CS 189 Introduction to Machine Learning
Summer 2019 Marc Khoury & Brijen Thananjeyan

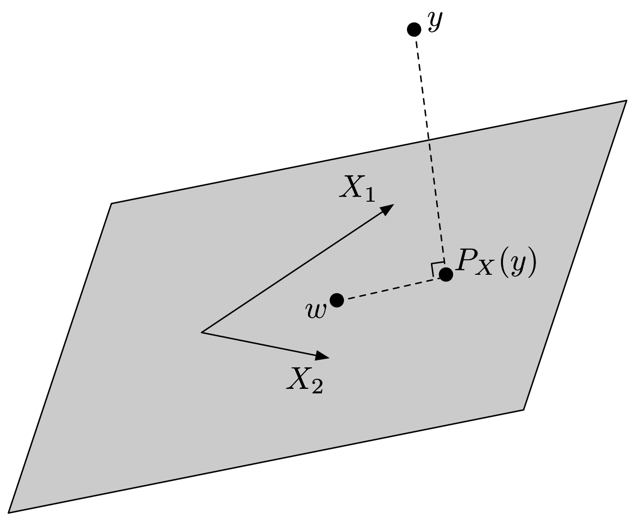
DIS6

1 Linear Regression, Projections and Pseudoinverses

We are given $X \in \mathbb{R}^{n \times d}$ where $n > d$ and $\text{rank}(X) = d$. We are also given a vector $y \in \mathbb{R}^n$. Define the orthogonal projection of y onto $\text{range}(X)$ as $P_X(y)$.

(a) Prove that $P_X(y) = \underset{w \in \text{range}(X)}{\arg\min} |y - w|^2$.

Solution:



Note that $|y-w|^2 = |y-P_X(y)+P_X(y)-w|^2 = |y-P_X(y)|^2+|P_X(y)-w|^2+2(y-P_X(y))^\top(P_X(y)-w)$. Now we can easily see from the figure above that $(y - P_X(y))$ is orthogonal to any vector in the columnspace of X. Hence $|y - w|^2 = |y - P_X(y)|^2 + |P_X(y) - w|^2 \geq |y - P_X(y)|^2$. This shows that $w = P_X(y)$.

Note that in lecture, we learned how to find $\hat{\theta}$ that minimizes the least squares loss $L(\theta) = |y - X\theta|^2$. In other words, we tried to find θ such that $X\theta$ is the vector in the columnspace of X that is closest to our response vector y. Hence, $P_X(y) = X\theta$.

(b) An orthogonal projection is a linear transformation. Hence, we can define $P_X(y) = Py$ for some projection matrix P. Specifically, given $1 \leq d \leq n$, a matrix $P \in \mathbb{R}^{n \times n}$ is said to be a rank-d orthogonal projection matrix if $\text{rank}(d) = P$, $P = P^\top$ and $P^2 = P$. Prove that P is a rank-d projection matrix if and only if there exists a $U \in \mathbb{R}^{n \times d}$ such that $P = UU^\top$ and $U^\top U = I$

Solution:

Since P is symmetric, $P = V\Sigma V^\top$ for some orthogonal $V \in \mathbb{R}^{n\times n}$ and real, diagonal $\Sigma \in \mathbb{R}^{n\times n}$. Let v be an eigenvector of P with eigenvalue λ. Then, $\lambda^2 v = P^2 v = Pv = \lambda v$, so that $\lambda \in \{0, 1\}$. Hence, the diagonals of Σ are binary-valued. Letting U denote the matrix whose columns correspond to the indicies i for which $\Sigma_{ii} = 1$, we have $P = V\Sigma V^\top = UU^\top$. Since P has rank d, there that there are d such 1-valued indices.

Since the columns of U are a subset of those of V, they are orthonormal, whence $U^\top U = I$.

Conversely, if $P = UU^\top$, then $P = P^\top$ trivially, and $P^2 = UU^\top UU^\top = UU^\top$. Moreover, P has rank at most d since $P = UU^\top$, and rank at least $\mathrm{rank}(PU) = \mathrm{rank}(UU^\top U) = \mathrm{rank}(U) = d$.

(c) Prove that if P is a rank d projection matrix, then $\mathrm{tr}(P) = d$.

Solution:

Approach 1: Using the trace trick $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, $\mathrm{tr}(P) = \mathrm{tr}(UU^\top) = \mathrm{tr}(U^\top U) = \mathrm{tr}(I_d) = d$.

Approach 2: $\mathrm{tr}(P)$ is the sum of the eigenvalues of P. As verified in Part (a), these lie in $\{0, 1\}$, and since $\mathrm{rank}(P) = d$, we must have that P has d eigenvalues equal to 1, and all others zero. Thus, $\mathrm{tr}(P) = 1 \cdot d = d$.

(d) Prove that if $X \in \mathbb{R}^{n\times d}$ and $\mathrm{rank}(X) = d$, then $X(X^\top X)^{-1}X^\top$ is a rank-d orthogonal projection matrix. What is the corresponding matrix U?

Solution:

Let $X = U\Sigma V^\top$ denote the SVD of X, with $U \in \mathbb{R}^{n\times d}$, and $\Sigma \in \mathbb{R}^{d\times d}$, and $V \in \mathbb{R}^{d\times d}$. Then, $X^\top X = V\Sigma^2 V^{top}$, and since $\mathrm{rank}(X) = d$, Σ^2 is invertible, with $X^\top X = V^\top \Sigma^{-2} V$. Hence,

$$X(X^\top X)^{-1}X^\top = U\Sigma V^\top (V\Sigma^{-2}V^\top)V\Sigma U^\top = U\Sigma\Sigma^{-2}\Sigma U = UU^\top,$$

which shows that $X(X^\top X)^{-1}X^\top$ is the projection matrix $UU^\top$, where U is the left singular-vectors matrix of X.

For the remainder of the problem set, we no longer assume that X is full rank.

(e) The Singular Value Decomposition theorem states that we can write any matrix X as

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

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 row space 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 row space 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 row space of $X^\top$ by the above. Hence, u_i are a basis for the columnspace of X.

(f) Define the Moore-Penrose pseudoinverse to be the matrix:

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

To what operator does the matrix $X^\dagger X$ correspond? What is $X^\dagger X$ if $\text{rank}(X) = d$? If $\text{rank}(X) = d$ and $n = d$?

Solution:

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

Hence, by our last homework we that $X^\dagger X$ is an orthogonal projection onto the span of v_i, which is precisely the row space of X. If $\text{rank}(X) = d$, 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 $n \geq d$, where $\text{rank}(X)$ is possibly less than d. As in problem 1, we will write the SVD of X as a sum of rank-one terms

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

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

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

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

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

Solution: The minimizers of the least squares objective are the solutions θ to the equation

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

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

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

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 minimizer is of the form $\theta_0 + \Delta$, where $\Delta \in$ nullspace$(X^\top X)$ = nullspace(X). Thus, for any other minimizer $\theta = \theta_0 + \Delta$

$$\begin{aligned}|\theta|^2 &= |\theta_0 + \Delta|^2 \\ &= |\theta_0|^2 + |\Delta|^2 + 2\theta_0^\top\Delta \\ &= |\theta_0|^2 + |\Delta|^2,\end{aligned}$$

where we use the fact that $\theta_0 \perp \Delta$, because the nullspace and rowspace of X are orthogonal. Hence, we conclude that $|\theta|^2$ is strictly greater than $|\theta_0|^2$ 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:

$$\widehat{\theta}_{LS,LN} = \sum_{i:\sigma_i>0} \frac{1}{\sigma_i} v_i(u_i^\top y), \tag{1}$$

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 minimizer of the least-squares objective.

Solution: The easiest way to go about this is to show that $\theta = \sum_{i:\sigma_i>0} \frac{1}{\sigma_i} v_i(u_i^\top y)$ is in the rowspace 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

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

To see that θ is in the rowspace 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

$$(X^\top X) = \sum_{i=1}^{d} \sigma_i^2 v_i v_i^\top$$

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

Therefore,

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

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:

(1) 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.

(2) Show that $(X^\top X)^\dagger X^\top = X^\dagger$.
*Hint: write everything out as a sum of rank-one terms.*A

Solution:

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

(3) 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 $P_X\theta = X^\dagger y$.

(4) 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,

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

3 SGD Convergence for Logistic Regression

In this problem, we will prove that gradient descent converges to a unique minimizer of the logistic regression cost function, binary cross-entropy. We will consider the case where we are minimizing this cost function for a single data point. For weights $w \in \mathbb{R}^d$, data $x \in \mathbb{R}^d$, and a label $y \in \{0, 1\}$, the logistic regression cost function is given by

$$J(w) = -y \log s(x \cdot w) - (1 - y) \log(1 - s(x \cdot w))$$

Where $s(\gamma) = 1/(1 + \exp(-\gamma))$ is the logistic function (also called the sigmoid). You may assume that $x \neq 0$.

(a) To start, write the gradient descent update function $G(w)$, which maps w to the result of a single gradient descent update with learning rate $\epsilon > 0$.

Solution: From lecture, we know $s'(\gamma) = s(\gamma)(1 - s(\gamma))$. Letting $z = s(w \cdot x)$, we get

$$\nabla_w J = -(y - z)x$$

Hence, the the gradient descent update is

$$G(w) = w + \epsilon(y - z)x$$

(b) Show that the cost function J has a unique minimizer w^* by proving that J is strictly convex. *Hint: how does this relate to the Hessian, $\nabla_w^2 J$?*

Solution: From the last part, the gradient is $-(y - s(w \cdot x))x$. Taking the gradient of this,

$$\nabla_w^2 J = z(1-z)xx^T$$

This is positive definite, since for any vector $v \in \mathbb{R}^d$, $v^T(z(z-1)xx^T)v = (z(z-1))(v^T x)^2 > 0$. This holds because $0 < z < 1$ and $x \neq 0$.

(c) Next, show that G is a *contraction*, which means that there is a constant $0 < \rho < 1$ such that, for any $w, w' \in \mathbb{R}^d$, $|G(w) - G(w')| < \rho|w - w'|$.
Hint: this is equivalent to showing that the gradient has bounded norm: $\|\nabla_w G(w)\| < \rho$

Solution: We compute the gradient of G:

$$\nabla_w G(w) = \nabla_w (w + \epsilon(y - z)w) = I - \epsilon z(1-z)xx^T$$

We know that $0 < z(1-z) < 1$. Setting $\epsilon < 1/|x|^2$ gives us that $0 < \|\epsilon z(1-z)xx^T\| < 1$. So, $\|\nabla_w G\| = \|I - \epsilon z(1-z)xx^T\| = \rho$ for a constant $0 < \rho < 1$.

(d) Finally, calling $w^{(t)}$ the t-th iterate of gradient descent, show that $|w^* - w^{(t)}| < \rho^t |w^* - w^{(0)}|$, so that $\lim_{t\to\infty} |w^* - w^{(t)}| = 0$.

Solution: For a single step:

$$|w^* - G(w)| = |G(w^*) - G(w)| < \rho|w^* - w|$$

So, for n steps, we get

$$|w^* - w^{(t)}| = |w^* - G(w^{(t-1)})| < \rho|w^* - G(w^{(t-2)})| < \rho^2|w^* - G(w^{(t-3)})| < \ldots < \rho^t|w^* - w^{(0)}|$$

