# Machine Learning Foundations Homework 3

**Problem 1** (60 points)**.**

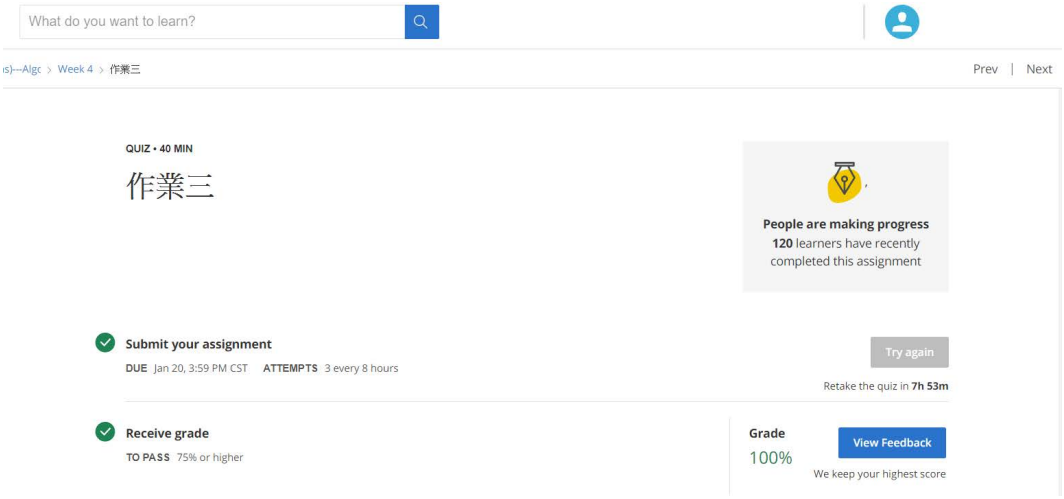



Figure 1: Q1 Screenshot (100 points). Please zoom in to see the details.

**Problem 2** (20 points)**.** When using SGD on the following error function and 'ignoring' some singular points that are not differentiable, prove or disprove that $\mathrm{err}(w) = \max(0, -yw^\top x)$ results in PLA.

***Sol.***

$$\nabla \mathrm{err}(w) = \begin{Bmatrix} 0, & yw^\top x \geq 0 \\ -yx, & yw^\top x < 0 \end{Bmatrix} \implies -\nabla \mathrm{err}(w) = \begin{Bmatrix} 0, & y = \mathrm{sgn}(w^\top x) \\ yx, & y \neq \mathrm{sgn}(w^\top x) \end{Bmatrix} = [\![ y \neq \mathrm{sgn}(w^\top x) ]\!] yx$$

SGD updates by

$$w_{t+1} \leftarrow w_t + \eta\theta(-y_n w_t^\top x_n)(y_n x_n) = w_t - \eta \nabla \mathrm{err}(w_t)$$

while PLA updates by

$$w_{t+1} \leftarrow w_t + [\![ y_n \neq \mathrm{sgn}(w_t^\top x_n) ]\!] y_n x_n.$$

Thus, $\eta = 1$ and $\mathrm{err}(w) = \max(0, -yw^\top x)$ make SGD PLA. □

**Problem 3** (20 points)**.** Write down the derivation steps of Question 9 of Homework 3 on Coursera.

***Sol.*** By definition, $\hat{E}_2(\Delta u, \Delta v) = \frac{1}{2}(\Delta u, \Delta v)^\top \nabla^2 E(u,v)(\Delta u, \Delta v) + (\Delta u, \Delta v)^\top \nabla E(u,v) + E(u,v)$. Since $\nabla^2 E(u,v)$ is positive definite, $\nabla^2 E(u,v)$ is invertible, and $\hat{E}_2(\Delta u, \Delta v)$ is differentiable and convex. Thus, $\hat{E}_2(\Delta u, \Delta v)$ attains global minimum at a stationary point if one exists.

$$0 = \frac{\partial \hat{E}_2(\Delta u, \Delta v)}{\partial(\Delta u, \Delta v)} = \nabla^2 E(u,v)(\Delta u, \Delta v) + \nabla E(u,v) \implies (\Delta u, \Delta v) = -(\nabla^2 E(u,v))^{-1} \nabla E(u,v)$$

□

**Problem 4** (20 points)**.** Write down the derivation steps of Question 16 of Homework 3 on Coursera.

***Sol.*** Given $h_y(x) = \dfrac{\exp(w_y^\top x)}{\sum_{k=1}^K \exp(w_k^\top x)}$, maximizing $\displaystyle\prod_{n=1}^N h_{y_n}(x_n)$ is minimizing $-\dfrac{1}{N}\displaystyle\sum_{n=1}^N \log h_{y_n}(x_n)$.

$$\begin{aligned} E_{\text{in}} &= -\frac{1}{N}\sum_{n=1}^N \log h_{y_n}(x_n) \\ &= -\frac{1}{N}\sum_{n=1}^N \left( \exp(w_{y_n}^\top x_n) - \log\left( \sum_{k=1}^K \exp(w_k^\top x_n) \right) \right) \\ &= \frac{1}{N}\sum_{n=1}^N \left( \log\left( \sum_{k=1}^K \exp(w_k^\top x_n) \right) - \exp(w_{y_n}^\top x_n) \right) \end{aligned}$$

$$\begin{aligned} \frac{\partial E_{\text{in}}}{\partial w_i} &= \frac{1}{N}\sum_{n=1}^N \left( \frac{\exp(w_i^\top x_n) x_n}{\sum_{k=1}^K \exp(w_k^\top x_n)} - [\![ y_n = i ]\!] \exp(w_{y_n}^\top x_n) x_n \right) \\ &= \frac{1}{N}\sum_{n=1}^N \Big( \big( h_i(x_n) - [\![ y_n = i ]\!] \exp(w_{y_n}^\top x_n) \big) x_n \Big) \end{aligned}$$

□

**Problem 5** (20 points)**.**

Write down the derivation steps of Question 11 of Homework 4 on Coursera.

***Sol.*** It is given that $X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{bmatrix}$, $y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$, $\tilde{X} = \begin{bmatrix} \tilde{x}_1^\top \\ \vdots \\ \tilde{x}_K^\top \end{bmatrix}$, $y = \begin{bmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_K \end{bmatrix}$, and

$$f(w) = \frac{1}{N+K}\left(\sum_{n=1}^{N}(y_n - w^\top x_n)^2 + \sum_{k=1}^{K}(\tilde{y}_k - w^\top \tilde{x}_k)^2\right).$$

Then,

$$\frac{\partial}{\partial w}\sum_{n=1}^{N}(y_n - w^\top x_n)^2 = \sum_{n=1}^{N} 2(y_n - w^\top x_n)(-x_n) = 2\sum_{n=1}^{N}(x_n^\top w - y_n)x_n$$

$$= 2\begin{bmatrix} x_1 & \cdots & x_N \end{bmatrix}\begin{bmatrix} x_1^\top w - y_1 \\ \vdots \\ x_N^\top w - y_N \end{bmatrix} = 2\begin{bmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{bmatrix}^\top \left(\begin{bmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{bmatrix} w - y\right) = 2X^\top(Xw - y).$$

Similarly,

$$\frac{\partial}{\partial w}\sum_{k=1}^{K}(\tilde{y}_k - w^\top \tilde{x}_k)^2 = 2\tilde{X}^\top(\tilde{X}w - \tilde{y}).$$

Since $f$ is differentiable, $w$ minimizes $f$ iff $w$ is a stationary point of $f$, i.e.,

$$\begin{aligned} & X^\top(Xw - y) + \tilde{X}^\top(\tilde{X}w - \tilde{y}) = 0 \\ \iff & (X^\top X + \tilde{X}^\top \tilde{X})w = x^\top y + \tilde{x}^\top \tilde{y} \\ \iff & w = (X^\top X + \tilde{X}^\top \tilde{X})^{-1}(X^\top y + \tilde{X}^\top \tilde{y}). \end{aligned}$$

□

**Problem 6** (20 points)**.**

Write down the derivation steps of Question 12 of Homework 4 on Coursera.

***Sol.*** Given Problem 5, define $g(w) = \frac{\lambda}{N}\|w\|^2 + \frac{1}{N}\|Xw - y\|^2$. Then, $w$ minimizes $g$ iff $w$ is a stationary point of $g$, i.e.,

$$\frac{2\lambda}{N}w + \frac{2}{N}X^\top(Xw - y) = 0 \iff (\lambda I + X^\top X)w = X^\top y \iff w = (\lambda I + X^\top X)^{-1}X^\top y.$$

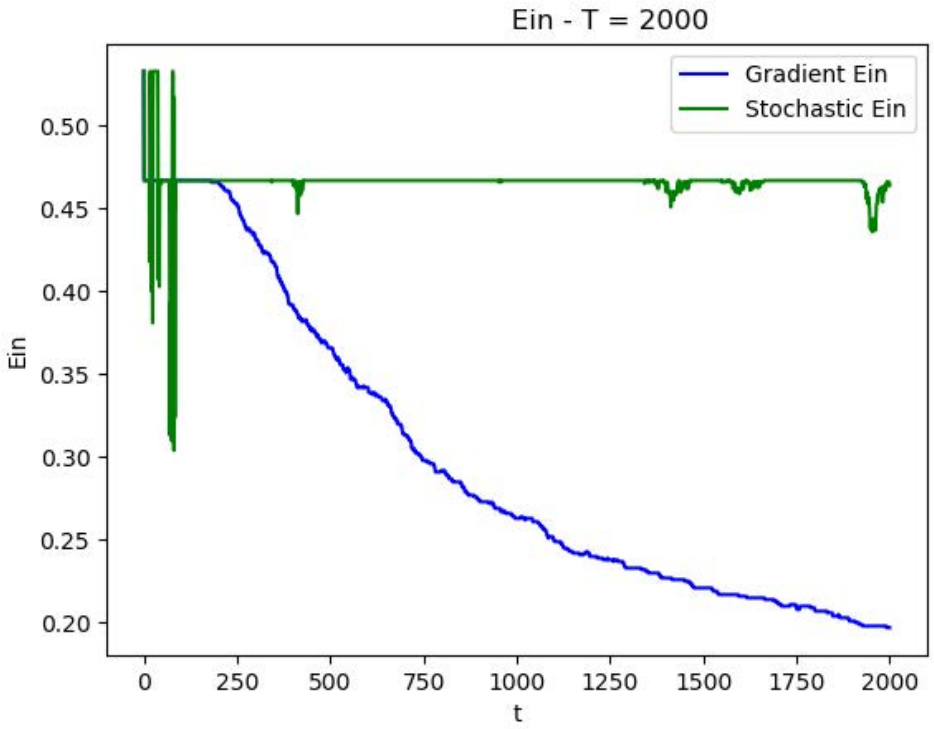
Thus, the minimizers of $f$ and $g$ agree iff

$$\begin{aligned}
&(X^\top X + \tilde{X}^\top \tilde{X})^{-1}(X^\top y + \tilde{X}^\top \tilde{y}) = (\lambda I + X^\top X)^{-1}X^\top y \\
\iff &(\lambda I + X^\top X)(X^\top y + \tilde{X}^\top \tilde{y}) = (X^\top X + \tilde{X}^\top \tilde{X})X^\top y \\
\iff &\lambda X^\top y + (\lambda I + X^\top X)\tilde{X}^\top \tilde{y} = \tilde{X}^\top \tilde{X} X^\top y \\
\iff &(X^\top X + \lambda I)\tilde{X}^\top \tilde{y} = (\tilde{X}^\top \tilde{X} - \lambda I)X^\top y.
\end{aligned}$$

Clearly, $\tilde{X} = \sqrt{\lambda}I$ forces $(\tilde{X}^\top \tilde{X} - \lambda I)X^\top y$ to 0, and $\tilde{y} = 0$ forces $(X^\top X + \lambda I)\tilde{X}^\top \tilde{y}$ to 0. Thus, $\tilde{X} = \sqrt{\lambda}I$ and $\tilde{y} = 0$ coerce the minimizers of $f$ and $g$ to agree. □

**Problem 7** (20 points)**.**

For Questions 19 and 20 of Homework 3 on Coursera, plot a figure that shows $E_{\text{in}}(w_t)$ as a function of $t$ for both the gradient descent version and the stochastic gradient descent version on the same figure. Describe your findings. Please print out the figure for grading.

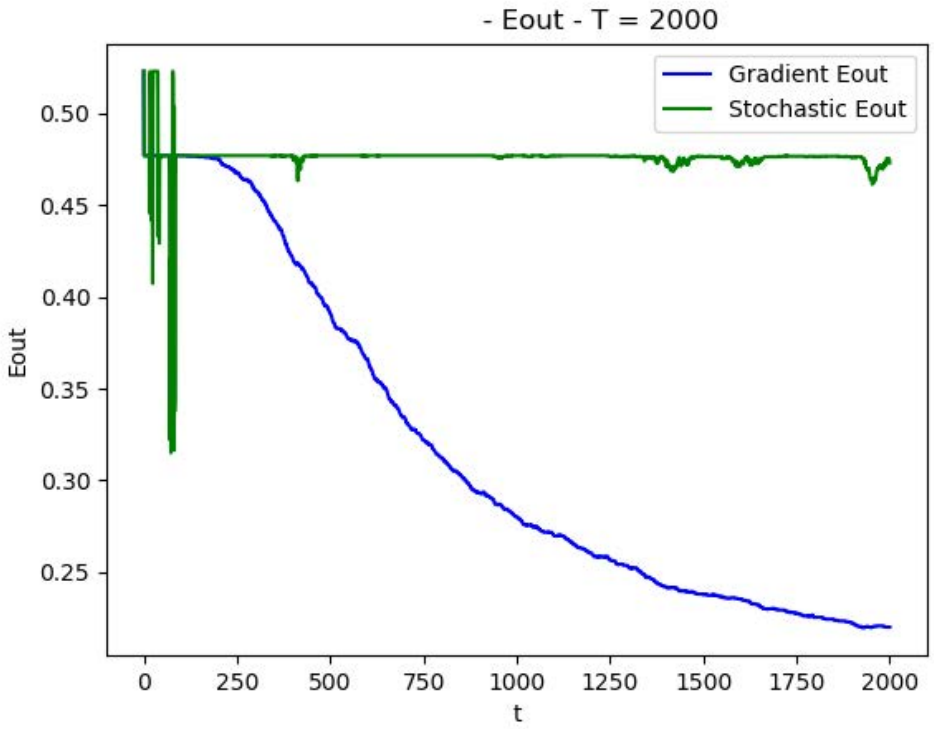



From this graph, we can see that from $w_0 = \mathbf{0}$, the training set error rate for gradient descent is above 0.5, quickly falling to around 0.47 and then dipping dramatically after $t = 250$, and finally stopping at 0.22 after 2000 iterations. This shows that using gradient descent, while significantly slower in terms of complexity, shows tangible improvement over each iteration.

In comparison, the training set error rate for stochastic gradient descent varies wildly for the first 100 or so iterations, before stabilizing afterwards at around 0.475, landing at approximately the same value after 2000 iterations. We can see that after some early volatility, $w$ eventually "finds" a stable vector with only minor fluctuations afterwards, though this vector does not show any major improvements at all over the course of 2000 iterations.

**Problem 8** (20 points)**.**

For Questions 19 and 20 of Homework 3 on Coursera, plot a figure that shows $E_{\text{out}}(w_t)$ as a function of $t$ for both the gradient descent version and the stochastic gradient descent version on the same figure. Describe your findings. Please print out the figure for grading




We can see similar trends as problem 7. Gradient descent shows clear improvement over the course of the training, whereas stochastic gradient descent shows stabilization (although the error rate leaves much to be desired).

**Problem 9** (20 points)**.** The SVD of a matrix $A \in \mathbb{R}^{p\times q}$ of rank $r$ is $A = U\Gamma V^\top$ such that $U \in \mathbb{R}^{p\times p}$ and $V \in \mathbb{R}^{q\times q}$ are orthogonal, i.e, $U^\top U = I_p$ and $V^\top V = I_q$, and

$$\Gamma = \begin{bmatrix} \Sigma & 0_{r\times(q-r)} \\ 0_{(p-r)\times r} & 0_{(p-r)\times(q-r)} \end{bmatrix} \in \mathbb{R}^{p\times q}$$

where $\Sigma \in \mathbb{R}^{r\times r}$ is diagonal with positive diagonal entries. Following previous notations, one defines

$$\Gamma^+ = \begin{bmatrix} \Sigma^{-1} & 0_{r\times(q-r)} \\ 0_{(p-r)\times r} & 0_{(p-r)\times(q-r)} \end{bmatrix}^\top = \begin{bmatrix} \Sigma^{-1} & 0_{r\times(p-r)} \\ 0_{(q-r)\times r} & 0_{(q-r)\times(p-r)} \end{bmatrix}$$

and the Moore-Penrose pseudo-inverse

$$A^\dagger = V\Gamma^+U^\top.$$

Given $X \in \mathbb{R}^{N\times(d+1)}$ of rank $\rho$ and $y \in \mathbb{R}^N$, show that $w_{\text{lin}} = X^\dagger y$ is a 2-norm-minimizing solution to

$$X^\top Xw = X^\top y. \tag{1}$$

P.S. The SVD formula given by TA is faulty.

***Sol.*** With the notations of the first part, one obtains the following.

$$\Gamma\Gamma^+ = \begin{bmatrix} \Sigma & 0_{r\times(q-r)} \\ 0_{(p-r)\times r} & 0_{(p-r)\times(q-r)} \end{bmatrix} \begin{bmatrix} \Sigma^{-1} & 0_{r\times(p-r)} \\ 0_{(q-r)\times r} & 0_{(q-r)\times(p-r)} \end{bmatrix} = \begin{bmatrix} I_r & 0_{r\times(p-r)} \\ 0_{(p-r)\times r} & 0_{(p-r)\times(p-r)} \end{bmatrix} = (\Gamma\Gamma^+)^\top$$

$$\Gamma^+\Gamma = \begin{bmatrix} \Sigma^{-1} & 0_{r\times(p-r)} \\ 0_{(q-r)\times r} & 0_{(q-r)\times(p-r)} \end{bmatrix} \begin{bmatrix} \Sigma & 0_{r\times(q-r)} \\ 0_{(p-r)\times r} & 0_{(p-r)\times(q-r)} \end{bmatrix} = \begin{bmatrix} I_r & 0_{r\times(q-r)} \\ 0_{(q-r)\times r} & 0_{(q-r)\times(q-r)} \end{bmatrix} = (\Gamma^+\Gamma)^\top$$

$$\Gamma^\top\Gamma\Gamma^+ = \begin{bmatrix} \Sigma^\top & 0_{r\times(p-r)} \\ 0_{(q-r)\times r} & 0_{(q-r)\times(p-r)} \end{bmatrix} \begin{bmatrix} I_r & 0_{r\times(p-r)} \\ 0_{(p-r)\times r} & 0_{(p-r)\times(p-r)} \end{bmatrix} = \begin{bmatrix} \Sigma^\top & 0_{r\times(p-r)} \\ 0_{(q-r)\times r} & 0_{(q-r)\times(p-r)} \end{bmatrix} = \Gamma^\top$$

$$\Gamma^+\Gamma\Gamma^+ = \begin{bmatrix} \Sigma^{-1} & 0_{r\times(p-r)} \\ 0_{(q-r)\times r} & 0_{(q-r)\times(p-r)} \end{bmatrix} \begin{bmatrix} I_r & 0_{r\times(p-r)} \\ 0_{(p-r)\times r} & 0_{(p-r)\times(p-r)} \end{bmatrix} = \begin{bmatrix} \Sigma^{-1} & 0_{r\times(p-r)} \\ 0_{(q-r)\times r} & 0_{(q-r)\times(p-r)} \end{bmatrix} = \Gamma^+$$

$$\begin{aligned}(AA^\dagger)^\top &= ((U\Gamma V^\top)(V\Gamma^+U^\top))^\top = (U(\Gamma\Gamma^+)U^\top)^\top = U(\Gamma\Gamma^+)^\top U^\top = U(\Gamma\Gamma^+)U^\top \\ &= (U\Gamma V^\top)(V\Gamma^+U^\top) = AA^\dagger\end{aligned}$$

$$\begin{aligned}(A^\dagger A)^\top &= ((V\Gamma^+U^\top)(U\Gamma V^\top))^\top = (V(\Gamma^+\Gamma)V^\top)^\top = V(\Gamma^+\Gamma)^\top V^\top = V(\Gamma^+\Gamma)V^\top \\ &= (V\Gamma^+U^\top)(U\Gamma V^\top) = A^\dagger A\end{aligned}$$

$$A^\top AA^\dagger = (V\Gamma^\top U^\top)(U\Gamma V^\top)(V\Gamma^+U^\top) = V(\Gamma^\top\Gamma\Gamma^+)U^\top = V\Gamma^\top U^\top = A^\top$$
$$A^\dagger AA^\dagger = (V\Gamma^+U^\top)(U\Gamma V^\top)(V\Gamma^+U^\top) = V(\Gamma^+\Gamma\Gamma^+)U^\top = V\Gamma^+U^\top = A^\dagger$$

Now introduce the notations of the second part.
Since $X^\top X w_{\text{lin}} = X^\top X X^\dagger y = X^\top y$, $w_{\text{lin}}$ is a solution to Eq. (1). Let $w$ be another solution to Eq. (1). Since $w_{\text{lin}} = X^\dagger y = X^\dagger X X^\dagger y = X^\dagger X w_{\text{lin}}$,

$$\begin{aligned}
{w_{\text{lin}}}^\top (w - w_{\text{lin}}) &= ((X^\dagger X)(X^\dagger X) w_{\text{lin}})^\top (w - w_{\text{lin}}) \\
&= {w_{\text{lin}}}^\top (X^\dagger X)^\top (X^\dagger X)^\top (w - w_{\text{lin}}) \\
&= {w_{\text{lin}}}^\top (X^\dagger X)(X^\dagger X)(w - w_{\text{lin}}) \\
&= {w_{\text{lin}}}^\top X^\dagger (X X^\dagger) X (w - w_{\text{lin}}) \\
&= {w_{\text{lin}}}^\top X^\dagger (X X^\dagger)^\top X (w - w_{\text{lin}}) \\
&= {w_{\text{lin}}}^\top X^\dagger (X^\dagger)^\top X^\top X (w - w_{\text{lin}}) \\
&= {w_{\text{lin}}}^\top X^\dagger (X^\dagger)^\top (X^\top X w - X^\top X w_{\text{lin}}) \\
&= {w_{\text{lin}}}^\top X^\dagger (X^\dagger)^\top (X^\top y - X^\top y) \\
&= 0.
\end{aligned}$$

Thus,

$$\begin{aligned}
w^\top w &= (w - w_{\text{lin}} + w_{\text{lin}})^\top (w - w_{\text{lin}} + w_{\text{lin}}) \\
&= (w - w_{\text{lin}})^\top (w - w_{\text{lin}}) + 2 {w_{\text{lin}}}^\top (w - w_{\text{lin}}) + {w_{\text{lin}}}^\top w_{\text{lin}} \\
&\geq 0 + 0 + {w_{\text{lin}}}^\top w_{\text{lin}}.
\end{aligned}$$

□

*Remark.* The derivation above seems ad hoc and lengthy, for the general scheme is obscured. Upon first contact with Eq. (1), namely, $X^\top X w = X^\top y$, the most natural attempt to solve for $w$ would be to compute $w_{\text{lin}} = (X^\top X)^\dagger X^\top y$; only to find out later that $(X^\top X)^\dagger X^\top = X^\dagger$, yielding $w_{\text{lin}} = X^\dagger y$, which the assignment designer considers a priori. In this case, the norm-minimizng property of $w_{\text{lin}}$ is manifested as a well-known fact. What follows constitutes the proof of $(X^\top X)^\dagger X^\top = X^\dagger$.

$$\begin{aligned}
(X^\top X)^\dagger X^\top &= (V \Gamma^\top \Gamma V^\top)^\dagger X^\top \\
&= V (\Gamma^\top \Gamma)^+ \Gamma^\top U^\top \\
&= V \begin{bmatrix} \Sigma^2 & 0_{\rho \times (d+1-\rho)} \\ 0_{(d+1-\rho)\times \rho} & 0_{(d+1-\rho)\times(d+1-\rho)} \end{bmatrix}^+ \Gamma^\top U^\top \\
&= V \begin{bmatrix} \Sigma^{-2} & 0_{\rho \times (d+1-\rho)} \\ 0_{(d+1-\rho)\times \rho} & 0_{(d+1-\rho)\times(d+1-\rho)} \end{bmatrix} \begin{bmatrix} \Sigma & 0_{\rho \times (N-\rho)} \\ 0_{(d+1-\rho)\times \rho} & 0_{(d+1-\rho)\times(N-\rho)} \end{bmatrix} U^\top \\
&= V \begin{bmatrix} \Sigma^{-1} & 0_{\rho \times (N-\rho)} \\ 0_{(d+1-\rho)\times \rho} & 0_{(d+1-\rho)\times(N-\rho)} \end{bmatrix} U^\top \\
&= V \Gamma^+ U^\top \\
&= X^\dagger
\end{aligned}$$