Exercise List: Strong convexity and smoothness

Robert M. Gower, Francis Bach and Hadrien Hendrikx

February 5th 2019

Time to get familiarized with convexity, smoothness and strong convexity, and finish the proof for gradient descent. Least-squares as a bonus special example.

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

1 Quick review of eigenvalues of symmetric matrices

Let S be a real squared symmetric matrix of size $d \times d$. Then, the spectral theorem states that S can be decomposed as

$$S = UDU^\top,$$

where $D = \text{diag}\left(\lambda_1, ..., \lambda_d\right)$ is a diagonal matrix and U is such that $UU^\top = U^\top U = I$. We can further assume that $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$. Values λ_i are called the *eigenvalues* of S and the columns of U are their associated *eigenvectors*. They are such that for all $i \in \{1, ..., d\}$,

$$SU_i = \lambda_i U_i. \tag{1}$$

The eigenvectors of S form an orthonormal *basis* of $\mathbb{R}^d$, meaning that any $x \in \mathbb{R}^d$, x can be written as $x = UU^\top x = \sum_{i=1}^d U_i(U_i^\top x)$, where the $(U_i^\top x)$ are the coefficients of x in the eigenbasis. In particular, if we note $\lambda_{\min}(S)$ and $\lambda_{\max}(S)$ the smallest and highest eigenvalues of S, they can be obtained as:

$$\lambda_{\min}(S) = \min_{x,\ \|x\|_2=1} x^\top Sx, \qquad \lambda_{\max}(S) = \max_{x,\ \|x\|_2=1} x^\top Sx.$$

2 Convexity

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

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y), \quad \forall x, y \in \mathbb{R}^d, \lambda \in [0,\, 1]. \tag{2}$$

or equivalently

$$\lambda_{\min}\left(\nabla^2 f(x)\right) \geq 0, \quad \forall x \in \mathbb{R}^d. \tag{3}$$

We say that f is μ–strongly convex if

$$\lambda_{\min}\left(\nabla^2 f(x)\right) \geq \mu, \quad \forall x \in \mathbb{R}^d. \tag{4}$$

Ex. 1 — We say that $\|\cdot\| \to \mathbb{R}_+$ 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 $f_i : \mathbb{R}^d \to \mathbb{R}$ 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 $\lambda_{\min}(A^\top A)$–strongly convex.

Part VI

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

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2, \quad \forall x, y \in \mathbb{R}^d. \tag{5}$$

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

$$\|\nabla f(x)\|_2^2 \geq 2\mu(f(x) - f(x^*)), \quad \forall x. \tag{6}$$

3 Smoothness

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

$$\forall x, y \in \mathbb{R}^d, \quad f(x) \leqslant f(y) + \nabla f(y)^\top (x-y) + \frac{L}{2}\|x-y\|_2^2, \tag{9}$$

or equivalently if f is twice differentiable then

$$\lambda_{\max}(\nabla^2 f(x)) \leq L, \quad \forall x \in \mathbb{R}^d. \tag{10}$$

Ex. 2 — *Part I*

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

Part II

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 III

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

Part IV

For given scalars $y_i \in \mathbb{R}$ and vectors $a_i \in \mathbb{R}^d$ for $i = 1, \dots, n$ 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 V

Let $A \in \mathbb{R}^{n\times d}$ be any matrix. Prove that $\frac{1}{2}\|Ax-b\|_2^2$ is $\lambda_{\max}(A^\top A)$–smooth.

Part VI

(BONUS) 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 $\frac{M}{n}\lambda_{\max}(A^\top A)$–smooth. With this result, can you find a better estimate of the smoothness constant of the logistic regression loss?
Hint : Show that $-\nabla^2 f(x) + \frac{M}{n}A^\top A$ is positive semidefinite.

Part VII

(Proof of convergence for strongly-convex smooth gradient descent)
(a) Consider $w_t = w_{t-1} - \frac{1}{L}\nabla f(w_{t-1})$ for f L-smooth. Show that

$$f(w_t) \leq f(w_{t-1}) - \frac{1}{2L}\|\nabla f'(w_{t-1})\|_2^2.$$

(b) Using the *Polyak–Lojasiewicz* condition, show that, for w_* a global minimizer of f,

$$f(w_t) \leq f(w_{t-1}) - \frac{\mu}{L}\big[f(w_{t-1}) - f(w_*)\big].$$

(c) Show that $f(w_t) - f(w_*) \leq (1-\mu/L)^t\big[f(w_0) - f(w_*)\big]$.

Part VIII

(BONUS) (Proof of convergence for smooth gradient descent)
(a) Show that, for w_* a global minimizer of f, $(w - w_*)^\top \nabla f(w) \geq f(w) - f(w_*)$.
(b) Consider $w_t = w_{t-1} - \frac{1}{L}\nabla f(w_{t-1})$ for f L-smooth. Show that, for w_* a global minimizer of f,

$$\|w_t - w_*\|_2^2 = \|w_{t-1} - w_*\|_2^2 + \frac{1}{L^2}\|\nabla f(w_{t-1})\|^2 - \frac{2}{L}\big[f(w_{t-1}) - f(w_*)\big].$$

(c) By linearly combining with $f(w_t) \leq f(w_{t-1}) - \frac{1}{2L}\|\nabla f'(w_{t-1})\|_2^2$, show that

$$t\big[f(w_t) - f(w_*)\big] + \frac{L}{2}\|w_t - w_*\|^2$$

is decreasing and conclude on the convergence of gradient descent.

Part IX

(BONUS) Co-coercivity. Let f be L-smooth, show that

$$\langle \nabla f(y) - \nabla f(x), y - x\rangle \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_2^2$$

Hint: Start by showing that $f(y) - f(x) \leq \langle \nabla f(y), y - x\rangle - \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2$, by considering the lower-bound of f at x and the upper-bound of f at y, both taken at a generic z.

4 Gradient descent

We will now solve the following ridge regression problem

$$w^* = \arg\min_{w\in\mathbb{R}^d} \left(\frac{1}{2n}\|X^\top w - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2 \stackrel{\text{def}}{=} f(w)\right), \tag{12}$$

using gradient descent.

Ex. 3 — Consider the Gradient descent method

$$w^{t+1} = w^t - \alpha \nabla f(w^t), \tag{13}$$

where

$$\alpha = \frac{1}{\lambda_{\max}(A)}, \tag{14}$$

is a fixed stepsize and

$$A \overset{\text{def}}{=} \tfrac{1}{n} X X^\top + \lambda I. \tag{15}$$

Part I

Show that the gradient $\nabla f(x)$ of (12) is given by

$$\nabla f(w) = Aw - b = A(w - w^*),$$

where w^* is the solution to (12) and

$$b \overset{\text{def}}{=} \tfrac{1}{n} X y.$$

Now that we have calculated the gradient, re-write the iterates (13) using this gradient.

Part II

Show or convince yourself that A as defined in (15) is positive semi-definite, that is

$$\langle Aw, w \rangle \geq 0, \quad \forall w \in \mathbb{R}^d, \tag{16}$$

and that

$$\lambda_{\max}(I - \alpha A) = 1 - \alpha\, \lambda_{\min}(A) = 1 - \frac{\lambda_{\min}(A)}{\lambda_{\max}(A)}. \tag{17}$$

Part III

Show that the iterates (13) converge to w^* according to

$$\|w^{t+1} - w^*\|_2 \leq \left(1 - \frac{\lambda_{\min}(A)}{\lambda_{\max}(A)}\right) \|w^t - w^*\|_2,$$

for all t. The number $(1 - \lambda_{\min}(A)/\lambda_{\max}(A))$ is known as the *rate of convergence.*

Hint 1: Subtract w^* from both sides of (13) and use the results from the previous two exercises.
Hint 2: Try and show that $b = Aw^*$!

Part IV

Let

$$\kappa(A) \stackrel{\text{def}}{=} \frac{\lambda_{\text{max}}(A)}{\lambda_{\text{min}}(A)},$$

which is known as the condition number of A. What happens to κ as $\lambda \to \infty$ and $\lambda \to 0$, respectively? What does this imply about the speed at which gradient descent converges to the solution?

Part V

(BONUS) Let us consider the extreme case where $\lambda = 0$. Consider the coordinate change $\hat{w} = P^{-1}w$, where $P \in \mathbb{R}^{d\times d}$ is invertible. With this coordinate change we can solve the problem in $\hat{w}$ given by

$$\hat{w}^* = \arg\min_{\hat{w}\in\mathbb{R}^d} \left(\frac{1}{2n}\|X^\top P\hat{w} - y\|_2^2 + \frac{\lambda}{2}\|P\hat{w}\|_2^2 \right), \tag{18}$$

then switch back the coordinate system to get the solution in w^* given by

$$w^* = P\hat{w}^*. \tag{19}$$

If we use gradient descent to solve (18), at what rate does it converge? To get the fastest rate possible, what should P be? Does the choice

$$P = \mathrm{diag}(XX^\top)^{-1}, \tag{20}$$

make sense?

Remark: The matrix P is known as the preconditioner and the particular choice given by (20) is a standard choice known as "feature scaling" and it is often used in machine learning.