10-725: Convex Optimization **Spring 2023**

Lecture 2: January 19

Lecturer: Siva Balakrishnan

2.1 Convex Functions

There are three characterizations of convexity that you should be familiar with:

1. **No Assumptions (Zeroth-Order):** This is the definition we discussed last time, i.e. f is convex if its domain is a convex set and, for any $x, y \in \text{dom}(f)$,
$$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta) f(y).$$
2. **Differentiable (First-Order):** Suppose our function f has a derivative (at all points in its domain) then, f is convex if its domain is a convex set and, for any $x, y \in \text{dom}(f)$,
$$f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle.$$
3. **Twice Differentiable (Second-Order):** A function f is convex, if its domain is a convex set and, for any $x \in \text{dom}(f)$,
$$\nabla^2 f(x) \succeq 0.$$

In your HW you will explore some connections between these definitions (in particular, showing that (3) $\implies$ (2) $\implies$ (1)). You can find the proof that (2) $\implies$ (1) in the BV textbook (but it will still likely be a part of your HW).

It is also worth noting that there is a definition analogous to (2) above in the case when the function is not differentiable everywhere.

2' **Non-Smooth:** A function f is convex if its domain is a convex set, and if at every point $x \in \text{dom}(f)$, there exists a vector g_x such that, for any $y \in \text{dom}(f)$,
$$f(y) \ge f(x) + \langle g_x, y - x \rangle.$$
It is worth noting that if f is differentiable at x, then there is only one vector which will satisfy the above definition and it will coincide with the usual gradient, i.e. $g_x = \nabla f(x)$.

Any g_x which satisfies the above property is called a *subgradient* of f at x. The set of all subgradients at a point x is called the *subdifferential* of f at x and it will be denoted as $\partial f(x)$.

Except for some very pathological functions (and only at the boundary of their domain) subgradients always exist. Formally, one can for instance show that a subgradient g_x of a convex function f at x exists if x is in the interior of their domain.

Notational Note: I will often stop adding the qualifiers "for $x, y \in \text{dom}(f)$". One way to make this precise (I, and most textbooks do this implicitly) is to allow f to be whats called an *extended* function, and define it to be ∞ outside its (effective) domain. This won't change any of its convexity properties, and things like the first and zeroth-order characterizations will now make sense for any $x, y \in \mathbb{R}^d$.

2.1.1 An example

Let us consider the quadratic function $f(x) = \frac{1}{2}x^T Q x + a^T x + b$ where $Q \succeq 0$.

Applying definition (3) is easiest, since $\nabla^2 f(x) = Q$ and this is PSD.

Now, let us try to apply definition (2). It is a differentiable function, with gradient $\nabla f(x) = Qx + a$. So we need to verify if,

$$\frac{1}{2}y^T Q y + a^T y + b \overset{?}{\geq} \frac{1}{2}x^T Q x + a^T x + b + \langle Qx + a, y - x \rangle.$$

Re-arranging we obtain that we need to check if,

$$\frac{1}{2}(y - x)^T Q (y - x) \geq 0,$$

which is certainly the case since $Q \succeq 0$.

Finally, let us try to apply definition (1). We see (after cancelling some terms) that we need to verify if for $0 \leq \theta \leq 1$,

$$\frac{1}{2}\left(\theta x + (1-\theta) y\right)^T Q \left(\theta x + (1-\theta) y\right) \overset{?}{\leq} \frac{\theta}{2} x^T Q x + \frac{1-\theta}{2} y^T Q y.$$

Now, use the fact (you should see how you might prove this fact) that, $x^T Q y \leq \frac{1}{2}\left[x^T Q x + y^T Q y\right]$ for PSD Q (this is the matrix analogue of the simple fact that $a \times b \leq (a^2 + b^2)/2$), to verify that the desired inequality holds.

2.2 More Examples of Convex Functions

Here are a few examples of convex functions:

1. $\exp(ax)$ is convex for any a over $\mathbb{R}$.

2. $\log x$ is concave on $\mathbb{R}_{++}$.

3. $a^T x + b$ is convex (and concave).

4. The least squares loss $\|Ax - b\|^2$ is convex (for any A, b).

5. Any norm is convex, i.e. $\|x\|$ is a convex function.

6. The spectral norm, and the trace norm of a matrix are convex, i.e. $\|X\|_{\mathrm{op}} = \sigma_1(X)$, $\|X\|_{\mathrm{tr}} = \sum_{i=1}^d \sigma_i(X)$ where $\sigma_i(X)$ denotes the i-th singular value of X.

7. **Convex Indicators:** If C is a convex set, then the indicator function (which is defined on the extended reals):

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

is convex.

2.3 Convexity and Monotonicity

One nice property of convex functions is that their gradients are monotone.

In 1D this is a simple thing to interpret, a monotone function is order preserving. A function which is monotone *increasing* has the property that if $x \geq y$ then $f(x) \geq f(y)$. One way to write this mathematically is to say that for any x, y, $(x - y) \times (f(x) - f(y)) \geq 0$.

The (sub)gradient of a convex function satisfies a multivariate analogue of this property. Particularly for any $x, y \in \mathrm{dom}(f)$, if f is convex we have that for any $g_x \in \partial f(x)$ and $g_y \in \partial f(y)$,

$$(x - y)^T (g_x - g_y) \geq 0.$$

To see this we observe that by the first-order characterization:

$$f(y) \geq f(x) + g_x^T (y - x),$$
$$f(x) \geq f(y) + g_y^T (x - y),$$

and summing these inequalities gives our desired result.

It turns out that there is a converse to the above characterization. If you have a differentiable function whose gradient is monotone, then it must be convex. This idea will likely be useful in your HW for verifying some of the equivalences.

2.4 Properties of Convex Functions

Here are a few properties of convex functions that will be useful:

1. A function is convex iff the univariate functions $g(t) = f(x + tv)$ are convex for any $v \in \mathbb{R}^d$, and for any $x \in \text{dom}(f)$.

2. A function is convex iff its epigraph,
$$\text{epi}(f) = \{(x, t) \in \text{dom}(f) \times \mathbb{R} : f(x) \leq t\}$$
is a convex set.

3. Convex functions satisfy Jensen's inequality. If f is convex, then for any random variable X supported on $\text{dom}(f)$, $f(\mathbb{E}[X]) \leq \mathbb{E}f(X)$.

2.5 Operations which Preserve Convexity

1. **Non-negative Linear Combination:** Suppose $f_1, \ldots, f_m$ are convex, then so is $\sum_{i=1}^m a_i f_i$ for any $a_1, \ldots, a_m \geq 0$.

2. **Pointwise Max:** If the collection of functions f_s for $s \in S$ are convex, then so is $g(x) = \sup_{s \in S} f_s(x)$.

3. **Partial Minimization:** If $g(x, y)$ is a convex function, and C is a convex set, then $f(x) = \min_{y \in C} g(x, y)$ is a convex function.

An Example:

1. Suppose C is an arbitrary set, consider $f(x) = \max_{y \in C} \|x - y\|$. f is convex. To see this, we can view f as a maximum of convex functions $f_y(x) = \|x - y\|$.

2. Let C be a convex set, then $f(x) = \min_{y \in C} \|x - y\|$ is a convex function. We can view this as a partial minimization of the function $g(x, y) = \|x - y\|$ which is a convex function (in (x, y)).

Function compositions:

1. **Affine Composition:** If f is convex then so is $g(x) = f(Ax + b)$.

2. **General Composition:** Suppose that $f = h \circ g$, where $g : \mathbb{R}^d \mapsto \mathbb{R}$, $h : \mathbb{R} \mapsto \mathbb{R}$, $f : \mathbb{R}^d \mapsto \mathbb{R}$. Then one can ask when f is convex. There are many cases to cover (see BV) but we'll simply study one, and try to understand where it comes from: f is convex if h is convex and nondecreasing, g is convex.

   To see this: imagine everything was twice differentiable, then by the chain rule

   $$f''(x) = h''(g(x))(g'(x))^2 + h'(g(x))g''(x).$$

   When h is convex and non-decreasing, h'' and h' are positive, and when g is convex, g'' is positive, so f'' is positive.

2.6 Smooth, Strongly Convex and Strictly Convex Functions

For this section, we will switch back to thinking about differentiable convex functions.

2.6.1 Smoothness

In optimization smoothness has a very particular meaning (it has a slightly different meaning in stats, and other areas of math). A function f is β-smooth, if its gradient is Lipschitz continuous with parameter β, i.e. for any $x, y \in \text{dom}(f)$,

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|.$$

There are several useful implications of smoothness that you will show in your HW, but we will briefly discuss now:

1. If f is β-smooth then the function $\frac{\beta}{2}\|x\|^2 - f(x)$ is convex. Typically, we would not expect $-f(x)$ to be convex (except when f is affine).

2. Another implication of smoothness, is that it implies a quadratic upper bound on the function, i.e. if f is β-smooth then,

   $$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2}\|y - x\|^2.$$

   To interpret this fix a point x. Convex functions always lie *above* their tangent lines. *Smooth* convex functions always lie *below* a parabola which passes through the point $(x, f(x))$ (defined by the RHS above).

3. Finally, if f is twice differentiable, then β-smoothness is equivalent to the condition that,

$$\nabla^2 f(x) \preceq \beta I_d.$$

Examples: It is worth briefly considering two examples (canonical examples of non-smooth and smooth convex functions):

1. **Absolute value:** Here we consider $f(x) = |x|$, and observe that at $x = 0$, it's impossible to seat a parabola at the origin which is always above the function. Roughly, a parabola must have close to zero derivative near its minimum, but the absolute value function has constant derivative near its minimum.

2. **Quadratic function:** Suppose we consider $f(x) = x^T Q x + a^T x + b$ where $Q \succeq 0$. Its now easy to see that this function has Hessian $2Q$, and consequently it satisfies smoothness for any $\beta \geq 2\lambda_{\max}(Q)$ (i.e. twice the largest eigenvalue of Q).

2.6.2 Strong Convexity

The twin assumption to smoothness is strong convexity. A function f is α-strongly convex, if the function $g(x) = f(x) - \frac{\alpha}{2}\|x\|^2$ is convex. As with smoothness there are several important implications of strong convexity that you will explore in your HW.

1. If f is strongly convex then an equivalent definition is that it satisfies the following inequality for any $x, y \in \text{dom}(f)$,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\alpha}{2}\|y - x\|^2.$$

   Again to interpret this, fix a point x, and observe that this expression tells us that a strongly-convex function is *above* a parabola which passes through the point $(x, f(x))$.

2. If f is twice differentiable, an equivalent characterization is that,

$$\nabla^2 f(x) \succeq \alpha I_d.$$

Examples:

1. **Absolute value:** Consider the same function as before. It is not strongly convex. For instance, if we consider $x = 1, y = 2$, then $f(y) - (f(x) + \nabla f(x)^T (y - x))$ is 0, so the definition can only hold with $\alpha = 0$.

2. **Quadratic function:** Once again using the second-order characterization of strong convexity we see that the quadratic function satisfies the definition of strong convexity for any $\alpha \leq 2\lambda_{\min}(Q)$.

It is possible to have strongly convex functions which are not smooth and vice versa, and it is worth trying to "draw" some examples to convince yourself of this.

2.6.3 Strict Convexity

Strict convexity is a "weakening" of strong convexity (we won't use it so much in this course but it's a useful concept to be aware of). A function f is *strictly* convex if either:

1. $f(\theta x + (1-\theta)y) < \theta f(x) + (1-\theta)f(y)$ for $0 < \theta < 1$.
2. $f(y) > f(x) + \nabla f(x)^T(y-x)$, for any $x \neq y$.

It is worth noting the second-order characterization doesn't work in the expected way, i.e. you can have twice-differentiable, strictly convex functions which don't satisfy the condition that $\nabla^2 f(x) \succ 0$. (As an example, think about the function x^4 at $x = 0$.)

2.7 Optimality Conditions

Here we will revisit some things we discussed briefly in the previous lecture. Here is the basic question. We are interested in solving a problem:

$$\min_{x \in C} f(x),$$

where f is a convex function, and C is a convex set. What can I say about a solution x^* to this problem?

1. **Unconstrained Case:** Suppose first that $C = \mathbb{R}^d$, and that $\mathrm{dom}(f) = \mathbb{R}^d$ then our characterization should be familiar to us from usual calculus classes.

   Theorem 2.1 *x^* is optimal, if (and only if) $0 \in \partial f(x^*)$.*

   Proof: If $0 \in \partial f(x^*)$, then from the first-order condition we know that,

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

Conversely, if x^* is optimal, then we know that, $f(y) \geq f(x^*) + g_{x^*}^T(y - x^*)$ for all y, when $g_{x^*} = 0$ and so we know that 0 is valid subgradient at x^*.

Notice an interesting aspect of this result – it does not require convexity, i.e. for *any* function f the condition that $0 \in \partial f(x^*)$ is necessary and sufficient for x^* to be a minimizer. ∎

2. **Constrained, Differentiable Case:** A feasible point x^* is optimal, if and only if $\nabla f(x^*)^T(y - x^*) \geq 0$ for all $y \in C$.

   We will only verify one direction of this (the other direction requires a bit of analysis to check). Suppose that, $\nabla f(x^*)^T(y - x^*) \geq 0$ for all $y \in C$, then from the first-order condition we have that,

   $$f(y) \geq f(x^*) + \nabla f(x^*)^T(y - x^*) \geq f(x^*),$$

   so x^* is optimal. If you recall the definition of the normal cone from last lecture, then you will see that this condition says that,

   $$-\nabla f(x^*) \in N_C(x^*).$$

3. **General, Constrained Case:** A feasible point x^* is optimal, if and only if $0 \in \partial f(x^*) + N_C(x^*)$. Here we are adding two sets, i.e. $C + D = \{y : y = u + v, u \in C, v \in D\}$.

   Again it's only easy to verify one direction of this, i.e. suppose that $0 \in \partial f(x^*) + N_C(x^*)$, this means that there are two vectors $u \in \partial f(x^*)$ and $v \in N_C(x^*)$ such that,

   $$u + v = 0.$$

   Now, we know that for any y which is feasible,

   $$\begin{aligned} f(y) &\geq f(x^*) + u^T(y - x^*) \\ &= f(x^*) - v^T(y - x^*). \end{aligned}$$

   Since $v \in N_C(x^*)$ we know that $v^T(y - x^*) \leq 0$ for every feasible y, and so we conclude that $f(y) \geq f(x^*)$.

2.7.1 Optimality Conditions for Projection

Here is a very basic/important problem. It arises in signal processing and statistics as a basic denoising scheme. For some convex set K, and observation y we would like to solve the constrained minimization problem,

$$\min_{x \in K} \frac{1}{2}\|y - x\|^2.$$

This finds the closest point in K to y, and is called the *projection* of y onto K. We will denote the solution x^* to the above program by $P_K(y)$.

Let us first write out the optimality conditions, and then use them to show a nice property of this projection operation. Since f is differentiable we have that,

$$0 \in x^* - y + N_C(x^*).$$

Equivalently, this means that, $(y - x^*)^T(a - x^*) \leq 0$ for all $a \in K$. This can be easily understood with a picture.

Theorem 2.2 *Projection onto a convex set is a contraction, i.e. for any pair of points* a, b,

$$\|P_K(a) - P_K(b)\| \leq \|a - b\|.$$

Proof: From the optimality conditions we have that for any $x \in K$,

$$(a - P_K(a))^T(x - P_K(a)) \leq 0$$
$$(b - P_K(b))^T(x - P_K(b)) \leq 0.$$

As a consequence we can see that,

$$(a - P_K(a))^T(P_K(b) - P_K(a)) \leq 0$$
$$(b - P_K(b))^T(P_K(a) - P_K(b)) \leq 0.$$

Adding these inequalities we obtain that,

$$(b - a + (P_K(a) - P_K(b)))^T(P_K(a) - P_K(b)) \leq 0.$$

Now, re-arranging and applying the Cauchy-Schwarz inequality, we see that,

$$\|P_K(a) - P_K(b)\|^2 \leq (a - b)^T(P_K(a) - P_K(b)) \leq \|a - b\|\|P_K(a) - P_K(b)\|,$$

which is our desired conclusion.

■