EECS 127/227AT **Fall 2020**
Discussion 3

1 Weighted least squares

You need funding for your startup, for which you do some research to understand how startup funding works. You have past funding data from K different startup incubators. Each of the incubators bases its response to any startup proposal that comes to it on the same n features. The data you have from startup i is thus a matrix $A_i \in \mathbb{R}^{m_i \times n}$ and a vector $b_i \in \mathbb{R}^{m_i}$, where m_i is the number of proposals made to incubator i for which you have data and, for each $1 \le j \le m_i$, the j-th row of A_i gives the features for the j-th proposal made to that incubator and the j-th coordinate of b_i gives the corresponding response from that incubator.

You aim to come up with an assessment of the best possible weight to assign to each feature to explain all this data, assuming that at each incubator the response to each proposal it gets is a weighted linear combination of the values of the features of that proposal. With this in mind, you come up with a least-squares methodology for penalizing the residuals of the fit as $\|A_i x - b_i\|_2^2$ at incubator i, but since you have more faith in the correctness of the data available to you from some of the incubators than others, you incorporate this prior knowledge by forming an overall weighted least squares objective given by

$$f(x) = \lambda_1 \|A_1 x - b_1\|_2^2 + \lambda_2 \|A_2 x - b_2\|_2^2 + \cdots + \lambda_K \|A_K x - b_K\|_2^2,$$

where $\lambda_i > 0$ for $i \in \{1, 2, \cdots, K\}$.

(a) Show that the overall objective function $f(x)$ can be expressed as a new least square objective $\|Ax - b\|_2^2$, where you will express A and b in terms of λ_i, A_i, b_i, $i \in \{1, 2, \cdots, K\}$.

(b) What is a criterion for existence of an unique solution for $\min_x f(x)$, assuming a solution exists, i.e. assuming that $b \in \mathcal{R}(A)$? Express your answer in terms of rows/columns of A as applicable.

(c) Show that any optimal solution, call it x^*, to $\min_x f(x)$ satisfies the equation

$$(\lambda_1 A_1^\top A_1 + \lambda_2 A_2^\top A_2 + \cdots + \lambda_K A_K^\top A_K) x^* = \lambda_1 A_1^\top b_1 + \lambda_2 A_2^\top b_2 + \cdots + \lambda_K A_K^\top b_K.$$

2 Moore-Penrose pseudoinverse, least squares and the SVD

Let $A \in \mathbb{R}^{m \times n}$ with rank r. We will assume that $r \geq 1$, i.e. A is not the zero matrix. We know that $r \leq \min(m, n)$. As you have seen in the lectures, A has a singular value decomposition (SVD) of the form

$$A = U\tilde{\Sigma}V^T = \begin{bmatrix} U_{\mathcal{R}} & U_{\mathcal{N}} \end{bmatrix} \begin{bmatrix} \Sigma & 0^{r \times (n-r)} \\ 0^{(m-r) \times r} & 0^{(m-r) \times (n-r)} \end{bmatrix} \begin{bmatrix} V_{\mathcal{R}}^T \\ V_{\mathcal{N}}^T \end{bmatrix},$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices, i.e. $U^T U = UU^T = I_m$ and $V^T V = VV^T = I_n$, and $\Sigma \in \mathbb{R}^{r \times r}$ is a diagonal matrix with the r singular values of A (which are all strictly positive) along the main diagonal in decreasing order. Further we have $U_{\mathcal{R}} \in \mathbb{R}^{m \times r}$, with its columns giving an orthonormal basis for $\mathcal{R}(A)$; $U_{\mathcal{N}} \in \mathbb{R}^{m \times (m-r)}$, with its columns giving an orthonormal basis for $\mathcal{N}(A^T)$; $V_{\mathcal{R}} \in \mathbb{R}^{n \times r}$, with its columns giving an orthonormal basis for $\mathcal{R}(A^T)$; and $V_{\mathcal{N}} \in \mathbb{R}^{n \times (n-r)}$, with its columns giving an orthonormal basis for $\mathcal{N}(A)$. (If either $r = m$ or $r = n$ or both, the corresponding matrices $U_{\mathcal{N}}$ and/or $V_{\mathcal{N}}$ will not need to be defined.)

We also know that the representation

$$A = U_{\mathcal{R}} \Sigma V_{\mathcal{R}}^T,$$

which follows from the above, is called a compact form SVD of A.

In the lectures we defined

$$A^\dagger := V_{\mathcal{R}} \Sigma^{-1} U_{\mathcal{R}}^T,$$

which is called the *Moore-Penrose pseudoinverse* of A.

In this problem we will explore some connections between the Moore-Penrose pseudoinverse, the least squares problem and the SVD.

(a) Suppose $A = a \in \mathbb{R}^{m \times 1}$ is a nonzero column vector. What is $A^\dagger$ (i.e. $a^\dagger$)?

(b) Suppose $A = a^T \in \mathbb{R}^{1 \times n}$ is a nonzero row vector. What is $A^\dagger$ (i.e. $a^{T\dagger}$)?

(c) Now consider the linear system $Ax = b$, to be solved for $x \in \mathbb{R}^n$ given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. In general this equation may not have a solution and even if it does, it may not have a unique solution. In the least squares methodology we attempt to find the "best possible approximate solution", using as a metric to measure how "good" a solution is the sum of the squares of the residual errors, i.e. $\|Ax - b\|_2^2$. (Here $b - Ax$ is called the *residual* and x is called the vector of *regression coefficients*.) If there are many solutions even with this approximation criterion in place (as would happen only if $\mathcal{N}(A)$ is non-trivial) we may further try to focus attention on the one for which $\|x\|_2$ is the smallest.

Based on the SVD representation of A, which we assume to have rank $r \geq 1$ as in the earlier parts of this problem, simplify $\|Ax - b\|_2^2$ so that it is in terms of $\Sigma, V_{\mathcal{R}}, U_{\mathcal{R}}$, and $U_{\mathcal{N}}$.

(d) Solve for an optimal x^* that minimizes $\|Ax - b\|_2^2$.

(e) What is the error at the optimal x^*?

(f) Assume now that $r = n$. Show that the residual error $b - Ax$ at any optimal choice of regression coefficients x (which we know must be of the form $x^* + z$ for some $z \in \mathcal{N}(A)$) satisfies the equation

$$\begin{bmatrix} I_m & A \\ A^T & 0^{n \times n} \end{bmatrix} \begin{bmatrix} b - Ax \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}$$

3 PCA and low-rank compression

We are given an $m \times n$ matrix $X = [x_1, \ldots, x_n]$, with $x_i \in \mathbb{R}^m$ for $i = 1, \ldots, n$ being the data points. (Thus, each feature corresponds to a row of X.) We assume that the data matrix is centered, in the sense that $x_1 + \ldots + x_n = 0$.

$C := \frac{1}{n} X X^\top$ is called the covariance matrix of the data, since the empirical variance of the data points along a direction $u \in \mathbb{R}^m$ is $u^\top C u$.

Consider the following three problems:

(P_1) Find a line going through the origin that maximizes the empirical variance of the collection of data points projected on the line. More explicitly, P_1 is the problem

$$\max_{u \in \mathbb{R}^m : u^\top u = 1} u^\top C u.$$

(P_2) Find a line going through the origin that minimizes the sum of squares of the distances from the points to their projections on this line. More explicitly, P_2 is the problem

$$\min_{u \in \mathbb{R}^m : u^\top u = 1} \sum_{i=1}^{n} \min_{\alpha_i \in \mathbb{R}} \|x_i - \alpha_i u\|_2.$$

(P_3) Find the best approximation of rank at most 1 to the data matrix in Frobenius norm. More explicitly, P_3 is the problem

$$\min_{X' : \mathrm{rank}(X') \leq 1} \|X - X'\|_F.$$

In this exercise, you are asked to show the equivalence between these three problems.

(a) Given $x_0, u \in \mathbb{R}^m$ such that $u^\top u = 1$, consider the problem of projecting a point $x \in \mathbb{R}^m$ on the line $\mathcal{L} := \{x_0 + \alpha u \ : \ \alpha \in \mathbb{R}\}$.

Show that the projected point z^* is given by

$$z^* = x_0 + \alpha^* u,$$

where we define

$$\alpha^* := (x - x_0)^\top u.$$

Further, show that the squared distance $\|z^* - x\|_2^2$ is equal to $\|x - x_0\|_2^2 - \alpha^{*2}$. (By the definition of the projection, this is the minimal squared distance from x to any point on the line $\mathcal{L}$.)

(b) Show that problems P_1, P_2 are equivalent.

(c) Show that P_3 is equivalent to P_1.

Hint: The data matrix is rank-one if and only if it can be expressed as the outer product of two nonzero vectors.