Optimization and Computational Linear Algebra for Data Science
Solutions to the final exam

December 17, 2019

Problem 1. True or false? [WITHOUT PROOF] The 4 statements are true.

Problem 2. True or false? [WITH PROOF]

(a) False. Take for instance $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^3$. f is differentiable and $f'(x) = 3x^2$ for all $x \in \mathbb{R}$. We have $f'(0) = 0$ but 0 is not a local minimum or a local maximum of f since $f(0) = 0$ and for all $x > 0$, $f(x) > 0$ and for all $x < 0$, $f(x) < 0$.

(b) True. f is convex, hence for all $t \in [0, 1]$ we have

$$f\left(tx + (1-t)x'\right) \leq tf(x) + (1-t)f(x') = 0.$$

Taking $t = 1/2$ proves the statement.

Problem 3. **(a)** The function $f(x) = \|Ax - y\|^2$ is convex (as seen in class or in homeworks). Hence x is a solution of (1) if and only if $\nabla f(x) = 0$, i.e.

$$2A^\mathsf{T} Ax - 2A^\mathsf{T} y = 0$$

Since $A^\mathsf{T} A$ is assumed to be invertible, this equation (which is equivalent to $A^\mathsf{T} Ax = A^\mathsf{T} y$) has a unique solution $x^* = (A^\mathsf{T} A)^{-1} A^\mathsf{T} y$. We conclude that (1) has a unique solution

$$x^* = (A^\mathsf{T} A)^{-1} A^\mathsf{T} y = (A^\mathsf{T} A)^{-1} A^\mathsf{T} (Ax_0 + w) = x_0 + (A^\mathsf{T} A)^{-1} A^\mathsf{T} w.$$

(b) The rank of A is equal to the number of non-zero singular values. Here $\mathrm{rank}(A) = m$ hence $\sigma_1, \dots, \sigma_m$ are all non-zero and because they are all (by definition of singular values) all non-negative, we get they are all positive and in particular $\sigma_m > 0$.

U and V are orthogonal, hence $U^\mathsf{T} U = \mathrm{Id}_n$ and $V^\mathsf{T} V = \mathrm{Id}_m$. Using these facts we compute

$$\begin{aligned}
(A^\mathsf{T} A)^{-1} A^\mathsf{T} &= \left(V \Sigma^\mathsf{T} U^\mathsf{T} U \Sigma V^\mathsf{T}\right)^{-1} V \Sigma^\mathsf{T} U^\mathsf{T} \\
&= \left(V \Sigma^\mathsf{T} \Sigma V^\mathsf{T}\right)^{-1} V \Sigma^\mathsf{T} U^\mathsf{T} \\
&= V \,\mathrm{Diag}(\sigma_1^{-2}, \dots, \sigma_m^{-2})\, V^\mathsf{T} V \Sigma^\mathsf{T} U^\mathsf{T} \\
&= V \,\mathrm{Diag}(\sigma_1^{-2}, \dots, \sigma_m^{-2}) \Sigma^\mathsf{T} U^\mathsf{T} \\
&= V \Sigma^+ U^\mathsf{T},
\end{aligned}$$

where $\Sigma^+ \in \mathbb{R}^{m \times n}$ is given by $\Sigma^+_{i,j} = 0$ for $i \neq j$ and $\Sigma^+_{i,i} = 1/\sigma_i$. Hence we get that the singular values of $(A^\mathsf{T} A)^{-1} A^\mathsf{T} = V \Sigma^+ U^\mathsf{T}$ are $\sigma_1^{-1}, \dots, \sigma_m^{-1}$.

(c) The spectral norm of $(A^\mathsf{T} A)^{-1} A^\mathsf{T}$ is therefore equal to σ_m^{-1}. Hence

$$\|x^* - x_0\| = \|(A^\mathsf{T} A)^{-1} A^\mathsf{T} w\| \leq \|(A^\mathsf{T} A)^{-1} A^\mathsf{T}\|_{\mathrm{Sp}} \|w\| = \frac{1}{\sigma_m} \|w\|,$$

where the inequality above follows from the definition from the spectral norm (see homeworks).

Problem 4. Let $f(x, y, z) = x^2 + y^2 + z^2$ and $g(x, y, z) = x + y + z - 1$. Both function are continuously differentiable and $\nabla g(x, y, z) = (1, 1, 1) \neq 0$ for all x, y, z. Therefore, there exists some $\lambda \in \mathbb{R}$ such that the solution (x^*, y^*, z^*) of (2) verifies:

$$\nabla f(x^*, y^*, z^*) + \lambda \nabla g(x^*, y^*, z^*) = 0.$$

Since $\nabla f(x^*, y^*, z^*) = 2(x^*, y^*, z^*)$ and $\nabla g(x^*, y^*, z^*) = (1, 1, 1)$ we get

$$x^* = y^* = z^* = -\frac{\lambda}{2}.$$

Furthermore, $x^* + y^* + z^* = 1$ because (x^*, y^*, z^*) is solution of (2). This gives

$$x^* = y^* = z^* = \frac{1}{3}.$$

Problem 5. Let $v_1, \dots, v_d$ be the right singular vectors of A. The vector b_i of the first k principal components of the point a_i is given by

$$b_i = \begin{pmatrix} \langle v_1, a_i \rangle \\ \vdots \\ \langle v_k, a_i \rangle \end{pmatrix}.$$

$(v_1, \dots v_d)$ is orthonormal and is therefore an orthonormal basis of $\mathbb{R}^d$. Hence

$$\|a_i\|^2 = \sum_{j=1}^{d} \langle a_i, v_j \rangle^2 \geq \sum_{j=1}^{k} \langle a_i, v_j \rangle^2 = \|b_i\|^2.$$

Problem 6. Let $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x) = x_1^2 + 4x_2^2 - 4x_1 - 8x_2 + 1$, for $x = (x_1, x_2) \in \mathbb{R}^2$.

(a) The Hessian matrix of f is given by

$$H_f(x) = \begin{pmatrix} 2 & 0 \\ 0 & 8 \end{pmatrix},$$

for all $x \in \mathbb{R}^2$. Hence the eigenvalues of $H_f(x)$ are 2 and 8 which are non-negative: $H_f(x)$ is positive semi-definite: f is therefore convex.

$$\begin{aligned} x \text{ is a global minimizer of } f &\iff \nabla f(x) = 0 \\ &\iff \begin{pmatrix} 2x_1 - 4 \\ 8x_2 - 8 \end{pmatrix} = 0 \\ &\iff x = (2, 1). \end{aligned}$$

f has therefore one unique minimizer $x^* = (2, 1)$.

(b)

$$\begin{aligned} w(t+1) &= x(t+1) - x^* \\ &= x(t) - \alpha \nabla f(x(t)) - x^* \\ &= w(t) - \alpha \begin{pmatrix} 2x_1(t) - 4 \\ 8x_2(t) - 8 \end{pmatrix}. \end{aligned}$$

Hence

$$\begin{cases} w_1(t+1) = w_1(t) - 2\alpha(x_1(t) - 2) = w_1(t) - 2\alpha w_1(t) = (1 - 2\alpha) w_1(t) \\ w_2(t+1) = w_2(t) - 8\alpha(x_2(t) - 1) = w_2(t) - 8\alpha w_2(t) = (1 - 8\alpha) w_2(t). \end{cases}$$

(c) From the previous question we deduce that

$$w_1(t) = (1 - 2\alpha)^t w_1(0) = (1 - 2\alpha)^t (-2) \quad \text{and} \quad w_2(t) = (1 - 8\alpha)^t w_2(0) = (1 - 8\alpha)^t (-1).$$

- if $0 < \alpha < 1/4$, then $(1 - 2\alpha) \in (0, 1)$ and $(1 - 8\alpha) \in (0, 1)$. Hence $w_1(t)$ and $w_2(t)$ go to zero as $t \to \infty$ which gives that $x(t)$ converge to x^*.

- if $\alpha \geq \frac{1}{4}$, then $1 - 8\alpha \leq -1$ and therefore $w_2(t) = -(1-8\alpha)^t$ does not go to zero as $t \to \infty$: $w(t)$ does not go to zero, hence gradient descent does not converge to x^*.

Problem 7. Let P_S be the orthogonal projection onto $S = \mathrm{Im}(A^\mathsf{T})$ and let $x = P_S(x^*)$. By contradiction, assume that x^* does not belong to S, hence $x \neq x^*$. Since $x \perp (x^* - x)$, the Pythagorean theorem gives

$$\|x^*\|^2 = \|x\|^2 + \|x^* - x\|^2 > \|x\|^2$$

because $\|x^* - x\| > 0$ since $x \neq x^*$. By definition $x = P_S(x^*)$, therefore $x^* - x$ is orthogonal to $S = \mathrm{Im}(A^\mathsf{T})$ and therefore to the rows of A. This gives

$$A(x^* - x) = 0 \quad \text{thus} \quad Ax^* = Ax.$$

We conclude that

$$f(Ax) + \lambda\|x\|^2 < f(Ax^*) + \lambda\|x^*\|^2$$

which is a contradiction with the fact that x^* is a global minimizer. We conclude that $x^* \in S$.

Problem 8. Let $x \in \mathbb{R}^n$ be an eigenvector of A associated to λ: $Ax = \lambda x$. Fix $i \in \{1, \dots nn\}$ such that $|x_i| \geq |x_j|$ for all $j \in \{1, \dots, n\}$. Looking at the i^{th} coordinate of $Ax = \lambda x$ gives

$$\sum_{j=1}^{n} A_{i,j} x_j = \lambda x_i.$$

Hence,

$$|\lambda||x_i| = \left| \sum_{j=1}^{n} A_{i,j} x_j \right| \leq \sum_{j=1}^{n} A_{i,j} |x_j| \leq |x_i| \sum_{j=1}^{n} A_{i,j} \leq |x_i| d$$

because the sum of the entries of the i^{th} row of the adjacency matrix A is equal to the degree of the node i which is assumed to be less than d. Since $|x_i| \neq 0$ (otherwise $x = 0$ which is not, by definition, an eigenvector of A), we conclude that $|\lambda| \leq d$.