This homework is due at 11 pm on 23 September 2022. Self grades are due at 11 pm on 30 September 2022.

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 printouts of the completed Jupyter notebooks for this homework.

1. PCA and low-rank compression

We have a data matrix \( X = \begin{bmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_n^\top \end{bmatrix} \) of size \( n \times m \) containing \( n \) data points\(^1\), \( x_1, x_2, \ldots, x_n \), with \( x_i \in \mathbb{R}^m \).

Note that \( x_i^\top \) is the \( i \)th 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:

\[
\operatorname{argmax}_{\vec{u} \in \mathbb{R}^m; \vec{u}^\top \vec{u} = 1} \vec{u}^\top C \vec{u}
\]

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:

\[
\operatorname{argmin}_{\vec{u} \in \mathbb{R}^m; \vec{u}^\top \vec{u} = 1} \sum_{i=1}^{n} \min_{v_i \in \mathbb{R}} \| \vec{x}_i - v_i \vec{u} \|_2^2
\]

Note that the vector projection of \( \vec{x}_i \) on \( \vec{u} \) is given by \( v^* \vec{u} \), where

\[
v^* = \operatorname{argmin}_{v \in \mathbb{R}} \| \vec{x}_i - v \vec{u} \|_2^2
\]

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

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

\[
\operatorname{argmin}_{Y: \text{rk}(Y) \leq 1} \| X - Y \|_F
\]

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.

\(^1\)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.
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 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 \( L = \{ \bar{x}_0 + v\bar{u} : v \in \mathbb{R} \} \), with \( \bar{x}_0 \in \mathbb{R}^m, \bar{u}^T\bar{u} = 1 \). Recall that the vector projection of a point \( \bar{x} \in \mathbb{R}^m \) on to the line \( L \) is given by \( \bar{z} = \bar{x}_0 + v^*\bar{u} \), where \( v^* \) is given by:

\[
v^* = \arg\min_v \| \bar{x}_0 + v\bar{u} - \bar{x} \|_2
\]  

(5)

Show that \( v^* = (\bar{x} - \bar{x}_0)^T\bar{u} \). Use this to show that the square of the distance between \( x \) and its vector projection on \( L \) is given by:

\[
d^2 = \| \bar{x} - \bar{x}_0 \|^2 - ((\bar{x} - \bar{x}_0)^T\bar{u})^2
\]  

(6)

(b) Show that \( P_2 \) is equivalent to \( P_1 \).

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

(c) Show that every matrix \( Y \in \mathbb{R}^{n \times m} \) with rank at most 1, can be expressed as \( Y = \bar{v}\bar{u}^T \) for some \( \bar{v} \in \mathbb{R}^n, \bar{u} \in \mathbb{R}^m \) and \( \| \bar{u} \| = 1 \).

(d) Show that \( P_3 \) is equivalent to \( P_2 \).

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

\[
\arg\min_{\bar{u} \in \mathbb{R}^m : \bar{u}^T\bar{u} = 1, \bar{v} \in \mathbb{R}^n} \| X - \bar{v}\bar{u}^T \|^2_F
\]

(7)

\( \text{Prove that this is equivalent to } P_2. \)
2. Quadratics and Least Squares

In this question, we will see that every least squares problem can be considered as minimization of a quadratic cost function; whereas not every quadratic minimization problem corresponds to a least-squares problem. To begin with, consider the quadratic function, \( f : \mathbb{R}^2 \to \mathbb{R} \) given by:

\[
f(\vec{w}) = \vec{w}^\top A\vec{w} - 2\vec{b}^\top \vec{w} + c \tag{8}
\]

where \( A \in \mathbb{S}^n_+ \) (set of symmetric positive semidefinite matrices in \( \mathbb{R}^{2 \times 2} \)), \( \vec{b} \in \mathbb{R}^2 \) and \( c \in \mathbb{R} \).

(a) Assume \( c = 0 \), and assume that setting \( \nabla f(\vec{w}) = 0 \) allows us to find the unique minimizer. Give a concrete example of a matrix \( A \succ 0 \) and a vector \( \vec{b} \) such that the point \( \vec{w}^\ast = \begin{bmatrix} -1 & 1 \end{bmatrix}^\top \) is the unique minimizer of the quadratic function \( f(\vec{w}) \).

(b) Assume \( c = 0 \). Give a concrete example of a matrix \( A \succeq 0 \), and a vector \( \vec{b} \) such that the quadratic function \( f(\vec{w}) \) has infinitely many minimizers and all of them lie on the line \( w_1 + w_2 = 0 \).

(HINT: Take the gradient of the expression and set it to zero. What needs to be true for there to be infinitely many solutions to the equation?)

(c) Assume \( c = 0 \). Let \( \vec{w} = \begin{bmatrix} 1 & 0 \end{bmatrix}^\top \). Give a concrete example of a non-zero matrix \( A \succeq 0 \) and a vector \( \vec{b} \) such that the quadratic function \( f(\alpha \vec{w}) \) tends to \( -\infty \) as \( \alpha \to \infty \). (HINT: Use the eigenvalue decomposition to write \( A = \sigma_1 \vec{u}_1 \vec{u}_1^\top + \sigma_2 \vec{u}_2 \vec{u}_2^\top \) and express \( \vec{w} \) in the basis formed by \( \vec{u}_1, \vec{u}_2 \).)

(d) Say that we have the data set \( \{(\vec{x}_i, y_i)\}_{i=1}^n \) of data points \( \vec{x}_i \in \mathbb{R}^d \) and values \( y_i \in \mathbb{R} \). Define \( X = \begin{bmatrix} \vec{x}_1 & \ldots & \vec{x}_n \end{bmatrix}^\top \) and \( \vec{y} = \begin{bmatrix} y_1 & \ldots & y_n \end{bmatrix} \). In terms of \( X \) and \( \vec{y} \), find a matrix \( A \), a vector \( \vec{b} \in \mathbb{R}^d \) and a scalar \( c \), so that we can express the sum of the square losses \( \sum_{i=1}^n (\vec{w}^\top \vec{x}_i - y_i)^2 \) as the quadratic function \( f(\vec{w}) = \vec{w}^\top A\vec{w} - 2\vec{b}^\top \vec{w} + c \).

(e) Here are three statements with regards to the minimization of a quadratic loss function:

i. It can have a unique minimizer.

ii. It can have infinitely many minimizers.

iii. It can be unbounded from below, i.e. there is some direction, \( \vec{w} \) so that \( f(\alpha \vec{w}) \) goes to \( -\infty \) as \( \alpha \to \infty \).

All three statements apply to general minimization of a quadratic cost function. Parts (a), (b) and (c) give concrete examples of quadratic cost functions where (i), (ii) and (iii) apply respectively. However, notice that statement (iii) cannot apply to the least squares problem as the objective is always positive. The least-squares problem can have infinitely many minimizers though. How? Consider the gradient of the least squares problem in part (d) at an optimal solution \( \vec{w}^\ast \):

\[
\nabla f(\vec{w}^\ast) = 2X^\top X\vec{w}^\ast - 2\vec{b} = 0. \tag{9}
\]

Therefore, the least squares problem only has multiple solutions if \( X^\top X \) is not full rank. This means that \( \text{rk}(X^\top X) = \text{rk}(X) < d \). Finally, the rank of \( X \) is less than \( d \) when the data points \( \{\vec{x}_i\}_{i=1}^n \) do not span \( \mathbb{R}^d \). This can happen when the number of data points \( n \) is less than \( d \) or when \( \{\vec{f}_i\}_{i=1}^d \) are linearly dependent where \( \vec{f}_i \) are the columns of \( X \), i.e., the features.

Indicate below that you have read and understood the discussion above.
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 $\nabla g$, containing the derivatives of components of $g$ with respect to the variables:

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

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:

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

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. $g(x) = x^\top Ax$

ii. $g(x) = \|Ax - b\|_2^2$

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. $g(x) = Ax$

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

iii. $g(x) = f(Ax + b)x$ 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. $g(x_1, x_2) = \frac{x_1^2}{4} + \frac{x_2^2}{9}$

ii. $g(x_1, x_2) = x_1x_2$

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

\[ f(x) = z^\top (I + xx^\top)^{-1} z. \]

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

(a) Show that,

\[ f(x) = \|z\|_2^2 - \frac{(z^\top x)^2}{1 + x^\top x}. \]  

**Hint:** prove that for every \( x \in \mathbb{R}^n \), we have

\[ (I + xx^\top)^{-1} = I - \frac{xx^\top}{1 + x^\top x}. \]

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

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

\[ \nabla h(x) = \frac{d(x)\nabla n(x) - n(x)\nabla d(x)}{(d(x))^2}. \]  

**Hint:** Start with the expression of \( f(x) \) from equation 10 in part a) and reduce the problem to that of computing the gradient of a function of the form \( h(x) = \frac{n(x)}{d(x)} \) with \( d(x) = 1 + x^\top x \).

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

(d) Let \( Q = zz^\top \). Compute the Hessian of \( f \) at a point \( x \). Again, we expect a symbolic answer in terms of \( Q \) and \( x \). Don’t consider partial derivatives with respect to components of \( x \).

**Hints:**

i. Let \( g : \mathbb{R}^n \to \mathbb{R}^n \) denote the gradient map, \( g(x) = \nabla f(x) \). First find the Jacobian of \( g \) denoted as \( Dg(x) \).

Recall the product rule for finding the Jacobian of \( g(x) = v(x)s(x) \), when \( g(x) \) is a product of a vector-valued function \( v(x) \) and a scalar valued function \( s(x) \). We have,

\[ Dg(x) = Dv(x)s(x) + v(x)Ds(x). \]

Note that the first term is a product of a matrix and a scalar while the second term is the product of a column vector with a row vector.

ii. Recall that the Hessian of \( f \) is related to the Jacobian of the gradient of \( f \) as \( \nabla^2 f(x) = (Dg(x))^\top \). Additionally if \( f \) is twice differentiable, the Hessian will be symmetric.
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.