Optimization and Computational Linear Algebra for Data Science
Final review problems

For review exercises on linear algebra, look at last year's final review exercises (available the course's website).

Problem 0.1. *Let $A \in \mathbb{R}^{n \times m}$. Let $\sigma_1(A)$ be the largest singular value of A. Show that*

$$\sigma_1(A) = \max_{\|x\|=1} \|Ax\|.$$

Problem 0.2. *Let $A \in \mathbb{R}^{n \times m}$. Show that $A^\top A$ and $AA^\top$ have the same non-zero eigenvalues.*

Problem 0.3 (True or false?). *For each of the following statement, say if they are true or false and justify your answer.*

- *For all $A \in \mathbb{R}^{n \times n}$, if λ is an eigenvalue of A then λ^2 is an eigenvalue of A^2.*
- *For all $A \in \mathbb{R}^{n \times n}$, if σ is a singular value of A then σ^2 is a singular value of A^2.*
- *For all symmetric matrix $A \in \mathbb{R}^{n \times n}$ the eigenvalues of A are singular values of A.*

Problem 0.4. *Let $A \in \mathbb{R}^{n \times m}$. Show that for all $u \in \mathrm{Im}(A)$ and for all $v \in \mathrm{Ker}(A^\top)$ we have*

$$\langle u, v \rangle = 0.$$

Problem 0.5. *Let $A \in \mathbb{R}^{n \times m}$. Show that if A has linearly independent columns, then $A^\dagger = (A^\top A)^{-1} A^\top$.*

Problem 0.6. *Which of the following functions $f : \mathbb{R}^n \to \mathbb{R}^n$ are convex? Justify your answer*

- $f(x) = \|x\|^2$.
- $f(x) = Ax$, *for some $A \in \mathbb{R}^{n \times n}$.*
- $f(x) = \sum_{i=1}^n x_i^3$.

Problem 0.7. *Which of the following subset S of $\mathbb{R}^n$ are convex? Justify your answer*

- $S = \{x \in \mathbb{R}^n \mid \|x\|_\infty \leq 1\}$.
- $S = \{x \in \mathbb{R}^n \mid \|x\|_1 \geq 1\}$.
- $S = \{x \in \mathbb{R}^n \mid \|Ax\| < 1\}$, *for some $A \in \mathbb{R}^{n \times n}$.*

Problem 0.8. *Show that we are performing PCA on n data points $a_1, \dots, a_n \in \mathbb{R}^d$ and keep only the first $k < d$ principal components of each point. We store the dimensionally reduced dataset in a $n \times k$ matrix B, where $B_{i,j}$ is the j^{th} principal component of the point a_i. Show that the columns of B are orthogonal.*

Problem 0.9 (True of false?). *For each of the following statement, say if they are true or false and justify your answer.*

- *If a continuous function $f : \mathbb{R} \to \mathbb{R}$ has a unique minimizer then f is convex.*
- *If a continuous function $f : \mathbb{R} \to \mathbb{R}$ is such that there exists x_0 such that f is decreasing on $(-\infty, x_0]$ and increasing on $[x_0, +\infty)$ then f is convex.*
- *A twice differentiable function $f : \mathbb{R} \to \mathbb{R}$ whose derivative f' is non-decreasing is convex.*

Problem 0.10. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex, differentiable function. Assume that there exist $x, y \in \mathbb{R}^n$ such that $\nabla f(x) = \nabla f(y) = 0$. Show that $\nabla f(\frac{1}{2}(x + y)) = 0$.*

Problem 0.11. *Assume that we are doing linear regression with the least-squares cost*

$$f(x) = \|Ax - y\|^2$$

where $A \in \mathbb{R}^{n\times d}$ and $y \in \mathbb{R}^n$. Should you normalize the dataset A (that is, should we divide each column of A by its norm) to get better results (smaller training error or smaller test error on new data points)?

Suppose that we now want to use the lasso and minimize

$$f(x) = \|Ax - y\|^2 + \lambda\|x\|_1$$

for some $\lambda > 0$. Is there any reason why you might want to normalize the dataset in that case?

Problem 0.12. *Compute the critical points of the following function and say if they are local minimizers, local maximizers or saddle points.*

$$f(x, y, z) = x^2 + y^2 - z^2 \quad \textit{and} \quad g(x, y) = 3x^2 + y^2 - 6x - 4y - 10.$$

Problem 0.13. *Solve the following constrained minimization problem (find all the solutions to these problems).*

1. *Minimize $x + y + z$ subject to $e^{-x} + e^{-y} + e^{-z} = 1$.*
2. *Minimize $x^2 + y^2 + z^2$ subject to $xyz = 1$.*

Problem 0.14. *Assume that we are doing standard gradient descent to minimize the least-square cost*

$$f(x) = \|Ax - y\|^2.$$

Assume that the columns of A are linearly dependent, meaning that $\mathrm{Ker}(A) \neq \{0\}$. At which speed should gradient descent converge to the minimum? If now $\mathrm{Ker}(A) = \{0\}$, at which speed should gradient descent converge? By speed, we only ask about the dependence in t, the number of iterations, of the gap $f(x_t) - \min f$, where x_t is the position of gradient descent after t iterations.

Problem 0.15. *Let $A \in \mathbb{R}^{n\times d}$. Assume that the columns of A are linearly independent. How many steps of Newton's method do you need to minimize*

$$\|Ax - y\|^2 \ ?$$

($y \in \mathbb{R}^n$ is a fixed vector). Justify your answer.

Problem 0.16. *When running stochastic gradient descent, what are upsides and downsides of having a rapidly decaying learning rate?*

Hints. Please only look at the hints if you have spent a reasonable time thinking about the problems!

1. Use the fact that $\|Ax\|^2 = x^\mathsf{T} A^\mathsf{T} Ax$ and then use the SVD decomposition of A to rewrite $A^\mathsf{T} A$.
2. Use the SVD of A.
3. (a) True (b) False (c) False (eigenvalues can be negative but singular values can not. The singular values of a symmetric matrix are the absolute value of its eigenvalues).
4. Use the definitions of kernel and image.
5. Use the SVD decomposition of A to compute $(A^\mathsf{T} A)^{-1} A^\mathsf{T}$ and see that it corresponds to the definition of †.
6. Convex, convex, not convex.
7. Convex, not convex, convex.
8. Express the columns of B using the left-singular vectors of the matrix A whose rows are the a_i.
9. False. False. True.
10. Show that $(x + y)/2$ is a global minimizer of f.
11. Normalizing the dataset is useless for ordinary least-squares, but can be useful for Lasso.
12. Compute gradient and Hessian.
13. Use Lagrange multipliers.
14. If the columns of A are linearly dependent, then f will be L-smooth but not strongly convex, hence the speed of gradient descent will be $O(1/t)$. If the columns of A are linearly independent then you can show that $f(x)$ is μ-strongly convex and L-smooth, for some $\mu, L > 0$. Hence the error of gradient descent will be $O(e^{-\rho t})$ after t steps, for some constant $\rho > 0$.
15. 1 step.
16. See lecture notes.