# Subgradients

Ryan Tibshirani
Convex Optimization 10-725

# Last time: gradient descent

Consider the problem

$$\min_x \; f(x)$$

for $f$ convex and differentiable, $\mathrm{dom}(f) = \mathbb{R}^n$. 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 $\nabla f$ is Lipschitz, gradient descent has convergence rate $O(1/\epsilon)$. Downsides:

- Requires $f$ differentiable
- Can be slow to converge

# Outline

Today: crucial mathematical underpinnings!

- Subgradients
- Examples
- Properties
- Optimality characterizations

# Subgradients

Recall that for convex and differentiable $f$,

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

That is, linear approximation always underestimates $f$

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

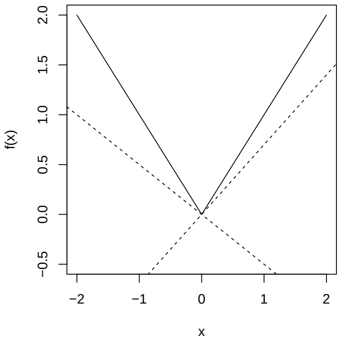
$$f(y) \geq f(x) + g^T (y - x) \quad \text{for all } y$$

- Always exists[1]
- If $f$ differentiable at $x$, then $g = \nabla f(x)$ uniquely
- Same definition works for nonconvex $f$ (however, subgradients need not exist)

[1]On the relative interior of $\mathrm{dom}(f)$

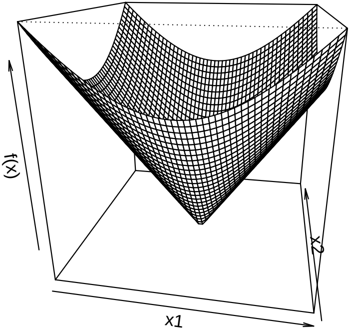
# Examples of subgradients

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



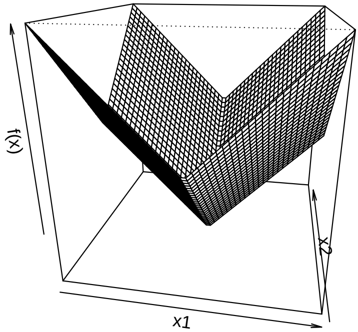

- For $x \neq 0$, unique subgradient $g = \mathrm{sign}(x)$
- For $x = 0$, subgradient $g$ is any element of $[-1, 1]$

Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_2$



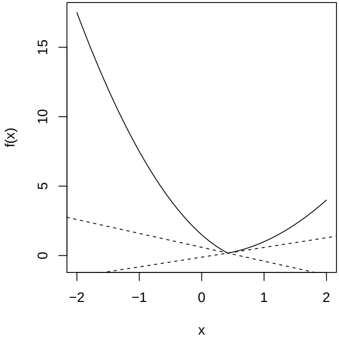

- For $x \neq 0$, unique subgradient $g = x/\|x\|_2$
- For $x = 0$, subgradient $g$ is any element of $\{z : \|z\|_2 \leq 1\}$

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 $[-1, 1]$

Consider $f(x) = \max\{f_1(x), f_2(x)\}$, for $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ convex, differentiable




- 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 line segment between $\nabla f_1(x)$ and $\nabla f_2(x)$

# Subdifferential

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

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

- Nonempty (only for convex $f$)
- $\partial f(x)$ is closed and convex (even 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 $I_C : \mathbb{R}^n \to \mathbb{R}$,

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

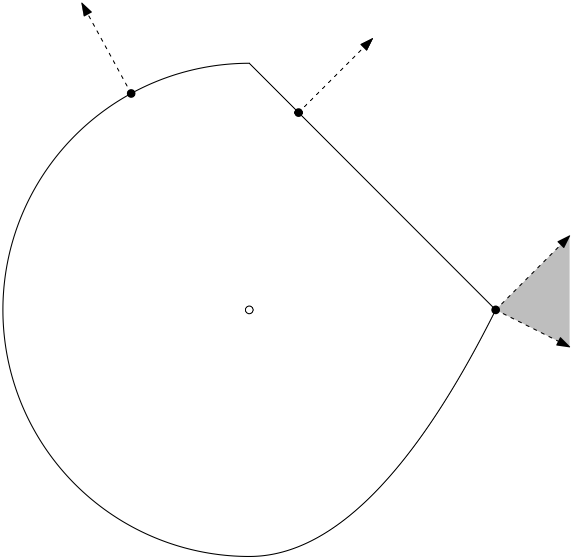
For $x \in C$, $\partial I_C(x) = \mathcal{N}_C(x)$, the normal cone of $C$ at $x$ is, recall

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

Why? By definition of subgradient $g$,

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

- 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 $f(x) = \max_{i=1,\ldots,m} f_i(x)$, then

$$\partial f(x) = \operatorname{conv}\left( \bigcup_{i: f_i(x) = f(x)} \partial f_i(x) \right)$$

convex hull of union of subdifferentials of active functions at $x$

- General composition: if

$$f(x) = h\big(g(x)\big) = h\big(g_1(x), \ldots, g_k(x)\big)$$

where $g : \mathbb{R}^n \to \mathbb{R}^k$, $h : \mathbb{R}^k \to \mathbb{R}$, $f : \mathbb{R}^n \to \mathbb{R}$, $h$ is convex and nondecreasing in each argument, $g$ is convex, then

$$\partial f(x) \subseteq \Big\{ p_1 q_1 + \cdots + p_k q_k : \\ p \in \partial h(g(x)),\ q_i \in \partial g_i(x),\ i = 1, \ldots, k \Big\}$$

- General pointwise maximum: if $f(x) = \max_{s \in S} f_s(x)$, then

$$\partial f(x) \supseteq \mathrm{cl}\Bigg\{ \mathrm{conv}\Bigg( \bigcup_{s : f_s(x) = f(x)} \partial f_s(x) \Bigg) \Bigg\}$$

Under some regularity conditions (on $S$, $f_s$), we get equality

- Norms: important special case. To each norm $\|\cdot\|$, there is a dual norm $\|\cdot\|_*$ such that

$$\|x\| = \max_{\|z\|_* \leq 1} z^T x$$

(For example, $\|\cdot\|_p$ and $\|\cdot\|_q$ are dual when $1/p + 1/q = 1$.) In fact, for $f(x) = \|x\|$ (and $f_z(x) = z^T x$), we get equality:

$$\partial f(x) = \mathrm{cl}\left\{\mathrm{conv}\left(\bigcup_{z: f_z(x) = f(x)} \partial f_z(x)\right)\right\}$$

Note that $\partial f_z(x) = z$. And if $z_1, z_2$ each achieve the max at $x$, which means that $z_1^T x = z_2^T x = \|x\|$, then by linearity, so will $t z_1 + (1-t) z_2$ for any $t \in [0, 1]$. Thus

$$\partial f(x) = \underset{\|z\|_* \leq 1}{\mathrm{argmax}}\; z^T x$$

# Optimality condition

For any $f$ (convex or not),

$$f(x^\star) = \min_x \; f(x) \quad \Longleftrightarrow \quad 0 \in \partial f(x^\star)$$

That is, $x^\star$ is a minimizer if and only if $0$ is a subgradient of $f$ at $x^\star$. This is called the subgradient optimality condition

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

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

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

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

is solved at $x$, for $f$ convex and differentiable, if and only if

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

Intuitively: says that gradient increases as we move away from $x$. How to prove it? First recast problem as

$$\min_x \; f(x) + I_C(x)$$

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

Observe

$$
\begin{aligned}
0 \in \partial\big(f(x) + I_C(x)\big) \\
&\iff 0 \in \{\nabla f(x)\} + \mathcal{N}_C(x) \\
&\iff -\nabla f(x) \in \mathcal{N}_C(x) \\
&\iff -\nabla f(x)^T x \geq -\nabla f(x)^T y \text{ for all } y \in C \\
&\iff \nabla f(x)^T (y - x) \geq 0 \text{ for all } y \in C
\end{aligned}
$$

as desired

Note: the condition $0 \in \partial f(x) + \mathcal{N}_C(x)$ is a fully general condition for optimality in convex problems. But it's 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

$$\min_\beta \ \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

where $\lambda \geq 0$. Subgradient optimality:

$$\begin{aligned} 0 \in \partial\Big(\frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\Big) & \\ \iff & \ 0 \in -X^T(y - X\beta) + \lambda\partial\|\beta\|_1 \\ \iff & \ X^T(y - X\beta) = \lambda v \end{aligned}$$

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

$$v_i \in \begin{cases} \{1\} & \text{if } \beta_i > 0 \\ \{-1\} & \text{if } \beta_i < 0 \\ [-1, 1] & \text{if } \beta_i = 0 \end{cases}, \quad i = 1, \ldots, p$$

Write $X_1, \ldots, X_p$ for columns of $X$. Then our condition reads:

$$\begin{cases} X_i^T(y - X\beta) = \lambda \cdot \mathrm{sign}(\beta_i) & \text{if } \beta_i \neq 0 \\ |X_i^T(y - X\beta)| \leq \lambda & \text{if } \beta_i = 0 \end{cases}$$

Note: subgradient optimality conditions don't lead to closed-form 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 $\beta_i = 0$ (used by screening rules, later?)

# Example: soft-thresholding

Simplfied lasso problem with $X = I$:

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

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

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

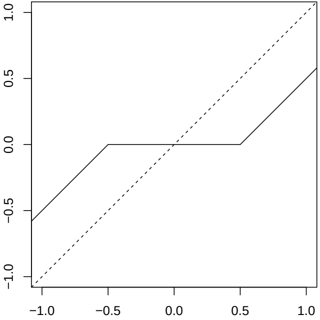
Check: from last slide, subgradient optimality conditions are

$$\begin{cases} y_i - \beta_i = \lambda \cdot \mathrm{sign}(\beta_i) & \text{if } \beta_i \neq 0 \\ |y_i - \beta_i| \le \lambda & \text{if } \beta_i = 0 \end{cases}$$

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 $|y_i| \le \lambda$, $\beta_i = 0$, and $|y_i - \beta_i| = |y_i| \le \lambda$

Soft-thresholding in one variable:



# Example: distance to a convex set

Recall the distance function to a closed, convex set $C$:

$$\mathrm{dist}(x, C) = \min_{y \in C} \ \|y - x\|_2$$

This is a convex function. What are its subgradients?

Write $\mathrm{dist}(x, C) = \|x - P_C(x)\|_2$, where $P_C(x)$ is the projection of $x$ onto $C$. It turns out that when $\mathrm{dist}(x, C) > 0$,

$$\partial \mathrm{dist}(x, C) = \left\{ \frac{x - P_C(x)}{\|x - P_C(x)\|_2} \right\}$$

Only has one element, so in fact $\mathrm{dist}(x, C)$ is differentiable and this is its gradient

We will only show one direction, i.e., that

$$\frac{x - P_C(x)}{\|x - P_C(x)\|_2} \in \partial \mathrm{dist}(x, C)$$

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

$$(x-u)^T(y-u) \le 0 \quad \text{for all } y \in C$$

Hence

$$C \subseteq H = \{y : (x-u)^T(y-u) \le 0\}$$

Claim:

$$\mathrm{dist}(y, C) \ge \frac{(x-u)^T(y-u)}{\|x-u\|_2} \quad \text{for all } y$$

Check: first, for $y \in H$, the right-hand side is $\le 0$

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

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

as desired

Using the claim, we have for any $y$

$$\begin{aligned}\mathrm{dist}(y,C) &\geq \frac{(x-u)^T(y-x+x-u)}{\|x-u\|_2} \\ &= \|x-u\|_2 + \left(\frac{x-u}{\|x-u\|_2}\right)^T(y-x)\end{aligned}$$

Hence $g = (x-u)/\|x-u\|_2$ is a subgradient of $\mathrm{dist}(x,C)$ at $x$

# References and further reading


- S. Boyd, Lecture notes for EE 264B, Stanford University, Spring 2010-2011
- R. T. Rockafellar (1970), "Convex analysis", Chapters 23–25
- L. Vandenberghe, Lecture notes for EE 236C, UCLA, Spring 2011-2012