EECS 127/227AT Optimization Models in Engineering
Spring 2020 Homework 4

This homework is due Friday, February 21, 2020 at 23:00 (11pm).
Self grades are due Friday, February 28, 2020 at 23:00 (11pm).

This version was compiled on 2020-02-15 07:51.

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 your IPython notebook with solutions saved as a PDF.

1. (Practice) Matrix norm

The Frobenius norm of a matrix A is defined as

$$\|A\|_F = \sqrt{\langle A, A\rangle}$$

where for two matrices $A, B \in \mathbb{R}^{m\times n}$, the canonical inner product defined over this space is $\langle A, B\rangle := \mathrm{Tr}(A^\top B) = \sum_{ij} A_{ij}B_{ij}$. The previous definition of the inner product is equivalent to interpreting the matrices A and B as vectors of length n^2 and taking the vector inner product of the respective n^2-dimensional vectors. The Cauchy-Schwarz inequality for the inner product follows in a straightforward way from the Cauchy-Schwarz inequality for vectors:

$$\langle A, B\rangle = \sum_{\substack{1\le i\le m\\ 1\le j\le n}} A_{ij}B_{ij} \le \left(\sum_{\substack{1\le i\le m\\ 1\le j\le n}} A_{ij}^2\right)^{1/2} \left(\sum_{\substack{1\le i\le m\\ 1\le j\le n}} B_{ij}^2\right)^{1/2} = \|A\|_F \|B\|_F .$$

(a) Show that the Frobenius norm satisfies all three properties of a norm.

(b) Write the Frobenius norm squared in terms of singular values.

(c) Express the Frobenius norm squared in terms of the ℓ_2-norm of the columns of A with $\vec{a}_i$ denoting column i. Concretely, prove $\|A\|_F^2 = \sum_{i=1}^n \|\vec{a}_i\|^2$ where $\vec{a}_i$ are the columns of A.

(d) A generalization of the least squares problem is to find a *matrix* X that most closely solves the problem $AX = B$. This is sometimes called the *matrix least squares problem*, and when X and B are vectors, reduces to the ordinary least squares problem you are familiar with. Formally, we can define the problem given $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{m\times k}$:

$$\min_{X\in\mathbb{R}^{n\times k}} \|AX - B\|_F$$

However, at this point we only know how to solve the vector least squares problem. Reformulate the above objective in terms of vector least squares problems that way we can solve it.

Hint: The result derived in part (iii) may be particularly useful in addition to the following fact:

$$\min_{\vec{a}_1,\vec{a}_2,\cdots\in\mathbb{R}^n} \|\vec{a}_1\|_2^2 + \|\vec{a}_2\|_2^2 + \cdots = \min_{\vec{a}_1\in\mathbb{R}^n} \|\vec{a}_1\|_2^2 + \min_{\vec{a}_2\in\mathbb{R}^n} \|\vec{a}_2\|_2^2 + \ldots$$

2. 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 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:

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

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:

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

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

$$\underset{Y:\operatorname{rank}(Y)\leq 1}{\operatorname{argmin}} \quad \|X - Y\|_F \tag{3}$$

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

$$v^* = \underset{v}{\operatorname{argmin}} \|\vec{x}_0 + v\vec{u} - \vec{x}\|_2$$

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

$$d^2 = \|\vec{x} - \vec{x}_0\|_2^2 - ((\vec{x} - \vec{x}_0)^\top\vec{u})^2$$

[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.

(b) Show that P_2 is equivalent to P_1.
Hint: Start with equation (2) and using the result from part (a) show that it is equivalent to equation (1).

(c) Show that every matrix $Y \in \mathbb{R}^{n\times m}$ with rank at most 1, can be expressed as $Y = \vec{v}\vec{u}^\top$ for some $\vec{v} \in \mathbb{R}^n$, $\vec{u} \in \mathbb{R}^m$ and $\|\vec{u}\| = 1$.
Hint: Use the SVD.

(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:

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

Prove that this is equivalent to equation (2).

3. 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:

$$f(\vec{w}) = \vec{w}^\top A\vec{w} - 2\vec{b}^\top\vec{w} + c$$

where $A \in \mathbb{S}_+^2$ (set of symmetric positive semidefinite matrices in $\mathbb{R}^{2\times2}$), $\vec{b} \in \mathbb{R}^2$ 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 $\vec{w}^* = \begin{bmatrix}-1 & 1\end{bmatrix}^\top$ 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 $\{(\vec{x}^{(i)}, y^{(i)})\}_{i=1,\dots,n}$ of features $\vec{x}^{(i)} \in \mathbb{R}^d$ and values $y^{(i)} \in \mathbb{R}$. Define $X = \begin{bmatrix}\vec{x}^{(1)} & \dots & \vec{x}^{(n)}\end{bmatrix}^\top$ and $\vec{y} = \begin{bmatrix}y^{(1)} & \dots & y^{(n)}\end{bmatrix}^\top$. 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 $\vec{w}^*$:

$$\nabla f(\vec{w}^*) = 2X^\top X\vec{w}^* - 2\vec{b} = 0.$$

Therefore, the least squares problem only has multiple solutions if $X^\top X$ is not full rank. This means that $\text{rank}(X^\top X) = \text{rank}(X) < d$. 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 $\{\vec{z}_i\}_{i=1}^d$ are linearly dependent where $\vec{z}_i$ are the columns of X.

To complete this subpart of the question, make sure you understand the discussion above and indicate that in the solution.

4. Proof of the Eckart-Young Theorem

Given a matrix $A \in \mathbb{R}^{m\times n}$ with singular value decomposition $A = U\Sigma V^\top$, define the matrix $A_k = \sum_{i=1}^k \sigma_i \vec{u}_i \vec{v}_i^\top$ where $\vec{u}$ and $\vec{v}$ denote the ith left and right singular vectors of A and σ_i denotes the ith singular value. Recall that the Eckart-Young Theorem states that:

$$A_k = \underset{\substack{B\in\mathbb{R}^{m\times n}\\ \text{rank}(B)\le k}}{\text{argmin}} \|A - B\|_2 \qquad \text{Spectral Norm Approximation}$$

$$A_k = \underset{\substack{B\in\mathbb{R}^{m\times n}\\ \text{rank}(B)\le k}}{\text{argmin}} \|A - B\|_F \qquad \text{Frobenius Norm Approximation.}$$

That is, the matrix A_k is the best rank-k approximation of A in both the Spectral and Frobenius norms. In the question, we will prove the Eckart-Young Theorem.

(a) Prove the Spectral Norm Approximation result from the Eckart-Young Theorem

(b) Prove the Frobenius Norm Approximation result from the Eckart-Young Theorem

5. (Practice) Ridge Regression

Prove that the optimal solution to the ridge regression problem:

$$\min_{\vec{w}\in\mathbb{R}^p} \|X\vec{w} - \vec{y}\|_2^2 + \lambda \|\vec{w}\|_2^2,$$

where $X \in \mathbb{R}^{n\times p}$, $\lambda > 0$ and $\vec{y} \in \mathbb{R}^n$, is given by:

$$\vec{w}^* = (X^\top X + \lambda I)^{-1} X^\top \vec{y}.$$

6. Variants on least sqaures: a playground

In this problem we will explore four different types of regression: Ordinary least squares, ridge regression, weighted least squares and Tikhonov regularization. This problem has an associated ipython notebook 'regression_playground.ipynb'. You will write lines of code and answer questions asked (marked with TODO) in the notebook itself.

7. Homework process

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