Labs
Optimization for Machine Learning
Spring 2022

EPFL
School of Computer and Communication Sciences
Martin Jaggi & Nicolas Flammarion
github.com/epfml/OptML_course

Problem Set 5 — Solutions (Recap on convexity and gradient descent algorithms.)

Gradient descent on a quadratic function. Consider the quadratic function $f(x) = \frac{1}{2}x^\top Ax + \langle b, x\rangle + c$, where A is a $d \times d$ symmetric matrix, $b \in \mathbb{R}^d$ and c in $\mathbb{R}$.

1. What are the minimal conditions on A, b and c that ensure that f is strictly convex ? For the rest of the exercise we assume that these conditions are fulfilled.

   Answer: f is a quadratic function which has a constant hessian equal to A since for all $x \in \mathbb{R}^d$, $\nabla^2 f(x) = \frac{1}{2}(A + A^\top) = A$ (last equality is because A is symmetric). f is therefore **strictly convex** iif $A \succ 0$. Note that we don't need any assumptions on b or c.

2. Is f strongly convex ?

   Answer: Since $A \succ 0$, all the eigenvalues are strictly positive, let μ denote its smallest one. This leads to $A \succeq \mu I_d$ and therefore $\nabla^2 f(x) - \mu I_d \succeq 0$ which means that $x \mapsto f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex. From Lemma 2.11 this is equivalent to f being strongly convex with parameter μ.

3. Prove that f has a unique minimum x^* and give its closed form expression.

   Answer: Since f is strongly convex, we have from Lemma 2.12 that f has a unique global minimum. Furthermore $\nabla f(x) = Ax + b$, setting it to 0 we get that $x^* = -A^{-1}b$ (note that A is indeed invertible since from question 1 we assume that $A \succ 0$).

4. Show that f can be rewritten as $f(x) = \frac{1}{2}(x - x^*)^\top A(x - x^*) + f(x^*)$

   Answer: Direct computations provide:

$$
\begin{aligned}
f(x) - f(x^*) &= \frac{1}{2}x^\top Ax + \langle b, x - x^*\rangle - \frac{1}{2}{x^*}^\top Ax^* \\
&= \frac{1}{2}x^\top Ax - \langle Ax^*, x - x^*\rangle - \frac{1}{2}{x^*}^\top Ax^* \\
&= \frac{1}{2}(x - x^*)^\top A(x - x^*)
\end{aligned}
$$

5. From an initial point $x_0 \in \mathbb{R}^d$, assume we run gradient descent with step-size $\gamma > 0$ on the function f. Show that the n^{th} iterate x_n satisfies $x_n = x^* + (I_d - \gamma A)^n\ (x_0 - x^*)$, where I_d is the $d \times d$ identity matrix.

   Answer: From the previous expression we get that $\nabla f(x) = A(x - x^*)$, hence one step of gradient descent corresponds to $x_n = x_{n-1} - \gamma\nabla f(x_{n-1}) = x_{n-1} - \gamma A(x_{n-1} - x^*)$. Therefore $x_n - x^* = (I_d - \gamma A)(x_{n-1} - x^*) = (I_d - \gamma A)^n(x_0 - x^*)$.

6. In which range must the step-size γ be so that the iterates towards x^* ?

   Answer: Since A is symmetric we can diagonalise A as $A = P \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{pmatrix} P^\top$, where $PP^\top = P^\top P = I_d$ and $\lambda_1 \geq \cdots \geq \lambda_d$. Therefore $(x_n - x^*) = P \begin{pmatrix} (1-\gamma\lambda_1)^n & & 0 \\ & \ddots & \\ 0 & & (1-\gamma\lambda_d)^n \end{pmatrix} P^\top(x_0 - x^*)$.

   In order to have that $x_n \underset{n\to\infty}{\longrightarrow} x^*$ we must have that $-1 < 1 - \gamma\lambda_i < 1$ for all i, i.e. $0 < \gamma < 2/\lambda_i$ for all i. This leads to having $\gamma \in (0, \frac{2}{L})$ where L corresponds to the largest eigenvalue of A (which also corresponds to the smoothness constant of our quadratic).