EECS 227C / STAT 260
Optimization algorithms and analysis
Spring 2017
Lecturer: Martin Wainwright

Further material on subgradients

In many applications, we are confronted with optimization problems involving functions that need not be not be differentiable. In the convex case, there is a natural generalization of differentiability, which leads us into subgradients and subdifferentials.

1 Basic definitions and properties

Consider a convex function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ taking values in the extended reals. The domain of the function f, or $\mathrm{dom}(f)$ for short, is the set of x for which $f(x) < +\infty$. For a given $x \in \mathbb{R}^d$, a subgradient of f at x is a vector $g \in \mathbb{R}^d$ such that

$$f(y) \geq f(x) + \langle g,\, y - x\rangle \qquad \text{for all } y \in \mathbb{R}^d.$$

In geometric terms, this means that the vector g specifies a supporting hyperplane to the epigraph of f at x. It is a natural generalization of the derivative for a convex function, since the gradient $\nabla f(x)$ satisfies this condition when f is differentiable.

We write $\partial f(x)$ to mean the collection of all subgradients of f at x. Some useful properties:

- For vectors x belonging to the (relative) interior of the domain, we are guaranteed that $\partial f(x)$ is a non-empty set. This can be shown by applying the supporting hyperplane theorem to the epigraph of the set (e.g., see Boyd and Vandenberghe for details).
- The subdifferential is a convex set (exercise for student).
- Differentiable case: we have $\partial f(x) = \{\nabla f(x)\}$ whenever f is differentiable at x.

The following provides us with the natural generalization of the zero-gradient optimality condition for convex optimization:

Theorem 1. *For a convex function f, we have $x^* \in \arg\min_{x\in\mathbb{R}^d} f(x)$ if and only if $0 \in \partial f(x^*)$.*

Proof. This claim is basically immediate from the definition. On one hand, if $0 \in \partial f(x^*)$, then the definition of sub-gradient guarantees that $f(y) \geq f(x^*) + \langle 0,\, y - x^*\rangle \geq f(x^*)$ for all y, which establishes the claim. On the other hand, if $f(y) \geq f(x^*)$ for all y, then we see that $0 \in \partial f(x^*)$. □

2 Subgradient calculus

The strong form of subgradient calculus refers to results that lead to a complete characterization of the entire subdifferential $\partial f(x)$. The weak form of subgradient calculus refer to rules that allow to us to compute a particular subgradient $g \in \partial f(x)$. In algorithmic contexts—such as when implementing a subgradient method—the "weak" results are often adequate.

The following results apply to convex functions whose domain is all of $\mathbb{R}^d$:

- Non-negative linear combinations: we have

$$\partial(\alpha_1 f_1 + \alpha_2 f_2)(x) = \alpha_1 \partial f_1(x) + \alpha_2 \partial f_2 \tag{1}$$

  for $\alpha_1, \alpha_2 \geq 0$.

- Affine transformations: for $f(x) = h(Ax + b)$, we have

$$\partial f(x) = A^T \partial h(Ax + b)$$

If we deal with convex functions that take the value ∞ at certain points of $\mathbb{R}^d$, then life becomes more complicated. For instance, the additivity property (1) may fail to hold unless we have an additional condition on the intersection of their domains—namely, that $\text{int}(\text{dom} f_1 \cap \text{dom} f_2)$ is non-empty. When this condition fails, we will see a problematic example in HW #4. On the other hand, it can be verified that we always have the inclusion

$$\partial f_1(x) + \partial f_2(x) \subseteq \partial(f_1 + f_2)(x).$$

This is often enough for algorithmic purposes, because it means that we can generate a subgradient vector for the sum $f_1 + f_2$ by computing $g_j \in \partial f_j(x)$ for $j = 1, 2$ and then forming the sum $g := g_1 + g_2$.

2.1 Danskin's theorem

Given a function $\phi : \mathbb{R}^d \to \mathbb{R}^m \to \mathbb{R}$ and compact set $\mathcal{Z} \subset \mathbb{R}^m$, consider the new function

$$f(x) := \max_{z \in \mathcal{Z}} \phi(x, z). \tag{2}$$

Danskin's theorem guarantees that under certain regularity conditions on ϕ, the function f is convex, and characterizes its subdifferential.

Here is a fairly general form of Danskin's theorem due to D. Bertsekas. In particular, suppose that the function $x \mapsto \phi(x, z)$ is convex and closed[1] for each $z \in \mathcal{Z}$. Suppose that f has a domain with non-empty interior, and that ϕ is continuous on $\text{int}(\text{dom}(f)) \times \mathcal{Z}$. For each $x \in \text{int}(\text{dom}(f))$, define the set $\mathcal{Z}^*(x) = \{z \in \mathcal{Z} \mid f(x) = \phi(x, z^*)\}$. Then we have

$$\partial f(x) = \text{conv}\Big(\bigcup_{z^* \in \mathcal{Z}^*(x)} \partial_x \phi(x, z^*) \Big).$$

Here $\partial_x \phi(\cdot, z^*)$ denotes the subdifferential of the function $x \mapsto \phi(\cdot, z^*)$. When ϕ is differentiable, then we have

$$\partial f(x) = \text{conv}\Big(\bigcup_{z^* \in \mathcal{Z}^*(x)} \nabla_x \phi(x, z^*) \Big).$$

[1]This means that its epigraph is closed

2.2 Finite maxima

We often come across functions defined in terms of finite maxima—that is,

$$f(x) := \max_{j=1,\ldots,N} g_j(x) \tag{3}$$

where $\{g_j\}_{j=1}^N$ are a collection of convex functions. It can be seen that f is always a convex function. Assuming that each g_j is also differentiable, let us verify that the subdifferential of f takes the form

$$\partial f(x) = \operatorname{conv}\Big\{\nabla g_j(x) \mid j \in J^*(x)\Big\} \qquad \text{where } J^*(x) = \{k \mid g_k(x) = f(x)\}.$$

This is actually a special case of Danskin's theorem. Define the function $\phi : \mathbb{R}^d \times \mathbb{R}^N \to \mathbb{R}$ via $\phi(x, z) = \sum_{j=1}^N z_j g_j(x)$, and note that

$$f(x) = \max_{z \in \mathcal{Z}} \phi(x, z), \qquad \text{where } \mathcal{Z} := \{z \in \mathbb{R}^N \mid z_j \in [0, 1], \textstyle\sum_{\ell=1}^N z_\ell = 1\}.$$

Noting that this set-up satisfies all the requirements of Danskin's theorem, the claim follows.

For future reference, we also note that a slight generalization of Danskin's theorem guarantees that, even when the g_j are not differentiable at x, then we still have

$$\partial f(x) = \bigcup_{j \in J^*(x)} \partial g_j(x).$$

2.3 Min-functions

Given a function $\varphi : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ that is jointly convex in (x, y), define the new function

$$f(x) = \inf_y \varphi(x, y) \tag{4}$$

Let's assume that $f(x) > -\infty$ for all x to avoid degeneracies. We claim that f is convex. Indeed, given two $x_1, x_2 \in \mathbb{R}^d$, let $y_1, y_2 \in \mathbb{R}^m$ be corresponding vectors such that $f(x_j) = \varphi(x_j, y_j)$ for $j = 1, 2$. (If the infimum is not achieved, we can find y_j that provides a δ-approximation to the infimum, work through the argument, and then take limits as $\delta \to 0$ at the end.) We then have

$$\begin{aligned} f\big(\lambda x_1 + (1-\lambda)x_2\big) &\leq \varphi\big(\lambda x_1 + (1-\lambda)x_2, \lambda y_1 + (1-\lambda)y_2\big) \\ &\leq \lambda\varphi(x_1, y_1) + (1-\lambda)\varphi(x_2, y_2) \\ &= \lambda f(x_1) + (1-\lambda) f(x_2). \end{aligned}$$

Lemma 1. *For a given x, suppose that there exists some y such that $f(x) = \varphi(x, y)$. Then for any subgradient $(g, 0) \in \partial\varphi(x, y)$, we have $g \in \partial f(x)$.*

Proof. For any pair (x', y'), we have

$$\begin{aligned} \varphi(x', y') &\geq \varphi(x, y) + \langle (g, 0), \, (x' - x, y' - y)\rangle \\ &= f(x) + \langle g, \, x' - x\rangle. \end{aligned}$$

This inequality holds for all y', we may take the infimum to conclude that

$$f(x') = \inf_y \varphi(x', y') \geq f(x) + \langle g, \, x' - x\rangle.$$

This inequality holds for all x', which shows that $g \in \partial f(x)$, as claimed. □

3 Some examples

Let us consider some examples to illustrate these properties.

3.1 Revisiting the absolute value function

We can compute the subdifferential of the ℓ_1-norm as a consequence of Danskin's theorem. In particular, defining the function $\phi(x, z) = xz$, note that we have

$$f(x) := |x| \;=\; \max_{|z|\leq 1} \phi(x, z).$$

It is easy to see that the conditions of Danskin's theorem are satisfied, and that $x\ \mathrm{sign}(x) = |x|$, so that the maximum is achieved at $z^* = \mathrm{sign}(x)$ for $x \neq 0$. When $x = 0$, the maximum is achieved for all $z^* \in [-1, 1]$, whence

$$\partial|x| = \begin{cases} \{\mathrm{sign}(x)\} & \text{if } x \neq 0 \\ [-1, 1] & \text{otherwise.} \end{cases}$$

3.2 Piecewise linear functions

A bit more generally, a piecewise linear function takes the form

$$f(x) = \max_{j=1,\ldots,N} \big(\langle a_j,\, x\rangle + b_j\big)$$

where each $(a_j, b_j) \in \mathbb{R}^d \times \mathbb{R}$ defines a hyperplane. For example, the function $f(x) = |x|$ is a very special case with $d = 1$, $N = 2$, $(a_1, b_1) = (1, 0)$ and $(a_2, b_2) = (-1, 0)$.

From Section 2.2, we have

$$\partial f(x) = \mathrm{conv}\,\big\{a_j \mid j \in J^*(x)\big\} \qquad \text{where } J^*(x) = \{k \mid \langle a_k,\, x\rangle + b_k = f(x)\}.$$

3.3 Indicator functions and their subdifferentials

Given a closed convex set $\mathcal{C}$, let us define the indicator function

$$\mathbb{I}_{\mathcal{C}}(x) := \begin{cases} 0 & \text{if } x \in \mathcal{C} \\ +\infty & \text{otherwise.} \end{cases} \tag{5}$$

It is easy to see that $\mathbb{I}_{\mathcal{C}}$ is a convex function. Let us characterize its subdifferential. For points x in the interior of $\mathcal{C}$, it is easy to see that $\partial\mathbb{I}_{\mathcal{C}}(x) = \{0\}$. The interesting cases are when x belongs to the boundary of $\mathcal{C}$. In this case, we have $g \in \partial\mathbb{I}_{\mathcal{C}}(x)$ if and only if

$$\mathbb{I}_{\mathcal{C}}(y) \geq 0 + \langle g,\, y - x\rangle \qquad \text{for all } y \in \mathbb{R}^d.$$

For $y \notin \mathcal{C}$, we have $\mathbb{I}_{\mathcal{C}}(y) = +\infty$, so this constraint is not meaningful. So it reduces to the condition $\langle g,\, y - x\rangle \leq 0$ for all $y \in \mathcal{C}$. This set of constraints defines a set known as the normal cone at x, viz.

$$\mathcal{N}_{\mathcal{C}}(x) = \big\{g \in \mathbb{R}^d \mid \langle g,\, y - x\rangle \leq 0 \quad \text{for all } y \in \mathcal{C}\big\}. \tag{6a}$$

Note that the normal cone at x is polar to the tangent cone at x given by

$$\mathcal{T}_{\mathcal{C}}(x) := \big\{y \in \mathbb{R}^d \mid y - x \in \mathcal{C}\big\}. \tag{6b}$$

The tangent cone corresponds to the set of directions that are locally feasible from x. The normal cone corresponds to the set of all directions that have non-positive inner product with any feasible direction in the tangent cone.

3.4 Distance from a given convex set

Given a closed convex set $\mathcal{C}$, we can define the function

$$\mathrm{dist}_{\mathcal{C}}(x) = \min_{y \in \mathcal{C}} \|x - y\|_2 \;=\; \|x - \Pi_{\mathcal{C}}(x)\|_2. \tag{7}$$

Let us use Lemma 1 to find an element of the subdifferential $\partial \mathrm{dist}_{\mathcal{C}}(x)$ for an $x \notin \mathcal{C}$. Introducing the function $\varphi(x, y) = \|x - y\|_2$, note that it is jointly convex in (x, y), and moreover we have at the the point $y = \Pi_{\mathcal{C}}(x)$, we have

$$\Big(\frac{x - \Pi_{\mathcal{C}}(x)}{\|x - \Pi_{\mathcal{C}}(x)\|_2}, 0\Big) \in \partial\varphi(x, \Pi_{\mathcal{C}}(x)).$$

From Lemma 1, we conclude that the vector

$$\frac{x - \Pi_{\mathcal{C}}(x)}{\|x - \Pi_{\mathcal{C}}(x)\|_2} \in \partial \mathrm{dist}_{\mathcal{C}}(x)$$

whenever $\mathrm{dist}_{\mathcal{C}}(x) > 0$.

If $\mathrm{dist}_{\mathcal{C}}(x) = 0$, then we have $0 \in \mathrm{dist}_{\mathcal{C}}(x)$ by the usual optimality condition.

3.5 Alternating projections onto convex sets

As an extension of the previous example, suppose that we are given a collection of convex sets $\{\mathcal{C}_j\}_{j=1}^N$, and we wish to find some point $x^* \in \cap_{j=1}^N \mathcal{C}_j$, assuming that the intersection is non-empty. Note that $x^* \in \cap_{j=1}^N \mathcal{C}_j$ if and only if $\mathrm{dist}_{\mathcal{C}_j}(x) = 0$ for all $j = 1, \ldots, N$. Thus, we can reformulate our problem in terms of the minimization problem

$$f(x) = \max_{j=1,\ldots,N} \mathrm{dist}_{\mathcal{C}_j}(x).$$

By our previous results on max-functions, this is a convex function. For any x such that $f(x) > 0$, we can find an element $g \in \partial f(x)$ as follows:

- choose an index j such that $\mathrm{dist}_{\mathcal{C}_j}(x) = f(x)$, meaning that $\mathcal{C}_j$ among the sets furthest away from x.
- compute the unit norm vector

$$g = \frac{x - \Pi_{\mathcal{C}_j}(x)}{\|x - \Pi_{\mathcal{C}_j}(x)\|_2}$$

We can thus run a subgradient method with this algorithm. In particular, if we use the step size $\alpha^\ell > 0$, then we have

$$x^{\ell+1} = x^\ell - \alpha^\ell g^\ell \;=\; x^\ell - \alpha^\ell \frac{x - \Pi_{\mathcal{C}_j}(x)}{\|x - \Pi_{\mathcal{C}_j}(x^\ell)\|_2} \qquad \text{for some } j \text{ such that } f(x^\ell) = \|x^\ell - \Pi_{\mathcal{C}_j}(x^\ell)\|_2.$$

If we use the handy step size $\alpha^\ell = f(x^\ell)$, then we have

$$x^{\ell+1} = \Pi_{\mathcal{C}_j}(x^\ell) \qquad \text{for some } j \text{ such that } f(x^\ell) = \|x^\ell - \Pi_{\mathcal{C}_j}(x^\ell)\|_2.$$

In the case of two sets (i.e., $\mathcal{C} = \mathcal{C}_1 \cap \mathcal{C}_2$), this method is exactly alternating projections onto convex sets. See Figure 3.5 for an illustration.

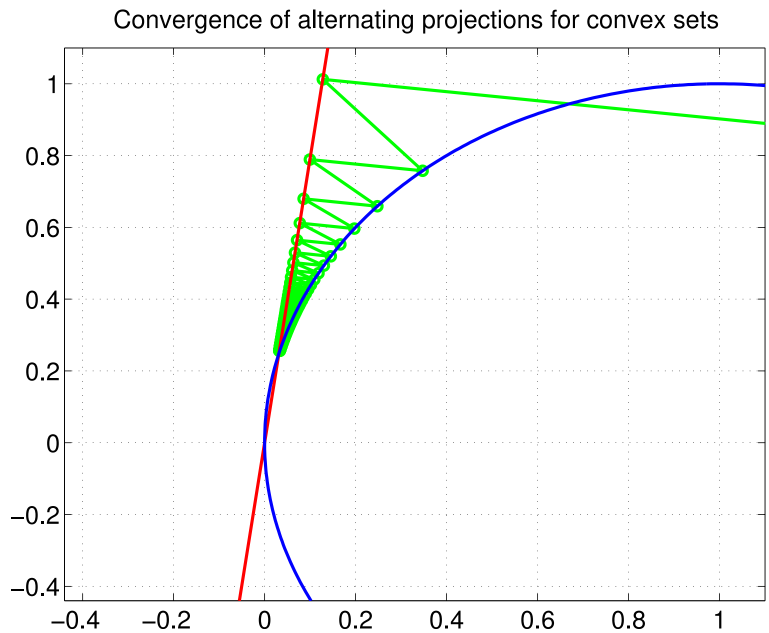

Figure 1: Illustration of the alternating projection method for finding a point $x^* \in \mathcal{C}_1 \cap \mathcal{C}_2$, with the halfspace $\mathcal{C}_1 = \{x \in \mathbb{R}^2 \mid -\sin(\theta)x_1 + \cos(\theta)x_2 \geq 0\}$ with $\theta = 1.4451$, and circle $\mathcal{C}_2 = \{x \in \mathbb{R}^2 \mid \|x - (1,0)\|_2 \leq 1\}$.

4 Consequences for constrained optimization

The zero-subgradient optimality condition stated in Theorem 1 has an interesting corollary for the problem of constrained optimization over a convex set. In particular, consider a problem of the form $\min_{x \in \mathcal{C}} g(x)$, where $\mathcal{C}$ is a closed convex set, and $g : \mathbb{R}^d \to \mathbb{R}$ is convex. In previous lectures, we have seen that the condition

$$\langle \nabla g(x^*),\, y - x^* \rangle \geq 0 \qquad \text{for all } y \in \mathcal{C} \tag{8}$$

is necessary and sufficient for the constrained optimality of x^*. This statement is actually a special case of Theorem 1.

Recalling the indicator function $\mathbb{I}_{\mathcal{C}}$ from equation (5), define the new convex function $f(x) = g(x) + \mathbb{I}_{\mathcal{C}}(x)$. Note that we have the equivalence $\min_{x \in \mathcal{C}} g(x) = \min_{x \in \mathbb{R}^d} f(x)$. Now from Example 3.3 and sub-gradient calculus (using the fact that $\text{dom}(g) = \mathbb{R}^d$), we know that

$$\partial f(x) = \nabla g(x) + \mathcal{N}_{\mathcal{C}}(x).$$

Hence the condition $0 \in \partial f(x^*)$ is equivalent to $-\nabla g(x^*) \in \mathcal{N}_{\mathcal{C}}(x^*)$, which is in turn equivalent to the original statement (8).