EECS 127/227AT **Fall 2020**
Homework 3
Homework 3 is due on Gradescope by Friday 9/25 at 11.59 p.m.

1 Ridge regression with bounded regressand noise

Consider the ridge regression problem in the case where our regressand measurements y are noisy and we have some bounds on this noise, as well as some specific knowledge about data matrix A.

Let the square matrix $A \in \mathbb{R}^{n\times n}$ have the singular value decomposition $A = U\Sigma V^\top$, and assume that all its singular values are strictly positive. We write $\sigma_{\min}(A)$ for its smallest singular value.

(a) Is A invertible? If so, write the singular value decomposition of A^{-1}.

(b) Suppose there is a ground truth vector $y \in \mathbb{R}^n$, but we only have access to a noisy measurement $\tilde{y} \in \mathbb{R}^n$ satisfying

$$\|\tilde{y} - y\|_2 \leq r,$$

for some $r > 0$. We wish to bound the distance between the solutions of the noisy regression $Ax = \tilde{y}$ and the true regression $Ax = y$.

Let $x^*(y)$ denote the solution of $Ax = y$. Show that

$$\max_{\tilde{y}:\|\tilde{y}-y\|_2\leq r} \|x^*(\tilde{y}) - x^*(y)\|_2 = \frac{r}{\sigma_{\min}(A)}$$

(c) What happens if the smallest singular value of A is very close to zero? Why is this problematic for finding our solution vector x^*?

(d) Now assume that we find the optimal value x^* via ridge regression, i.e., we compute

$$x^*_\lambda(\tilde{y}) = \arg\min_x \|Ax - \tilde{y}\|_2^2 + \lambda\|x\|_2^2,$$

for some chosen value $\lambda \in \mathbb{R}$. Compute $x^*_\lambda(\tilde{y})$, our optimal solution vector (now parameterized by λ), by solving this optimization problem.

(e) Show that for all $\lambda > 0$ we have

$$\max_{\tilde{y}:\|\tilde{y}-y\|_2\leq r} \|x^*_\lambda(\tilde{y}) - x^*_\lambda(y)\|_2 \leq \frac{r}{2\sqrt{\lambda}}.$$

How does the value of λ affect the sensitivity of your solution $x^*_\lambda(y)$ to noise in y?

Hint: For every $\lambda > 0$, we have

$$\max_{\sigma>0} \frac{\sigma}{\sigma^2 + \lambda} = \frac{1}{2\sqrt{\lambda}}.$$

(You need not show this; this optimization problem can be solved by setting the derivative of the objective function to 0 and solving for σ.)

2 Regression Playground

For this problem you will be implementing various types of regression techniques in this notebook and answering some questions along the way.
First, please run through the setup section in the notebook.

(a) Please implement Ordinary Least Squares in the notebook.

(b) i. Show that the optimal solution to the ridge regression problem:

$$\min_{w \in \mathbb{R}^m} \|Xw - y\|_2^2 + \lambda \|w\|_2^2,$$

where $X \in \mathbb{R}^{n \times m}$, $\lambda > 0$ and $y \in \mathbb{R}^n$, is given by: $w^* = (X^T X + \lambda I)^{-1} X^T y$.
Note: This part of the problem is identical to Problem 1(d) in slightly different notation, so you can just appeal to your solution for that part if you wish.

ii. Implement Ridge Regression in the jupyter notebook.

(c) i. Implement Weighted Least Squares in the Jupyter notebook. What is the intuition behind it?

ii. Comment on the difference between WLS and OLS w/ Bad Data: which one performed better and why?

(d) Implement Tikhonov regularization.

3 More fun with the SVD: The Eckart-Young-Mirsky Theorem

Consider a matrix $A \in \mathbb{R}^{m \times n}$ with singular value decomposition $A = U\Sigma V^\top$. We will think of the matrix as having $\min(m,n)$ singular values of which $(\min(m,n) - \text{rank}(A))$ are zero. The singular values of the matrix will therefore be $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(m,n)} \geq 0$, where, if $r := \text{rank}(A)$, we have $\sigma_1, \ldots, \sigma_r$ being strictly positive and $\sigma_{r+1}, \ldots, \sigma_{\min(m,n)}$ being zero. The column vectors $u_1, \ldots, u_{\min(m,n)}$ (i.e. the first $\min(m,n)$ columns of the $m \times m$ orthogonal matrix U) are called the left singular vectors and the column vectors $v_1, \ldots, v_{\min(m,n)}$ (i.e. the first $\min(m.n)$ columns of the $n \times n$ orthogonal matrix V) are called the right singular vectors. Note that here Σ is an $m \times n$ matrix. You can check that $A = \sum_{i=1}^{\min(m,n)} \sigma_i u_i v_i^T$.

For $0 \leq k \leq \min(m,n)$, define the matrix $A_k := \sum_{i=1}^{k} \sigma_i u_i v_i^\top$. The Eckart-Young-Mirsky theorem comes in two parts, one for the spectral norm and one for the Frobenius norm states, and states that

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

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

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

(a) Prove the spectral norm approximation part of the Eckart-Young-Mirksy theorem.
Hint: First try to see what $\|A - A_k\|_2$ simplifies to, after you have done this you should show that for any arbitrary matrix $B \in R^{m \times n}$, of rank k, we have $\|A - B\|_2 \geq \|A - A_k\|_2$. Then justify why this proves the theorem.

(b) **Optional**: Prove the Frobenius norm approximation part of the Eckart-Young-Mirsky theorem.

Hint: For any matrix, $C \in \mathbb{R}^{m \times n}$, we have:

$$\sigma_r(C) = \|C - C_{r-1}\|_2,$$

for all $1 \leq r \leq \min(m,n)$, where $C_i := \sum_{j=1}^{i} \sigma_j(C) u_j(C) v_j(C)^T$ for $0 \leq i \leq \min(m,n)$. Here $\sigma_1(C) \geq \sigma_2(C) \geq \ldots \geq \sigma_{\min(m,n)}(C)$ are the singular values of C, while $u_1(C), \ldots, u_{\min(m.n)}(C)$ are the corresponding left singular vectors and $v_1(C), \ldots, v_{\min(m.n)}(C)$ are the corresponding right singular vectors.

4 Schur complements

(a) Let $A \in \mathbb{S}^n$, $B \in \mathbb{S}^m_{++}$ and $X \in \mathbb{R}^{n\times m}$. Consider the symmetric block diagonal matrix

$$M := \begin{bmatrix} A & X \\ X^T & B \end{bmatrix}.$$

The matrix

$$S := A - XB^{-1}X^T$$

is called the *Schur complement* of A in M.

i. Show that

$$M \succeq 0 \Leftrightarrow S \succeq 0.$$

Hint: Consider the congruence transformation $C^T MC$ of M by the nonsingular block matrix

$$C := \begin{bmatrix} I_n & 0^{n\times m} \\ -B^{-1}X^T & I_m \end{bmatrix},$$

where I_n and I_m denote the $n \times n$ and the $m \times m$ identity matrix respectively.

ii. Show that

$$M \succ 0 \Leftrightarrow S \succ 0.$$

(b) Let $L_n(x_1, \ldots, x_n)$ denote the $n \times n$ symmetric, tri-diagonal matrix

$$L_n(x_1, \ldots, x_n) := \begin{bmatrix} x_1 & 1 & & & & & \\ 1 & x_2 & 1 & & & & \\ & 1 & x_3 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & x_{n-1} & 1 \\ & & & & 1 & x_n \end{bmatrix}.$$

Here $x_1, \ldots, x_n$ are real variables. It is convenient to write x for the column vector $\begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix}^T$. Consider the problem $\mathcal{P}_n$, which is

$$\begin{array}{rl} \min\limits_{x} & x_1 + \ldots + x_n \\ \text{subject to:} & L_n(x_1, \ldots, x_n) \succeq 0, \end{array}$$

and let p^*_n denote the value of this problem.

i. Show that if $n \geq 2$ then we must have $x_n > 0$ at every feasible point $x \in \mathbb{R}^n$ for problem $\mathcal{P}_n$. (Here $x \in \mathbb{R}^n$ is said to be feasible for $\mathcal{P}_n$ if it satisfies the constraints, i.e. if $L_n(x_1, \ldots, x_n) \succeq 0$.)

ii. Using Schur complements, show for $n \geq 2$ that the constraint in the problem $\mathcal{P}_n$ can be written as

$$L_{n-1}(x_1, \ldots, x_{n-2}, \bar{x}_{n-1}) \succeq 0, \quad x_n > 0,$$

where $\bar{x}_{n-1} := x_{n-1} - 1/x_n$.

iii. Using the new variable $\bar{x}_{n-1}$, show by induction on n that for all $n \geq 1$ we have $p_n^* = 2(n-1)$.

iv. Show that p_n^* (the value of problem $\mathcal{P}_n$) is achieved for $n \geq 2$ at the unique optimal point given by $x_1^* = x_n^* = 1$, $x_i = 2$, $i = 2, \ldots, n-1$, while p_1^* is achieved at the unique optimal point given by $x_1^* = 0$.

5 Graph Isoperimetry

In this exercise, we consider a simple undirected graph $G = (V, E)$, where $V = \{1, \ldots, n\}$ is the set of vertices, $E \subseteq V \times V$ is the set of oriented edges, and the corresponding Laplacian matrix is $L \in \mathbb{R}^{n \times n}$. Recall that for $i, j \in V$, we have $(i, j) \in E \Leftrightarrow (j, i) \in E$, i.e. each unordered pair of vertices $\{i, j\}$ with $j \neq i$ for which i is connected to j (which is usually what is called an edge in graph theory) is being counted twice in E, once as (i, j) and once as (j, i). Further, if d_i denotes the degree of vertex $i \in V$, i.e. the number of edges or equivalently half the number of oriented edges incident at the vertex, the Laplacian matrix $L \in \mathbb{R}^{n \times n}$ is defined as

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

A *cut* of G is, by definition, a subset of the vertices $S \subseteq V$. We think of such a a cut S as partitioning the set of vertices of G into the two subsets S and $V \setminus S$. The *volume* of the cut S is the number of vertices in the smaller of the two pieces, i.e.

$$\mathrm{vol}(S) \coloneqq \min\{|S|, |V \setminus S|\}.$$

The *surface area* of the cut S is the number of edges crossing the cut, i.e.

$$\mathrm{surf}(S) \coloneqq |\{(i, j) \in E \mid i \in S \wedge j \in V \setminus S\}|.$$

Recall that for each edge that both $(i, j) \in E$ and $(j, i) \in E$, so $\mathrm{surf}(S)$ counts each edge crossing S exactly once. We will consider only *nontrivial cuts*, i.e. cuts where $S \neq \emptyset$ and $S \neq V$. So we assume that $n \geq 2$.

We are interested in the ratio between the surface area and the volume of a nontrivial cut, given by

$$h(S) \coloneqq \frac{\mathrm{surf}(S)}{\mathrm{vol}(S)}.$$

In particular, we want to find the nontrivial cut of least surface area relative to volume:

$$h(G) \coloneqq \min_{S \subset V, S \neq \emptyset, S \neq V} h(S).$$

The quantity $h(G)$ is known as the *isoperimetric number* of the graph G.

(a) Instead of working directly with $h(G)$, we will use a more convenient quantity, also defined only for nontrivial cuts, given by

$$\phi(S) \coloneqq \frac{n \,\mathrm{surf}(S)}{|S||V \setminus S|},$$

and we let $\phi(G) \coloneqq \min_{S \subset V, S \neq \emptyset, S \neq V} \phi(S)$

Show that $\phi(G) \leq 2h(G)$.

(b) Let the eigenvalues (including multiplicities) of L be $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. Recall from the previous homework that if G is connected, then $\mathrm{rank}(L) = n - 1$, and consequently $\lambda_1 = 0$.

Conversely, suppose G is disconnected; that is, there exists nontrival $S \subset G$ such that $\mathrm{surf}(S) = 0$. Show that $\lambda_2 = 0$.

Hint: Recall that for $x \in \mathbb{R}^n$ we have

$$x^\top L x = \frac{1}{2} \sum_{(i,j) \in E} (x_i - x_j)^2.$$

(c) For any cut $S \subseteq V$, we can encode S as a zero/one vector $x \in \{0,1\}^n$ where

$$x_i = \begin{cases} 1 & i \in S \\ 0 & i \notin S. \end{cases}.$$

Moreover, any $x \in \{0,1\}^n$ corresponds to a cut $S = \{i \in [n] \mid x_i = 1\}$.

Let $J \in \mathbb{R}^{n \times n}$ denote the all-ones matrix. Show that for any nontrivial cut $S \subset V$ we have

$$\phi(S) = \frac{nx^\top Lx}{x^\top (nI - J)x}.$$

(d) By the previous part of the problem, we can write

$$\phi(G) = \min_{x \in \{0,1\}^n : x \neq 0,1} \frac{nx^\top Lx}{x^\top (nI - J)x}. \tag{1}$$

Note that we exclude the vectors 0 and 1 from the minimization, as they correspond to the trivial cuts.

Show that a very similar optimization problem gives us a formula for λ_2 as

$$\lambda_2 = \min_{x \in \mathbb{R}^n : x \neq 0, x \perp 1} \frac{nx^\top Lx}{x^\top (nI - J)x}. \tag{2}$$

where $x \perp 1$ denotes that the vector $x \in \mathbb{R}^n$ is orthogonal to the all ones vector.

(e) Conclude that $\lambda_2 \leq \phi(G) \leq 2h(G)$.

Remark: This exercise shows that $\lambda_2/2$ is a lower bound on the isoperimetric number $h(G)$, which is the smallest ratio between surface area and volume for a nontrivial cut in G. Intuitively, $h(G)$ is related to *how connected* the graph G is—if G is poorly connected, it is easy to separate a large portion of the vertices while cutting only a few edges. However, if G is well-connected, then any cut with large volume will also cut many edges. There is also a corresponding upper bound on $h(G)$ involving $\sqrt{\lambda_2}$, but we will not cover it here.