Math 273a: Optimization
Subgradients of convex functions

Made by: Damek Davis
Edited by Wotao Yin

Department of Mathematics, UCLA
Fall 2015

online discussions on piazza.com
Subgradients

Assumptions and Notation

1. $f : \mathbb{R}^n \rightarrow \mathbb{R}^n \cup \{\infty\}$ is a closed proper convex function, i.e.

   $\text{epi} f = \{(x, t) \in \mathbb{R}^n \times \mathbb{R} | f(x) \leq t\}$

   is closed and convex.

2. The effective domain of $f$ is

   $\text{dom} f = \{x \in \mathbb{R}^n : f(x) < \infty\}$

3. The function $f$ is proper, i.e. $\text{dom} f \neq \emptyset$.

4. A raised * (e.g., $x^*$) means global minimum of some function.
Recall: a convex function of $C^1$ obeys

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall x, y \in \mathbb{R}^n$$
Non-$C^1$ convex functions

For all each $\bar{x} \in \mathbb{R}^n$,

$$\partial f(\bar{x}) := \{ g \in \mathbb{R}^n : f(y) \geq f(\bar{x}) + \langle g, y - \bar{x} \rangle \}$$
Which functions have subgradients?

- If \( f \in C^1 \), then \( \nabla f(x) \in \partial f(x) \).
- In fact, \( \partial f(x) = \{\nabla f(x)\} \).

**Proof:** if \( g \) is a subgradient, then for \( y \in \mathbb{R}^n \)

\[
\langle \nabla f(x), y \rangle = \lim_{t \to 0} \frac{f(x + ty) - f(x)}{t} \\
\geq \lim_{t \to 0} \frac{\langle g, x + ty - x \rangle}{t} \\
= \langle g, y \rangle.
\]

Change \( y \) to \( -y \), and the inequality still holds. \( \implies \langle \nabla f(x), y \rangle = \langle g, y \rangle \).

Plugging in standard basis vectors \( \implies \nabla f(x) = g \).

- Next, the general case.
Which functions have subgradients?

Theorem (Nesterov’03 Thm 3.1.13)

Let $f$ be a closed convex function and $x_0 \in \text{int}(\text{dom}(f))$. Then $\partial f(x_0)$ is a nonempty bounded set.

Proof uses supporting hyperplanes of epigraph to show existence, and local lipschitz continuity of convex functions to show boundedness.

---

The converse

Lemma (Nesterov’03 Lm 3.1.6)

*If \( \partial f(x) \neq \emptyset \) for all \( x \in \text{dom}(f) \), then \( f \) is convex.*

Proof.

\[ x, y \in \text{dom}(f), \alpha \in [0,1], y_\alpha = (1 - \alpha)x + \alpha y = x + \alpha(y - x), \ g \in \partial f(y_\alpha). \]

\[
\begin{align*}
    f(y) & \geq f(y_\alpha) + \langle g, y - y_\alpha \rangle = f(y_\alpha) + (1 - \alpha)\langle g, y - x \rangle \quad (1) \\
    f(x) & \geq f(y_\alpha) + \langle g, x - y_\alpha \rangle = f(y_\alpha) - \alpha \langle g, y - x \rangle \quad (2)
\end{align*}
\]

- Multiply equation (1) by \( \alpha \) and equation (2) by \( (1 - \alpha) \).
- Add them together to get

\[
\alpha f(y) + (1 - \alpha) f(x) \geq f(y_\alpha)
\]
Technicality: \( \text{int}(\text{dom}(f)) \)

We cannot relax the assumption \( x \in \text{int}(\text{dom}(f)) \) to \( x \in \text{dom}(f) \).

Example:

\[
f : [0, +\infty) \rightarrow \mathbb{R}. \quad f(x) = -\sqrt{x}.
\]

\( \text{dom}(f) = [0, +\infty) \) but

\[
\partial f(0) = \emptyset.
\]
Compute subgradients: general rules

► Smooth functions: \( \partial f(x) = \{ \nabla f(x) \} \).

► Composition with affine: \( \phi(x) = f(A(x) + b) \)

\[
\partial \phi(x) = A^T \partial f(A(x) + b).
\]

► Positive sums: \( \alpha, \beta > 0, f(x) = \alpha f_1(x) + \beta f_2(x) \).

\[
\partial f(x) = \alpha \partial f_1(x) + \beta \partial f_2(x)
\]

► Maximums: \( f(x) = \max_{i \in \{1, \ldots, n\}} \{ f_i(x) \} \)

\[
\partial f(x) = \text{conv}\{ \partial f_i(x) | f_i(x) = f(x) \}
\]
Examples

**f(x) = |x|**.

\[
\partial f(x) = \begin{cases} 
\{\text{sign}(x)\} & x \neq 0; \\
[-1, 1] & \text{otherwise}
\end{cases}
\]

\[f(z) = |z|\]

\[\partial f(x)\]

---

\(^3\)figure taken from Boyd and Vandenberghe, 

Examples

\( f(x) = \sum_{i=1}^{n} |\langle a_i, x \rangle - b_i| \). Define

\[
I_-(x) = \{ i | \langle a_i, x \rangle - b_i < 0 \} \\
I_+(x) = \{ i | \langle a_i, x \rangle - b_i > 0 \} \\
I_0(x) = \{ i | \langle a_i, x \rangle - b_i = 0 \}.
\]

Then

\[
\partial f(x) = \sum_{i \in I_+(x)} a_i - \sum_{i \in I_-(x)} a_i + \sum_{i \in I_0(x)} [-a_i, -a_i]
\]

\( f(x) = \max_{i \in \{1, \ldots, n\}} x_i \). Then

\[
\partial f(x) = \text{conv}\{e_i | x_i = f(x)\} \\
\partial f(0) = \text{conv}\{e_i | i \in \{1, \ldots, n\}\}
\]
Examples

▸ $f(x) = \|x\|_2$. $f$ is differential away from 0, so:

$$\partial f(x) = \frac{x}{\|x\|_2} \quad x \neq 0.$$ 

At 0, go back to subgradient equation:

$$\|y\|_2 \geq 0 + \langle g, y - 0 \rangle$$

Thus, $g \in \partial f(0)$, if, and only if, $\frac{\langle g, y \rangle}{\|y\|_2} \leq 1$ for all $y \neq 0$. Thus, $g$ is in the dual ball $B_2^*(0, 1) = B_2(0, 1)$.

▸ This is a common pattern!
How to compute subgradients: Examples

$\n f(x) = \|x\|_\infty = \max_{i \in \{1, \ldots, n\}} |x^{(i)}|.
\n\partial f(x) = \text{conv}\{-e_i, e_i\} |x^{(i)}| = f(x) \quad x \neq 0.

Going back to subgradient equation

$\|y\|_\infty \geq 0 + \langle g, y \rangle$

Thus, $g \in \partial f(0)$, if, and only if, $\frac{\langle g, y \rangle}{\|y\|_\infty} \leq 1$ for all $y \neq 0$. Thus, $\partial f(0)$ is the dual ball to the $l_\infty$ norm: $B_1(0, 1)$. 
$f(x) = \|x\|_1 = \sum_{i=1}^{n} |x_i|$. Then

\[
\partial f(x) = \sum_{x_i > 0} e_i - \sum_{x_i < 0} e_i + \sum_{x_i = 0} [-e_i, e_i]
\]

for all $x$. Then

\[
\partial f(0) = \sum_{i=1}^{n} [-e_i, e_i] = B_\infty(0, 1).
\]
Semi-continuity

**Definition (upper semi-continuity)**

A point-to-set mapping $M$ is upper semi-continuous at $x'$ if any convergent sequence $(x^k, s^k) \to (x', s')$ satisfying $s^k \in M(x^k)$ for each $k$ also obeys

$$s' \in M(x').$$

**Interpretation:** if (i) $x^k \to x'$ and (ii) for each $x^k$ you can find $s^k \in M(x^k)$ so that $s^k \to s'$, then $s' \in M(x')$.

Lower semi-continuity is essentially the opposite.

**Definition (lower semi-continuity, lsc)**

A point-to-set mapping $M$ is lower semi-continuous if any convergent sequence $x^k \to x'$ and $s' \in M(x')$, there exists sequence $s^i \in M(x^{k_i})$ such that

$$s^i \to s'.$$
Lemma

Let $f$ be a proper convex function. $\partial f$ is upper semi-continuous, and $\partial f(x)$ is a convex set.

Proof: Take the limit of

$$f(y) \geq f(x^k) + \langle s^k, y - x^k \rangle, \quad s^k \in \partial f(x^k).$$

The second part is a direct result of linearity of $\langle \cdot, y - x \rangle$.

However, if $f(x) = |x|$, the $\partial f$ is not lower semi-continuous at $x = 0$.  

Directional derivative versus Subgradient

Assume that $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a proper, closed (thus lsc), and convex. Then

1. the directional derivative is well defined for every $d \in \mathbb{R}^n$:

$$f'(x; d) = \lim_{\alpha \downarrow 0} \frac{f(x + \alpha d) - f(x)}{\alpha}.$$ 

2. the directional derivatives at $x$ bound all the subgradient projections

$$\partial f(x) = \{p \in \mathbb{R}^n : f'(x; d) \geq \langle p, d \rangle, \ \forall \ d \in \mathbb{R}^n\}.$$ 

3. directional derivatives are extreme subgradients:

$$f'(x; d) = \max\{\langle p, d \rangle : p \in \partial f(x)\}.$$
0 subgradient $\implies$ minimum

- Suppose that $0 \in \partial f(x)$.
- If $f$ is smooth and convex, $0 \in \partial f(x) = \{\nabla f(x)\} \implies \nabla f(x) = 0$.
- In general: If $0 \in \partial f(x)$, then

$$f(y) \geq f(x) + \langle 0, y - x \rangle$$

$$= f(x)$$

for all $y \in \mathbb{R}^n$.

$\implies x$ is a minimum!

- Converse also true: $f(y) \geq f(x^*) + 0 = f(x^*) + \langle 0, y - x^* \rangle$. 
Subgradient method

Iteration:

\[ x^{k+1} \leftarrow x^k - \alpha^k p^k \]

where \( p^k \in \partial f(x^k) \).

Applications:

- find \( x^* \in \bigcap_{i=1}^m C_i \) by \( \min f(x) = \max \{ \text{dist}(x, C_1), \ldots, \text{dist}(x, C_m) \} \)
- minimize non-smooth convex functions, e.g., SVM with hinge loss
- dual ascent method (typically, non-smooth), dual decomposition

Step size and convergence: assumption \( \|p^k\| \leq G \) uniformly

- fix \( \alpha^k \equiv \alpha \). While \( k < O(\frac{1}{\alpha^2 G^2}) \), \( f_{\text{best}}^k - f^* \leq O(\frac{1}{\alpha k}) \).
  
  Larger \( \alpha \Rightarrow \) faster, less accurate. Smaller \( \alpha \Rightarrow \) slower, more accurate.

- diminishing \( \alpha_k \): \( \lim \alpha_k \to 0 \) and \( \sum_k \alpha_k = \infty \), then \( f_{\text{best}}^k - f^* \leq O(\frac{1}{\sqrt{k}}) \).
A negative subgradient may not be a descent direction!

Consider

\[ f(x) = |x_1| + 2|x_2| \]

At \( x = (1, 0) \),

\[ \partial f(x) = \{(1, \alpha)^T : \alpha \in [-2, 2]\}. \]

\( d = -(1, 2)^T \in -\partial f(x) \) but for any small \( \alpha > 0 \),

\[ f(x + \alpha d) = |1 - \alpha| + 2|\alpha| > 1 = f(x). \]

Consequences:

- iterative monotonicity of \( f^k \) is not generally guaranteed
- line search (highly effective in gradient descent) may not help here
Theorem (Nesterov’03 Thm 3.1.16)

Let \( x_0 \in \mathbb{R}^n \). Then all \( g \in \partial f(x_0) \) define supporting hyperplanes to the lower level set \( \mathcal{L}_f(f(x_0)) = \{x|f(x) < f(x_0)\} \):

\[
\langle g, x_0 - x \rangle \geq f(x_0) - f(x) \geq 0.
\]

Thus, each \( g \in \partial f(x_0) \) cuts the search space for \( x^* \) in half:

\[
\langle g, x_0 - x^* \rangle \geq 0
\]

Theorem

Define \( \mathcal{H}_f = \{ \text{affine function } h \text{ such that } h(x) \leq f(x) \ \forall \ x \in \mathbb{R}^n \} \). Then

\[
f(x) = \sup\{h(x) : h \in \mathcal{H}_f\}.
\]

These results motivate the cutting plane method.
Cutting plane method: a demonstration

Cutting plane method: a demonstration

Cutting plane method: a demonstration

Cutting plane method: a demonstration

Cutting plane method: a demonstration

Cutting plane method: a demonstration

\[9\]

Cutting plane method: a demonstration

Cutting plane method: a demonstration

Cutting plane method: a demonstration

Initialize compact set $C$ containing the minimizer, tolerance $\epsilon > 0$, $k = 1$, $x^1 \in C$, and $h_0 = -\infty$.

Iterate:

1. compute $p^k \in \partial f(x^k)$;
2. construct piece-wise affine function
   \[ h_k(x) = \max\{h_{k-1}(x), f(x^k) + \langle p^k, x - x^k \rangle\}; \]
3. find $x^{k+1} \in \arg\min_{x \in C} h_k(x)$;
4. compute $\epsilon_k = f(x^k) - h_k(x^k)$;
5. if $\epsilon_k < \epsilon$, STOP; otherwise, continue with $k \leftarrow k + 1$. 
Remarks:

- at every iteration, $x^{k+1} \in \arg \min_{x \in C} h_k(x)$ is an LP since $h_k$ is piece-wise maximum of affine functions
- the LP' size increases with $k$!
- the stopping condition is reliable
  if $f(x^k) - h_k(x^k) < \epsilon$, then since

$$\min h_k(x) \leq \min f(x)$$

for any $k$, we have

$$f(x^k) \leq h_k(x^k) + \epsilon = \min h_k(x) + \epsilon \leq \min f(x) + \epsilon.$$  

It is more reliable than the subgradient method, which often uses unreliable $\|p^k\|$.

- but, it possibly takes big, zig-zagging steps
Bundle methods

A bundle is referred to as \( \{x^k, f(x^k), p^k\} \) where \( p^k \in \partial f(x^k) \).

Initialize: tolerance \( \epsilon > 0, \gamma \in (0, 1), k = 1, \hat{x}^1 = x^1 \in C \), and \( h_0 = -\infty \).

Iterate:

1. compute \( p^k \in \partial f(x^k) \);

2. construct piece-wise affine function

\[
  h_k(x) = \max\{h_{k-1}(x), f(x^k) + \langle p^k, x - x^k \rangle\};
\]

3. find \( x^{k+1} \in \arg\min_{x \in C} h_k(x) + \frac{\mu_k}{2} \|x - \hat{x}^k\|^2 \);

4. compute \( \epsilon_k = f(\hat{x}^k) - [h_k(x^{k+1}) + \frac{\mu_k}{2} \|x^{k+1} - \hat{x}^k\|^2] \);

5. if \( \epsilon_k < \epsilon \), STOP; else, continue;

6. if \( f(\hat{x}^k) - f(x^{k+1}) \geq m\epsilon_k \), then serious step \( \hat{x}^{k+1} \leftarrow x^{k+1} \); else, null step \( \hat{x}^{k+1} \leftarrow \hat{x}^k \);

7. \( k \leftarrow k + 1 \).
Remarks:

- Bundle algorithm (BA) is a stabilized cutting plane algorithm
- next iterate is closer to the current $\hat{x}^k$ to avoid drastic moves
- let $K_s := \{k : a \text{ serious step is taken at iteration } k\}$
- step 6 ensures strictly decreasing $f(\hat{x}^k), k \in K_{\text{serious}}$
- the presented BA is a basic version; several enhancements exist
- convergence under non-growing # of constraints
- has convergence analysis assuming bounded $\mu_k$

<table>
<thead>
<tr>
<th>Method</th>
<th>Per-itr cost</th>
<th>Iteration #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgradient</td>
<td>Update a vector</td>
<td>Very high</td>
</tr>
<tr>
<td>Cutting plane</td>
<td>LP</td>
<td>Medium/High</td>
</tr>
<tr>
<td>Bundle method</td>
<td>QP</td>
<td>Small/Medium</td>
</tr>
</tbody>
</table>
Karush-Kuhn-Tucker conditions

Theorem (Kuhn-Tucker, Nesterov’03 Thm 3.1.17)

Let $f_i, i = 0, \cdots, m$, be $C^1$ convex functions.

► Suppose there exists $\bar{x}$ such that $f_i(\bar{x}) < 0$, for $i = 1, \cdots, m$.

Then a point $x^*$ is a solution to

$$\minimize_{x \in \mathbb{R}^n} f_0(x) \quad \text{s.t.} \quad f_i(x) < 0, \quad i = 1, \cdots, m,$$

if, and only if, there exists $\lambda_i \geq 0$, such that

$$\nabla f_0(x^*) + \sum_{\{i \mid f_i(x) = 0\}} \lambda_i \nabla f_i(x^*) = 0$$
Karush-Kuhn-Tucker conditions

Proof.

Suppose $x^*$ is a solution to (3), and $f^* := f(x^*)$.

Define

$$
\phi(x) = \max_{x \in \mathbb{R}^n} \{f(x) - f^*, f_1(x), \ldots, f_m(x)\}.
$$

Suppose $\hat{x}$ is a minimum of $\phi$. If $\phi(\hat{x}) < 0$, then $f(\hat{x}) < f^*$ and $f_i(\hat{x}) < 0$. Thus, $\hat{x}$ produces a strictly smaller objective value than $x^*$ does. It is also feasible. This is clearly a contradiction.

Thus, $\phi(x^*) = 0$, and $x^*$ is the minimum of $\phi$.

(Cont.)
Karush-Kuhn-Tucker conditions

Proof (Cont.)

How can we get an expression for $x^*$?

Subgradients! $x^*$ is a minimum of $\phi$ if, and only if,

$$0 \in \partial \phi(x^*) = \text{conv}\{\nabla f_i(x^*) : i \in I_0\},$$

where $I_0 = \{0\} \cup \{i | f_i(x^*) = 0\}$.

This is true if, and only if, there exists $\alpha_i \geq 0$, $i \in I_0$, such that

$$\alpha_0 + \sum_{i \in I_0} \alpha_i = 1 \quad \text{and} \quad \alpha_0 \nabla f(x^*) + \sum_{i \in I_0} \alpha_i \nabla f_i(x^*) = 0.$$ 

If $\alpha_0 \neq 0$, we’re done (divide by $\alpha_0$).

(Cont.)
Karush-Kuhn-Tucker conditions

Proof (Cont.)

- Suppose that $\alpha_0 = 0$. Remember $\bar{x}$, in the interior of the feasible set?

\[
\sum_{i \in I_0} \alpha_i f_i(\bar{x}) = \sum_{i \in I_0} \alpha_i (f_i(\bar{x}^*) + \langle \nabla f(\bar{x}^*), \bar{x} - \bar{x}^* \rangle) = 0.
\]

But $f_i(\bar{x}) < 0$ for all $i \in I_0$, and there exists $\alpha_i > 0$.

- This is a contradiction

\[
\implies \lambda_i = \frac{\alpha_i}{\alpha_0} \geq 0.
\]
The expression

\[ L(x, \lambda) := f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \]

is called the Lagrangian of (3). Note that we restrict \( \lambda \geq 0 \).

Lagrangian is a relaxation: If \( f_i(x) \leq 0 \) \( i = 1, \ldots, m \), then \( L(x, \lambda) < f_0(x) \).

For fixed \( \lambda \geq 0 \),

\[ \inf_{x} L(x, \lambda) \leq f_0(x^*) . \]

Note that

\[ \sup_{\lambda \geq 0} L(x, \lambda) = \begin{cases} f_0(x) & \text{if } f_i(x) < 0 \text{ for all } i > 0; \\ \infty & \text{otherwise} \end{cases} \]

Thus,

\[ \inf_{x} \sup_{\lambda \geq 0} L(x, \lambda) = f_0(x^*) . \]
If we set
\[ \nabla_x L(x, y) = \nabla f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) = 0 \]
we get a constraint for a characterization of \( x_\lambda \) for each \( \lambda \geq 0 \).

The previous theorem shows that \( x^* = x_{\lambda x^*} \) for some \( \lambda \geq 0 \).

In particular, it showed that \( \lambda_i = 0 \) for all \( f_i(x) < 0 \). Thus,
\[
L(x^*, \lambda_{x^*}) = f_0(x^*) + \sum_{i=1}^{n} \lambda_{x^*,i} f_i(x^*) \\
= f_0(x^*)
\]
Strong duality

If we take the supremum

\[ f_0(x^*) = L(x^*, \lambda x^*) \]

\[ = \sup_{\lambda \ge 0} L(x_{\lambda}, \lambda) \]

\[ = \sup_{\lambda \ge 0} \inf_x L(x, \lambda) \]

\[ \le \inf_x \sup_{\lambda \ge 0} L(x, \lambda) \]

\[ = f_0(x^*). \]

i.e.

\[ \sup_{\lambda \ge 0} \inf_x L(x, \lambda) = \inf_x \sup_{\lambda \ge 0} L(x, \lambda) \]

This is called strong duality.

Key to result is the existence of \( \bar{x} \) such that \( f_i(\bar{x}) < 0 \). (Slater’s condition)
Strong duality

Strong duality says problem (3) is equivalent to the dual problem:

\[
\sup_{\lambda \geq 0} \inf_x L(x, \lambda)
\]

Introduce \( g(\lambda) = \inf_x L(x, \lambda) \), which is called the dual function.

Since \( g \) is the infimum of a family of linear functions (infimum over \( x \), linear in \( \lambda \)), \( g(\lambda) \) is concave, regardless of the structure of \( f_0 \).

Sometimes, the dual problem is easier to solve than original problem. It takes care of constraints \( f_i(x) < 0, i > 0 \), implicitly, though introduce constraints \( \lambda \geq 0 \).

We’ll come back to duality at a later lecture.