Section 04: Solutions

1. Convexity

Convexity is defined for both sets and functions. For today we'll focus on discussing the convexity of functions.

Definition 1 (Convex functions)**.** *A function $f\colon \mathbb{R}^d \to \mathbb{R}$ is* ***convex*** *on a set $A \subseteq \mathbb{R}^d$ if for all $x, y \in A$ and $\lambda \in [0,1] \subset \mathbb{R}$:*

$$f\left(\lambda x + (1-\lambda)y\right) \leq \lambda f(x) + (1-\lambda)f(y)\ .$$

When this definition holds with the inequality being reversed, then f is said to be *concave*. From the definition, it is clear that a function f is convex if and only if $-f$ is concave.

(a) Why do we care whether a function is convex or not?

Solution:

Convex functions are useful because their local minima are always global minimum. Many numerical methods or algorithms find a local minima while in machine learning we are typically interested in finding the global minimum. To see why local minima of a convex function is a global minima: let x^* be a local minimizer for a convex function f, and suppose there exists a point $x_0 \neq x^*$ such that $f(x_0) < f(x^*)$. Now note that because f is convex, there exists some $\lambda \in (0,1)$ such that for $y := \lambda x^* + (1-\lambda)x_0$ and $f(y) \geq f(x^*)$ (i.e., y is sufficiently close to x^*). We now have a contradiction:

$$f(y) \leq \lambda f(x^*) + (1-\lambda)f(x_0) < f(x^*) \leq f(y)\ .$$

In words, a line segment between any arbitrary point x_0 and a local minimizer x^* should be entirely above the function by definition of convexity, ensuring that $f(x_0) < f(x^*)$ cannot happen.

(b) Which of the following functions are convex? (Hint: draw a picture)
(i) $x \mapsto |x|$ on $\mathbb{R}$, (ii) $x \mapsto \cos(x)$ on $\mathbb{R}$, (iii) $x \mapsto x^\top x$ on $\mathbb{R}^d$ for any $d \in \mathbb{N}$.

Solution:

The functions $x \mapsto |x|$ and $x \mapsto x^\top x$ are both convex on their entire domain. The function $x \mapsto \cos(x)$ is not convex on $\mathbb{R}$ since we can draw a line at two points (from say $\frac{\pi}{2}$ to $2\pi + \frac{\pi}{2}$) that is not entirely above the function.

Proof that $x \mapsto |x|$ is convex on $\mathbb{R}$:

$$f\left(\lambda x + (1-\lambda)y\right) = |\lambda x + (1-\lambda)y| \leq \lambda|x| + (1-\lambda)|y|\ .$$

Proof that $x \mapsto x^\top x$ is convex on $\mathbb{R}^d$ for any $d \in \mathbb{N}$:
We begin by examining the definition: whenever $\lambda \in [0,1]$, we have

$$\begin{aligned}
\left(\lambda x + (1-\lambda)y\right)^\top \left(\lambda x + (1-\lambda)y\right) &= \lambda^2 x^\top x + (1-\lambda)^2 y^\top y + 2\lambda(1-\lambda)x^\top y \\
&= \lambda(1-(1-\lambda))x^\top x + (1-\lambda)(1-\lambda)y^\top y + 2\lambda(1-\lambda)x^\top y \\
&= \lambda x^\top x + (1-\lambda)y^\top y - \lambda(1-\lambda)\left(x^\top x - 2x^\top y + y^\top y\right) \\
&= \lambda x^\top x + (1-\lambda)y^\top y - \lambda(1-\lambda)(x-y)^\top(x-y) \\
&\leq \lambda x^\top x + (1-\lambda)y^\top y,
\end{aligned}$$

where the inequality holds because $(x-y)^\top(x-y) = \|x-y\|_2^2 \geq 0$. So our function is convex. Note the in problem 2.c we show that sum of convex functions is convex, and since $x^\top x = \sum_{i=1}^{d} x_i^2$ (Problem 1.b.iii) we can use this property here by noting that each summand is a convex function.

(c) Can a function be both convex and concave on the same set? If so, give an example. If not, describe why not.

Solution:

Affine functions (i.e. functions such that $f(\lambda x + (1-\lambda)y) = \lambda f(x) + (1-\lambda)f(y)$) are both convex and concave.

2. Other Methods for Checking Convexity

Using the definition to check whether a function is convex or not can be a tedious task in many situations. Some basic methods that can help us achieve the task in an efficient way are introduced below:

- For a differentiable function f, examine $f(y) \geq f(x) + \nabla f(x)^\top (y-x)$ for any x, y in the domain of f.
- For twice differentiable functions f, examine $\nabla^2 f(x) \succeq 0$ (i.e., the Hessian is positive semi-definite in the domain of f).
- Non-negative weighted sum of convex functions is again convex.
  If for some $n \in \mathbb{N}$, $f_1, f_2, \ldots, f_n$ are convex functions on a set, then for all non-negative scalars $\alpha_i \geq 0$ for $i \in \{1, 2, \ldots, n\}$, $\sum_{i=1}^n \alpha_i f_i$ is also convex.
- Composition with affine function perserves convexity.
  If f is a convex function, and g is an affine function, then $f \circ g$ is convex. For example if $g(x) = Ax + b$ for some matrix A and vector b, then $x \mapsto (f \circ g)(x) = f(g(x)) = f(Ax+b)$ is also convex.
- Point-wise maximum and supremum.
  If $f\colon (x,y) \mapsto f(x,y)$ is convex in x for each y, then $x \mapsto g(x) := \sup_y f(x,y)$ is convex.

Note: there are even more such methods, which are covered in a convex optimization course or textbook.

(a) If f is differentiable, then f is convex if and only if $f(y) \geq f(x) + \nabla f(x)^\top (y-x)$ for any x, y in the domain of f. A geometric interpretation of this characterization is that any tangent hyperplane of a convex function f must lie entirely below f. One interesting application of this characterization is one of the most important inequalities in probability and statistics: the Jensen's inequality. Show that if X is a random variable, then $\mathbb{E}f(X) \geq f(\mathbb{E}(X))$ when f is convex.

Solution:

Let $\mu = \mathbb{E}(X)$, then since f is convex, we have

$$f(X) \geq f(\mu) + \nabla f(\mu)^\top (X - \mu)$$

with probability 1. This means that taking expectation on both sides preserves the inequality: $\mathbb{E}f(X) \geq f(\mu) = f(\mathbb{E}X)$.

(b) If f is twice differentiable with convex domain, then f is convex if and only if

$$\nabla^2 f(x) \succeq 0 \, ,$$

for any x in the domain of f. Use this method to show that the objective function in linear regression is convex.

Solution:

Let $f(w) = (y - Xw)^\top (y - Xw)$, then

$$\nabla^2 f(w) = 2X^\top X \ ,$$

which is clearly positive semi-definite.

(c) Suppose f is convex, then $g(x) := f(Ax + b)$ is convex. Use this method to show that $\|Ax + b\|_1$ is convex (in x), where $\|z\|_1 = \sum_i |z_i|$.

Solution:

With this method, we only need to show the convexity of $\|x\|_1$. This is true from definition by observing that

$$\|\lambda x + (1-\lambda)y\|_1 = \sum_i |\lambda x_i + (1-\lambda)y_i| \le \sum_i \lambda |x_i| + (1-\lambda)|y_i| = \lambda \|x\|_1 + (1-\lambda)\|y\|_1 \ ,$$

where the inequality holds because of triangular inequality for the absolute value function.

(d) Suppose you know that f_1 and f_2 are convex functions on a set A. The function $x \mapsto g(x) := \max\{f_1(x), f_2(x)\}$ is also convex on A. In general, if $f\colon (x, y) \mapsto f(x, y)$ is convex in x for each y, then $x \mapsto g(x) := \sup_y f(x, y)$ is convex. Use this method to show that the largest eigenvalue of a matrix X, $\lambda_{\max}(X)$, is convex in X (Using the definition of convexity would make this question quite difficult).

Solution:

Consider $f(v, X) := v^\top X v$, then for each v, we have

$$f(v, \lambda X + (1-\lambda)Y) = \lambda f(v, X) + (1-\lambda) f(v, Y) \ ,$$

suggesting that $f\colon (v, X) \mapsto f(v, X)$ is convex in X for each v. Then $g(X) := \lambda_{\max}(X) = \sup_{\|v\|_2 = 1} f(v, X)$ is convex in X using this method.

(e) Does the same result hold for $h(x) := \min\{f_1(x), f_2(x)\}$? If so, give a proof. If not, provide convex functions f_1, f_2 such that h is not convex.

Solution:

No, consider $f_1(x) = x^2$, $f_2(x) = (x-1)^2$. Then $h(0) = h(1) = 0$, but $h(0.5) = 0.25$, so $h(0.5 \cdot 0 + 0.5 \cdot 1) = 0.25 > 0 = 0.5 \cdot h(0) + 0.5 \cdot h(1)$. So the minimum of two convex functions is not convex in general.

3. Gradient Descent

We will now examine gradient descent algorithm and study the effect of learning rate $\alpha \ge 0$ on the convergence of the algorithm. Recall from lecture that Gradient Descent takes on the form of $x_{t+1} = x_t - \alpha \nabla f(x_t)$.

(a) Assume that $f\colon \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, and additionally,

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\| \ \text{ for any } x, y \in \mathbb{R}^n \ ,$$

i.e., ∇f is Lipschitz continuous with constant $L > 0$
Show that: Gradient descent with fixed step size $\eta \le \frac{1}{L}$ satisfies

$$f(x^{(k)}) - f(x^*) \le \frac{\|x^{(0)} - x^*\|^2}{2\eta k} \ ,$$

i.e., gradient descent has convergence rate $O\left(\frac{1}{k}\right)$.
Hints:

(i) ∇f is Lipschitz continuous with constant $L > 0 \implies f(y) \leq f(x) + \nabla f(x)(y-x) + \frac{L}{2}\|y-x\|^2$ for all x, y.

(ii) f is convex $\implies f(x) \leq f(x^*) + \nabla f(x)^\top (x - x^*)$, where x^* is the local minima that the gradient descent algorithm is converging to.

(iii) $2\eta\nabla f(x)^\top(x-x^*) - \eta^2\|\nabla f(x)\|^2 = \|x-x^*\|^2 - \|x - \eta\nabla f(x) - x^*\|^2$

Solution:

Proof:
For any positive integer k, $x^{(k)} = x^{(k-1)} - \eta\nabla f\left(x^{(k-1)}\right)$, according to the gradient descent algorithm.
By hint (1), we have

$$\begin{aligned}
f\left(x^{(k)}\right) &\leq f\left(x^{(k-1)}\right) + \nabla f\left(x^{(k-1)}\right)^\top \left(x^{(k)} - x^{(k-1)}\right) + \frac{L}{2}\left\|x^{(k)} - x^{(k-1)}\right\|^2 \\
&= f\left(x^{(k-1)}\right) - \eta\left\|\nabla f\left(x^{(k-1)}\right)\right\|^2 + \frac{L}{2}\eta^2\left\|\nabla f\left(x^{(k-1)}\right)\right\|^2 \\
&\leq f\left(x^{(k-1)}\right) + \left(-\eta + \frac{\eta}{2}\right)\left\|\nabla f\left(x^{(k-1)}\right)\right\|^2 \quad (\because \eta \leq L^{-1}) \\
&= f\left(x^{(k-1)}\right) - \frac{\eta}{2}\left\|\nabla f\left(x^{(k-1)}\right)\right\|^2 \\
&\leq f\left(x^*\right) + \nabla f\left(x^{(k-1)}\right)^\top \left(x^{(k-1)} - x^*\right) - \frac{\eta}{2}\left\|\nabla f\left(x^{(k-1)}\right)\right\|^2 \quad \text{(By hint (2))} \\
&= f\left(x^*\right) + \frac{1}{2\eta}\left(2\eta\nabla f\left(x^{(k-1)}\right)^\top\left(x^{(k-1)} - x^*\right) - \eta^2\left\|\nabla f\left(x^{(k-1)}\right)\right\|^2\right) \\
&\leq f\left(x^*\right) + \frac{1}{2\eta}\left(\left\|x^{(k-1)} - x^*\right\|^2 - \left\|x^{(k-1)} - \eta\nabla f(x^{(k-1)}) - x^*\right\|^2\right) \quad \text{(By hint(3))} \\
&= f\left(x^*\right) + \frac{1}{2\eta}\left(\left\|x^{(k-1)} - x^*\right\|^2 - \left\|x^{(k)} - x^*\right\|^2\right).
\end{aligned}$$

Hence, we have

$$f\left(x^{(k)}\right) - f\left(x^*\right) \leq \frac{1}{2\eta}\left(\left\|x^{(k-1)} - x^*\right\|^2 - \left\|x^{(k)} - x^*\right\|^2\right).$$

Adding up from 1 to k, we obtain

$$\begin{aligned}
\sum_{i=1}^{k}\left(f(x^{(i)}) - f(x^*)\right) &\leq \frac{1}{2\eta}\sum_{i=1}^{k}\left(\left\|x^{(i-1)} - x^*\right\|^2 - \left\|x^{(i)} - x^*\right\|^2\right) \\
\implies \sum_{i=1}^{k} f\left(x^{(i)}\right) - kf\left(x^*\right) &\leq \frac{1}{2\eta}\left(\left\|x^{(0)} - x^*\right\|^2 - \left\|x^{(k)} - x^*\right\|^2\right) \\
&\leq \frac{1}{2\eta}\left\|x^{(0)} - x^*\right\|^2
\end{aligned}$$

Since $f\left(x^{(k)}\right) \leq f\left(x^{(k-1)}\right)$, $f\left(x^{(k)}\right) \leq \frac{1}{k}\sum_{i=1}^{k} f\left(x^{(i)}\right)$.

Hence,

$$f\left(x^{(k)}\right) - f\left(x^*\right) \leq \frac{1}{2k\eta}\left\|x^{(0)} - x^*\right\|^2.$$