HAL open science

Basic theory in nondifferentiable optimization. A tutorial and algorithmic oriented approach

Claude Lemarechal

▶ To cite this version:

Claude Lemarechal. Basic theory in nondifferentiable optimization. A tutorial and algorithmic oriented approach. RR-0181, INRIA. 1982. inria-00076377

HAL Id: inria-00076377
https://inria.hal.science/inria-00076377v1

Submitted on 24 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

IRIA

CENTRE DE ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tél. 954 90 20

Rapports de Recherche

N° 181

BASIC THEORY IN NONDIFFERENTIABLE OPTIMIZATION

A TUTORIAL AND ALGORITHMIC ORIENTED APPROACH

Claude LEMARÉCHAL

Décembre 1982

BASIC THEORY IN NONDIFFERENTIABLE OPTIMIZATION
A TUTORIAL AND ALGORITHMIC ORIENTED APPROACH

Claude Lemaréchal, INRIA (France)

ABSTRACT.

This paper gives the state of the art concerning the optimization of nondifferentiable functions. It contains essentially two distinct parts. The first part concerns the study of locally Lipschitz functions, Clarke's subdifferential and related topics; various classes of nondifferentiable functions are studied in connection with this theory, serving as illustrations. The second part is devoted to the convex case, with a study of the approximate subdifferential.

RESUME.

Cet article présente l'état de l'art dans l'optimisation des fonctions non différentiables. Il contient essentiellement deux parties. La première concerne l'étude des fonctions localement lipschitziennes, le gradient généralisé de Clarke et sujets connexes; diverses classes de fonctions localement lipschitziennes, données à titre illustratif, sont étudiées relativement à cette théorie. La deuxième partie est consacrée au cas convexe, avec une étude du sous-différentiel approché.

PAPIER RECUPERÉ ET.

INTRODUCTION.

These notes form a part of a course taught by the author at the Federal University of Rio de Janeiro (UFRJ) in August-September 1981, in the framework of a collaboration between this University and INRIA. The author gratefully acknowledges support from the Institute of Mathematics and COPPE, at UFRJ.

We have summarized here what we believe to be the state of the art in optimization theory of nonsmooth functions. Practically none of the results mentioned here is new. Most of them can be found in various papers on the subject, which are mentioned when necessary. Our constant care has been to show the geometric intuition behind these results, rather than their proofs, which can be found in the relevant literature.

Other review papers on the subject exist, see for example [22], but the present development is mainly motivated by two considerations: pedagogy (results whose knowledge brings more understanding) and algorithmic applications (results which happen to be important for the existing algorithms - however these notes do not contain any algorithmic development: other papers are devoted to this subject, one of them is [15]).

Nondifferentiable optimization deals with functions which are not continuously differentiable. For such functions f, one asks questions such as minimize f(x), or find a point x such that $f(x) \leq 0$ and one has to study their local behaviour, i.e. to find how the concept of derivative can be properly generalized.

Note that this has nothing to do with "optimization without calculating derivatives". To avoid confusion, one replaces sometimes "nondifferentiable" by "nonsmooth".

Unfortunately, the above definition (f not differentiable) is negative and means actually nothing but "optimization of any function" (a differentiable function is just a particular example of a non differentiable function). Hence one needs to be more specific. At present, there is no general consensus on exactly what to do, except when f is assumed convex; our aim here is to present the Clarke theory [5], which is well developed, and which appears useful in applications. Then we study various classes of functions for which Clarke's theory is particularly well suited; among them are convex functions, whose approximate subdifferential appears essential for applications.

1. THE CLARKE GENERALIZED GRADIENT OR PERIDIFFERENTIAL

We assume that the function f under consideration maps the usual Euclidian space R^n into R, and that it is locally Lipchitz:

For any bounded set S in R^n, there exists $L > 0$ such that

$$|f(x)-f(y)| \leq L\ |x-y| \qquad \forall x,y \text{ in } S$$

1.1. Construction

From an old theorem due to Rademacher (see [24]) f has a gradient (vector of partial derivatives) almost everywhere. Therefore, given $x \in R^n$ even if $\nabla f(x)$ does not exists, any neighborhood of x contains a point y where $\nabla f(y)$ exists. When this neighborhood shrinks to $\{x\}$, $\nabla f(y)$ is bounded (by Lipschitz property) therefore it has at least one cluster point. Consider all possible cluster points, obtained for all possible such y's. These form a set defined by

$$M(x) = \{ \lim \nabla f(y) : \nabla f(y) \text{ exists}, y \to x, \nabla f(y) \text{ has a limit } \}$$

Then we consider the convex hull of M(x)

$$\partial f(x) = \{ \Sigma_i u_i g_i : g_i \in M(x), u_i \geq 0, \Sigma u_i = 1 \}$$

Clarke calls this set the generalized gradient. We do not like this terminology, and we prefer to call a generalized gradient an element of $\partial f(x)$. The whole set $\partial f(x)$ can be called the Clarke's subdifferential, or <u>peridifferential</u> of f at x. This latter wording, proposed by Penot, is particularly suggestive because $\partial f(x)$ represents the behaviour of f around x.

When f has a continuous gradient at x, then $M(x) = \partial f(x)$ = the singleton $\nabla f(x)$. In this case, the change $f(x+z)-f(x)$ can

be approximated by $(\nabla f(x),z)$, a linear function of z.

When $M(x)$ is not a singleton, the change $f(x+z)-f(x)$ cannot be approximated by a linear function.

1.2. Basic properties

1.2.1. From the Lipschitz property, $M(x)$ is nonempty and bounded. It is closed by definition (any limit of g's in $M(x)$ is a "limit of limit", still a limit and belongs to $M(x)$). Thus, $\partial f(x)$ is a <u>nonempty, closed, bounded, convex</u> set.

1.2.2. For the same sort of reason the multi-valued mapping $M(x)$ is closed, i.e. if $x_k \to x$, $g_k \in M(x_k)$ and $g_k \to g$, then $g \in M(x)$, and so is the multi-valued mapping ∂f. Observe that, from Lipschitz property, ∂f is locally bounded (i.e. $\partial f(x)$ is bounded uniformly with respect to x varying in a bounded set). Thus, in the above sequence g_k, we can always extract a convergent subsequence.

From properties 1.2.1 and 1.2.2, we say that f is an <u>upper-semi-continuous</u> multi-valued mapping.

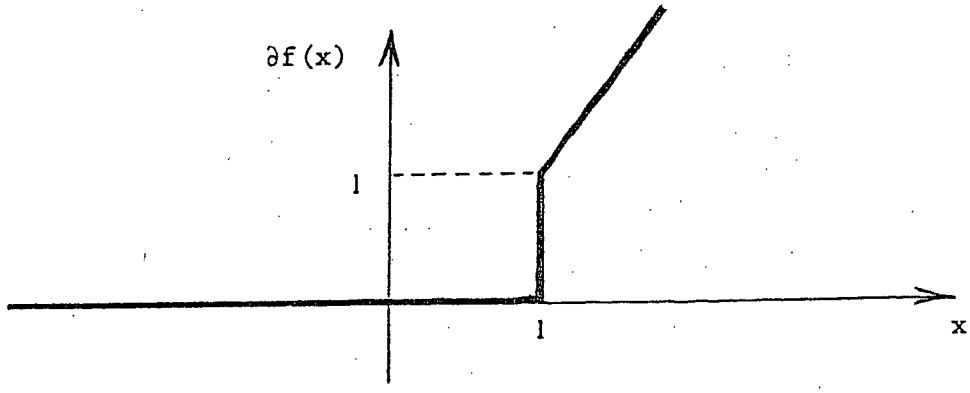
N.B. Upper-semi-continuity is a stronger property than closedness: consider, for $x \in [0,1]$, $F(0) = 0$, $F(x) = 1/x$ for $x > 0$. This mapping (a function) is closed but it would not be sensible to say that it is upper-semi-continuous (see [2]).

Upper-semi-continuity is the best continuity property we can obtain for $\partial f(x)$. In particular it cannot be lower-semi-continuous, which would mean: if $x_k \to x$ and if $g \in \partial f(x)$, then there exists a sequence $g_k \in \partial f(x_k)$ tending to g. Example: for n=1, consider

$$f(x) = \begin{cases} 0 & \text{if } x \leq 1 \\ x^2 - x & \text{if } x \geq 1 \end{cases}$$

Then $M(x) = \partial f(x) = \begin{cases} 0 & \text{if } x < 1 \\ 2x - 1 & \text{if } x > 1 \end{cases}$

so $M(1) = \{0,1\}$ and $\partial f(1) = [0,1]$.

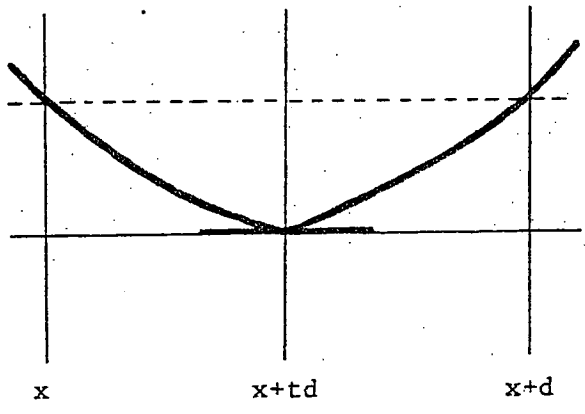


Observe that $g = 1/2 \in \partial f(1)$ but cannot be approximated by any derivative of f. On the other hand, any element of M(x) can be, by definition, approximated by such a derivative. Therefore it may seem artificial to take its convex hull, but this is motivated by the following property.

1.2.3. An important property is the generalization of the <u>mean value theorem</u>:

$$\left\{\begin{array}{l}\text{There exist } t \in]0,1[\text{ and } g \in \partial f(x+td) \text{ such that} \\ f(x+d) - f(x) = (g,d)\end{array}\right.$$

which can be illustrated geometrically by the next picture.



Note that the "slope" g may not belong to M(x+td), but it is between the two half derivatives of f(x+td) considered as a function of the scalar t.

N.B. The peridifferential is not the only possible set that can be defined to generalize the concept of gradient. However, Lebourg [12] has proved that, if one defines another "subdifferential", say δf which satisfies both upper-semi-

continuity and mean value theorem, then $\partial f(x) \subset \partial f(x)$. Thus, the Clarke subdifferential is minimal in some sense.

1.2.4. If f is multiplied by a scalar k, so are its gradients, and so is $M(x)$; therefore $\partial[kf](x) = k\ \partial f(x)$, if we interpret the latter as the set obtained by multiplying by k every element of $\partial f(x)$.

On the other hand, it is not true that $\partial[f_1 + f_2](x) = \partial f_1(x) + \partial f_2(x)$, if we interpret the latter as the set obtained by adding every element of $\partial f_1(x)$ to every element of $\partial f_2(x)$. Take for example $f_1(x) = |x|$, $f_2(x) = -|x|$, so that $\partial f_1(0) = \partial f_2(0) = [-1,+1]$. Then $[f_1 + f_2](x) = 0$, $\partial[f_1 + f_2](0) = 0$ whereas $\partial f_1(0) + \partial f_2(0) = [-2,+2]$. The trouble is that the sum of two sets is "bigger" than each of the sets, unless the other is a singleton.

Actually, the addition law is:

$$\partial[f_1 + f_2](x) \subset \partial f_1(x) + \partial f_2(x)$$

1.2.5. The previous example shows that a <u>linear operation</u> does not preserve the peridifferential but increases it. The same inclusion holds for more general operations [17]: let $f_1, f_2, \ldots, f_m$ be locally Lipschitz functions, and let E be a locally Lipschitz function on R^m. Then consider the composition

$$f(x) = E\ [f_1(x), f_2(x), \ldots, f_m(x)]$$

This f is a locally Lipschitz function, so it has a peridifferential.

Given x, call $y \in R^m$ the vector $f_1(x), f_2(x), \ldots, f_m(x)$. Compute $w \in \partial E(y)$ and for each i, compute $g_i \in \partial f_i(x)$. Then form the matrix J whose rows are $g_1, \ldots, g_m$ and, analogously to the gradient of composite functions, form the vector J w.
Then consider the set of all such vectors when w, $g_1, g_2, \ldots g_m$ vary in their respective sets:

$$N(x) = \{ J\,w \;:\; g_i \in \partial f_i(x),\ w \in \partial E(y) \}$$

Then

$$\partial f(x) \subset \operatorname{conv} N(x)$$

1.3. <u>A fundamental relation</u>

In the smooth case, the gradient serves to estimate changes $f(x+z)-f(x)$. In particular the differential quotient $[f(x+td)-f(x)]/t$, obtained for fixed d and $t \downarrow 0$, tends to $(\nabla f(x),d)$.

In the nonsmooth case, the differential quotient may not converge to a limit (although it is bounded, from Lipschitz property) so it is necessary to do something else. In Clarke's theory, it is convenient to define the Clarke <u>generalized derivative</u>:

$$(1) \qquad f^{o}(x,d) = \limsup \{ [f(x'+td)-f(x')]/t \;:\; x' \to x,\ t \downarrow 0 \}$$

Roughly speaking, $f^{o}(x,d)$ (which exists because of Lipschitz property) is an upper bound on the differential quotient in the neighborhood of x.

The reason for introducing (1) is that we have the following formula:

$$(2) \qquad f^{o}(x,d) = \max \{ (g,d) \;:\; g \in \partial f(x) \}$$

where, of course, "max" is due to the fact that f^{o} involves a "lim sup". This implies in particular that $f^{o}(x,d)$ is convex with respect to d.

Note that, if f has a continuous gradient at x, then

$$f^{o}(x,d) = (\nabla f(x),d)$$

so f^{o} really generalizes the usual derivative.

Thus we have defined a pair $\{\partial f(x), f^{o}(x,d)\}$. It would be equivalent to define first $f^{o}(x,d)$ by (1), thus obtaining a convex and positively homogeneous function in d, and then to define $\partial f(x)$ by (2).

This a general scheme in nondifferentiable optimization: one defines a pair {generalized gradient, generalized directional derivative} and there is a choice:

- either to define first a convex set generalizing the gradient, and then define its support function which generalizes the directional derivative (this is the present approach)
- or to define a convex positively homogeneous function generalizing the directional derivative, and then define the convex set having this function as its support function.

1.4. Descent directions

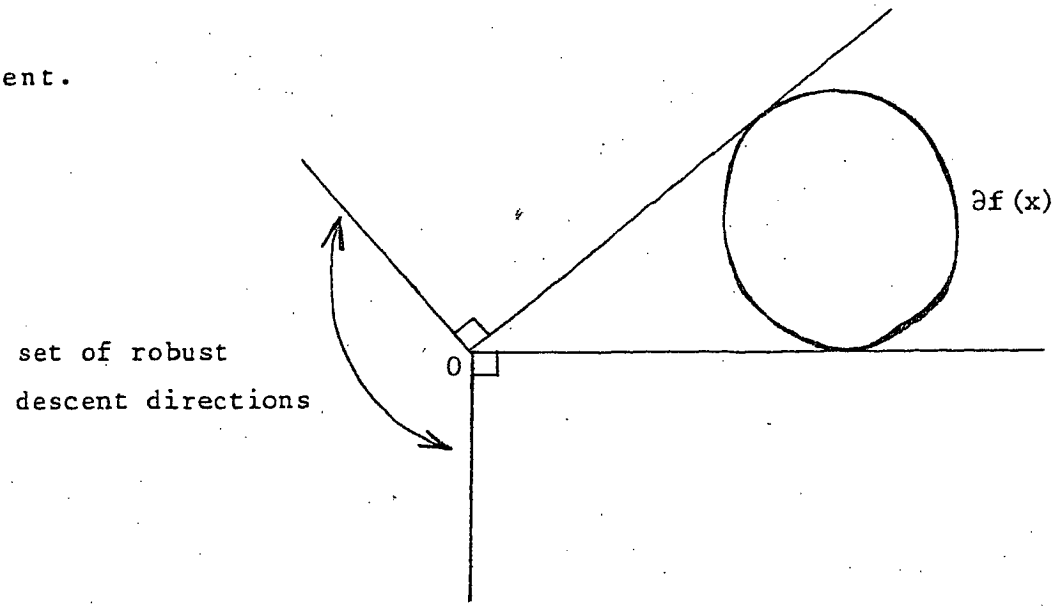
For numerical algorithms solving optimization problems, it is convenient to consider vectors d such that $f(x+td) < f(x)$ for $t > 0$ small enough. Such vectors are called descent directions. They are not easy to characterize mathematically. Therefore we relate these directions to generalized gradients and derivatives, and we say that d is a robust descent direction in Clarke's sense if:

$$f^{o}(x,d) < 0 \quad \text{or equivalently} \quad (g,d) < 0 \qquad \forall g \in \partial f(x)$$

(such directions are also called directions of uniform descent in [22]; the reason is that d is still a descent direction at any x' close enough to x, as can be seen from the definition (1)).

Thus, the set of robust descent directions forms an open convex cone, conjugate to the cone generated by $\partial f(x)$. The set of mere descent directions contains this cone. The broader $\partial f(x)$, the smaller this cone. As an extreme case, if $\partial f(x) = \{\nabla f(x)\}$ is a singleton, then this cone is the half space opposite to the

gradient.



1.5. Optimality conditions

The following result gives a necessary condition for optimality of a locally Lipschitz function :

(3) } If $\underline{x}$ is a local minimum then $0 \in \partial f(\underline{x})$.

Observing that a robust descent direction defines a hyperplane separating $\partial f(x)$ strictly from $\{0\}$, (3) actually says that there is no robust descent direction issued from $\underline{x}$ (otherwise there would be a hyperplane separating 0 from $f(\underline{x})$); this must be true if $\underline{x}$ is to be a local minimum.

If $0 \in \partial f(x)$ we say that x is critical, or stationary.

The above necessary condition works also with constraints: consider the problem

$$\begin{cases} \min f(x) \\ h_j(x) \leq 0 \quad j=1,\dots,m \end{cases}$$

and consider the set of active constraints

$$J(x) = \{ j \in \{1,\dots,m\} \ : \ h_j(x) = 0 \}$$

In the smooth case, the John condition says that 0 belongs to the convex hull of the set $\{\nabla f(x) \ ; \nabla h_j(x), \ j \in J(x)\}$ for any local solution.

Here, when f and the h_j's are locally Lipschitz, we have the same kind of condition [6]:

$$\left\{\begin{array}{l} \exists u_0 \geq 0,\ u_j \geq 0\ j \in J(x) \quad \text{not all zero} \\ \text{and } g_0 \in \partial f(x),\ g_j \in \partial h_j(x)\ j \in J(x) \\ \text{such that} \quad u_0 g_0 + \Sigma\, u_j g_j = 0 \end{array}\right.$$

In the smooth case it is known that, if a qualification hypothesis is satisfied by x, then the u_0 multiplying $\nabla f(x)$ is strictly positive, so it can be taken to 1, to obtain the Kuhn Tucker optimality condition.

Here, a suitable qualification condition can be written [9]:

$$\left\{\ \exists d \text{ such that } h_j^o(x,d) < 0 \quad \forall j \in J(x)\right.$$

which implies the Kuhn Tucker form of optimality condition for locally Lipschitz functions:

$$\left\{\ 0 \in \partial f(x) + \Sigma\ u_j\, \partial h_j(x)\right.$$

1.6. <u>Is the peridifferential a suitable generalization</u> ?

Because robustness restricts the class of descent directions, there may be many critical points which are not local minima: in the above "proof" of (3) the insertion of the word "robust" in front of "descent direction" may destroy a lot of information , and this is one of the difficulties in nondifferentiable optimization.

It has been said that, instead of the pair $\{\partial f(x), f^o(x,d)\}$, one can define other generalizations of directional derivatives, say $f^{\cdot}(x,d)$, which define other generalized gradients, say $\partial f(x)$, through

$$f^{\bullet}(x,d) = \max \{ (g,d) \ : \ g \in \delta f(x) \}$$

To every such definition is associated a set of robust descent directions: those such that $f^{\bullet}(x,d) < 0$, and a set of stationary points: those such that $0 \in \delta f(x)$.

When δf and $f^{\bullet}$ are larger, robustness is more restrictive and stationarity tells less: there are more stationary points. Therefore, one usually wishes to define generalizations of derivatives such that δf and $f^{\bullet}$ are as small as possible, so that the set of critical points which are not minimal is as small as possible. Two examples show why it is often considered that the Clarke subdifferential is too large, or equivalently that robustness in the sense of Clarke is too restrictive.

a) Take $f(x) = -|x|$. For $x = 0$, we have $\partial f(0) = [-1,+1]$ which contains 0, so $x = 0$ is critical. At this point $f^{o}(0,1) = f^{o}(0,-1) = 1$, none of the two possible directions ($d=1$ and $d=-1$) is a robust descent direction. Yet, the directional derivatives exist and we have $f'(0,1) = f'(0,-1) = -1$, both are descent directions.

The trouble illustrated by this example is that $f^{o}(x,d)$ depends on the behaviour of f outside the set $\{x+td \ : \ t > 0 \}$.

b) The above example shows that $0 \in \partial f(x)$ cannot filter critical points which are local maxima. It may be argued that this does not matter since, in the smooth case, the situation is the same (cf. $f(x)=-x^2$). However the following example is more subtle:

$$f(x) = x + x^2 \sin 1/x \quad \text{if } x > 0 \quad , \quad f(x) = x \quad \text{if } x \leq 0$$

For $x > 0$ we have $f'(x) = 1 + 2x \sin 1/x - \cos 1/x$, so f is locally Lipschitz. Observe that $M(0) = \partial f(0) = [0,2]$, so $x = 0$ is

critical; this is really a paradox because $f'(0)$ exists and is equal to 1.

The argument here is different from that of a) above: because f is smooth for $x \leq 0$, $\partial f(0)$ depends only on the behaviour of f for $x > 0$. In other words $f^{\circ}(0,1)$ depends only on $f(0+t.1)$ for $t > 0$.

Thus, admitting that Clarke's theory leads to too large estimates, one can try and and sharpen them. For example Penot [19] defines the generalized derivative

$$f^{\cdot}(x,d) = \liminf \{ [f(x+td') - f(x)]/t \ : \ d' \to d, \ t \downarrow 0 \}$$

Observe the difference with (2): one takes the lim inf instead of lim sup, and it is d that varies instead of x. Therefore the optimality condition $f^{\cdot}(x,d) \geq 0 \quad \forall\, d$ is rather sharp.
The resulting theory works without Lipschitz assumption, but it requires $f^{\cdot}$ to be convex in d, in order to define an associated subdifferential, which happens to enjoy some nice properties. However it does not seem to fit properly to requirements for the development of numerical algorithms: it is not easy to find a Penot-critical point numerically, without additional hypotheses.

We conclude this section with a big question mark: the peridifferential seems too large, although it seems as small as possible (see 1.2.3; upper-semi-continuity, in particular, is a key for numerical algorithms). In fact, for numerical algorithms, the peridifferential itself is too small and needs to be further enlarged so as to contain the gradients of f in a fixed neighborhood of x, and not only their limits as this neighborhood shrinks to x. Therefore it is not an easy matter to find a fully convenient extension of the derivative.

2. SOME PARTICULARIZATIONS

In the previous sections, we have seen that $\partial f(x)$ is too large, but cannot be easily reduced. Another way to remedy its defect is to make additional assumptions on f, so as to eliminate the nasty counter examples. Thus, in this section, we will consider various subclasses of locally Lipschitz functions, and use them as further illustrations of the Clarke theory.

2.1. Directional derivatives

Consider the differential quotient of f at x in the direction d:

$$[\ f(x+td) - f(x)\ ]\ /\ t \quad , \quad \text{defined for } t > 0$$

Because it is bounded, it has cluster points when $t \downarrow 0$, which are all lower than $f^o(x,d)$ (set $x' = x$ in (1)).

Now, symmetrically to (1), define

$$f_o(x,d) = \liminf \{\ [f(x'+td)-f(x')]/t \ :\ t \downarrow 0,\ x' \to x\ \}$$

Then we have

$$f_o(x,d) = (-f)^o(x,d) = \min \{\ (g,d)\ :\ g \in \partial f(x)\ \}$$

and all the cluster points of the differential quotient are between f_o and f^o so, from convexity of $\partial f(x)$, each of them is equal to some (g,d), $g \in \partial f(x)$.

N.B. Clarke's generalized derivative is not a symmetric concept, and it may seem more sensible (see Penot's subdifferential) to take the lim inf in (1).
The reason is that f^o is convex in d, and this is important (minimization itself is not a symmetric concept!).

We say that f has a directional derivative $f'(x,d)$ at x in the direction d if all these cluster points are equal, i.e. if

$$[f(x+td)-f(x)]/t \text{ has a limit } f'(x,d) \text{ when } t \downarrow 0$$

and then, there exists $g(d) \in \partial f(x)$ with $f'(x,d) = (g(d),d)$. Of course, $g(d)$ usually depends on d; if it does not, f is actually Gateaux differentiable ($f'(x,d)$ is then linear in d, in particular $f'(x,-d) = -f'(x,d)$).

Existence of directional derivatives may sound rather restrictive, yet it holds for all the classes of Lipschitz functions that are commonly considered. This is rather fortunate because a numerical algorithm (especially a line search) will probably run into serious trouble if the differential quotient varies unpredictably for small stepsizes.

2.2. Semi-smoothness

A Lipschitz function f is said semi-smooth [17] if, given x and d:

(4) For any sequences $t_k \downarrow 0$, $d_k \to d$, $g_k \in \partial f(x+t_k d_k)$, the number (g_k,d) tends to one limit.

In this definition, the point $x+t_k d_k$ tends to x following a curve tangent to d. The sequence g_k may have several cluster points (unless f has a gradient at x), and semi smoothness requires that its projection onto d has only one cluster point. In other words, all the cluster points of g_k have to lie in some hyperplane orthogonal to d.

This definition may sound somewhat artificial. However, it is motivated by numerical considerations, and we proceed now to explain where it comes from.

To implement a certain class of numerical algorithms (the so called bundle methods [13]) one needs from f the following property [3]: given x and d

(5) $\limsup \{[f(x+td)-f(x)]/t - (g,d) : t \downarrow 0, g \in \partial f(x+td)\} \leq 0$

In words this means that, when $t \downarrow 0$, all the "derivatives" (g,d) (whichever g is chosen in $\partial f(x+td)$) are asymptotically as large as the differential quotient. This property is called <u>upper-semi-differentiability</u> by Bihain [3].

Now, the lim sup of a sum is smaller than the sum of lim sups, so (5) is true if

$$(6)\qquad \begin{aligned} &\limsup \{ [f(x+td)-f(x)]/t \ : \ t \downarrow 0 \} \leq \\ &\leq \liminf \{ (g,d) \ : \ g \in \partial f(x+sd),\ s \downarrow 0 \} \end{aligned}$$

The reason for introducing (6) is that it is much more tractable than (5). In fact (6) implies that the differential quotient has exactly one limit (i.e. the directional derivative exists), and that the $\leq$ in (6) holds as an equality (the reason is that, for each t, the mean value theorem implies the existence of an $s < t$ such that $[f(x+td)-f(x)]/t = (g(x+sd),d)$). In other words, (6) is equivalent to assume that f has a directional derivative and

$$(7)\qquad f'(x,d) = \liminf \{ (g,d) \ : \ g \in \partial f(x+sd),\ s \downarrow 0 \}$$

Functions satisfying (6) or (7) are called <u>weakly upper-semi-smooth</u> by Mifflin. Weak upper-semi-smoothness is more restrictive than upper-semi-differentiability. In particular (6) implies the existence of directional derivative, while (5) does not. It is interesting to note that the existence of a directional derivatives is just what makes the difference: if a function satisfies (5) and has directional derivatives, then it satisfies (6) (this is again due to the mean value theorem).

Now, if the sequence on the right of (7) is allowed to have several cluster points, it is a matter of chance if they are all larger than the directional derivative. For example, if f satisfies (7), $-f$ does not. Therefore, there is no big loss of

generality in requiring that this sequence has only one cluster point.

This defines a <u>weakly-semi-smooth</u> function, a function which satisfies

$t_k \downarrow 0$ and $g_k \in \partial f(x+t_k d)$ imply that (g_k, d) has a limit

(and this limit is the directional derivative, which has to exist).

This definition is rather natural: it simply means that the derivative <u>does not oscillate</u> along line segments.

Finally, passing from weak semi-smoothness (derivative continuous on line segments) to semi-smoothness (derivative continuous on curves having a tangent) yields stability under chaining: a semi smooth composition of semi smooth functions is semi smooth, i.e.:

(8) Let $f_1, \ldots, f_m$ be semi smooth functions. Let E be a semi smooth function on R^m. Then the function $f(x) = E(f_1(x), \ldots, f_m(x))$ is semi-smooth.

but this result is false for weakly semi smooth functions.

In fact the result that can be proved is stronger than (8). We know (see section 1.2.5) that, for a composite function, one can compute the product of gradients to obtain a certain set $N(x)$ whose convex hull contains $\partial f(x)$. The proof of [17, Theorem 5] really says: for any sequences t_k, d_k as defined in (4), and for any sequence $g_k \in N(x+t_k d_k)$, (g_k, d) is convergent.
This is useful in applications, and stronger than (4) because g_k may not be in $\partial f(x+t_k d_k)$.

Finally note that one could define symmetrically lower-

semi-differentiable and lower-semi-smooth functions (although they seem to be of little use) and also semi-differentiable functions.

2.3. Quasi-differentiability and subdifferential regularity.

Motivated by the derivation of optimality conditions, Pschenichnyi [20] has defined the class of quasi-differentiable functions: those functions which have directional derivatives, and for which there exists a set, say $\delta f(x)$, such that

$$f'(x,d) = \max \{ (g,d) \ : \ g \in \delta f(x) \}$$

There are essentially two motivations for this definition: first f' is convex in d; second if $0 \in \delta f(x)$, then $f'(x,d) \geq 0$ for any d, so x really looks like a local minimum (in this class of functions, any descent direction is actually robust).

However, this definition it not easy to understand, because it simultaneously defines the subdifferential and the class of functions. Hence we find it convenient to restrict the class, assuming that $\delta f(x)$ is really the peridifferential. We say that the locally Lipschitz function f is quasi-differentiable in Clarke sense, or subdifferentially regular [22], if

$$f'(x,d) \text{ exists and is equal to } f^{o}(x,d)$$

Thus, subdifferential regularity is motivated by theoretical reasons while upper-semi-smoothness is motivated by numerical algorithms. It is interesting to note that these two definitions are antagonistic because requiring $f'(x,d) = f^{o}(x,d)$ is the best way to ensure that (7) is violated: if $t \downarrow 0$ and $g \in \partial f(x+td)$, any cluster point of (g,d) will be below $f^{o}(x,d) = f'(x,d)$. On the contrary, a function satisfying $f' = f_{o}$ (thus having concave directional derivatives!) would be automatically

upper-semi-smooth. This illustrates some of the difficulties that can be encountered to match theory and practice.

Thus, semi-smoothness and subdifferential regularity are not much related to each other (except that subdifferential regularity has a tendency to yield "weak-lower-semi-smoothness"). In one dimension, it is easy to construct a subdifferentially regular function which is not semi-smooth: take for $x \geq 0$ a function $f(x)$ between $-2x^2$ and $-x^2$, whose slopes alternate infinitely often on -1 and 0. Finding a semi-smooth function which is not subdifferentially regular is also possible, although it requires to define f outside a single direction.

Finally we mention an interesting property: suppose x is such that $0 \in \text{int}\ \partial f(x)$ (a property slightly stronger than mere stationarity). Then $\partial f(x)$ contains a ball of nonzero radius, so it is easy to show that $f^o(x,d) > 0$ for any non zero d. If, in addition, f is subdifferentially regular, then x is a <u>strict</u> local minimum. Hence, in some rather general situations, there exist first order sufficient optimality conditions. This, of course, is only possible in nonsmooth optimization; in the smooth case, sufficient optimality conditions can only be given by a second order analysis.

2.4. Max functions

Consider $h(x,y)$, a function depending on two (groups of) variables. Suppose it is continuous with respect to $y \in Y$, where Y is a compact set. Then the function

$$f(x) = \max \{ h(x,y) \ : \ y \in Y \}$$

is called a <u>max function</u>. Its regularity will depend on the regularity assumptions of h with respect to x.

N.B. The above hypotheses with respect to y are mainly present to ensure that the max exists. Symmetrically to max functions, we have the min functions, but they are not all that interesting; max functions are much more often encountered in practice: in penalty, duality and all sorts of decomposition, in game theory etc...

For each x we define the set of optimal y's:

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

and we have the following sequence of results:

2.4.1. If h is locally Lipschitz with respect to x, uniformly for $y \in Y$, then f is <u>locally Lipschitz</u> (with respect to x). Thus, f has a peridifferential, which can in general be expressed in terms of the peridifferential of the underlying funtion h [5], [17], namely:

(9) $\left\{ \begin{array}{l} \text{If, in addition, } \partial_x h \text{ is upper semi continuous} \\ \text{jointly with respect to x and y, then} \\ \partial f(x) = \operatorname{conv} \{ \partial_x h(x,y) \ : \ y \in Y(x) \}. \end{array} \right.$

In other words, to generate subgradients of f at x, we just have to compute the subgradients of h at those y's which are optimal at x.

2.4.2. If, in addition, $\nabla_x h$ exists and is continuous jointly with respect to x and y, we say that f is <u>lower C^1</u> [23]. A direct consequence of (9) is

$\left\{ \begin{array}{l} \text{If f is a lower } C^1 \text{ function, then} \\ f^o(x,d) = \max \{ (\nabla_x h(x,y),d) \ : \ y \in Y(x) \} \end{array} \right.$

a result which seems due to Danskin [7]. From this, we obtain [8], [17]:

$\left\{ \text{A lower } C^1 \text{ function is subdifferentially regular.} \right.$

It is not difficult to see why: because f is obtained by maximizing h, we have

$f(x+td) \geq h(x+td,y) = h(x,y) + t\,(\nabla_x h(x,y),d) + o(t)$

and this is true for any y. In particular, for any y in Y(x) we have $h(x,y) = f(x)$, so the differential quotient of f is (to first order) greater than $f^o(x,d)$.

Another property, more delicate to prove [8], [17], is:

{ A lower C^1 function is semi-smooth.

(we will give in 2.5.2 an explanation of this result).

This implies in particular that the maximum of a finite number of C^1 functions (Y finite) is semi-smooth, i.e. the function from R^m to R defined by $E(y_1,\ldots,y_m) = \max\{y_j : j=1,\ldots,m\}$ is semi-smooth.

From (4), we see also that, if f is semi smooth, −f is also semi smooth. Therefore an upper C^1 function is semi-smooth (an upper C^1 function being a min of C^1 functions or equivalently minus a lower C^1 funtion).

As a result, if $h_i(x,y)$ are m functions having a gradient with respect to x jointly continuous in x and y, then the functions $f_i(x) = \max_y h_i(x,y)$ are semi-smooth and because of (8), the function $f(x) = \min f_i(x)$ is semi smooth. However this property is not conserved for a function such as

$$f(x) = \min_{z\in Z} \max_{y\in Y} h(x,y,z)$$

when Z is also an infinite set. Counter examples can be found (Polak).

2.4.3. A particular case is when Y(x) is a singleton, y(x) say. This happens for example if h is strictly concave with respect to y.

Then from (9) f has a gradient at x, which is just $\nabla_x h(x,y(x))$. This may sound paradoxical because we have a priori

$$\nabla f(x) = \nabla_x h(x,y(x)) + \nabla_y h(x,y(x))\, y'(x)$$

(note that this expansion is very unformal because we have made no assumption on the regularity of y(x) - although it is continuous because the mapping Y(x) is upper semi continuous). In fact the second term is zero because transversality conditions imply that, at the optimal y, the gradient $\nabla_y h$ is orthogonal to any feasible move dy. In particular, if y(x) is unconstrained, its optimality implies $\nabla_y h = 0$.

Thus failure to differentiability of a lower C^1 function implies that Y(x) is not a singleton. Heuristically this must happen "almost never", which explains partly the initial Rademacher theorem, at least for this class of functions.

For example, if Y is a polyhedron and h linear in y (its coefficients depending on x) Y(x) is usually a single extreme point, except for those x's for which the level hyperplanes of h are parallel to a face of Y. Then $\partial f(x)$ is the convex hull of the image by $\nabla_x h$ of that face; it may be rather complicated except when h is also linear in x, in which case ∂f is a polyhedron: the image under the linear transformation $\nabla_x h$ of the optimal polyhedron Y(x).

On the other hand, if h is really nonlinear in y, then it can have several local maxima and differentiability fails only when several such local maxima have the same (maximal) h-value. Still Y(x) is usually a finite set, which implies that $\partial f(x)$ is usually a polyhedron, as the convex hull of a finite number of points (we can say from these observations that the structure of f is somewhat simpler when h is nonlinear in y).

2.5. Special examples of max-functions

Max functions form a very important class in optimization, so it is worth examining them further by studying separately some of their subclasses.

2.5.1. Convex functions. One knows that a convex function (when it is finite everywhere) is locally Lipschitz, that the subdifferential of convex analysis is defined by

$$\partial f(x) = \{ g \ : \ f(y) \geq f(x) + (g, y-x) \quad \forall y \}$$

It is also known [21] that this subdifferential is the peridifferential. Inverting the role of x and y, we can write

$$f(x) = \max \{ f(y) + (g, x-y) \ : \quad \forall y, \quad \forall g \in \partial f(y) \}$$

i.e. a convex function is a max of linear functions (this is somewhat unformal because the set of $\{y,g\}$ over which one maximizes is not compact). However, convex functions are so important that we will study them separately in Section 3.

2.5.2. The case where Y is finite. When Y is finite in 2.4, we more conveniently write f in the form

$$f(x) = \max \{ f_i(x) \ : \ i = 1, \ldots, m \}$$

where each f_i is smooth. There is no big difference with the general case 2.4, but things are now easier to understand.

The space is divided into regions R_i in which $f_i = f$. If x is interior to a particular R_i, then $f = f_i$ in a neighborhood included in R_i, so the gradient of f is that of f_i and thus is continuous.

The intersection of two regions R_i and R_j is a portion of the smooth surface defined by $f_i = f_j$. Its tangent plane at x is the set of $\{x+d\}$ where d satisfies

$$f_i(x) + (\nabla f_i(x), d) = f_j(x) + (\nabla f_j(x), d)$$

i.e.

$$(\nabla f_i(x), d) = (\nabla f_j(x), d)$$

In other words the tangent plane is the set of increments that make the same scalar product with all active gradients.

A point where several f_i are active is called a <u>kink</u>. To tend to a kink, one can stay in either R_i or R_j and this explains (9), which becomes

$$M(x) \subset \{\nabla f_i(x) \ : \ f_i(x) = f(x)\}$$

(note that it is an inclusion rather than an equality because there may be at x an active f_i which is not active elsewhere: see for example max $\{x, 0, -x\}$).

Thus we see why such functions are semi-smooth: when a point of the form x+td tends to x, either the curve x+td is contained in a single region, then no problem, or this curve is tangent to a kinky surface, but then $(\nabla f_i(x), d)$ tends to the same value, independent of the i defining the surface.

One can consider the <u>finite minimax problem</u>, which consists in minimizing this type of f. It is clearly equivalent to

$$\begin{cases} \min v \\ v \geq f_i(x) \quad i=1,\dots,m \end{cases}$$

where $v \in R$ is an extra variable. The John conditions for a local minimum are

$$\begin{cases} \exists\, u_i \geq 0 \ i = 0,\dots,m \quad u_i \text{ not all zero, such that} \\ \Sigma_{i=1}^m \ u_i = u_0 \quad \text{(stationarity in v)} \\ \Sigma_{i=1}^m \ u_i \ \nabla f_i(x) = 0 \quad \text{(stationarity in x)} \end{cases}$$

$$\begin{cases} v \geq f_i(x) & \text{(feasibility)} \\ u_i\,[v-f_i(x)] = 0 & \text{(transversality condition).} \end{cases}$$

They imply $u_0 > 0$, i.e. every local optimum is qualified and there hold the Kuhn-Tucker conditions as well. If one takes $u_0 = 1$, one obtains the standard Clarke's optimality condition (note that $f_i(x) < v = f(x)$ implies $u_i = 0$).

When the f_i's are also C^2, one can compute a "second order directional derivative" defined by

$$2\,f''(x,d) = \lim \{\, [f(x+td)-f(x)-tf'(x,d)]/t^2 \;:\; t \downarrow 0 \,\}$$

Call

$$I(x) = \{\, i \;:\; f_i(x) = f(x) \,\}$$

so $f'(x,d) = \max \{\, (f_i(x),d) \;:\; i \in I(x) \,\}$. Call also

$$I(x,d) = \{\, i \in I(x) \;:\; (\nabla f_i(x),d) = f'(x,d) \,\}$$

Then we have

$$f''(x,d) = \max \{\, (\nabla^2 f_i(x)d,d) \;:\; i \in I(x,d) \,\}$$

The reason is that, for t small enough we may suppose $I(x+td) \subset I(x,d)$ and then

$$2[f(x+td)-f(x)-tf'(x,d)]/t^2 = \max \{(\nabla^2 f_i(x)d,d) + o(t^2)/t^2\}$$

Based on this second order directional derivative, one can also define second order optimality conditions for the finite minimax problem.

2.5.3. Sums of max functions. The minimax problem is clearly that of the best ℓ_∞ approximation. The problem of best ℓ_1 approximation consists in minimizing a function having the form

$$f(x) = \Sigma_{i=1}^{m} |f_i(x)|$$

This is a sum of locally Lipschitz functions. For each i, one has (see 2.4.1)

$$\partial|f_i|(x) = \begin{cases} g_i = \text{sign}\,[f_i(x)]\,\nabla f_i(x) & \text{or} \\ [-1,+1]\,\nabla f_i(x) & \text{if } f_i(x) = 0 \end{cases}$$

Although it is not true in general (see 1.2.4) we do have here that $\partial f(x) = \Sigma\ \partial|f_i|(x)$ (see [10]); this is due to the combination of two properties: each f_i is continuously differentiable and the function from R^m to R: $E(y) = |y|_1$ is convex.

Therefore, calling I the set of i such that $f_i(x) = 0$, and $\bar{I}$ its complement,

$$\partial f(x) = \Sigma_{\bar{I}}\ g_i + \Sigma_I\ u_i\ \nabla f_i(x)$$

where each u_i describes the segment $[-1,+1]$.

Note that the structure of ∂f is much more complex than in the finite minimax case: in the latter, ∂f had at most m extreme points; here it may have 2^m.

An optimality condition can easily be derived by expressing that a value of $\{u_i : i \in I\}$ gives the zero vector. This condition can be recovered by restating the finite ℓ_1 problem as

$$\begin{cases} \min\ \Sigma\ v_i \\ f_i(x) \leq v_i \\ -f_i(x) \leq v_i \end{cases}$$

which has now m extra variables (instead of 1 in the finite minimax problem).

2.5.4. Piecewise linear functions. If we suppose the f_i's linear in 2.5.2 above, we obtain functions of the form

$$f(x) = \max \{ (g_i, x) + b_i \quad : \quad i = 1, \ldots, m \}$$

where the b_i's are real numbers, the g_i's are vectors in R^n. Such functions are called <u>polyhedral</u>, or piecewise linear convex (their epigraph is a convex polyhedron). A min of a finite number of linear functions would be called piecewise linear concave.

For these functions, the subdifferential is still a convex polyhedron, but whose extreme points are taken from a finite set of m elements. To each kinky surface corresponds a fixed polyhedron, having at most as many extreme points as the number of functions for which the maximum is obtained. The kinks now form hyperplanes defined by linear equations such as

$$(g_i, x) + b_i = (g_j, x) + b_j$$

For a fixed d, the directional derivative $f'(x,d)$ is piecewise constant - and discontinuous! - in x.

2.6. <u>A trivial specialization</u>

We give a last example for pedagogical purpose, as it illustrates the danger of defining at the same time a class of functions and their subdifferential (see 2.3). Norkin [18] defines <u>generalized differentiable</u> functions, which satisfy

> There exists a closed, convex and upper semi continuous mapping δf such that there exists $g \in \delta f(y)$ satisfying $f(y) = f(x) + (g, y-x) + o(x,y,g)$ for all x and y, where $o/|y-x|$ tends to 0 when $y \to x$, uniformly with respect to $g \in \delta f(y)$.

He shows that these functions are locally Lipschitz, and then he shows that all the classes 2.1 to 2.5 (plus many others) are generalized differentiable. He also gives an algorithm to find a stationary point.

Now take a locally Lipschitz functions; its peridifferential is locally bounded by a constant, say L. For any x define $\delta f(x) = \{g : |g| \leq L\}$, i.e. define the Norkin subdifferential to be constantly the ball of radius L. For any x and y, the mean value theorem says that

$$\exists\, g \in \delta f(x): f(y) = f(x) + (g, y-x)$$

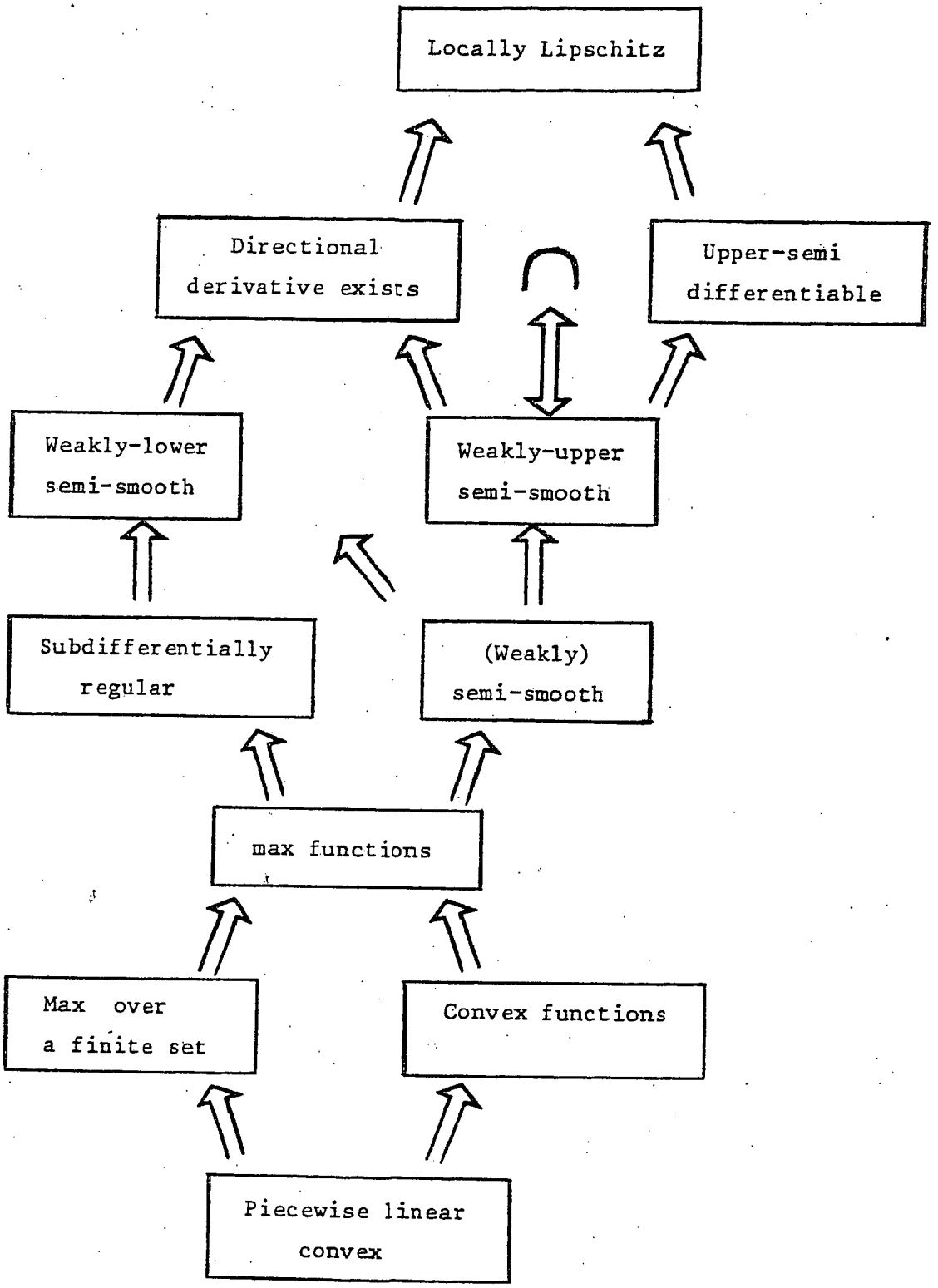
i.e. any Lipschitz function is generalized differentiable (take o = 0 in the above definition). The new definition brings nothing. Of course, stationarity does not tell much: every point is stationary in the present sense.

As mentioned in 2.3, it is not reasonable to let $\delta f(x)$ be any set. Suppose now that we take ∂f as the Norkin subdifferential, to define (as we did in 2.3) a generalized differentiable function in Clarke sense:

$$\left\{\begin{array}{l} \exists\, g \in \partial f(y): f(y) = f(x) + (g, y-x) + o(x,y,g) \\ \text{where } o(x,y,g)/|y-x| \to 0 \text{ when } y \to x, \\ \text{uniformly with respect to } g \in \partial f(y) \end{array}\right.$$

For such a function, particularizing y to x + td with d fixed and $t \downarrow 0$, we have that $[f(x+td)-f(x)]/t - (g,d)$ tends to 0, so a generalized differentiable function in Clarke sense is upper semi differentiable (we have even more: replace lim sup by lim and $\leq$ by = in (5)).

To conclude this section, we give a flow-chart showing how the classes we have considered are related to each other.


Locally Lipschitz
Directional derivative exists
Upper-semi differentiable
Weakly-lower semi-smooth
Weakly-upper semi-smooth
Subdifferentially regular
(Weakly) semi-smooth
max functions
Max over a finite set
Convex functions
Piecewise linear convex

3. THE CONVEX CASE - APPROXIMATE SUBDIFFERENTIALS

In this section f is assumed convex and we refer to [21] for an extensive study of such functions. For simplicity, we will suppose that f is finite everywhere, in which case it is locally Lipschitz. One can also consider generalized convex functions which can take on value $+\infty$ and which are Lipschitz on the relative interior of their effective domain.

Convexity implies that f is a max function (actually a max of linear functions) hence it is subdifferentially regular and semi-smooth.

It is known that the differential quotient is an increasing function of t, so the now familiar formula (2) gives here

$$(10)\quad \inf_t\ [f(x+td)-f(x)]/t = f'(x,d) = f^{o}(x,d) = \max_g\ (g,d)$$

and $\partial f(x) = \{\ g\ :\ f(y) \geq f(x) + (g,y-x)\ \}$.

Note that symmetrically, we would have for a concave function $\sup_t \ldots = f'(x,d) = f_o(x,d) = \min_g \ldots$ and the subdifferential (in fact the superdifferential) would be $\{\ g\ :\ f(y) \leq f(x) + (g,y-x)\ \}$.

Thus, in the convex case, $f^{o}(x,d) \geq 0$ implies that $f(x+td) \geq f(x)\ \ \forall\ t > 0$. There is no ambiguity in saying that d is a descent direction, and any descent direction is robust. Necessary conditions for a minimum are also sufficient, and any local minimum is also global. The set of minima is convex.

A difficulty with the peridifferential is that, for numerical purposes, it is too small, whereas it is too large for theoretical purposes - see the end of section 1.6. This difficulty is still present in the convex case.

In the convex case however, one can define an enlargement of the subdifferential, which is very suitable for numerical algorithms.

3.1. <u>The approximate subdifferential</u> - Definition

For the convex function f, and $\varepsilon \geq 0$ we define the ε-subdifferential:

$$(11) \qquad \partial_\varepsilon f(x) = \{ g \ : \ f(y) \geq f(x) - \varepsilon + (g, y-x) \quad \forall y \in R^n \}$$

We will call <u>ε-subgradients</u> the elements of $\partial_\varepsilon f(x)$. It is easy to see that this set increases when ε increases, while it just reduces to the ordinary subdifferential $\partial f(x)$ when $\varepsilon = 0$.

This set has an interpretation in the graph space, i.e. the set of couples $\{y,z\}$ in $R^n \times R$.

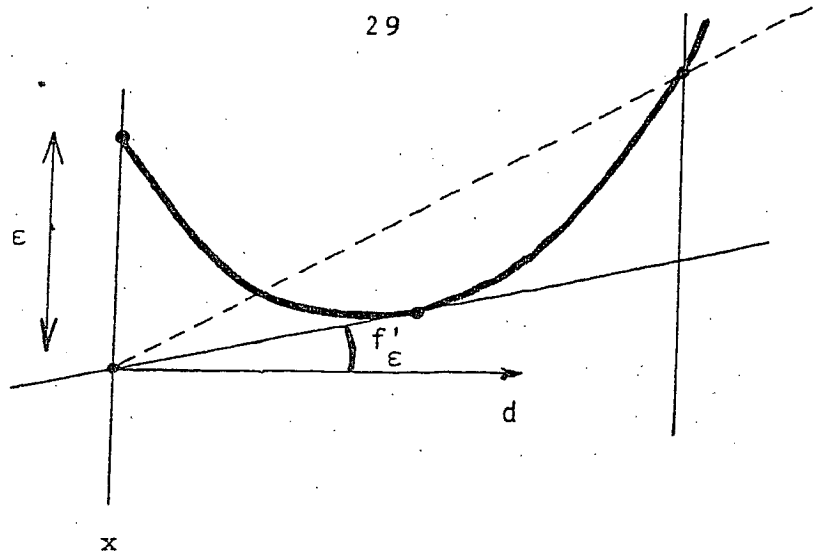
In this space, the equation $z = f(y)$ defines a surface, the graph of f and, for each vector g, the equation $z = f(x) - \varepsilon + (g, y-x)$ defines a non vertical hyperplane passing through the point $\{ x , f(x) - \varepsilon \}$.

To say that $g \in \partial_\varepsilon f(x)$ is to say that this hyperplane is below the graph of f.

Now restrict the study to points $y = x+td$, $t \geq 0$. The differential quotient $[f(x+td)-(f(x)-\varepsilon)]/t$ is the slope of the straight line joining $\{x+td, f(x+td)\}$ to $\{x, f(x)-\varepsilon\}$.

When t varies, this slope has a minimum which is the maximal slope passing through $\{x, f(x)-\varepsilon\}$ and staying below the graph of f. This means that the following formula holds:

$$(12) \qquad \left\{ \begin{array}{l} \inf \{ [f(x+td)-f(x)+\varepsilon]/t \ : \ t > 0 \} = \\ = \max \{ (g,d) \ : \ g \in \partial_\varepsilon f(x) \} = f'_\varepsilon(x,d) \end{array} \right.$$



and this defines $f'_\varepsilon(x,d)$, the <u>ε-directional derivative</u>, which is also increasing when ε increases, and reduces to the ordinary directional derivative when $\varepsilon = 0$.

3.2. <u>Some examples</u>

3.2.1. Linear and positively homogeneous functions. If f is linear, the geometric interpretaion of section 3.1 shows that $\partial_\varepsilon f(x)$ is just the gradient of f, independent of ε and of x.

In fact we can say more: suppose that f is <u>positively homogeneous</u> at x in the direction d, i.e.

$$f(x+td) = f(x) + kt \qquad \forall\ t > 0$$

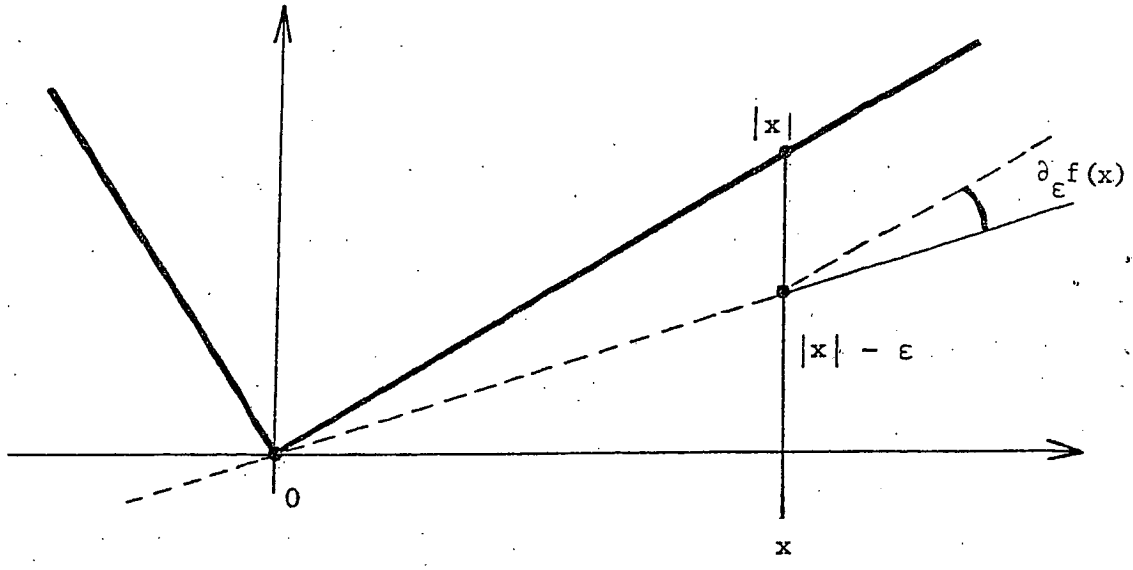
(k, a coefficient depending on d, is actually $f'(x,d)$). Then, for any ε, $f'_\varepsilon(x,d)$ is clearly $f'(x,d)$ i.e. $f'_\varepsilon(x,d) = (g,d)$ for some g in $\partial f(x)$. If the above property holds for any d, then $\partial_\varepsilon f(x) = \partial f(x)$ for any ε.

3.2.2. For n=1 consider $f(x) = |x|$. We have that

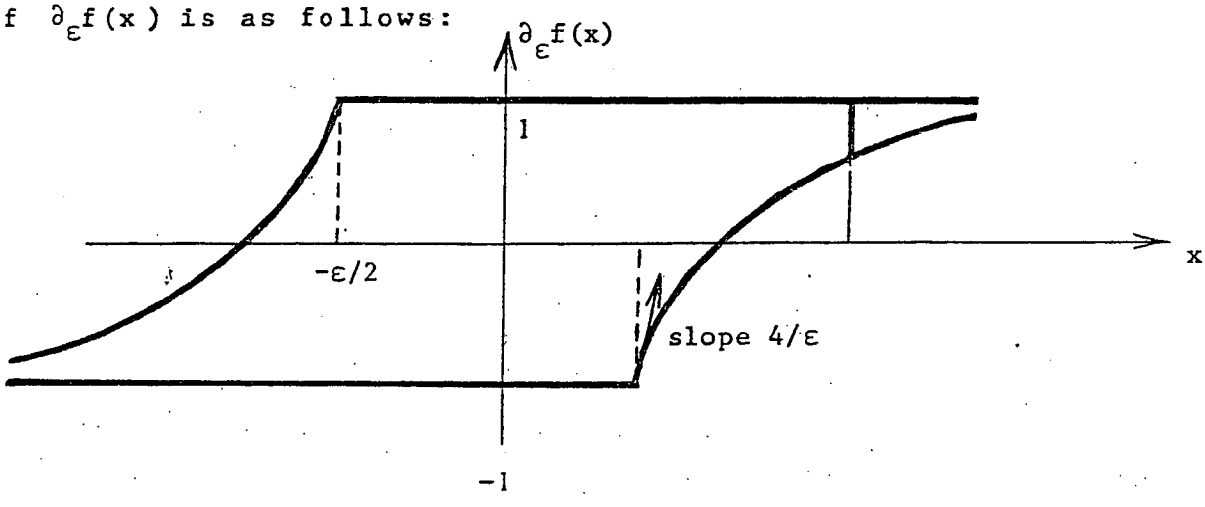
$$\partial f(x) = \begin{cases} 1 & \text{if } x > 0 \\ [-1,+1] & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

and $\partial_\varepsilon f(0) = [-1,+1] = \partial f(0)$ for any $\varepsilon \geq 0$ (this is in conformity with 3.2.1). For $x \neq 0$, $\partial f(x)$ is a singleton but $\partial_\varepsilon f(x)$ (which is a convex set, i.e. a segment) has non zero length. The following

picture illustrates the graph of f and the two extreme slopes of the segment $\partial_\varepsilon f(x)$.



Using definition (11), easy calculations show that the graph of $\partial_\varepsilon f(x)$ is as follows:



For $|d| = 1$, $f'_\varepsilon(x,d)$ is one of the two end points of the segment $\partial_\varepsilon f(x)$. When $x \geq \varepsilon/2$ for example, $f'(x,-1) = \varepsilon/x - 1$ and $f'(x,1) = 1$.

3.2.3. <u>Quadratic functions</u>. Let A be a symmetric positive semi definite matrix, and $f(x) = 1/2(x,Ax) + (b,x)$. Plugging the value of f in the definition (11), it is just a matter of algebra to see that

(13) $$\partial_\varepsilon f(x) = \{ Ay + b \ : \ 1/2\ (A(y-x),y-x) \leq \varepsilon \}$$

which shows that $\partial_\varepsilon f(x)$ is the set of gradients of f in a neighborhood of x; this neighborhood contains the kernel of A (translated at x). If A is non singular, it defines a metric, whose ball of radius ε is the neighborhood in question. Then (13) can equivalently be written

$$\partial_\varepsilon f(x) = \{ g \ : \ 1/2 \ (A^{-1}(g-\nabla f(x)), g-\nabla f(x)) \leq \varepsilon \}$$

which is also a neighborhood of $\nabla f(x)$: the ball of radius ε for the metric of A^{-1}. Finally, using the definition (12) shows that

$$f'_\varepsilon(x,d) = f'(x,d) + \sqrt{2\,\varepsilon\,(d,Ad)}$$

3.2.4. Indicator functions. Although it is not locally Lipschitz, consider the function

$$\delta(x) = \begin{cases} 0 & \text{if} \quad (a,x) \leq b \\ +\infty & \text{if} \quad (a,x) > b \end{cases}$$

where $a \in R^n$, $b \in R$ are given. This function is called the <u>indicator</u> function of the set $\{x : (a,x) \leq b\}$ (which may be interpreted as a feasible set in some linearly constrained optimization problem).

Formula (11), when specialized to the present situation, reduces to requiring

$$0 \geq \delta(x) + (g,y-x) - \varepsilon \qquad \forall\ y \text{ such that } (a,y) \leq b$$

This inequality is impossible to obtain if $(a,x) > b$ (then $\delta(x) = +\infty$). For x feasible, we decompose $y - x$ as $ka + v$, $k \in R$ and v orthogonal to a; the feasible y's are described by $k \leq [b-(a,x)]/|a|^2$. Then $\partial_\varepsilon\, \delta(x)$ is the set of g such that, for any v orthogonal to a and for any $k \leq [b-(a,x)]/|a|^2$, $(g,ka+v) \leq \varepsilon$. One can see that g has to be a positive multiple of a and in summary

$$\partial_\varepsilon \delta(x) = \begin{cases} \emptyset & \text{if } (a,x) > b \\ [0,k_\varepsilon]\, a & \text{if } (a,x) \leq b \end{cases}$$

where $k_\varepsilon = \varepsilon/[b-(a,x)]$ (with the understanding that if $(a,x) = b$, $\partial_\varepsilon \delta(x)$ is the half line ka, $k \geq 0$).

3.3. Fundamental properties

3.3.1. Just as the ordinary subdifferential, $\partial_\varepsilon f(x)$ is a closed, bounded, convex set. It is also locally bounded, of closed graph, and hence upper semi continuous (if f is no longer locally Lipschitz, as in example 3.2.4 above, these properties hold for $x \in$ int dom f and everything can happen when $x \notin$ int dom f; if $x \notin$ dom f, $\partial_\varepsilon f(x)$ is void). Concerning boundedness we have the following statement:

(14) $\begin{cases} \text{Let } L(r) \text{ be a Lipschitz constant on the r-ball around x.} \\ \text{Then, for any } g \in \partial_\varepsilon f(x),\ |g| \leq L(r) + \varepsilon/r \end{cases}$

(to prove it, take in (11) $y = x + tg$, $t|g| = r$).

3.3.2. As suggested in exampe 3.2.2 above, the mapping $\partial_\varepsilon f$ is continuous i.e. contrary to ∂f, it is also lower semi continuous. Actually we have more: it is Lipschitz continuous, namely: for x and y varying in a bounded set, there exists L such that

$$D\,(\partial_\varepsilon f(x)\,,\ \partial_\varepsilon f(y)\,) \leq L/\varepsilon\ |x-y|$$

where D denotes the Hausdorff distance between two sets (it is difficult to give a father to this very important result, which has several simple proofs).

N.B. The Hausdorff distance is defined as follows: let

$$d(a,B) = \inf\{\ |a-b| \ :\ b \in B\ \}$$

be the distance between a point a and a set B. Then the number

$$D_{AB} = \sup\ \{d(a,B) : a \in A\}$$

represents how much a given set A is larger than B. In particular it is 0 if $A \subset cl\, B$. The Hausdorff distance between the two sets A and B is then defined by

$$D(A,B) = \max \{ D_{AB} , D_{BA} \}.$$

3.3.3. We recall the definition of <u>conjugacy</u>: given a function f (not necessarily convex) one can form another function f*, also from R^n to R (but now R^n is interpreted as its own dual) by

$$f^*(g) = \sup \{ (g,x) - f(x) \ : \ x \in R^n \}$$

By definition, for any pair x and g, we have $f(x) + f^*(g) \geq (g,x)$. As a sup of linear function (compare with Section 2.4) f* is convex and its subdifferential is the convex hull of the optimal points, i.e. those for which there holds $f(x) + f^*(g) = (g,x)$. In particular, if f is convex, this optimal set is convex and it is equivalent to say

$$\begin{cases} x \in f^*(g) \\ g \in f(x) \\ f(x) + f^*(g) = (g,x) \end{cases}$$

Note that f* is not necessarily finite (hence locally Lipschitz) although it is so if f is coercive i.e. $f(x)/|x| \to +\infty$ if $|x| \to \infty$.

Now from (11), we see that $g \in \partial_\varepsilon f(x)$ if and only if

$$(g,y) - f(y) \leq (g,x) - f(x) + \varepsilon \qquad \forall\, y$$

hence $\partial_\varepsilon f(x)$ can equivalently be defined by

$$\partial_\varepsilon f(x) = \{ g \ : \ f(x) + f^*(g) \leq (g,x) + \varepsilon \}.$$

i.e. it can be described by a single convex inequality, for which the Slater condition holds (provided $\partial f(x)$ is not void).

3.3.4. Because one knows the conjugate of a number of composite functions, the above property allows to describe approximate subdifferentials of <u>composite</u> functions. Several results along this line can be found in [11].

First, as can be seen from (11), if f is multiplied by k > 0, and if ε is <u>also</u> multiplied by k, then $\partial_\varepsilon f(x)$ is multiplied by k. Therefore

$$\partial_\varepsilon[kf](x) = k\ \partial_{\varepsilon/k} f(x).$$

Also, it is not difficult to see from (11) that $\partial_\alpha f_1(x) + \partial_\beta f_2(x) \subset \partial_{\alpha+\beta}[f_1+f_2](x)$

To obtain a reverse inclusion, one must consider all the pairs having the same sum and there holds the following formula:

$$\partial_\varepsilon[f_1+f_2](x) = \cup \{ \partial_\alpha f_1(x) + \partial_{\varepsilon-\alpha} f_2(x) \ : \alpha \in [0,\varepsilon] \}$$

Finally for a max function: let h(x,z) be convex in x and consider $f(x) = \max_z h(x,z)$; for any z take $g_z \in \partial_x h(x,z)$, so

$$f(y) \geq h(y,z) \geq h(x,z) + (g_z, y-x)$$

Taking z such that $h(x,z) \geq f(x) - \varepsilon$, we obtain (11), which shows that the ε-subdifferential of f at x contains the subgradients of h, computed at ε-maximizers at x. However the converse inequality is far from being true and the correct formula is actually rather complicated [11], [25], as it involves sums over z of sets such as $\partial_{\varepsilon(z)/t(z)} h(x,z)$. Because the approximate subdifferential is constant for linear functions, this formula becomes much simpler in the piecewise linear case: let $f(x) = \max f_i(x)$, where each $f_i(x) = (g_i, x) + b_i$ is linear. Then $\partial_\varepsilon f(x)$ is the set of convex combinations $\Sigma\, u_i\, g_i$, where u describes the set

$$u_i \geq 0 \qquad \Sigma\, u_i = 1 \qquad \Sigma\, u_i\, [f(x) - f_i(x)] \leq \varepsilon$$

When $\varepsilon = 0$ we recover formula (9). But it is important to note that when $\varepsilon > 0$, $\partial_\varepsilon f(x)$ depends on the gradient of f_i for some i's which have nothing to do with the value of f at x.

3.4. Sets of ε-descent directions and ε-optimality conditions

Just as in section 1.4, we have the cone of ε-descent directions, i.e. the set of d such that $f'_\varepsilon(x,d) < 0$. From (12), d is an ε-descent direction if and only if

$$f(x+td) < f(x) - \varepsilon \qquad \text{for some } t > 0$$

(note that if $\varepsilon > 0$, this inequality is certainly not obtained for $t \downarrow 0$, as would be the case for $\varepsilon = 0$). As in secton 1.5, to say that $0 \in \partial_\varepsilon f(x)$ is to say that there does not exist any ε-descent direction, i.e. x minimizes f within ε, as is obvious by setting $g = 0$ in (11).

This approximate optimality condition can also be used in constrained optimization. Consider

$$\min f(x) \qquad x \in C$$

where C is a convex set. This is clearly equivalent to minimizing without constraints the function $f(x) + \delta(x)$, where δ is the indicator of C, as seen in Section 3.2.4. Knowing the formulae for the approximate subdifferential of an indicator and of a sum, we can get approximate optimality conditions for the above problem. In the simpler problem

$$\begin{cases} \min f(x) \\ (a_i, x) \leq b_i \qquad i = 1,\dots,m \end{cases}$$

C is a polyhedron, whose indicator is the sum of the indicators of each linear constraint and, using the results mentioned in Section 3.3.4, we obtain [25]:

In the above problem, a feasible point x satisfies
$f(y) \geq f(x) - \varepsilon$ for any feasible y
if and only if there exists α, $g \in \partial_\alpha f(x)$
and $u_1 \geq 0, \ldots, u_m \geq 0$ such that
$g + \Sigma u_i a_i = 0$
$\alpha + \Sigma u_i [b_i - (a_i, x)] \leq \varepsilon$

The first condition corresponds to stationarity (0 belongs to a certain approximate subdifferential of the Lagrange function, see the end of Section 1.5) and the second condition corresponds to complementary slackness, but it does not imply that $u_i = 0$ if the corresponding constraint is inactive !.

3.5. Approximate subdifferentials constructed from subgradients

The results of this section mainly come from [14]. Perhaps the most important property for numerics is that $\partial_\varepsilon f(x)$ "catches" the subgradients of a neighborhood of x, as suggested by (13). We do have in general that, if y is close enough to x, then any subgradient at y is an ε-subgradient at x (provided $\varepsilon \geq 0$). It is even possible to specify what exactly the expression "close enough" means, and the following formula holds:

(15) Let ε, α, x and y be given, and suppose that $g \in \partial_\alpha f(y)$
If $f(y) \geq f(x) + (g, y-x) - (\varepsilon - \alpha)$ then $g \in \partial_\varepsilon f(x)$

This result is really trivial: from $f(z) \geq f(y) + (g, z-y) - \alpha$ we deduce immediately

$$f(z) \geq f(x) + (g, z-x) + [f(y) - f(x) + (g, x-y)] - \alpha$$

and the claim is visible.

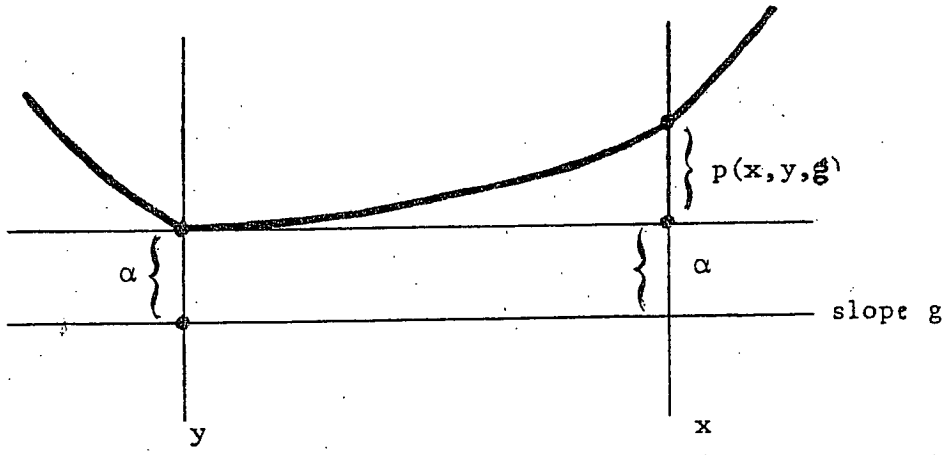
The main use of this result is when $\alpha = 0$, i.e. when $g \in \partial f(y)$. Note that if (15) does not hold with $\alpha = 0$, then g cannot be in $\partial_\varepsilon f(x)$ (apply the definition (11)). Hence we have the following corollary:

(16) $\begin{cases} \text{Let } g \in \partial f(y). \text{ Then } g \in \partial_\varepsilon f(x) \text{ if and only if} \\ f(y) \geq f(x) + (g, y-x) - \varepsilon \end{cases}$

An instructive interpretation of this result is as follows: associated with the triple x, y, g where $g \in \partial f(y)$, consider the number:

(17) $$p(x,y,g) = f(x) - [\, f(y) + (g, x-y) \,]$$

It is the error that is made at x when f is replaced by the linear function of slope g, passing through $\{y, f(y)\}$. If $g \in \partial f(y)$, this number is always positive. Then, $g \in \partial_\varepsilon f(x)$ if and only if $p(x,y,g) \leq \varepsilon$. This simple fact can be illustrated in one dimension:



Thus we see that a subgradient at y is also a p-subgradient at x, where p is $p(x,y,g)$ of (17). What (15) tells in addition to (16) is that, if one makes an error α when linearizing f at y, this same error is conserved at x.

Given x, and in view of the property of "absorption" mentioned at the beginning of this paragraph, we can consider the set of y whose subgradients are in $\partial_\varepsilon f(x)$. However, since $\partial f(y)$ may contain several points, this set is ambiguous and we consider two sets. One is

$$V_\varepsilon(x) = \{\, y \;:\; \partial f(y) \subset \partial_\varepsilon f(x) \,\}$$

which, because of (16) and (17), can equivalently be written

$$V_\varepsilon(x) = \{ y \ : \ p(x,y,g) \leq \varepsilon \quad \forall\, g \in \partial f(y) \}$$

Clearly it is a neighborhood of x. If L is a local Lipschitz constant for f, we have that $p(x,y,g) \leq 2L|y-x|$ therefore $V_\varepsilon(x)$ contains the ball of radius $\varepsilon/2L$.

The second set, which is larger, is

$$\bar{V}_\varepsilon(x) = \{ y \ : \ \partial f(y) \cap \partial_\varepsilon f(x) \neq \emptyset \}$$
$$= \{ y \ : \ p(x,y,g) \leq \varepsilon \text{ for some } g \in \partial f(y) \}$$

This notation is used because it can be shown that $\bar{V}_\varepsilon(x)$ is really the closure of $V_\varepsilon(x)$.

We observe that if f is C^1, $\partial f(y)$ is a singleton (p is uniquely defined by x and y) so $V_\varepsilon(x)$ and $\bar{V}_\varepsilon(x)$ coincide. If f is quadratic, (13) shows that $V_\varepsilon(x)$ is the ball of radius ε for the metric of A. In this case, p itself induces a metric because $p(x,y,g) = 1/2\ (x-y, A(x-y))$

Neglecting the difference between a set and its closure, we thus have defined in the general case a neighborhood $(V_\varepsilon(x))$ whose image by ∂f is contained in $\partial_\varepsilon f(x)$. Because $\partial_\varepsilon f(x)$ is closed, we actually have

$$\overline{\partial f\ [V_\varepsilon(x)]} \subset \partial_\varepsilon f(x).$$

Conversely, we may ask if this inclusion is tight, i.e.: is it possible to generate all $\partial_\varepsilon f(x)$ by computing subgradients at appropriate points ? The answer is given by a theorem of Broensted and Rockafellar [4]:

Suppose f is lower semi continuous.
Let x and $g \in \partial_\varepsilon f(x)$ be given.

$$\left\{\begin{array}{l}\text{For any } \alpha > 0, \text{ there exist } y_\alpha \text{ and } g_\alpha \in \partial f(y_\alpha) \\ \text{such that } |y_\alpha - x| \leq \alpha \text{ and } |g_\alpha - g| \leq \varepsilon/\alpha.\end{array}\right.$$

Thus, by letting $\alpha \downarrow 0$, g can be described by a limit of g_α and we obtain

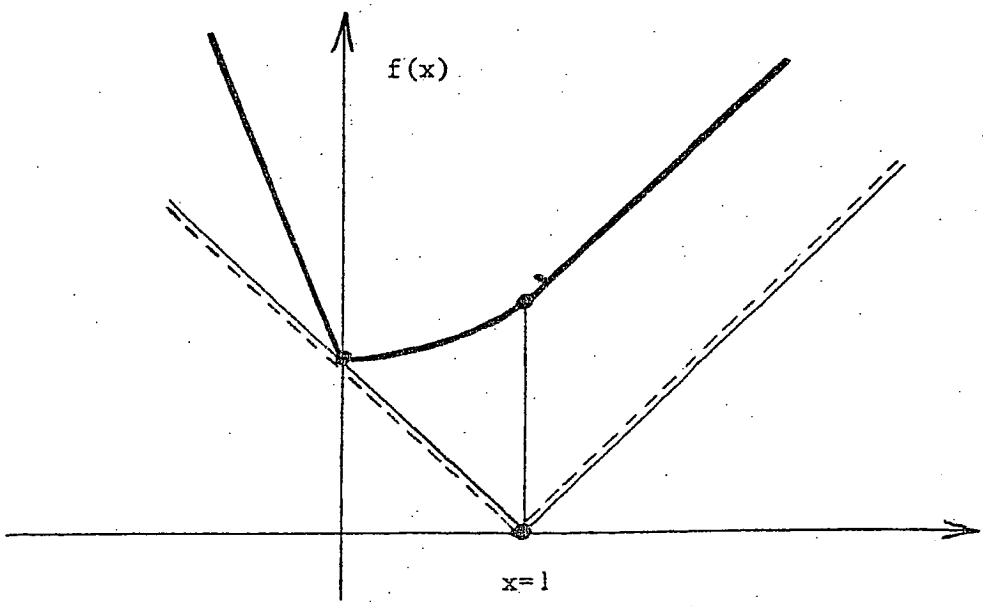
$$\partial_\varepsilon f(x) \subset \overline{\partial f[\bar{V}_\varepsilon(x)]}$$

Note that lower semi continuity of f (which holds for a locally Lipschitz function) is essential: take for example n=1 and

	for	$x > 0$	$x = 0$	$x < 0$
define	$f(x) =$	x	1	$+\infty$
so	$\partial f(x) =$	$\{1\}$	$\emptyset$	$\emptyset$

For $x > 0$, $\partial_\varepsilon f(x)$ exists and is the non zero segment $[1-\varepsilon/x, 1]$, while $\partial f(R) = \{1\}$ is only able to build one point of this segment.

On the other hand, the result can be illustrated by $f(x) = \max\{1-2x, \sqrt{1+x^2}\}$ (see the graph of f below).



Take $x = 1$, for which $f(x) = \sqrt{2}$, and take $\varepsilon = \sqrt{2}$. Geometric inspection and algebraic calculations show that:

$$\partial_\varepsilon f(x) = [-1,+1] \qquad V_\varepsilon(x) =]0,+\infty[\qquad \bar{V}_\varepsilon(x) = [0,+\infty[$$

$$\partial f[V_\varepsilon(x)] =]0,1[\subset [-1,+1] = \partial_\varepsilon f(x) \subset [-2,+1] = \overline{\partial f[\bar{V}_\varepsilon(x)]}$$

This example shows that failure to differentiability may cause the inclusions squarring $\partial_\varepsilon f(x)$ to be strict. Note that $V_\varepsilon(1)$ remains constant when ε varies from $\sqrt{2} - 1$ to $\sqrt{2} + 1$.

This example shows also that $V_\varepsilon(x)$ is not necessarily bounded and we conclude this section with an example showing that it is not necessarily convex: in 2 dimensions take $f(x,y) = \max\{0 , x^2+y^2-1\}$ and consider the point $x = 1$, $y = 1$. For $\varepsilon = 1/2$ we see that $0 \notin \partial_{1/2} f(1,1)$ so $V_{1/2}(1,1)$ cannot contain any point where $f = 0$ (in which case 0 is a subgradient). Actually one can show that $V_{1/2}(1,1)$ is described by the two inequalities

$$(x-1)^2 + (y-1)^2 \leq 1/2 \qquad x^2 + y^2 > 1$$

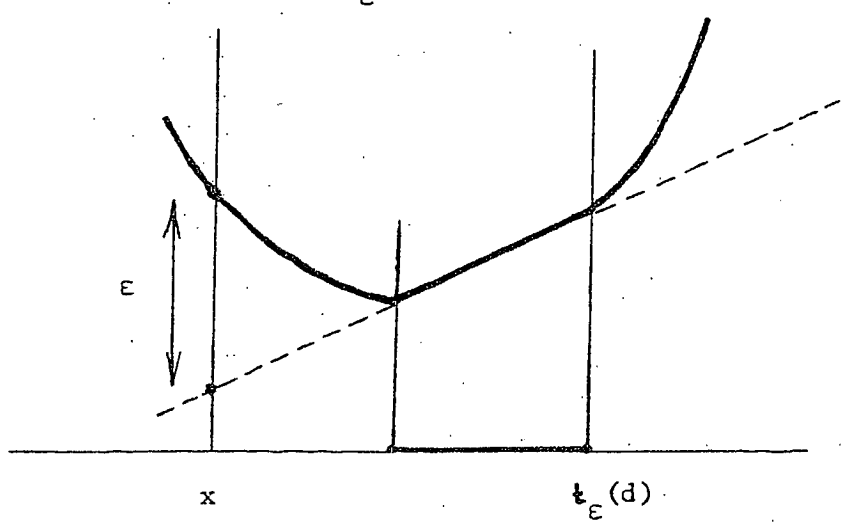
3.6. Second differentiability

In this section we suppose $\varepsilon > 0$ and we mention some recent results concerning $f'_\varepsilon(x,d)$ as a function of x.

In view of (12), we may define for each given x, d and ε:

$$T_\varepsilon(d) = \{ t > 0 \ : \ [f(x+td)-f(x)+\varepsilon]/t = f'_\varepsilon(x,d) \}$$

$$t_\varepsilon(d) = \sup \{ t \ : \ t \in T_\varepsilon(d) \}$$



The above picture illustrates these definitions. It shows that $T_\varepsilon(d)$ is a convex segment, which may be void or unbounded, in which case we set $t_\varepsilon(d) = +\infty$ (this typically happens when $f(x+td)$ has an asymptote for $t \to \infty$ and ε is large enough).

Not only $0 \notin T_\varepsilon(x)$ but the optimization problem in (12) is safe, in the sense that $T_\varepsilon(x)$ is "uniformly (with respect to x) far from 0":

(18) $\begin{cases} \text{Let } L \text{ be a Lipschitz constant on the unit ball around } x. \\ \text{Then for any } t \in T_\varepsilon(d), \text{ we have } t|d| \geq \varepsilon / (2L + \varepsilon) \end{cases}$

To prove it, take $t|d| < \varepsilon/(2L+\varepsilon) \leq 1$, so $f(x+td) \geq f(x) - Lt|d|$ which implies

$[\, f(x+td) - f(x) + \varepsilon \,] / t \geq \varepsilon/t - L|d| > (L + \varepsilon)\,|d|$

and the latter is larger than $f'_\varepsilon(x,d)$ because of (14).

It is interesting to note that $T_\varepsilon(d)$ is related to $V_\varepsilon(x)$ of Section 3.5: it has been mentioned that $V_\varepsilon(x)$ may not be convex. However its intersection with any direction d is convex, it is a star-set, i.e. if $x+td \in \bar{V}_\varepsilon(x)$ for $t > 0$ then $x+t'd$ $V_\varepsilon(x)$ for any $t' \in [0,t[$. We can even say more: for each direction d, the point $x + t_\varepsilon(d)d$ is on the boundary of $V_\varepsilon(x)$ and we have the following result

$$\{\, t \geq 0 \;:\; x + td \in \bar{V}_\varepsilon(x) \,\} = [0, t_\varepsilon(d)]$$

which shows in particular that $T_\varepsilon(d)\, d \subset [0, t_\varepsilon(d)]\, d \cap \bar{V}_\varepsilon(x)$.

Now, using (18), we can replace $t > 0$ in (12) by $t|d| \geq \varepsilon/(2L+\varepsilon)$, and this explains why $f'_\varepsilon(x,d)$ is locally Lipschitz in x, as obtained by minimizing a sum of locally Lipschitz functions. Unfortunately, computing its peridifferential by simple composition rules gives only an over estimate (see Section 1.2.5).

However we can study the differential quotient in a given direction z:

$$[\ f'_\varepsilon(x+sz,d) - f'_\varepsilon(x,d)\ ]\ /\ s$$

and, if by any chance it has a limit when $s \downarrow 0$, this limit can conveniently be called $f''_\varepsilon(x,d;z)$. Such a study is made easier using the definition 3.3.3 of $\partial_\varepsilon f(x)$:

$$f'_\varepsilon(x,d) = \max_g \{\ (g,d)\ :\ f^*(g) - (x,g) \leq \varepsilon - f(x)\ \}$$

Call $G_\varepsilon(d)$ (a convex subset of $\partial_\varepsilon f(x)$) and $U_\varepsilon(d)$ (a non negative segment) the primal and dual solution sets of this convex single-constrained optimization problem. They are interestingly related to $T_\varepsilon(d)$ through the following result:

$$\begin{cases} G_\varepsilon(d) = \{\ g \in \partial f(x+td)\ :\ t \in T_\varepsilon(d),\ (g,d) = f'_\varepsilon(x,d)\ \} \\ t \in T_\varepsilon(d) \text{ and } t < t_\varepsilon(d) \text{ imply } \partial f(x+td) \subset G_\varepsilon(d) \\ U_\varepsilon(d) = 1/T_\varepsilon(d) \text{ with the convention that } 0 = 1/+\infty \end{cases}$$

Then the following result can be proved [1], [16]:

$$\begin{cases} f''(x,d;z) \text{ exists and is given by the formula} \\ \max_{g \in G_\varepsilon(d)} \ \min_{u \in U_\varepsilon(d)}\ u\ [(g,z) - f'(x,z)]. \end{cases}$$

where the min and max operations commute (note that the operand is linear in g and u, and the fields are convex, one at least is compact). We can also replace $u \in U_\varepsilon(d)$ by $1/t \in 1/T_\varepsilon(d)$. Of course, $f''_\varepsilon(x,d;z)$ is positively homogeneous with respect to d and to z (note in particular that, when d is multiplied by k, $G_\varepsilon(d)$ is not changed and $T_\varepsilon(d)$ is divided by k).

The mapping $d,p \rightarrow f''_\varepsilon(x,d;p)$ is rather ugly but this formula simplifies in some situations

a) Suppose that $T_\varepsilon(d)$ is a singleton for which $\partial f(x+t_\varepsilon(d)d)$ is also a singleton. Then we obtain

$$f''_\varepsilon(x,d;z) = [\ f'(x+t_\varepsilon(d)d,z) - f'(x,z)\ ]\ /\ t_\varepsilon(d)$$

which looks like a differential quotient of $f'(.,z)$ in the direction d with stepsize $t_\varepsilon(d)$. In particular, if f is quadratic, direct computations using 3.2.3 show that

$$t_\varepsilon(d) = \sqrt{2\varepsilon\ /\ (d,Ad)}\ ,\ G_\varepsilon(d) = \nabla f(x) + t_\varepsilon(d)\ Ad$$
$$f''_\varepsilon(x,d;z) = (Ad,z) \quad \text{for any } \varepsilon > 0$$

b) When $z = d$, $(g,z) = f'_\varepsilon(x,d)$ for any g in $G_\varepsilon(d)$ so

$$f''_\varepsilon(x,d;d) = [f'_\varepsilon(x,d) - f'(x,d)]/t_\varepsilon(d)$$

which is always positive and positively homogeneous of degree 2 with respect to d; it is 0 if and only if $f'_\varepsilon(x,d) = f'(x,d)$ i.e. f is positively homogeneous at x in the direction d (see Section 3.2.1).

In particular, if f is quadratic, $f''_\varepsilon(x,d;d)$ is convex in d and defines a metric if A is positive definite.

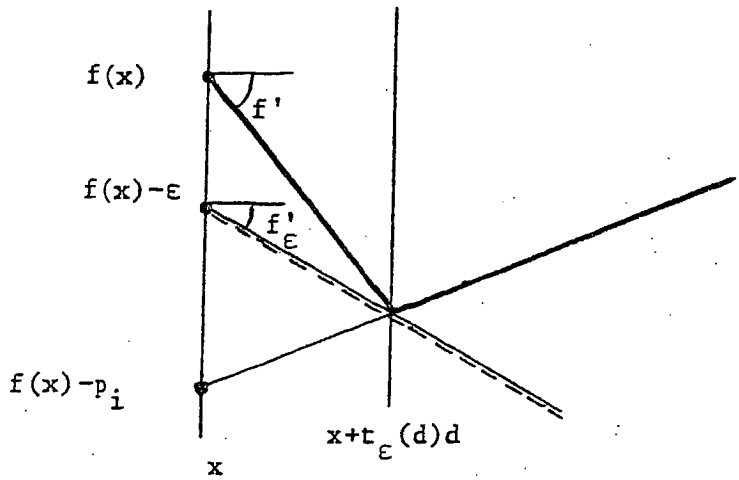
c) Another interesting case is when f is piecewise linear. It is convenient to write a piecewise linear function as follows:

$$f(x+d) = f(x) + \max\ \{\ -p_i + (g_i,d)\ :\ i = 1,\ldots,m\ \}$$

where the p_i's are m non negative numbers and the g_i's are m vectors. We denote I the set of i such that $p_i = 0$ so that

$$f'(x,d) = \max\ \{\ (g_i,d)\ :\ i \in I\ \}$$

We suppose ε small enough, namely strictly smaller than $\min\ \{p_i\ :\ i \in I\}$. Then the singleton $T_\varepsilon(d)$ is the abscissa of the first kink of f in the direction d.



From the definition of $t_\varepsilon(d)$,

$$f[x+t_\varepsilon(d)d] - f(x) = t_\varepsilon(d)\, f'(x,d) = -p_i + t_\varepsilon(d)\,(g_i,d)$$

for some $i \in I$ and the picture clearly shows the relations

$$t_\varepsilon(d) = \min \{ p_i \,/\, [(g_i,d)-f'(x,d)] \;:\; i \notin I\,,\ (g_i,d) > f'(x,d) \}$$

so

$$1 \,/\, t_\varepsilon(d) = \max \{ [(g_i,d) - f'(x,d)] \,/\, p_i \;:\; i \notin I \}$$

(the second condition in the max can be dropped because it is certainly redundant). Furthermore

$$t_\varepsilon(d)\, f'(x,d) = -\varepsilon + t_\varepsilon(d)\, f'_\varepsilon(x,d)$$

hence

$$f''_\varepsilon(x,d;d) = [f'_\varepsilon(x,d) - f'(x,d)] \,/\, t_\varepsilon(d) = \varepsilon \,/\, t^2(d) =$$
$$= \varepsilon \max{}^2 \{ [(g_i,d) - f'(x,d)] \,/\, p_i \;:\; i \notin I \}$$

This is the square of a max of positive functions, which is convex in d if the underlying functions are convex, i.e. if $-f'(x,d)$ (which is concave) is linear, i.e. if f has a gradient at x, i.e. if I is a singleton. Otherwise, it is easy to find counter-examples showing that, contrary to the quadratic case, the set of d such that $f''_\varepsilon(x,d;d) \leq 1$ is neither bounded nor convex.

Because of the relations between Sections 3.5 and 3.6, $f''_\varepsilon(x,.;.)$ and $V_\varepsilon(x)$ are intimately related. We leave as an exercise the study at $x = 0$, $y = 0$, $\varepsilon < 1$ of the 2 dimensional counter-example

$$f(x,y) = \max \{ x, y, -1 + 2x + 2y \}$$

for which

$$V_\varepsilon(0) = \{ x,y \ : \ x + 2y < 1 \ \underline{\text{or}} \ 2x + y < 1 \}$$

while, setting $d = (d_1,d_2)$,

$$f''(0,d;d)/\varepsilon = \begin{cases} 0 & \text{if } d_1 + 2d_2 \leq 0 \text{ or } 2d_1 + d_2 \leq 0 \\ [\, 2(d_1+d_2) - \max(d_1,d_2) \,]^2 & \text{otherwise} \end{cases}$$

REFERENCES

[1] A. Auslender, "On the differential properties of the support function of the epsilon-subdifferential of a convex function", Mathematical Programming 24,3(1982) 257-268.

[2] C. Berge, Espaces topologiques et fonctions multivoques (Dunod, Paris, 1966).

[3] A. Bihain, "Optimization of upper semi differentiable functions", Journal of Optimization Theory and Applications (to appear).

[4] A. Broensted and R.T. Rockafellar, "On the subdifferentiability of convex functions", Proceedings of the American Mathematical Society 16(1965) 605-611.

[5] F.H. Clarke, "Generalized gradients and applications", Transactions of the American Mathematical Society 205(1975) 247-262.

[6] F.H. Clarke, "A new approach to Lagrange multipliers", Mathematics of Operations Research 1,2(1976).

[7] J.M. Danskin, "The theory of maxmin with applications", SIAM Journal on Applied Mathematics 14,4(1966) 641-655.

[8] A. Feuer, "An implementable mathematical programming algorithm for admissible fundamental functions", Ph.D. Thesis, Department of Mathematics, Columbia University (New York, 1974).

[9] J.B. Hiriart Urruty, "On optimality conditions in nondifferentiable programming", Mathematical Programming 14(1978) 73-86.

[10] J.B. Hiriart Urruty, "Optimality conditions for discrete non linear norm-approximatio problems", in: A. Auslender, W. Oettli and J. Stoer, eds., Optimization and optimal control, Lecture Notes in Control and Information Sciences 30 (Springer Verlag, Heidelberg, 1981).

[11] J.B. Hiriart Urruty, "Epsilon-subdifferential calculus", in: Convex analysis and optimization, Research Notes in Mathematics Series 57 (Pitman Publishers, 1982).

[12] G. Lebourg, "Valeur moyenne pour gradient généralisé", Comptes Rendus Académie des Sciences Paris 281,A(1975) 795-797.

[13] C. Lemaréchal, "Bundle methods in nonsmooth optimization", in: C. Lemaréchal and R. Mifflin, eds., Nonsmooth optimization (Pergamon Press, Headington Hill Hall, England, 1978) pp. 79-102.

[14] C. Lemaréchal, "Extensions diverses des méthodes de gradient et applications", Thèse d'Etat, UER Mathématiques de la Décision, University of Paris IX (Paris, 1980).

[15] C. Lemaréchal, "Nondifferentiable optimization", in: L.C.W. Dixon, E. Spedicato and G.P. Szego, eds., Nonlinear optimization (Birkhauser, Boston, 1980) pp. 149-199.

[16] C. Lemaréchal and E.A. Nurminskii, "Sur la différentiabilité de la fonction d'appui du sous diférentiel approché", Comptes Rendus Académie des Sciences Paris 290,A(1980) 855-858.

[17] R. Mifflin, "Semi smooth and semi convex functions in constrained optimization", SIAM Journal on Control 15,6(1977) 959-972.

[18] V.I. Norkin, "Generalized differentiable functions", Cybernetics 16,1(1980) 10-12.

[19] J.P. Penot, "Calcul sous différentiel et optimisation", Journal of Functional Analysis 27,2(1978) 248-276.

[20] B.N. Pshenichnyi, Necessary conditions for an extremum (Marcel Dekker, New York, 1971).

[21] R.T. Rockafellar, Convex analysis (Princeton University Press, Princeton N.J., 1970).

[22] R.T. Rockafellar, The theory of subgradients and its applications to problems of optimization, Lecture Notes, University of Montréal (1978). French translation: La théorie des sous gradients et ses applications à l'optimisation, Collection de la Chaire Aisenstadt, Les Presses de l'Université de Montréal (C.P. 6128, succ. A, Montréal H3C 3J7, Canada).

[23] R.T. Rockafellar, Favorable classes of Lipschitz continuous functions in subgradient optimization, in: E.A. Nurminskii, ed., Progress in nondifferentiable optimization (IIASA, CP-82-S8, 1982) pp. 125-143.

[24] S. Saks, Theory of the integral (Hafner Publishing Co., New York 1937).

[25] J.J. Strodiot, V.H. Nguyen and N. Heukemes, "Epsilon-optimal solutions in nondifferentiable convex programming and some related questions", Mathematical Programming (to appear).

Imprimé en France

par

l'Institut National de Recherche en Informatique et en Automatique