Convex Functions: Properties, Examples and Gradient Descent

Orestis Plevrakis

1 Properties of Convex Functions

In the last lecture, we defined convex functions. The following two theorems give some equivalent definitions.

Theorem 1. *Let K be a convex subset of $\mathbb{R}^n$ and $f : K \to \mathbb{R}$. Then, f is convex if and only if all its restrictions on lines are convex, i.e., for all $x_0 \in K$, $v \in \mathbb{R}^n$, we have that the univariate function $g(t) = f(x_0 + tv)$, with domain $dom(g) = \{t \in \mathbb{R} \mid x_0 + tv \in K\}$, is convex.*

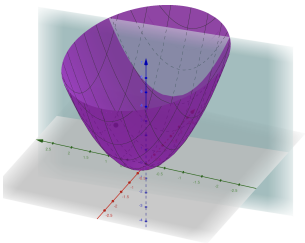


Figure 1: Example of restriction on a line of a convex function.

The proof follows immediately from the definition.

Proof. First of all, we observe that for any $x_0 \in K$, $v \in \mathbb{R}^n$, the set $\{t \in \mathbb{R} \mid x_0 + tv \in K\}$ is a convex subset of $\mathbb{R}$ (why?), and so it is an interval[1].

Suppose f is convex, and let $x_0 \in K$, $v \in \mathbb{R}^n$, g the corresponding restriction of f. Then, for any $t_1, t_2 \in \text{dom}(g)$ and $\theta \in [0, 1]$, we have

$$\begin{aligned} g((1-\theta)t_1 + \theta t_2) = f\left(x_0 + ((1-\theta)t_1 + \theta t_2)\, v\right) &= f\left((1-\theta)(x_0 + t_1 v) + \theta(x_0 + t_2 v)\right) \\ &\leq (1-\theta) f(x_0 + t_1 v) + \theta f(x_0 + t_2 v) \\ &= (1-\theta) g(t_1) + \theta g(t_2) \end{aligned}$$

Suppose all the restrictions of f on lines are convex. Let $x, y \in K$, $\theta \in [0, 1]$. Set $x_0 = x$, $v = y - x$, and let $g(t) = f(x + t(y - x))$. Then,

$$f((1-\theta)x + \theta y) = g(\theta) = g((1-\theta)\cdot 0 + \theta \cdot 1) \leq (1-\theta) g(0) + \theta g(1) = (1-\theta) f(x) + \theta f(y)$$

□

Theorem 2. *Let K be an open convex subset of $\mathbb{R}^n$, and $f : K \to \mathbb{R}$.*

[1] Note that $\mathbb{R}$ and singleton-sets are considered to be intervals.

- *If f is C^1, then f is convex if and only if for all $x, x_0 \in K$, $f(x) \geq f(x_0) + \nabla f(x_0) \cdot (x - x_0)$.*
- *If f is C^2, then f is convex if and only if for all $x \in K$, $\nabla^2 f(x) \succcurlyeq 0$.*

Observe that for $n = 1$, the theorem is well-known: the first condition says that the tangent lines are below the graph, and the second says that the second derivative is nonnegative. To prove Theorem 2, we will take for granted that it holds for $n = 1$.[2] The idea of the proof is to use Theorem 1 to reduce the problem in one dimension, where we know that the theorem holds.

Proof. Since K is convex, the domain of any restriction of f on a line is an interval. Since K is also open, this interval will be open (why?).

- Suppose f is convex. Let $x_0, x \in K$. Consider the function $g(t) = f(x_0 + t(x - x_0))$, with $\text{dom}(g) = \{t \in \mathbb{R} \mid x_0 + t(x - x_0) \in K\}$. From Theorem 1, g is convex, and so

  $$g(1) \geq g(0) + g'(0)(1 - 0)$$

  which gives that $f(x) \geq f(x_0) + \nabla f(x_0) \cdot (x - x_0)$. Now, suppose that for all $x, x_0 \in K$, $f(x) \geq f(x_0) + \nabla f(x_0) \cdot (x - x_0)$. Let $z_0 \in K, v \in \mathbb{R}^n$. We will show that the function $g(t) = f(z_0 + tv)$, with domain $\text{dom}(g) = \{t \in \mathbb{R} \mid z_0 + tv \in K\}$, is convex. It suffices to prove that for any $t_0, t \in \text{dom}(g)$,

  $$g(t) \geq g(t_0) + g'(t_0)(t - t_0)$$

  and by substituting:

  $$f(z_0 + tv) \geq f(z_0 + t_0 v) + \nabla f(z_0 + t_0 v) \cdot v \ (t - t_0)$$

  which is true. Since z_0, v where chosen arbitrarily, Theorem 1 implies that f is convex.

- Suppose f is convex. Let $x_0 \in K, v \in \mathbb{R}^n$, and consider the restriction of f on the induced line: $g(t) = f(x_0 + tv)$. Then, g is convex, and so $0 \leq g''(0) = v^\top \nabla^2 f(x_0) v$. Since v was chosen arbitrarily, we get $\nabla^2 f(x_0) \succcurlyeq 0$. Now, suppose that for all $x \in K$, we have that $\nabla^2 f(x) \succcurlyeq 0$. Let $x_0 \in K, v \in \mathbb{R}^n$. Then, for any t such that $x_0 + tv \in K$, we have $\frac{d^2}{dt^2} f(x_0 + tv) = v^\top \nabla^2 f(x_0 + tv) v \geq 0$

□

2 The simplest convex functions

We revisit the simple multivariate functions from the previous lecture, to check whether they are convex. First of all, the affine functions $f(x) = c^\top x + b$ are convex, since $\nabla^2 f(x) = 0 \succcurlyeq 0$.

[2] If you are curious to see a proof that includes the $n = 1$ case, check these notes.

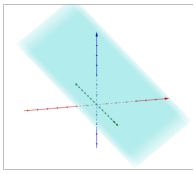
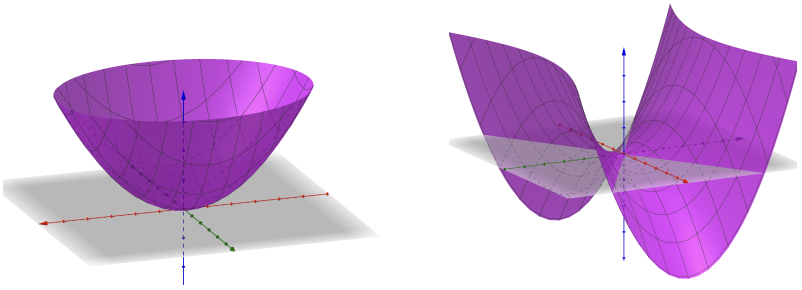


Figure 2: The graphs of two-dimensional affine functions are planes.

Let's look at the quadratic functions now: $f(x) = x^\top Ax + c^\top x + b$ (remember that here A is symmetric). In the previous lecture we also saw that $\nabla^2 f(x) = 2A$, and so f is convex if and only if $A \succcurlyeq 0$. To get a sense of how convex and non-convex quadratic functions look like, consider the two-dimensional quadratic: $f(x_1, x_2) = \lambda_1 x_1^2 + \lambda_2 x_2^2$, which corresponds to diagonal A, $c = 0$ and $b = 0$. Observe that f is convex if and only if $\lambda_1, \lambda_2 \geq 0$. Here are figures of convex and non-convex versions of f:



(a) $0 < \lambda_1 < \lambda_2$ (b) $\lambda_1 < 0 < \lambda_2$

Figure 3: A convex and a non-convex example. The x_1 axis is colored in red.

3 Operations that preserve convexity

The following theorem often enables us to quickly show that a function is convex.

Theorem 3. *Let $f, g : K \to \mathbb{R}$ convex functions. Then,*

- *$f + g$ is convex.*
- *For any $\alpha \geq 0$, αf is convex.*
- *For any $A \in \mathbb{R}^{n \times m}, b \in \mathbb{R}^n$, the function $g(x) = f(Ax + b)$, defined on every x such that $Ax + b \in K$, is convex.*

Notice that the third part says that the composition of an affine function with a convex function is convex. Caution: it is not true in general that the composition of two convex functions is convex.

Proof. They all follow immediately from the definition. We prove the third. First of all, the domain of g is convex (why?). Let $x, y \in \mathrm{dom}(g)$ and $\theta \in [0, 1]$. We have

$$\begin{aligned} g((1-\theta)x + \theta y) = f(A((1-\theta)x + \theta y) + b) &= f((1-\theta)(Ax+b) + \theta(Ay+b)) \\ &\leq (1-\theta) f(Ax+b) + \theta f(Ay+b) \end{aligned}$$

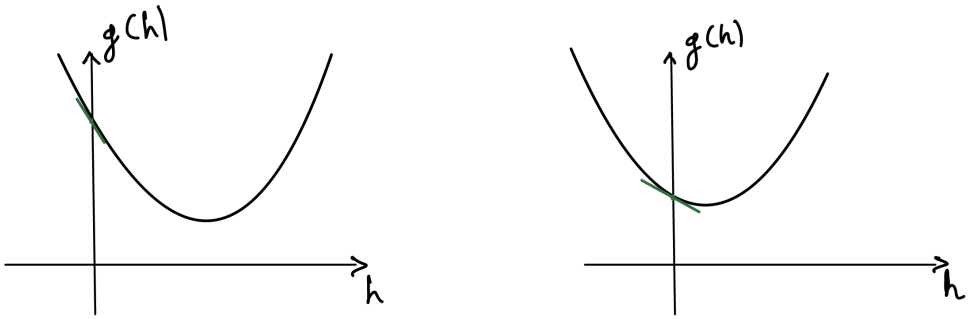
□

4 Gradient Descent

A natural iterative algorithm for minimizing a (not necessarily convex) function f is the following: in each step, choose the direction at which f decreases at a maximum rate. Then, make a small step in that direction, and iterate. It is easy to see that for non-convex functions such an algorithm can easily get stuck around local minima. So, let's suppose that f is convex. How do we find the steepest direction? Suppose that at step t, we are at point x_t, and let v be a unit vector. The directional derivative along v is $\frac{d}{dh} f(x_t + hv)\big|_{h=0} = \nabla f(x_t) \cdot v$, and if $\nabla f(x_t) \neq 0$, the derivative is minimized at $v_t := -\frac{\nabla f(x_t)}{\|\nabla f(x_t)\|}$, with minimum value $\frac{d}{dh} f(x_t + hv_t)\big|_{h=0} = -\|\nabla f(x_t)\|$. Thus, the iteration is

$$x_{t+1} = x_t + \eta_t v_t$$

where $\eta_t > 0$ is the stepsize. How to choose η_t? From Theorem 1, $g(h) = f(x_t + hv_t)$ is convex, and the following pictures suggest that η_t should be larger, for larger $|g'(0)|$. Thus, a reasonable choice



is to set η_t proportional to $|g'(0)|$, which equals $\|\nabla f(x_t)\|$. So, we choose $\eta_t = \eta\|\nabla f(x_t)\|$, for some constant $\eta > 0$ independent of time. The resulting iteration is

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

and the algorithm is called Gradient Descent (GD).[3]

4.1 Smoothness

We want GD to decrease the value of the function in each iteration. This is not possible for every convex function. Consider the one-dimensional example of $f(x) = |x|$. No matter how small η is, if we are close enough to zero, the next step will increase the function. The problem here is not the non-differentiability, as we can construct a "smoothed" $|x|$: The problem is that the derivative changes too quickly, and this manifests in the second derivative of the smoothed version, which will be very large for x close to zero. To exclude these problematic cases, we define as *β-smooth* those univariate twice-differentiable convex functions $f(x)$, that satisfy $f''(x) \leq \beta$ in their whole domain. For multivariate functions:

Definition 4. Let $f : K \to \mathbb{R}$ a C^2 convex function, and let $\beta > 0$. We say that f is β-smooth, if all its second directional derivatives along unit vectors are at most β, i.e., for all $x_0 \in K$ and unit vectors $v \in \mathbb{R}^n$, we have $v^\top \nabla^2 f(x_0) v \leq \beta$.

[3] There are variations of GD where η actually decreases with t, but we will not get into those.

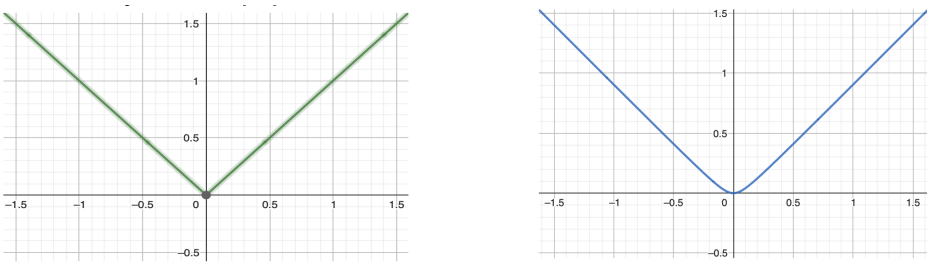


Figure 4: $f(x) = |x|$ and its smoothed version.

In Problem 1 of the second homework, we saw that if A is symmetric, then its maximum eigenvalue is given by the formula $\lambda_{\max}(A) = \max_{\|v\|=1} v^\top Av$. Thus, a C^2 convex function f is β-smooth if and only if for every x_0 in its domain, we have $\lambda_{\max}(\nabla^2 f(x_0)) \leq \beta$. The following theorem gives an equivalent definition.

Theorem 5. *Let $f : K \to \mathbb{R}$ be C^2 and convex. Let $\beta > 0$. Then, f is β-smooth if any only if for any $x, y \in K$, $\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$.*

We will not prove this theorem, but you can freely use it in the problems if you need to.

4.2 Taylor's theorem

If a convex function is β-smooth, we can bound how much its graph deviates locally from its tangent hyperplane. This can be deduced from Taylor's theorem:

Theorem 6. *Let $f : K \to \mathbb{R}$ a C^2 function, and let $x_0 \in K$. Then, for every $x \in K$ such that the line segment $[x_0, x] \subseteq K$, we have*

$$f(x) = f(x_0) + \nabla f(x_0) \cdot (x - x_0) + \frac{1}{2}(x - x_0)^\top \nabla^2 f(\xi)(x - x_0)$$

for some $\xi \in [x_0, x]$.

Observe that in the above theorem, $\frac{1}{2}(x - x_0)^\top \nabla^2 f(\xi)(x - x_0)$ is the error for approximating $f(x)$ with its linearization at x_0. For β-smooth functions, the error is

$$\frac{1}{2}\|x - x_0\|^2 \left(\frac{x - x_0}{\|x - x_0\|}\right)^\top \nabla^2 f(\xi) \left(\frac{x - x_0}{\|x - x_0\|}\right) \leq \frac{\beta}{2}\|x - x_0\|^2 \tag{1}$$

4.3 Convergence of Gradient Descent

In this lecture and in the beginning of the next, we will prove this theorem:

Theorem 7. *Let $f : \mathbb{R}^n \to \mathbb{R}$ a C^2, convex, β-smooth function, which attains its minimum at a $x_* \in \mathbb{R}^n$. Suppose we run Gradient Descent on f with $\eta = 1/\beta$, starting from a point $x_1 \in \mathbb{R}^n$. Then, for any $\epsilon > 0$, there exists a $t_0 = O(\beta\|x_1 - x_*\|^2/\epsilon)$, such that for all $t \geq t_0$, we have $f(x_t) \leq f(x_*) + \epsilon$.*

The iterations of Gradient Descent are given by $x_{t+1} = x_t - \frac{1}{\beta}\nabla f(x_t)$. The proof starts by applying Theorem 6 together with inequality 1, to prove that $f(x_t)$ is decreasing:

$$f(x_{t+1}) \leq f(x_t) + \nabla f(x_t) \cdot (x_{t+1} - x_t) + \frac{\beta}{2}\|x_{t+1} - x_t\|^2$$

Since $x_{t+1} - x_t = -\frac{1}{\beta}\nabla f(x_t)$, we get

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2\beta}\|\nabla f(x_t)\|^2$$