Exercises: intro to optimization and gradient descent

1 Convexity: general results

1.1

Show that a sum of smooth functions is smooth. What is the corresponding smoothness constant?

Show that the sum of strongly convex functions is strongly convex. What is the corresponding strong convexity constant ?

1.2

Show that $x \to \|x\|$ is convex, where $\|\cdot\|$ is any norm on $\mathbb{R}^d$.

1.3

Let $f : \mathbb{R}^d \to \mathbb{R}$ convex. Show that $g(x) = f(Ax + b)$ is convex, where $A \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^d$. If f is μ-strongly convex, is g strongly convex? If so, what is a strong convexity constant of g? If f is L-smooth, is g smooth? If so, what is a smoothness constant of g?

Hint: You can demonstrate, and then use the fact that $\sigma_{\min}(AB) \geq \sigma_{\min}(A)\sigma_{\min}(B)$ and $\sigma_{\max}(AB) \leq \sigma_{\max}(A)\sigma_{\max}(B)$ for two square matrices A, B.

1.4

Let $h_1, \dots, h_n : \mathbb{R} \to \mathbb{R}$ some convex function, $X \in \mathbb{R}^{n\times p}$ and define

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} h_i(\langle x_i, w\rangle),$$

where $x_i \in \mathbb{R}^p$ is the n-th row of X. Assume that the h_i are such that $\sup_{t\in\mathbb{R}} h_i''(t) = M < +\infty$. Show that f is smooth, and determine a smoothness constant.

2 Convexity / non-convexity of matrix functions

2.1

Let $m \in \mathbb{R}$ and define $f(x) = \frac{1}{2}(x - m)^2$, $g(a, b) = \frac{1}{2}(ab - m)^2$. What are the gradient/ Hessian of these functions? Are these functions convex ?

2.2

Determine the set of points a, b such that $\nabla^2 g(a, b)$ is positive. What do you observe at the minimum? Could we have predicted this?

2.3

Let $M \in \mathbb{R}^{p\times p}$ and define $f(X) = \frac{1}{2}\|X - M\|_F^2$, $g(A, B) = \frac{1}{2}\|AB - M\|_F^2$ where $A, B \in \mathbb{R}^{p\times p}$. What are the gradient/ Hessian of these functions? Are these functions convex ?

Hint: here, it is convenient to write the Hessians as linear operators. For instance for f, we can write $\nabla^2 f(X)(U) = \dots$ where $\dots$ is a linear function of $U \in \mathbb{R}^{p\times p}$.

3 Polyak-Lojasciewicz inequality

Let $f : \mathbb{R} \to \mathbb{R}$ be a μ-strongly convex function. Let x^* its arg-minimum. Show that f verifies the Polyak-Lojasciewicz inequality:

$$\forall x \in \mathbb{R}^d,\ \ f(x) - f(x^*) \leq \frac{1}{2\mu}\|\nabla f(x)\|^2$$

4 Gradient descent in a simple case

We let $p \geq 0$, and consider a vector $b \in \mathbb{R}^p$ and a matrix $A \in \mathbb{R}^{p\times p}$. We assume that A is a symmetric matrix with positive eigenvalues $\lambda_{\max} = \lambda_1 \geq \cdots \geq \lambda_p = \lambda_{\min}$. We define the following *quadratic* objective function:

$$f(x) = \frac{1}{2}x^\top Ax - b^\top x$$

Exercise 1: Show that this function is convex, and that its gradient is given by $\nabla f(x) = Ax - b$. Find the analytical expression of its minimizer x^*, and of $f(x^*)$.

We now consider the sequence of iterates of gradient descent with a step size $\rho > 0$, starting from $x_0 = 0$:

$$\text{For } n \geq 0 :\ \ x_{n+1} = x_n - \rho\nabla f(x_n)$$

Exercise 2: Obtain a closed form expression for x_n. Hint : what recursion does the sequence $y_n = x_n - x^*$ satisfy?

We now use the spectral decomposition of A, and write

$$A = U^\top DU$$

where $D = \text{diag}(\lambda_1, \dots, \lambda_p)$ contains the eigenvalues of A and $U \in \mathbb{R}^{p\times p}$ contains the eigenvectors of A. We recall that $UU^\top = U^\top U = I_p$.

Exercise 3: Define $z_n = U(x_n - x^*)$. Show that z_n is given by

$$z_n = (I_p - \rho D)^n z_0$$

Give a condition on ρ for this sequence to converge to 0.

In the following, we assume that $\rho = \frac{1}{\lambda_{\max}}$.

Exercise 4: Demonstrate that $\|x_n - x^*\| \leq (1 - \frac{\lambda_{\min}}{\lambda_{\max}})^n\|x^*\|$.

This is what we call *linear* convergence, and $1 - \frac{\lambda_{\min}}{\lambda_{\max}}$ is the rate of convergence.

The quantity $\kappa = \frac{\lambda_{\min}}{\lambda_{\max}}$ is called the *conditioning* of the matrix A, and, by extension, of the function f. This number is always between 0 and 1. The closer it is to one, the faster gradient descent converges.

Here, if for instance $\kappa = \frac{1}{2}$, then the convergence is very fast: $\|x_n - x^*\| \leq \frac{1}{2^n}\|x^*\|$, every iteration halves the error. However, in some cases we can have some very poorly conditioned problems.

Exercise 5: Assume that $\kappa = \frac{1}{1000}$, and that $\|x^*\| = 1$. How many iterations of gradient descent are needed to reach an error $\|x_n - x^*\| \leq \frac{1}{10}$? and to get $\|x_n - x^*\| \leq \frac{1}{100}$?

In these badly conditioned case, it would be useful to obtain a bound on the error that does not depend on the conditioning of the problem. To get such a bound, we look at another measure of the error, $f(x_n) - f(x^*)$.

Exercise 6: Show that for all x, $f(x) - f(x^*) = \frac{1}{2}(x - x^*)^\top A(x - x^*)$. Deduce a closed form formula for $f(x_n) - f(x^*)$.

We are now ready to give a bound that does not depend on the conditioning of the problem:

Exercise 7: Show that for all $\mu \in [0, 1]$ and all n we have $(1-\mu)^{2n}\mu \leq \frac{1}{2n+1}$. Deduce that

$$f(x_n) - f(x^*) \leq \frac{1}{\rho(2n+1)}\|x^*\|^2$$

This is what we call *sub-linear* convergence. Note that this rate of convergence does not get worse when $\lambda_{\min}$ goes to 0: it does not depend on the conditioning of the problem.