10-425/625: Introduction to Convex Optimization (Fall 2023)

Lecture 5: Properties of Convex Functions

Instructor:[1] *Matt Gormley* *September 13, 2023*

5.1 Properties of Convex Functions

5.1.1 Convexity and Monotonicity

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

5.1.1.1 Monotonicity in 1D

Definition 5.1 (Monotone Increasing Function)**.** *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,

$$(x - y) \times (f(x) - f(y)) \geq 0.$$

5.1.1.2 Monotonicity of Gradients

For a differentiable convex function f, the multivariate analogue is that for any $x, y \in \text{dom}(f)$:

$$(x - y)^T(\nabla f(x) - \nabla f(y)) \geq 0.$$

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

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), (y - x)\rangle \\ f(x) &\geq f(y) + \langle \nabla f(y), (x - y)\rangle, \end{aligned}$$

and summing these inequalities gives our desired result: $\langle x - y, \nabla f(x) - \nabla f(y)\rangle \geq 0$. ■

[1] These notes were originally written by Siva Balakrishnan for 10-725 Spring 2023 (original version: here) and were edited and adapted for 10-425/625.

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

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

Proof: 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: $(x-y)^T(g_x-g_y) \geq 0$.
■

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.

5.1.2 Other Properties

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

1. A function is convex iff its epigraph,

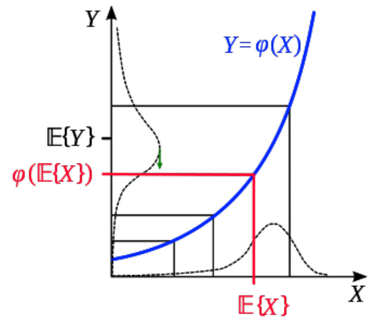
   $$\text{epi}(f) = \{(x, t) \in \text{dom}(f) \times \mathbb{R} : f(x) \leq t\}$$

   is a convex set. *In simpler terms:* if you take all the points that lie above a function, those form a convex set.

   Interesting note: There is an connection here between the supporting hyperplane of this epigraph (set) and subgradients, most easily shown with a picture.

2. 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 $x \in \text{dom}(f)$.

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



5.2 Smooth, Strongly Convex and Strictly Convex Functions

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

5.2.1 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 $x = 0$.)

For a strictly convex function, we are guaranteed that its minimizer is *unique* if it exists. That is, a strictly convex function has at most one local minimum.

Background: (Continuous, Lipschitz continuous)

Definition 5.2 (Continuous Function). *A function* $f : \mathbb{R}^n \to \mathbb{R}$ *is* continuous at a point $y \in \mathrm{dom}(f)$ *iff*

$$f(y) \text{ exists}$$
$$\lim_{x \to y} f(x) \text{ exists}$$
$$\lim_{x \to y} f(x) = f(y).$$

Intuitively, this means the function consists of one curve without any breaks over the reals. If a function $f : \mathbb{R}^n \to \mathbb{R}$ is continuous over all points $y \in \mathbb{R}^n$, then it is a continuous function.

Definition 5.3 (Continuously Differentiable)**.** *A function $f : \mathbb{R}^n \to \mathbb{R}$ is* continuously differentiable *if its gradient $\nabla f(x)$ exists and each of its partial derivatives $\frac{\partial f(x)}{\partial x_i}$ is a continuous function at all points x.*

Definition 5.4 (Lipschitz Continuous)**.** *A 1D function $f : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous if there exists a constant $L \in \mathbb{R}$ such that for all $x, y \in \mathbb{R}$:*

$$|f(x) - f(y)| \leq L|x - y|$$

That is, the difference of the rate of change of the function from the beginning to the end of some interval is bounded by a constant factor of the interval size, for all size intervals.

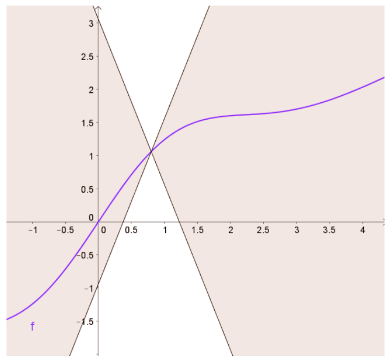
More generally, a function $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz continuous *if there exists a constant $L \in \mathbb{R}$ such that for all $x, y \in \mathbb{R}$:*

$$\|f(x) - f(y)\| \leq L\|x - y\|$$

The above holds for any norm $\|\cdot\|$, but we can assume we're working with the ℓ-2 norm. We say that such a function is L-Lipschitz.

We can understand Lipschitz continuity at a point x geometrically, by considering two cones: an upper cone and a lower cone sitting above and below $f(x)$ at x respecitively. The two cones are defined by all lines whose slope obeys $\frac{\|f(x)-f(y)\|}{\|x-y\|} \leq L$. Roughly, if for all x the function never enters the upper or lower cones then it must be L-Lipschitz.

A smaller Lipschitz constant L means a wider pair of cones, indicating slower growth or change. Conversely, a larger Lipschitz constant L means steeper cones, indicating faster growth or change.



Any function that is Lipschitz continuous is also continuous.

Example 5.5. Some examples of functions that are *not* Lipschitz continuous are those that grow very rapidly, such as $f(x) = \exp(x)$ and $f(x) = x^2$, both of which become arbitrarily steep as $x \to \infty$.

5.2.2 Smoothness

In optimization smoothness has a very particular meaning (it has a slightly different meaning in stats, and other areas of math).

Definition 5.6 (β-Smooth)**.** *A function f is β-smooth, if its gradient is Lipschitz continuous with parameter β, i.e. for any $x, y \in dom(f)$,*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta \|x - y\|_2.$$

There are several useful implications of smoothness that we will briefly discuss now:

1. 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 (i.e. $f(y) \geq f(x) + \nabla f(x)^T (y - x)$). *Smooth* convex functions always lie *below* a parabola which passes through the point $(x, f(x))$ (defined by the RHS above).

2. Suppose x^* is a minimum of a β-smooth function f, then for all $y \in \text{dom}(f)$

$$\|\nabla f(y)\|_2 \leq \beta\|y - x\|_2$$

That is, if we are at a point y that is close to the minimum x^*, then the gradient at y, $\nabla f(y)$ must also be small. So any algorithm we have that follows the gradients of the functions should intuitively slow down as it approaches the minimum.

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

$$0 \preceq \nabla^2 f(x) \preceq \beta I_d.$$

where the lower bound $0 \preceq$ comes from convexity of f and the upper bound $\preceq \beta I_d$ comes from β-smoothness of f.

4. 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).

Segue... Next time we'll pick up with some examples of β-smooth functions and then look at strong convexity.