1 Linear Regression, Projections and Pseudoinverses

(a) $|y-w|^2 = |y-P_X(y)+P_X(y)-w|^2 = |y-P_X(y)|^2 + |P_X(y)-w|^2 + 2(y-P_X(y))^\top(P_X(y)-w)$. Since $y-P_X(y)$ is orthogonal to any vector in $\mathrm{range}(X)$, we have $|y-w|^2 = |y-P_X(y)|^2 + |P_X(y)-w|^2$. It minimizes when $w = P_X(y)$. **QED.**

(b)

$\Leftarrow$ If $P = UU^\top$, then $P = P^T$ trivially holds. $P^2 = U(U^\top U)U^\top = UU^\top = P$. Since, $P = UU^\top$, we have $\mathrm{rank}(P) \le \mathrm{rank}(U) \le d$. $\mathrm{rank}(P) \ge \mathrm{rank}(U^\top PU) = \mathrm{rank}(I) = d$. Therefore, we have $\mathrm{rank}(P) = d$.

$\Rightarrow$ Since P is symmetric, we have $P = Q\Lambda Q^\top$, where $Q \in \mathbb{R}^{n\times n}$ is orthogonal and $\Lambda \in \mathbb{R}^{n\times n}$ is diagonal and real. Let λ be a eigenvalue of P and v be its corresponding eigenvector. Then we have $\lambda^2 v = P^2 v = Pv = \lambda v$. So the eigenvalues must be 0 or 1. Let U be the matrix of subset of columns of Q whose corresponding eigenvalues are 1. Since $\mathrm{rank}(P) = d$, there are d such columns. Since $P = \sum_{i=1}^n \lambda_i q_i q^\top$, we have $P = UU^\top$ with $U \in \mathbb{R}^{n\times d}$. Since U is orthogonal, we have $U^\top U = I$. **QED.**

(c) $\mathrm{tr}(P)$ is the sum of all the eigenvalues of P. we have proved in part (b) that all the eigenvalues are either 0 or 1, and since $\mathrm{rank}(P) = d$, we have $\mathrm{tr}(P) = d$.

(d) Let $X = U\Sigma V^\top$ be the SVD of X, where $U \in \mathbb{R}^{n\times d}, \Sigma \in \mathbb{R}^{d\times d}, V \in \mathbb{R}^{d\times d}$.

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

This proves that $X(X^\top X)^{-1}X^\top$ is a rand-d orthogonal projection matrix. The corresponding matrix U is the matrix of the left singular vectors of X. **QED.**

(e)

Because the row space of X is the orthogonal complement of the null space of X. It is sufficient to show for any $v \in \mathbb{R}^n$, we have $Xv = 0 \Leftrightarrow \forall i, v_i^\top v = 0$.

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

Since $\sigma_i u_i$'s are independent vectors, $Xv = 0 \Leftrightarrow \forall i, v_i^\top v = 0$. **QED.**

(f)

We can write X as the reduced form of the SVD: $X = U_r \Sigma_r V_r^T$. Then by definition of the Moore-Penrose pseudoinverse of X is $X^+ = V_r \Sigma_r^{-1} U_r^\top$. Then we have $X^+ X = V_r V_r^T$, which is the orthogonal projection matrix onto the row space of X. If $\mathrm{rank}(X) = d$. Then $X^+ X = I$. If $\mathrm{rank}(X) = d$ and $n = d$, then $X X^+ = U_r U_r^\top = I$. So X^+ is the inverse of X, i.e. $X^+ = X^{-1}$.

2 The Least Norm Solution

(a)

If θ is a minimizer of $|X\theta - y|^2$, then θ satisfies $X^\top X\theta = X^\top y$. θ can be written as $\theta_0 + \delta$ where θ_0 is in the row space of X and δ is in the null space. Note that δ has no impact on $||y - X\theta||$, since $X\delta = 0$. However, it affects $||\theta||^2 = ||\theta_0||^2 + ||\delta||^2$. So the minimus norm solution of $X^\top X\theta = X^\top y$ is θ_0 which lies in the rowspace of X, i.e., it has a zero-component in the nullspace of X.

(b)

$$\hat{\theta}_{LS,LN} = \sum_{i:\delta_i>0} \frac{1}{\delta_i} v_i(u_i^\top y) = V_r\Sigma_r^{-1}U_r^\top y$$

It is obvious that $\hat{\theta}_{LS,LN}$ is in the rowspace of X, since $\hat{\theta}_{LS,LN}$ is in the column space of V_r and the columns of V_r are an orthonormal basis for the rowspace of X. To show that $\hat{\theta}_{LS,LN}$ satisfies the normal equation, note that $X^\top X\hat{\theta}_{LS,LN} = V_r\Sigma_r(U_r^\top U_r)\Sigma_r(V_r^\top V_r)\Sigma_r^{-1}U_r^\top y = X^\top y$. **QED.**

(c)

1. $(X^\top X)^+(X^\top X)$ is the orthogonal projection onto the rowspace of X.
2. $(X^TX)^+X^T = V_r\Sigma_r^{-2}V_r^\top V_r\Sigma_r U_r^\top = X^+$
3. $P_X\theta = (X^\top X)^+(X^\top X)\theta = (X^\top X)^+X^\top y = X^+y$
4. From (a) we know $\hat{\theta}_{LS,LN}$ lies in the rowspace of X, so $P_X\hat{\theta}_{LS,LN} = \hat{\theta}_{LS,LN} = X^+y$

3 SGD Convergence for Logistic Regression

(a)

By the definition of gradient descent,

$$G(w) = w - \epsilon \cdot \Delta_w J$$

Let $z = s(x \circ w)$, then $\Delta_w z = z(1-z)x$, and $\Delta_w J = -(y-z)x$. So $G(w) = w + \epsilon(y - z)x$.

(b)

$\Delta_w^2 J = \Delta_w zx = z(1-z)xx^\top$, which positive definite. Because for any $v \in \mathbb{R}^d$ and $v \neq 0$, $v^\top \Delta_w^2 Jv = z(1-z)|x^\top v|^2 > 0$, since $0 < z < 1$ and $x \neq 0$. This proves that J is strictly convex.

(c)

$$\Delta_w G(w) = I - \epsilon z(1-z)xx^\top$$

Since $0 < z(1-z) < 1$. Setting $\epsilon < \frac{1}{|x|^2}$ gives us $||\Delta_w G(w)|| < \rho$ for some $0 < \rho < 1$.

(d)

From (c), $|G(w^*) - G(w^t)| < \rho|w^* - w^t|$. Then we have $|w^* - w^{t+1}| < \rho|w^* - w^t|$. By telescoping, we have $|w^* - w^t| < \rho^t|w^* - w^0|$.