Math 273a: Optimization
Subgradient Methods

Instructor: Wotao Yin
Department of Mathematics, UCLA
Fall 2015

online discussions on piazza.com
Recall: For $\bar{x} \subseteq \mathbb{R}^n$,

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

- If $f(x)$ is differentiable, then $\partial f(x) = \{\nabla f(x)\}$ is a singleton.
- For any $x, y, g_x \in \partial f(x)$, and $g_y \in \partial f(y)$, we have

$$\langle g_x - g_y, x - y \rangle \geq 0. \quad (\partial f \text{ is a monotonic point-to-set map})$$
Which functions have subgradients?

**Theorem (Nesterov 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.

---

\[ (g, -1) \]

---

\[ 1 \]

---

\[ \text{epi } f \]

---

\[ \text{dom}(f) \]

---

\[ \text{int}(\text{dom}(f)) \]

---

\[ \partial f(x_0) \]

---

\[ \text{This and next Slide taken from Damek's lecture. Figure taken from Boyd and Vandenberghe, http://see.stanford.edu/materials/lsocoee364b/01-subgradients_notes.pdf.} \]
The converse

**Lemma (Nesterov Lm 3.1.6)**

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

**Proof.**

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

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

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

\[
\alpha f(y) + (1 - \alpha) f(x) \geq f(y_\alpha).
\]
Compute subgradients: general rules

- **Differentiable functions:** $\partial f(x) = \{\nabla f(x)\}$.

- **Composition with affine map:** $\phi(x) = f(A(x) + b)$
  \[
  \partial \phi(x) = A^T \partial f(A(x) + b).
  \]

- **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:** $f(x) = \max_{i \in \{1, \ldots, n\}} \{f_i(x)\}$
  \[
  \partial f(x) = \text{conv}\{\partial f_i(x) \mid f_i(x) = f(x)\}
  \]
Examples

- \( f(x) = |x| \).

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

- $f(\mathbf{x}) = \sum_{i=1}^{n}|\langle \mathbf{a}_i, \mathbf{x} \rangle - b_i|$. Define
  
  $I_-(\mathbf{x}) = \{i|\langle \mathbf{a}_i, \mathbf{x} \rangle - b_i < 0\}$
  
  $I_+(\mathbf{x}) = \{i|\langle \mathbf{a}_i, \mathbf{x} \rangle - b_i > 0\}$
  
  $I_0(\mathbf{x}) = \{i|\langle \mathbf{a}_i, \mathbf{x} \rangle - b_i = 0\}$.

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

- $f(\mathbf{x}) = \max_{i \in \{1, \ldots, n\}} x_i$. Then
  
  $\partial f(\mathbf{x}) = \text{conv}\{\mathbf{e}_i | x_i = f(\mathbf{x})\}$
  
  $\partial f(0) = \text{conv}\{\mathbf{e}_i | i \in \{1, \cdots, n\}\}$
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!
Examples

- \( f(x) = \|x\|_\infty = \max_{i \in \{1, \ldots, n\}} |x^{(i)}| \).

\[
\partial f(x) = \text{conv}\{g^{(i)} | g^{(i)} \in \partial |x^{(i)}|, |x^{(i)}| = f(x)\}, \quad x \neq 0.
\]

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: \( B_1(0, 1) \).
Examples

- \( f(x) = \|x\|_1 = \sum_{i=1}^{n} |x_i| \). 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

\[
\partial f(0) = \sum_{i=1}^{n} [-e_i, e_i] = B_\infty(0, 1).
\]
Optimality condition: 0 subgradient $\iff$ 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: $f(y) \geq f(x^*) + 0 = f(x^*) + \langle 0, y - x^* \rangle$. 
The subgradient method

Iteration:

\[ x^{k+1} \leftarrow x^k - \alpha_k g^k \]

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

Questions:

- Applications?
- Are \( f(x^k) \) and \( \|x^k - x^*\| \) monotonic?
- How to choose \( \alpha_k \)?
Applications

- Finding a point in the intersection of closed convex sets

\[
\text{minimize } f(x) = \max\{\text{dist}(x, C_1), \ldots, \text{dist}(x, C_1)\}
\]

Subgradient: if \( f(x) = \text{dist}(x, C_j) \) and \( x \notin C_j \), then

\[
g = \frac{x - \text{Proj}_{C_j}(x)}{\|x - \text{Proj}_{C_j}(x)\|} \in \partial_x \text{dist}(x, C_j).
\]

- Minimizing non-smooth convex functions, e.g., piece-wise linear convex functions

- Dual subgradient method (generalizes the Uzawa algorithm), dual decomposition

(more to come in this lecture)
Convergence overview

- Typically, convergence is established by identifying a monotonically nonincreasing sequence, such as $f(x^k) - f^*$ and $\|x^k - x^*\|^2$

- However, since the subgradient $g(x)$ is not continuous in $x$, neither sequence is monotonic

- Instead, we will define the total descent and use it to bound $f(x^k)$

- The choice of step sizes $\alpha_k$ is critical.
Monotonicity of $f(x^k)$?

- The definition

$$f(y) \geq f(x) + \langle g, y - x \rangle, \quad \forall g \in \partial f(x)$$

yields

$$f(x^{k+1}) \leq f(x^k) - \langle g^{k+1}, x^k - x^{k+1} \rangle = f(x^k) - \alpha_k \langle g^{k+1}, g^k \rangle.$$  

- It is generally difficult to estimate $\langle g^{k+1}, g^k \rangle$ since $g$ is not continuous. (No matter how close $x^{k+1}$ is to $x^k$, their subgradients $g^{k+1}$ and $g^k$ can be very different.) Therefore, we cannot guarantee $f(x^{k+1}) < f(x^k)$.

- **Note:** Taking the implicit iteration $x^{k+1} = x^k - \alpha_k g^{k+1}$ (the proximal method), we would ensure $f(x^{k+1}) \leq f(x^k)$. It is more expensive to compute though.
Monotonicity of $\|x^k - x^*\|^2$

- Let us assume that $x^*$ exists. Then

\[
\|x^{k+1} - x^*\|^2 = \|(x^k - x^*) - \alpha_k g^k\|^2
= \|x^k - x^*\|^2 - 2\alpha_k \langle g^k, x^k - x^* \rangle + \alpha_k^2 \|g^k\|^2.
\]

- To have monotonicity: $\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2$, we need

\[
-2\alpha_k \langle g^k, x^k - x^* \rangle + \alpha_k^2 \|g^k\|^2 \leq 0 \iff \langle g^k, x^k - x^* \rangle \geq \frac{\alpha_k}{2} \|g^k\|^2.
\]

- However, even at $x^k \approx x^*$, $g^k$ may not vanish.

- Therefore, $\|x^k - x^*\|^2$ is generally not monotonic unless
  - $\|g^k\| < G$ (commonly assumed for subgradient method), and
  - $\alpha_k = O(\|x^k - x^*\|)$, which is unrealistic to ensure since $x^*$ is unknown.
However, it is often easy to have an estimate on $f^*$. For example, $f^* \geq 0$ in many applications.

The definition

\[ f(y) \geq f(x) + \langle g, y - x \rangle \quad \forall g \in \partial f(x) \]

yields

\[-\alpha_k \langle g^k, x^k - x^* \rangle \leq -\alpha_k (f(x^k) - f^*) \leq 0.\]

**Interpretation:** the term $-\alpha_k \langle g^k, x^k - x^* \rangle$ guarantees a sufficient descent by at least $-\alpha_k (f(x^k) - f^*)$

However, the ascending term $\alpha_k^2 \|g^k\|^2$ can be as large as $\alpha_k^2 G^2$. 
After substitution, we get the bound
\[
\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - 2\alpha_k (f(x^k) - f^*) + \alpha_k^2 \|g^k\|^2,
\]
Telescopic sum over \(k\) gives
\[
\|x^{k+1} - x^*\|^2 = \|x^0 - x^*\|^2 - 2 \sum_{i=0}^{k} \alpha_i (f(x^i) - f^*) + \sum_{i=0}^{k} \alpha_i^2 \|g^i\|^2.
\]
Therefore,
\[
\|x^{k+1} - x^*\|^2 + 2 \sum_{i=0}^{k} \alpha_i (f(x^i) - f^*) \leq \|x^0 - x^*\|^2 + \sum_{i=0}^{k} \alpha_i^2 \|g^i\|^2
\]
which bounds the total descent \(\sum_{i=0}^{k} \alpha_k (f(x^i) - f^*)\) by the total ascent \(\sum_{i=0}^{k} \alpha_i^2 \|g^i\|^2\). Clearly, \(\alpha_k\) play the key role.
Step size and convergence

- By

\[
\|x^{k+1} - x^*\|^2 + 2 \sum_{i=0}^{k} \alpha_i (f(x^i) - f^*) \leq \|x^0 - x^*\|^2 + \sum_{i=0}^{k} \alpha_i^2 \|g^i\|^2
\]

and letting

\[
f_{\text{best}}^k = \min \{ f(x^i) : i = 0, 1, \ldots, k \}
\]

(thus, \( f_{\text{best}}^k - f^* \leq f(x^i) - f^*, \ i \leq k \))

\[
\|g\| \leq G \quad (\text{equivalent to Lip. } f: |f(x) - f(y)| \leq G \|x - y\| \quad \forall x, y)
\]

we have

\[
f_{\text{best}}^k - f^* \leq \frac{\|x^0 - x^*\|^2 + G^2 \sum_{i=0}^{k} \alpha_i^2}{2 \sum_{i=0}^{k} \alpha_i}.
\]
We need unbounded $\sum_{i=0}^{k} \alpha_i$ and bounded $\sum_{i=0}^{k} \alpha_i$ as $k \to \infty$.

- To have $f_{\text{best}}^k - f^* \to 0$, we require, for example,
  - $\sum_{i=0}^{k} \alpha_i = \infty$ and $\sum_{i=0}^{k} \alpha_i^2 \leq \infty$;
  - or more weakly, $\sum_{i=0}^{k} \alpha_i = \infty$ and $\lim \alpha_k \to 0$. (the truncation trick)

Otherwise, $f_{\text{best}}^k \not\to f^*$ in general.
Fixed step size

- **Fixing** $\alpha_k \equiv \alpha$, we get

  $$f_{\text{best}}^k - f^* \leq \frac{||x^0 - x^*||^2 + G^2 \sum_{i=0}^{k} \alpha^2}{2 \sum_{i=0}^{k} \alpha} = \frac{||x^0 - x^*||^2}{2\alpha(k+1)} + \frac{\alpha G^2}{2}.$$ 

- $f_{\text{best}}^k - f^* \to \alpha G^2 / 2 = O(\alpha)$.

- **while in the early stage**

  $$k < \frac{||x^0 - x^*||^2}{\alpha^2 G^2},$$

  we have

  $$\frac{||x^0 - x^*||^2}{2\alpha(k+1)} + \frac{\alpha G^2}{2} \leq \frac{||x^0 - x^*||^2}{\alpha(k+1)}$$

  and thus the (non-asymptotic, conditional) rate of convergence $O\left(\frac{1}{\alpha k}\right)$.

- larger $\alpha \implies$ faster convergence, lower final accuracy

- smaller $\alpha \implies$ slower convergence, higher final accuracy
Fixed step “length” \( \alpha_k = \alpha / \| g^k \| \)

- We have
  \[
  f^k_{\text{best}} - f^* \leq \frac{\| x^0 - x^* \|^2 + \sum_{i=0}^{k} \alpha_k^2 \| g^k \|^2}{2 \sum_{i=0}^{k} \alpha_k} = \frac{\| x^0 - x^* \|^2 G}{2\alpha(k + 1)} + \frac{\alpha G}{2}.
  \]

- \( f^k_{\text{best}} - f^* \to \alpha G/2 = O(\alpha) \), slightly better than with a fixed step size.

- while
  \[
  k < \frac{\| x^0 - x^* \|^2}{\alpha^2},
  \]
  we have the (non-asymptotic, conditional) rate of convergence \( O\left( \frac{G}{\alpha_k} \right) \).

- larger \( \alpha \) \( \implies \) faster convergence, lower final accuracy

- smaller \( \alpha \) \( \implies \) slower convergence, higher final accuracy
Polyak step size

- Assume that $f^*$ is known (not $x^*$ though). Example: $f^* = 0$ in projection problems.) Set:

$$
\alpha_k = \frac{f(x^k) - f^*}{\|g^k\|} = \arg \min \|x^k - x^*\|^2 - 2\alpha_k(f(x^k) - f^*) + \alpha_k^2\|g^k\|^2
$$

- Then,

$$
\|x^k - x^*\|^2 - 2\alpha_k(f(x^k) - f^*) + \alpha_k^2\|g^k\|^2 = \|x^k - x^*\|^2 - \frac{(f(x^k) - f^*)^2}{\|g^k\|^2}
$$

(so $\|x^k - x^*\|^2$ is monotonic) and thus after adding over $k$,

$$
\|x^{k+1} - x^*\|^2 \leq \|x^0 - x^*\|^2 - \frac{1}{G} \sum_{i=0}^{k} (f(x^i) - f^*)^2.
$$

- Finally,

$$
f_{\text{best}}^k - f^* \leq \frac{\|x^0 - x^*\|\sqrt{G}}{\sqrt{k + 1}} = O\left(\frac{1}{\sqrt{k}}\right).
$$
Oracle optimality

For an $O\left(\frac{1}{\sqrt{k}}\right)$ method to guarantee $f_{\text{best}}^k - f^* \leq \epsilon$, we need $O\left(\frac{1}{\epsilon^2}\right)$ iterations.

Is this tight for the subgradient method? Answer: Yes.

Suppose $x^{k+1}$ is computed by an arbitrary method as

$$x^{k+1} \in x^0 + \text{span}\{g^0, g^1, \ldots, g^k\}$$

where the oracle gives

- arbitrary $g^k \in \partial f(x^k)$ and
- $f(x^k)$.

Theorem (Nesterov Thm 3.2.1)

There is a nonsmooth convex function with $\|g\| \leq G$ uniformly so that the above method obeys

$$f(x^k) - f(x^*) \geq \frac{\|x^0 - x^*\|G}{2(1 + \sqrt{k + 1})}. $$
The subgradient algorithm

- $f$ is a proper closed convex function.

- **Problem:** minimize $\min_x f(x)$

- **Algorithm:** pick any starting point $x^1$
  - pick $g^k \in \partial f(x^k)$
  - set $\alpha_k$ ($\alpha_0/k$, fixed size, fixed length, or Polyak)
  - $x^{k+1} \leftarrow x^k - \alpha_k g^k$
  - $k \leftarrow k + 1$

(monitor $f(x^k)$ and $f_{\text{best}}^k$, especially if using fixed size, fixed length)
Variant: projected subgradient method

- $f$ is a proper closed convex function.
- $C$ is a nonempty closed convex set.
- Problem:
  \[
  \min_{x} f(x), \quad \text{subject to } x \in C.
  \]
- pick $g^k \in \partial f(x^k)$ and $\alpha_k$, and apply
  \[
  x^{k+1} \leftarrow \text{Proj}_C(x^k - \alpha_k g^k)
  \]
since projection is \textit{nonexpansive},

\[ \| \text{Proj}_C(x) - \text{Proj}_C(y) \| \leq \| x - y \|, \quad \forall x, y \in \mathbb{R}^n \]

the analysis remains the same.

\[
\| x^{k+1} - x^* \| ^2 = \| \text{Proj}_C(x^k - \alpha_k g^k) - \text{Proj}_C(x^*) \| ^2 \\
\leq \| (x^k - \alpha_k g^k) - x^* \| ^2 \\
= \| (x^k - x^*) - \alpha_k g^k \| ^2 \\
= \| x^k - x^* \| ^2 - 2\alpha_k \langle g^k, x^k - x^* \rangle + \alpha_k^2 \| g^k \| ^2 \\
= \ldots .
\]
Summary for subgradient methods

- **Universal.** It handles non-differentiable convex problems and, in particular, the dual problem of linearly constrained convex problems (later lectures)

- **No monotonicity** for either $f(x^k)$ or $\text{dist}(x^k, X^*)$ except for Polyak’s step size (requiring known $f^*$)

- Convergence relies on **uniformly bounded subgradient** (or Lipschitz $f$)

- Rate of convergence $f_{\text{best}}^k - f^*$ depends on the step size
  - Constant step size (or length) does not guarantee $f_{\text{best}}^k \rightarrow f^*$
  - If we need $f_{\text{best}}^k \rightarrow f^*$, use diminishing step sizes; the rate is at best $O(1/\sqrt{k})$

- Convergence is quite slow (but the method is widely applicable)

- Some non-smooth problems have better rates by other methods, e.g., prox-linear iteration, operator splitting, dual smoothing (later lectures)