A GRADUATE SEMINAR REPORT
ON

DESCENT THEORY OF CONVEX MINIMIZATION

By
AYALEW GETACHEW

Advisor
Prof. Dr. rer. nat. habil. R. Deumlich

School of Graduate Studies
Addis Ababa University

June 1998
Addis Ababa

A SEMINAR REPORT

ON

DESCENT THEORY OF CONVEX MINIMIZATION

BY

AYALEW GETACHEW

ADVISOR

PROF.DR.RER.NAT.HABIL.R. DEUMLICH

ADDIS ABABA UNIVERSITY

JUNE 1998

Table of Contents

Acknowledgment

This Seminar would have been impossible without the fatherly guidance and day to day follow up of my most respected advisor Prof. Dr. rer. nat. habl. R. Deumlich, who has helped me a lot to make this report a reality.

Ayalew Getachew

June, 1998

Addis Ababa University

Preface

Steepest Descent method is one of the methods of unconstrained function minimization. Studying minimization of unconstrained function is useful since some of the most powerful and convenient methods of solving constrained minimization problems (such as Lagrange- method) involve the transformation of the problem into one of unconstrained minimization.

Since the main objective of optimization is to minimize (or maximize) optimization problems, one should hjave a way of tackling unconstrained function minimization, and so among the major methods of unconstrained function minimization methods of Steepest Descent is the one.

This Seminar Report is a completion of the two Seminars I have delivered for qualification for M.Sc. in Mathematics. Thus in this Seminar paper I have attempted to present the basic definitions and properties of Descent Theory for Convex Minimization.

I have divided the paper into two parts:

I. Preliminaries - a review of definitions.

II. Descent Theory for Convex Minimization.

Through out this paper the norm considered is the Euclidean norm.

$\|.\| = < ., . >^{\frac{1}{2}}$

PART ONE
PRELIMINARIS (REVIEW)

1.1 Convex Sets and Convex Functions

DEFINITION 1.1.1: A set $K \subseteq \Re^n$ is said to be convex if and only if $\lambda x + (1-\lambda)y \in K$ for each $x, y \in K$ and $\lambda \in [0, 1]$.

DEFINITION 1.1.2: A function $f : \Re^n \to \Re \cup \{+\infty\}$, not identically $+\infty$, is said to be convex if for each $x, y \in \Re^n$ and $\lambda \in [0, 1]$ there holds $f(\lambda x + (1-\lambda)y \leq \lambda f(x) + (1-\lambda)f(y)$.

1.2 Supporting hyperplane and support function

DEFINITION 1.2.1: Let $K \subseteq X = \Re^n$ and $H = \{x \in X : < u, x >= \alpha, u \neq 0$ be a hyperplane. H is said to be a supporting hyperplane of K if and only if

1. $< u, x > \geq \alpha \quad \forall x \in K.$

2. $H \cap K \neq \phi.$

DEFINITION 1.2.2: Let S be a non empty set in $\Re^n$. Let $x \in \Re^n$. The function $\sigma_S : \Re^n \to \Re$ defined by

$$\sigma_S(x) := \sup\{< s, x >: s \in S\} \quad .$$

is called the support function of S.

. A support function is closed and sublinear.

Denotation: $\sigma_S(d)$: Support function of S at $d \in \Re^n$.

1.3 Directional derivative

DEFINITION 1.3.1: Let x and d be fixed in $\Re^n$ and $f : \Re^n \to \Re$ convex. The right side directional derivative of f at x in the direction d is

$$f'_+(x, d) := \lim_{t \downarrow 0} \frac{f(x+td)-f(x)}{t} \quad \text{for } t > 0.$$

1.4 Subdifferential

DEFINITION 1.4.1: (Subdifferential I).

The Subdifferential $\partial f(x)$ of f at x is non-empty compact convex set of $\Re^n$ whose support function is $f'_+(x,.)$, i.e.
$\partial f(x) := \{s \in \Re^n : \langle s.d \rangle \leq f'_+(x.d) \text{ for all } d \in \Re^n\}$ where
$f: \Re^n \to \Re$ is convex.

. A vector s is called the subgradient of f at x.

DEFINITION 1.4.2: (Subdifferential II)

Let $f: \Re^n \to \Re$ be convex. The subdifferential of f at $x \in \Re^n$ is a set of vectors $s \in \Re^n$ satisfying

$$f(y) \geq f(x) + \langle s, y - x \rangle \text{ for all } y \in \Re^n .$$

1.5 Kink

DEFINITION 1.5.1: A point at which $\partial f(x)$ has more than one element, i.e. at which f is not differentiable is called a kink (sharp corner of f).

1.6 Cone

DEFINITION 1.6.1: A cone K is a set in $\Re^n$ such that the "open" half line $\{\alpha x : \alpha > 0\}$ is entirely contained in K whenever $x \in K$.

DEFINITION 1.6.2: The polar of a cone of K is

$$K^\circ := \{s \in \Re^n : \langle s, x \rangle \leq 0 \text{ for all } x \in K\} .$$

DEFINITION 1.6.3: Normal cone

1. The direction $s \in \Re^n$ is said to be normal to $C \subseteq \Re^n$ at $x \in C$ when $\langle s, y - x \rangle \leq 0$ *for all* $y \in C$.

2. The set of all such directions is called normal cone to C at x, denoted by $N_C(x)$.

. The tangent cone is the polar of Normal Cone.

1.7 Saddle Point

Let X and Y be two non-empty sets and consider the given function

$$l : X \times Y \to \Re$$

Suppose we want to minimize l with respect to X, and maximize with respect to Y.

Consider the following functions

$$T(x) := \{\bar{y} \in Y : l(x,\bar{y}) = \sup_{y \in Y} l(x,y)\}$$

$$S(y) := \{\bar{x} \in X : l(\bar{x},y) = \inf_{x \in X} l(x,y)\}$$

This defines two multi-functions, $T : X \to Y$ and $S : Y \to X$ whose graphs are subsets of $X \times Y$ and $Y \times X$ respectively.

DEFINITION 1.7.1: A couple $(\bar{x},\bar{y}) \in X \times Y$ is said to be a saddle point of l on $X \times Y$ when

$$\bar{y} \in T(\bar{x}) \text{ and } \bar{x} \in S(\bar{y})$$

. A further definition is given by

$$l(\bar{x},y) \le l(x,\bar{y}).$$

PART TWO
DESCENT THEORY FOR CONVEX MINIMIZATION

1. Descent Directions and Steepest -Descent Schemes

1.1 Basic Definitions

Let $f: \Re^n \to \Re$ be convex, consider

$$(P): \quad f(x) \to \min, x \in \Re^n$$

Descent method is the method of solving an unconstrained optimization problem. In this method, we start from an intiial trial x_1 and iteratively move towards an optimum point according to the following rule:

1. We compute a direction of move $d_k \in \Re^n$ for $X_k \in \Re^n$.
2. We compute stepsize $t_k > 0, t_k \in \Re^+$ and then $x_k + t_k d_k$ such that $f(x_k + t_k d_k) < f(x_k)$ at each iteration.

Definition 1.1.1: Let $f: \Re^n \to \Re$ convex. A vector $d \in \Re^n, d \neq 0$ is said to be a descent direction of f at x if it is satisfies: $\exists t > 0$ such that $f(x+td) < f(x)$.

Theorem 1.1.2: A descent direction is equivalently stated by any one of the following properties:

$$\left.\begin{array}{l} f'_x(x,d) < 0 \\ \sigma_{\partial f(x)}(d) < 0 \\ < s, d > \ < 0 \ \ for\ all\ s \in \partial f(x) \end{array}\right\} \qquad (1.1.1)$$

Proof: Suppose that $d \neq 0$ is a descent direction
Assume $f'_+(x,d) \geq 0$ (proof by contradiction).

$$\Rightarrow \quad \lim_{t \downarrow 0} \frac{d(x+td)-f(x)}{t} \geq 0$$

$$\Rightarrow \quad \frac{f(x+td)-f(x)}{t} \geq 0 \ \ (\text{since } \varphi(t) = \frac{f(x+td)-(f(x)}{t} \text{ is monotone increasing})$$

$$\Rightarrow \quad \frac{f(x+td)-f(x)}{t} \geq 0 \qquad \forall t > 0$$

$\Rightarrow$ d is not a descent direction. But this is a contradiction to our supposition.

Therefore $f'_+(x,d) < 0$.

Suppose

$f'_+(x,d) < 0$

$$f'_+(x,d) < 0 \quad \Rightarrow \quad \lim_{t \downarrow 0} \frac{f(x+td)-f(x)}{t} < 0$$

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

Hence d is a descent direction.

And we have $\sigma_{\partial f(x)}(d) < 0 \Leftrightarrow \sup\{< s,d >: s \in \partial f(x)\} < 0 \Leftrightarrow f'_+(x,d) < 0$ by definition. //

When the gradient $\nabla f(x)$ happens to exist, the subdifferential reduces to the singleton $\{\nabla f(x)\}$ which generate half-line and the set of descent directions expands to the (open) half space opposite to $\nabla f(x)$.

Geometrically, a descent direction corresponds to a hyperplane separeting the two closed convex sets $\partial f(x)$ and $\{0\}$ strictly when $0 \notin \partial f(x)$.

Denote the subspace orthogonal to a given $d \in \Re^n (d \neq 0)$ by

$$d^{\perp} := H = \{z \in \Re^n :< z,d > \ = 0\}$$

Then this d defines a descent direction when $\partial f(x)$ lies entirely in the open half space limited by $d^{\perp}$ and opposite to d. (See Fig 1.1.1) The dashed line which passes between 0 and $\partial f(x)$, is such a separating hyperplane.

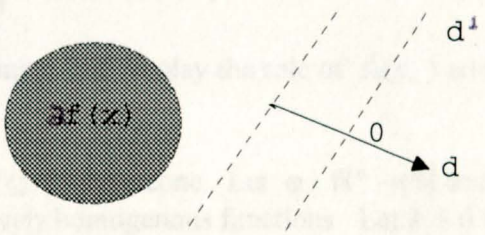

Fig 1.1.1 Descent direction and separating hyperplane

Theorem 1.1.3 A vector d is descent direction if for $\alpha \in [f'_+(x,d), 0)$, the hyperplane

$$H = \{z \in \Re^n : < z, d > = \alpha\}$$

separates $\partial f(x)$ and $\{0$ strictly, or in otherwords

$< s, d > \leq \alpha < 0$ *for all* $s \in \partial f(x)$ *and for* $\alpha \in [f'_+(x,d), 0]$ **(1.1.2)**

Proof: From the definition of subdifferential we have $< s, d > \leq f'_+(x,d)$ *for all* $s \in \partial f(x)$.

Therefore $< s, d > \leq f'_+(s,d) \leq \alpha < 0$ [Since $\alpha \in (f'_+(x,d), 0)$]

$\Rightarrow$ $< s, d > < < s', d >$ *for all* $s \in \partial f(x)$ and for all $s' \in \{0\} \in \Re^n$.

$\Rightarrow$ $\sup < s, d > < \inf < s, d >$

Hence $\partial f(x)$ and $\{0\}$ can be strictly separated. //

Definition 1.1.4: Let $\|.\|$ be a Euclidean norm on $\Re^n$. A normalized steepest descent direction of f at x, associated with $\|.\|$ is a solution of the problem

$$(P_d): \ f'_+(x,d) \to \min, d \in S \qquad \textbf{(1.1.3)}$$

$$S = \{d \in \Re^n : \|d\| = 1\}$$

or equivalently

$$(P_d) \qquad \max < s, d > \to \min, d \in S \quad \textbf{(1.1.4)}$$

$$S = \{d \in \Re^n : \|d\| = 1\}$$

A vector $d \in \Re^n$ is said to be a non-normalized steepest descent direction if $\|d\| \neq 1$ and $d' = \frac{d}{\|d\|}$ is a solution of (1.1.3).

In the following φ and ν play the role of $f'_+(x,.)$ and $\|\ \|$ respectively.

Proposition 1.1.5: Let $N \subseteq \Re^n$ be a cone. Let $\varphi : \Re^n \to \Re$ and $\nu : \Re^n \to \Re$ be two positively homogenous functions. Let $k > 0$ be given and let D_k be the set of solutions of

$$P_k \qquad \varphi_{(d)} \to \min, d \in S \qquad \textbf{(1.1.5)}$$

$$S = \{d \in N : \nu(d) = k\}$$

then $D_k = kD_1$ *for all* $k > 0$ where D_1 is the set of solution of

$$(P_1): \quad \varphi(d) \to \min, d \in S$$

$$S = \{d \in N : \nu(d) = 1\}$$

Proof: Take arbitrary $\lambda > 0$ and suppose that d solves

$$(\varphi_\lambda): \quad \varphi(d) \to \min, d \in S$$

$$S = \{d \in N : \nu(d) = \lambda\}$$

Then $d \in N, \nu(d) = \lambda$ and $\varphi(d') \geq \varphi(d)$ for all $d' \in N$ with $\nu(d') = \lambda$ which can be written (multiplying by k and using homogenity)

$$\varphi(kd') \geq \varphi(kd) \text{ for all } d' \in N \text{ with } \nu(kd') = k\lambda$$

Take $d'' \in N$ arbitrary with $\nu(d'') = k\lambda$ and set $d' := \frac{d''}{k} \in N$

Then $\varphi(d'') \geq \varphi(kd)$ *for all* $d'' \in N$ with $\nu(d'') = k\lambda$.

In other words

$$kD_\lambda \subset D_{k\lambda} \text{ for all } k > 0 \text{ and } \lambda > 0 \qquad \textbf{(1.1.6)}$$

Since $\frac{1}{k} > 0$ for $k > 0$, there also holds

$$\tfrac{1}{k} D_\lambda \subset D_k^\lambda \text{ for all } k, \lambda > 0 \qquad \textbf{(1.1.7)}$$

If we take $\lambda = 1$ in (1.1.6) and $\lambda = k$ in (1.1.7) we obtain

$$D_k = kD_1 \quad //$$

We observe that (1.1.3) has at least one optimal solution, because it consists of minimizing the continuous $f'_+(x, .)$ on the compact unit ball. With this in mind, we can specify more accurately our algorithmic scheme.

Algorithm 1.1.6 (Steepest-Descent Scheme)

Start from some $x_1 \in \Re^n$. Set $k = 1$

Step 1 **(Stopping Criterion)** If $0 \in \partial f(x)$ stop

Step 2 **(Direction Finding)** For the norm $\| \ \|$ take d_k solving (1.1.3) or (1.1.4)

Step 3 **(Line Search)** Find a stepsize $t_k > 0$ and new iterate $x_{k+1} = x_k + t_k d_k$ such that $f(x_{k+1}) < f(x_k)$.

step 4 **(Loop)** Replace k by $k+1$ and loop to step 1.

A stop in step 1 means that x_k is optimal.

1.2 Solving the Direction Finding problem

Consider the steepest descent problem

$$(p_d): \qquad f'_+(x,d) \to \min, d \in S$$

$$S = \{d \in \Re^n : \|d\| \le 1\}$$

with its nonlinear and nonconvex constraint. This problem must be solved at each execution of step 2 in Algorithm 1.1.6. Let us consider one given iteration $x \in \Re^n$. The question is how to find a steepest descent direction at a given $x \in \Re^n$.

As an alternative to (1.1.3) consider the nicer, "convexified," problem

$$(P'_d): \qquad f'_+(x,d) \to \min, d \in S'$$

$$S' = \{d \in \Re^n : \|d\| \le 1\}$$

Or equivalently

$$(P'_d): \qquad \max_{s \in \partial f(x)} < s, d > \to \min, d \in S' \qquad \textbf{(1.2.1)}$$

$$S' = \{d \in \Re^n : \|d\| \le 1\}$$

The next result makes precise the difference between (1.2.1) and (1.1.3).

Theorem 1.2.1 The solution set of (1.1.3) and (1.2.1) have the following properties.

1. Either $f'_+(x,d) \geq 0$ which means x minimizes f.

2. Or if $0 \in \partial f(x)$, the solution set of (1.2.1) and (1.1.3) coincide.

Proof:

1. The minimal objective value is never strictly positive since it cannot be larger than $f'_+(x,d) = 0$. But $f'_+(x,d) \geq 0 \Rightarrow$ there is no descent direction.

   i.e. $f(x+td) \geq f(x)$ *for all* $d \in \Re^n, t > 0$

   $\Rightarrow$ $f(y) \geq f(x)$

   $\Rightarrow$ x minimizes f which means $0 \in \partial f(x)$

2. Suppose now $0 \notin \partial f(x)$. So there exists $\bar{d}$ with $\|\bar{d}\| \leq 1$ and $f'_+(x,\bar{d}) < 0$. By the positive homogenity of $f'_+(x,.)$ we may just assume $\|d\| = 1$. Thus $f'_+(x,\bar{d}) < 0$ both in (1.2.1) and (1.1.3) and $d = 0$ cannot be optimal in any of these problems. Then consider and arbitrary d with $f'_+(x,d) < 0$ and $\|d\| < 1$. Set $d' = \frac{d}{\|d\|}$

   $$f'_+(x,d') = f'_+\tfrac{(x,d)}{\|d\|} < f'(x,d)$$

   In other words, d' is feasible in (1.2.1) and strictly better than d, which is therefore cannot be optimal. This means that (1.2.1) is not changed if it is feasible set is restricted to $\|d\| = 1$ which is the feasible set in (1.1.3). Therefore the solution set of (1.1.3) and (1.2.1) coincide. //

Theorem 1.2.2 Let $\hat{S}$ be the solution set of

$$(P_s): \qquad \|s\| \to \min . s \in \partial f(x) \qquad \textbf{(1.2.2)}$$

Take arbitrary $\hat{s} \in \hat{S}$. Then the solution of (1.2.1) are the solution of

$$(P_d): \qquad <-\hat{s}, d> \ \to \ \max d \in S' \qquad \textbf{(1.2.3)}$$

$$S' = \{d \in \Re^n : \|d\| \leq 1\}$$

that lie in the normal cone $N_{\partial f(x)}(\hat{s})$ to $\partial f(x)$ at $\hat{s}$.

Proof: Consider the set $\tilde{S} \times D$ of saddle-points of the bilinear function $(s,d) \rightarrow < s,d >$ over the product of compact convex sets $\partial f(x)$ and

$B = \{d \in \Re^n : \|d\| \leq 1\}$

$\hat{s} \in \hat{S} \subset \partial f(x)$ *and* $\hat{d} \in D \subset B$ if and only if

$$< s,d > \leq < \hat{s},\hat{d} > \leq < \hat{s},\hat{d} > \text{ for all } S \in \partial f(x) \text{ and } d \in B \qquad \textbf{(1.2.4)}$$

From the theory of saddle points $\tilde{S}$ is exactly the solution set of
$\max_{s \in \partial f(x)} \min_{d \in B} < s,d > \Leftrightarrow \max_{s \in \partial f(x)} \{-\max_{d \in B} < -s,d >\}$

$$\Leftrightarrow \max_{s \in \partial f(x)} \{-\|d\|\} \Leftrightarrow \max_{s \in \partial f(x)} \{-\|s\|\}$$

which means that

$x \in \tilde{S} \Leftrightarrow x \in \hat{S}$. Hence $\tilde{S} = \hat{S}$

We know also that D is exactly the solution set of (1.2.3) but from (1.2.4), $\tilde{d} \in D$ if and only if, given $\tilde{s} \in \partial f(x)$ the following two properties hold:

$$< s,\hat{d} > \leq < \hat{s},d > \text{ for all } s \in \partial f(x) \qquad [d \in N_{\partial f(x)}(\hat{s})]$$

$$< \hat{s},\hat{d} > \leq < \hat{s},\hat{d} > \text{ for all } d \in B \qquad [d \text{ solves (1.2.3)}].$$

This result indicates how to solve the converxified steepest descent problem (1.2.1).

Corollary 1.2.3: The following statements are equivalent

i) $\hat{d}$ solves (1.2.1) and $\hat{s}$ solves (1.2.2)

ii) $\left\|\hat{d}\right\| \leq 1, \hat{s} \in \partial f(x)$ and there holds

$$< \hat{s},\hat{d} > = -\left\|\hat{s}\right\| = \sigma_{\partial f(x)}(\hat{d}) \qquad [= f'_+(x,\hat{d})] \qquad \textbf{(1.2.5)}$$

Proof: By Theorem 1.2.2 (i) holds if and only if $\hat{d}$ solves (1.2.3)

Hence $< -\hat{s},\hat{d} > = \left\|-\hat{s}\right\|$

$\Rightarrow \quad < \hat{s}, \hat{d} > = -\|\hat{s}$ and

$d \in N_{\partial f(x)}(\hat{s})$ *i.e.* $< s, \hat{d} > \leq < \hat{s}, \hat{d} >$ *for all* $s \in \partial f(x)$.

Therefore $< \hat{s}, \hat{d} > = -\|\hat{s}\| = \sigma_{\partial f(x)}(\hat{d})$.

Hence (ii) //

The computation of an $\hat{s} \in \hat{S}$ in Theorem 1.2.2 is a familiar enough problem $\hat{s}$ is the best approximation of the origin in the $\|.\|$ sense with regard to a compact convex set $\partial f(x)$.

From corollary 1.2.3 the solutions of (1.2.1) are then the solutions of the system

The following problem may be considered as more handy than (1.2.6)

$$\left| \begin{array}{l} d \leq 1 \\ < \hat{s} > = -\|\hat{s}\| \\ < s, d > \leq < \hat{s}, d > \text{ for all } s \in \partial f(x) \end{array} \right. \qquad \textbf{(1.2.6)}$$

$$(P_d): \quad \|d\| \to \min, d \in S \qquad \textbf{(1.2.7)}$$

$S = \{ d \in \Re^n :< \hat{s}, d > = -\|\hat{S}\|, < s.d > \leq < \hat{s}, d > \text{ for all } s \in \partial f(s)$

i.e $d \in N_{\partial f(x)}(\hat{s})\}$

This problem is a convex minimization problem with one affine enequality constraints and a possibly infinite number of linear constraints.

Proposition 1.2.4: Let $\hat{s}$ solve (1.2.2). The solutions of (1.2.7) solve (1.2.1). Conversly, (1.2.1) and (1.2.7) have the same solution set if $\hat{s} \neq 0$.

Proof: Because (1.2.1) has a solution, (1.2.6) does have a solution. The optimal value in (1.2.7) is therefore not greater than 1, and any optimal solution of (1.2.7) solves (1.2.1), this is corollary 1.2.3.

If $\hat{s} \neq 0$, all solutions of (1.2.1) have norm 1 (this is Theorem 1.2.1). Hence all the solutions of the equivalent problems (1.2.3) and (1.2.6) have norm 1. We calculate that the minimal value in (1.2.7) is exactly 1 and (1.2.7) is really equivalent to (1.2.6) = (1.2.1).

Therefore (1.2.7) and (1.2.1) have the same solution. //

If $\hat{s}=0$, we observe that (1.2.7) has a unique solution $\hat{d}=0$. Yet, (1.2.1) or (1.2.6) may have non zero solutions unless $N_{\partial f(x)}(0)=\{0\}$, i.e. $0 \in int\partial f(x)$. This confirms that (1.2.7) is not exactly equivalent to (1.2.1). When $\hat{s}\neq 0$, the steepest-descent directions are upto normalization, the solution of

$$(P_d): \ \|d\| \to \min, d \in S \qquad \textbf{(1.2.8)}$$

$$S=\{d \in \Re^n : < s,d >= -1, < s-\hat{s},d > \ \leq 0 \ for\ all\ \ s \in \partial f(x)\} .$$

In summary, to perform steps 1 and 2 in the Steepest-Descent Algorithm 1.2.6 one has to

- solve (1.2.2), a projection problem
- check that it has a non-zero solution 3 (otherwise stop)
- solve (1.2.6), (1.2.7) or (1.2.8)

Remark 1.2.5 The constraint $d \in N_{\partial f(x)}(\hat{s})$ (i.e. $< s,d > \ \leq \ < \hat{s},d \quad \forall s \in \partial f(x)$) may really trouble if $\partial f(x)$ is complicated enough.

Suppose the problem

$$(P_d): \ \|d\| \to \min .d \in S$$

$$S: \{d \in \Re^n : < \hat{s},d > \ = -1, \hat{s} \ \in \partial f(x \text{ solving } 1.2.2\}$$

has unique solution $\hat{d}$. Then this $\hat{d}$ has to lie in $N_{\partial f(x)}(\hat{s})$ and solve (1.2.7). (Otherwise (1.2.7) and (1.2.1) would have no solution !) in this case, the last line of constraints in (1.2.7) or (1.2.8) can be neglected.

.3 Examples for some particular cases

Let us see some particular implication of the results of previous section:

Example 1.3.1

Let $X=\Re^2, f: \Re^2 \to \Re$ convex. Let

$$\partial f(x) = conv\{(0,\tfrac{3}{2}),(3,0)\} \qquad \textbf{(1.3.1)}$$

Find descent direction d of f at x.

Solution:

Let $s \in \partial f(x)$. $s = (s_1, s_2)$. $\partial f(x)$ can be rewritten as

$$\partial f(x) = \{s \in \Re^2 : 2s_2 - s_1 - 3 = 0, \ \ s_1 \geq 0, s_2 \geq 0\}$$

Therefore (1.2.2) can be written as

$$(P): \quad \|s\| \to \min, s \in \partial f(x)$$

$$\partial f(x) = \{s \in \Re^2 : 2s_2 + s_1 - 3 = 0, \ \ s_1 \geq 0, s_2 \geq 0\}$$

$L(s, \lambda) = s_1^2 + s_2^2 + \lambda(2s_2 + s_1 - \frac{3}{2}), \qquad [s_1 \geq 0, s_2 \geq 0]$

Kuhn Tucker Conditions

1. $\frac{\partial L}{\partial s_1} = 2s_1 - \frac{\lambda}{2} = 0$
2. $\frac{\partial L}{\partial s_2} = 2s_2 + \lambda = 0$
3. $2s_2 + s_1 - 3 = 0$
4. $s_1 \geq 0, s_1 \geq 0$

Case 1: $\lambda = 0$

$(1)(2) \Rightarrow \qquad s_1 = 0, \ \ s_2 = 0$

But this contradicts condition 3.

Case 2: $\lambda \neq 0$

$(1) \Rightarrow \ \lambda = 4s_1$

$(2) \Rightarrow \ \lambda = -2s_2$

$\Rightarrow \ 4s_1 + 2s_2 = 0 \qquad ...(5)$

Hence combining (5) and (3) we get

$$\begin{aligned} 4s_1 + 2s_2 &= 0 \\ s_1 + 2s_2 &= 3 \\ \hline 5s_1 &= 3 \\ s_1 = \tfrac{3}{5},\ s_2 &= \tfrac{6}{5} \end{aligned}$$

$\hat{s} = (\frac{3}{5}, \frac{6}{5})$ is a solution of p.

To find d consider the optimization problem.

$$(P_d): \ \|d\| \to \min, \quad d \in S$$

$$S = \{d \in \Re^n : < \hat{s}, d > = -1 \quad < s, d > \ \leq \ < \hat{s}, d \quad \forall s \in \partial f(x)\} \ .$$

Or

$$(P_d): \ \sqrt{d_1^2 + d_2^2} \to \min, d \in S$$

$$S = \{d \in \Re^2 : \ \tfrac{3}{5}d_1 + \tfrac{6}{5}d_2 = -1, \ \ 2d_1 - d_2 = 0\}$$

Solving this we get

$$d_1 = -\tfrac{1}{3}, \ \ d_2 = -\tfrac{2}{3}$$

Hence $d = (d_1, d_2) = (-\frac{1}{3}, -\frac{2}{3})$

1.4 Conclusion

In section 1 of this paper it is shown that computing the direction in the steepest-descent algorithm 1.1.6 amounts to solving two optimization problems: first the projection (1.1.2) and then (1.2.8). It is now necessary to ask the question: is this a constructive way of computing a steepest descent direction ?

Both problems (1.2.2) and (1.2.8) involve the structure of norming and the subdifferential. However, (1.2.7) can be considered as the easier problem, because its complexity depend less on $\partial f(x)$. For example, suppose we solve (1.2.3), which does not involve the subdifferential, and obtain a unique solution. Then this solution is the required steepest-descent direction.

On the other hand, (1.2.2) is usually impossible to solve, unless there is some structure in f (which is not under our control !). Three instances are mentioned below in which such a managable structure exist.

Case 1: The subdifferential is a compact convex polyhedron characterized as a convex

hull : $s_1, s_2, ..., s_m$ are given in $\Re^n$ and, Δm being the unit simplex,

$$\partial f(x) = conv\{s_1, ..., s_m\} = \{s = \sum_{j=1}^{m} \alpha_j s_j : d \in \Delta m\}$$

Then (1.2.2) is the convex quadratic minimization problem with m variables d_j.

$$\frac{1}{2}\left\| \sum_{j=1}^{m} \alpha_j s_j \right\|^2 \rightarrow \min, d \in \Delta m$$

Case 2: The subdifferential is a convex polyhedron (assumed compact) characterized by its supporting hyperplanes: m nonzero vectors $v_1, v_2, ..., v_m$ are given in $\Re^n$, together with m numbers $r_1, r_2, ..., r_m$ and

$$\partial f(x) := \{s : < s, r_i > \leq r_j \ for \ \ j = 1, 2, ..., m\}$$

Then (1.2.2) is again a convex quadratic minimization problem, but this time witn n variables and m inequality constraints

$$\min\{\tfrac{1}{2}\|s\|^2 : < s, v_j > \leq r_j \ for \ \ j = 1, 2, ..., m\}$$

Where $R : \Re^m \rightarrow \Re^n$ is a given linear mapping. Then (1.2.2) can be written as

$$\tfrac{1}{2}\|Rz + c\|^2 \rightarrow \min \ : z \in S$$

$$S = \{z \in \Re^m : \tfrac{1}{2}\|z\|^2 \leq \tfrac{1}{2}\}$$

2. The Finite Minimax Problems

Let

$$f(x) := \max\{f_j(x) : j = 1, 2, ...p\} \qquad \textbf{(2.0.1)}$$

where each $f_i : \Re^n \to \Re$ is convex and (continuously) differentiable and p is some given positive natural number.

The optimization problem

$$(P) : \quad f(x) \to \min, x \in \Re^n$$

is called a finite minimax problem.

It is assumed throughout that all the functions f_j are available together with their gradients.

2.1 The Steepest-Descent Method for Finite Minimax Problem

For each $x \in \Re^n$, the set

$$J(x) := \{j : f_j(x) = f(x)\} \qquad \textbf{(2.1.1)}$$

is called the active index set at x. For $j \in J(x)$ the function f_j and $\nabla f_j(x)$ are called respectively the active functions and active gradient (at x).

Now let us see the following important fact:

with the notation (2.0.1), (2.1.1)
$\partial f(x) = Conv\{u\partial_x f(x) : j \in j(x)\}$

with this fact in mind the following theorem is stated.

Theorem 2.1.1 The function f of (2.0.1) is convex. For given x, its subdifferential is the convex hull of the active gradient at x :

$$\partial f(x) = Conv\{\nabla f_j(x) : j \in J(x)\}$$

Proof: The second assertion follows from the above fact. We prove the first assertion, i.e., the convexity of f.

Let $J_1(z) := \{j : f_j(\lambda x + (1-\lambda)y) = f(\lambda x + (1-\lambda)y\}$ where $z = \lambda x + (1-\lambda)y$

$x \in \Re^n, y \in \Re^n, \lambda \in [0, 1]$

$$f(\lambda x + (1-\lambda)y) = f_j(\lambda x + (1-\lambda)y, \qquad j \in J_1(x)$$

$$\leq \lambda f_j(x) + (1-\lambda) f_j(y) \qquad (\text{convexity of } f_j, j \in J_1(x))$$

$$\leq \lambda f(x) + (1-\lambda) f(y) \qquad \text{Since } f(x) \geq f_j(x)\ \forall_j \in J(x)$$

Hence f is convex.

Thus, the subdifferential of such an f is a compact convex polyhedron having at most P extrem points. An actual computation of this polyhedron exactly amounts to performing the following which can be done by computer program:

- finding all the active indicies j at the given x.
- doing some ordinary differential calculus to compute the corresponding gradients.
- by Theorem 2.1.1 the subdifferential is then the set of all convex combinations of the gradient:

$$\partial f(x) = \{ \sum_{j \in J(x)} \alpha_j \nabla f_j(x) : \sum_{j \in J(x)} \alpha_j = 1, \alpha_j \geq 0 \text{ for } j \in J(x)\} \qquad \textbf{(2.1.2)}$$

Proposition 2.1.2 A necessary and sufficient condition for x to minimize f defined by (2.0.1) is that there exist coefficients $\alpha_j, J \in (x)$ satisfying.

$$\alpha_j \geq 0 \text{ for } j \in J(x), \sum_{j \in J(x)} \alpha_j = 1, \sum_{j \in J(x)} \alpha_j \nabla f_j(x) = 0 \ .$$

Proof: By definition of subdifferential x minimizes f if and only if $0 \in \partial f(x)$.
By Theorem 2.1.1.

$0 \in \partial f(x)$ iff there exist coefficients $\alpha_j, \alpha_j \geq 0$ *for*

$$j \in J(x), \sum_{j \in J(x)} \alpha_j = 1 \quad , \quad 0 = \sum_{j \in J(x)} \alpha_j \nabla f_j(x) \ . \ \textit{//} \ .$$

Proposition 2.1.3 A vector $d \in \Re^n$ is a descent direction if it satisfies the finite set of inequalities

$$< d, \nabla f_j(x) > \ < 0 \ \textit{for all} \quad j \in J(x) \qquad \textbf{(2.1.3)}$$

Actually, there holds

$$f'_+(x,d) = \max\{< \nabla f_j(x), d >: j \in J(x\} \qquad \textbf{(2.1.4)}$$

Proof: Let $t > 0$ be small enough. By the continuity of each f_j, those indices not in $J(x)$ do not count at $x + td$, i.e.

$$f(x+td) > f_j(x+td) \ \textit{for all} \ j \in J(x)$$

This means that $J(x+d) \subset J(x)$. So we can replace (2.0.1) by

$$f(x+td) = \max\{f_j(x+td) : j \in J(x)\} \qquad \textbf{(2.1.5)}$$

valid around x, Now a first order development of f_j yield

$$f(x+td) = \max\{f_j(x) + t < \nabla f_j(x), d > + \varepsilon_j(t) : j \in J(x)\}$$

$$= f(x) + t \max\{< \nabla f_j(x), d > + \varepsilon_j(t) : j \in J(x)\} \ \text{where}$$

$$\varepsilon_j(t) \to 0 \ \text{ for } t \downarrow 0$$

i.e, $\lim\limits_{t \downarrow 0} \frac{f(x+td)-f(x)}{t} = \lim\limits_{t \downarrow 0} [\max \ < \nabla f_j(x), d > + \varepsilon_j(x)]$

$$\Rightarrow \quad \lim_{t \downarrow 0} \frac{f(x+td)-f(x)}{t} = \max\{< \nabla f_j(x), d >: j \in J(x)\}$$

$$\Rightarrow \quad f'_+(x,d) = \max\{< \nabla f_j(x), d >: j \in J(x)\}$$

And d is descent if and only if $f'_+(x,d) < 0$

$$f'_+(x,d) < 0 \ \Leftrightarrow \ \max\{< \nabla f_j(x), d >: j \in J(x)\} < 0$$

$$\Leftrightarrow \ < \nabla f_j(x), d > \ < 0 \ \textit{for all} \ j \in J(x) \ . \ \textit{//}$$

Algorithm 2.1.4 (Steepest-Descent, Finite Minimax Problem)

Let the initial point $x_1 \in \Re^n$ and the tolerance $\delta > 0$ be given.

Step 1: **(Stopping Criterion)**
Solve the optimization problem

$$(P): \quad \left\| \sum_{j \in J(x)} \alpha_j \nabla f_j(x) \right\| \to \min, \alpha \in \nabla \qquad \textbf{(2.1.6)}$$

$$\Delta = \{\alpha \in \Re \Big| \sum_{j \in J(x)} \alpha_j = 1, \alpha_j \geq 0 \quad for \quad j \in J(x)\}$$

Let $s_k := \sum_{j \in J(x)} \alpha_j \nabla f_j(x_k)$ be the result. If $\|s_k\| \leq \delta$ stop.

Step 2: **(Direction Finding)** Solve

$$(P_d): \quad \|d\| \to \min, d \in S$$

$$S = \{d \in \Re^n < s_k, d > \; = -1, < \nabla f_j(x_k), d > \; \leq -1 \; for \; all \; j \in J(x)\}$$

Let $d_k \neq 0$ be the solution.

Step 3: **(Line Search)** Find a stepsize $t_k > 0$ and a new iterate $x_{k+1} = x_x + t_k d_k$ such that

$$f(x_k + t_k d_x) < f(x_k) \quad .$$

Step 4: **(Loop)** Replace k by $k+1$ and loop to step 1.

Consider the problem

$$(P): \quad \frac{1}{2} \left\| \sum_{j \in J(x)} \alpha_j \nabla f_j(x) \right\|^2 \to \min, \quad \alpha \in \Delta, x \in \Re^n \qquad \textbf{(2.1.7)}$$

$$\Delta = \{\alpha \in R : \sum_{j \in J(x)} \alpha_j = 1, \quad \alpha_j \geq 0 \; for \; all \; j \in J(x)\}$$

(2.1.7) may have several solutions, but they all make up the same vector $\sum_{j \in J(x)} \hat{\alpha}_j \nabla f_j(x) = \hat{s}$ for any solution $\hat{\alpha}$ of (2.1.7) which is the best approximation of $0 \in \Re^n$ with regard to the convex hull of active gradients of f at x.

Proposition 2.1.7 The above projection $\hat{s}$ is a unique convex combination

$$s \in conv\{\nabla f_j(x) : j \in J(x) \text{ satisfying}$$

$$< s, \nabla f_j(x) > \geq \|s\|^2 \text{ for all } j \in J(x) \qquad \textbf{(2.1.8)}$$

Equality holds in (2.1.8) for all j such that there is some $\hat{\alpha}$ solving (2.1.7) and having $\alpha_j > 0$.

Proof: From $\hat{s} = \sum_{j \in J(x)} \hat{\alpha} \nabla f_j(x)$, $\hat{s}$ is clearly a convex combination of active gradients of f at x where $\hat{\alpha}$ solves (2.1.7). Hence $\hat{s} \in \partial f(x)$. Since $\hat{\alpha}_j$ is a solution of (2.1.7)

$\hat{s} = \sum_{j \in J(x)} \hat{\alpha}_j \nabla f_j(x)$ is a solution of

(P) : $\frac{1}{2}\|s\|^2 \rightarrow \min, s \in \partial f(x)$

Hence $\hat{s}$ satisfy (2.1.8). The subdifferential, $\partial f(x)$ is convex, $\hat{s}$ is unique. To show equality in (2.1.8) assume that strict inequality holds in (2.1.8) for some j_0 with α_{j_0} with $\alpha_{j_0} > 0$ for some $\hat{\alpha}$ solving (2.1.7) (Proof by contradiction)

For j_0

$$< \hat{s}, \hat{\alpha}_{j_0} \nabla f_{j_0} > > \hat{\alpha}_{j_0} \left\|\hat{s}\right\|^2$$

$$< \hat{s}, \sum_{\substack{j \neq j_0 \\ j \in J(x)}} \hat{\alpha} \nabla f_j(x) > + < \hat{s}, \hat{\alpha}_{j_0} \nabla f_{j_0}(x) > > \hat{\alpha}_{j_0} \left\|\hat{s}\right\|^2 + \hat{s}, \sum_{\substack{j \neq j_0 \\ j \in J(x)}} \hat{\alpha} \nabla f_j(x) >$$

$$\Rightarrow < \hat{s}, \sum_{\substack{j \neq j_0 \\ j \in J(x)}} \hat{\alpha}_j \nabla f_j(x) + \alpha_{j_0} \nabla f_{j_0}(x) > > \alpha_{j_0} \left\|\hat{s}\right\|^2 + \sum_{\substack{j \neq j_0 \\ j \in J(x)}} < \hat{s}, \nabla f_j(x) >$$

$$\geq \hat{\alpha}_{j_0} \left\|\hat{s}\right\|^2 + \sum_{j \neq j_0} \alpha_j \left\|\hat{s}\right\|$$

$$\Rightarrow < \hat{s}, \hat{s} > > \left\|\hat{s}\right\| \text{ which is a contradiciton.}$$

Therefore strict inequality holds for no j_0 with $\alpha_{j_0 > 0}$. //

2.2 Non-convergence of Steepest Descent Method

Counter Example

Let us consider a counter example which illustrate that the steepest -descent method may not be convergent.

Consider the following functions of $x = (\varepsilon, \eta) \in \Re^2$

i.e. $f_j : \Re^2 \to \Re \qquad j = 1, 2, 3, 4, 5$

$$\begin{aligned} f_1(x) &:= -100 \\ f_2(x) &:= 2\varepsilon + 3\eta \\ f_3(x) &:= -2\varepsilon + 3\eta \\ f_4(x) &:= 5\varepsilon + 2\eta \\ f_5(x) &:= -5\varepsilon + 2\eta \end{aligned} \qquad \textbf{(2.2.1)}$$

Set

$$f(x) := \max\{f_j(x) : j = 1, 2, 3, 4, 5\}$$

Let us concentrate on the region $\eta \geq 0$, in which f is non-negative and f_1 does not count. Then ∇f fails to exist on the half lines

$$L_1 := \{x : 0 \leq \eta = 3\varepsilon\} \quad \textit{where } f_2 = f_4$$

$$L_2 := \{x : 0 \leq \eta = -3\varepsilon\} \quad \textit{where } f_3 = f_5$$

$$L_3 := \{x : \varepsilon = 0\} \quad \textit{where } f_2 = f_3$$

This is illustrated by Fig 2.2.1, which shows, a level set of f, then the three critical lines and four possible values for ∇f. The minimal value of f is clearly -100, attained for sufficiently negative values of η.

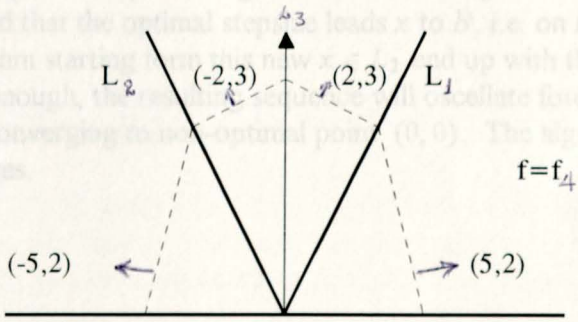

Fig. 2.2.1 A counter-example for steepest-descent

Let the current point x of Algorithm 2.1.4 be in the first quadrant such that f_∂ is active; suppose for example that $x \in L_1$. The steepest descent direction is $-\nabla f_2(x)$, i.e. (-2,-3) and one easily observes that it is still (-2,-3) even $x \notin L_1$.

Since

$$16 = < \nabla f_2(x), \nabla f_4(x) > \; > \|\nabla f_\partial(x)\|^2 = 13$$

by the above proposition $\nabla f_2(x)$ is the best approximation of $0 \in \Re^2$ with regard to $[\nabla f_2(x), \nabla f_4(x)]$.

Fig. 2.2.2 shows descent direction d. It is simple to observe that the straight line $x + \Re d$ passes above the origin; the reason is that $\frac{3}{2}$, the slope of $\nabla f_2(x)$, is smaller than 3 the slope of L_1.

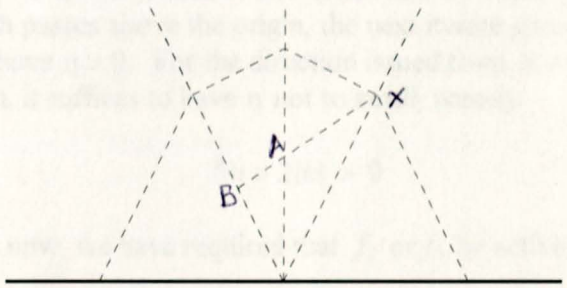

Fig. 2.2.2 Non-convergence of the steepest descent

The one dimensional function $t \to f(x + td)$ is piecewise affine with two kinks (= sharp corners) A and B. If we take a "reasonable" stepsize along this direction, for example the optimal stepsize. Fig. 2.2.3 shows that f is decreasing along the segment AB, and that the optimal stepsize leads x to B, i.e. on L_2. By symmetry, the same Algorithm starting form this new $x \in L_2$ end up with the next iterate on L_1. Clearly enough, the resulting sequence will oscellate forever between L_1 and L_2 and converging to non-optimal point $(0, 0)$. The algorithm is subject to zigzages.

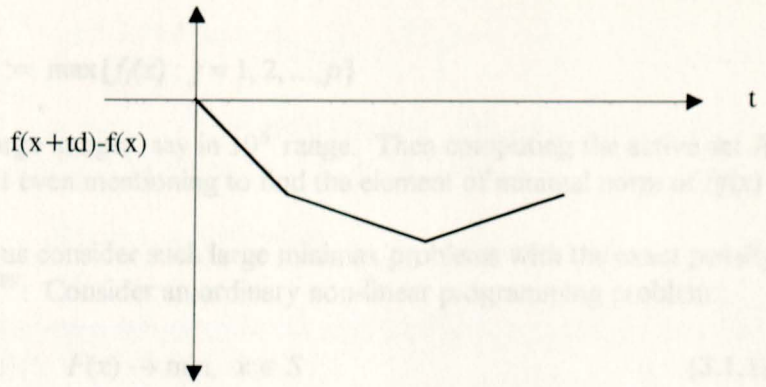

Fig. 2.2.3 Objective value along steepest-descent

Remark 2.2.1 The whole idea of this construction is that, for each iterate, the direction search passes above the origin, the next iterate given by the line search, will then have $\eta > 0$. For the direction issued from $x = (\varepsilon, \eta)$ to pass above the origin, it suffices to have η not to small, namely

$$5\eta > 2|\varepsilon| > 0 \qquad \textbf{(2.2.2)}$$

up to now, we have required that f_2 or f_3 be active at each iterate, i.e.

$$\eta \geq 3\|\varepsilon\| (> 0) \qquad \textbf{(2.2.3)}$$

which implies (2.2.2). If (2.2.3) doesnot hold, the direction become (5,-2) or (-5,-2). Redrawing Fig. 2.2.2 we see the optimal step size would lead the next iterate to A. The next direction would be vertical, pointing directly down to the optimal set; the counter-example would disappear.

3. The practical value of descent schemes

Section 2 of this paper was mainly devoted to the zigzagging phenomenon, common to all steepest descent methods. Another problem is that the practical implementation of such methods could be difficult. The full subdifferential had to be computed. The aim of present section is to show that, in many situations, such a computation is not convenient.

3.1 Lagre Minimax Problems

Let

$$f(x) := \max\{f_j(x) : j = 1, 2, ..., p\}$$

but p is a large integer, say in 10^6 range. Then computing the active set $J(x)$ is difficult, not even mentioning to find the element of minimal norm of $\partial f(x)$ at x.

Let us consider such large minimax problems with the exact penalty techniques[APP]: Consider an ordinary non-linear programming problem:

$$(P): \quad F(x) \to \min, \quad x \in S \qquad \textbf{(3.1.1)}$$

$$S = \{x \in \Re^n : c_j(x) \leq 0 \text{ for } j = 1, 2, ..., p\}$$

where both $F : \Re^n \to \Re$ and $c_j : \Re^n \to \Re$, convex smooth functions, and p is an extremely large number. Known methods for constrained optimization become impractical in this situation. The penalty idea is to transform the problem by aggregating the constraints into the objective function.

Let us choose the penalty coefficient π large enough

Then

$$(P): \quad F(x) + \pi \max\{0, c_1(x), c_2(x), ..., c_p(x)\} \to \min, x \in \Re^n \qquad \textbf{(3.1.2)}$$

From this one observes that the max operation involves $p + 1$ terms.

Remark 3.1.1 **Consider the function**

$$\Re^n \quad x \to F_\pi(x) := F(x) + \pi \sum_{j=1}^{p} \max\{0, c_j(x)\} \qquad \textbf{(3.1.3)}$$

This function can be put in the minimax framework of section 2.

$$F_\pi(x) = F(x) + \pi \max\{\varepsilon_1 c_1(x), + \varepsilon_2 c_2(x) + ... + \varepsilon_p c_p(x) : \varepsilon_j \in \{0, 1\}$$

So F_π is max of 2^p functions. To characterize $\partial F_\pi(x)$ from this last expression, denote by

---

APP: See Appendix

$J(x) := \{j : c_j(x) = 0\}$ the set of active indices a given x, and by

$$s_0(x) := \nabla F(x) + \pi \sum_{\{f : c_j(x) > 0\}} \nabla c_j(x)$$

the "smooth part" of the differentiation. The subdifferential of F_π at x is therefore the convex hull of $2^{|J(x)|}$ points:

$$\partial F_\pi(x) = s_0(x) + \pi Conv\{\sum_{j \in J(x)} \varepsilon_1 \nabla c_j(x) : \varepsilon_j \in \{0, 1\}\}$$

computing a steepest descent direction is now a convex minimization problem with $|J(x)|$ variable α_j:

$$(P): \quad \left\| s_0(x) + \pi \sum_{j \in J(x)} \alpha_j \nabla c_j(x) \right\| \rightarrow \min, \ \alpha_j \in S$$

$$S = \{\alpha \in \Re \, 0 \le \alpha_j \le 1 \ for \quad j \in J(x)\}$$

3.2 Infinite Minimax Problems

Consider

$$f(x) := \max\{h(x, y) : y \in Y\} \qquad \textbf{(3.2.1)}$$

where h is convex in x and smooth on the compact set Y . Among such problems, are the frequently encountered semi-infinite programs (= optimization problems with finite variable and with infinitly many constraints.) Consider

$$F(x) \rightarrow \min, \ x \in S$$

$$S = \{x \in \Re^n : g(x, t) \le 0 \quad \forall t \in T\}$$

Where T is compact interval in $\Re$, and

$$g(x, t) := \sum_{i=1}^{n} \varepsilon_i \varphi_i(t) - \varphi_0(t) = [\varphi(t)]^T x - \varphi_0(t)$$

$$x \in \Re^n, \ x = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$$

and $\varphi_i : T \rightarrow \Re \qquad i = 0, 1, \ldots, n$ continuous.

Clearly to compute the whole subdifferential $\partial f(x)$ is very difficult in (3.2.1), which amounts to computing the whole, potentially infinite, set of maximal $y's$ at x. On the other hand, computing some subgradient of f in (3.2.1) is also hard: the underlying maximization has no reason to be easy.

3.3 Smooth but Stiff Functions

Section 3.1 and Section 3.2 of this paper dealt with problems in which computing the subdifferential was difficult, or even impossible. In other field of applications, computing it is simply meaningless. This concerns objective functions whose gradient varies rapidly, although continuously.

There is no clear-cut between functions that are smooth and functions that are not. Between these two classes, there is a rather fuzzy boundary of stiff functions, for which it is not clear that whether smooth optimization or non smooth optimization is better suited.

A non-smooth function f can be regarded as the limiting case of twice differentiable function say g , whose second derivatives grow unboundedly at some points (the kink of f).

The question is:

> When is a smooth function so stiff that an algorithm tailored for smooth functions will become inefficient ?

The following experiment illustrates how vague the class of stiff function is.

The objective function in (2.0.1) can be written

$$f(x) = \max_{\alpha \in \Delta p} \sum_{j=1}^{p} \alpha_j f_j(x) \ . \qquad \textbf{(3.3.3)}$$

Where ΔP is the unit simplex.[APP]

APP: See Appendix

Now let us take $\pi > 0$ and set

$$\varphi^{\pi}(x, \alpha) := \sum_{j=1}^{p} \alpha_j f_j(x) + \pi \sum_{j=1}^{p} \log$$

For small $\pi > 0$, φ^{π} approximates the maximand in (3,3,3) and the function

$$f^{\pi}(x) : = \max\{\varphi^{\pi}(x, d) : \sum_{j=1}^{p} \alpha_j = 1\} \qquad \textbf{(3.3.4)}$$

approximates f.

For $\alpha_j \in \Delta p$, $\log \alpha_j \leq 0$. Hence we have

$$f^{\pi}(x) = \max\{\varphi^{\pi}(x, \alpha) : \sum_{i=1}^{p} \alpha_j = 1\}$$

$$\leq \sum_{\substack{\alpha=1 \\ \alpha \in \Delta p}}^{p} \alpha_j f_j(x) \leq \max_{\alpha \in \Delta p} \sum_{j=1}^{p} \alpha_j f_j(x) = f(x)$$

$$\text{and } \lim_{\pi \to 0} f^{\pi}(x) = f(x)$$

In contrast with (3.3.3), the α-problem (3.3.4) making up f^{π} has a unique maximal solution, so the resulting f^{π} is now differentiable. Now we have to minimize with respect x a smooth $f^{\pi}(x)$, instead of the nonsmooth f(x).

Test Problem 3.3.3 (MAXQUAD)

Let $f_j : \Re^n \to \Re$ be given by

$$f_j(x) := \tfrac{1}{2} x^T A_j x + b_j^T(x) + c \qquad \textit{for } J = 1, \ldots, p$$

where each f_j is symmetric positive definite $n \times n$ matrix and b_j an n-vector; c real number.

Let $x = (\varepsilon_1, \ldots, \varepsilon_n) \in \Re^n$. $A_j = (\alpha_{ik})_{i,\, k = 1, 2, \ldots, n}$

$$f_j(x) = \tfrac{1}{2}(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) \begin{pmatrix} \sum_{i=1}^{n} \alpha_{1i} \varepsilon_i \\ \vdots \\ \sum_{i=1}^{n} \alpha_{ni} \varepsilon_i \end{pmatrix}_j + (b_1, b_2, \ldots, b_n)_j \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} + c$$

$$\nabla f_j(x) = \begin{pmatrix} \frac{\partial f_j(x)}{\partial \varepsilon_1} \\ \vdots \\ \frac{\partial f_j(x)}{\partial \varepsilon_n} \end{pmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \alpha_{1i}\varepsilon_i \\ \vdots \\ \sum_{i=1}^{n} \alpha_{ni}\varepsilon_i \end{bmatrix}_j + \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}_j$$

$$= A_j(x) + b_j$$

Let $f(x) := \max\{f_j(x) : j = 1,2,3,4,5\}$ (*)

In the specific example called MAXQUAD, $n = 10, p = 5, c = 0.8414$ and the minimal value is $f(\bar{x}) = 0$. At the unique minimal $\bar{x}$, the understanding functions have characterstics listed in Table 3.3.1.

Table 3.3.1

j	1	2	3	4	5
$f_j(\bar{x})$	0	0	0	0	-298
$\|\nabla f_j(\bar{x})\|$	6	14	40	500	10^{-4}

Now, the objective function f of (*) can be approximated by the smooth function (3.3.4) which inturn can be minimized by and "ordinary" method.

Table 3.3.2 displays the behaviour of various algorithm for different values of π. Each entry of Table 2.3.2 contains number of iterations to reach to three exact digits, and between parentheses, the corresponding number of f^π - and ∇f^π - evaluations. The last two rows indicate the value of f^π at its minimum x^π and the corresponding values of f. All methods were started from the same initial iterate and used the same line search.

π	100	10	1	10^{-1}	10^{-2}	10^{-3}	0
Steepest descent	2(6)	21(35)	30(57)	59(97)	358(457)	$\infty(\infty)$	failed
Conjugate gradient	5(15)	10(23)	13(33)	20(50)	77(222)	69(194)	failed
Nonsmooth	3(16)	6(10)	12(20)	15(22)	24(46)	24(54)	17(44)
$f^\pi(x^\pi)$	-1,117	-121	-14	-1.68	-0.21	-0.02	0
$f(x^\pi)$	184	8.7	1.6	0.17	0.01	0	0

Table 3.3.2 What is a smooth function ?

This table makes clear enough the danger of believing that a "smooth method" is automatically appropriate for a smooth function.

Appendix

1. Exact Penalty

Let $f: \Re^n \to \Re$ be convex function. Consider

$$(P) \ : \ f(x) \to \min, \ x \in C$$

$$C \subseteq \Re^n, \ C \ \textit{Convex}$$

Let P satisfy

$$p(x) = \begin{cases} 0 & \textit{if } x \in C \\ P(x) > 0 & \textit{if } x \notin C \end{cases} \qquad (*)$$

Consider

$$(P_\pi) \quad f(x) + \pi P(x) \to \min, \ x \in \Re^n \qquad (**)$$

Definition: Let p has a nonempty solution set. A penalty function P satisfying $(*)$ is said have the exact penalty property if there is $\pi > 0$ such that (p_π) has a solution belonging to C.

2. Unit Simplex

Any collection of numbers $\{\alpha_1, ..., \alpha_k\}$ satisfying

$$\alpha_i \geq 0 \ \textit{ for } \ i = 1, ..., k \ \textit{ and } \ \sum_{i=1}^{k} \alpha_i = 1$$

is called the unit simplex of $\Re^k$.

References

Jean - Baptiste Hiriart-Urruty, and

Claude Lemarechal: Convex Analysis and Minimization Algorithms I.
Springer - Verlag (1993)