PROBLEM 12.2.

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

$$\begin{cases} L = \lambda_1 \text{ largest eigenvalue of } M \\ \mu = \lambda_d \text{ smallest eigenvalue of } M \end{cases}$$

(a) $H_f(x) = M$ for all $x \in \mathbb{R}^n$.

$\Rightarrow L = \lambda_1 \geq \lambda_{max}(H_f(x))$ for any $x \in \mathbb{R}^n$

$\Rightarrow f$ is L-smooth.

$\Rightarrow \mu = \lambda_d \leq \lambda_{min}(H_f(x))$ for any $x \in \mathbb{R}^n$

$\Rightarrow f$ is μ-strongly convex

Since f is strongly convex, it is strictly convex (cf HW)

$\Rightarrow$ it has a unique minizer (if a minimizer exists)

$\nabla f(x) = 0 \Rightarrow Mx - b = 0$

$\Rightarrow x = M^{-1} b$ (since $\lambda_d > 0$ M is full rank.)

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

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

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

(c) $\beta = 1/L$

By induction we get from the previous question

$$(x_t - x^*) = (Id - \frac{1}{L} M)^t (x_0 - x^*)$$

$$\| x^t - x^* \| = \| (Id - \frac{1}{L} M)^t (x_0 - x^*) \|$$

$$\leq \underbrace{\| (Id - \frac{1}{L} M)^t \|_{sp}}_{\text{spectral norm}} \| x_0 - x^* \|$$

thm 10.3(a)

$\| (Id - \frac{1}{L} M)^t \|_{sp} =$ largest eigenvalue of $(Id - \frac{1}{L} M)^t$

$$\text{Spectrum}\left((Id - \frac{1}{L} M)^t \right) = \left\{ (1 - \frac{\lambda_d}{L})^t, \dots (1 - \frac{\lambda_1}{L})^t \right\}$$

largest.

Yet $\lambda_d = \mu \Rightarrow$ results.

(d) $w_t = x_t - x^*$

$$(\overset{w_t}{x_t - x_*}) = (Id - \frac{1}{L} M)^t (\overset{w_0}{x_0 - x_*})$$

take scalar product with v_i:

$$\langle w_t, v_i \rangle = \langle \underbrace{(Id - \frac{1}{L} M)^t}_{\text{symmetric matrix}} w_0, v_i \rangle$$

$$\alpha_i(t) = \; = w_0^T (Id - \frac{1}{L} M)^t v_i$$

eigenvector of $(Id - \frac{1}{L} M)^t$ with eigenvalue

$$\left(1-\frac{\lambda_i}{L}\right)^t.$$

$$= \left(1-\frac{\lambda_i}{L}\right)^t \underbrace{w_0^T v_i}_{\alpha_i(0)}$$

$$\Rightarrow \alpha_i(t) = \left(1-\frac{\lambda_i}{L}\right)^t \alpha_i(0).$$

(e) $\alpha_i \rightarrow$ coordinate over v_i of the difference $(x_t - x^*)$

$\longrightarrow$ and the rate of convergence will depend on how big is $\frac{\lambda_i}{L}$

$\rightarrow$ the bigger lambda the faster the convergence.

(f) Using the coordinates in the bases $(v_1, \dots, v_d)$

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

$$= \sqrt{\sum_{i=1}^{d} \alpha_i(t)^2}$$

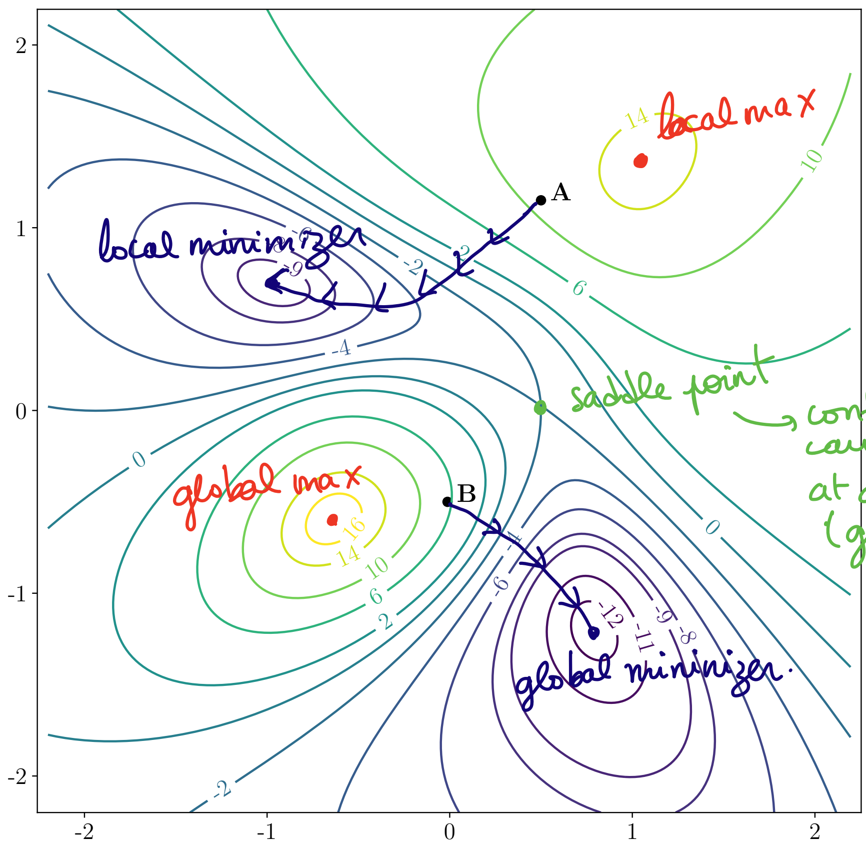
$$= \sqrt{\sum_{i=1}^{d} \left(1-\frac{\lambda_i}{L}\right)^{2t} \alpha_i(0)^2}$$

FOR 11.4 $\rightarrow$ look at last year's homework solution posted on Piazza.

Rules:

- Unless otherwise stated, all answers must be mathematically justified.
- Partial answers will be graded.
- Your submission has to be uploaded to Gradescope. In Gradescope, indicate the page on which each problem is written.
- You can work in groups but each student must write his/her/their own solution based on his/her/their 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 ($\star$) 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 (`mgabrie@nyu.edu`) or to stop at the office hours.

Problem 11.1 (2 points)**.** *The following plot shows the contour lines of a function $f : \mathbb{R}^2 \to \mathbb{R}$.*



→ contour lines can only cross at a critical point (gradients ⊥ contour lines)

(a) *Give (approximately) the coordinates of the global/local minimizers/maximizers, saddle points of f.*

(b) *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 11.2 (5 points)**.** *Let $M \in \mathbb{R}^{d\times d}$ be a positive definite 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. Since we assume that M is positive definite (i.e. all its eigenvalues are positive). We let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0$ 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 $t \geq 0$,*

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

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

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

(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*

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

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 $t \geq 0$*

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

Problem 11.3 (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 → 'print' → 'Save as pdf').

Problem 11.4 ($\star$). *We take exactly the same setting of Problem 11.2, but we now consider gradient descent with momentum:*

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

for $t \geq 1$, where we take

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

Show 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). Using this relation, show 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, but that may depend on x_0, x_1, μ and L (a precise expression of C_i is not expected). 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.

(Help: If you need to get a refresher about what is a linear recurrence, you can check this for instance: `https://www.usna.edu/Users/cs/roche/courses/f19sm242/get.php?f=slides5_8.pdf`. And you should not be afraid of seeing complex numbers showing up.)