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. **Solution:** The rowspace is the span of the rows of X, the columnspace is the span of the columns of X, and nullspace is the set of vectors v such that $Xv = 0$.

(b) Check the following facts:

   (a) The rowspace of X is the columnspace of $X^\top$, and vice versa. **Solution:** The rows of X are the columns of $X^\top$, and vice versa.

   (b) The nullspace of X and the rowspace of X are orthogonal complements. **Solution:** v is in the nullsapce of X if and only if $Xv = 0$, which is true if and only if for every row X_i of X, $\langle X_i, v\rangle = 0$. This is precisely the condition that v is perpedicular to each row of X. This means that v is in the nullspace of X if and only if v is in the orthogonal complement of the span of the rows of X, i.e. the orthogonal complement of the rowspace of X.

   (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 $v^\top X^\top X v = 0$.* **Solution:** If v is in the nullspace of X, then $X^\top X v = X^\top 0 = 0$. On the other hand, if v is in the nullspace of $X^\top X$, then $v^\top X^\top X v = v^\top 0 = 0$. Then, $v^\top X^\top X v = \|Xv\|_2^2 = 0$, which implies that $Xv = 0$.

   (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.* **Solution:** $X^\top X$ is symmetric, and therefore its rows and columns are the same; hence, its columspaces and rowspaces are the same. By the previous problem, the nullspace of $X^\top X$ is equal to the nullspace of X, therefore. Thus,

$$\text{rowspace}(X) = \text{nullspace}(X)^\perp = \text{nullspace}(X^\top X)^\perp = \text{rowspace}(X^\top X),$$

   where $()^\perp$ denotes orthogonal complement.

(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* $X^\top$)

Solution: Since $\{v_i : \sigma_i > 0\}$ are an orthonormal, it suffices to show that their span is the row space of X. Since the rowspace is the orthogonal complement of the nullspace of X, it suffices to show that $v \in \mathrm{span}(\{v_i : \sigma_i > 0\})^\perp$ if and only if then $Xv = 0$. We have that

$$Xv = \sum_{i:\sigma_i>0} \sigma_i u_i (v_i^\top v).$$

Since $\sigma_i u_i$ are all linearly independent, $Xv = 0$ if and only if $(v_i^\top v) = 0$ for all i, as needed.

The the second part,

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

which means that u_i are a basis for the rowspace of $X^\top$ by the above. Hence, u_i are a basis for the columnspace of X.

(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 $\mathrm{rank}(X) = d$? What $X^\dagger X$ if $\mathrm{rank}(X) = d$ and $n = d$? **Solution:**

$$\begin{aligned} X^\dagger X &= \sum_{i:\sigma_i>0} \sigma_i^{-1} v_i u_i^\top \sum_{j:\sigma_j>0} \sigma_j u_j v_i^\top \\ &= \sum_{i:\sigma_i>0} \sum_{j:\sigma_j>0} \sigma_j \sigma_i^{-1} u_i^\top u_j \cdot v_i v_j^\top \\ &= \sum_{i:\sigma_i>0} \sum_{j:\sigma_j>0} \sigma_j \sigma_i^{-1} \mathbf{I}(i = j) \cdot v_i v_j^\top \\ &= \sum_{i:\sigma_i>0} v_i v_i^\top. \end{aligned}$$

Hence, by our last homework we that $X^\dagger X$ is an orthogonal projection onto the span of v_i, which is precisely the rowspace of X. If $\mathrm{rank}(X) = d$, then $X^\dagger X = I$, and thus if $d = n$, $X^\dagger = X^{-1}$.

2 The Least Norm Solution

Let $X \in \mathbb{R}^{n \times d}$, where $n \geq d$, but where $\mathrm{rank}(X)$ is possibly less than d. As in problem 1, we will write 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. **Solution:** The minimizers of the least squares objective are precisely the solutions θ to the equation we have that

$$X^\top X\theta = X^\top y$$

In particular, for any single solution $\overline{\theta}$, we can write $\overline{\theta} = \theta_0 + \Delta$, where θ_0 is in the rowsapce of X and Δ is in the nullspace of $\mathrm{nullspace}(X^\top X) = \mathrm{nullspace}(X)$; this follows since the rowspace of X and the nullspace of X are orthogonal complements. Moreover, we find that

$$X^\top X\theta_0 = X^\top X(\overline{\theta} - \Delta) = X^\top X(\overline{\theta}) = X^\top y,$$

so θ_0 is also a minimizer of the least squares objective.

Note that since the minimizers are the solution to a linear system, and θ_0 is one such solution, then any other minimzer is of the form $\theta_0 + \Delta$, where $\Delta \in \mathrm{nullspace}(X^\top X) = \mathrm{nullspace}(X)$. Thus, for any other minimizer $\theta = \theta_0 + \Delta$

$$\begin{aligned}
\|\theta\|_2^2 &= \|\theta_0 + \Delta\|_2^2 \\
&= \|\theta_0\|^2 + \|\Delta\|_2^2 + 2\langle\theta_1, \Delta\rangle \\
&= \|\theta_0\|^2 + \|\Delta|_2^2,
\end{aligned}$$

where we use the fact that $\theta_0 \perp \Delta$, because the nullspace and rowspace of X are orthogonal. Hence, we conclude that $\|\theta\|_2^2$ is strictly greater than $\|\theta_0\|_2^2$ unless $\Delta = 0$,i.e. $\theta = \theta_0$. It follows that θ_0 is precisely the least norm least squares solution.

(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.

Solution: The easiest was to go about this is to show that $\theta = \sum_{i:\sigma_i>0} \frac{1}{\sigma_i} v_i \langle u_i, y\rangle$ is in the row space of X, and that θ satisfies the normal equations $X^\top X\theta = X^\top\theta$. By the previous problem, this implies that $\theta = \widehat{\theta}_{LN,LS}$. Recall from the SVD-theorem that

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

To see that θ is in the row space of X, observe that θ is a linear combination of v_i for $i : \sigma_i > 0$. Each v_i is in the rowspace of X, by problem 1.

Next, we show that θ satisfies the normal equation

$$(X^\top X)\theta = X^\top y$$

Using the SVD theorem, we can write

$$(X^\top X) = \sum_{i=1}^{d} \sigma_i^2 v_i v_i^\top$$
$$X^\top y = \sum_{i=1}^{d} \sigma_i v_i \langle u_i, y\rangle$$

Therefore,

$$\begin{aligned}
(X^\top X)\theta &= (\sum_{i=1}^{d} \sigma_i^2 v_i v_i^\top)(\sum_{j:\sigma_j>0} \sigma_j^{-1} v_i \langle u_i, y\rangle) \\
&= \sum_{i=1}^{d} \sum_{j:\sigma_j>0} v_i \cdot (\sigma_i^2 \sigma_j^{-1}) \cdot \langle v_i, v_j\rangle \cdot \langle u_i, y\rangle \\
&= \sum_{i=1}^{d} \sum_{j:\sigma_j>0} v_i \cdot (\sigma_i^2 \sigma_j^{-1}) \mathbf{I}(i=j) \langle u_i, y\rangle \\
&= \sum_{i:\sigma_i>0} v_i (\sigma_i^2 \sigma_i^{-1}) \langle u_i, y\rangle \\
&= \sum_{i:\sigma_i>0} v_i \sigma_i \langle u_i, y\rangle,
\end{aligned}$$

which is precisely $X^\top y$.

(c) We give another solution to finding a form for $\widehat{\theta}_{LS,LN}$ using the pseudoinverse. 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* $X \leftarrow X^\top X$ **Solution:** By problem 1, $(X^\top X)^\dagger (X^\top X)$ is the orthogonal projection onto the rowspace of $X^\top X$, which is precisely the rowspace of X.
- Show that $(X^\top X)^\dagger X^\top = X^\dagger$ *Hint: write everything out as a sum of rank-one terms* **Solution:**

$$\begin{aligned}
(X^\top X)^\dagger X^\top &= \sum_{i:\sigma_i>0} \sigma_i^{-2} v_i v_i^\top \sum_j \sigma_j v_j u_j^\top \\
&= \sum_j \sum_{i:\sigma_i>0} \frac{\sigma_j}{\sigma_i^2} \langle v_j, v_i\rangle \cdot v_i u_j^\top
\end{aligned}$$

$$= \sum_j \sum_{i:\sigma_i>0} \frac{\sigma_j}{\sigma_i^2} \mathbf{I}(i=j) \cdot v_i u_j^\top$$
$$= \sum_{i:\sigma_i>0} \sigma_i^{-1} v_i u_j^\top = X^\dagger$$

- 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. **Solution:** Any least squares solution satisfies

$$X^\top X\theta = X^\top y$$

Multiply by $(X^\top X)^\dagger$, which gives

$$(X^\top X)^\dagger (X^\top X)\theta = (X^\top X)^\dagger X^\top y.$$

Using the previous part, this simplies to $P_X\theta = X^\dagger y$.

- Conclude that

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

Verify that this is consistent with your answer to the previous part of the problem. **Solution:** Since $\widehat{\theta}_{LS,LN}$ lies in the rowspace of X, we have $\widehat{\theta}_{LS,LN} = P_X \widehat{\theta}_{LS,LN} = X^\dagger y$. Moreover,

$$X^\dagger y = (\sum_{i:\sigma_i>0} \sigma_i^{-1} v_i u_i^\top) y = \sum_{i:\sigma_i>0} \sigma_i^{-1} \langle u_i, y_i \rangle v_i.$$

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

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

Solution: First, we have that

$$\begin{aligned}
\|\widehat{\theta}_\lambda\|_2^2 &= \left\langle \sum_{i=1}^{d} \frac{\sigma_i}{\sigma_i^2+\lambda} v_i \langle u_i, y\rangle, \sum_{j=1}^{d} \frac{\sigma_j}{\sigma_j^2+\lambda} v_j \langle u_j, y\rangle \right\rangle \\
&= \sum_{1\le i,j\le d} (\frac{\sigma_i}{\sigma_i^2+\lambda})(\frac{\sigma_j}{\sigma_j^2+\lambda}) \langle u_j, y\rangle \langle u_i, y\rangle \langle v_i, v_j\rangle \\
&= \sum_{1\le i\le d} (\frac{\sigma_i}{\sigma_i^2+\lambda})^2 \langle u_i, y\rangle^2. \\
&= \sum_{i:\sigma_i>0} (\frac{\sigma_i}{\sigma_i^2+\lambda})^2 \langle u_i, y\rangle^2.
\end{aligned}$$

(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.* **Solution:** If the least norm least squares solution is 0, then

$$\sum_{i:\sigma_i>0} \sigma_i^{-1} \langle u_i, y_i\rangle v_i = 0,$$

, which means that $\langle u_i, y_i\rangle = 0$ for each $i : \sigma_i = 0$, because v_i are linearly independent. Hence, $\widehat{\theta}_\lambda = 0$ by plugging into the formula for $\widehat{\theta}_\lambda = 0$.

(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 $(0, \infty)$. **Solution:** If $\widehat{\theta}_\lambda \neq 0$, then at least one of the terms $\langle u_i, y\rangle^2$ is strictly greater than zero. Thus,

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

is a non-trivial nonnegative weighted combination of terms $(\frac{\sigma_i}{\sigma_i^2+\lambda})^2$, which are positive and strictly decreasing in λ.

(d) Show that

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

Solution: Even though the limit of the ridge-regression objective as $\lambda \to 0$ is the least squares objective, this does not immediately guarantee that limit of the ridge solution is the least squares solution. Instead, lets use the form

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

Since limits commute with sums, we have

$$\lim_{\lambda \to 0} \widehat{\theta}_\lambda = \sum_{i=1}^{n} v_i \langle u_i, y \rangle \cdot (\lim_{\lambda \to 0} \frac{\sigma_i}{\sigma_i^2 + \lambda})$$

Now,

$$\lim_{\lambda \to 0} \frac{\sigma_i}{\sigma_i^2 + \lambda} = \begin{cases} 0 & \sigma_i = 0 \\ \sigma_i^{-1} & \sigma_i > 0 \end{cases}.$$

Thus,

$$\lim_{\lambda \to 0} \widehat{\theta}_\lambda = \sum_{i:\sigma_i > 0} \sigma_i^{-1} v_i \langle u_i, y \rangle,$$

which we have shown above is the least norm solution.

(e) In light of the above, why do you think that people describe the ridge regression as "controlling the complexity" of the solution $\widehat{\theta}_\lambda$ **Solution:** We see that increasing the ridge parameter λ shrinks the norm of $\widehat{\theta}_\lambda$, and that even as $\lambda \to 0$, $\widehat{\theta}_\lambda$ picks out the least norm least squares solution.