A Standard material on convexity

Definition A.1 A set S in $\mathbb{R}^n$ is said to be *convex* if for every $x_1, x_2 \in S$ the line segment $\{\lambda x_1 + (1-\lambda)x_2 : 0 \leq \lambda \leq 1\}$ belongs to S.

For instance, a hyperplane $S = \{x \in \mathbb{R}^n : p^t x = \alpha\}$ or a ball $S = \{x \in \mathbb{R}^n : |x - x_0| \leq \beta\}$ are examples of convex sets. However, the sphere $S = \{x \in \mathbb{R}^n : |x - x_0| = \beta\}$ provides an example of a set that is not convex ($\beta > 0$). It is easy to see that arbitrary intersections of convex sets are again convex; also finite sums of convex sets are convex again.

Theorem A.2 (strict point-set separation [1, Thm. 2.4.4]) *Let S be a nonempty closed convex subset of $\mathbb{R}^n$ and let $y \in \mathbb{R}^n \backslash S$. Then there exists $p \in \mathbb{R}^n$, $p \neq 0$, such that*

$$\sup_{x \in S} p^t x < p^t y.$$

Proof. It is a standard result that there exists $\hat{x} \in S$ such that $\sup_{s \in S} |y - s| = |y - \hat{x}|$ (consider a suitable closed ball around y and apply the theorem of Weierstrass [1, Thm. 2.3.1]). By convexity of S, this means that for every $x \in S$ and every $\lambda \in (0, 1]$

$$|y - (\lambda x + (1-\lambda)\hat{x})|^2 \geq |y - \hat{x}|^2.$$

Obviously, the expression on the left equals

$$|y - \hat{x} - \lambda(x - \hat{x})|^2 = |y - \hat{x}|^2 - 2\lambda(y - \hat{x})^t(x - \hat{x}) + \lambda^2 |x - \hat{x}|^2,$$

so the above inequality amounts to

$$2\lambda(y - \hat{x})^t(x - \hat{x}) \leq \lambda^2 |x - \hat{x}|^2$$

for every $x \in S$ and every $\lambda \in (0, 1]$. Dividing by $\lambda > 0$ and letting λ go to zero then gives

$$(y - \hat{x}) \cdot (x - \hat{x}) \leq 0 \text{ for all } x \in S.$$

Set $p := y - \hat{x}$; then $p \neq 0$ (note that $p = 0$ would imply $y \in S$). We clearly have $p^t x \leq p^t \hat{x}$. Also, we have now $p^t \hat{x} > p^t y$, for otherwise $(y - \hat{x})^t(\hat{x} - y) \geq 0$ would imply $y = \hat{x} \in S$, which is impossible. QED

For our next result, recall that $\partial S := \text{cl} S \cap \text{cl}(\mathbb{R}^n \backslash S) = \text{cl} S \backslash \text{int S}$ denotes the boundary of a set $S \subset \mathbb{R}^n$.

Theorem A.3 (supporting hyperplane [1, Thm. 2.4.7]) *Let S be a nonempty convex subset of $\mathbb{R}^n$ and let $y \in \partial S$. Then there exists $q \in \mathbb{R}^n$, $q \neq 0$, such that*

$$\sup_{x \in cl\ S} q^t x \leq q^t y.$$

In geometric terms, $H := \{x \in \mathbb{R}^n : q^t x = q^t y\}$ is said to be a supporting hyperplane *for S at y: the hyperplane H contains the point y and the set S (as well as* cl S*) is contained the halfspace $\{x \in \mathbb{R}^n : p^t x \leq p^t y\}$.*

Proof. Let $Z := \text{cl } S$; then $\partial S \subset \partial Z$ (exercise). Of course, Z is closed and it is easy to show that Z is convex (use limit arguments). So there exists a sequence (y_k) in $\mathbb{R}^n \backslash Z$ such that $y_k \to y$. By Theorem A.2 there exists for every k a nonzero vector $p_k \in \mathbb{R}^n$ such that

$$\sup_{x \in Z} p_k^t x < p_k^t y_k.$$

Division by $|p_k|$ turns this into

$$\sup_{x \in Z} q_k^t x < q_k^t y_k,$$

where $q_k := p_k/|p_k|$ belongs to the unit sphere of $\mathbb{R}^n$. This sphere is compact (Bolzano-Weierstrass theorem), so we can suppose without loss of generality that (q_k) converges to some q, $|q| = 1$ (so q is nonzero). Now for every $x \in Z$ the inequality $q_k^t x < q_k^t y_k$, which holds for all k, implies

$$q^t x = \lim_k q_k^t x \leq \lim_k q_k^t y_k = q^t y,$$

and the proof is finished. QED

Theorem A.4 (set-set separation [1, Thm. 2.4.8]) *Let S_1, S_2 be two nonempty convex sets in $\mathbb{R}^n$ such that $S_1 \cap S_2 = \emptyset$. Then there exist $p \in \mathbb{R}^n$, $p \neq 0$, and $\alpha \in \mathbb{R}$ such that*

$$\sup_{x \in S_1} p^t x \leq \alpha \leq \inf_{y \in S_2} p^t y.$$

In geometric terms, $H := \{x \in \mathbb{R}^n : p^t x = \alpha\}$ is said to be a separating hyperplane *for S_1 and S_2: each of the two convex sets is contained in precisely one of the two halfspaces $\{x \in \mathbb{R}^n : p^t x \leq \alpha\}$ and $\{x \in \mathbb{R}^n : p^t x \geq \alpha\}$.*

Proof. It is easy to see that $S := S_1 - S_2$ is convex. Now $0 \notin S$, for otherwise we get an immediate contradiction to $S_1 \cap S_2 = \emptyset$. W distinguish now two cases: (i) $0 \in \text{cl } S$ and (ii) $0 \notin \text{cl } S$.

In case (i) we have $0 \in \partial S$, so by Theorem A.3 we then have the existence of a nonzero $p \in \mathbb{R}^n$ such that

$$p^t z \leq 0 \text{ for every } z \in S = S_1 - S_2, \tag{2}$$

i.e., for every $z = x - y$, with $x \in S_1$ and $y \in S_2$. This gives $p^t x \leq p^t y$ for all $x \in S_1$ and $y \in S_2$, whence the result.

In case (ii) we apply Theorem A.2 to get immediately (2) as well. The result follows just as in case (i). QED

Theorem A.5 (strong set-set separation [1, Thm. 2.4.10]) *Let S_1, S_2 be two nonempty closed convex sets in $\mathbb{R}^n$ such that $S_1 \cap S_2 = \emptyset$ and such that S_1 is bounded. Then there exist $p \in \mathbb{R}^n$, $p \neq 0$, and $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}$ such that*

$$\sup_{x \in S_1} p^t x \leq \alpha < \beta \leq \inf_{y \in S_2} p^t y.$$

Proof. As in the previous proof, it is easy to see that $S := S_1 - S_2$ is convex. Now S is also seen to be closed (exercise). As in the previous proof, we have $0 \notin S$. We can now apply Theorem A.2 to get the desired result, just as in case (ii) of the previous proof. QED

Subgradients

Ryan Tibshirani
Convex Optimization 10-725/36-725

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

Downsides:

- Requires f differentiable $\leftarrow$ next lecture
- Can be slow to converge $\leftarrow$ 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

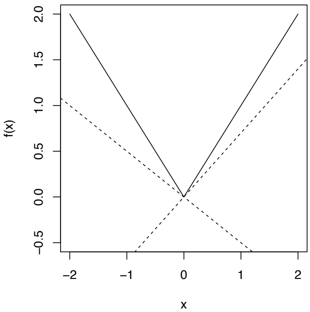
A subgradient of a convex function f at x is any $g \in \mathbb{R}^n$ such that

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

- 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)

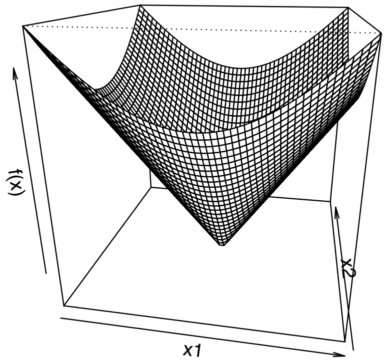
Examples of subgradients

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



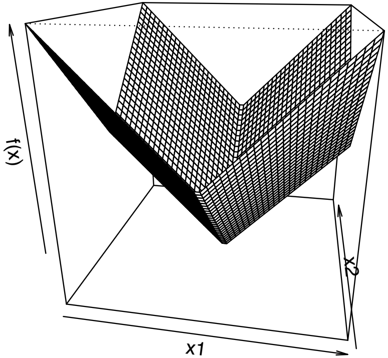
- For $x \neq 0$, unique subgradient $g = \mathrm{sign}(x)$
- For $x = 0$, subgradient g is any element of $[-1, 1]$

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



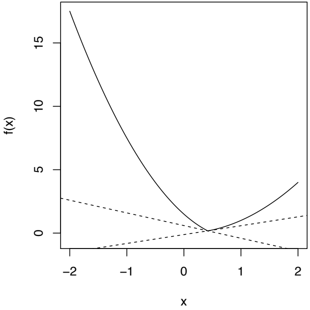
- For $x \neq 0$, unique subgradient $g = x/\|x\|_2$
- For $x = 0$, subgradient g is any element of $\{z : \|z\|_2 \leq 1\}$

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



- For $x_i \neq 0$, unique ith component $g_i = \mathrm{sign}(x_i)$
- For $x_i = 0$, ith 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 $I_C : \mathbb{R}^n \to \mathbb{R}$,

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

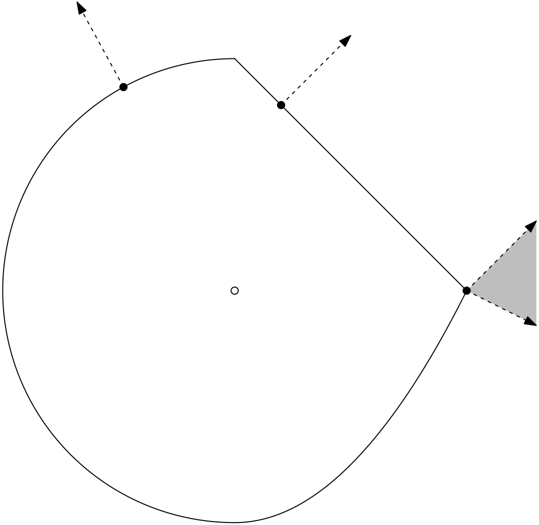
For $x \in C$, $\partial I_C(x) = \mathcal{N}_C(x)$, the normal cone of C at x, recall

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

Why? By definition of subgradient g,

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

- 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

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

  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

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

and under some regularity conditions (on S, f_s), we get an equality above

- 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),

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

I.e., $x^\star$ is a minimizer if and only if 0 is a subgradient of f at $x^\star$. 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 that for f convex and differentiable, the problem

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

is solved at x 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 see this? First recast problem as

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

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

But

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

as desired

Note: the condition $0 \in \partial f(x) + \mathcal{N}_C(x)$ is a fully general condition for optimality in 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:

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

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 $X_1, \dots X_p$ for columns of X. Then subgradient optimality 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: 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

Simplfied 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_λ is the soft-thresholding operator:

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

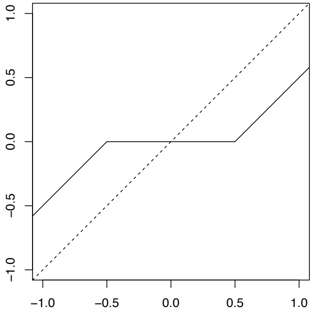
Check: from last slide, subgradient optimality conditions are

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

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 $|y_i| \leq \lambda$, $\beta_i = 0$, 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:

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

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. Then 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

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

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 ≤ 0

Now for $y \notin H$, we have $(x-u)^T(y-u) = \|x-u\|_2\|y-u\|_2 \cos\theta$ where θ 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