University of California, Berkeley Lecturer: Benjamin Recht

CS227C/STAT260 Convex Optimization and Approximation: Optimization for Modern Data Analysis

Feburary 24, 2015 Notes: Benjamin Recht

Lecture 10: Subgradients

Up until now, we have been assuming that our convex functions are everywhere differentiable. Indeed, we have been a bit cavalier with this fiat. Consider the SVM loss function

$$f(x) = \max(1 - x, 0)$$

This function is differentiable everywhere except at $x = 1$. But what do we do at 1 if we want to compute a derivative. 1 is clearly a minimizer of f, but what general rule can prove such a thing. And if we had the function

$$f(x) = \max(-2x, -x, x - 2)$$

then we have a function that is not differentiable at $x = \{0, 1, 2\}$.

We can obviously create a long list of interesting functions to optimize that are convex but fail to be differentiable. *Every norm* is not differentiable at $x = 0$. More exotically, the maximum eigenvalue of a symmetric matrix is convex, but not differentiable (consider the example where the maximum eigenvalue has multiplicity greater than 1).

In this lecture we turn to analyzing such non-smooth convex functions. In the next lecture we will show that they can be optimized in polynomial time, albeit at a very slow rate of $O(\epsilon^{-2})$. But nonsmooth functions also play a major role in constrained optimization, and we will see that understanding their structure will be critical when we turn to constrained problems.

1 Subgradients

Recall our first order condition for convex, differentiable functions. For all z and x we had the relation

$$f(z) \geq f(x) + \nabla f(x)^T (z - x)\,.$$

This condition asserts that every convex function can be lower bounded at every point by an affine function. It turns out that this notion generalizes to non-smooth functions as well.

We say that g is a *subgradient* of f at x if

$$f(z) \geq f(x) + g^T (z - x)$$

for all z. This is precisely a generalization of the first-order convexity conditions. Under very mild conditions, every convex function has at least one subgradient.

Before turning to the abstract analysis of subgradients, we can already see that this notion buys us something very powerful. Indeed, suppose f has at least one subgradient at every point. Then $x_\star$ is a global minimizer of f if and only if 0 is a subgradient of f at x. The proof of this wholly trivial. 0 is a subgradient if and only if

$$f(z) \geq f(x_\star) + 0^T (y - x)\,.$$

Note that if f is convex and differentiable, then there is unique subgradient at every x and it is equal to $\nabla f(x)$. We can easily verify this in the one dimensional case. Obviously, $\nabla f(x)$ is a subgradient. Conversely, let g be any subgradient. Note that $f(x+h) \geq f(x) + g^T h$. Rearranging both sides and taking the limit as h goes to zero verifies

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \geq g$$

A similar argument shows

$$-f'(x) = \lim_{h \to 0} \frac{f(x-h) - f(x)}{h} \geq -g$$

Together, these show that $f'(x) = g$.

We will prove this fact in the multidimensional case when we turn to directional derivatives.

2 Directional Derivatives

Though convex functions might not be differentiable at every point, they are differentiable *in every direction.* The *directional derivative* of f at x in direction v is defined as

$$f'(x; v) = \lim_{\alpha \downarrow 0} \frac{f(x + \alpha v) - f(x)}{\alpha}$$

Note that if f is differentiable, $f'(x; v) = \nabla f(x)^T v$. Moreover, $f'(x; v) = -f'(x; -v)$ when f is differentiable. This equality need not hold when f is merely convex and not smooth. But we will show that the limit always exists. This is rather surprising, and a deep property of convex functions.

To proceed, let's begin with one dimensional functions. Let $f : I \to \mathbb{R}$ where I is an interval $[a, b]$ with $a = -\infty$ and $b = \infty$ allowed. If $x < y < z$, then we have the relationship

$$\frac{f(y) - f(x)}{y - x} \leq \frac{f(z) - f(x)}{z - x} \leq \frac{f(z) - f(y)}{z - y} \tag{1}$$

This can be verified by drawing a picture. But the proof is also an immediate consequence of the definition of convex functions.

For $x > a$ and $x \leq b - \alpha$, we can define

$$s^+(x, \alpha) := \frac{f(x + \alpha) - f(x)}{\alpha}$$

which provides a measure of the slope of the function as we move in the positive direction away from x. Note that if $0 < \alpha_1 < \alpha_2$,

$$s^+(x, \alpha_1) \leq s^+(x, \alpha_2)\,. \tag{2}$$

This can be verified by using $y = x + \alpha_1$ and $z = x + \alpha_2$ in equation (1).

Now, we can define

$$f^+(x) = \lim_{\alpha \downarrow 0} \frac{f(x + \alpha) - f(x)}{\alpha} = \inf_{\alpha > 0} \frac{f(x + \alpha) - f(x)}{\alpha}\,.$$

By (2), either $f^+(x)$ is finite, or it is equal to $-\infty$. Similarly, we can define

$$s^-(x,\alpha) := \frac{f(x) - f(x-\alpha)}{\alpha}$$

and

$$f^-(x) = \lim_{\alpha \downarrow 0} \frac{f(x) - f(x-\alpha)}{\alpha} = \sup_{\alpha > 0} \frac{f(x) - f(x-\alpha)}{\alpha}\,.$$

$f^-(x)$ is either finite or equal to ∞. By convention, we'll set $f^-(a) = -\infty$ and $f^+(b) = \infty$, as we can't properly define these limits. We can collect some properties of these limits in the following

Proposition 1 *1. $f^-(x) \le f^+(x)$ for all $x \in I$.*

2. $x \in (a,b)$ implies that $f^+(x)$ and $f^-(x)$ are both finite.

3. If $a \le x < z \le b$, then $f^+(x) < f^-(z)$.

4. f^- and f^+ are both nondecreasing

Proof

1. This is clear at the endpoints by definition. For $t > a$,

   $$s^-(t,\alpha) \le s^+(t,\alpha)\,.$$

   This follows by using $x = t - \alpha$, $y = t$, and $z = t + \alpha$ in (1). Taking limits completes the argument.

2. $f^-(x) \ge s^-(x,\alpha)$ for all α, so it is bounded below. Similarly, $f + (x)$ is bounded above. But then part (a) implies that both must be finite.

3. Using (1) with $y = \frac{1}{2}(z+x)$, we have

   $$s^+(x, \tfrac{1}{2}(z+x)) \le s^-(z, \tfrac{1}{2}(z+x))\,.$$

   Using $f^+(x) \le s^+(x, \frac{1}{2}(z+x))$ and $f^-(z) \ge s^-(z, \frac{1}{2}(z+x))$ completes the proof.

4. This follows from (a) and (c).

■

Note that by our definition, $f^+(x) = f'(x;1)$ and $f^-(x) = -f'(x;-1)$ This proposition proves that the directional derivative exists in both directions for one-dimensional convex functions. Moreover, both $f'(x;1)$ and $-f'(x;-1)$ are nondecreasing functions.

For the multidimensional case, choose $v \in \mathbb{R}^d$ and define $F(t) = f(x+tv)$. F is a convex function and

$$f'(x;v) = \lim_{\alpha \downarrow 0} \frac{f(x+\alpha v) - f(x)}{\alpha} = \lim_{\alpha \downarrow 0} \frac{F(\alpha - F(0)}{\alpha} = F^+(0)$$

Hence, the directional derivative exists in all directions. Moreover, $-f'(x;-v) \le f'(x;v)$ from our proposition.

3 The subdifferential

The collection of subgradients of f at x is called the *subdifferential* of f at x and is denoted $\partial f(x)$. We will establish that the subdifferential of convex functions are nonempty, convex, and compact. (**Note for fans of convex analysis:** we will skirt issues of properness and regularity in this section by assuming that f is finite everywhere.)

Proposition 2 *g is a subgradient of f at x if and only if*

$$f'(x;y) \geq \langle g, y \rangle$$

for all y. Moreover,

$$f'(x;y) = \sup_{g \in \partial f(x)} \langle g, y \rangle$$

This proposition asserts that the function mapping y to $f'(x;y)$ is the support function for the subdifferential of f at x.

Proof [Rockafellar] By the definition of the subgradient,

$$\frac{f(x+\alpha y) - f(x)}{\alpha} \geq \langle g, y \rangle$$

for all $\alpha > 0$. Using (2), this proves the first assertion of the theorem. For the second, note that $f'(x;y)$ is a convex function of y. It is also homogeneous in the sense that

$$f'(x;\alpha y) = \alpha f'(x;y)$$

for $\alpha > 0$. Thus,

$$f'(x;y) = \sup_u \ \langle u, y \rangle - \varphi^*(u)$$

where $\varphi^*(u)$ denotes the Fenchel conjugate of $f'(x;y)$ as a function of y. That is, the closure of $f'(x;y)$ is equal to the double conjugate of $f'(x;y)$. Homogeneity of the directional derivative shows that $\varphi^*(u)$ must equal 0 if $\varphi^*(u)$ is defined (and ∞ otherwise). In particular,

$$\varphi^*(u) = \sup_y \langle u, y \rangle - f'(x;y) \, .$$

Hence, $\varphi^*(u) = 0$ if and only if $f'(x;y) \geq \langle u, y \rangle$ for all y if and only if $u \in \partial f(x)$. This completes the proof. ∎

(**Another note for the convex analysis junkies:** This assumes that $f'(x;y)$ is closed and proper. But this can be shown because we are assuming $f(x)$ is finite everywhere. See Rockafellar Thm 23.4.)

Proposition 3 *Let f be a convex function. Then $\partial f(x)$ is nonempty, convex, and compact for all x.*

Proof Existence follows from Proposition 2. It is immediate from the definition that the convex combination of two subgradients is a subgradient. Moreover, the subdifferential must be closed as it is the intersection of set of inequalities. This point is a bit subtle, but note that

$$\partial f(x) = \{g \;:\; \langle a, g\rangle \le b \;\text{ whenever } a = z - x\,,\; b = f(z) - f(x) \;\text{ for some }\; z \in \mathbb{R}^d\}\,.$$

To prove boundedness, note that the directional derivative is finite everywhere. ■

Proposition 2 also lets us show that if f is convex and differentiable, then the $\nabla f(x)$ is the unique subgradient of f at x. To see this, note that

$$f'(x; y) = \nabla f(x)^T y$$

for all y. If there were two subgradients $g_1 \neq g_2$, then there would be a direction y such that

$$\langle g_1, y\rangle > \langle g_2, y\rangle\,.$$

Hence, for this direction, $f'(x; y) \neq -f'(x; -y)$.

3.1 Subdifferential calculus

We can enumerate some properties important of the subdifferential. Essentially, these properties are what lets us form a "subgradient calculus" from which we can compute subgradients from simple primitives.

Proposition 4 *For $\alpha > 0$, $\partial\,(\alpha f)\,(x) = \alpha\partial f(x)$.*

Proof Immediate from the definition. ■

Proposition 5 *If $h(x) = f(Ax + b)$, then $\partial h(x) = A^T \partial f(Ax + b)$.*

Proof [Bertsekas, 4.2.5] From the definition of directional derivatives, it follows that

$$h'(x; y) = f'(Ax + b; Ay)\,.$$

Let $g \in \partial f(Ax + b)$. Then

$$\langle g, z\rangle \le f'(Ax + b; z)\,.$$

This implies that

$$\langle A^T g, y\rangle = \langle g, Ay\rangle \le h'(x; y)$$

for all y. Thus $A^T \partial f(Ax + b) \subset \partial h(x)$. Conversely, suppose $v \in \partial h(x)$ but $v \notin A^T \partial f(Ax + b)$. Since the set $\Omega := A^T \partial f(Ax + b)$ is compact, there exists a hyperplane strictly separating g from Ω. That is, there is a vector y and a scalar β such that

$$y^T (A^T g) < \beta < y^T v$$

for all $g \in \partial f(Ax + b)$. This means that

$$\sup_{g \in \partial f(Ax+b)} (Ay)^T g < y^T v\,.$$

Using Proposition 2, this means

$$h'(x,y) = f'(Ax+b; Ay) < y^T v$$

but this contradicts the fact that $v \in \partial h(x)$. ■

A similar argument gives the following:

Proposition 6 $\partial(f_1+f_2)(x) = \partial f_1(x) + \partial f_2(x) = \{g+h \ : \ g \in \partial f_1(x), \ h \in \partial f_2(x)\}$.

Proof Certainly the "$\supseteq$" direction is an immediate consequence of the definition. The other inclusion is pretty much the same separating hyperplane argument we just made. Suppose there exists $g \in \partial(f_1+f_2)(x)$ but $g \neq \partial f_1(x)+\partial f_2(x)$. Then there exists a hyperplane strictly separating g and $\partial f_1(x) + \partial f_2(x)$:

$$y^T(g_1+g_2) < b < y^T g$$

for all $g_1 \in \partial f_1(x)$, $g_2 \in \partial f_2(x)$. Taking suprema, we see

$$(f_1+f_2)'(x;y) = f_1'(x;y) + f_2'(x;y) = \sup_{g_1 \in \partial f_1(x)} \langle y, g_1 \rangle + \sup_{g_2 \in \partial f_2(x)} \langle y, g_2 \rangle < y^T g$$

contradicting the assertion that $g \in \partial(f_1+f_2)(x)$. ■

3.2 The max function

We conclude this section with one of the most important tools in subdifferential calculus: the subgradient of the max-function. Let I be a compact subset of $\mathbb{R}^n$. Let $\varphi : \mathbb{R}^d \times I \to \mathbb{R}$ be continuous and assume $\varphi(\cdot; i)$ is convex for all $i \in I$. Then

$$f(x) = \max_{i \in I} \varphi(x,i)$$

f is clearly convex as it is the pointwise maximum of convex functions.

Let $\varphi'(x,i;y)$ denote the directional derivative of $\varphi(\cdot,i)$ at x in the direction y. For a fixed x, let $I_{\max}(x)$ denote the set of maximizing points:

$$I_{\max}(x) := \{j \ : \ \varphi(x,j) = \max_{i \in I} \varphi(x,i)\}\,.$$

The following proposition characterizes the subdifferential of this max-function

Theorem 7 (Danskin's Theorem) .

$$f'(x,y) = \max_{i \in I_{\max}(x)} \varphi'(x,i;y)\,.$$

If $\varphi(\cdot,i)$ is a differentiable function of x for all $i \in I$ and $\nabla_x\varphi(x,\cdot)$ is continuous on I for all x then

$$\partial f(x) = \operatorname{conv}\{\nabla_x \varphi(x,i) \ : \ i \in I_{\max}(x)\}$$

Proof [Bertsekas] Using the definition of f, we have that for any $i \in I_{\max}(x)$,

$$\frac{f(x+\alpha y)-f(x)}{\alpha} \geq \frac{\varphi(x+\alpha y,i)-\varphi(x,i)}{\alpha}$$

because $f(x+\alpha y) \geq \varphi(x+\alpha y, i)$ and $f(x)=\varphi(x,i)$. Taking the limit as α tends to zero proves $f'(x;y) \geq \varphi'(x,i;y)$. Since this is true for all $i \in I_{\max}(x)$,

$$f'(x;y) \geq \sup_{i \in I_{\max}(x)} \varphi'(x,i;y)\,.$$

For the reverse inequality, Let α_k be a decreasing sequence on positive scalars whose limit is 0. For each k, let i_k be some element of $I_{\max}(x+\alpha_k y)$. Since I is compact, there is a subsequence of $\{i_k\}$ that converges to $\hat{i} \in I$. Without loss of generality, assume that i_k converges to $\hat{i}$ (you could just subsample the sequence). Then we have

$$\varphi(x+\alpha_k y, i_k) \geq \varphi(x+\alpha_k y, i)$$

for all $i \in I$. Taking the limit of both sides shows $\varphi(x,\hat{i}) \geq \varphi(x,i)$ for all $i \in I$. This means that $\hat{i} \in I_{\max}(x)$, and we have

$$\begin{aligned}
f'(x;y) &\leq \frac{f(x+\alpha_k y)-f(x)}{\alpha_k}\\
&= \frac{\varphi(x+\alpha_k y, i_k)-\varphi(x,\hat{i})}{\alpha_k}\\
&\leq \frac{\varphi(x+\alpha_k y, i_k)-\varphi(x,i_k)}{\alpha_k}\\
&= -\frac{\varphi(x+\alpha_k y+\alpha_k(-y), i_k)-\varphi(x+\alpha_k y,i_k)}{\alpha_k}\\
&\leq -\varphi'(x+\alpha_k y, i_k; -y)\\
&\leq \varphi'(x+\alpha_k y, i_k; y)
\end{aligned}$$

Letting k go to infinity, we have

$$f'(x;y) \leq \lim_{k\to\infty} \varphi'(x+\alpha_k y, i_k; y) \leq \varphi'(x,\hat{i};y)\,.$$

This proves the first part of the theorem. For the second part, note that for all $i \in I_{\max}(x)$,

$$\begin{aligned}
f(z) &= \max_{j \in I} \varphi(z,j)\\
&\geq \varphi(z,i)\\
&\leq \varphi(x,i) + \nabla_x \varphi(x,i)^T(z-x)\\
&= f(x) + \nabla_x \varphi(x,i)^T(z-x)
\end{aligned}$$

proving that $\varphi(x,i) \in \partial f(x)$. This proves the "$\supseteq$" direction. For the reverse direction, we use our now familiar separating hyperplane argument.

Suppose there is a $v \in \partial f(x)$ that is not in $\operatorname{conv}\{\nabla_x \varphi(x,i) \ : \ i \in I_{\max}(x)\}$. Now, $I_{\max}(x)$ is a compact set because $\varphi(x,\cdot)$ is continuous. Moreover, $\operatorname{conv}\{\nabla_x \varphi(x,i) \ : \ i \in I_{\max}(x)\}$ must be compact. Thus, there exists a y and scalar β such that

$$\langle v, y\rangle > \beta > \nabla_x \varphi(x,i)^T y$$

for all $i \in I_{\max}(x)$. But this means that

$$\langle v, y\rangle > \max_{i \in I_{\max}(x)} \nabla_x \varphi(x,i)^T y = f'(x;y)$$

which is a contradiction by Proposition 2. ■

4 Extended Real-Valued Functions

We say that a function $f : \mathbb{R}^n \to \mathbb{R} \cup \infty$ is *an extended real-valued function.* This is a useful construction for constrained optimization, but one needs to be a bit careful when playing with these functions, as certain properties are not inherited from their real counterparts.

The most common extended real valued function is the indicator function, $\mathbb{I}_C$, of a convex set C

$$\mathbb{I}_C(x) = \begin{cases} 0 & x \in C \\ \infty & \text{otherwise} \end{cases} . \tag{3}$$

Note that convex extended real-valued functions form a convex cone. The sum of two indicator functions is simply the indicator of their intersection. We can even define subgradients of extended convex function. Note that for an indictor function, the subdifferential need not be convex. As a simple example, let $[a,b] \subset \mathbb{R}$ be an interval. Then,

$$\partial I_{[a,b]}(a) = (-\infty, 0]$$

This is simple to check. If, $x \in [a,b]$, and $g \geq 0$,

$$I_{[a,b]}(x) = 0 \geq 0 + g(x-a) = I_{[a,b]}(x) + g(x-a)\,.$$

Similarly, if $x \notin [a,b]$,

$$I_{[a,b]}(x) = \infty \geq 0 + g(x-a) = I_{[a,b]}(x) + g(x-a)\,.$$

Many calculations are simple manipulations inequalities using ∞ in this manner. However, some facts are more difficult. For example, $\partial f_1(x) + \partial f_2(x)$ may not equal $\partial(f_1 + f_2)(x)$. Let $C_1 = \{(x,y) \ : \ x^2 + y^2 \leq 1\} \subset \mathbb{R}^2$ and $C_2 = \{(x,y) \ : \ x = 1\} \subset \mathbb{R}^2$. Then $C_1 \cap C_2 = \{(1,0)\} =: C_3$ and

$$I_{C_1} + I_{C_2} = I_{C3}$$

Now, at $(1,0)$,

$$\begin{aligned} \partial I_{C_1}(1,0) &= \{(1,0)\} \\ \partial I_{C_2}(1,0) &= \{(g,0) \ : \ g \in \mathbb{R}\} \\ \partial I_{C_3}(1,0) &= \mathbb{R}^2 \end{aligned}$$

4.1 Constrained Optimization

Let C be a convex set and let $\mathbb{I}_C$ denote its indicator function. What's the subdifferential of $\mathbb{I}_C(x)$ for $x \in C$? By definition $g \in \partial \mathbb{I}_C(x)$ if and only if

$$\mathbb{I}_C(y) \geq \mathbb{I}_C(x) + g^T(y - x) \tag{4}$$

for all y. This is equivalent to

$$\partial \mathbb{I}_C(x) = \{g \;:\; g^T(x - y) \geq 0 \;\forall y \in C\} \tag{5}$$

for $x \in C$. This set is often called the *normal cone* of C at x.

Consider the constrained optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in C \end{array} \tag{6}$$

for smooth, convex f. Then $x_\star$ is optimal if and only if $-\nabla f(x_\star) \in \partial \mathbb{I}_C(x_\star)$. In other words, $x_\star$ is optimal if and only if the directional derivative $f'(x_\star; y) = \nabla f(x_\star)^T(y - x_\star)$ is nonnegative for all $y \in C$. This is precisely the minimum principle derived last lecture.