matrix $\frac{\partial f(\hat{x})}{\partial x}$ such that $Df(\hat{x};h) = \frac{\partial f(\hat{x})}{\partial x} h$ for all $h \in I\!R^n$; $\frac{\partial f(\hat{x})}{\partial x}$ is called a *Jacobian* matrix.

When $f : I\!R^n \to I\!R$ is differentiable, we use the notation $\nabla f(x) = \frac{\partial f(x)^T}{\partial x}$, and call $\nabla f(\cdot)$ the *gradient* of $f(\cdot)$.

Proposition 2.1: Suppose that the function $f : I\!R^n \to I\!R^m$ has a differential $Df(\hat{x};h)$ at $\hat{x}$. Then the ijth component of the Jacobian $\frac{\partial f(\hat{x})}{\partial x}$ is the partial derivative $\frac{\partial f^i(\hat{x})}{\partial x^j}$.

Proof: Set $h = te_j$, where e_j is the jth unit vector in $I\!R^n$. Then $\frac{\partial f(\hat{x})}{\partial x} te_j = t\left[\frac{\partial f(\hat{x})}{\partial x}\right]_{\cdot j}$, the jth column of $\frac{\partial f(\hat{x})}{\partial x}$, and hence, from (2.1), for $j = 1,2,\ldots,m$,

$$\lim_{t\to 0} \frac{\left|f^i(\hat{x}+te_j)-f^i(\hat{x})-t\left[\frac{\partial f(\hat{x})}{\partial x}\right]_{ij}\right|}{t} = 0 \tag{2.2}$$

i.e., $\left[\dfrac{\partial f(\hat{x})}{\partial x}\right]_{ij} = \dfrac{\partial f^i(\hat{x})}{\partial x^j}$. ▪

Definition 2.2: We say that $f: \mathbb{R}^n \to \mathbb{R}^m$ is *locally Lipschitz continuous* at $\hat{x}$ if there exist $L \in [0,\infty)$, $\hat{\rho} > 0$ such that

$$\|f(x) - f(x')\| \le L\|x - x'\| \quad \forall x, x' \in B(\hat{x},\hat{\rho}) \tag{2.3a}$$

▪

Exercise 2.1: Suppose that $f: \mathbb{R}^n \to \mathbb{R}^m$ has a continuous differential $Df(\cdot,\cdot)$ in a neighborhood of $\hat{x}$. Show that f is locally Lipschitz continuous at $\hat{x}$. ▪

It should be noted that the existence of partial derivatives does not ensure the existence of a differential (see e.g. Apostol p. 103 [Apo 1]). Thus consider the function

$$\begin{aligned} f(x,y) &= x + y \ \text{ if } x = 0 \text{ or } y = 0 \\ f(x,y) &= 1 \ \textit{ otherwise} \end{aligned} \tag{2.3b}$$

In this case

$$\frac{\partial f(0,0)}{\partial x} = \lim_{t\to 0} \frac{f(t,0) - f(0,0)}{t} = 1 \tag{2.4a}$$

$$\frac{\partial f(0,0)}{\partial y} = \lim_{t\to 0} \frac{f(0,t) - f(0,0)}{t} = 1 \tag{2.4b}$$

but the function is not even continuous at (0,0). In view of this, the following result is of interest (see Apostol p. 118 [Apo 1]).

Proposition 2.2: Consider a function $f: \mathbb{R}^n \to \mathbb{R}^m$ such that the partial derivatives $\dfrac{\partial f^i(x)}{dx^j}$ exist in a neighborhood of $\hat{x}$, for $i = 1,2,\ldots,n$, $j = 1,2,\ldots,m$. If these partial derivatives are *continuous* at $\hat{x}$, then the differential $Df(\hat{x};h)$ exists. ▪

The following *chain rule* holds.

Proposition 2.3: Suppose that $f:\mathbb{R}^n \to \mathbb{R}^m$ is defined by $f(x)=h(g(x))$ with both $h:\mathbb{R}^l \to \mathbb{R}^m$ and $g:\mathbb{R}^n \to \mathbb{R}^l$ differentiable. Then

$$\frac{\partial f(\hat{x})}{\partial x} = \frac{\partial h(g(\hat{x}))}{\partial x} \frac{\partial g(\hat{x})}{\partial x} \tag{2.5}$$

■

We make frequent use of Taylor's formula with remainder up to order 2. It comes in two forms: in terms of an intermediate point, and in integral form (see Apostol p. 124 [Apo 1] and Dieudonné p. 186 [Die 1]. Also, refer to Apostol p. 124 [Apo 1] for exposition on higher order differentials). We denote by $D^k f(\cdot;\cdot)$ the differential of order k of $f(\cdot)$.

Proposition 2.4: Consider a function $f:\mathbb{R}^n \to \mathbb{R}$. Suppose that $f(\cdot)$ has continuous partial derivatives of order p at each point x of $\mathbb{R}^n$. Then for any $x,y \in \mathbb{R}^n$

$$\begin{aligned} f(y) - f(x) &= \sum_{k=1}^{p-1} \frac{1}{k!} D^k f(x;y-x) \\ &+ \frac{1}{p!} D^p f(z;y-x) \end{aligned} \tag{2.6a}$$

for some $z=x+t(y-x)$, $t \in [0,1]$. ■

When $p=1$, we recognize (2.6a) as being simply the mean value theorem. For $p=2$, $D^2 f(x;y-x)=\langle y-x, \frac{\partial^2 f(x)}{\partial x^2}(y-x)\rangle$, where $\frac{\partial^2 f(x)}{\partial x^2}$ is a matrix of second partial derivatives, i.e., $\left[\frac{\partial^2 f(x)}{\partial x^2}\right]_{ij} = \frac{\partial^2 f(x)}{\partial x^i \partial x^j}$.

For functions $f:\mathbb{R}^n \to \mathbb{R}^m$, with $m>1$, formula (2.6a) is not valid since there is no z of the form stated that works for all the components of $f(\cdot)$. Instead we use the following result (see Dieudonné p. 186 [Die 1]).

Proposition 2.5: Consider a function $f:\mathbb{R}^n \to \mathbb{R}^m$. Suppose that $f(\cdot)$ has continuous partial derivatives of order p at each point x of $\mathbb{R}^n$. Then for any $x,y \in \mathbb{R}^n$,

$$f(y) - f(x) = \sum_{k=1}^{p-1} \frac{1}{k!} D^k f(x; y-x)$$
$$+ \frac{1}{(p-1)!} \int_0^1 (1-s)^{p-1} D^p f(x+s(y-x); y-x) ds \qquad (2.6b)$$

Proof: We shall prove (2.6b) only for $p \leq 2$. For $p=1$, consider the function $g(s)=f(x+s(y-x))$. Then

$$g(1) = f(y), \; g(0) = f(x) \text{ and}$$
$$g(1) - g(0) = \int_0^1 g'(s) ds$$
$$= \int_0^1 Df(x+s(y-x); y-x) ds \qquad (2.7a)$$

which completes the proof for $p=1$.

Next, let $p=2$. Then we have

$$g''(s)(1-s) = \frac{d}{ds}[g'(s)(1-s) + g(s)] \qquad (2.7b)$$

Integrating (2.7b) from 0 to 1 we get

$$g(1) - g(0) - g'(0) = \int_0^1 (1-s) g''(s) ds \qquad (2.7c)$$

which, on rearranging, we recognize as being

$$f(y) - f(x) = \langle \nabla f(x), y-x \rangle + \int_0^1 (1-s) D^2 f(x+s(y-x); (y-x)) ds$$

after substitution for $g(s)$. ▪

Finally, we define directional derivatives which may exist even when a function fails to have a differential.

Definition 2.3: Let $f: \mathbb{R}^n \to \mathbb{R}^m$. We define the *directional derivative* of $f(\cdot)$ at a point $\hat{x} \in \mathbb{R}^n$ in the direction $h \in \mathbb{R}^n$ $(h \neq 0)$ by

$$df(\hat{x}; h) \triangleq \lim_{\substack{t \downarrow 0 \\ t > 0}} \frac{f(\hat{x}+th) - f(\hat{x})}{t} \qquad (2.8)$$

if this limit exists. Note that $t > 0$ is required. ▪

Exercise 2.2: Suppose that $f : \mathbb{R}^n \to \mathbb{R}^m$ has a differential at $\hat{x}$. Show that for any h, the directional derivative $df(\hat{x};h)$ exists and is given by

$$df(\hat{x};h) = Df(\hat{x};h) = \frac{\partial f(x)}{\partial x} h$$

▪

As we shall see later, directional derivatives play a very important part in the theory of optimization.

3. Convexity

Convexity is an enormous subject (e.g. see Rockafellar [Roc 1]). We collect here only a few essential results that we need in optimization. We begin with convex sets.

Definition 3.1: A set $S \subset \mathbb{R}^n$ is said to be *convex* if for any $x', x'' \in S$ and $\lambda \in [0,1]$, $[\lambda x' + (1-\lambda)x''] \in S$. ▪

Exercise 3.1: Suppose $S \subset \mathbb{R}^n$ is convex. Let $\{x_i\}_{i=1}^{k}$ be points in S and let $\{\mu^i\}_{i=1}^{k}$ be scalars such that $\mu^i \geq 0$ for $i=1,2,\ldots,k$ and $\sum_{i=1}^{k} \mu^i = 1$. Show that

$$\left(\sum_{i=1}^{k} \mu^i x_i\right) \in S \tag{3.1}$$

▪

Definition 3.2: Let S be a subset of $\mathbb{R}^n$. We say that coS is the *convex hull* of S if it is the smallest convex set containing S. ▪

Proposition 3.1 (Caratheodory): Let S be a subset in $\mathbb{R}^n$. If $\bar{x} \in coS$, then there exists at most $(n+1)$ distinct points $\{x_i\}_{i=1}^{n+1}$, in S such that $\bar{x} = \sum_{i=1}^{n+1} \mu^i x_i$, $\mu^i \geq 0$, $\sum_{i=1}^{n+1} \mu^i = 1$.

Proof: Clearly, (Rockaffellar, Theorem 2.3 p. 12 [Roc 1])

$$coS = \{x \mid x = \sum_{i=1}^{k_x} \mu^i x_i \ , x_i \in S \ , \mu^i \geq 0 \ , \sum_{i=1}^{k_x} \mu^i = 1, \ k_x \in \mathbb{N}\} \tag{3.2}$$

where $\mathbb{N} \triangleq \{0,1,2,3,\ldots.\}$. Now suppose that

$$\bar{x} = \sum_{i=1}^{\bar{k}} \bar{\mu}^i x_i$$

with $\bar{\mu}^i \geq 0$, $i=1,2,\ldots,\bar{k}$, $\sum_{i=1}^{\bar{k}} \bar{\mu}^i = 1$. Thus, the following system of equations is satisfied:

$$\sum_{i=1}^{\bar{k}} \bar{\mu}^i \begin{bmatrix} x_i \\ 1 \end{bmatrix} = \begin{bmatrix} \bar{x} \\ 1 \end{bmatrix} \tag{3.4}$$

with $\bar{\mu}^i \geq 0$. Suppose that $\bar{k} > n+1$. Then there exist coefficients α^j , $j=1,2,\ldots,\bar{k}$, not all zero, such that

$$\sum_{i=1}^{\bar{k}} \alpha^i \begin{bmatrix} x_i \\ 1 \end{bmatrix} = 0 \tag{3.5}$$

Adding (3.5) multiplied by ϑ to (3.4) we get

$$\sum_{i=1}^{\bar{k}} (\bar{\mu}^i + \vartheta\alpha^i) \begin{bmatrix} x_i \\ 1 \end{bmatrix} = \begin{bmatrix} \bar{x} \\ 1 \end{bmatrix} \tag{3.6}$$

Suppose (w.l.o.g.) that at least one $\alpha^i < 0$. Then there exists a $\bar{\vartheta} > 0$ such that $\bar{\mu}^j + \bar{\vartheta}\alpha^j = 0$ for some j while $\bar{\mu}^i + \bar{\vartheta}\alpha^i \geq 0$ for all other i. Thus we have succeeded in expressing $\bar{x}$ in terms of $\bar{k}-1$ vectors in S. Clearly, these reductions can go on as long as $\bar{x}$ is expressed in terms of more than $(n+1)$ vectors in S. Q.E.D. ▪

Definition 3.3: Let S_1, S_2 be any two sets in $\mathbb{R}^n$. We say that the hyperplane

$$H = \{x \in \mathbb{R}^n \mid \langle x, v \rangle = \alpha\} \tag{3.7}$$

separates S_1 and S_2 if

$$\langle x, v \rangle \geq \alpha \quad \forall\, x \in S_1 \tag{3.8a}$$

$$\langle y, v \rangle \leq \alpha \quad \forall\, y \in S_2 \tag{3.8b}$$

The separation is said to be *strict* if one of the inequalities (3.8a), (3.8b) is satisfied strictly. ▪

Proposition 3.2 (Hahn-Banach): Let S_1, S_2 be two convex sets in $\mathbb{R}^n$ such that $S_1 \cap S_2 = \emptyset$. Then there exists a hyperplane which separates S_1 and S_2. Furthermore, if *Sub* 1 and S_2 are closed and either S_1 or S_2 is compact, then then the separation can be made strict. ∎

Proposition 3.3: Suppose that $S \subset \mathbb{R}^n$ is closed and convex and $0 \notin S$. Let

$$\hat{x} = arg\ \min\{\|x\|^2 \mid x \in S\} \tag{3.9}$$

Then

$$H = \{x \mid \langle \hat{x}, x \rangle = \|\hat{x}\|^2\} \tag{3.10}$$

separates S from 0, i.e., $\langle \hat{x}, x \rangle \geq \|\hat{x}\|^2$ for all $x \in S$.

Proof: Let $x \in S$ be arbitrary. Then, since S is convex, $[\hat{x} + \lambda(x - \hat{x})] \in S$ for all $\lambda \in [0,1]$. By definition of $\hat{x}$, we must have

$$\begin{aligned} 0 < \|\hat{x}\|^2 &\leq \|\hat{x} + \lambda(x - \hat{x})\|^2 \\ &= \|\hat{x}\|^2 + 2\lambda\langle \hat{x}, x - \hat{x} \rangle + \lambda^2\|x - \hat{x}\|^2 \end{aligned} \tag{3.11a}$$

Hence, for all $\lambda \in (0,1]$

$$0 \leq 2\langle \hat{x}, x - \hat{x} \rangle + \lambda\|x - \hat{x}\|^2 \tag{3.11b}$$

Letting $\lambda \to 0$ we get the desired result. ∎

Definition 3.4: Suppose $S \subset \mathbb{R}^n$ is convex. We say that $H = \{x \mid \langle x - \bar{x}, v \rangle = 0\}$ is a *support hyperplane* to S through $\bar{x}$ with *inward (outward) normal* v if $\bar{x} \in \bar{S}$ (the closure of S) and

$$\langle x - \bar{x}, v \rangle \geq 0 (\leq 0) \quad \forall x \in S \tag{3.12}$$

∎

Proposition 3.4: A closed convex set is equal to the intersection of the half spaces which contain it.

Proof: Let C be a closed convex set and A the intersection of half spaces containing C. Then clearly $C \subset A$. Now suppose $\bar{x} \notin C$. Then there exists a support hyperplane H which separates strictly $\bar{x}$ and C, i.e., $\bar{x}$ does not belong to one subspace containing C, i.e., $\bar{x} \notin A$. Hence $C^c \subset A^c$ which leads to the conclusion that $A \subset C$. ▪

Next we turn to convex functions. For an example see Fig. 3.1.

Definition 3.4: A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *convex* if its epigraph is convex, i.e., if for any $x', x'' \in \mathbb{R}^n$ and $\lambda \in [0,1]$,

$$f(\lambda x' + (1-\lambda)x'') \leq \lambda f(x') + (1-\lambda) f(x'') \tag{3.13}$$

▪

A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be *concave* if $-f(\cdot)$ is convex.

Proposition 3.5: Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is convex. Then $f(\cdot)$ is continuous. (For a proof, see Berge p. 193 [Ber 1]). ▪

The following property can be deduced from Fig. 3.1.

Proposition 3.6: Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable. Then $f(\cdot)$ is convex if and only if

$$f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle \quad \forall\, x, y \in \mathbb{R}^n \tag{3.14}$$

Proof: $\Rightarrow$ Suppose $f(\cdot)$ is convex. Then for any $x, y \in \mathbb{R}^n$, $\lambda \in [0,1]$,

$$f(x + \lambda(y-x)) \leq (1-\lambda) f(x) + \lambda f(y) \tag{3.15}$$

Rearranging (3.15) we get

$$\frac{f(x + \lambda(y-x)) - f(x)}{\lambda} \leq f(y) - f(x) \quad \forall\, \lambda \in [0,1] \tag{3.16}$$

Taking the limit as $\lambda \to 0$ we get (3.14).

$\Leftarrow$ Suppose (3.14) holds. Then for any $\lambda \in [0,1]$, $x, y \in \mathbb{R}^n$

$$f(y) - f(x + \lambda(y-x)) \geq \langle \nabla f(x + \lambda(y-x)), y - x \rangle (1-\lambda) \tag{3.17a}$$
$$f(x) - f(x + \lambda(y-x)) \geq \langle \nabla f(x + \lambda(y-x)), y - x \rangle (-\lambda) \tag{3.17b}$$

Multiplying (3.17a) by λ, (3.17b) by $(1-\lambda)$ and adding, we get (3.15), i.e., $f(\cdot)$ is convex. ▪

Proposition 3.7: Suppose that $f:\mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable. Then $f(\cdot)$ is convex if and only if the Hessian (second derivative) matrix $\frac{\partial^2 f(x)}{\partial x^2}$ is positive semi-definite for all $x \in \mathbb{R}^n$.

Proof: $\Rightarrow$ Suppose $f(\cdot)$ is convex. Then for any $x, y \in \mathbb{R}^n$, because of Propositions 3.6 and 2.5

$$0 \le f(y) - f(x) - \langle \nabla f(x), y - x \rangle$$
$$= \int_0^1 (1-s)\langle y - x, \frac{\partial^2 f(x + s(y-x))}{\partial x^2}(y-x)\rangle ds \tag{3.17}$$

Hence, dividing by $\|y-x\|^2$ and letting $y \to x$, we obtain that $\frac{\partial^2 f(x)}{\partial x^2}$ is positive semi-definite.

$\Leftarrow$ Suppose that $\frac{\partial^2 f(x)}{\partial x^2}$ is positive semi-definite for all $x \in \mathbb{R}$. Then it follows directly from the equality in (3.17) and Proposition 3.6 that $f(\cdot)$ is convex. ▪

Exercise 3.2: Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable, it attains its infimum and $\frac{\partial^2 f}{\partial x^2}(x) > 0$, $\forall x \in \mathbb{R}^n$. Show that the level sets that for some $m > 0$, $\langle y, \frac{\partial^2 f}{\partial x^2}(x) y \rangle \ge M \|y\|^2$ for all $x, y \in \mathbb{R}^n$.

Exercise 3.3: Suppose $f^i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, 2, \ldots, m$ are convex. Show that

$$\psi^1(x) \triangleq \max_i f^i(x)$$
$$\psi^2(x) \triangleq \sum_{i=1}^m f^i(x)$$

are both convex. ▪

Definition 3.5: Let $S \subset \mathbb{R}^n$ be convex and compact. We define the *support functional* $\sigma_S : \mathbb{R}^n \to \mathbb{R}$ by

$$\sigma_S(h) \triangleq \max\{\langle h, x \rangle \mid x \in S\} \tag{3.19}$$

Proposition 3.8: Consider $\sigma_S(\cdot)$ as defined by (3.19) with S convex and compact. Then

a) $\sigma_S(\cdot)$ is positive homogeneous, i.e., $\forall\ \lambda \geq 0$,

$$\sigma_S(\lambda h) = \lambda \sigma_S(h) \tag{3.20}$$

b) $\sigma_S(\cdot)$ is subadditive, i.e., $\forall\ h_1, h_2$,

$$\sigma_S(h_1 + h_2) \leq \sigma_S(h_1) + \sigma_S(h_2) \tag{3.21}$$

c) $\sigma_S(\cdot)$ is convex.

Proof:

a) This is immediate.

b) Let $\bar{x} \in S$ be such that $\sigma_S(h_1+h_2) = \langle h_1+h_2, \bar{x} \rangle = \langle h_1, \bar{x} \rangle + \langle h_2, \bar{x} \rangle$. It follows from (3.19) that $\sigma_S(h_i) \geq \langle h_i, \bar{x} \rangle$, for $i = 1,2$. Hence (3.21) follows.

c) Let $h_1, h_2 \in \mathbb{R}^n \cdot \lambda \in [0,1]$ be given. Then

$$\begin{aligned}\sigma_S(\lambda h_1 + (1-\lambda) h_2) &\leq \sigma_S(\lambda h_1) + \sigma_S((1-\lambda) h_2) \\ &= \lambda \sigma_S(h_1) + (1-\lambda)\sigma_S(h_2)\end{aligned}$$

which shows that $\sigma_S(\cdot)$ is convex. ∎

Exercise 3.4: Let $S \subset \mathbb{R}^n$ be convex and compact. Suppose that for a given $h \in \mathbb{R}^n$ $x_h \in S$ is such that $\sigma_S(h) = \langle h, x_h \rangle$. Show that

$$\langle x - x_h, h \rangle \leq 0 \ \ \forall\ x \in S \tag{3.22}$$

i.e., $\{x \in \mathbb{R}^N \mid \langle x, h \rangle = \langle x_h, h \rangle\}$ is a support hyperplane to S with *outward* normal h. ∎

Since by Proposition 3.5 S is the intersection of all the closed half spaces containing it and $\sigma_S(h)$ characterizes such a half space, it should be possible to describe a closed convex set by means of its support function.

Exercise 3.4: Prove the following result.

Proposition 3.9: Let $\sigma: \mathbb{R}^n \to \mathbb{R}$ be a positive homogeneous, subadditive function. Then the set

$$C = \{x \in \mathbb{R}^n \mid \langle x,h \rangle \leq \sigma(h) \quad \forall\, h \in \mathbb{R}^n\} \tag{3.23}$$

is nonempty, convex, compact and $\sigma(\cdot)$ is the support function for C. [Hint: use the fact that $x \in C$ if and only if (-1,x) defines an outward normal to a support hyperplane of the epigraph of $\sigma(\cdot)$ at some point $(\sigma(h),h)$] ▪

Minimax theorems play an important role both in game theory and in optimization. The following one is among the best known.

Theorem 3.1 (Von Neumann): Let $f: \mathbb{R}^n x \mathbb{R}^m \to \mathbb{R}$ be such that $f(x,y)$is convex in x and concave in y and let $X \subset \mathbb{R}^n, Y \subset \mathbb{R}^m$ be compact convex sets. Then

$$\min_{x \in X} \max_{y \in Y} f(x,y) = \max_{y \in Y} \min_{x \in X} f(x,y) \tag{3.24}$$

▪

(For a proof see [Ber 1]).

It is easy to extend the Von Neumann Theorem to the case where either X or Y is unbounded, as follows.

Corollary 3.1: Let $f: \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be such that $f(x,y)$ is convex in x and concave in y and let Y be a compact, convex set in $\mathbb{R}^m$. If $\max_{y \in Y} f(x,y)$ has compact level sets, then

$$\min_{x \in \mathbb{R}^n} \max_{y \in Y} f(x,y) = \max_{y \in Y} \min_{x \in \mathbb{R}^n} f(x,y) \tag{3.25}$$

▪

The result for Y unbounded is obtained by assuming that $\min_{x \in X} f(x,y)$ has compact level sets.

The minimax theorems lead to the following important results.

Proposition 3.10: Let S be a compact convex set in $\mathbb{R}^n$ and let

$B=\{h \in \mathbb{R}^n \mid \|h\| \le 1\}$. Then, with $\sigma_S(\cdot)$ the support function of S,

$$\min_{h \in B} \sigma_S(h) = -\min_{x \in S} \|x\| \tag{3.26a}$$

and

$$\min_{h \in \mathbb{R}^n} \{½ \|h\|^2 + \sigma_S(h)\} = -\min_{x \in S} ½ \|x\|^2 \tag{3.26b}$$

Proof: By definition of $\sigma_S(\cdot)$,

$$\min_{h \in B} \sigma_S(h) = \min_{h \in B} \max_{x \in S} \langle h,x \rangle \tag{3.27}$$

Since B,S are convex and compact and $\langle h,x \rangle$ is convex-concave, by the Van Neumann Theorem we get

$$\min_{h \in B} \sigma_S(h) = \max_{x \in S} \min_{h \in B} \langle x,h \rangle \tag{3.28}$$

Now $\min_{h \in B} \langle x,h \rangle$ is solved by $h=-x/\|x\|$. Hence, substituting in (3.28) we get

$$\begin{aligned} \min_{h \in B} \sigma_S(h) &= \max_{x \in S} -\|x\| \\ &= -\min_{x \in S} \|x\| \end{aligned} \tag{3.29}$$

Next, by Corollary 3.1,

$$\begin{aligned} &\min_{h \in \mathbb{R}^n} \{½\|h\|^2 + \sigma_S(h)\} \\ &= \min_{h \in \mathbb{R}^n} \max_{x \in S} \{½\|h\|^2 + \langle h,x \rangle\} \\ &= \max_{x \in S} \min_{h \in \mathbb{R}^n} \{½\|h\|^2 + \langle h,x \rangle\} \end{aligned} \tag{3.30}$$

Now $\min_{h \in \mathbb{R}^n} \{½\|h\|^2+\langle h,x \rangle\}$ is solved by $h=-x$ (by taking derivatives and setting them to zero). Substituting into (3.30) we obtain

$$\begin{aligned} \min_{h \in \mathbb{R}^n} \{½\|h\|^2 + \sigma_S(h)\} &= \max_{x \in S} -½\|x\|^2 \\ &= -\min_{x \in S} ½\|x\|^2 \end{aligned} \tag{3.31}$$

Q.E.D. ▪

The following obvious corollary plays an important role in the development of optimality conditions for optimization problems.

Corollary 3.2: Let S be a compact convex set in $\mathbb{R}^n$. Then $\sigma_S(h) \geq 0$ for all $h \in \mathbb{R}^n$ if and only if $0 \in S$. ▪

Exercise 3.5: Prove the following.

Proposition 3.11: Let C, D be two convex, compact subsets in $\mathbb{R}^n$. Then $C \subset D$ if and only if $\sigma_C(h) \leq \sigma_D(h)$ for all $h \in \mathbb{R}^n$. ▪

4. Max Functions

Max functions play a central role in optimization problems that arise in engineering design. They are also a particularly tractable kind of nondifferentiable functions. We establish some of their most important properties [Dan 1].

Proposition 4.1: Let $\varphi:\mathbb{R}^n \mathrm{x} \mathbb{R}^m \to \mathbb{R}$ be continuous and $Y:\mathbb{R}^n \to 2^{\mathbb{R}^m}$ u.s.c. Then

$$\psi(x) \triangleq \max\{\varphi(x,y) \mid y \in Y(x)\} \tag{4.1}$$

is u.s.c.

Proof: Let $x_i \to \hat{x}$ as $i \to \infty$, be arbitrary and let $y_i \in Y(x_i)$ be such that $\psi(x_i)=\varphi(x_i,y_i)$. Since $Y(\cdot)$ is u.s.c. and $x_i \to \hat{x}$, $\{y_i\}$ is bounded and hence, since $\varphi(\cdot,\cdot)$ is continuous, $\overline{\lim}\varphi(x_i,y_i)$ exists. Suppose y_i, $i \in K \subset \{0,1,\dots\}$ is such that $\overline{\lim}\varphi(x_i,y_i)=\lim_{i\in K}\varphi(x_i,y_i)$ and $y_i \xrightarrow{K} y^*$. Then $y^* \in Y(\hat{x})$ by u.s.c. of $Y(\cdot)$ and hence

$$\psi(\hat{x}) \ge \varphi(\hat{x},y^*) = \lim_{i\in K}\varphi(x_i,y_i) = \overline{\lim}\,\psi(x_i) \tag{4.2}$$

which completes our proof. ▪

Corollary 4.1: Consider φ and Y as in Proposition 4.1 and suppose that $Y(x)$ is continuous. Then $\psi(x)$ is continuous.

Proof: We only need to show that $\psi(\cdot)$ is l.s.c. under the stronger assumption on $Y(\cdot)$. For the sake of contradiction, suppose there is a point $\hat{x} \in \mathbb{R}^n$ and a sequence $x_i \to \hat{x}$ as $i \to \infty$ such that

$$\lim\psi(x_i) < \psi(\hat{x}) \tag{4.3}$$

Suppose that $\psi(\hat{x})=\varphi(\hat{x},\hat{y})$ with $\hat{y} \in Y(\hat{x})$. Let $y_i \in Y(x_i)$ be such that $\psi(x_i)=\varphi(x_i,y_i)$ and let $\hat{y}_i = \mathit{argmin}\{\|y-\hat{y}\|^2 \mid y \in Y(x_i)\}$. Then, since $Y(\cdot)$ and $\varphi(\cdot,\cdot)$ are continuous, $\hat{y}_i \to \hat{y}$ as $i \to \infty$, so that $\lim\varphi(x_i,\hat{y}_i)=\varphi(\hat{x},\hat{y})$. Hence there exists an i_0 such that $\varphi(x_i,\hat{y}_i) > \psi(x_i)$, which contradicts the definition of $\psi(x_i)$. ▪

Proposition 4.2: Consider the function

$$\psi(x) = \max_{y}\{\varphi(x,y) \mid y \in Y(x)\} \tag{4.4}$$

with $\varphi: \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ continuous and $Y: \mathbb{R}^n \to 2^{\mathbb{R}^m}$. Let

$$\hat{Y}(x) \triangleq \{y \in Y(x) \mid \psi(x) = \varphi(x,y)\} \tag{4.5}$$

Then $\hat{Y}(\cdot)$ is u.s.c.

Proof: Clearly $\hat{Y}(\cdot)$ is bounded on bounded sets and $\hat{Y}(x)$ is compact because $Y(x)$ is compact and $\varphi(x,\cdot)$ is continuous. By Proposition 1.4 we only need to show that $\overline{Lim}\,\hat{Y}(x_i) \subset \hat{Y}(\hat{x})$ for any sequence $\{x_i\}_{i=0}^{\infty}$ converging to a point $\hat{x}$. Suppose this is false, i.e., there exists a point $\hat{x}$ and a sequence $x_i \to \hat{x}$ such that for $y_i \in \hat{Y}(x_i)$ we have $y_i \to \hat{y} \not\in \hat{Y}(\hat{x})$. But this means that $\psi(x_i) = \varphi(x_i, y_i) \to \varphi(\hat{x},\hat{y}) < \psi(\hat{x})$, which contradicts the continuity of $\psi(\cdot)$ (Corollary 4.1). ▪

Next we turn to max functions of the form (4.4) with $\varphi(x,y)$ differentiable in x and $\nabla_x \varphi(x,y)$ continuous. First consider the simplest case where $Y = \{y_1, y_2, \cdots y_m\}$. Letting $f^i(x) = \varphi(x, y_i)$, $i = 1,2,\ldots,m$, (4.4) becomes

$$\psi(x) = \max_{i \in \underline{m}} f^i(x) \tag{4.6}$$

where

$$\underline{m} \triangleq \{1,2,\ldots,m\} \tag{4.7}$$

Drawing the graph of the function $\psi(x+\lambda h)$, for fixed x $h \in \mathbb{R}^n$, which is a function of λ only, we obtain Fig. 4.1 and conclude that $\psi(x)$ is not differentiable everywhere. However, its directional derivative seems to exist and should be equal to the steepest slope of the "active functions", i.e., if we denote $I(x) = \{i \in \underline{m} \mid \psi(x) = f^i(x)\}$, then

$$\begin{aligned} d\psi(x;h) &= \max_{i \in I(x)} df^i(x;h) \\ &= \max_{i \in I(x)} \langle \nabla f^i(x), h \rangle \end{aligned} \tag{4.8}$$

We now show that this result is true in general.

Proposition 4.3: Consider the function

$$\psi(x) = \max\{\varphi(x,y) \mid y \in Y\} \tag{4.9a}$$

where $\varphi: \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is differentiable in x, $\nabla_x \varphi(x,y)$ is continuous in both arguments, and $Y \subset \mathbb{R}^m$ is compact. Then for any $\hat{x}, h \in \mathbb{R}^n$,

$$d\psi(\hat{x};h) = \max_{y \in \hat{Y}(\hat{x})} \langle \nabla_x \varphi(\hat{x},y), h \rangle \tag{4.9b}$$

where $\hat{Y}(\cdot)$ is defined as in (4.5).

Proof: Since $\varphi(x,y)$ is continuously differentiable in x and Y is compact, $\varphi(\cdot,y)$ is uniformly locally Lipschitz continuous. Hence for x', x in a bounded set,

$$\begin{aligned} \psi(x') - \psi(x) &= \varphi(x',y') - \varphi(x,y) \\ &= [\varphi(x',y') - \varphi(x,y')] + [\varphi(x,y') - \varphi(x,y)] \\ &\le L\|x' - x\| \end{aligned} \tag{4.10}$$

where $y' \in \hat{Y}(x')$, $y \in \hat{Y}(x)$ and L is the Lipschitz constant for $\varphi(\cdot,y)$. (Clearly, since $y \in \hat{Y}(x)$, $\varphi(x,y') - \varphi(x,y) \le 0$). Since we can interchange x' and x in (4.10), we conclude that $\psi(\cdot)$ is locally Lipschitz continuous. Hence both $\varliminf_{t \downarrow 0} \frac{\psi(x+th)-\psi(x)}{t}$ and $\varlimsup_{t \downarrow 0} \frac{\psi(x+th)-\psi(x)}{t}$ must exist. Now,

$$\begin{aligned} &\varliminf_{t \downarrow 0} \frac{\psi(x+th)-\psi(x)}{t} \\ &\ge \varliminf_{t \downarrow 0} \max_{y \in \hat{Y}(x)} \frac{\varphi(x+th,y)-\psi(x)}{t} \end{aligned} \tag{4.11a}$$

since $\hat{Y}(x) \subset Y(x)$. Since $\varphi(x,y) = \psi(x)$ for all $y \in \hat{Y}(x)$, we obtain from Proposition 2.5 that

$$\begin{aligned} &\varliminf_{t \downarrow 0} \frac{\psi(x+th)-\psi(x)}{t} \\ &\ge \varliminf_{t \downarrow 0} \max_{y \in \hat{Y}(x)} \int_0^1 \langle \nabla_x \varphi(x+sth,y), h \rangle ds \\ &= \max_{y \in \hat{Y}(x)} \langle \nabla_x \varphi(x,y), h \rangle \end{aligned} \tag{4.11b}$$

where we have made use of the fact that the max function in (4.11b) is continuous in t and h for x fixed. Next,

$$\overline{\lim_{t\downarrow 0}}\frac{\psi(x+th)-\psi(x)}{t}$$

$$=\overline{\lim_{t\downarrow 0}}\max_{y\in Y}\frac{\varphi(x+th,y)-\psi(x)}{t}$$

$$=\overline{\lim_{t\downarrow 0}}\max_{y\in\hat{Y}(x+th)}\frac{\varphi(x+th,y)-\psi(x)}{t} \tag{4.12}$$

$$=\overline{\lim_{t\downarrow 0}}\max_{y\in\hat{Y}(x+th)}\{\int_0^1\langle\nabla_x\varphi(x+sth,y),h\rangle ds+\frac{\varphi(x,y)-\psi(x)}{t}\} \tag{4.13}$$

Now, $\varphi(x,y)\le\psi(x)$ for all $y\in Y$; $\hat{Y}(\cdot)$ is u.s.c. and $\nabla_x\varphi(\cdot,\cdot)$ is continuous. Hence the max of the integral in (4.13) is u.s.c. in t, for x and h fixed, and we get

$$\overline{\lim_{t\downarrow 0}}\frac{\psi(x+th)-\psi(x)}{t}$$

$$\le\max_{y\in\hat{Y}(x)}\langle\nabla_x\varphi(x,y),h\rangle \tag{4.14}$$

Hence the desired result follows. ▪

The following result is obvious.

Corollary 4.2: Let $\psi(\cdot)$ be defined as in (4.9) and suppose that the assumptions of Proposition 4.3 hold. If $\hat{x}\in\mathbb{R}^n$ is such that $\hat{Y}(\hat{x})=\{\hat{y}\}$ a singleton, then $\psi(\cdot)$ has a gradient at $\hat{x}$, with $\nabla\psi(\hat{x})=\nabla_x\varphi(\hat{x},\hat{y})$. ▪

Before we can proceed further we must establish a result in the theory of linear cost optimization problems.

Lemma 4.1: Let m,m' be defined by

$$m\triangleq\max\{\langle c,x\rangle\,|\,x\in X\} \tag{4.15a}$$

$$m'\triangleq\max\{\langle c,x\rangle\,|\,x\in coX\} \tag{4.15b}$$

where $X\subset\mathbb{R}^n$ is a compact set and $c\in\mathbb{R}^n$ is given. Then $m=m'$.

Proof: Since $X \subset coX$, we must have $m' \geq m$. Let $m' = \langle c, \hat{x} \rangle$, $\hat{x} \in coX$. By Caratheodory's Theorem (Proposition 3.1), $\hat{x} = \sum_{i=1}^{n+1} \mu^i x_i$, with $x_i \in X$ and $\mu^i \geq 0$, $\sum_{i=1}^{n+1} \mu^i = 1$. Hence,

$$m' = \langle c, \hat{x} \rangle = \sum_{i=1}^{n+1} \mu^i \langle c, x_i \rangle \leq \left[\sum_{i=1}^{n+1} \mu^i \right] \langle c, x_k \rangle$$
$$= \langle c, x_k \rangle \leq m \,, \tag{4.15c}$$

where $\langle c, x_k \rangle = \max_{i \in \underline{n+1}} \langle c, x_i \rangle$ and $\underline{n+1} \triangleq \{1,2,\ldots,n+1\}$. This completes our proof. ▪

Proposition 4.4: Consider the function $\psi(x)$ defined in (4.9), with assumptions as in Proposition 4.3. Let

$$\partial\psi(x) \triangleq co\{\nabla_x \varphi(x,y) \mid y \in \hat{Y}(x)\} \tag{4.15}$$

Then the directional derivative $d\psi(x;\cdot)$ is the support function for $\partial\psi(x)$.

Proof: By (4.9b), for any $h \in I\!R^n$,

$$d\psi(x;h) = \max_{y \in \hat{Y}(x)} \langle \nabla_x \varphi(x,y), h \rangle$$
$$= \max\{\langle h, z \rangle \mid z = \nabla_x \varphi(x,y),\ y \in \hat{Y}(x)\}$$
$$= \max\{\langle h, z \rangle \mid z \in \partial\psi(x)\} \tag{4.16}$$

by Lemma 4.1, which completes our proof. ▪

Exercise 4.1: Determine whether Proposition 4.3 remains valid when the constant set Y is replaced by a continuous set valued map Y(x). ▪

Exercise 4.2: Consider the function $\psi(x) = \max\{\varphi(x,y) \mid y \in Y\}$ where $\varphi: I\!R^n \times I\!R^m \to I\!R$ is continuous in $x, \nabla_x \varphi(x,y)$ exists and is continuous, and $Y \subset I\!R^m$ is compact. Show that a steepest descent direction for ψ can be computed in two ways by showing that

$$\underset{h}{\mathrm{argmin}}\{d\psi(x;h) \mid \|h\| \leq 1\} = \underset{h}{\mathrm{argmin}}\{\tfrac{1}{2}\|h\|^2 + d\psi(x;h)\} \tag{4.17}$$

▪

5. First Order Optimality Conditions: The Differentiable Case.

We shall now develop first order optimality conditions for two "differentiable" optimization problems: one unconstrained and one with inequality constraints. Optimality conditions for problems with both equality and inequality constraints then follow by extension.

Definition 5.1: Consider the problem P:$\min\{f(x) \mid x \in X\}$ where $f: \mathbb{R}^n \to \mathbb{R}$ is continuous and $X \subset \mathbb{R}^n$. We say that $\hat{x}$ is a *global solution* to P if $f(\hat{x}) \leq f(x)$ $\forall$ $x \in X$. We say that $\hat{x}$ is a *local solution* to P if there exists a $\hat{\rho} > 0$ such that $f(\hat{x}) \leq f(x)$ $\forall$ $x \in X$ such that $\|x - \hat{x}\| < \hat{\rho}$. ▪

Proposition 5.1: Consider the problem

$$\min\{f(x) \mid x \in \mathbb{R}^n\} \tag{5.1}$$

where $f: \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. Suppose that $\hat{x}$ is a global solution to (5.1), then $\nabla f(\hat{x}) = 0$.

Proof: Suppose $\hat{x}$ is a global solution to (5.1). Then we must have

$$df(\hat{x}; h) \geq 0 \quad \forall\, h \in \mathbb{R}^n \tag{5.2}$$

for otherwise there would be a direction $\hat{h}$ such that

$$df(\hat{x}; \hat{h}) \triangleq \lim_{t \downarrow 0} \frac{f(\hat{x} + t\hat{h}) - f(\hat{x})}{t} < 0 \tag{5.3}$$

and hence for a finite $\hat{t} > 0$, $f(\hat{x} + t\hat{h}) < f(\hat{x})$ would hold. Now, since $f(\cdot)$ is differentiable,

$$df(\hat{x}; h) = \langle \nabla f(\hat{x}), h \rangle \tag{5.4}$$

and hence (5.2) can hold for all $h \in \mathbb{R}^n$ if and only if $\nabla f(\hat{x}) = 0$. ▪

The following result is obvious.

Corollary 5.1: Consider the problem (5.1) under conditions stated in Proposition

5.1. Suppose that $\hat{x}$ is a local solution to (5.1). Then $\nabla f(\hat{x})=0$. ▪

The following result is suggested by Fig. 5.1 for the simple case where there is only a finite number of inequality constraints. Note that for the "active" gradients in Fig. 5.1 the origin is moved to the optimal point $\hat{x}$.

Proposition 5.2: Consider the problem

$$\min\{f(x) \mid \varphi(x,y) \le 0 \ \forall\, y \in Y\} \tag{5.5}$$

where $f: \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, $\varphi: \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is continuous and continuously differentiable in x (i.e., $\nabla_x \varphi(x,y)$ is continuous), and $Y \subset \mathbb{R}^m$ is compact. If $\hat{x}$ is a local solution to (5.5), then

$$\begin{gathered} 0 \in co\{\nabla f(\hat{x});\ \nabla_x \varphi(\hat{x},y)\,,\ y \in \hat{Y}(\hat{x})\} \\ \text{if } \psi(\hat{x}) = 0 \\ 0 = \nabla f(\hat{x}) \ \textit{otherwise} \end{gathered} \tag{5.6}$$

where, as before,

$$\psi(\hat{x}) \triangleq \max\{\varphi(x,y) \mid y \in Y\} \tag{5.7a}$$

$$\hat{Y}(x) \triangleq \{y \in Y \mid \varphi(x,y) = \psi(x)\} \tag{5.7b}$$

Proof: Let

$$\begin{aligned} F(x) &\triangleq \max\{f(x) - f(\hat{x}), \psi(x)\} \\ &= \max\{f(x) - f(\hat{x});\ \varphi(x,y),\ y \in Y\} \end{aligned} \tag{5.8}$$

Note that $F(\hat{x})=0$, since $\psi(\hat{x}) \le 0$ and that $F(x) \ge 0$ for all $x \in B(\hat{x},\rho)$, for some $\rho > 0$, because $f(x) - f(\hat{x}) \ge 0$ when $\psi(x) \le 0$ and $x \in B(\hat{x},\rho)$. Hence $\hat{x}$ is a global minimizer of $F(x)$. Since $F(\cdot)$ is directionally differentiable by Proposition 4.3, we must have

$$dF(\hat{x};h) \ge 0 \ \forall\, h \in \mathbb{R}^n \tag{5.9}$$

since the existence of a $\hat{h} \ne 0$ such that $dF(\hat{x};\hat{h}) < 0$ implies that $F(\hat{x}+t\hat{h}) < F(\hat{x}) < 0$ for some $t > 0$, which is clearly impossible.

Now, by Proposition 4.4

$$dF(\hat{x};h) = \max\{\langle z,h\rangle \mid z\in Z(\hat{x})\} \tag{5.10}$$

where

$$Z(\hat{x}) \triangleq co\,\{\nabla f(\hat{x});\ \nabla_x\varphi(\hat{x},y),\ y\in Y^*(\hat{x})\} \tag{5.11}$$

with $Y^*(\hat{x})=\hat{Y}(\hat{x})$ if $\psi(\hat{x})=0$ and $Y^*(\hat{x})$ the empty set otherwise. It now follows from Corollary 3.2 that $0\in Z(\hat{x})$, i.e., that (5.6) holds. Q.E.D. ▪

Exercise 5.1: Prove the following.

Corollary 5.2: Suppose that $\hat{x}$ solves (5.5) and that the assumptions of Proposition 5.2 are satisfied. Then there exist at most $(n+2)$ points in the set $Z(\hat{x})$: $\nabla f(\hat{x})$, $\nabla_x\varphi(\hat{x},y_i)$ $i=1,2,\ldots,n+1$, such that

$$\mu^0\nabla f(\hat{x}) + \sum_{i=1}^{n+1}\mu^i\nabla_x\varphi(\hat{x},y_i) = 0 \tag{5.12}$$

where $\mu^i\geq 0$ for $i=0,1,\ldots,n+1$, $\sum_{i=0}^{n+1}\mu^i=1$. Furthermore, if either $\psi(\hat{x})<0$ or $\psi(\hat{x})=0$ and $0\notin co\{\nabla_x\varphi(\hat{x},y)\mid y\in\hat{Y}(\hat{x})\}$, then $\mu^0>0$. ▪

Exercise 5.2: Use the fact that an equation $h(x)=0$ can be replaced by the two inequalities $h(x)\leq 0$, $-h(x)\leq 0$, to prove the following result.

Proposition 5.3: Consider the problem

$$\min\{f(x)\mid h(x)=0,\ \varphi(x,y)\leq 0\ \forall\, y\in Y\} \tag{5.13}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ and $h:\mathbb{R}^n\to\mathbb{R}^l$ are continuously differentiable and $\varphi:\mathbb{R}^n\times\mathbb{R}^m$ is continuous and continuously differentiable in x (i.e., $\nabla_x\varphi(x,y)$ is continuous),, and $Y\subset\mathbb{R}^m$ is compact. If $\hat{x}$ solves (5.13) then for some $\psi\in\mathbb{R}^l$, $y_i\in Y^*(\hat{x})$, $i=1,2,\ldots n+1$, and $\mu^0,\mu^1,\cdots,\mu^{n+1}\geq 0$ such that $(\mu,\psi)\neq 0$ (where $\mu=(\mu^0,\mu^1,\ldots,\mu^{n+1})$)

$$\mu^0 \nabla f(\hat{x}) + \sum_{i=1}^{n+1} \mu^i \nabla_x \varphi(\hat{x}, y_i) + \frac{\partial h(\hat{x})^T}{\partial x} \psi = 0 \tag{5.14}$$

■

Exercise 5.3: Develop conditions which ensure that $\mu^0 \neq 0$ in (5.14). ■

6. Nondifferentiable Analysis and Optimization

We now turn to real valued functions on $\mathbb{R}^n$ which are assumed to be only locally Lipschitz continuous (l.L.c.) (see Definition 2.2). Functions within this category that are particularly important in engineering design are the max functions discussed in Section 4, eigenvalues and singular values of various system matrices [Pol 6], and max min max functions discussed in [Pol 7], in connection with tolerancing and tuning problems. We begin by stating a key property of l.L.c. functions, the Rademacher Theorem [Ste 1].

Proposition 6.1: Suppose $f:\mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz continuous. Then $\nabla f(x)$ exists for almost all $x \in \mathbb{R}^n$. ▪

The following results are culled from the book by F. H. Clarke [Cla 1]. First, a l.L.c. function may fail to have directional derivatives at a point $x \in \mathbb{R}^n$. This has led to the following extension of the concept of directional derivative.

Definition 6.2: Let $f:\mathbb{R}^n \to \mathbb{R}$ be l.L.c. We defined the *generalized directional derivative* of $f(\cdot)$ at $x \in \mathbb{R}^n$ in the direction $h \in \mathbb{R}^n$ by

$$d_0 f(x;h) \triangleq \overline{\lim_{\substack{t \downarrow 0 \\ y \to x}}} \frac{f(y+th)-f(y)}{t} \tag{6.1}$$

▪

Since there exist $\varepsilon > 0$, $L > 0$ such that $|f(y+th)-f(y)| \leq tL\|h\|$, for all $y \in B(x,\varepsilon)$, $0 \leq t < \varepsilon$, it is clear that $d_0 f(x;h)$ is well defined.

Exercise 6.1: Let $\varphi:\mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be a continuous function such that $\nabla_x \varphi(x,y)$ exists and is continuous and let Y be a compact subset of $\mathbb{R}^m$. Let

$$\psi(x) \triangleq \max \{\varphi(x,y) \mid y \in Y\} \tag{6.2a}$$

$$\zeta(x) = \min \{\varphi(x,y) \mid y \in Y\} \tag{6.2b}$$

show that for any $x, h \in \mathbb{R}^n$,

$$d\psi(x;h) = d_0\psi(x;h) \tag{6.3a}$$

$$d\zeta(x;h) \le d_0\zeta(x;h) \tag{6.3b}$$

Hint: Use Proposition 4.3. ▪

Proposition 6.1: The generalized directional derivative $d_0 f(x;h)$ of a l.L.c. function $f: I\!R^n \to I\!R$, defined by (6.1), has the following properties:

a) $h \to d_0 f(x;h)$ is (i) positive homogeneous and (ii) subadditive on $I\!R^n$.

b) If L is a local Lipschitz constant for $f(\cdot)$ at x, then for any $h \in I\!R^n$

$$|d_0 f(x;h)| \le L\|h\| \tag{6.4}$$

c) $(x,h) \to d_0 f(x;h)$ is u.s.c.

d) $h \to d_0 f(x;h)$ is Lipschitz continuous with constant L, where L is a local Lipschitz constant for $f(\cdot)$ at x.

e) For any $h \in I\!R^n$, $d_0 f(x;-h) = d_0(-f)(x;h)$.

Proof:

a) (i) For any $\lambda > 0$, and $x, h \in I\!R^n$,

$$\begin{aligned} d_0 f(x;\lambda h) &= \overline{\lim_{\substack{t \downarrow 0 \\ y \to x}}} \frac{f(y+t\lambda h) - f(x)}{t} \\ &= \lambda \overline{\lim_{\substack{t \downarrow 0 \\ y \to x}}} \frac{f(y+t\lambda h) - f(x)}{t\lambda} \\ &= \lambda d_0 f(x;h) \end{aligned} \tag{6.5}$$

which shows that $d_0 f(x;\cdot)$ is positive homogeneous.

a) (ii) For any $x, h_1, h_2 \in I\!R^n$,

$$\begin{aligned} d_0 f(x;h_1+h_2) &= \overline{\lim_{\substack{t \downarrow 0 \\ y \to x}}} \frac{f(y+t(h_1+h_2)) - f(y)}{t} \\ &= \overline{\lim_{\substack{t \downarrow 0 \\ y \to x}}} \Big\{ \frac{f(y+th_1+th_2)) - f(y+th_1)}{t} \\ &\quad + \frac{f(y+th_1) - f(y)}{t} \Big\} \end{aligned}$$

$$\leq d_0 f(x;h_2) + d_0 f(x;h_1) \tag{6.6}$$

which shows that $d_0 f(x;\cdot)$ is subadditive.

b) Since $|f(y+th)-f(y)| \leq Lt\|h\|$ for any $\|y-x\|$ and t sufficiently small, (6.4) follows.

c) Let $x_i \to x$ and $h_i \to h$ as $i \to \infty$. We must show that $\overline{\lim} d_0 f(x_i,h_i) \leq d_0 f(x,h)$. By definition of $\overline{\lim}$, for every i there exists $y_i \in \mathbb{R}^n$ and $t_i > 0$ such that $t_i \downarrow 0$ as $i \to \infty$, $\|y_i - x_i\| + t_i \leq \frac{1}{i}$ and

$$\begin{aligned} d_0 f(x_i;h_i) - \frac{1}{i} &\leq \frac{f(y_i + t_i h_i) - f(y_i)}{t_i} \\ &= \frac{f(y_i + t_i h) - f(y_i)}{t_i} + \frac{f(y_i + t_i h_i) - f(y_i + t_i h)}{t_i} \end{aligned} \tag{6.7}$$

Hence,

$$\begin{aligned} \overline{\lim} d_0 f(x_i;h_i) &\leq \overline{\lim}\{\frac{1}{i} + \frac{f(y_i + t_i h) - f(y_i)}{t_i} \\ &\quad + L\|h_i - h\|\} \\ &\leq d_0 f(x;h) \end{aligned} \tag{6.8}$$

which shows that $d_0 f(\cdot;\cdot)$ is u.s.c.

d) For any y in a neighborhood of $x \in \mathbb{R}^n, h_1, h_2 \in \mathbb{R}^n$ and t sufficiently small, we have

$$f(y + th_1) - f(y) \leq f(y + th_2) - f(y) + Lt\|h_1 - h_2\| \tag{6.9}$$

Hence

$$d_0 f(x;h_1) \leq d_0 f(x;h_2) + L\|h_1 - h_2\| \tag{6.10}$$

Since we can interchange h_1 and h_2 in (6.10), it follows that

$$|d_0 f(x;h_1) - d_0 f(x;h_2)| \leq L\|h_1 - h_2\| \tag{6.11}$$

which shows that $d_0 f(x;\cdot)$ is Lipschitz continuous.

e) For any $h \in \mathbb{R}^n$,

$$d_0 f(x;-h) = \overline{\lim_{\substack{t \downarrow 0 \\ y \to x}}} \frac{f(y - th) - f(y)}{t} \tag{6.12}$$

Let $z = y - th$, then (6.12) becomes

$$\begin{aligned} d_0 f(x;-h) &= \overline{\lim_{\substack{t \downarrow 0 \\ z \to x}}} \frac{(-f)(z + th) - (-f(z))}{t} \\ &= d_0(-f)(x;h) \end{aligned} \tag{6.13}$$

which completes our proof. ▪

By Proposition 3.9, the generalized directional derivative $d_0 f(x;\cdot)$ can be used to define a convex set for which it is the support function.

Definition 6.1: Let $f : \mathbb{R}^n \to \mathbb{R}$ be l.L.c. We define the *generalized gradient* of $f(\cdot)$ at x by

$$\partial f(x) \triangleq \{\xi \in \mathbb{R}^n \mid d_0 f(x;h) \geq \langle \xi, h \rangle, \ \forall\, h \in \mathbb{R}^n\} \tag{6.14}$$

▪

Exercise 6.2

Let $\varphi : \mathbb{R}^n \mathrm{x} \mathbb{R}^m \to \mathbb{R}$ be a continuous function such that $\nabla_x \varphi(x,y)$ exists and is continuous and let Y be a compact subset of $\mathbb{R}^m$. Let

$$\psi(x) \triangleq \max \{\varphi(x,y) \mid y \in Y\} \tag{6.2a}$$

We now elucidate the reasons for calling the set $\partial f(x)$ the generalized gradient of $f(\cdot)$. First, suppose that $f(\cdot)$ is differentiable at x. Then, $d_0 f(x;h) = df(x;h) = \langle \nabla f(x), h \rangle$ for any $h \in \mathbb{R}^n$. By definition (6.14), for any $\xi \in \partial f(x)$

$$\langle \nabla f(x) - \xi,\ h \rangle \geq 0 \ \ \forall\, h \in \mathbb{R}^n \tag{6.15}$$

Hence we must have $\nabla f(x) - \xi = 0$ for all $\xi \in \partial f(x)$, i.e., $\partial f(x) = \{\nabla f(x)\}$. Next, suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is l.L.c. and convex. Then its epigraph is convex and, at any point $(\hat{x}, f(\hat{x}))$ the epigraph has one or more support hyperplanes, with normal

$(-1,\xi)\in\mathbb{R}^{n+1}$, such that

$$\langle(-1,\xi),\ (f(x)-f(\hat{x}),\ x-\hat{x})\rangle \le 0 \quad \forall\, x\in\mathbb{R}^n \tag{6.16}$$

as shown in Figure 6.1. Hence

$$\langle\xi,\ (x-\hat{x})\rangle \le f(x)-f(\hat{x}) \quad \forall\, x\in\mathbb{R}^n \tag{6.17}$$

Now let $x=\hat{x}+th$, for any $h\in\mathbb{R}^n$, $t>0$. Then we get

$$\langle\xi,h\rangle \le \overline{\lim_{t\downarrow 0}}\ \frac{f(\hat{x}+th)-f(\hat{x})}{t} \le d_0 f(\hat{x};h) \tag{6.17b}$$

i.e., $\xi\in\partial f(\hat{x})$. Finally we have

Proposition 6.2: Suppose that $f:\mathbb{R}^n\to\mathbb{R}$ is l.L.c. with constant L in a ball centered on $\hat{x}$. Then

a) $\partial f(\hat{x})$ is nonempty, convex and compact, and $\|\xi\|\le L$ for all $\xi\in\partial f(\hat{x})$.

b) For every $h\in\mathbb{R}^n$,

$$d_0 f(\hat{x};h) = \max\{\langle\xi,h\rangle \mid \xi\in\partial f(x)\} \tag{6.18}$$

c)

$$\partial f(x) = G(x) \triangleq co \lim_{x_i\to x}\{\nabla f(x_i)\} \tag{6.19}$$

where the convex hull is taken over all sequences $\{x_i\}$, such that the $\nabla f(x_i)$ exist for all $i\in\mathbb{N}$ and $\{\nabla f(x_i)\}_{i=1}^{\infty}$ converges.

d) $\partial f(x)$ is u.s.c.

Proof:

a) The fact that $\partial f(x)$ is nonempty, convex and compact follows from Proposition 3.9. Next we have by definition of $\partial f(x)$ that for any $\xi\in\partial f(x)$

$$\langle\xi,h\rangle \le d_0 f(x;h) \quad \forall\, h\in\mathbb{R}^n \tag{6.20}$$

It now follows from (6.4) that for $h=\xi$,

$$\|\xi\|^2 \le d_0 f(x;\xi) \le L\|\xi\| \tag{6.21}$$

which shows that $\|\xi\| \le L$.

b) This follows directly from Proposition 3.9.

c) Let $x_i \to \hat{x}$ as $i \to \infty$ be such that $\nabla f(x_i)$ exists for all $i \in \mathbb{N}$ and $\nabla f(x_i) \to \hat{\xi}$ as $i \to \infty$. Then for any $h \in \mathbb{R}^n$, $t_i \downarrow 0$ as $i \to \infty$,

$$\begin{aligned} d_0 f(\hat{x};h) &\ge \lim_{i\to\infty} \frac{f(x_i+t_ih) - f(x_i)}{t_i} \\ &= \lim_{i\to\infty}[\langle \nabla f(x_i),h\rangle + o(t_i)] \\ &= \langle \hat{\xi},h\rangle \end{aligned} \tag{6.22}$$

Hence $\hat{\xi} \in \partial f(\hat{x})$. It remains to show that $\partial f(\hat{x})$ is contained in $G(\hat{x})$ the convex hull of the gradient limits. We shall make use of Proposition 3.11. We note that $G(\hat{x})$ defined by (6.19) is convex and compact. Hence to show that $\partial f(\hat{x}) \subset G(\hat{x})$ we only need to show that $d_0 f(\hat{x};h) \le \sigma_{G(\hat{x})}(h)$ for all $h \in \mathbb{R}^n$, where $\sigma_{G(\hat{x})}(\cdot)$ is the support function of $G(\hat{x})$. Let $\varepsilon > 0$ be arbitrary and $h \in \mathbb{R}^n$ be given. We denote by $X \subset \mathbb{R}^n$ the set of measure 0 where $\nabla f(x)$ does not exist. Let

$$\alpha \triangleq \overline{\lim_{y\to\hat{x}}} \{\langle \nabla f(y),h\rangle \mid y \notin X\} \tag{6.23}$$

By definition of $\overline{\lim}$, there exists a $\delta > 0$ such that if $\|y-\hat{x}\| < \delta$ and $y \notin X$, then $\langle \nabla f(y),h\rangle \le \alpha+\varepsilon$. Furthermore, for almost all $y \in B(\hat{x},\delta)$, the gradient $\nabla f(y+sh)$ exists for almost all s. Hence, for sufficiently small t and almost all y such that $\|\hat{x}-y\| < \frac{\delta}{2}$ (by an extension of Proposition 2.5)

$$\begin{aligned} f(y+th) - f(y) &= \int_0^1 t\langle \nabla f(y+sth),h\rangle ds = \int_0^t \langle \nabla f(y+sh),h\rangle ds \\ &\le t(\alpha+\varepsilon) \end{aligned} \tag{6.24}$$

because $\|y+sh-\hat{x}\| < \delta$ for $t \ge s \ge 0$ sufficiently small. Since $f(\cdot)$ is continuous, sets of measure zero can be discarded in computing $\overline{\lim}$, and hence

$$d_0 f(\hat{x},h) = \overline{\lim_{\substack{t\downarrow 0 \\ y\to\hat{x}}}} \frac{f(y+th) - f(y)}{t} \le \alpha + \varepsilon \tag{6.25}$$

Since (6.24) holds for all $\varepsilon>0$, we conclude that $d_0 f(\hat{x};h) \le \alpha = \max\{\langle h,\xi\rangle \mid \xi memberco \{\lim_{x_i\to\hat{x}} \nabla f(x_i),\ x_i \notin X\}$,

i.e., that (6.19) holds.

d) Since $\partial f(x)$ is compact for all x and bounded on bounded sets, to prove that it is u.s.c., we only need to show that if $x_i \to \hat{x}$ as $i\to\infty$, then $\overline{Lim}\,\partial f(x_i) \subset \partial f(\hat{x})$ (Proposition 1.4). Thus, suppose that $x_i \to \hat{x}$ as $i\to\infty$ and $\xi_i \in \partial f(x_i)$ are such that $\xi_i \to \hat{\xi}$. Since $d_0 f(\cdot;h)$ is u.s.c., for any $h \in \mathbb{R}^n$,

$$d_0 f(\hat{x};h) \ge \overline{\lim}\, d_0 f(x_i;h) \ge \overline{\lim}\langle \xi_i,h\rangle = \langle \hat{\xi},h\rangle \tag{6.26}$$

Hence $\hat{\xi} \in \partial f(\hat{x})$, which completes our proof. ▪

Exercise 6.3

Let $\varphi: \mathbb{R}^n \mathrm{x} \mathbb{R}^m \to \mathbb{R}$ be a continuous function such that $\nabla_x \varphi(x,y)$ exists and is continuous and let Y be a compact subset of $\mathbb{R}^m$. Let

$$\psi(x) \triangleq \max\{\varphi(x,y) \mid y \in Y\} \tag{6.27a}$$

Show that

$$\partial\psi(x) = \underset{y\in\hat{Y}(x)}{\mathrm{co}}\{\nabla_x \varphi(x,y)\} \tag{6.27b}$$

where

$$\hat{Y}(x) = \{y \in Y \mid \varphi(x,y) = \psi(x)\} \tag{6.27c}$$

Proposition 6.3: Suppose that $f_1, f_2: \mathbb{R}^n \to \mathbb{R}$ are l.L.c. Then for any $x \in \mathbb{R}^n$,

$$\partial[f_1 + f_2](x) \subset \partial f_1(x) + \partial f_2(x) \tag{6.28a}$$

Proof: Clearly, for all x, $h \in \mathbb{R}^n$,

$$d_0(f_1 + f_2)(x;h) \le d_0 f_1(x;h) + d_0 f_2(x;h) \tag{6.28b}$$

Hence for all $h \in \mathbb{R}^n$,

$$\max\{\langle \xi,h\rangle \mid \xi \in \partial(f_1 + f_2)(x)\}$$
$$\le \max\{\langle \xi_1,h\rangle \mid \xi_1 \in \partial f_1(x)\} + \max\{\langle \xi_2,h\rangle \mid \xi_2 \in \partial f_2(x)\}$$

$$= \max \{\langle \xi_1+\xi_2, h\rangle \mid \xi_1 \in \partial f_1(x),\ \xi_2 \in \partial f(x_2)\} \tag{6.29}$$

The desired result now follows from Proposition 3.11. ▪

Proposition 6.4: Suppose that $f^1, f^2, \cdots, f^m : \mathbb{R}^n \to \mathbb{R}$ are l.L.c. and let

$$\psi(x) \triangleq \max_{j \in \mathbf{m}} f^j(x) \tag{6.30}$$

Then

$$\partial\psi(x) \subset co\,\{\partial f^j(x)\}_{j \in I(x)} \tag{6.31}$$

where $I(x) \triangleq \{j \in \underline{m} \mid f^j(x) = \psi(x)\}$ and $\underline{m} \triangleq \{i, 2, \ldots, m\}$.

Proof: First, given any x, $h \in \mathbb{R}^n$, there exists $\hat{\rho} > 0$, $\hat{t} > 0$ such that if $j \notin I(x)$, then $j \notin I(y+th)$ for al $t \in [0, \hat{t}]$, $\|y-x\| \leq \hat{\rho}$. Hence $I(y+th) \subset I(x)$ for all such y and t. Therefore, for such y, t,

$$\begin{aligned} \frac{\psi(y+th) - \psi(y)}{t} &= \max_{j \in \mathbf{m}} \frac{f^j(y+th) - \psi(y)}{t} \\ &= \max_{j \in I(x)} \frac{f^j(y+th) - f^j(y)}{t} \end{aligned} \tag{6.32}$$

since $\overline{\lim}$ and max are interchangeable operations, we get that

$$d_0\psi(x;h) \leq \max_{j \in I(x)} d_0 f^j(x;h) \tag{6.33}$$

that is,

$$\begin{aligned} \max \{\langle \xi, h\rangle \mid \xi \in \partial\psi(x)\} &\leq \max_{j \in I(x)} \max \{\langle \xi_j, h\rangle \mid \xi_j \in \partial f^j(x)\} \\ &= \max \{\langle \xi, h\rangle \mid \xi \in co\,\{\partial f^j(x)\}_{j \in I(x)}\} \end{aligned} \tag{6.34}$$

which, in light of Proposition 3.11, completes our proof. ▪

It is also possible to establish a chain rule. We shall present only the simplest case.

Proposition 6.5: Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is l.L.c. and, for any $x, y \in \mathbb{R}^n$ given, let $g : \mathbb{R} \to \mathbb{R}$ be defined by

$$g(t) \triangleq f(x+t(y-x)) \tag{6.35}$$

Then for any $t \in \mathbb{R}$

$$\partial g(t) \subset \{\gamma \mid \gamma = \langle \xi, y-x \rangle, \xi \in \partial f(x_t)\} \triangleq G(t) \tag{6.36}$$

where $x_t = x + t(y-x)$.

Proof: Clearly, $g(\cdot)$ is l.L.c. Since the sets on the left and the right of (6.36) are intervals, to establish (6.36), it suffices to prove (see Proposition 3.11) that for $h = \pm 1$,

$$\max\{\gamma h \mid \gamma \in \partial g(t)\} \le \max\{\gamma h \mid \gamma \in G(t)\} \tag{6.37}$$

Now, the left hand side of (6.37) is just $d_0 g(t;h)$. Hence

$$\begin{aligned}
\max\{\gamma h \mid \gamma \in \partial g(t)\} &= d_0 g(t;h) \\
&= \overline{\lim_{\substack{s \to t \\ \lambda \downarrow 0}}} \frac{g(s+\lambda h) - g(s)}{\lambda} \\
&= \overline{\lim_{\substack{s \to t \\ \lambda \downarrow 0}}} \frac{f(x+(s+\lambda h)(y-x)) - f(x+s(y-x))}{\lambda} \\
&\le \overline{\lim_{\substack{z \to x_t \\ \lambda \downarrow 0}}} \frac{f(z+\lambda h(y-x)) - f(z)}{\lambda} \\
&= d_0 f(x_t; h(y-x)) \\
&= \max\{\langle \xi, h(y-x) \rangle \mid \xi \in \partial f(x_t)\} \\
&= \max\{h\gamma \mid \gamma \in G(t)\}
\end{aligned} \tag{6.38}$$

which completes our proof. ▪

More generally, we can prove the following result, which, again, can be generalized to vector valued functions.

Proposition 6.5: Let $h: \mathbb{R}^n \to \mathbb{R}^m$ and $g: \mathbb{R}^m \to \mathbb{R}$ be l.L.c. (for $h(\cdot)$ componentwise) and let $f: \mathbb{R}^n \to \mathbb{R}$ be defined by

$$f(x) \triangleq g(h(x)) \tag{6.39}$$

Then

$$\partial f(x) \subset \overline{co}\{\sum_{i=1}^{m} \gamma^i \eta_i \mid \eta_i \in \partial h^i(x), \gamma^i \in \partial g(h(x))\} \tag{6.40}$$

■

The last result in nondifferentiable analysis that we wish to establish is the Lebourg Mean Value Theorem. However, its proof requires a knowledge of optimality conditions, which we will therefore present first.

Proposition 6.6: Consider the problem

$$\min\{f(x) \mid x \in \mathbb{R}^n\} \tag{6.41}$$

where $f:\mathbb{R}^n \to \mathbb{R}$ is l.L.c. If $\hat{x}$ solves (6.41) (global or local solution) then $0 \in \partial f(\hat{x})$.

Proof: Suppose that $\hat{x}$ solves (6.41). Then we must have for all $h \in \mathbb{R}^n$

$$d_0 f(\hat{x};h) = \max\{\langle \xi, h \rangle \mid \xi \in \partial f(\hat{x})\} \geq 0 \tag{6.42}$$

For suppose that there exists an $h \in \mathbb{R}^n$ such that $d_0 f(\hat{x};h) < 0$. It then follows from the definition of $d_0 f(\hat{x};h)$ that there exists a $\hat{t} > 0$ (sufficiently small for the local solution case) such that $f(\hat{x}+\hat{t}h) < f(\hat{x})$ which contradicts the optimality of $\hat{x}$. It now follows from Corollary 3.2 that $0 \in \partial f(\hat{x})$. Q.E.D. ■

Proposition 6.7: Consider the problem

$$\min\{f(x) \mid g^j(x) \leq 0,\ j \in \underline{m}\} \tag{6.43}$$

where $f:\mathbb{R}^n \to \mathbb{R}$, $g^j:\mathbb{R}^n \to \mathbb{R}$, $j \in \underline{m}$ are l.L.c.

If $\hat{x}$ solves (6.43), then

$$0 \in co\ \{\partial f(\hat{x}); \partial g^j(\hat{x}), j \in I^*(\hat{x})\} \tag{6.44}$$

where $I^*(\hat{x}) \triangleq \{j \in \underline{m} \mid g^j(\hat{x}) = 0\}$.

Proof: (c.f. proof of Proposition 5.2). Let $F:\mathbb{R}^n \to \mathbb{R}$ be defined by

$$F(x) \triangleq \max\{f(x) - f(\hat{x}); g^j(x), j \in \underline{m}\} \tag{6.45}$$

Clearly, $F(\cdot)$ is l.L.c. and $F(x) \geq 0$ for all $x \in I\!R^n$. Hence, since $F(\hat{x})=0$, $\hat{x}$ is a global minimizer for $F(\cdot)$ and hence we must have $0 \in \partial F(\hat{x})$. Making use of (6.31) we obtain (6.44). Q.E.D. ▪

An extension of the result in (6.31) leads to the following extension of Proposition 6.7.

Proposition 6.8: Consider the problem

$$\min \{f(x) \mid \varphi(x,y) \leq 0 \quad \forall\, y \in Y\} \tag{6.46}$$

where $f: I\!R^n \to I\!R$ and $\varphi: I\!R^n x I\!R^m \to I\!R$ are l.L.c., and Y is a compact subset of $I\!R^m$. If $\hat{x}$ solves (6.46), then

$$0 \in co\ \{\partial f(\hat{x}); \partial_x \varphi(\hat{x},y),\ y \in Y^*(\hat{x})\} \tag{6.47}$$

where $Y^*(\hat{x}) = \{y \in Y \mid \varphi(\hat{x},y)=0\}$. ▪

Finally, we present the Lebourg Mean Value Theorem.

Proposition 6.9 (Mean Value Theorem): Let $f: I\!R^n \to I\!R$ be l.L.c. Then, given any $x,y \in I\!R^n$,

$$f(y) - f(x) = \langle \xi_s, y-x \rangle \tag{6.48}$$

for some $\xi_s \in \partial f(x+s(y-x))$, $s \in (0,1)$.

Proof: Consider the function $h: I\!R \to I\!R$ defined by

$$h(t) \triangleq f(x + t(y-x)) + t[f(x) - f(y)] \tag{6.49}$$

Then $h(0)=f(x)$, $h(1)=f(x)$, so that $h(0)=h(1)$. Clearly, $h(\cdot)$ must have either a local min or a local max for some $s \in (0,1)$. Hence, for some $s \in (0,1)$,

$$0 \in \partial h(s) \in G(s) + [f(x) - f(y)] \tag{6.50}$$

where we have made use of Proposition 6.5 and the definition (6.36) for $G(s)$. But (6.50) is equivalent to (6.48) and hence our proof is complete. ▪

7. Semi-Infinite Optimization Algorithms I.

We now turn to a class of optimization problems which correspond to an important class of engineering design problems. We shall consider in detail only the simplest problem in this class since it captures all the essential features of this class. Thus, consider the problem

$$\min\{f(x) \mid \varphi(x,y) \le 0 \ \ \forall\, y \in Y\} \tag{7.1}$$

where $f:\mathbb{R}^n \to \mathbb{R}$ and $\varphi:\mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ are locally Lipschitz continuous (l.L.c.) and Y is a compact subset of $\mathbb{R}^m$ (a more general problem would have many inequality constraints). Quite often, in engineering applications, Y is an interval on the real line. If we define

$$\psi(x) \triangleq \max\{\varphi(x,y) \mid y \in Y\} \tag{7.2}$$

we can express (7.1) in the equivalent form

$$\min\{f(x) \mid \psi(x) \le 0\} \tag{7.3}$$

We recall that first order optimality conditions for the problem (7.2) were given in Proposition 6.8. In this section we turn to the development of algorithms for solving problems of the form (7.2). All the algorithms that we will present can be thought of as being evolved from the method of steepest descent for unconstrained differentiable optimization. We therefore begin by recalling this method.

Consider the problem

$$\min\{\psi(x) \mid x \in \mathbb{R}^n\} \tag{7.3}$$

where $\psi:\mathbb{R}^n \to \mathbb{R}$ is continuously differentiable.

Algorithm 7.1: (Differentiable Steepest Descent)

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

Step 0: Set $i=0$.

Step 1: Compute the *search direction*

$$h_i = h(x_i) \triangleq arg\ \min\{\tfrac{1}{2}\|h\|^2 + d\psi(x_i;h)\,|\,h\in I\!R^n\}$$
$$= -\nabla\psi(x_i) \tag{7.4}$$

Step 2: Compute the *step size*

$$\lambda_i \in \lambda(x_i) \triangleq arg\ \min_{\lambda\geq 0}\psi(x_i+\lambda h_i) \tag{7.5}$$

Step 3: Update:

$$x_{i+1} = x_i + \lambda_i h_i \tag{7.6}$$

Replace i by $i+1$ and go to Step 1. ▪

Theorem 7.1: Consider a sequence $\{x_i\}_{i=0}^{\infty}$ constructed by Algorithm 7.1. If $x_i \to^K \hat{x}$ as $i\to\infty$ ($i\in K\subset\{0,1,2,...\}$) then $\nabla\psi(\hat{x})=0$.

Proof: Suppose that $\nabla\psi(\hat{x})\neq 0$. Then

$$d\psi(\hat{x};h(\hat{x})) = -\|\nabla\psi(\hat{x})\|^2 < 0 \tag{7.7}$$

Hence any $\hat{\lambda}\in\lambda(\hat{x})$ satisfies $\hat{\lambda}>0$ and

$$\psi(\hat{x}+\hat{\lambda}h(\hat{x})) - \psi(\hat{x}) = -\hat{\delta} < 0 \tag{7.8}$$

Since $h(\cdot)$ is continuous by assumption, the function $\psi(x+\hat{\lambda}h(x))-\psi(x)$ is continuous in x and hence there exists an i_0 such that for all $i\in K$, $i\geq i_0$,

$$\psi(x_i+\lambda_i h_i) - \psi(x_i) \leq \psi(x_i+\hat{\lambda}h(x_i)) - \psi(x_i) \leq -\frac{\hat{\delta}}{2} \tag{7.9}$$

Now, by construction, $\{\psi(x_i)\}_{i=0}^{\infty}$ is monotone decreasing and $\psi(x_i)\to^K\psi(\hat{x})$ as $i\to\infty$ by continuity of $\psi(\cdot)$; we must therefore have that $\psi(x_i)\to\psi(\hat{x})$ as $i\to\infty$. But this contradicts (7.9). Hence we must have had $\nabla\psi(\hat{x})=0$. ▪

We must point out at this time that practical algorithms do not use the stepsize rule (7.5), but the much more efficient *Armijo stepsize* [Pol 1], rule which uses two parameters $\alpha, \beta \in (0,1)$ and which is defined by

$$\lambda_i \triangleq \max \{\lambda \mid \lambda = \beta^k, k \in \mathbb{N},$$
$$f(x_i + \lambda h_i) - f(x_i) \leq -\lambda\alpha \|h_i\|^2\} \tag{7.10}$$

where $\mathbb{N} = \{0,1,2,3,\ldots\}$. The geometry of this stepsize rule is given in Fig. 7.1.

The convergence analysis of the algorithms 7.1, modified to accept the Armijo stepsize rule is somewhat more complex than in Theorem 7.1 and is left as an exercise for the reader (alternatively, see [Pol 1]).

Now suppose that $\psi(\cdot)$ in (7.3) is only l.L.c. Since in this case the directional derivative $d\psi(x;h)$ need not exist, a first attempt at generalizing Algorithm 7.1 to the nondifferentiable case would consist of replacing $d\psi(x_i;h)$ in (7.4) by $d_0\psi(x_i;h)$. This amounts to computing the search direction according to the formula

$$\begin{aligned} h_i = h(x_i) &\triangleq arg \min_{h \in \mathbb{R}^n} \{½\|h\|^2 + d_0\psi(x_i;h)\} \\ &= arg \min_{h \in \mathbb{R}^n} \max_{\xi \in \partial\psi(\xi)} \{½\|h\|^2 + \langle \xi, h \rangle\} \\ &= arg \max_{\xi \in \partial\psi(\xi)} \min_{h \in \mathbb{R}^n} \{½\|h\|^2 + \langle \xi, h \rangle\} \\ &= -arg \min \{½\|h\|^2 \mid h \in \partial\psi(x)\} \end{aligned} \tag{7.11}$$

where we have interchanged the min and max operations on the authority of Corollary 3.1 and have eliminated the min on the basis that if h_ξ solves $\min\{½\|h\|^2 + \langle \xi, h \rangle \mid h \in \mathbb{R}^n\}$ then $h_\xi = -\xi$, so that $½\|h_\xi\|^2 + \langle \xi, h_\xi \rangle = -½\|h_\xi\|^2$.

Because $\partial\psi(\cdot)$ is not continuous, $h(x)$, as defined by (7.11), is not continuous. Hence it is not possible to simply mimic the proof of theorem 7.1 in trying to show that the extended algorithm is convergent in the sense that $x_i \to^K \hat{x}$ implies that $0 \in \partial\psi(\hat{x})$. In fact, there are known counter examples which show that the accumulation points $\hat{x}$ constructed by the extension of Algorithm 7.1 using (7.11) fail to satisfy $0 \in \partial\psi(\hat{x})$. Clearly, a much more sophisticated approach than using (7.11) is needed.

To try to obtain some intuitive insight into techniques for generating continuous search directions, let us examine the simple case where $\psi(x)=\max_{j\in\underline{m}} f^j(x)$, with the $f^j:\mathbb{R}^n\to\mathbb{R}$ continuously differentiable. In this case, $\partial\psi(x)=\mathrm{co}\,\{\nabla f^j(x)\}_{j\in I(x)}$, where $I(x)=\{j\in\underline{m}\,|\,\psi(x)-f^j(x)=0\}$. Since the index set $I(x)$ can change abruptly, it is clear that $\partial\psi(x)$ is not continuous. Now, if $\hat{x}$ is a minimizer of $\psi(\cdot)$ over $\mathbb{R}^n$, then we have $0\in\partial\psi(\hat{x})$, i.e., for some $\mu^j\geq 0$ such that $\sum_{j\in I(x)}\mu^j=1$, we have $\sum_{j\in I(x)}\mu^j\nabla f^j(\hat{x})=0$. A commonly used trick to avoid introducing the index set $I(x)$ into the optimality condition is to express it in the equivalent form of two equations

$$\sum_{j=1}^{m}\mu^j\nabla f^j(\hat{x}) = 0 \tag{7.12a}$$

$$\sum_{j=1}^{m}\mu^j(\psi(\hat{x}) - f^j(\hat{x})) = 0 \tag{7.12b}$$

with the $\mu^j\geq 0$ such that $\sum_{j=1}^{m}\mu^j=1$. Since $\mu^j\geq 0$ and $\psi(\hat{x})-f^j(\hat{x})\geq 0$, (7.12b) implies that $\mu^j=0$ for all $j\notin I(x)$. Now, (7.12a) and (7.12b) state that 0 is an element of the set $\overline{G}\psi(\hat{x})\subset\mathbb{R}^{n+1}$ defined by

$$\begin{aligned}\overline{G}\psi(\hat{x}) &\triangleq \mathrm{co}\ \{\overline{\xi}_j\in\mathbb{R}^{n+1}\ |\\ \overline{\xi}_j &= (\psi(\hat{x}) - f^j(\hat{x}),\nabla f^j(\hat{x})), j\in\underline{m}\}\end{aligned} \tag{7.12c}$$

where we have abused notation in denoting vectors in $\mathbb{R}^{n+1}$ as $\overline{\xi}=(\xi^0,\xi)$ with $\xi\in\mathbb{R}^n$. Rather interestingly, the set valued map $\overline{G}\psi(\cdot)$ is continuous and hence, $\overline{h}(x)=(h^0(x),h(x))$, with $h(x)\in\mathbb{R}^n$, defined by $\overline{h}(x)\triangleq \arg\min\,\{\tfrac{1}{2}\|\overline{h}\|^2\,|\,\overline{h}\in\overline{G}\psi(x)\}$, is also continuous. The *principle of wishful thinking* suggests that $h(x)$ must be a "good" continuous search direction for solving $\min\{\psi(x)\,|\,x\in\mathbb{R}^n\}$. We shall now establish an axiomatic structure for utilizing this guess. In the next section we will present a more complicated approach which leads to computationally more efficient algorithms.

Definition 7.1: Let $\psi:\mathbb{R}^n \to \mathbb{R}$ be l.L.c. We shall say that $\bar{G}\psi:\mathbb{R}^n \to 2^{\mathbb{R}^{n+1}}$ is an *augmented convergent direction finding* (a.c.d.f.) map for $\psi(\cdot)$ if:

a) $\bar{G}\psi(\cdot)$ is continuous (i.e., both u.s.c. and l.s.c.) and $\bar{G}\psi(x)$ is convex for all $x\in\mathbb{R}^n$.

b) For any $x\in\mathbb{R}^n$, if $\bar{\xi}=(\xi^0,\xi)\in\mathbb{R}^{n+1}$ is an element of $\bar{G}\psi(x)$, then $\xi^0\geq 0$.

c) For any $x\in\mathbb{R}^n$, a point $\bar{\xi}=(0,\xi)$ is an element of $\bar{G}\psi(x)$ if and only if $\xi\in\partial\psi(x)$. ▪

Proposition 7.1: Suppose that $\psi:\mathbb{R}^n\to\mathbb{R}$ is l.L.c. and $\bar{G}\psi(\cdot)$ is c.d.f. map for $\psi(\cdot)$. Then for any $x\in\mathbb{R}^n$,

a) $0\in\partial\psi(x) \Leftrightarrow 0\in\bar{G}\psi(x)$

b) The functions $\vartheta:\mathbb{R}^n\to\mathbb{R}$ and $\bar{h}:\mathbb{R}^n\to\mathbb{R}^{n+1}$ defined by

$$\vartheta(x) \triangleq \min\{\tfrac{1}{2}\|\bar{\xi}\|^2 \mid \bar{\xi}\in\bar{G}f(x)\} \tag{7.13a}$$

$$\bar{h}(x) \triangleq -\mathit{arg}\ \min\{\tfrac{1}{2}\|\bar{\xi}\|^2 \mid \bar{\xi}\in\bar{G}f(x)\} \tag{7.13b}$$

are both continuous and $\vartheta(x)=0 \Leftrightarrow 0\in\partial\psi(x)$.

c) Writing $\bar{h}(x)=(h^0(x),h(x))$, with $h(x)\in\mathbb{R}^n$,

$$d_0\psi(x;h(x)) \leq -\vartheta(x) \tag{7.13c}$$

Proof:

a) $\Rightarrow$ Suppose $0\in\partial\psi(x)$. Then, by Caratheodory's theorem (Proposition 3.1), there exist at most $(n+1)$ vectors $\xi_i\in\partial\psi(x)$, $i=1,2,\ldots,n+1$, such that for some $\mu^i\geq 0$, $\sum_{i=1}^{n+1}\mu^i=1$, $\sum_{i=1}^{n+1}\mu^i\xi_i=0$. Now, the vectors $\bar{\xi}_i=(0,\xi_i)\in\bar{G}\psi(x)$ by definition and $\sum_{i=1}^{n+1}\mu^i\bar{\xi}_i=0$ which proves that $0\in\bar{G}\psi(x)$.

$\Leftarrow$ Suppose that $0\in\bar{G}\psi(x)$. Then there exist (by Caratheodory) at most $n+2$ vectors $\bar{\xi}_i\in\bar{G}\psi(x)$ such that $\sum_{i=1}^{n+2}\mu^i\bar{\xi}_i=0$ with $\mu^i\geq 0$ and $\sum_{i=1}^{n+2}\mu^i=1$. Now, $\bar{\xi}_i=(\xi_i^0,\xi_i)$

and by b) of Definition 7.1, $\xi_i^0 \geq 0$. Since $\sum_{i=1}^{n+2} \mu^i \xi_i^0 = 0$, we must have $\mu^i \xi_i^0 = 0$ for $i=1,2,\ldots,n+2$. Hence, for all i such that $\mu^i > 0$, $\xi_i^0 = 0$ and hence $\xi_i \in \partial\psi(x)$. We conclude that $0 \in \partial\psi(x)$.

b) Since $\overline{G}f(\cdot)$ is continuous, it follows from Corollary 4.1 and Proposition 4.2 that $\vartheta(x)$ is continuous and $\overline{h}(x)$ is u.s.c. Since the solution of (7.13a) is unique, it follows that $\overline{h}(x)$ is a point-to-point map and hence continuous.

c) By definition (7.13b) we have

$$\langle -\overline{h}(x), \overline{\xi} \rangle \geq \tfrac{1}{2} \| \overline{h}(x) \|^2 = \vartheta(x) \quad \forall\, \overline{\xi} \in \overline{G}f(x) \tag{7.14}$$

Now suppose that $\overline{\xi} = (0, \xi)$, so that $\xi \in \partial\psi(x)$. Then

$$\langle -\overline{h}(x), \overline{\xi} \rangle = \langle -h(x), \xi \rangle \geq \vartheta(x) \tag{7.15}$$

consequently, we have

$$d_0\psi(x; h(x)) = \max_{\xi \in \partial\psi(x)} \langle h(x), \xi \rangle \leq -\vartheta(x) \tag{7.16}$$

which completes our proof. ▪

Exercise 7.1: Suppose that $\psi: \mathbb{R}^n \to \mathbb{R}$ is defined by

$$\psi(x) \triangleq \max_{j \in \mathbf{m}} f^j(x) \tag{7.17}$$

where $f^j: \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable functions. Let

$$\overline{G}\psi(x) \triangleq \underset{j \in \mathbf{m}}{co} \left\{ \begin{pmatrix} \psi(x) - f^j(x) \\ \nabla f^j(x) \end{pmatrix} \right\} \tag{7.18}$$

Show that this set is an augmented convergent direction finding map. ▪

Exercise 7.2: Suppose that $\psi: \mathbb{R}^n \to \mathbb{R}$ is defined by

$$\psi(x) = \max \{ \varphi(x,y) \mid y \in Y \} \tag{7.19}$$

with $\varphi: \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ continuous and $\nabla_x \varphi(x,y)$ continuous and $Y \subset \mathbb{R}^m$ compact. Let

$$\bar{G}\psi(x) \triangleq \underset{y\in Y}{co}\left\{\begin{bmatrix}\psi(x)-\varphi(x,y)\\ \nabla_x\varphi(x,y)\end{bmatrix}\right\} \tag{7.20}$$

Show that this set is an a.c.d.f. map. ▪

We shall now see that if we modify the search direction computation in (7.4) as shown below, then we can mimic the proof of Theorem 7.1.

Algorithm 7.2: (Nondifferentiable Steepest Descent) (Requires an a.c.d.f. map $\bar{G}f(\cdot)$).

Data: $x_0 \in I\!R^n$.

Step 0: Set $i=0$.

Step 1: Compute the *search direction* h_i as the last n elements of $\bar{h}(x_i)$ defined in (7.12b).

Step 2: Compute the *step length*

$$\lambda_i \in \lambda(x_i) \triangleq arg\ \min_{\lambda\geq 0}\psi(x_i+\lambda h_i) \tag{7.21}$$

Step 3: Update $x_{i+1}=x_i+\lambda_i h_i$; replace i by $i+1$ and go to step 1. ▪

Theorem 7.2: Consider a sequence $\{x_i\}_{i=0}^{\infty}$ constructed by Algorithm 7.2. If $x_i \to^K \hat{x}$ as $i\to\infty$ ($i\in K\subset\{0,1,2,...\}$) then $0\in\partial\psi(\hat{x})$.

Proof: Suppose that $x_i \to^k \hat{x}$ and $0\notin\partial\psi(\hat{x})$. Then $0\notin\bar{G}\psi(\hat{x})$ and hence

$$d_0\psi(\hat{x},h(\hat{x})) \leq -\vartheta(\hat{x}) < 0 \tag{7.22}$$

Hence, for the stepsize $\hat{\lambda}>0$ computed at $\hat{x}$, we must have that

$$\psi(\hat{x} + \hat{\lambda}h(\hat{x})) - \psi(\hat{x}) = -\hat{\delta} < 0 \tag{7.23}$$

Since $\psi(\cdot)$ and $h(\cdot)$ are both continuous (Proposition 7.1), it follows that there exists an i_0 such that for all $i\in K$, $i\geq i_0$.

$$\psi(x_{i+1}) - \psi(x_i) \leq \psi(x_i + \hat{\lambda}h_i) - \psi(x_i) \leq -\frac{\hat{\delta}}{2} \tag{7.24}$$

Now $\{\psi(x_i)\}_{i=0}^{\infty}$ is monotonically decreasing and $\psi(x_i) \to^K \psi(\hat{x})$ as $i \to \infty$. Hence $\psi(x_i) \to \psi(\hat{x})$ as $i \to \infty$. But this contradicts (7.24), and hence we must have $0 \in \partial\psi(\hat{x})$. ▪

The main objection to the use of a.c.d.f. maps is that they usually turn out to be complex, as in (7.20), so that the computation of $\bar{h}(x)$ is next to impossible. However, they have been known to be used (with a slightly modified direction computation) in optimal control. We give a relatively simple example.

Example 7.1: Suppose we are given a dynamical system

$$\dot{z}(t) = f(z(t), u(t)),\ z(0) = z_0 \tag{7.25}$$

where $f: I\!R^n \times I\!R \to I\!R^m$ is continuously differentiable, and suppose that we are required to find a control $u(t)$ such that $g(z(t)) \le 0$ for all $t \in [0,1]$, with $g: I\!R^n \to I\!R$ continuously differentiable. First, denoting the solution of (7.25) by $z^u(t)$, we define

$$\varphi(u,t) \triangleq g(z^u(t)) \tag{7.26}$$

and

$$\psi(x) \triangleq \max_{t \in [0,1]} \varphi(u,t) \tag{7.27}$$

Clearly, this is no longer a problem in $I\!R^n$. We can either assume that $u(t)$ is piecewise constant (with at most n discontinuities) , which reduces the problem to $I\!R^n$, or else assume that $u \in L_\infty[0,1]$, which leads us to produce a formal (but justifiable) extension of our results to an infinite dimensional space. We elect to do the latter.

First, we define $\bar{G}\psi(u)$ as in (7.20). To obtain an expression for $\langle \nabla_u \varphi(u,t), \delta u \rangle_2$, where $\langle \cdot,\cdot \rangle_2$ is the L_2 scalar product, we note, formally, that to first order (in $L_\infty[0,1]$)

$$\varphi(u+\delta u,t) - \varphi(u) = g(z^{u+\delta u}(t)) - g(z^u(t))$$
$$\approx \frac{\partial g}{\partial z}(z^u(t))\delta z(t)$$
$$= \langle \nabla_u \varphi(u,t), \delta u \rangle_2 \tag{7.28}$$

where

$$\delta \dot{z}(t) = \frac{\partial f}{\partial z}(z^u(t),u(t))\delta z(t) + \frac{\partial f}{\partial u}(z^u(t),u(t))\delta u(t)$$
$$\delta z(0) = 0 \tag{7.29}$$

Next, replacing $\partial\psi(x)$ by $\overline{G}\psi(x)$ in (7.11), we find that $\bar{h}(u)$, defined in (7.13b), is computable by solving (via an extension of Corollary 3.1)

$$\bar{h}(u) = arg \max_{\xi \in \overline{G}\psi(u)} \min_{\delta\overline{u} \in L_\infty^2[0,1]} \{ \tfrac{1}{2} \|\delta\overline{u}\|_2^2 + \langle \xi, \delta\overline{u} \rangle_2 \} \tag{7.30}$$

where $\|\cdot\|_2$ is the L_2 norm and $\delta\overline{u}=(\delta u^0,\delta u)$ assumes values in $\mathbb{R}^2$; $\delta u^0(\cdot)$ is an artificial control variable. Next let P denote the set of all probability measures on $[0,1]$, i.e., $\mu \in P$ is an integrable function such that $\mu(t) \geq 0$ for all $t \in [0,1]$ and $\int_0^1 \mu(t)dt = 1$. Then, a vector $\xi \in \overline{G}\psi(u)$, which is a convex combination of vectors of the form $(\psi(u)-\varphi(u,t), \nabla_u\varphi(u,t))$, has the form $(\int_0^1 \mu(t)[\psi(u)-\varphi(u,t)]dt,$ $\int_0^1 \mu(t)\nabla_u\varphi(u,t)dt)$. Hence (7.30) becomes

$$\bar{h}(u) = arg \max_{\mu \in P} \min_{\delta\overline{u} \in L_\infty^2[0,1]} \{ \tfrac{1}{2}\int_0^1 [\delta u^0(t)^2 + \delta u(t)^2]dt$$
$$+ \int_0^1 \mu(t)[\psi(u) - g(z^u(t))]dt + \int_0^1 \mu(t)\frac{\partial g}{\partial z}(z^u(t))\delta z(t)dt \mid \delta\dot{z}(t) \tag{7.31}$$

We see that the inner (min) problem is a simple linear quadratic optimal control problem solvable by the Pontryagin Maximum Principle [Pon 1]. The outer (max) problem is solvable by means of dual algorithms. Thus we see that the algorithm that we have described is extendable to optimal control as well. ▪

Finally we turn to the problem (7.3). We assume that an algorithm of the form of Algorithm 7.2 has been used to find a point x_0 such that $\psi(x_0)\leq 0$. Note that if $\min_{x\in\mathbb{R}^n}\psi(x)<0$, then such an x_0 is obtained in a finite number of iterations. We need to postulate a continuous set valued map $\overline{G}_{f,\psi}(x)$ such that $0\in\overline{G}_{f,\psi}(x)$ holds if and only if the optimality condition (6.44) holds. We proceed by extension from the unconstrained case.

Definition 7.2: Let $f:\mathbb{R}^n\to\mathbb{R}$, $\psi:\mathbb{R}^n\to\mathbb{R}$ be l.L.c and let $F\triangleq\{x\in\mathbb{R}^n\,|\,\psi(x)\leq 0\}$. We shall say that $\overline{G}_{f,\psi}:\mathbb{R}^n\to 2^{\mathbb{R}^{n+1}}$ is an *augmented convergent direction finding map* for (7.3) if

a) $\overline{G}_{f,\psi}(\cdot)$ is continuous and $\overline{G}_{f,\psi}(x)$ is convex for all $x\in\mathbb{R}^n$.

b) For any $x\in F$, if $\overline{\xi}=(\xi^0,\xi)\in\mathbb{R}^{n+1}$ is an element of $\overline{G}_{f,\psi}(x)$, then $\xi^0\geq 0$.

c) For any $x\in F$, a point $\overline{\xi}=(0,\xi)$ is an element of $\overline{G}_{f,\psi}(x)$ if and only if either $\psi\in\partial f(x)$ or $\psi\in\mathrm{co}\{\partial f(x),\,\partial\psi(x)\}$ and $\psi(x)=0$.

d) For any $x\in F$, such that $\psi(x)<0$, a point $\overline{\xi}=(-\psi(x),\xi)$ is an element of $\overline{G}_{f,\psi}(x)$ for all $\xi\in\partial\psi(x)$. ▪

Proposition 7.2: Suppose that $f,\psi:\mathbb{R}^n\to\mathbb{R}$ are l.L.c. functions and that $\overline{G}_{f,\psi}(\cdot)$ is an a.c.d.f. map for (7.3). Then for any $x\in\mathbb{R}^n$ such that $\psi(x)\leq 0$,

a) (i) if $\psi(x)<0$, $0\in\partial f(x)\iff 0\in\overline{G}_{f,\psi}(x)$.

(ii) if $\psi(x)=0$, $0\in co\,\{\partial f(x),\partial\psi(x)\}\iff 0\in\overline{G}_{f,\psi}(x)$.

b) The functions $\vartheta:\mathbb{R}^n\to\mathbb{R}$ and $\overline{h}:\mathbb{R}^n\to\mathbb{R}^{n+1}$ defined by

$$\vartheta(x)\triangleq\min\{\tfrac{1}{2}\|\overline{\xi}\|\,|\,\overline{\xi}\in\overline{G}_{f,\psi}(x)\}\tag{7.32a}$$

$$\overline{h}(x)\triangleq -arg\ \min\{\tfrac{1}{2}\|\overline{\xi}\|\,|\,\overline{\xi}\in\overline{G}_{f,\psi}(x)\}\tag{7.32b}$$

are both continuous.

c) Writing $\overline{h}(x)=(h^0(x),h(x))$, with $h(x)\in\mathbb{R}^n$, we have

$$-h^0(x)\psi(x) + d_0\psi(x;h(x)) \le -\vartheta(x) \tag{7.33a}$$

$$d_0 f(x;h(x)) \le -\vartheta(x) \tag{7.33b}$$

Proof:

a) $\Rightarrow$ if $0 \in \partial f(x)$, then $0 \in \overline{G}_{f,\psi}(x)$ because of c) in definition 7.2.

Now suppose that $0 \in \overline{G}_{f,\psi}(x)$. Then, because of c) in definition 7.2, we must have $0 \in \partial f(x)$.

b) The continuity of $\vartheta(\cdot)$ and $\overline{h}(\cdot)$ follows from Corollary 4.1 and the fact that the argmin in (7.32b) is a singleton.

c) By definition (7.32b), $\overline{h}(x)$ satisfies

$$\langle -\overline{h}(x), \overline{\xi} \rangle \ge \tfrac{1}{2}\|\overline{h}(x)\|^2 = \vartheta(x) \quad \forall\, \overline{\xi} \in \overline{G}_{f,\psi}(x) \tag{7.34}$$

Now, let $\overline{\xi} = (0,\xi) \in \overline{G}_{f,\psi}(x)$ be such that $\xi \in \partial f(x)$. Then we get

$$h^0(x)0 + \langle h(x), \xi \rangle \le -\vartheta(x) \tag{7.35}$$

Maximizing (7.35) over $\xi \in \partial f(x)$ we obtain (7.33b).

Next, suppose that $\xi \in \partial \psi(x)$. Then $(-\psi(x), \xi) \in \overline{G}_{f,\psi}(x)$ and hence (7.34) yields

$$-h^0(x)\psi(x) + \langle h(x), \xi \rangle \le -\vartheta(x) \tag{7.36}$$

Maximizing (7.36) over $\xi \in \partial\psi(x)$, we obtain (7.33a). Q.E.D. ▪

Exercise 7.3: Suppose that $f: \mathbb{R}^n \to \mathbb{R}$ and $\varphi: \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ are differentiable functions and that $\psi(x) \triangleq \max\{\varphi(x,y) \mid y \in Y\}$ with $Y \subset \mathbb{R}^m$ compact. Show that the map

$$\overline{G}_{f,\psi}(x) \triangleq \text{co} \left\{ \begin{bmatrix} 0 \\ \nabla f(x) \end{bmatrix}; \begin{bmatrix} -\varphi(x,y) \\ \nabla_x \varphi(x,y) \end{bmatrix}, y \in Y \right\} \tag{7.37}$$

satisfies the assumptions of definition 7.2. ▪

We conclude this section by stating an algorithm model for solving (7.3) and giving a proof of its convergence.

Algorithm 7.3: (Constrained Nondifferentiable Optimization) (Requires an a.c.d.f.

map $\overline{G}_{f,\psi}(\cdot)$).

Data: $x_0 \in I\!R^n$ such that $\psi(x_0) \le 0$.

Step 0: Set $i=0$.

Step 1: Compute the search direction h_i as the last n elements of $\bar{h}(x_i)$, defined in (7.32b).

Step 2: Compute the step size

$$\lambda_i \in \lambda(x_i) \triangleq arg \min_{\lambda \ge 0} \{ f(x_i + \lambda h_i) \mid \psi(x_i + \lambda h_i) \le 0 \} \tag{7.38}$$

Step 3: Update: $x_{i+1} = x_i + \lambda_i h_i$, replace i by $i+1$ and go to step 1. ▪

Theorem 7.3: Suppose that $f, \psi : I\!R^n \to I\!R$ are l.L.c., that $\overline{G}_{f,\psi}(\cdot)$ is an a.c.d.f. map for (7.3). If $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm 7.3 and $x_i \stackrel{K}{\to} \hat{x}$ as $i \to \infty (K \subset \{0,1,2,...\})$ then $\psi(\hat{x}) \le 0$ and $0 \in \overline{G}_{f,\psi}(\hat{x})$. (i.e., $\hat{x}$ satisfies the first order condition of optimality).

Proof: To obtain a contradiction, suppose that $0 \notin \overline{G}_{f,\psi}(\hat{x})$. We consider two cases. (Clearly, since $\psi(x_i) \le 0$ for all i, we must have $\psi(\hat{x}) \le 0$).

a) $\psi(\hat{x}) < 0$. Then, since $\vartheta(\hat{x}) > 0$, (see 7.32a), we have from (7.33b) that

$$d_0 f(\hat{x}; h(\hat{x})) ,= -\vartheta(\hat{x}) < 0 \tag{7.39}$$

Consequently, since $f(\cdot)$, $\psi(\cdot)$ and $h(\cdot)$ are continuous, there exist a $\hat{\rho} > 0$, a $\hat{\lambda} > 0$, and a $\hat{\delta} > 0$ such that

$$f(x + \hat{\lambda} h(x)) - f(x) \le -\hat{\delta} \tag{7.40a}$$

$$\psi(x + \hat{\lambda} h(x)) \le 0 \tag{7.40b}$$

for all $x \in B(\hat{x}, \hat{\rho})$. Hence, since $x_i \to^K \hat{x}$ as $i \to \infty$, there exists an i_0 such that

$$f(x_{i+1}) - f(x_i) \le -\hat{\delta} \quad \forall\, i \ge i_0,\ i \in K \tag{7.41}$$

Now $\{f(x_i)\}_{i=0}^{\infty}$ is monotone decreasing and $f(x_i) \to^K f(\hat{x})$ because $f(\cdot)$ is continuous, hence $f(x_i) \to f(\hat{x})$. But this contradicts (7.42) and hence we must have $0 \in \overline{G}_{f,\psi}(\hat{x})$.

b) $\psi(\hat{x})=0$. In this case, since $0 \notin \overline{G}_{f,\psi}(\hat{x})$, it follows from (7.33a) that

$$d_0\psi(\hat{x};h(\hat{x})) \le -\vartheta(\hat{x}) < 0 \tag{7.41}$$

holds in addition to (7.39). It now follows from the continuity of $f(\cdot)$, $\psi(\cdot)$ and $h(\cdot)$ that for some $\hat{\rho}>0$, $\hat{\lambda}>0$, $\hat{\delta}>0$ (7.40a), (7.40b) hold for all $x \in B(\hat{x},\hat{\rho})$. Hence, we obtain a contradiction as for case a). ▪

This concludes our exposition of a first approach to the construction of semi-infinite optimization algorithms. While the approach is simple, it results in unacceptably difficult search direction finding problems. Our second approach will therefore be to reduce this computational difficulty at the expense of an increase in algorithmic complexity.

8. Semi-Infinite Optimization Algorithms II.

We continue with the problem

$$\min \{ f(x) \mid \psi(x) \le 0 \} \tag{8.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $\psi : \mathbb{R}^n \to \mathbb{R}$ are both l.L.c. In particular, we are interested in the case where

$$\psi(x) \triangleq \max_{y \in Y} \varphi(x, y) \tag{8.2}$$

where $\varphi : \mathbb{R}^n \mathrm{x} \mathbb{R}^m \to \mathbb{R}$ is l.L.c. and Y is a compact subset of $\mathbb{R}^m$.

As in the preceding section, we begin by first considering the problem of finding a ***feasible point***, i.e., finding an $x \in \mathbb{R}^n$ such that $\psi(x) \le 0$ by solving the problem

$$\min \{ \psi(x) \mid x \in \mathbb{R}^n \} \tag{8.3}$$

In Section 7 we found that if we used $\partial\psi(x)$ to compute search directions, we could not prove convergence. On the other hand, when we embedded $\partial\psi(x)$ in an augmented convergent direction finding (a.c.d.f.) map $\overline{G}\psi(x)$, we could prove convergence. However for $\psi(\cdot)$ as in (8.2), the computation of the search direction involved the solution of an infinite quadratic programming problem which, at best, is extremely hard to carry out. To develop our intuition, we turn again to the simple case where

$$\psi(x) = \max_{j \in \mathbf{m}} f^j(x)$$

with $f^j : \mathbb{R}^n \to \mathbb{R}$ continuously differentiable. For this case we have

$$\partial\psi(x) = co \, \{ \nabla f^j(x) \}_{j \in I(x)} \tag{8.4}$$

where $I(x) \triangleq \{ j \in \underline{m} \mid f^j(x) = \psi(x) \}$ and we can define an augmented convergent direction finding map by (see Exercise 7.1)

$$\overline{G}\psi(x) \triangleq \underset{j \in \mathbf{m}}{co} \left\{ \begin{pmatrix} \psi(x) - f^j(x) \\ \nabla f^j(x) \end{pmatrix} \right\} \tag{8.5}$$

Our first attempt at reducing dimensionality of the search direction finding problem may consist in discarding from (8.5) all the vectors such that $\psi(x)-f^j(x)>\varepsilon>0$ for some ε and somehow adjust $\varepsilon>0$ as the computation proceeds. This yields a candidate map

$$\overline{G}_\varepsilon\psi(x) \triangleq \underset{j\in I_\varepsilon(x)}{co}\left\{\begin{pmatrix}\psi(x)-f^j(x)\\ \nabla f^j(x)\end{pmatrix}\right\} \tag{8.6}$$

where

$$I_\varepsilon(x) = \{j\in \underline{m} \mid \psi(x) - f^j(x) \le \varepsilon\} \tag{8.7}$$

Indeed, this is a perfectly good starting point for a convergent direction finding map. However, our knowledge of methods of feasible directions (see [Pol 1], [Pol 5], [Gon 2]) leads us to guess that once ε is introduced, we no longer need to retain the values of $\psi(x)$ and $f^j(x)$ in $\overline{G}_\varepsilon\psi(x)$, i.e., that a convergent direction finding map with set values in $\mathbb{R}^n$ rather than in $\mathbb{R}^{n+1}$ can be used. The obvious candidate for such a c.d.f. map is

$$G_\varepsilon\psi(x) \triangleq co\ \{\nabla f^j(x)\}_{j\in I_\varepsilon(x)}\,,\ \varepsilon\ge 0 \tag{8.8}$$

We proceed again on the basis of the principle of wishful thinking, which leads us to believe that we must be right and construct an axiomatic structure which abstracts the properties of the set in (8.8).

Definition 8.1: We shall say that $\{G_\varepsilon\psi(\cdot)\}_{\varepsilon\ge 0}$ where $G_\varepsilon\psi:\mathbb{R}^n\to 2^{\mathbb{R}^n}$, is a family of *convergent direction finding* (c.d.f.) *maps* for the l.L.c. function $\psi:\mathbb{R}^n\to\mathbb{R}$ if

a) For all $x\in\mathbb{R}^n$, $\partial\psi(x) = G_0\psi(x)$.

b) For all $x\in\mathbb{R}^n$, $0\le\varepsilon<\varepsilon' \Longrightarrow G_\varepsilon\psi(x)\subset G_{\varepsilon'}\psi(x)$.

c) For any $\varepsilon\ge 0$ and $x\in\mathbb{R}^n$, $G_\varepsilon\psi(x)$ is convex.

d) For any $\varepsilon\ge 0$, $G_\varepsilon\psi(x)$ is bounded on bounded sets.

e) $G_\varepsilon\psi(x)$ is u.s.c. in (ε,x) at $(0,\hat{x})$ for all $\hat{x}\in\mathbb{R}^n$.

f) 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 $\hat{\xi}\in\partial\psi(\hat{x})$ and any $x\in B(\hat{x},\hat{\rho})$, there exists an $\xi\in\widehat{G_{\hat{\varepsilon}}}\psi(x)$ such that $\|\xi-\hat{\xi}\|\le\hat{\delta}$. ▪

We note that the property f) above is that $G_\varepsilon\psi(\cdot)$ is "almost" l.s.c., which is quite close to continuity.

Before we proceed, we shall prove that $G_\varepsilon\psi(\cdot)$ as defined in (8.8) indeed has the properties specified in Definition 8.1.

Proposition 8.1: Suppose that $f^j:\mathbb{R}^n\to\mathbb{R}$, $j\in\underline{m}$ are continuous differentiable functions. Then the family of maps $\{G_\varepsilon\psi(\cdot)\}_{\varepsilon\ge0}$ defined by (8.8) is a family of c.d.f. maps.

Proof:

a) Clearly $G_0\psi(x)=\partial\psi(x)$ for all $x\in\mathbb{R}^n$.

b) Since $0\le\varepsilon<\varepsilon'\Rightarrow I_\varepsilon(x)\subset I_{\varepsilon'}(x)$ we must have $G_\varepsilon\psi(x)\subset G_{\varepsilon'}\psi(x)$.

c) $G_\varepsilon\psi(x)$ is convex by definition.

d) For any $\varepsilon\ge0$, $G_\varepsilon\psi(x)\subset co\{\nabla f^j(x)\}_{j\in\underline{m}}$. Since the $\nabla f^j(\cdot)$ are all continuous, it follows that $G_\varepsilon\psi(x)\subset co\{\nabla f^j(x)\}_{j\in\underline{m}}$ is bounded on bounded sets for any $\varepsilon\ge0$.

e) Consider the point $(0,\hat{x})$. If $j\in\underline{m}$ is such that $j\notin I_0(\hat{x})$, then $\psi(\hat{x})-f^j(\hat{x})>0$. Hence there exists a $\hat{\rho}>0$ and an $\hat{\varepsilon}>0$ such that $j\notin I_\varepsilon(x)$ for all $x\in B(\hat{x},\hat{\rho})$, $\varepsilon\in[0,\hat{\varepsilon}]$; i.e., for all $x\in B(\hat{x},\hat{\rho})$ and $\varepsilon\in[0,\hat{\varepsilon}]$, $I_0(\hat{x})\supset I_\varepsilon(x)$. Hence, if $\varepsilon_i\to0$ and $x_i\to\hat{x}$ as $i\to\infty$ are arbitrary sequences then, since the $\nabla f^j(x)$ are continuous, we must have $\overline{Lim}G_{\varepsilon_i}\psi(x_i)\subset G_0\psi(x_i)$, i.e., $G_\varepsilon\psi(x)$ is u.s.c. at $(0,\hat{x})$.

f) Let $\hat{x}$, $\hat{\varepsilon}>0$ and $\hat{\delta}>0$ be given. First, since $j\in I_0(\hat{x})\Rightarrow\psi(\hat{x})-f^j(\hat{x})=0$, there exists a $\rho_1>0$ such that $\psi(x)-f^j(\hat{x})\le\varepsilon$ for all $x\in B(\hat{x},\rho_1)$ and $j\in I_0(\hat{x})$, i.e., $I_0(\hat{x})\subset I_{\hat{\varepsilon}}(x)$ for all $x\in B(\hat{x},\rho_1)$. Next, there exitst a $\hat{\rho}\in(0,\rho_1]$ such that $\|\nabla f^j(x)-\nabla f^j(\hat{x})\|\le\hat{\delta}$ for all $x\in B(\hat{x},\hat{\rho})$ and $j\in\underline{m}$. Hence, if $\hat{\xi}=\sum_{j\in I_0(\hat{x})}\hat{\mu}^j\nabla f^j(\hat{x})$ with

$\hat{\mu}^j \geq 0$ and $\sum_{j \in I_0(\hat{x})} \hat{\mu}^j = 1$ is any point in $\partial\psi(\hat{x})$, then for any $x \in B(\hat{x},\hat{\rho})$ there exists a $\xi = \sum_{j \in I_0(\hat{x})} \hat{\mu}^j \nabla f^j(x)$ in $G_{\hat{\varepsilon}}\psi(x)$ such that

$$\|\xi - \hat{\xi}\| = \| \sum_{j \in I_0(\hat{x})} \hat{\mu}^j (\nabla f^j(x) - \nabla f^j(\hat{x}))\| \leq \hat{\delta}$$

■

Exercise 8.1: Prove the following.

Lemma 8.1: Suppose that $\hat{G_\varepsilon}\psi(\cdot)$, $\hat{\varepsilon}>0$, is an element of a family of c.d.f. maps, defined as above. Then for any $\hat{x} \in \mathbb{R}^n$, $\hat{\delta}>0$, there exists a $\hat{\rho}>0$ such that for any $x', x'' \in B(\hat{x},\hat{\rho})$ and any $\xi' \in \partial\psi(x')$ there exists an $\xi'' \in \hat{G_\varepsilon}\psi(x'')$ such that $\|\xi'' - \xi'\| \leq \hat{\delta}$.

[*Hint:* use f) in Definition 8.1 and the fact that $\partial\psi(\cdot)$ is u.s.c.]. ■

The purpose for the augmentation of $\partial\psi(x)$ in the construction of $G_\varepsilon\psi(x)$ is to provide us with a "look ahead" property which should enable us to detect "corners" in the equal cost contours of $\psi(\cdot)$. Hence, the most naive such augmentation is to let

$$G_\varepsilon\psi(x) \triangleq co\ \{\partial\psi(x')\}_{x' \in B(x,\varepsilon)} \tag{8.9}$$

Obviously, this definition is quite unattractive for the case of max functions. However, it may have merrit in the case of less structured semi-smooth functions, see, e.g., [Pol 7].

Exercise 8.2: Consider $\psi(\cdot)$ defined by (8.2) and suppose that $\nabla_x\varphi(x,y)$ and $\nabla_y\varphi(x,y)$ exist and are continuous. Let

$$G_\varepsilon\psi(x) \triangleq co\ \{\nabla_x\varphi(x,y)\}_{y \in \hat{Y}_\varepsilon(x)},\ \varepsilon \geq 0 \tag{8.10a}$$

where

$$\begin{aligned} \hat{Y}_\varepsilon(x) \triangleq \{y \in Y \mid \psi(x) - \varphi(x,y) \leq \varepsilon \\ \text{and } y \text{ is a local maximizer of } \varphi(x,\cdot) \text{ in } Y\} \end{aligned} \tag{8.10b}$$

Suppose that $\hat{Y}_0(x)$ is a finite set for all $x \in \mathbb{R}^n$. Show that (8.10a) defines a family of c.d.f. maps for $\psi(\cdot)$ (Hint: see [Pol 3], [Gon 2]). ▪

Exercise 8.3: Let $Q(x)$ be an $n \times n$ symmetric, positive semi-definite, complex valued matrix whose elements are continuously differentiable in x and let its eigenvalues be denoted by $\lambda^1(x) \geq \lambda^2(x) \geq \cdots \geq \lambda^n(x)$. Define $\psi(x) \triangleq \lambda^1(x)$. For any $\varepsilon \geq 0$, let $U_\varepsilon(x)$ be a matrix of ordered ortho-normal eigenvectors corresponding to the eigenvalues $\lambda^j(x)$ of $Q(x)$ such that $\psi(x) - \lambda^j(x) \leq \varepsilon$. Show that

$$G_\varepsilon \psi(x) \triangleq co \ \{\xi \in \mathbb{R}^n \mid \xi^i = \langle U_\varepsilon(x) z ; \frac{\partial Q(x)}{\partial x^i} U_\varepsilon(x) z \rangle \quad i = 1,2,\ldots,n \ , \ \|z\| = 1\} \tag{8.11}$$

defines a family of c.d.f. maps for $\psi(\cdot)$ Hint: see [Pol 3], [Pol 6]). ▪

Finally, we need to define a feedback law for decreasing ε in a family of c.d.f. maps.

Definition 8.2: Let $\{G_\varepsilon \psi(x)\}_{\varepsilon \geq 0}$ be a family of c.d.f. maps. Let $\alpha \in (0,1)$. We define the ε *-search direction* at $x \in \mathbb{R}^n$ by

$$h_\varepsilon(x) \triangleq -arg \ \min \{ \tfrac{1}{2} \|h\|^2 \mid h \in G_\varepsilon \psi(x) \} \tag{8.12}$$

and the ε adjustment law by

$$\varepsilon(x) \triangleq \max \{ \varepsilon \in E \mid \|h_\varepsilon(x)\|^2 \geq \varepsilon \} \tag{8.13}$$

where

$$E \triangleq \{0, 1, \alpha, \alpha^2, \alpha^3, \cdots \} \tag{8.14}$$

▪

Before we continue, it is worthwhile to pause and re-examine the ε -search directions $h_\varepsilon(x)$ defined by (8.12). We can define an *ε-generalized directional derivative* by

$$d_\varepsilon\psi(x;h) \triangleq \max_{\xi\in G_\varepsilon\psi(x)}\langle\xi,h\rangle \tag{8.14a}$$

Then we find that, because $\partial\psi(x)\subset G_\varepsilon\psi(x)$,

$$d_0\psi(x,h) \leq d_\varepsilon\psi(x;h) \tag{8.14b}$$

so that any h which makes $d_\varepsilon\psi(x;h)$ negative is a descent direction for $\psi(\cdot)$. Also, it is easy to see that

$$h_\varepsilon(x) = arg\ \min_{h\in\mathbb{R}^n}\{½\|h\|^2 + d_\varepsilon\psi(x;h)\} \tag{8.14c}$$

which, by comparison with (7.11) shows that we are fairly close to the most naive extension of the method of steepest descent to the nondifferentiable case, except that, now, hopefully, we have generated some near continuity properties.

Remark 8.1: In practice, it is common to add a second parameter $\delta>0$ to the definition of $\varepsilon(x)$, using the test $\|h_\varepsilon(x)\|^2\geq\delta\varepsilon$, which enables us to "balance" the computation better. ▪

We are ready to state an algorithm model and establish its convergence.

Algorithm 8.1: [Requires a family of c.d.f. maps $\{G_\varepsilon\psi(\cdot)\}_{\varepsilon\geq 0}$ and $\alpha\in(0,1)$ for (8.13)]

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

Step 1: Compute $\varepsilon(x_i)$ according to (8.13) and the *search direction* $h_i=h_{\varepsilon(x_i)}(x_i)$ according to (8.12).

Step 2: Compute the *stepsize*

$$\lambda_i \in \lambda(x_i) \triangleq arg\ \min_{\lambda\geq 0}\psi(x_i + \lambda h_i) \tag{8.15}$$

Step 3: Update: $x_{i+1}=x_i+\lambda_i h_i$; replace i by $i+1$ and go to Step 1. ▪

Lemma 8.2: For every $\hat{x}\in\mathbb{R}^n$ such that $0\not\in\partial\psi(\hat{x})$, there exist a $\hat{\rho}>0$ and $\hat{\varepsilon}\in E$, $\hat{\varepsilon}>0$ such that $\varepsilon(x)\geq\hat{\varepsilon}$ for all $x\in B(\hat{x},\hat{\rho})$.

Proof: Since $G_\varepsilon\psi(x)$ is u.s.c. in (ε,x) at $(0,\hat{x})$, it follows that $\|h_\varepsilon(x)\|^2$ is l.s.c. in (ε,x) at $(0,\hat{x})$. Since $\|h_0(\hat{x})\|>0$, it now follows from the l.s.c. of $\|h_\varepsilon(x)\|^2$ at $(0,\hat{x})$ that there exist a $\hat{\rho}>0$ and an $\hat{\varepsilon}>0$, $\hat{\varepsilon}\in E$, such that $\|h_\varepsilon(x)\|^2-\alpha\varepsilon\geq 0$ for all $x\in B(\hat{x},\hat{\rho})$ and $0\leq\varepsilon\leq\hat{\varepsilon}$. Hence $\|h_{\hat{\varepsilon}}(x)\|^2-\hat{\varepsilon}\geq 0$ for all $x\in B(\hat{x},\hat{\rho})$. But this implies that $\varepsilon(x)\geq\hat{\varepsilon}$ for all $x\in B(\hat{x},\hat{\rho})$. Q.E.D. ▪

Theorem 8.1: Suppose $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by algorithm 8.1 in minimizing a l.L.c. function $\psi:\mathbb{R}^n\to\mathbb{R}$. If $x_i\xrightarrow{K}\hat{x}$ as $i\to\infty$ $(K\subset\{0,1,2,\ldots,\})$, then $0\in\partial\psi(\hat{x})$.

Proof: Suppose that $0\notin\partial\psi(\hat{x})$ for the sake of obtaining a contradiction. Then $\varepsilon(\hat{x})>0$ and by Lemma 8.2, there exist i_0 and $\hat{\varepsilon}>0$ such that $\varepsilon(x_i)\geq\hat{\varepsilon}>0$ for all $i\geq i_0$, $i\in K$. Now, by the mean value theorem of Lebourg (Proposition 6.9),

$$\psi(x_i+\lambda h_i)-\psi(x_i)=\lambda\langle h_i,\xi_{i\lambda}\rangle \tag{8.16}$$

where $\xi_{i\lambda}\in\partial\psi(x_i+s\lambda h_i)$, for some $s\in(0,1)$. Now, since $G_\varepsilon\psi(x)$ is bounded on bounded sets and $G_\varepsilon\psi(x)\subset G_1\psi(x)$, we have that there exists a $b<\infty$ such that for all $i\in K$, $i\geq i_0$ $0<\hat{\varepsilon}\leq\varepsilon(x_i)\leq\|h(x_i)\|^2\leq b^2$. Referring to Lemma 8.1, let $\hat{\rho}>0$ be such that for any $x',x''\in B(\hat{x},\hat{\rho})$, and any $\xi'\in\partial\psi(x')$ there exists a $\xi''\in G_{\hat{\varepsilon}}(x'')$ such that $b\|\xi'-\xi''\|\leq\hat{\varepsilon}/2$. Hence there exists an $i_1\geq i_0$ and a $\hat{\lambda}>0$ such that for all $i\geq i_1$, $i\in K$, both $x_i\in B(\hat{x},\hat{\rho})$, and $(x_i+s\hat{\lambda}h_i)\in B(\hat{x},\hat{\rho})$ for all $s\in(0,1)$ and for $\xi_{i\hat{\lambda}}\in\partial\psi(x_i+s\hat{\lambda}h_i)$ there exists a $\xi_{i\hat{\lambda}}'\in G_{\hat{\varepsilon}}\psi(x_i)\subset G_{\varepsilon(x_i)}$ such that $\|\xi_{i\hat{\lambda}}-\xi_{i\hat{\lambda}}'\|b\leq\hat{\varepsilon}/2$. Substituting in (8.16) we now obtain, for all $i\geq i_1$, $i\in K$

$$\begin{aligned}
&\psi(x_i+\lambda_i h_i)-\psi(x_i)\leq\\
&\psi(x_i+\hat{\lambda}h_i)-\psi(x_i)\\
&=\hat{\lambda}[\langle h_i,\xi_{i\hat{\lambda}}'\rangle+\langle h_i,\xi_{i\hat{\lambda}}-\xi_{i\hat{\lambda}}'\rangle]\\
&\leq\hat{\lambda}[-\|h_i\|^2+\langle h_i,\xi_{i\hat{\lambda}}-\xi_{i\hat{\lambda}}'\rangle\|]\\
&\leq\hat{\lambda}[-\hat{\varepsilon}(x_i)+\|h_i\|\,\|\xi_{i\hat{\lambda}}-\xi_{i\hat{\lambda}}'\|]\\
&\leq-\hat{\lambda}\hat{\varepsilon}/2<0
\end{aligned} \tag{8.17}$$

Now, $\{\psi(x_i)\}_{i=0}^{\infty}$ is monotone decreasing and $\psi(x_i)\to^K\psi(\hat{x})$ since $\psi(\cdot)$ is continuous. But this implies that $\psi(x_i)\to\psi(\hat{x})$ as $i\to\infty$, contradicting (8.17). Hence we must have had $0\in\partial\psi(\hat{x})$. Q.E.D. ▪

Next, we develop an algorithm model for solving (8.1) under the assumption that we have an $x_0\in I\!R^n$ such that $\psi(x_0)\leq 0$. As we have indicated earlier, whenever $\min\{\psi(x)\mid x\in I\!R^n\}<0$ such an x_0 can be computed by means of a finite number of iterations of Algorithm 8.1.

Definition 8.3: Let $\{G_\varepsilon f(\cdot)\}_{\varepsilon\geq 0}$ and $\{G_\varepsilon\psi(\cdot)\}_{\varepsilon\geq 0}$ be given families of c.d.f. maps for the l.L.c. functions $f(\cdot)$ and $\psi(\cdot)$ in (8.1). We define the family of (phase II) c.d.f. maps $\{G^\varepsilon_{f,\psi}(\cdot)\}_{\varepsilon\geq 0}$ for (8.1) by setting

$$G^\varepsilon_{f,\psi}(x) \triangleq G_\varepsilon f(x) \text{ if } \psi(x) < -\varepsilon \tag{8.18a}$$

$$G^\varepsilon_{f,\psi}(x) \triangleq co\ \{G_\varepsilon f(x),\ G_\varepsilon\psi(x)\} \text{ if } \psi(x) \geq -\varepsilon \tag{8.18b}$$

Next, we define

$$h_\varepsilon(x) \triangleq arg\ \min\ \{½\|h\|^2 \mid h\in G^\varepsilon_{f,\psi}(x)\} \tag{8.19}$$

and

$$\varepsilon(x) \triangleq \max\ \{\varepsilon\in E \mid \|h_\varepsilon(2)\|^2 \geq \varepsilon\} \tag{8.20}$$

where E was defined in (8.14). ▪

Algorithm 8.2: [Requires $\{G_\varepsilon f(\cdot)\}_{\varepsilon\geq 0}$, $\{G_\varepsilon\psi(\cdot)\}_{\varepsilon\geq 0}$ families of c.d.f. maps for $f(\cdot)$ and $\psi(\cdot)$; $\alpha\in(0,1)$ for the set ε in (8.14)].

Data: $x_0\in I\!R^n$ such that $\psi(x_0)\leq 0$.

Step 0: Set $i=0$.

Step 1: Compute $\varepsilon(x_i)$ according to (8.20) and the *search direction* $h_i=-h_{\varepsilon(x_i)}(x_i)$, according to (8.19).

Step 2: Compute the *stepsize*

$$\lambda_i \in \lambda(x_i) \triangleq arg \min_{\lambda \geq 0} \{ f(x_i + \lambda h_i) \mid \psi(x_i + \lambda_i h_i) \leq 0 \} \tag{8.21}$$

Step 3: Update: $x_{i+1} = x_i + \lambda_i h_i$, replace i by $i+1$ and go to Step 1. ▪

Exercise 8.4: Prove the following.

Lemma 8.3:

a) For every $\hat{x} \in \mathbb{R}^n$ such that $0 \notin G^0_{f,\psi}(\hat{x})$, there exist a $\hat{\rho}>0$ and an $\hat{\varepsilon} \in E$, $\hat{\varepsilon}>0$ such that $\varepsilon(x) \geq \hat{\varepsilon}$ for all $x \in B(\hat{x}, \hat{\rho})$.

b) Suppose that $\hat{x}$ solves (8.1), then $\varepsilon(\hat{x})=0$. ▪

Theorem 8.2: Suppose $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm 8.2 in solving (8.1), with $f(\cdot)$, $\psi(\cdot)$ l.L.c. If $x_i \to^K \hat{x}$ as $i \to \infty$ ($K \subset \{0,1,2,\ldots\}$), then $\psi(\hat{x}) \leq 0$ and $0 \in G^0_{f,\psi}(\hat{x})$ (and $\varepsilon(\hat{x})=0$).

Proof: First, since $\psi(x_i) \leq 0$ for all i, we must have $\psi(\hat{x}) \leq 0$ for any accumulation point $\hat{x}$ of $\{x_i\}$. For the sake of contradiction, suppose that $x_i \to^K \hat{x}$ and $0 \notin G^0_{f,\psi}(\hat{x})$. Then $\varepsilon(\hat{x})>0$ and, by Lemma 8.3, there is an i_0 and an $\hat{\varepsilon} \in E$ such that $\varepsilon(x_i) \geq \hat{\varepsilon} > 0$ for all $i \geq i_0$ and $i \in K$.

a) Suppose that $\psi(\hat{x})<0$. Then, by essentially repeating the arguments of the proof of Theorem 8.1 we can show that there is an $i_1 \geq i_0$ and a $\hat{\lambda}>0$ such that for all $i \in K$, $i \geq i_1$, $\psi(x_i + \hat{\lambda} h_i) \leq 0$ while

$$f(x_i + \hat{\lambda} h_i) - f(x_i) \leq -\hat{\lambda}\hat{\varepsilon}/2 \tag{8.22}$$

since $f(x_{i+1}) - f(x_i) \leq f(x_i + \hat{\lambda} h_i) - f(x_i)$, we are led to a contradiction, exactly as in the proof of Theorem 8.1.

b) Suppose that $\psi(\hat{x})=0$. Then there exists an $i_2 \geq i_0$ such that for all $i \in K$, $i \geq i_2$, $\psi(x_i) \geq -\hat{\varepsilon} \geq -\varepsilon(x_i)$ so that $G^{\varepsilon(x_i)}_{f,\psi}(x_i)$ is given by (8.18b). Consequently,

$$h_i = arg\ \min \{\tfrac{1}{2}\|h\|^2 + \max \{d_{\varepsilon(x_i)} f(x_i;h),$$
$$d_{\varepsilon(x_i)}\psi(x_i;h)\}\} \tag{8.23}$$

and is a descent direction for both $f(\cdot)$ and $\psi(\cdot)$. Repeating again the arguments in the proof of Theorem 8.1, this time both for $f(\cdot)$ and for $\psi(\cdot)$, we conclude that there is an $i_3 \geq i_2$ and a $\hat{\lambda} \geq 0$ such that for all $i \in K$, $i \geq i_3$,

$$f(x_i + \hat{\lambda} h_i) - f(x_i) \leq -\hat{\lambda}\hat{\varepsilon}/2 \tag{8.24}$$

$$\psi(x_i + \hat{\lambda} h_i) - \psi(x_i) \leq -\hat{\lambda}\hat{\varepsilon}/2 \tag{8.25}$$

which, clearly, leads to a contradiction of the fact that $f(x_i) \to f(\hat{x})$ as $i \to \infty$. Hence our proof is complete. ▪

So far, we have assumed that to solve (8.1) we use one algorithm to obtain a feasible starting point x_0 (such that $\psi(x_0) \leq 0$) and then apply a second algorithm to optimize the design. There are two disadvantages to this: (i) two codes need to be written and used, and data must be transferred from one to the other; and (ii) the phase I process (computation of a feasible initial design) can produce bad initial designs since it pays no attention to the cost. We now show that these disadvantages can be mitigated by constructing a phase I-phase II algorithm for solving (8.1).

Let $\zeta : \mathbb{R} \to \mathbb{R}$ be a continuous, monotonic increasing function such that $\zeta(0)=0$ and $\zeta(t) \to \infty$ as $t \to \infty$. We define the phaseI-phase II c.d.f. maps as having values in $2^{\mathbb{R}^{n+1}}$.

Definition 8.4: Let $\{G_\varepsilon f(\cdot)\}_{\varepsilon \geq 0}$ and $\{G_\varepsilon \psi(\cdot)\}_{\varepsilon \geq 0}$ be given families of c.d.f. maps for the l.L.c. functions $f(\cdot)$ and $\psi(\cdot)$ in (8.1), let $\psi(x)_+ \triangleq \max\{0, \psi(x)\}$ and $\gamma > 0$ be given. We define the family of phase I-phase II c.d.f. maps $\{\overline{G}^\varepsilon_{f,\psi}(\cdot)\}_{\varepsilon \geq 0}$ for (8.1) by setting

$$\overline{G}^\varepsilon_{f,\psi}(x) = \{\overline{\xi} \in \mathbb{R}^{n+1} \mid \overline{\xi} = (0, \xi),\ \xi \in G_\varepsilon f(x)\}$$

$$\text{if } \psi(x) < -\varepsilon \tag{8.26a}$$

$$\overline{G}^{\varepsilon}_{f,\psi}(x) = co\{\overline{\xi}\in I\!R^{n+1} \mid \overline{\xi} = (\zeta(\gamma\psi(x)_+)\,,\,\xi)\,,with\ \xi\in G_\varepsilon f(x)\ or\ \overline{\xi} = (0\,,\,\xi)\,,with\ \xi\in G_\varepsilon\psi(x)\}$$

$$\text{if } \psi(x) \geq -\varepsilon \tag{8.26b}$$

Next, we define

$$\overline{h}_\varepsilon(x) = (h^0_\varepsilon(x), h_\varepsilon(x) \triangleq argmin\{\tfrac{1}{2}\|h\|^2 \mid h\in\overline{G}^{\varepsilon}_{f,\psi}(x)\} \tag{8.27}$$

and

$$\varepsilon(x) \triangleq \max\{\varepsilon\in E \mid \|\overline{h}_\varepsilon(x)\|^2 \geq \varepsilon\} \tag{8.28}$$

Exercise 8.5: Show that $\overline{h}_\varepsilon(x)$ defined by (8.27) satisfies

$$\max\{d_\varepsilon f(x;h) - \sqrt{\gamma\psi(x)_+}h^0,\ d_\varepsilon\psi(x;h)\} \tag{8.29}$$

where $d_\varepsilon f(x;h) \triangleq \max_{\xi\in G_\varepsilon f(x)}\langle\xi,h\rangle$ and $d_\varepsilon\psi(x;h) \triangleq \max_{\xi\in G_\varepsilon\psi(x)}\langle\xi,h\rangle$. ▪

Note the effect of $\psi(x)$ on $\overline{h}_\varepsilon(x)$. When $\psi(x)_+$ is large, then $h_\varepsilon(x) \cong argmin\{\frac{1}{\varepsilon}\|h\|^2 \mid h\in G_\psi(x)\}$. When $\psi(x)\leq 0$, $h_\varepsilon(x)$ is the same as computed in Algorithm 8.2 (phase II). When $\psi(x)>0$ and decreases to zero, the effect of the cost ($G_\varepsilon f(x)$) on $h_\varepsilon(x)$ becomes progressively more pronounced. For the case where $f(cdot)$ and $\psi(\cdot)$ are differentiable, this effect is illustrated in the figure below.

We can now state a phase I-phase II algorithm for solving (8.1). (Note that in [Pol 3], a less efficient phase I-phase II scheme is described).

Algorithm 8.3: [Requires $\{G_\varepsilon f(\cdot)\}_{\varepsilon\geq 0}$, $\{G_\varepsilon\psi(\cdot)\}_{\varepsilon\geq 0}$ families of c.d.f. maps for $f(\cdot)$ and $\psi(\cdot)$; $\alpha\in(0,1)$ for the set E in (8.14)].

Data: $x_0\in I\!R^n$.

Step 0: Set $i=0$.

Step 1: Compute $\varepsilon(x_i)$ according to (8.28) and the search direction

$h_i = -h_{\varepsilon(x_i)}(x_i)$ according to (8.27).

Step 2: Compute the step size

$$\lambda_i \in \lambda_\psi(x_i) \triangleq \underset{\lambda \geq 0}{argmin}\, \psi(x_i + \lambda h_i) \ \text{ if } \ \psi(x_i) > 0$$
$$\lambda_i \in \lambda_f(x_i) \triangleq \underset{\lambda \geq 0}{argmin}\{ f(x_i + \lambda h_i) \mid \psi(x_i + \lambda h_i) \leq 0 \} \ \text{ if } \ \psi(x_i) \leq 0 \tag{8.30}$$

Step 3: Update: $x_{i+1} = x_i + \lambda_i h_i$, replace i by $i+1$ and go to step 1. ▪

Exercise 8.6: Prove the following.

Theorem 8.3: Suppose $\{x_i\}_{i=0}^{\infty}$ is a sequence constructed by Algorithm 8.3 in solving (8.1) with $f(\cdot)$, $\psi(\cdot)$, l.L.c. and suppose that $0 \notin G_0\psi(x)$ for all $x \in \mathbb{R}^n$ such that $\psi(x) \geq 0$. If $x_i \overset{k}{\to} \hat{x}$ as $i \to \infty$ ($K \subset \{0,1,2,...\}$), then $\psi(\hat{x}) \leq 0$ and $0 \in \bar{G}^0_{f,\psi}(\hat{x})$. ▪

This concludes our brief exposition of semi-infinite algorithm theory. It is worth noting that the step size rules of exact minimization that we have presented can be replaced by much more efficient Armijo-like step size rules without affecting the convergence properties of the algorithms. The important question of approximations in evaluating max functions in the execution of algorithms was not touched in these notes. The interested reader is referred to [Gon 2], [Pol 1] and [Tra 1] for details on implementation of conceptual algorithms. Finally, we point out that when the maximalization is over a multidimensional set or when the maximand is not differentiable, it may be more convenient to decompose an optimization problem by means of outer approximation techniques. For details see [Gon 1], [Pol 7].

Acknowledgments

Research sponsored in part by the Office of Naval Research Contract N00014-83-K-0602, and the Air Force Office of Scientific Research Grant AFOSR-83-0361.

9. REFERENCES.

[Apo 1] Apostol, T. M., *Mathematical Analysis*, Addison-Wesley, Reading, Mass., 1960.

[Ber 1] Berge, C., *Topological Spaces*, MacMillan, New York, N.Y., 1963.

[Cla 1] Clarke, F. H., *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, N.Y., 1983.

[Dan 1] Danskin, J. M., The Theory of Max Min, Springer-Verlag, New York, N.Y., 1967.

[Die 1] Dieudonne, J., *Foundations of Modern Analysis*, Academic Press, New York, N.Y., 1960.

[Gon 1] Gonzaga C., and E. Polak, " On constraint dropping schemes and optimality functions for a class of outer approximatins algorithms", *SIAM J. Control and Optimization*, Vol. 17, No.4, 1979, pp 477-493.

[Gon 2] 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.

[Pol 1] Polak, E., *Computational Methods in Optimization*, Academic Press, 1971.

[Pol 2] Polak, E., "Semi-Infinite Optimization in Engineering Design", pp. 236-245 in *Semi-Infinite Programming and Applications*, A. V. Fiacco and K. O. Kortanek eds., Springer-Verlag New York, N.Y., Vol. 215, *Lecture Notes in Economics and Mathematical Systems*, 1983

[Pol 3] Polak, E., and D. Q. Mayne, "Algorithm models for nondifferentiable optimization", University of California, Electronics Research Laboratory Memo UCB/ERL No. M82/34, 5/10/82, to appear in SICOPT.

REFERENCES

[Pol 4] Polak, E., D.Q. Mayne and Y. Wardi, "On the Extension of Constrained Optimization Algorithms from Differentiable to Nondifferentiable Problems", *SIAM J. Cont. & Opt.*, Vol. 21, No.2, pp 179,204, 1983.

[Pol 5] Polak, E., R. Trahan and D. Q. Mayne, "Combined phase I - phase II methods of feasible directions", *Mathematical Programming*, Vol. 17, No. 1, pp 32-61, 1979.

[Pol 6] Polak, E., and Y. Wardi, "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.

[Pol 7] 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.

[Pon 1] Pontryagin, L. S., V. G. Boltianskii, R. V. Gamkrelidze and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Addison-Wesley Co. Reading Mass., 1962.

[Roc 1] Rockafellar, R. T., *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

[Ste 1] Stein, E.M., *Singular integrals and differentiability properpties of functions*, Princeton Math Series, No. 30, Princeton University Press, Princeton, NJ, 1970. MR 44 #7280.

[Tra 1] Trahan R., and E. Polak, "A derivative free algorithm for a class of infinitely constrained problems", *IEEE Transactions on Automatic Control*, Vol. AC-25, No.1 1979, pp. 54-62.

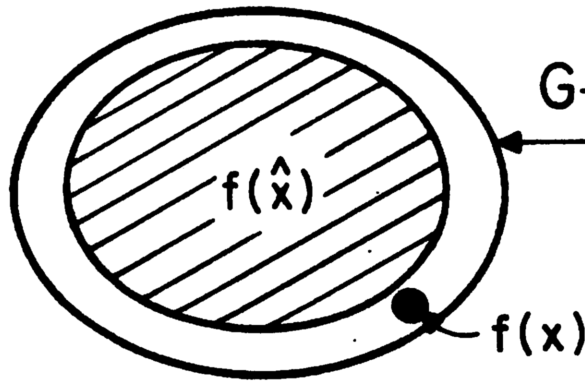

Fig. 1.1. f(•) is u.s.c.

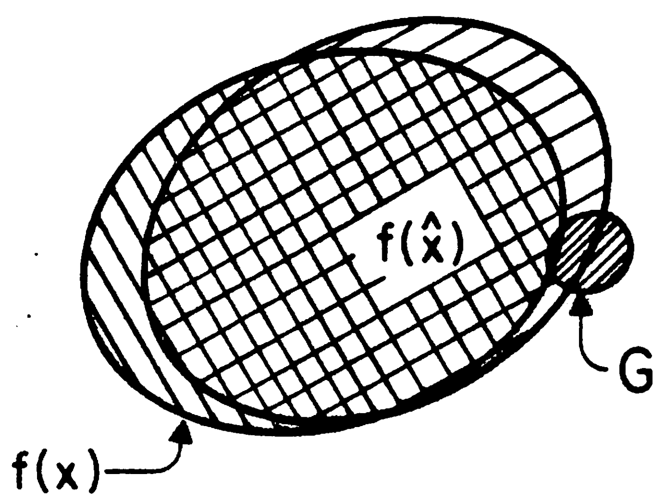

Fig. 1.2. f(•) is l.s.c.

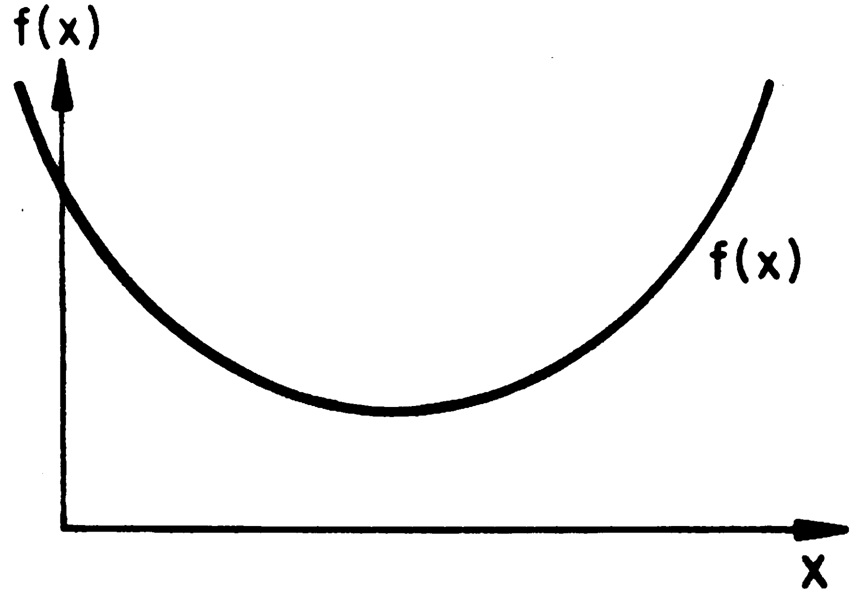

Fig. 3.1. A convex function.

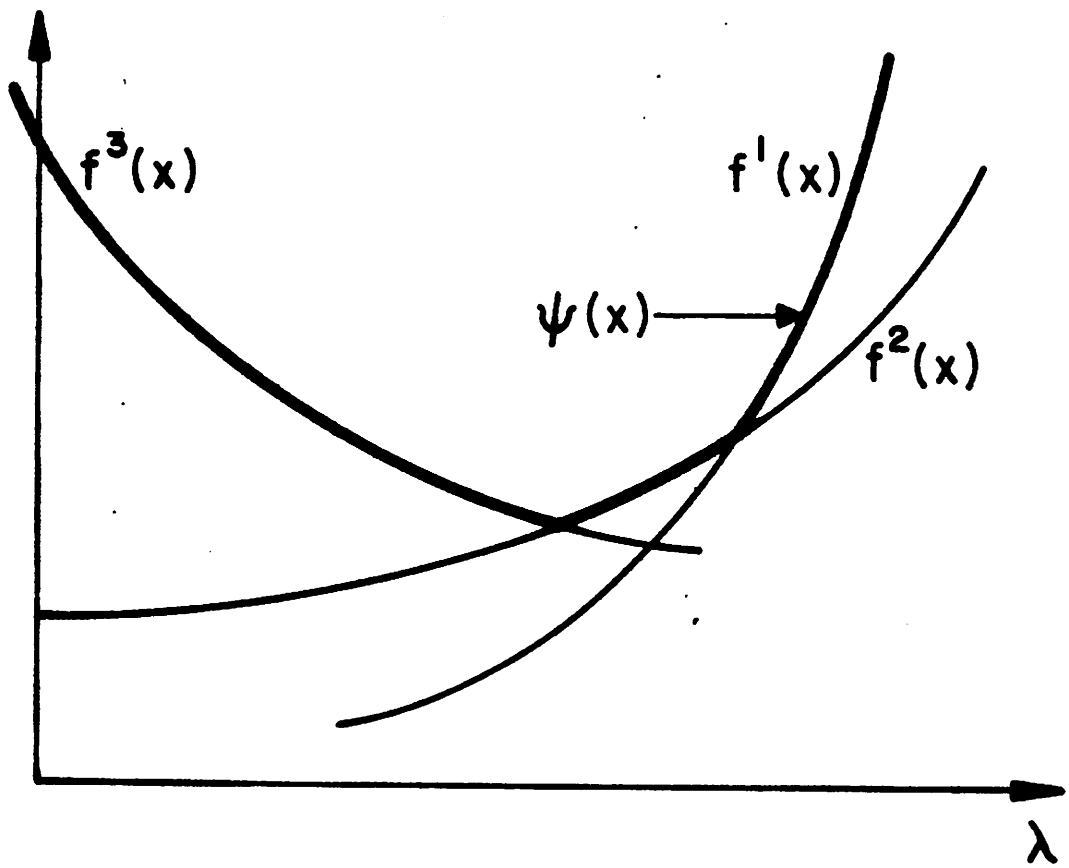

Fig. 4.1. Graph of $\psi(x) = \max_{j \in \underline{3}} f^j(x)$.

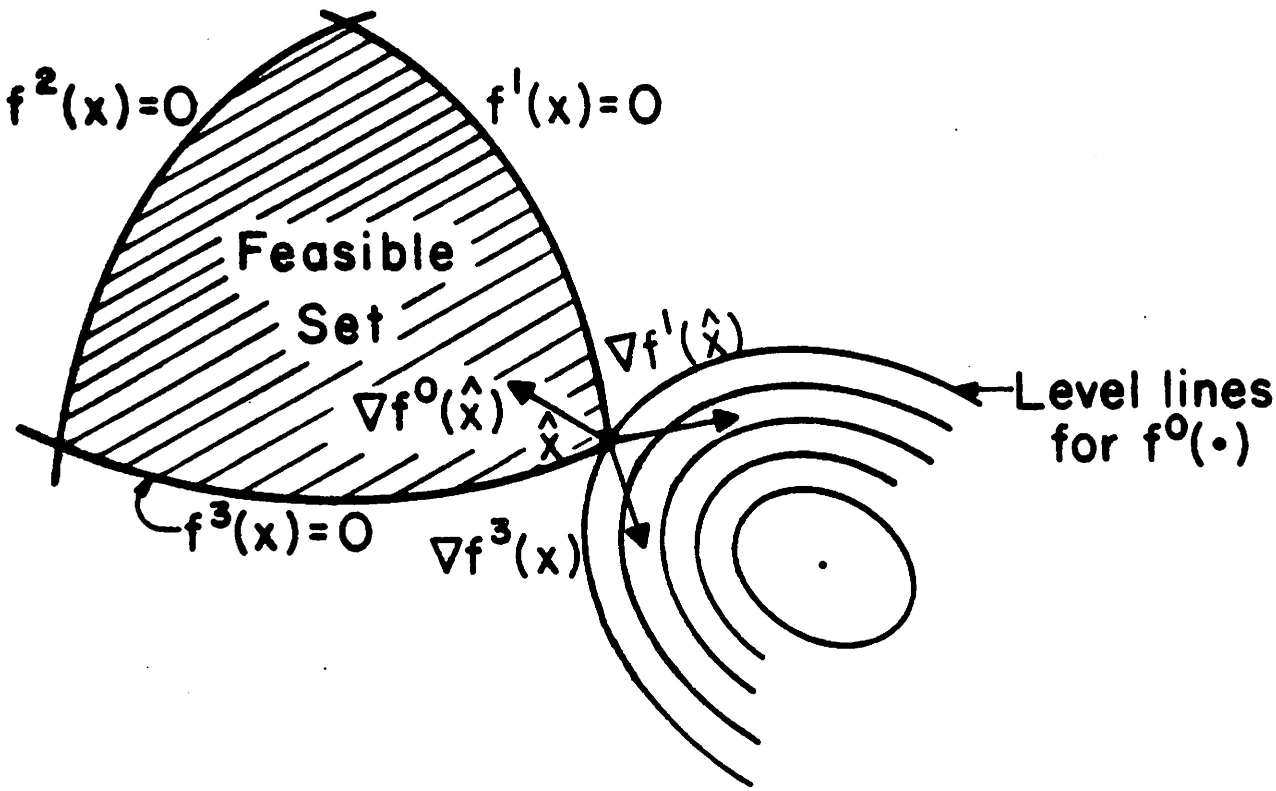

Fig. 5.1. $\hat{x}$ solves $\min \{f^0(x) | f^j(x) \leq 0,\ j = 1,2,3\}$

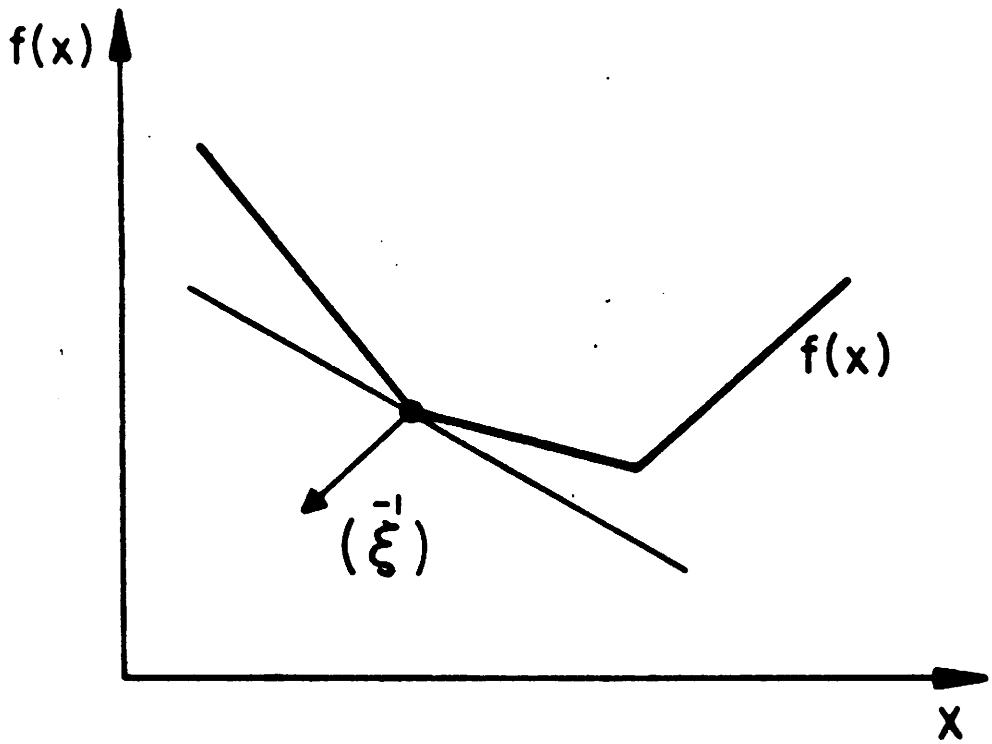

Fig. 6.1. Subgradients of a convex function.

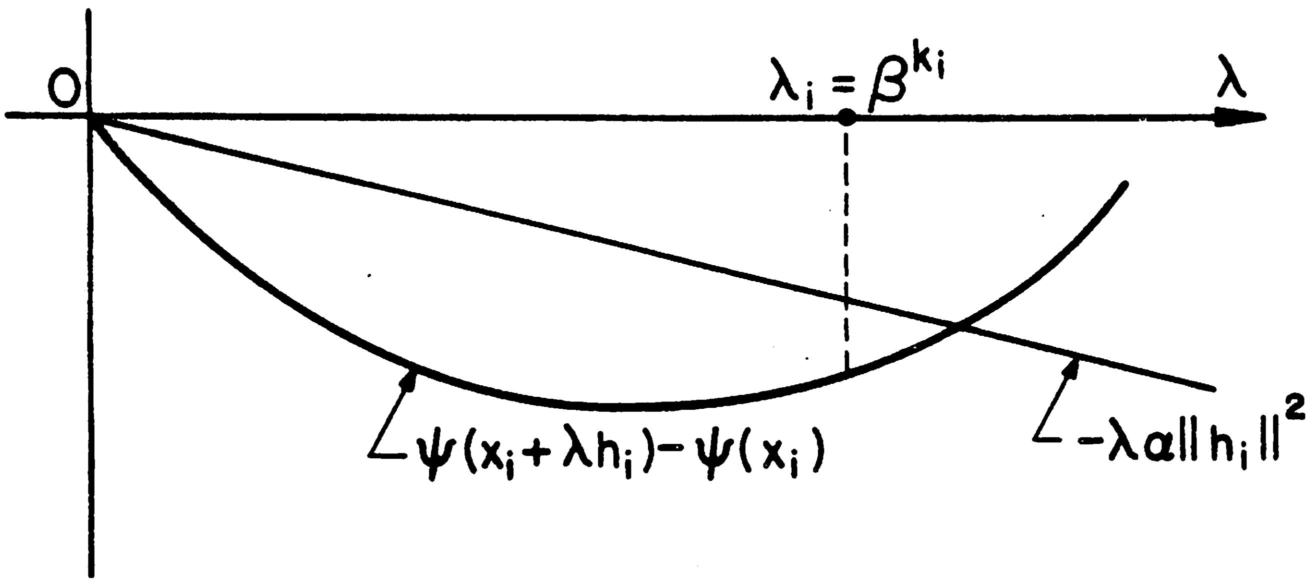

Fig. 7.1. The Armijo step size rule.

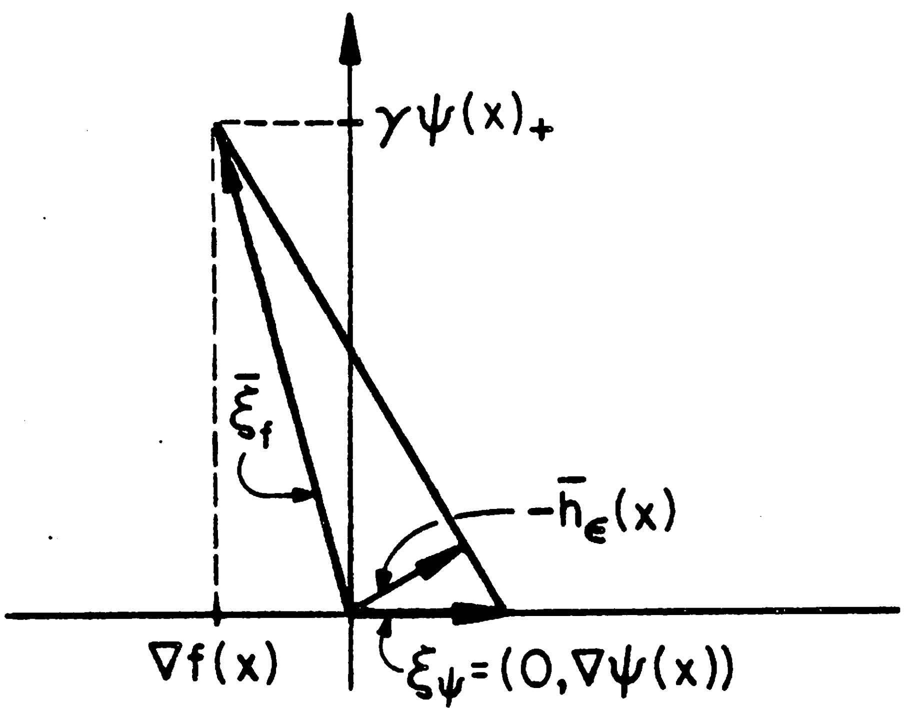

Fig. 8.1. Effect of cost on search direction for $\psi(x) > 0$.