MATH-329 Nonlinear optimization
Exercise session 3: Convexity

Instructor: Nicolas Boumal
TAs: Andrew McRae, Andreea Musat

Document compiled on September 24, 2024

1. Strict and strong.

1. Give an example of a function $f\colon \mathbb{R} \to \mathbb{R}$ which is convex but not strictly convex.
2. Give an example of a function $f\colon \mathbb{R} \to \mathbb{R}$ which is strictly convex but not strongly convex.
3. Give an example of a function $f\colon \mathbb{R} \to \mathbb{R}$ which is strictly convex yet not bounded below.
4. Give an example of a function $f\colon \mathbb{R} \to \mathbb{R}$ which is strictly convex and bounded below yet does not have a minimum.

Answer.

1. Any convex function with an affine portion (some line segment) will be convex but not strictly convex, e.g. $f(x) = |x|$.
2. An example can be $f(x) = e^x$. In fact, by the theorem on the characterization of convexity for twice differentiable functions (see lecture notes), we have:

$$\begin{aligned} f''(x) > 0 \text{ for all } x \in \mathbb{R} \quad &\implies \quad f \text{ is strictly convex} \\ f''(x) \geq \mu > 0 \text{ for all } x \in \mathbb{R} \quad &\iff \quad f \text{ is } \mu\text{-strongly convex} \end{aligned}$$

   In our case, $f''(x) = e^x > 0$ but $f''(x) \to 0$ as $x \to -\infty$.
3. Consider $f(x) = x + e^x$. This is strictly convex since $f''(x) = e^x > 0$ yet $f(x) \to -\infty$ as $x \to -\infty$.
4. Once again, an example is $f(x) = e^x$. It is strictly convex and lower bounded by zero, but the lower bound is never reached: the function does not have a minimum, but an infimum of 0.

■

2. Quadratic functions. Let $\mathcal{E} = \mathbb{R}^n$ with the usual inner product and let $f(x) = \frac{1}{2}x^\top Ax + b^\top x + c$.

1. Show that f is convex if and only if $A \succeq 0$.
2. Show that f is strictly convex if and only if $A \succ 0$.
3. Show that f is μ-strongly convex if and only if $A \succeq \mu I$. What is the best choice of μ is terms of A?

Answer. We rely on the theorem on the characterization of convexity for twice differentiable functions (see lecture notes). The Hessian of f is given by $\nabla^2 f(x) = A$ for all $x \in \mathcal{E}$. Therefore, direct application of the theorem gives:

1. The function f convex if and only if $\nabla^2 f(x) = A \succeq 0$.
2. We know that if $\nabla^2 f(x) = A \succ 0$ then f is strictly convex. Here, the converse is also true. To show this we use the theorem characterizing convexity for differentiable functions (see lecture notes):

$$
\begin{aligned}
& & & f \text{ strictly convex} \\
&\iff & \forall x, y \in \mathbb{R}^n, \quad & f(y) > f(x) + \langle \nabla f(x), y - x \rangle \\
&\iff & \forall x, y \in \mathbb{R}^n, \quad & \frac{1}{2}y^\top Ay + b^\top y + c > \frac{1}{2}x^\top Ax + b^\top x + c + (Ax + b)^\top (y - x) \\
&\iff & \forall x, y \in \mathbb{R}^n, \quad & \frac{1}{2}y^\top Ay + \frac{1}{2}x^\top Ax > x^\top Ay \\
&\iff & \forall x, y \in \mathbb{R}^n, \quad & \frac{1}{2}(y - x)^\top A(y - x) > 0 \\
&\iff & \forall v \in \mathbb{R}^n, \quad & v^\top Av > 0 \\
&\iff & & A \succ 0.
\end{aligned}
$$

3. The definition of strong convexity gives that f is μ-strongly convex if and only if $\nabla^2 f(x) = A \succeq \mu I$ for some $\mu > 0$. The best strong convexity constant is μ^*, which gives us the tightest quadratic lower bound on the function. In other words it is the largest $\mu \in \mathbb{R}$ such that $A - \mu I \succeq 0$. Denoting the eigenvalues of A as $\{\lambda_i\}_{i=1}^n$, the eigenvalues of $A - \mu I$ are $\{\lambda_i - \mu\}_{i=1}^n$ and we want

$$\lambda_i - \mu \geq 0, \quad \forall i = 1, \ldots, n.$$

Therefore $\mu^* = \min \lambda_i$, the smallest eigenvalue of A.

■

3. Jensen's inequality. Let $\mathcal{E}$ be a linear space. Let $f \colon \mathcal{E} \to \mathbb{R}$ be a convex function. Show that for all $x_1, \ldots, x_n \in \mathcal{E}$ and any $\lambda_1, \ldots, \lambda_n \geq 0$ such that $\lambda_1 + \cdots + \lambda_n = 1$ we have

$$f(\lambda_1 x_1 + \cdots + \lambda_n x_n) \leq \lambda_1 f(x_1) + \cdots + \lambda_n f(x_n).$$

Hint: proceed by induction on n.

We call the quantity $\lambda_1 x_1 + \cdots + \lambda_n x_n$ a convex combination of the points $x_1, \ldots, x_n$. The result that you proved shows that if X is a discrete random variable taking values $x_1, \ldots, x_n \in \mathcal{E}$ with probabilities $p_1, \ldots, p_n$ respectively, then we have

$$f(\mathbf{E}[X]) \leq \mathbf{E}[f(X)],$$

where $\mathbf{E}$ denotes mathematical expectation. This inequality generalizes to any random variable X.

Answer. The result holds for $n = 1, 2$. Suppose that it holds at order $n \in \mathbb{N}$. Let $\lambda_1, \ldots, \lambda_{n+1} \geq 0$ such that $\lambda_1 + \cdots + \lambda_{n+1} = 1$ and $x_1, \ldots, x_{n+1} \in \mathcal{E}$. Let $\bar{\lambda} = \lambda_1 + \cdots + \lambda_n$. Then we have

$$\begin{aligned} f(\lambda_1 x_1 + \cdots + \lambda_{n+1} x_{n+1}) &= f\Big(\bar{\lambda}\Big(\frac{\lambda_1}{\bar{\lambda}} x_1 + \cdots + \frac{\lambda_n}{\bar{\lambda}} x_n\Big) + \lambda_{n+1} x_{n+1}\Big) \\ &\leq \bar{\lambda} f\Big(\frac{\lambda_1}{\bar{\lambda}} x_1 + \cdots + \frac{\lambda_n}{\bar{\lambda}} x_n\Big) + \lambda_{n+1} f(x_{n+1}) \\ &\leq \bar{\lambda}\Big(\frac{\lambda_1}{\bar{\lambda}} f(x_1) + \cdots + \frac{\lambda_n}{\bar{\lambda}} f(x_n)\Big) + \lambda_{n+1} f(x_{n+1}) \\ &= \lambda_1 f(x_1) + \cdots + \lambda_{n+1} f(x_{n+1}), \end{aligned}$$

where we used the convexity of f for the first inequality (as $\bar{\lambda} + \lambda_{n+1} = 1$) and the induction hypothesis for the second inequality. ■

4. Log-sum-exp. *If you prefer you can consider the following exercise with $k = 2$.*

1. Show that the *log-sum-exp function* is convex from $\mathbb{R}^k$ to $\mathbb{R}$ ($t > 0$ is a fixed, real parameter):

$$f(x) = t \log\left(\sum_{i=1}^{k} e^{x_i/t}\right). \tag{1}$$

This function is often used in applications because it is a smooth approximation of the maximum function. Indeed:

2. With $\bar{x} = \max_i x_i$ show that

$$\bar{x} \leq f(x) = \bar{x} + t \log\left(\sum_{i=1}^{k} e^{\frac{x_i - \bar{x}}{t}}\right) \leq \bar{x} + t \log(k). \tag{2}$$

Thus, the smaller t is, the better the approximation. However:

3. From an optimization perspective (for example, if we plan to use gradient descent), can you see a reason why we should not take t too small?

Note: on a computer, it is necessary to use expression (2) rather than expression (1) to compute f (and its derivatives). Indeed, expression (1) can lead to overflow when t is small because it involves computing exponentials of possibly large numbers. In contrast, expression (2) only involves exponentials of nonpositive numbers. Still, even with expression (2), evaluating f and its derivatives can get tricky numerically when t is small.

Answer.

1. There are several ways to do it. The function f is smooth and its gradient is given by $\nabla f(x) = u/s$ where we let $s = \sum_{i=1}^{k} \exp(x_i/t)$ and

$$u = \begin{bmatrix} \exp(x_1/t) \\ \vdots \\ \exp(x_n/t) \end{bmatrix}.$$

We deduce that

$$\frac{\partial^2}{\partial_i \partial_j} f(x) = \frac{1}{ts^2} \begin{cases} \exp(x_i/t) \sum_{k \neq i} \exp(x_k/t) & \text{if } i = j, \\ -\exp(x_i/t) \exp(x_j/t) & \text{otherwise.} \end{cases}$$

So the Hessian can be written as follows:

$$\nabla^2 f(x) = \frac{1}{ts^2} \big(s \operatorname{diag}(u) - uu^\top \big),$$

where the diag operator transforms a vector into a diagonal matrix. The matrix $s \operatorname{diag}(u) - uu^\top$ is diagonally dominant so it is positive semidefinite. We conclude that $\nabla^2 f(x)$ is positive semidefinite and that f is convex.

This way to proceed is error prone. Another possibility is to prove that for all $x, y \in \mathbb{R}^k$ and $\theta \in [0, 1]$ we have

$$f((1-\theta)x + \theta y) \leq (1-\theta) f(x) + \theta f(y).$$

To do so we use Hölder's inequality, which states that

$$\Big| \sum_{i=1}^{n} x_i y_i \Big| \leq \Big(\sum_{i=1} |x_i|^p \Big)^{1/p} \Big(\sum_{i=1} |y_i|^q \Big)^{1/q}$$

for all vectors $x, y \in \mathbb{R}^k$ whenever p and q are positive numbers such that $1/p + 1/q = 1$. This is a generalization of Cauchy–Schwarz' inequality.

Alternatively, you may have seen in the lectures that when a function f is continuous, it is sufficient to show that there exists a single $\theta \in (0, 1)$ such that for all $x, y \in \mathbb{R}^k$

$$f((1-\theta)x + \theta y) \leq (1-\theta) f(x) + \theta f(y)$$

in order to conclude that f is convex. We will show this property with $\theta = 1/2$. For all $x, y \in \mathbb{R}^k$ we have

$$\begin{aligned} f(x/2 + y/2) &= t \log \Big(\sum_{i=1}^{k} \exp(x_i/2t + y_i/2t) \Big) \\ &= t \log \Big(\sum_{i=1}^{k} \exp(x_i/2t) \exp(y_i/2t) \Big) \\ &\leq t \log \bigg(\Big(\sum_{i=1}^{k} \exp(x_i/t) \Big)^{1/2} \Big(\sum_{i=1}^{k} \exp(y_i/t) \Big)^{1/2} \bigg) \\ &= \frac{t}{2} \log \Big(\sum_{i=1}^{k} \exp(x_i/t) \Big) + \frac{t}{2} \log \Big(\sum_{i=1}^{k} \exp(y_i/t) \Big) \\ &= f(x)/2 + f(y)/2, \end{aligned}$$

where the inequality comes from Cauchy–Schwarz.

2. Let $x \in \mathbb{R}^k$ and $\bar{x} = \max_i x_i$. Then

$$\begin{aligned} f(x) &= t \log \left(\sum_{i=1}^{k} \exp((x_i - \bar{x} + \bar{x})/t) \right) \\ &= t \log \left(\exp(\bar{x}/t) \sum_{i=1}^{k} \exp((x_i - \bar{x})/t) \right) \\ &= \bar{x} + t \log \left(\sum_{i=1}^{k} \exp((x_i - \bar{x})/t) \right) \\ &\leq \bar{x} + t \log(k), \end{aligned}$$

because each term in the sum is less than 1.

3. The Lipschitz constant of ∇f explodes when t goes to zero. Moreover we know from our analysis of gradient descent that when the Lipschitz constant is large the progress can be slower. You can visualize what is happening by plotting the function for $k = 2$. You will see that as t goes to zero f becomes a better approximation of the max function, which is not differentiable and certainly does not have Lipschitz continuous gradients.

■

5. Norms. Let $\mathcal{E}$ be a Euclidean space.

1. Show that any norm on $\mathcal{E}$ is convex.

2. Show that any squared norm on $\mathcal{E}$ is convex.

Interestingly, a norm is never differentiable at $x = 0$. Do you see why? However:

3. Let $\langle \cdot, \cdot \rangle$ be an inner product on $\mathcal{E}$ and $\| \cdot \|$ the associated norm (that is, $\|x\| = \sqrt{\langle x, x \rangle}$ for all $x \in \mathcal{E}$). Prove that the squared norm $x \mapsto \|x\|^2$ is differentiable.

A norm may not be differentiable if it is not derived from an inner product. Can you come up with an example?

Answer.

1. Let $\| \cdot \|$ be a norm on $\mathcal{E}$. Remember that a norm is a map $\mathcal{E} \to \mathbb{R}$ such that

   (a) $\|\alpha x\| = |\alpha| \|x\|$ for all $\alpha \in \mathbb{R}, x \in \mathcal{E}$,

   (b) $\|x + y\| \leq \|x\| + \|y\|$,

   (c) $\|x\| \geq 0$ for all $x \in \mathcal{E}$ and $\|x\| = 0 \iff x = 0$.

Let $x, y \in \mathcal{E}$ and $t \in [0,1]$. Then

$$\begin{aligned}\|(1-t)x + ty\| &\leq \|(1-t)x\| + \|ty\| \\ &= |1-t|\|x\| + |t|\|y\| \\ &= (1-t)\|x\| + t\|y\|\end{aligned}$$

where we used the two first axioms of the norm and the fact that $t, 1-t \geq 0$.

2. We let g be a norm on $\mathcal{E}$ and $f(x) = x^2$. We know from the previous question that g is convex and nonnegative. Also, f is convex and nondecreasing on $[0, +\infty[$, which is the range of g. So we conclude that $f \circ g$ is convex.
   Alternatively we could also show the convexity using the definition.

3. Let $f\colon x \mapsto \langle x, x\rangle$ be the squared norm. For all $x, u \in \mathcal{E}$ and $t \in \mathbb{R}$ we have

$$f(x + tu) = f(x) + 2t\langle x, u\rangle + O(t^2).$$

   We deduce that

$$\lim_{t\to 0} \frac{1}{t}(f(x+tu) - f(x)) = 2\langle x, u\rangle .$$

   The limit exists for all u so we conclude that f is differentiable.

■