EECS 127/227AT Optimization Models in Engineering
Spring 2019 Homework 3

Due date: 2/21/19, 23:00 (11 pm). Please LaTeX or handwrite your homework solution and submit an electronic version. Self grades are due at 2/28/19 at 23:00 (11 pm).

1. **Interpreting the data matrix** In several areas such as machine learning, statistics and data analysis you come across a data matrix X. Sometimes this matrix has dimensions $\mathbb{R}^{m\times n}$ while other times it has dimensions $\mathbb{R}^{n\times m}$ and it can get really confusing as to what exactly it represents. In this problem, we describe a way of interpreting the data matrix.

   First, what exactly is a data matrix? As the name suggests, it is a collection of data points. Suppose you are collecting data about courses offered in EECS department in Fall 2018. Each course has certain attributes or features that you are interested in. Possible examples of features are the number of students in the course, the number of GSIs in the course,the number of units the course is worth, the size of the classroom that the course was taught in, the difficulty rating of the course in numerical (1-5) scale and so on. Suppose there were a total of 20 courses and for each course we have 10 features. Then we have 20 data points, with each data point being a 10-dimensional vector. We can arrange this in a matrix of size 20×10, where each row corresponds to values of different features for the same point, and each column corresponds to values of same feature for different points.

   Generalizing this, suppose we have n data points with each point containing values for m features (i.e each point lies in m-dimensional space) then our data matrix X would be of size $n \times m$, i.e. $X \in \mathbb{R}^{n\times m}$. We can interpret X in the following two ways:

   (a) $X = \begin{bmatrix} \leftarrow x_1^\top \rightarrow \\ \leftarrow x_2^\top \rightarrow \\ \vdots \\ \leftarrow x_n^\top \rightarrow \end{bmatrix}$.

   Here $x_i \in \mathbb{R}^m,\ i = 1, 2, \ldots, n$, and $x_i^\top$ is a row vector that contains values of different features for the ith data point.

   (b) $X = \begin{bmatrix} \uparrow & \uparrow & \ldots & \uparrow \\ x^1 & x^2 & \ldots & x^m \\ \downarrow & \downarrow & \ldots & \downarrow \end{bmatrix}$.

   Here $x^j \in \mathbb{R}^n, j = 1, 2, \ldots, m$ and x^j is a column vector that contains values of the the jth feature for different data points. Note that in several places you will encounter the case where the columns are referred to as $x_1, x_2, \ldots$ instead but it is important to understand the context and be clear on what the column represents.

   Consider the matrix X as describe above. We explore how we can manipulate the data matrix to get some desirable properties.

   (a) Suppose we want to compute a vector that contains the mean value for each feature. What is the length of the vector containing mean value of the features? Which of the following python commands will give us the mean value of the features:

i. feature_means = numpy.mean(X, axis = 0)
ii. feature_means = numpy.mean(X, axis = 1)

(b) Suppose we want to compute the standard deviation of each feature. What is the dimension of the vector containing standard deviation of the features? Which of the following python commands will give us the standard deviation of the features:

i. feature_stddevs = numpy.std(X, axis = 0)
ii. feature_stddevs = numpy.std(X, axis = 1)

(c) Suppose we want every feature *"centered"*, i.e. the feature is zero mean. How would you achieve this?

(d) Suppose we want every feature *"standardized"*, i.e the feature is zero mean and has unit variance. How would you achieve this?

(e) Another metric of interest is the covariance matrix, which tells us how different features are related to each other. What is the size of the covariance matrix?:

i. $n \times n$
ii. $m \times m$.

Hint: Is the number of features m or n?

(f) For rest of the problem assume that the data matrix is centred so every feature is zero mean. Let C denote the covariance matrix. Show that C can be represented in the following ways:

$$C = \frac{X^\top X}{n}$$
$$C = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top.$$

Recall that $x_i^\top$ is the ith row of X.

(g) Let the (i, j) entry of C (c_{ij}) denote the covariance between feature i and feature j. Then which of the following is true?

i. $c_{ij} = \frac{1}{m}(x^i)^\top x^j$
ii. $c_{ij} = \frac{1}{n}(x^i)^\top x^j$.

Recall that x^i is the ith column of X.

(h) Recall that our data points are the rows of X and these lie in a $m-$ dimensional space. Suppose we are interested in taking projection of the points on a one-dimensional subspace in $\mathbb{R}^m$ spanned by the unit vector u. Sometimes this is referred to informally as "Projecting points along direction u". Then which of the following is true:

i. $u \in \mathbb{R}^m$
ii. $u \in \mathbb{R}^n$

Hint: Think about how many points we have and what dimension a single point lies in.

(i) Note there are three different interpretations of the term "projection" and these are used interchangeably with abuse of notation which can make it confusing at times.
Consider vectors a and b in $\mathbb{R}^n$. Let b be unit norm (i.e $b^\top b = 1$). Then we have:

i. The **vector projection** of a on b is given by $(a^\top b)b$. Note that is a vector in $\mathbb{R}^n$.
ii. The **scalar projection** of a on b is given $a^\top b$. This is a scalar but can take both positive and negative values.

iii. The **projection length** of a on b is given by $|a^\top b|$, and is the absolute value of the scalar projection.

Recall that our data points are the rows of X. Suppose we want to obtain a column vector, $z \in \mathbb{R}^n$ containing scalar projections of points along the direction given by the unit vector u. Show that this is given by,

$$z = Xu.$$

(j) Now we are interested in calculating empirical variance of the scalar projections along direction u. Show that this can be calculated as,

$$Var(z) = \frac{1}{n} u^\top X^\top X u = u^\top C u.$$

Hint: What is the empirical mean of z, μ_z? Recall that X is assumed to be centered.

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, $x_1, x_2, \ldots, x_n$, 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 the 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 $P1$ solves the problem:

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

with $C = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top$ 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:

$$\underset{u \in \mathbb{R}^m \,:\, u^T u = 1}{\operatorname{argmin}} \sum_{i=1}^{n} \min_{v_i \in \mathbb{R}} \|x_i - v_i u\|_2^2. \tag{2}$$

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

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

Note that loosely speaking, two problems are said to be "equivalent" if solution of one can be "easily" translated to the solution of 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 line $\mathcal{L} = \{x_0 + vu : v \in \mathbb{R}\}$, with $x_0 \in \mathbb{R}^m, u^\top u = 1$. Recall that the vector projection of a point $x \in \mathbb{R}^m$ on to the line $\mathcal{L}$ is given by $z = x_0 + v^* u$, where v^* is given by,

$$v^* = \underset{v}{\operatorname{argmin}} \ \|x_0 + vu - x\|_2.$$

Show that $v^* = (x - x_0)^\top u$. Use this to show that the square of the distance between x and its vector projection on $\mathcal{L}$ is given by,

$$d^2 = \|x - x_0\|_2^2 - ((x - x_0)^\top u)^2.$$

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

(c) Show that every rank one matrix $Y \in \mathbb{R}^{n \times m}$ can be expressed as $Y = vu^\top$ for some $v \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$. *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:

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

Prove that this is equivalent to equation (2).

3. SVD Transformation Recall that a matrix can be viewed as a linear operator. In this problem we will interpret the linear map corresponding to a matrix $A \in \mathbb{R}^{n \times n}$ by looking at its singular value decomposition, $A = UDV^\top$. Recall that here $U, D, V \in \mathbb{R}^{n \times n}$ and U, V are unitary matrices while D is a diagonal matrix. We will first look at how $V^\top, D$ and U each separately transform the unit circle $C = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}$ and then look at their effect as a whole. This problem has an associated jupyter notebook, "svd_transformation.ipynb" that contains several parts (b,c,d,e) of the problem. These subparts can be answered in the notebook itself in the space provided and can be submitted as a pdf using the 'Download as pdf' feature that jupyter notebook supports.

(a) Multiplying by $V^\top$ can be interpreted as a change of basis. Show that $V^\top x$ represents x in the basis defined by the columns of V.

For rest of the problem we restrict ourselves to the case where $A \in \mathbb{R}^{2 \times 2}$ and move to the jupyter notebook.

4. PCA and Senate Voting Data In this problem, we look at Senate voting data. The data is contained in a $n \times m$ data matrix X, where each row corresponds to a Senator, and each column to a bill. Each entry of X is either $1, -1$ or 0 depending on whether the senator voted for, against or abstained.

(a) We want to assign a score to each senator based on their voting pattern. For this let us pick $a \in \mathbb{R}^m$, and a scalar b and define the score for senator i as:

$$f(x_i, a, b) = x_i^\top a + b, \quad i = 1, 2, \ldots, n.$$

Note that in our notation the rows of X are $x_i^\top$ and thus each $x_i^\top$ is a row vector of length m.

Let us denote by $f(X, a, b)$, the column vector of length n, obtained by stacking the scores for each senator. Then,

$$z = f(X, a, b) = Xa + b\mathbf{1} \in \mathbb{R}^n$$

where $\mathbf{1}$ is a vector with all entries equal to 1. Let us denote the mean value of z by $\mu_z = \frac{1}{n}\mathbf{1}^\top z$. Further let $\mu_x^\top$ denote the row vector corresponding to mean of each column of X. Then

$$\begin{aligned}\mu_z &= \frac{1}{n}\sum_{i=1}^{n} f(x_i, a, b) \\ &= a^\top \mu_x + b.\end{aligned}$$

Then the empirical variance of the scores can be obtained as:

$$\begin{aligned}Var(f(X, a, b)) &= Var(z) \\ &= \frac{1}{n}(z - \mu_z\mathbf{1})^\top(z - \mu_z\mathbf{1}) \\ &= \frac{1}{n}(Xa + b\mathbf{1} - a^\top\mu_x\mathbf{1} - b\mathbf{1})^\top(Xa + b\mathbf{1} - a^\top\mu_x\mathbf{1} - b\mathbf{1}) \\ &= \frac{1}{n}(Xa - \mathbf{1}\mu_x^\top a)^\top(Xa - \mathbf{1}\mu_x^\top a) \\ &= \frac{1}{n}a^\top(X - \mathbf{1}\mu_x^\top)^\top(X - \mathbf{1}\mu_x^\top)a\end{aligned}$$

We observe that the variance of score functions depends on "centered data" and does not depend on b.

For the remainder of the problem we assume that the data has been "centered" (i.e mean of each column of X is zero) and set $b = 0$ so that $\mu_z = 0$.
This leads us to the simpler formula,

$$Var(f(X, a)) = \frac{1}{n}a^\top X^\top Xa.$$

Suppose we restrict a to have unit-norm. Then,find a that maximizes $Var(f(X, a))$. What is the value of the maximum variance?

(b) From the previous part what can you say about how senators vote as compared to their party average? Compute the variance of the scores and comment on the projections along a for the following two cases:

i. a = a_mean_red, the average of rows of X corresponding to 'Red' senators.

ii. a = a_mean_blue, the average of rows of X corresponding to 'Blue' senators.

(c) Recall from the first question of this homework on interpreting the data matrix that we can compute the variance of scalar projections, z, of the the data points (rows of X) along a unit direction u as,

$$Var(z) = u^\top Cu,$$

where $C = \frac{X^\top X}{n}$. Can you express the variance along the first two principal components of PCA, a_1, a_2 in terms of eigenvalues of C? What is the sum of variance along a_1 and a_2. Plot the data projected on the plane spanned by a_1 and a_2.

(d) Suppose we want to find the bills that are most/least contentious. That is bills for which there is most variability in voting pattern among senators and bills for which the voting is almost unanimous. We can do this in the following two ways:

i. For each basis vector associated with a bill (That is vector of all zeros with 1 in the entry corresponding to the bill), we can compute score using the basis vector as a, and look at variance of the score. This is equivalent to looking at variance of columns of X.

ii. We can look at those bills corresponding to highest and lowest absolute values in the first principal component. This is equivalent to taking inner-products of the first principal component with each of the basis vectors corresponding to the bill and picking the ones with highest/lowest absolute value of inner-products.

Fill in the code in the Jupyter notebook and comment on your observations in the space provided in the notebook.

(e) Finally we can classify senators as most/least "extreme" based on the absolute value of their scores along the first principal component. Comment on your observations in the space provided in the Jupyter notebook.

5. Matrix Norms

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

$$\|A\|_p := \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}.$$

Note that for our purposes, we use max instead of sup.

(a) First we show an equivalent way of defining the induced norm. Prove that,

$$\|A\|_p = \max_{\|x\|_p = 1} \|Ax\|_p.$$

(b) Next 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:

i. $\|A\|_1$ is the maximum absolute column sum of A,

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

ii. $\|A\|_\infty$ is the maximum absolute row sum of A,

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

iii. $\|A\|_2 = \sigma_{\max}(A)$, the maximum singular value of A.

(c) Prove that the induced matrix norm satisfy the sub-multiplicative property,

$$\|AB\|_p \leq \|A\|_p \|B\|_p ,$$

where A and B are appropriately sized matrices. If we view matrices as linear maps acting on input then the sub-multiplicative property allows us to prove properties of composition of these maps.
Hint: First prove that for a matrix A and vector v, we have $\|Av\|_p \leq \|A\|_v \|v\|_p$.

(d) Recall for a vector $x \in \mathbb{R}^n$, we had equivalence of different norms, i.e we had the following inequalities relating different norms,

$$\frac{1}{\sqrt{n}}\|x\|_2 \leq \|x\|_\infty \leq \|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2 \leq n\|x\|_\infty.$$

Similarly we can prove equivalence of norms for induced matrix norms for $A \in \mathbb{R}^{m\times n}$. This gives us the following set of inequalities:

$$\frac{1}{\sqrt{n}} \|A\|_\infty \leq \|A\|_2 \leq \sqrt{m} \|A\|_\infty$$
$$\frac{1}{\sqrt{m}} \|A\|_1 \leq \|A\|_2 \leq \sqrt{n} \|A\|_1 .$$

Equivalence of norms is a useful property to have since some norms are easier to find bounds for than others. For instance the $\|A\|_1$ is much easier to compute than $\|A\|_2$ and using $\|A\|_1$ we can find bounds for $\|A\|_2$. Using the inequalities, find a and b such that,

$$a \|A\|_1 \leq \|A\|_\infty \leq b \|A\|_1 .$$

(e) Another useful norm is the Frobenius norm defined as,

$$\|A\|_F = \sqrt{\text{trace}(A^\top A)}.$$

Show that we have equivalence between the Frobenius norm and induced 2-norm by proving that,

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{r} \|A\|_2 ,$$

where r is the rank of A.

6. Connected Graphs and Laplacians

We are given a graph as a set of vertices in $V = \{1, \ldots, n\}$, with an edge joining any pair of vertices in a set $E \subseteq V \times V$. We assume that the graph is undirected (without arrows), meaning that $(i, j) \in E$ implies $(j, i) \in E$. We define the Laplacian matrix by

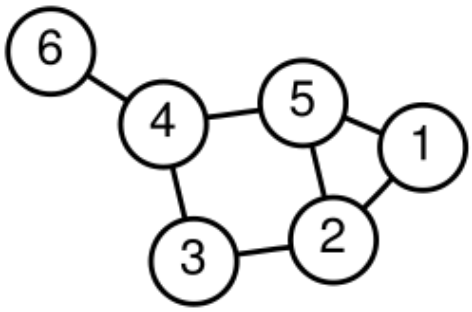

Figure 1: Example of an undirected graph.

$$L_{ij} = \begin{cases} -1 & \text{if } (i,j) \in E, \\ d(i) & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

Here, $d(i)$ is the number of edges adjacent to vertex i. For example, $d(4) = 3$ and $d(6) = 1$ for the graph in Figure 1.

(a) Form the Laplacian for the graph shown in Figure 1.

(b) Turning to a generic graph, show that the Laplacian L is symmetric.

(c) Show that L is positive-semidefinite, proving the following identity, valid for any $u \in \mathbb{R}^n$:

$$u^\top L u = q(u) \doteq \frac{1}{2} \sum_{(i,j)\in E} (u_i - u_j)^2.$$

Hint: find the values $q(e_k)$, $q(e_k \pm e_l)$, for two unit vectors e_k, e_l such that $(k,l) \in E$.

(d) Show that 0 is always an eigenvalue of L, and exhibit an eigenvector.

(e) The graph is said to be connected if there is a path joining any pair of vertices. Show that if the graph is connected, then the zero eigenvalue is simple, that is, the dimension of the nullspace of L is 1. *Hint:* prove that if $u^\top L u = 0$, then $u_i = u_j$ for every pair $(i,j) \in E$.