Student Name: Mingqian Liao
Student ID: 3038745426

Deep Neural Networks(EECS182)
Homework1

---

1. Bias-Variance Tradeoff Review

(a)

Assume that $Y = f(x) + \epsilon$

$$\begin{aligned}
E[(Y-\hat{f})] &= E[(f+\epsilon-\hat{f})^2] \\
&= E[(f+\epsilon-\hat{f}+E[\hat{f}]-E[\hat{f}])^2] \\
&= E[(f-E[\hat{f}])^2] + E[\epsilon^2] + E[(E[\hat{f}]-\hat{f})^2] + 2E[(f-E[\hat{f}])\epsilon] + 2E[\epsilon(E[\hat{f}]-\hat{f})] \\
&+ 2E[(E[\hat{f}]-\hat{f})(f-E[\hat{f}])] \\
&= E[(f-E[\hat{f}])^2] + E[\epsilon^2] + E[(E[\hat{f}]-\hat{f})^2] \\
&= E[(f-E[\hat{f}])^2] + Var[\epsilon] + Var[f] \\
&= Bias[\hat{f}]^2 + Var[\epsilon] + Var[f] \\
&= Bias[\hat{f}]^2 + \sigma^2 + Var[f]
\end{aligned}$$

(b)

$$\begin{aligned}
E[\hat{\theta}] &= E[(X^TX)^{-1}X^TY] \\
&= E[(X^TX)^{-1}X^T(X\theta+\epsilon)] \\
&= E[(X^TX)^{-1}X^TX\theta + (X^TX)^{-1}X^T\epsilon] \\
&= E[\theta] + E[(X^TX)^{-1}X^T]E[\epsilon] \\
&= \theta
\end{aligned}$$

Therefore, the bias of $\hat{\theta}$ is 0.

Next we compute the covariance of $\hat{\theta}$.

$$\begin{aligned}
E[(\hat{\theta}-\theta)(\hat{\theta}-\theta)^T] &= E[(X^TX)^{-1}X^T\epsilon\epsilon^T((X^TX)^{-1}X^T)^T] \\
&= (X^TX)^{-1}X^TE[\epsilon\epsilon^T]X((X^TX)^{-1})^T \\
&= (X^TX)^{-1}X^TI_n((X^TX)^{-1}X^T)^T \\
&= (X^TX)^{-1}
\end{aligned}$$

With a particular input x, the bias is still 0, and the variance is

$$\begin{aligned}
Var[x^T(\hat{\theta}-\theta)] &= E[x^T(\hat{\theta}-\theta)(\hat{\theta}-\theta)x] \\
&= x^T(X^TX)^{-1}x
\end{aligned}$$

2. Least Squares and the Min-norm problem from the Perspective of SVD

(a)

Let $L = \|Xw - y\|^2$, then

$$\begin{aligned} dL &= 2(Xw - y)^T X dw \\ \implies \frac{\partial L}{\partial w} &= 2X^T(Xw - y) \end{aligned}$$

Let $\frac{\partial L}{\partial w} = 0$, then

$$w = (X^T X)^{-1} X^T y$$

(b)

$$\begin{aligned} w &= (X^T X)^{-1} X^T y \\ &= (V\Sigma^T \Sigma V^T)^{-1} V \Sigma^T U^T y \\ &= V(\Sigma^T \Sigma)^{-1} V^T V \Sigma^T U^T y \\ &= V(\Sigma^T \Sigma)^{-1} \Sigma^T U^T y \\ &= V \Sigma^\dagger U^T y \end{aligned}$$

(c)

When we left-multiply X by our $A = V\Sigma^\dagger U^T$, since $X = U\Sigma V^T$ we get

$$AX = V\Sigma^\dagger U^T U \Sigma V^T = I$$

(d)

Since it is a optimization problem with constraints $Xw - y = 0$, we can use the Lagrange method to solve it. Let $L(w, \lambda) = \|w\|^2 + \lambda(Xw - y)$, we have

$$\frac{\partial L}{\partial w} = 2w - X^T \lambda = 0 \tag{1a}$$

$$\frac{\partial L}{\partial \lambda} = 2w - X^T \lambda \tag{1b}$$

From formula1a and 1b, we can obtian that $y = \frac{1}{2} X X^T \lambda$, so that $\lambda = 2(XX^T)^{-1} y$, plug it into formula1a, we can obtain

$$w = X^T (XX^T)^{-1} y$$

(e)

By means of $X = U\Sigma V^T$ and (d), we can obtain that

$$\begin{aligned} w &= X^T (XX^T)^{-1} y \\ &= V\Sigma^T U^T (U\Sigma V^T V \Sigma^T U^T)^{-1} y \\ &= V\Sigma^T U^T U (\Sigma\Sigma^T)^{-1} U^T y \\ &= V\Sigma^T (\Sigma\Sigma^T)^{-1} U^T y \\ &= V\Sigma^\dagger U^T y \end{aligned}$$

Awesome SVD Lagrange!!!

(f)

From (e), we can obtian that $B = V\Sigma^{\dagger}U^T$, iif we right-multiply X by B, we can get

$$\begin{aligned} XB &= U\Sigma V^T V\Sigma^{\dagger}U^T \\ &= I \end{aligned}$$

3. The 5 Interpretations of Ridge Regression

(a)

Let $L = \|y - Xw\|_2^2 + \lambda\|w\|^2$, then

$$\begin{aligned} dL &= d((y - Xw)^T)(y - Xw) + (y - Xw)^T d(y - Xw) + \lambda d\lambda w^T w + \lambda w^T dw \\ &= -2(y - Xw)^T X dw + 2\lambda w^T dw \end{aligned}$$

Therefore,

$$\frac{\partial L}{\partial x} = 2\lambda w^T - 2X^T(y - Xw)^T$$

Let $\frac{\partial L}{\partial x} = 0$, then

$$w = (X^T X + \lambda I)X^T y$$

(b)

$$\begin{aligned} w &= (X^T X + \lambda I)^{-1} X^T y \\ &= (V\Sigma^T \Sigma V^T + \lambda I)^{-1} V\Sigma^T U^T y \\ &= V(\Sigma^T\Sigma + \lambda I)^{-1}\Sigma\widetilde{y} \\ \Longrightarrow \widetilde{w} &= (\Sigma^T\Sigma + \lambda I)^{-1}\Sigma\widetilde{y} \\ \Longrightarrow \widetilde{w}[i] &= \frac{\sigma_i}{\sigma_i^2 + \lambda}\widetilde{y}[i] \end{aligned}$$

If $\lambda \gg \sigma_i^2$, then the value of $\widetilde{w}[i]$ is very small, close to zero, which is like $\frac{\sigma}{\lambda}\widetilde{y}[i]$

If $\lambda \ll \sigma_i^2$, then the value of $\widetilde{w}[i]$ is like $\frac{1}{\sigma_i^2}\widetilde{y}[i]$, the same as the unregularized case.

(c)

Since $Y = Xw + \sqrt{\lambda}N$, therefore $y_i = X_i^T w + \sqrt{\lambda}N_i$, so $y_i \sim N(X_i^T w, \lambda)$, then

$$
\begin{aligned}
MAP(w|Y=y) &= \arg\max_w L(w|Y=y) \\
&= \arg\max_w \frac{L(w,y)}{L(y)} \\
&= \arg\max_w L(w) * L(y|w) \\
&= \arg\max_w \frac{1}{\sqrt{2\pi}} e^{-\frac{\|w\|^2}{2}} * \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\lambda}} e^{\frac{-(y_i - X_i^T w)^2}{2\lambda}} \\
&= \arg\max_w -\frac{\|w\|^2}{2} + \sum_{i=1}^{n} \frac{-(y_i - X_i^T w)^2}{2\lambda} \\
&= \arg\max_w -\frac{\|w\|^2}{2} + \frac{-(y - Xw)^2}{2\lambda} \\
&= (y - Xw)^2 + \arg\min_w \lambda\|w\|^2
\end{aligned}
$$

(d)

We can directly use the OLS step to conclude that

$$
\begin{aligned}
w &= (\hat{X}^T \hat{X})^{-1} \hat{X}^T \hat{y} \\
&= (\begin{bmatrix} X & \sqrt{\lambda} I_d \end{bmatrix} \begin{bmatrix} X \\ \sqrt{\lambda} I_d \end{bmatrix})^{-1} \begin{bmatrix} X & \sqrt{\lambda} I_d \end{bmatrix} \begin{bmatrix} y \\ 0_d \end{bmatrix} \\
&= (X^T X + \lambda I_d)^{-1} X y
\end{aligned}
$$

(e)

Since when conducting the calculation, the $\sqrt{\lambda} I$ part of $\hat{X}$ will become 0 because the last n dimensions of the $d+n$ is 0 in y, which will serve as an ridge regression that constrait the $\lambda\|w\|$ to zero, where w is the first d coordinates of η^*, and it forms the minimizer of (1), since the rest rows of η yield 0.

(f)

We can use the Moore-Penrose Pseudo inverse to find a min-norm solution to the problem. Combining the OLS soluion, we get

$$
\begin{aligned}
\eta &= \begin{bmatrix} X^T \\ \sqrt{\lambda} I \end{bmatrix} (\begin{bmatrix} X & \lambda I \end{bmatrix} \begin{bmatrix} X^T \\ \sqrt{\lambda} I \end{bmatrix})^{-1} y \\
\implies w &= X^T (X X^T + \lambda I)^{-1} y
\end{aligned}
$$

$$
\begin{aligned}
\hat{w} &= X^T(XX^T+\lambda I)^{-1}y \\
&= (X^TX+\lambda I)^{-1}(X^TX+\lambda I)X^T(XX^T+\lambda I)^{-1}y \\
&= (X^TX+\lambda I)^{-1}(X^TXX^T+\lambda X^T)(XX^T+\lambda I)^{-1}y \\
&= (X^TX+\lambda I)^{-1}X^T(XX^T+\lambda I)(XX^T+\lambda I)^{-1}y \\
&= (X^TX+\lambda I)^{-1}X^Ty
\end{aligned}
$$

(g)

The $\lambda \longrightarrow$ means that the $(X^TX+\lambda I)^{-1}$ converges to the zero matrix, makes $w=0$

(h) When the $\lambda \longrightarrow 0$, and when the matrix is tall, then $w=(X^TX)^{-1}X^Ty$; when the matrix is wide, then $w=X^T(X^TX)^{-1}y$

4. General Case Tikhonov Regularization

(a)

Let $L=\|W_1(Ax-b)\|_2^2+\|W_2(x-c)\|_2^2$, then we have

$$
\begin{aligned}
dl &= 2(W_1(Ax-b))^TW_1Adx+2(W_2(x-c))^T(W_2(x-c)) \\
&= 2[(W_1(Ax-b))^TW_1A+(W_2(x-c))^TW_2]dx \\
\Longrightarrow \frac{\partial l}{\partial x} &= 2((A^TW_1^TW_1A+W_2^TW_2)x-A^TW_1^TW_1b-W_2^TW_2c)
\end{aligned}
$$

Set $\frac{\partial l}{\partial x}=0$, so that we can derive

$$
x=(A^TW_1^TW_1A+W_2^TW_2)^{-1}(A^TW_1^TW_1b+W_2^TW_2c)
$$

(b)

$$
C=\begin{bmatrix} W_A \\ W_2 \end{bmatrix}
$$

$$
d=\begin{bmatrix} W_1b \\ W_2c \end{bmatrix}
$$

then,

$$
\begin{aligned}
x^* &= (C^TC)^{-1}C^Td \\
&= x=(A^TW_1^TW_1A+W_2^TW_2)^{-1}(A^TW_1^TW_1b+W_2^TW_2c)
\end{aligned}
$$

This formula is in agreement with the previous part.

(c)

I will choose the W_1 that makes $W_1^TW_1=I$, and choose the W_2 so that $W_2^TW_2=\lambda I$, and $c=0$

5. Coding Fully Connected Networks

(a) Yes, it will take longer to train the five layer net than three layer net.

6. Visualizing features from local linearization of neural nets
7. Homework Process and Study Group
   (a) The tutorial of numpy
   (b) Name: Chuanchen SID: 3038743333
   (c) Approximately 16 hours