CS 189 Introduction to Machine Learning
Fall 2018 DIS2

1 The Pseudoinverse

Let $X \in \mathbb{R}^{n\times d}$. We do not assume that X is full rank.

(a) Give the definition of the rowspace, columnspace, and nullspace of X.

(b) Check the following facts:

  (a) The rowspace of X is the columnspace of $X^\top$, and vice versa.

  (b) The nullspace of X and the rowspace of X are orthogonal complements.

  (c) The nullspace of $X^\top X$ is the same as the nullspace of X. *Hint: if v is in the nullspace of $X^\top X$, then $v^\top X^\top X v = 0$.*

  (d) The columspace and rowspace of $X^\top X$ are the same, and are equal to the rowspace of X. *Hint: Use the relationship between nullspace and rowspace.*

(c) Recall from the SVD theorem that we can write any matrix X as

$$X = \sum_{i=1}^{\min\{d,n\}} \sigma_i u_i v_i^\top = \sum_{i:\sigma_i>0} \sigma_i u_i v_i^\top$$

where $\sigma_i \geq 0$, and $\{u_i\}$ and $\{v_i\}$ are an orthonormal. Show that

  (a) $\{v_i : \sigma_i > 0\}$ are an orthonormal basis for the rowspace of of X

  (b) Similarly, $\{u_i : \sigma_i > 0\}$ are an orthonormal basis for the columnspace of X (*Hint: consider $X^\top$*)

(d) Define the Moore-penrose pseudoinverse to be the matrix:

$$X^\dagger = \sum_{i:\sigma_i>0} \sigma_i^{-1} v_i u_i^\top,$$

What is the matrix $X^\dagger X$, what operator does it correspond to? What is $X^\dagger X$ if $\operatorname{rank}(X) = d$? What $X^\dagger X$ if $\operatorname{rank}(X) = d$ and $n = d$?

2 The Least Norm Solution

Let $X \in \mathbb{R}^{n\times d}$, where $n \geq d$, but where $\operatorname{rank}(X)$ is possibly less than d. As in problem 1, we will right the SVD of X as a sum of rank-one terms

$$X = \sum_{i:\sigma_i>0} u_i \sigma_i v_i^\top,$$

In this problem, our goal will to provide an explicit expression for the *least-norm* least-squares estimator, defined as :

$$\widehat{\theta}_{LS,LN} := \arg\min_{\theta}\{\|\theta\|_2^2 : \theta \text{ is a minimizer of } \|X\theta - y\|_2^2\},$$

where $\theta \in \mathbb{R}^d$ and $y \in \mathbb{R}^n$.

(a) Show that $\widehat{\theta}_{LS,LN}$ is the unique minimizer of $\|X\theta - y\|_2^2$ which lies in the rowspace of X. Try not to use the SVD.

(b) Show that $\widehat{\theta}_{LS,LN}$ has the following form:

$$\widehat{\theta}_{LS,LN} = \sum_{i:\sigma_i > 0} \frac{1}{\sigma_i} v_i \langle u_i, y \rangle, \tag{1}$$

Solve this problem by directly checking that the above expression for $\widehat{\theta}_{LS,LN}$ is in the rowspace of X, and satisfies the necessary optimality condition to be a minimzer of the least-squares objective.

(c) We give another solution to finding a form for $\widehat{\theta}_{LS,LN}$. Again, write out the SVD decomposition of X as

$$X = \sum_{i:\sigma_i > 0} u_i \sigma_i v_i^\top,$$

Now follow these steps:

- What is the operator $(X^\top X)^\dagger (X^\top X)$? *Hint: pattern match with the last part of Problem 1, where* $X \leftarrow X^\top X$
- Show that $(X^\top X)^\dagger X^\top = X^\dagger$ *Hint: write everything out as a sum of rank-one terms*
- Show that any minimizer of the least squares objective satisfies

$$P_X \theta = X^\dagger y,$$

where P_X is the orthogonal projection onto the rowspace of X.

- Conclude that

$$\widehat{\theta}_{LS,LN} = X^\dagger y.$$

Verify that this is consistent with your answer to the previous part of the problem.

3 The Ridge Regression Estimator

Recall the ridge estimator for $\lambda > 0$,

$$\widehat{\theta}_\lambda := \arg\min_{\theta} \|X\theta - y\|_2^2 + \lambda\|\theta\|_2^2,$$

Let

$$X = \sum_i \sigma_i u_i v_i^\top$$

be the SVD decomposition as given in the previous two problems. On the homework, you will show that

$$\widehat{\theta}_\lambda = \sum_{i=1}^{d} \frac{\sigma_i}{\sigma_i^2 + \lambda} v_i \langle u_i, y \rangle$$

You should use this decomposition in this problem.

(a) Show that

$$\|\widehat{\theta}_\lambda\|_2^2 = \sum_{i:\sigma_i > 0} (\frac{\sigma_i}{\sigma_i^2 + \lambda})^2 \langle u_i, y \rangle^2.$$

(b) Recall the least-norm least squares solution is $\widehat{\theta}_{LN,LS}$ from Problem 2. Show that if $\widehat{\theta}_{LN,LS} = 0$, then $\widehat{\theta}_\lambda = 0$ for all $\lambda > 0$. *Hint: use the formula for the least norm least squares solution from Problem 2.*

(c) Show that if $\widehat{\theta}_{LN,LS} \neq 0$, then the map $\lambda \mapsto \|\widehat{\theta}_\lambda\|_2^2$ is strictly decreasing and strictly positive on $(0, \infty)$.

(d) Show that

$$\lim_{\lambda \to 0} \widehat{\theta}_\lambda \to \widehat{\theta}_{LS,LN}.$$

(e) In light of the above, why do you think that people describe the ridge regression as "controlling the complexity" of the solution $\widehat{\theta}_\lambda$