

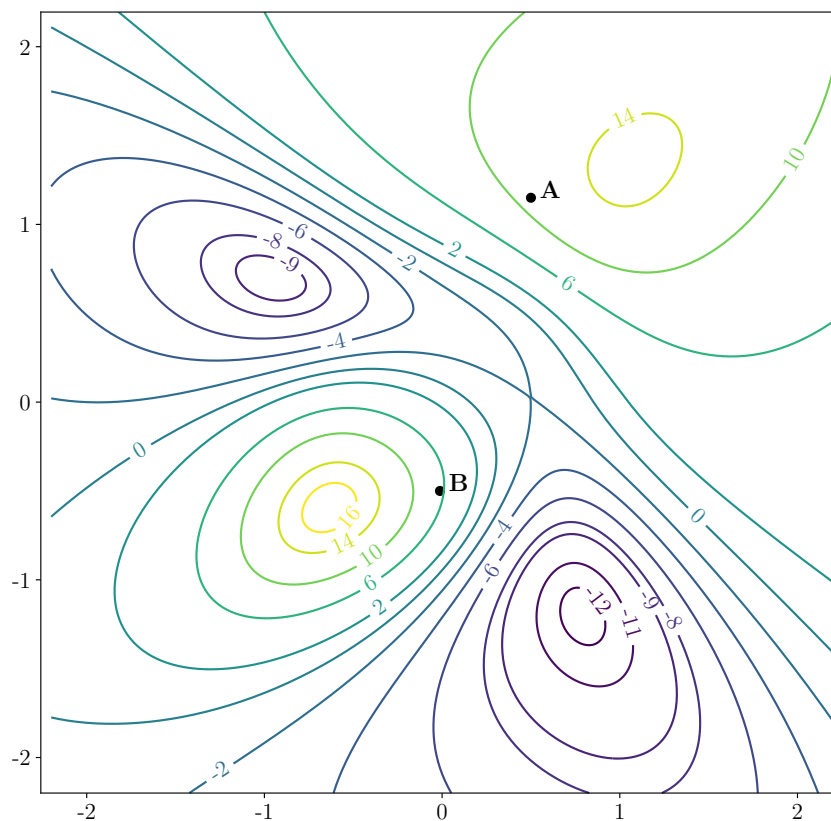
Optimization and Computational Linear Algebra for Data Science

Homework 12: Gradient descent

Due on December 13, 2019

-
- Unless otherwise stated, all answers must be mathematically justified.
 - Partial answers will be graded.
 - You can work in groups but each student must write his/her own solution based on his/her own understanding of the problem. Please list on your submission the students you work with for the homework (this will not affect your grade).
 - Problems with a (★) are extra credit, they will not (directly) contribute to your score of this homework. However, for every 4 extra credit questions successfully answered your lowest homework score get replaced by a perfect score.
 - If you have any questions, feel free to contact me (lm4271@nyu.edu) or to stop at the office hours.
-

Problem 12.1 (2 points). *The following plot shows the contour lines of a function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$.*



- Give (approximately) the coordinates of the global/local minimizers/maximizers, saddle points of f .*
- Assume that we run gradient descent to minimize f . Will gradient descent converge to the global minimizer of f when initialized at point **A** ? at point **B** ?*

Problem 12.2 (3 points). Let $M \in \mathbb{R}^{d \times d}$ be a positive semidefinite matrix, $b \in \mathbb{R}^d$ and $c \in \mathbb{R}$. We aim at minimizing the quadratic function

$$f(x) = \frac{1}{2}x^\top Mx - \langle x, b \rangle + c$$

using gradient descent. We assume that M is positive definite (i.e. all its eigenvalues are positive). We let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d > 0$ be its eigenvalues and let v_1, \dots, v_d be an orthonormal basis of \mathbb{R}^d consisting of associated eigenvectors ($Mv_i = \lambda_i v_i$ for all i). We write $L = \lambda_1$ and $\mu = \lambda_d$.

We consider standard gradient descent with constant step-size β :

$$x_{t+1} = x_t - \beta \nabla f(x_t).$$

(a) Show that f is L -smooth, μ -strongly convex and that $x^* = M^{-1}b$ is the unique minimizer of f .

(b) We now study the convergence of gradient descent to x^* . Show that for all $t \geq 0$,

$$x_{t+1} - x^* = (\text{Id} - \beta M)(x_t - x^*).$$

(c) From now, we set $\beta = 1/L$. Deduce from the previous question that for all $t \geq 0$

$$\|x_t - x^*\| \leq \left(1 - \frac{\mu}{L}\right)^t \|x_0 - x^*\|.$$

(d) We would like now to have something more precise than the error bound of the previous question. We define $w_t \stackrel{\text{def}}{=} x_t - x^*$. Let

$$\alpha_1(t) = \langle v_1, w_t \rangle, \dots, \alpha_d(t) = \langle v_d, w_t \rangle$$

be the coordinates of w_t in the orthonormal basis (v_1, \dots, v_d) . For $i \in \{1, \dots, d\}$, express $\alpha_i(t)$ in terms of t, λ_i, L and $\alpha_i(0)$.

(e) Using the previous question, justify the following sentence:

« Gradient descent converges towards the minimizer faster in directions given by the eigenvectors of the Hessian of f corresponding to large eigenvalues than in directions corresponding to eigenvectors with small eigenvalues. »

(f) Show that for all $t \geq 0$

$$\|x_t - x^*\| = \sqrt{\sum_{i=1}^d \left(1 - \frac{\lambda_i}{L}\right)^{2t} \langle v_i, x_0 - x^* \rangle^2}.$$

Problem 12.3 (3 points). In this problem, you will implement and compare gradient descent with or without momentum to minimize the Ridge cost function:

$$f(x) = \frac{1}{2}\|Ax - y\|^2 + \frac{\lambda}{2}\|x\|^2.$$

All the instructions and questions are in the Jupyter notebook `gradient_descent.ipynb`.

It is intended that you code in Python and use the provided Jupyter Notebook. Please only submit a pdf version of your notebook (right-click \rightarrow 'print' \rightarrow 'Save as pdf').

Problem 12.4 (★). We take exactly the same setting of Problem 12.2, but we now consider gradient descent with momentum: $x_1 = x_0 = 0$ (we take them equal to zero to simplify the computations) and for $t \geq 1$:

$$x_{t+1} = x_t - \beta \nabla f(x_t) + \gamma(x_t - x_{t-1})$$

where we take

$$\beta = \frac{4}{(\sqrt{L} + \sqrt{\gamma})^2} \quad \text{and} \quad \gamma = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^2.$$

Show now that the $\alpha_i(t) \stackrel{\text{def}}{=} \langle v_i, x_t - x^* \rangle$ satisfy a second order linear recurrence relation (as a sequence indexed by t). Solve this equation to obtain that for all $t \geq 0$

$$\alpha_i(t) \leq C_i \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^t$$

where C_i is a constant that does not depend on t . Deduce that for all $t \geq 0$

$$\|x_t - x^*\| \leq C \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^t$$

where C is a constant that does not depend on t .

