Exercise List: Properties and examples of convexity and smoothness

Robert M. Gower.

October 1, 2018

Time to get familiarized with convexity, smoothness and a bit of strong convexity.

Notation: For every $x, y, \in \mathbb{R}^d$ let $\langle x, y\rangle \overset{\text{def}}{=} x^\top y$ and let $\|x\|_2 = \sqrt{\langle x, x\rangle}$.

Let $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ be the smallest and largest singular values of A defined by

$$\sigma_{\min}(A) \overset{\text{def}}{=} \min_{x\in\mathbb{R}^d} \frac{\|Ax\|_2}{\|x\|_2} \quad \text{and} \quad \sigma_{\max}(A) \overset{\text{def}}{=} \max_{x\in\mathbb{R}^d} \frac{\|Ax\|_2}{\|x\|_2}. \tag{1}$$

Thus clearly

$$\frac{\|Ax\|_2^2}{\|x\|_2^2} \leq \sigma_{\max}(A)^2, \quad \forall x \in \mathbb{R}^d. \tag{2}$$

Let $\|A\|_F^2 \overset{\text{def}}{=} \operatorname{Tr}\left(A^\top A\right)$ denote the Frobenius norm of A. Finally, a result you will need, for every symmetric matrix G the $L2$ induced matrix norm can be equivalently defined by

$$||G||_2 = \sigma_{\max}(G) = \sup_{x\in\mathbb{R}^d,\, x\neq 0} \frac{|\langle Gx, x\rangle|}{\|x\|_2^2} = \max_{x\in\mathbb{R}^d,\, x\neq 0} \frac{\|Gx\|_2}{\|x\|_2}. \tag{3}$$

1 Convexity

We say that a twice differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y), \quad \forall x, y \in \mathbb{R}^d, \lambda \in [0,\, 1]. \tag{4}$$

or equivalently

$$v^\top \nabla^2 f(x) v \geq 0, \quad \forall x, v \in \mathbb{R}^d. \tag{5}$$

We say that f is μ–strongly convex if

$$v^\top \nabla^2 f(x) v \geq \mu \|v\|_2^2, \quad \forall x, v \in \mathbb{R}^d. \tag{6}$$

Ex. 1 — We say that $\|\cdot\| \rightarrow \mathbb{R}_+$ is a norm over $\mathbb{R}^d$ if it satisfies the following three properties

1. **Point separating:** $\|x\| = 0 \Leftrightarrow x = 0, \forall x \in \mathbb{R}^d$.
2. **Subadditive:** $\|x + y\| \leq \|x\| + \|y\|, \forall x, y \in \mathbb{R}^d$
3. **Homogeneous:** $\|ax\| = |a|\|x\|, \forall x \in \mathbb{R}^d, a \in \mathbb{R}$.

Part I

Prove that $x \mapsto \|x\|$ is a convex function.

Part II

For every convex function $f : y \in \mathbb{R}^m \mapsto f(y)$, prove that $g : x \in \mathbb{R}^d \mapsto f(Ax - b)$ is a convex function, where $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$.

Part III

Let $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex for $i = 1, \ldots, n$. Prove that $\sum_{i=1}^n f_i$ is convex.

Part IV

For given scalars $y_i \in \mathbb{R}$ and vectors $a_i \in \mathbb{R}^d$ for $i = 1, \ldots, m$ prove that the *logistic regression* function $f(x) = \frac{1}{n}\sum_{i=1}^n \ln(1 + e^{-y_i\langle x, a_i\rangle})$ is convex.

Part V

Let $A \in \mathbb{R}^{n\times d}$ have full column rank. Prove that $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ is $\sigma^2_{\min}(A)$–strongly convex.

Part VI

Now suppose that the function $f(x)$ is μ–strongly convex, that is, it satisfies

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|_2^2, \quad \forall x, y \in \mathbb{R}^d. \tag{7}$$

Prove that $f(x)$ satisfies the *Polyak–Lojasiewicz* condition, that is

$$\|\nabla f(x)\|_2^2 \geq 2\mu(f(x) - f(x^*)), \quad \forall x. \tag{8}$$

Answer (Ex. I) — Let $x, y \in \mathbb{R}^d$ and $\lambda \in [0, 1]$. It follows that

$$\begin{aligned} \|\lambda x + (1-\lambda)y\| &\overset{\text{item 2}}{\leq} \|\lambda x\| + \|(1-\lambda)y\| \\ &\overset{\text{item 3}}{\leq} \lambda\|x\| + (1-\lambda)\|y\|. \quad \blacksquare \end{aligned}$$

Answer (Ex. II) — Let $x, y \in \mathbb{R}^d$ and $\lambda \in [0,\, 1]$. It follows that

$$\begin{aligned} g(\lambda x + (1-\lambda)y) \quad &= \quad f(A(\lambda x + (1-\lambda))y - b) \\ &= \quad f(\lambda(Ax-b) + (1-\lambda)(Ay-b)) \\ &\overset{f \text{ is conv.}}{=} \quad \lambda f(Ax-b) + (1-\lambda)f(Ay-b). \quad \blacksquare \end{aligned} \tag{9}$$

Answer (Ex. III) — Immediate through either definition.

Answer (Ex. IV) — From exercise V we need only prove that $f(x) = \ln(1 + e^{-y\langle x,w\rangle})$ is convex for a given $y \in \mathbb{R}$ and $w \in \mathbb{R}^d$. From exercise II we need only prove that $\phi(\alpha) = \ln(1+e^{\alpha})$ is convex, since $x \mapsto -y\,\langle x, w\rangle$ is a linear function. The convexity of $f(\alpha)$ now follows by differentiating once

$$\phi'(\alpha) = \frac{e^{\alpha}}{1+e^{\alpha}},$$

then differentiating again

$$\phi''(\alpha) = \frac{e^{\alpha}}{1+e^{\alpha}} - \frac{e^{2\alpha}}{(1+e^{\alpha})^2} = \frac{e^{\alpha}}{(1+e^{\alpha})^2} \geq 0, \quad \forall \alpha. \tag{10}$$

We can now call upon the definition (5), but since $\alpha \in \mathbb{R}$ is a scalar, the above already proves that $\phi(\alpha)$ is convex.

Answer (Ex. V) — Differentiating twice we have that

$$\nabla^2 f(x) = A^\top A.$$

Consequently

$$v^\top \nabla^2 f(x) v = v^\top A^\top A v = \|Av\|_2^2 \geq \sigma_{\min}(A)^2 \|v\|_2^2.$$

Answer (Ex. VI) — Multiplying (7) by minus and substituting $y = x^*$ we have that

$$\begin{aligned} f(x) - f(x^*) \quad &\leq \quad \langle \nabla f(x), x - x^* \rangle - \frac{\mu}{2}\|x^* - x\|_2^2 \\ &= \quad -\frac{1}{2}\|\sqrt{\mu}(x-x^*) - \frac{1}{\sqrt{\mu}}\nabla f(x)\|_2^2 + \frac{1}{2\mu}\|\nabla f(x)\|_2^2 \\ &\leq \quad \frac{1}{2\mu}\|\nabla f(x)\|_2^2. \end{aligned}$$

2 Smoothness

We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is L–smooth if

$$\|\nabla f(x) - \nabla f(y)\| \quad \leq \quad L\|x - y\| \tag{11}$$

or equivalently if f is twice differentiable then

$$v^\top \nabla^2 f(x) v \leq L\|v\|_2^2, \quad \forall x, v \in \mathbb{R}^d. \tag{12}$$

Ex. 2 — *Part I*

Prove that $x \mapsto \frac{1}{2}\|x\|^2$ is 1–smooth.

Part II

Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice differentiable and L–smooth. Show that

$$\sigma_{\max}(\nabla^2 f(x)) = \|\nabla^2 f(x)\|_2 \leq L.$$

Part III

For every twice differentiable L–smooth function $f : y \in \mathbb{R}^n \mapsto f(y)$, prove that $g : x \in \mathbb{R}^d \mapsto f(Ax - b)$ is a smooth function, where $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$. Find the smoothness constant of g.

Part IV

Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable and L_i–smooth for $i = 1, \ldots, n$. Prove that $\frac{1}{n}\sum_{i=1} f_i$ is $\sum_{i=1} \frac{L_i}{n}$–smooth.

Part V

For given scalars $y_i \in \mathbb{R}$ and vectors $a_i \in \mathbb{R}^d$ for $i = 1, \ldots, n$ prove that the *logistic regression* function $f(x) = \frac{1}{n}\sum_{i=1}^n \ln(1 + e^{-y_i\langle x, a_i\rangle})$ is smooth. Find the smoothness constant!

Part VI

Let $A \in \mathbb{R}^{n \times d}$ be any matrix. Prove that $\|Ax - b\|_2^2$ is $\sigma_{\max}^2(A)$–smooth.

Part VII

Let $M > 0$ be a positive constant. Let $f(x) = \frac{1}{n}\sum_{i=1}^n \phi_i(a_i^\top x)$ where $\phi_i : \mathbb{R} \to \mathbb{R}$ is a scalar function such that $\phi_i''(t) \leq M$ for all $t \in \mathbb{R}$. Prove that $f(x)$ is $\frac{M}{n}\sigma_{\max}^2(A)$–smooth. With this result, can you find a better estimate of the smoothness constant of the logistic regression loss?
Hint 1: ...

Answer (Ex. I) — Clearly $\nabla^2 \frac{1}{2}\|x\|^2 = I$ and thus follows from definition (11).

Answer (Ex. II) — Using that the induced norm for symmetric matrices is given by

$$\|\nabla^2 f(x)\|_2 = \sup_{v\neq 0} \frac{|v^\top \nabla^2 f(x) v|}{\|v\|_2^2} \overset{(12)}{\leq} \sup_{v\neq 0} \frac{L\|v\|_2^2}{\|v\|_2^2} = L.$$

Answer (Ex. III) — Differentiating $g(x)$ once gives

$$\nabla g(x) = A^\top \nabla f(Ax-b).$$

First we prove the claim using the definition (11). Indeed note that

$$\begin{array}{rcl} \|\nabla g(x) - \nabla g(y)\|_2 & = & \|A^\top\left(\nabla f(Ax-b) - \nabla f(Ay-b)\right)\|_2 \\ & \leq & \|A^\top\|_2 \|\nabla f(Ax-b) - \nabla f(Ay-b)\|_2 \\ & \overset{\text{smooth. of } f}{\leq} & L\|A^\top\|_2 \|Ax-b-(Ay-b)\|_2 \\ & \leq & L\|A^\top\|_2\|A\|_2\|x-y\|_2. \end{array}$$

This the smoothness parameter is given by $L\|A\|_2^2$ where we used that $\|A^\top\|_2 = \|A\|_2$. This completes the proof.
We can also prove the claim using (12). Differentiating again we have that

$$\nabla^2 g(x) = A^\top \nabla^2 f(Ax-b) A.$$

Consequently

$$\|\nabla^2 g(x)\|_2^2 \leq \|A\|_2^2 \|\nabla^2 f(Ax-b)\|_2^2 \leq L\|A\|_2^2.$$

We could further tighten this by considering the smoothness constant of f restricted to the set $\{x \,|\, Ax-b\}$ which might be smaller then $\mathbb{R}^d$.

Answer (Ex. IV) — Clearly

$$\nabla^2 (\tfrac{1}{n}\sum_{i=1}^n f_i(x)) = \tfrac{1}{n}\sum_{i=1}^n \nabla^2 f_i(x) \preceq \tfrac{1}{n}\sum_{i=1}^n L_i I.$$

You can also prove this using the definition (11) and applying repeatedly the subadditivity of the norm.

Answer (Ex. V) — First note that from (10) we can see that the function $\phi(\alpha) = \ln(1+e^\alpha)$ is at least 1–smooth. Consequently from exercise II the function $f_i(x) = \ln(1+e^{-y_i\langle x, a_i\rangle})$ is $y_i^2\|a_i\|_2^2$–smooth. Finally from exercise III the logistic regression function is $\sum_{i=1}^n \frac{y_i^2\|a_i\|_2^2}{n}$–smooth.
But this is not the tightest smoothness constant, as we will see in the next two exercises!

Answer (Ex. VI) — Differentiating twice we have that

$$\nabla^2 f(x) = A^\top A.$$

Consequently

$$v^\top \nabla^2 f(x) v = v^\top A^\top A v \leq \|Av\|_2^2 \leq \sigma_{\max}(A)^2 \|v\|_2^2.$$

Answer (Ex. VII) — By analysing directly the Hessian of $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ we see that

$$\nabla^2 f(x) = A^\top \Phi(x) A,$$

where $\Phi(x) = \mathrm{diag}(\phi_1''(a_1^\top x), \ldots, \phi_n''(a_n^\top x))$, Consequently

$$\|\nabla^2 f(x)\|_2 = \frac{1}{n}\|A^\top \Phi(x) A\|_2 \leq \frac{1}{n}\|A\|_2^2 \|\Phi(x)\|_2 \leq M\|A\|_2^2 \overset{(1)}{=} \frac{M}{n}\sigma_{\max}(A)^2.$$

For the logistic function, note that $\phi''(a_i^\top x) = \frac{e^\alpha}{(1+e^\alpha)^2}$, where $\alpha = -y_i \left\langle a_i, x \right\rangle$. Furthermore

$$\phi''(\alpha) = \frac{e^\alpha}{(1+e^\alpha)^2} \leq \frac{1}{4}, \quad \forall \alpha. \tag{13}$$

Consequently a better estimate of the smoothness constant is given by

$$L \leq \frac{\sigma_{\max}(A)^2}{4n}.$$

This is a much tighter smoothness constant and the one that is used in practice.