Homework 4

This homework is due at 11 PM on February 16, 2023.

Submission Format: Your homework submission should consist of a single PDF file that contains all of your answers (any handwritten answers should be scanned), as well as a printout of your completed Jupyter notebook(s).

1. **PCA and low-rank compression**

   We have a data matrix $X = \begin{bmatrix} \vec{x}_1^\top \\ \vec{x}_2^\top \\ \vdots \\ \vec{x}_n^\top \end{bmatrix}$ of size $n \times d$ containing n data points[1], $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n$, with $\vec{x}_i \in \mathbb{R}^d$. Note that $\vec{x}_i^\top$ is the ith row of X. Assume that the data matrix is centered, i.e. each column of X is zero mean. In this problem, we will show equivalence between the following three problems:

   (P_1) Finding a line going through the origin that maximizes the variance of the scalar projections of the points on the line. Formally P_1 solves the problem:

   $$\underset{\vec{u} \in \mathbb{R}^d : \vec{u}^\top \vec{u} = 1}{\operatorname{argmax}} \quad \vec{u}^\top C \vec{u} \tag{1}$$

   with $C = \frac{1}{n} \sum_{i=1}^{n} \vec{x}_i \vec{x}_i^\top$ denoting the covariance matrix associated with the centered data.

   (P_2) Finding a line going through the origin that minimizes the sum of squares of the distances from the points to their vector projections. Formally P_2 solves the minimization problem:

   $$\underset{\vec{u} \in \mathbb{R}^d : \vec{u}^\top \vec{u} = 1}{\operatorname{argmin}} \sum_{i=1}^{n} \min_{v_i \in \mathbb{R}} \left\| \vec{x}_i - v_i \vec{u} \right\|_2^2 . \tag{2}$$

   Note that the vector projection of $\vec{x}$ on $\vec{u}$ is given by $v^\star \vec{u}$, where

   $$v^\star = \underset{v \in \mathbb{R}}{\operatorname{argmin}} \left\| \vec{x} - v \vec{u} \right\|_2^2 , \tag{3}$$

   and we will show that $v^\star = \langle \vec{x}, \ \vec{u} \rangle$ in part (a).

   (P_3) Finding a rank-one approximation to the data matrix. Formally P_3 solves the minimization problem:

   $$\underset{Y : \operatorname{rk}(Y) \leq 1}{\operatorname{argmin}} \ \left\| X - Y \right\|_F . \tag{4}$$

   Note that loosely speaking, two problems are said to be “equivalent" if the solution of one can be “easily" translated to the solution of the other. Some form of “easy" translations include adding/subtracting a constant or some quantity depending on the data points.

   Note the significance of these results. P_1 is finding the first principal component of X, the direction that maximizes variance of scalar projections. P_2 says that this direction also minimizes the distances between the points to their vector projections along this direction. If we view the distances as errors in approximating the

[1] Data matrices are sometimes represented as above, and sometimes as the transpose of the matrix here. Make sure you always check this, and recall that based on the definition of the data matrix, the definition of the covariance matrix also changes.

points by their projections along a line, then the error is minimized by choosing the line in the same direction as the first principal component. Finally P_3 tells us that finding a rank one matrix to best approximate the data matrix (in terms of error computed using Frobenius norm) is equivalent to finding the first principal component as well!

(a) Consider the line $\mathcal{L} = \{\vec{x}_0 \ + a\vec{u} : a \in \mathbb{R}\}$, with $\vec{x}_0 \in \mathbb{R}^d, \vec{u}^\top \vec{u} = 1$. Recall that the vector projection of a point $\vec{x} \in \mathbb{R}^d$ on to the line $\mathcal{L}$ is given by $\vec{z} = \vec{x}_0 + a^\star \vec{u}$, where $a^\star$ is given by:

$$a^\star = \underset{a}{\operatorname{argmin}} \left\| \vec{x}_0 + a\vec{u} - \vec{x} \right\|_2 . \tag{5}$$

Show that $a^\star = (\vec{x} - \vec{x}_0)^\top \vec{u}$. Use this to show that the square of the distance between x and its vector projection on $\mathcal{L}$ is given by:

$$\left\| \vec{x} - \vec{z} \right\|_2^2 = \left\| \vec{x} - \vec{x}_0 \right\|_2^2 - ((\vec{x} - \vec{x}_0)^\top \vec{u})^2. \tag{6}$$

(b) Show that P_2 is equivalent to P_1.

HINT: Start with P_2 and using the result from part (a) show that it is equivalent to P_1.

(c) Show that every matrix $Y \in \mathbb{R}^{n \times d}$ with rank at most 1, can be expressed as $Y = \vec{v}\vec{u}^\top$ for some $\vec{v} \in \mathbb{R}^n$, $\vec{u} \in \mathbb{R}^d$ and $\left\| \vec{u} \right\|_2 = 1$.

(d) Show that P_3 is equivalent to P_2.

HINT: Use the result from part (c) to show that P_3 is equivalent to:

$$\underset{\vec{u} \in \mathbb{R}^d : \vec{u}^\top \vec{u} = 1, \vec{v} \in \mathbb{R}^n}{\operatorname{argmin}} \left\| X - \vec{v}\vec{u}^\top \right\|_F^2 \tag{7}$$

Prove that this is equivalent to P_2.

2. Operator Norms

For a matrix $A \in \mathbb{R}^{m \times n}$, the *induced norm* or *operator norm* $\|A\|_p$ is defined as

$$\|A\|_p \doteq \max_{\vec{x} \neq \vec{0}} \frac{\|A\vec{x}\|_p}{\|\vec{x}\|_p}. \tag{8}$$

In this problem, we provide a characterization of the induced norm for certain values of p. Let a_{ij} denote the (i, j)-th entry of A. Prove the following:

(a) $\|A\|_1$ is the maximum absolute column sum of A,

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{m} |a_{ij}|. \tag{9}$$

HINT: Write $A\vec{x}$ as a linear combination of the columns of A to obtain $\|A\vec{x}\|_1 = \|\sum_{i=1}^{n} x_i \cdot \vec{a}_i\|_1$, where $\vec{a}_i$ denotes the i-th column of A. Then apply triangle inequality to terms within the sum.

(b) **(OPTIONAL)** $\|A\|_\infty$ is the maximum absolute row sum of A,

$$\|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}|. \tag{10}$$

HINT: First write $\|A\vec{x}\|_\infty = \max_i \left|\sum_{j=1}^{n} a_{ij} x_j\right|$. Then apply triangle inequality and use the fact that $|x_j| \leq \max_i |x_i|, \ \forall j$.

(c) $\|A\|_2 = \sigma_{\max}\{A\}$, the maximum singular value of A. *HINT: Consider connecting $\|A\|_2^2$ to a particular Rayleigh coefficient.*

3. Gradients, Jacobian matrices and Hessians

The *Gradient* of a scalar-valued function $g : \mathbb{R}^n \to \mathbb{R}$, is the column vector of length n, denoted as ∇g, containing the derivatives of components of g with respect to the variables:

$$(\nabla g(\vec{x}))_i = \frac{\partial g}{\partial x_i}(\vec{x}),\ i = 1, \ldots n.$$

The *Hessian* of a scalar-valued function $g : \mathbb{R}^n \to \mathbb{R}$, is the $n \times n$ matrix, denoted as $\nabla^2 g$, containing the second derivatives of components of g with respect to the variables:

$$(\nabla^2 g(\vec{x}))_{ij} = \frac{\partial^2 g}{\partial x_i \partial x_j}(\vec{x}),\ \ i = 1, \ldots, n,\ \ j = 1, \ldots, n.$$

The *Jacobian* of a vector-valued function $g : \mathbb{R}^n \to \mathbb{R}^m$ is the $m \times n$ matrix, denoted as Dg, containing the derivatives of components of g with respect to the variables:

$$(Dg)_{ij} = \frac{\partial g_i}{\partial x_j},\ \ i = 1, \ldots, m,\ \ j = 1, \ldots, n.$$

For the remainder of the class, we will repeatedly have to take gradients, hessians and jacobians of functions we are trying to optimize. This exercise serves as a warm up for future problems.

(a) Compute the gradients and Hessians for the following functions:

i. $g_1(\vec{x}) = \vec{x}^\top A\vec{x}$

ii. $g_2(\vec{x}) = \|A\vec{x} - b\|_2^2$

iii. $g_3(\vec{x}) = \log\left(\sum\limits_{i=1}^{m} e^{x_i}\right)$

iv. $g_4(\vec{x}) = \log\left(\sum_{i=1}^{m} e^{a_i^\top \vec{x} - b_i}\right)$

v. $g_5(\vec{x}) = e^{\|A\vec{x} - b\|_2^2}$

Consider the case now where all vectors and matrices above are scalar; do your answers above make sense? (No need to answer this in your submission)

(b) Compute the Jacobians for the following maps

i. $g(\vec{x}) = A\vec{x}$

ii. $g(\vec{x}) = f(\vec{x})\vec{x}$ where $f : \mathbb{R}^n \mapsto \mathbb{R}$ is once-differentiable

iii. $g(\vec{x}) = f(A\vec{x} + b)\vec{x}$ where $f : \mathbb{R}^n \mapsto \mathbb{R}$ is once differentiable and $A \in \mathbb{R}^{n \times n}$

(c) Plot/hand-draw the level sets of the following functions:

i. $g(x_1, x_2) = \dfrac{x_1^2}{4} + \dfrac{x_2^2}{9}$

ii. $g(x_1, x_2) = x_1 x_2$

Also point out the gradient directions in the level-set diagram. Additionally, compute the first and second order Taylor series approximation around the point $(1, 1)$ for each function and comment on how accurately they approximate the true function.

4. A quadratic-fractional function and its gradient

The goal of this problem is to compute gradients and hessians without explicitly writing out the partial derivatives. As we will see later in the class, most optimization problems can be reformulated as quadratic functionals (such as the famous least squares objective function) and it is essential that you are able to compute symbolic gradients of these without having to write out the partial derivatives.

In this problem, we consider the function $f : \mathbb{R}^n \to \mathbb{R}$, with values for $\vec{x} \in \mathbb{R}^n$ given by

$$f(\vec{x}) = \vec{z}^\top (I + \vec{x}\vec{x}^\top)^{-1} \vec{z}.$$

Here, $\vec{z} \in \mathbb{R}^n$ is given and non-zero. Note that $(I + \vec{x}\vec{x}^\top)$ is invertible for all $\vec{x}$, as you know from the shift property of eigenvalues shown in lecture.

(a) Show that,

$$f(\vec{x}) = \|\vec{z}\|_2^2 - \frac{(\vec{z}^\top \vec{x})^2}{1 + \vec{x}^\top \vec{x}}. \tag{11}$$

Hint: prove that for every $\vec{x} \in \mathbb{R}^n$, we have

$$(I + \vec{x}\vec{x}^\top)^{-1} = I - \frac{\vec{x}\vec{x}^\top}{1 + \vec{x}^\top \vec{x}}.$$

(b) Compute the gradient of f at a point $\vec{x}$. We expect a symbolic answer in terms of $\vec{x}, Q \doteq \vec{z}\vec{z}^\top$, and $\lambda(\vec{x}) \doteq \frac{\vec{x}^\top Q \vec{x}}{1+\vec{x}^\top \vec{x}}$. We don't expect answers expressed in terms of components of $\vec{x}$ and $\vec{z}$ computed by taking partial derivatives with respect to components of $\vec{x}$.

Recall the quotient rule for finding the gradient of $h(\vec{x}) = \frac{n(\vec{x})}{d(\vec{x})}$, where $n(\vec{x})$ and $d(\vec{x})$ are scalar-valued functions. We have,

$$\nabla h(\vec{x}) = \frac{d(\vec{x}) \nabla n(\vec{x}) - n(\vec{x}) \nabla d(\vec{x})}{(d(\vec{x}))^2}. \tag{12}$$

Hint: Start with the expression of $f(\vec{x})$ from equation 11 in part a) and reduce the problem to that of computing the gradient of a function of the form $h(\vec{x}) = \frac{n(\vec{x})}{d(\vec{x})}$ with $d(\vec{x}) = 1 + \vec{x}^\top \vec{x}$.

(c) Describe the set of points for which the gradient is zero.

5. **Homework Process**

   With whom did you work on this homework? List the names and SIDs of your group members.

   NOTE: If you didn't work with anyone, you can put "none" as your answer.