Subgradients

Barnabas Poczos & Ryan Tibshirani
Convex Optimization 10-725/36-725

Recall gradient descent

We want to solve

$$\min_{x \in \mathbb{R}^n} \; f(x),$$

for f convex and differentiable

Gradient descent: choose initial $x^{(0)} \in \mathbb{R}^n$, repeat

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

If ∇f Lipschitz, gradient descent has convergence rate $O(1/k)$

Downsides:

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

Outline

Today:

- Subgradients
- Examples
- Subgradient rules
- Optimality characterizations

Subgradients

Remember that for convex $f : \mathbb{R}^n \to \mathbb{R}$,

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

I.e., linear approximation always underestimates f

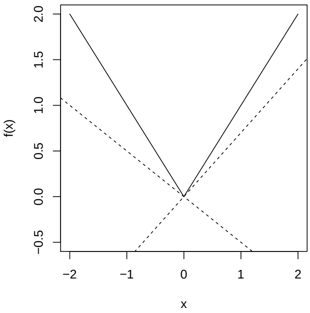
A subgradient of convex $f : \mathbb{R}^n \to \mathbb{R}$ at x is any $g \in \mathbb{R}^n$ such that

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

- 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)

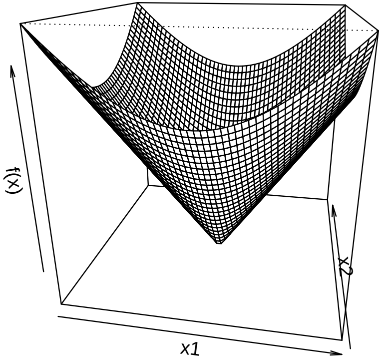
Examples

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



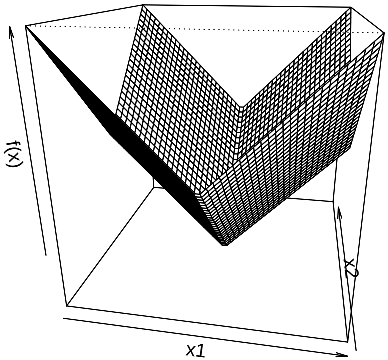
- For $x \neq 0$, unique subgradient $g = \mathrm{sign}(x)$
- For $x = 0$, subgradient g is any element of $[-1, 1]$

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



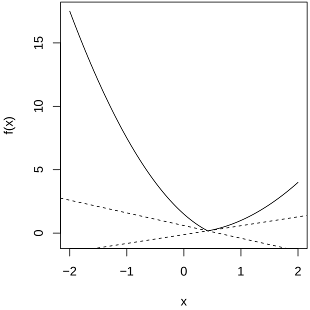
- For $x \neq 0$, unique subgradient $g = x/\|x\|_2$
- For $x = 0$, subgradient g is any element of $\{z : \|z\|_2 \leq 1\}$

Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_1$



- For $x_i \neq 0$, unique ith component $g_i = \text{sign}(x_i)$
- For $x_i = 0$, ith component g_i is an element of $[-1, 1]$

Let $f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$ be convex, 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 $I_C : \mathbb{R}^n \to \mathbb{R}$,

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

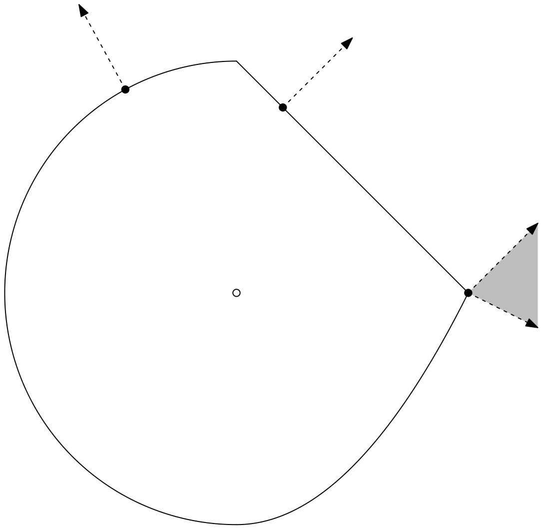
For $x \in C$, $\partial I_C(x) = \mathcal{N}_C(x)$, the normal cone of C at x,

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

Why? Recall 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

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

  the convex hull of union of subdifferentials of all active functions at x

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

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

and under some regularity conditions (on $\mathcal{S}$, f_s), we get $=$

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

$$\|x\|_p = \max_{\|z\|_q \le 1} z^T x$$

Hence

$$\partial f(x) = \Big\{y : \|y\|_q \le 1 \text{ and } y^T x = \max_{\|z\|_q \le 1} z^T x\Big\}$$

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),

$$f(x^\star) = \min_{x \in \mathbb{R}^n} f(x) \quad \Longleftrightarrow \quad 0 \in \partial f(x^\star)$$

I.e., $x^\star$ is a minimizer if and only if 0 is a subgradient of f at $x^\star$

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

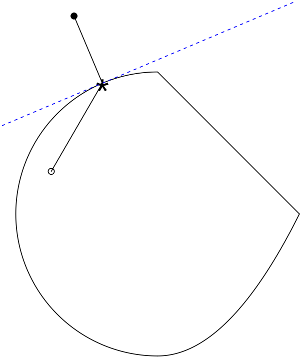
$$f(y) \geq f(x^\star) + 0^T(y - x^\star) = f(x^\star)$$

Note implication for differentiable case, where $\partial f(x) = \{\nabla f(x)\}$

Projection onto a convex set

Given closed, convex set $C \subseteq \mathbb{R}^n$, and a point $y \in \mathbb{R}^n$, we define the projection operator onto C as

$$P_C(x) = \underset{x \in C}{\operatorname{argmin}} \; \|y - x\|_2$$



Optimality characterization: $x^\star = P_C(y)$ if and only if

$$\langle y - x^\star, x^\star - x \rangle \geq 0 \quad \text{for all } x \in C$$

Sometimes called variational inequality

How to see this? Note that $x^\star = P_C(y)$ minimizes the criterion

$$f(x) = \frac{1}{2}\|y - x\|_2^2 + I_C(x)$$

where I_C is the indicator function of C. Hence we know this is equivalent to

$$0 \in \partial f(x^\star) = -(y - x^\star) + \mathcal{N}_C(x^\star)$$

i.e.,

$$y - x^\star \in \mathcal{N}_C(x^\star)$$

which exactly means

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

Soft-thresholding

Lasso problem can be parametrized as

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

where $\lambda \geq 0$. Consider simplified problem with $X = I$:

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

Claim: solution of simple problem is $\hat{\beta} = S_\lambda(y)$, where S_λ is the soft-thresholding operator,

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

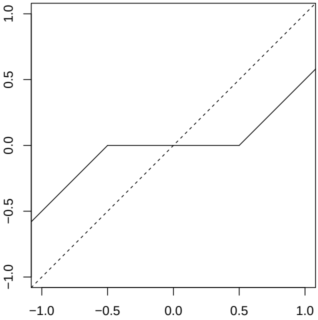
Why? Subgradients of $f(\beta) = \frac{1}{2}\|y - \beta\|_2^2 + \lambda\|\beta\|_1$ are

$$g = \beta - y + \lambda s,$$

where $s_i = \text{sign}(\beta_i)$ if $\beta_i \neq 0$ and $s_i \in [-1, 1]$ if $\beta_i = 0$

Now just plug in $\beta = S_\lambda(y)$ and check that we can get $g = 0$

Soft-thresholding in one variable:



References

- 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