Subgradients

Ryan Tibshirani
Convex Optimization 10-725/36-725
Consider the problem

$$\min f(x)$$

for $f$ convex and differentiable, $\text{dom}(f) = \mathbb{R}^n$. Gradient descent:
choose initial $x^{(0)} \in \mathbb{R}^n$, repeat

$$x^{(k)} = x^{(k-1)} - t_k \cdot \nabla f(x^{(k-1)}), \quad k = 1, 2, 3, \ldots$$

Step sizes $t_k$ chosen to be fixed and small, or by backtracking line search

If $\nabla f$ Lipschitz, gradient descent has convergence rate $O(1/\epsilon)$

Downsides:

- Requires $f$ differentiable ← next lecture
- Can be slow to converge ← two lectures from now
Outline

Today: crucial mathematical underpinnings!

• Subgradients
• Examples
• Subgradient rules
• Optimality characterizations
Subgradients

Remember that for convex and differentiable $f$, 

$$f(y) \geq f(x) + \nabla f(x)^T(y - x) \quad \text{for all } x, y$$

I.e., linear approximation always underestimates $f$

A subgradient of a convex function $f$ at $x$ is any $g \in \mathbb{R}^n$ such that

$$f(y) \geq f(x) + g^T(y - x) \quad \text{for all } y$$

- Always exists
- If $f$ differentiable at $x$, then $g = \nabla f(x)$ uniquely
- Actually, same definition works for nonconvex $f$ (however, subgradients need not exist)
Examples of subgradients

Consider $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(x) = |x|

- For $x \neq 0$, unique subgradient $g = \text{sign}(x)$
- For $x = 0$, subgradient $g$ is any element of $[-1, 1]$
Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_2$

- For $x \neq 0$, unique subgradient $g = x/\|x\|_2$
- For $x = 0$, subgradient $g$ is any element of $\{z : \|z\|_2 \leq 1\}$
Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_1$

- For $x_i \neq 0$, unique $i$th component $g_i = \text{sign}(x_i)$
- For $x_i = 0$, $i$th component $g_i$ is any element of $[-1, 1]$
Let $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable, and consider $f(x) = \max\{f_1(x), f_2(x)\}$

- For $f_1(x) > f_2(x)$, unique subgradient $g = \nabla f_1(x)$
- For $f_2(x) > f_1(x)$, unique subgradient $g = \nabla f_2(x)$
- For $f_1(x) = f_2(x)$, subgradient $g$ is any point on the line segment between $\nabla f_1(x)$ and $\nabla f_2(x)$
Subdifferential

Set of all subgradients of convex $f$ is called the subdifferential:

$$\partial f(x) = \{ g \in \mathbb{R}^n : g \text{ is a subgradient of } f \text{ at } x \}$$

- $\partial f(x)$ is closed and convex (even for nonconvex $f$)
- Nonempty (can be empty for nonconvex $f$)
- If $f$ is differentiable at $x$, then $\partial f(x) = \{ \nabla f(x) \}$
- If $\partial f(x) = \{ g \}$, then $f$ is differentiable at $x$ and $\nabla f(x) = g$
Connection to convex geometry

Convex set $C \subseteq \mathbb{R}^n$, consider indicator function $I_C : \mathbb{R}^n \to \mathbb{R}$,

$$I_C(x) = I\{x \in C\} = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \notin C \end{cases}$$

For $x \in C$, $\partial I_C(x) = \mathcal{N}_C(x)$, the normal cone of $C$ at $x$, recall

$$\mathcal{N}_C(x) = \{g \in \mathbb{R}^n : g^T x \geq g^T y \text{ for any } y \in C\}$$

Why? By definition of subgradient $g$,

$$I_C(y) \geq I_C(x) + g^T (y - x) \text{ for all } y$$

- For $y \notin C$, $I_C(y) = \infty$
- For $y \in C$, this means $0 \geq g^T (y - x)$
Subgradient calculus

Basic rules for convex functions:

- **Scaling**: $\partial(af) = a \cdot \partial f$ provided $a > 0$
- **Addition**: $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$
- **Affine composition**: if $g(x) = f(Ax + b)$, then
  \[
  \partial g(x) = A^T \partial f(Ax + b)
  \]
- **Finite pointwise maximum**: if $f(x) = \max_{i=1,...,m} f_i(x)$, then
  \[
  \partial f(x) = \text{conv}\left(\bigcup_{i : f_i(x) = f(x)} \partial f_i(x)\right)
  \]
  the convex hull of union of subdifferentials of all active functions at $x$
• **General pointwise maximum**: if \( f(x) = \max_{s \in S} f_s(x) \), then

\[
\partial f(x) \supseteq \operatorname{cl}\left\{ \operatorname{conv}\left( \bigcup_{s : f_s(x) = f(x)} \partial f_s(x) \right) \right\}
\]

and under some regularity conditions (on \( S, f_s \)), we get =

• **Norms**: important special case, \( f(x) = \|x\|_p \). Let \( q \) be such that \( 1/p + 1/q = 1 \), then

\[
\|x\|_p = \max_{\|z\|_q \leq 1} z^T x
\]

Hence

\[
\partial f(x) = \operatorname{argmax}_{\|z\|_q \leq 1} z^T x
\]
Why subgradients?

Subgradients are important for two reasons:

- **Convex analysis**: optimality characterization via subgradients, monotonicity, relationship to duality
- **Convex optimization**: if you can compute subgradients, then you can minimize (almost) any convex function
Optimality condition

For any $f$ (convex or not),

$$f(x^*) = \min f(x) \iff 0 \in \partial f(x^*)$$

i.e., $x^*$ is a minimizer if and only if 0 is a subgradient of $f$ at $x^*$. This is called the subgradient optimality condition.

Why? Easy: $g = 0$ being a subgradient means that for all $y$

$$f(y) \geq f(x^*) + 0^T(y - x^*) = f(x^*)$$

Note the implication for a convex and differentiable function $f$, with $\partial f(x) = \{\nabla f(x)\}$
Derivation of first-order optimality

Example of the power of subgradients: we can use what we have learned so far to derive the first-order optimality condition. Recall that for $f$ convex and differentiable, the problem

$$\min f(x) \text{ subject to } x \in C$$

is solved at $x$ if and only if

$$\nabla f(x)^T(y - x) \geq 0 \quad \text{for all } y \in C$$

Intuitively says that gradient increases as we move away from $x$

How to see this? First recast problem as

$$\min f(x) + I_C(x)$$

Now apply subgradient optimality: $0 \in \partial(f(x) + I_C(x))$
But

\[ 0 \in \partial (f(x) + I_C(x)) \]

\[ \iff 0 \in \{\nabla f(x)\} + N_C(x) \]

\[ \iff -\nabla f(x) \in N_C(x) \]

\[ \iff -\nabla f(x)^T x \geq -\nabla f(x)^T y \text{ for all } y \in C \]

\[ \iff \nabla f(x)^T (y - x) \geq 0 \text{ for all } y \in C \]

as desired

Note: the condition \( 0 \in \partial f(x) + N_C(x) \) is a fully general condition for optimality in a convex problem. But this is not always easy to work with (KKT conditions, later, are easier)
Example: lasso optimality conditions

Given \( y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p} \), lasso problem can be parametrized as:

\[
\min_{\beta \in \mathbb{R}^p} \frac{1}{2} \| y - X \beta \|_2^2 + \lambda \| \beta \|_1
\]

where \( \lambda \geq 0 \). Subgradient optimality:

\[
0 \in \partial \left( \frac{1}{2} \| y - X \beta \|_2^2 + \lambda \| \beta \|_1 \right)
\]

\( \iff \) \( 0 \in -X^T (y - X \beta) + \lambda \partial \| \beta \|_1 \)

\( \iff \) \( X^T (y - X \beta) = \lambda v \)

for some \( v \in \partial \| \beta \|_1 \), i.e.,

\[
v_i \in \begin{cases} 
1 & \text{if } \beta_i > 0 \\
-1 & \text{if } \beta_i < 0 \ , \ i = 1, \ldots, p \\
[-1, 1] & \text{if } \beta_i = 0
\end{cases}
\]
Write $X_1, \ldots, X_p$ for columns of $X$. Then subgradient optimality reads:

\[
\begin{align*}
X_i^T(y - X\beta) &= \lambda \cdot \text{sign}(\beta_i) \quad \text{if } \beta_i \neq 0 \\
|X_i^T(y - X\beta)| &\leq \lambda \quad \text{if } \beta_i = 0
\end{align*}
\]

Note: the subgradient optimality conditions do not directly lead to an expression for a lasso solution ... however they do provide a way to check lasso optimality

They are also helpful in understanding the lasso estimator; e.g., if $|X_i^T(y - X\beta)| < \lambda$, then $\beta_i = 0$
Example: soft-thresholding

Simplified lasso problem with $X = I$:

$$\min_{\beta \in \mathbb{R}^n} \frac{1}{2} \| y - \beta \|_2^2 + \lambda \| \beta \|_1$$

This we can solve directly using subgradient optimality. Solution is $\beta = S_{\lambda}(y)$, where $S_{\lambda}$ is the soft-thresholding operator:

$$[S_{\lambda}(y)]_i = \begin{cases} 
  y_i - \lambda & \text{if } y_i > \lambda \\
  0 & \text{if } -\lambda \leq y_i \leq \lambda, \ i = 1, \ldots, n \\
  y_i + \lambda & \text{if } y_i < -\lambda 
\end{cases}$$

Check: from last slide, subgradient optimality conditions are

$$\begin{cases} 
  y_i - \beta_i = \lambda \cdot \text{sign}(\beta_i) & \text{if } \beta_i \neq 0 \\
  |y_i - \beta_i| \leq \lambda & \text{if } \beta_i = 0 
\end{cases}$$
Now plug in $\beta = S_\lambda(y)$ and check these are satisfied:

- When $y_i > \lambda$, $\beta_i = y_i - \lambda > 0$, so $y_i - \beta_i = \lambda = \lambda \cdot 1$
- When $y_i < -\lambda$, argument is similar
- When $|y_i| \leq \lambda$, $\beta_i = 0$, and $|y_i - \beta_i| = |y_i| \leq \lambda$

Soft-thresholding in one variable:
Example: distance to a convex set

Recall the distance function to a convex set \( C \):

\[
\text{dist}(x, C) = \min_{y \in C} \|y - x\|_2
\]

This is a convex function. What are its subgradients?

Write \( \text{dist}(x, C) = \|x - P_C(x)\|_2 \), where \( P_C(x) \) is the projection of \( x \) onto \( C \). Then when \( \text{dist}(x, C) > 0 \),

\[
\partial \text{dist}(x, C) = \left\{ \frac{x - P_C(x)}{\|x - P_C(x)\|_2} \right\}
\]

Only has one element, so in fact \( \text{dist}(x, C) \) is differentiable and this is its gradient.
We will only show one direction, i.e., that

\[
\frac{x - P_C(x)}{\|x - P_C(x)\|_2} \in \partial \text{dist}(x, C)
\]

Write \( u = P_C(x) \). Then by first-order optimality conditions for a projection,

\[
(u - x)^T(y - u) \geq 0 \quad \text{for all } y \in C
\]

Hence

\[
C \subseteq H = \{y : (u - x)^T(y - u) \geq 0\}
\]

Claim: for any \( y \),

\[
\text{dist}(y, C) \geq \frac{(x - u)^T(y - u)}{\|x - u\|_2}
\]

Check: first, for \( y \in H \), the right-hand side is \( \leq 0 \)
Now for \( y \notin H \), we have \((x - u)^T(y - u) = \|x - u\|_2 \|y - u\|_2 \cos \theta\)
where \( \theta \) is the angle between \( x - u \) and \( y - u \). Thus

\[
\frac{(x - u)^T(y - u)}{\|x - u\|_2} = \|y - u\|_2 \cos \theta = \text{dist}(y, H) \leq \text{dist}(y, C)
\]

as desired.

Using the claim, we have for any \( y \)

\[
\text{dist}(y, C) \geq \frac{(x - u)^T(y - x + x - u)}{\|x - u\|_2}
\]

\[
= \|x - u\|_2 + \left( \frac{x - u}{\|x - u\|_2} \right)^T (y - x)
\]

Hence \( g = (x - u)/\|x - u\|_2 \) is a subgradient of \( \text{dist}(x, C) \) at \( x \).
References and further reading

- S. Boyd, Lecture notes for EE 264B, Stanford University, Spring 2010-2011
- L. Vandenberghe, Lecture notes for EE 236C, UCLA, Spring 2011-2012