Machine Learning from Data CSCI 4100
Assignment 6
Jae Park (RIN: 661994900)

Exercise 3.4

a) Given $y = w^{*T}X + \epsilon$, and from LFD pg.86, $\hat{y} = X(X^TX)^{-1}X^Ty$. Substituting y, we get

$$\hat{y} = X(X^TX)^{-1}X^T(w^{*T}X + \epsilon)$$
$$= \mathrm{X}(\mathrm{X}^TX)^{-1}X^T(Xw^* + \epsilon)$$
$$= \mathrm{X}(\mathrm{X}^TX)^{-1}(X^TX)w^* + X(X^TX)^{-1}X^T\epsilon$$
$$= \mathrm{Xw}^* + X(X^TX)^{-1}X^T\epsilon$$

and we know $H = X(X^TX)^{-1}X^T$ from LFD (3.6), so

$$= Xw^* + H\epsilon$$

b) $\hat{y} - y = (Xw^* + H\epsilon) - (Xw^* + \epsilon) = H\epsilon - \epsilon = (H - I)\epsilon$
So, the matrix of interest is $(H - I)$

c)

$$E_{in}(w_{lin}) = \frac{1}{N}\|\hat{y} - y\|^2$$
$$= \frac{1}{N}\|(H - I)\epsilon\|^2$$
$$= \frac{1}{N}\epsilon^T(H - I)^T \cdot (H - I)\epsilon$$

From **Exercise 3.3** a and c, we can take a couple of facts; $H = X(X^TX)^{-1}X^T$ and it is symmetric, also $(I - H)^K = I - H$. With these in mind,

$$= \frac{1}{N}\epsilon^T(H - I)^T(H - I)\epsilon$$
$$= \frac{1}{N}\epsilon^T(H - I)^T(H - I)\epsilon$$
$$= \frac{1}{N}\epsilon^T(H - I)^2\epsilon$$
$$= \frac{1}{N}\epsilon^T(I - H)^2\epsilon$$
$$= \frac{1}{N}\epsilon^T(I - H)\epsilon$$

d)

$$\mathbb{E}_{\mathcal{D}}[E_{in}(w_{lin})] = \mathbb{E}_D\left[\frac{1}{N}\epsilon^T(I - H)\epsilon\right]$$
$$= \mathbb{E}_{\mathcal{D}}\left[\frac{1}{N}\ \epsilon^T\epsilon - \frac{1}{N}\epsilon^TH\epsilon\right]$$
$$= \mathbb{E}_{\mathcal{D}}\left[\frac{1}{N}\epsilon^T\epsilon\right] - \mathbb{E}_{\mathcal{D}}\left[\frac{1}{N}\epsilon^TH\epsilon\right]$$
$$= \frac{1}{N}\mathbb{E}_{\mathcal{D}}[\epsilon^T\epsilon] - \frac{1}{N}\mathbb{E}_{\mathcal{D}}[\epsilon^TH\epsilon]$$

ϵ is a noise term with zero mean and σ^2 variance, defined by the problem.
We know that $\mathbb{E}_{\mathcal{D}}\left[\epsilon^T\epsilon\right]$ term is equal to $N\sigma^2$.

$\mathbb{E}_{\mathcal{D}}\left[\epsilon^T H\epsilon\right] = \sigma^2 \cdot tr(H)$. $\epsilon^T H\epsilon$ forms a diagonal matrix composed of σ^2's and H, since ϵ is independent with zero mean and σ^2 variance.
Trace of H is given as $(d+1)$ from **Exercise 3.3** (c)

$$\therefore \mathbb{E}_{\mathcal{D}}\left[\epsilon^T H\epsilon\right] = \sigma^2(d+1).$$

So,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[E_{in}(w_{lin})] &= \frac{1}{N}\mathbb{E}_{\mathcal{D}}[\epsilon^T\epsilon] - \frac{1}{N}\mathbb{E}_{\mathcal{D}}[\epsilon^T H\epsilon] \\
&= \frac{1}{N}(N\sigma^2 - \sigma^2(d+1)) \\
&= \sigma^2 - \frac{\sigma^2}{N}(d+1) \\
&= \sigma^2(1 - \frac{d+1}{N}) \text{ for } N \geq d+1
\end{aligned}$$

e)

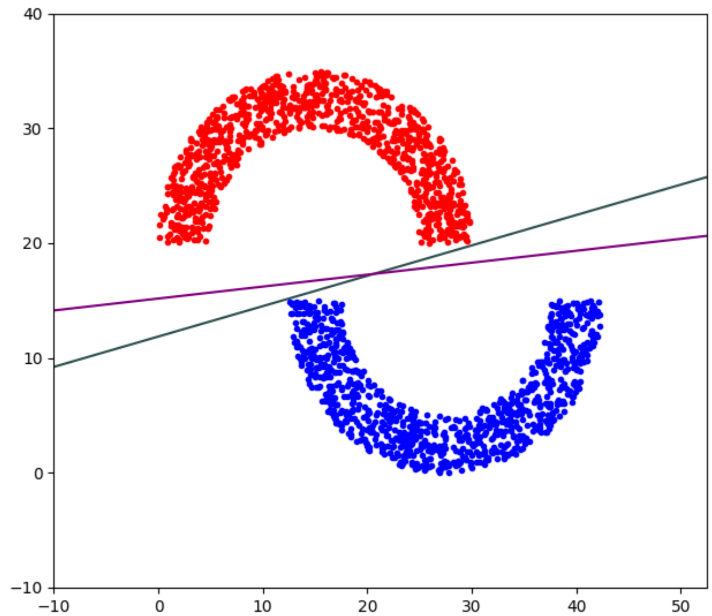
$$\begin{aligned}
\mathbb{E}_{\mathcal{D},\epsilon'}[E_{test}(w_{lin})] &= \mathbb{E}_{\mathcal{D},\epsilon'}\left[\frac{1}{N}\left\|H\epsilon - \epsilon'\right\|^2\right] \\
&= \frac{1}{N}\mathbb{E}_{\mathcal{D},\epsilon'}\left[\left\|H\epsilon - \epsilon'\right\|^2\right] \\
&= \frac{1}{N}\mathbb{E}_{\mathcal{D},\epsilon'}\left[(H\epsilon - \epsilon')^T(H\epsilon - \epsilon')\right] \\
&= \frac{1}{N}\mathbb{E}_{\mathcal{D},\epsilon'}\left[((\epsilon^T H^T)H\epsilon - (\epsilon^T H^T)\epsilon' - (\epsilon')^T H\epsilon + (\epsilon')^T\epsilon'\right]
\end{aligned}$$

because H is symmetric, $H^T H = H^2 = H \ldots$ according to **Exercise 3.3**$(a),(b)$.

$$\begin{aligned}
&= \frac{1}{N}\mathbb{E}_{\mathcal{D},\epsilon'}\left[(\epsilon^T H^T H\epsilon - \epsilon^T H^T\epsilon' - (\epsilon')^T H\epsilon + (\epsilon')^T\epsilon'\right] \\
&= \frac{1}{N}\mathbb{E}_{\mathcal{D},\epsilon'}\left[(\epsilon^T H\epsilon - \epsilon^T H^T\epsilon' - (\epsilon')^T H\epsilon + (\epsilon')^T\epsilon'\right] \\
&= \frac{1}{N}\mathbb{E}_{\mathcal{D},\epsilon'}\left[(\epsilon^T H\epsilon - \epsilon^T H^T\epsilon' - (\epsilon^T H^T\epsilon')^T + (\epsilon')^T\epsilon'\right] \\
&\cdots \\
&= \frac{1}{N}\times\left\{\mathbb{E}_{\mathcal{D},\epsilon'}[\epsilon^T H\epsilon] + \underbrace{\mathbb{E}_{\mathcal{D},\epsilon'}[-\epsilon^T H^T\epsilon']}_{=0} + \underbrace{\mathbb{E}_{\mathcal{D},\epsilon'}[-(\epsilon^T H^T\epsilon')^T]}_{=0} + \mathbb{E}_{\mathcal{D},\epsilon'}[(\epsilon')^T\epsilon']\right\} \\
&= \frac{1}{N}\times\left\{\sigma^2(d+1) + 0 + 0 + N\sigma^2\right\} \\
&= \sigma^2\left(1 + \frac{d+1}{N}\right)
\end{aligned}$$

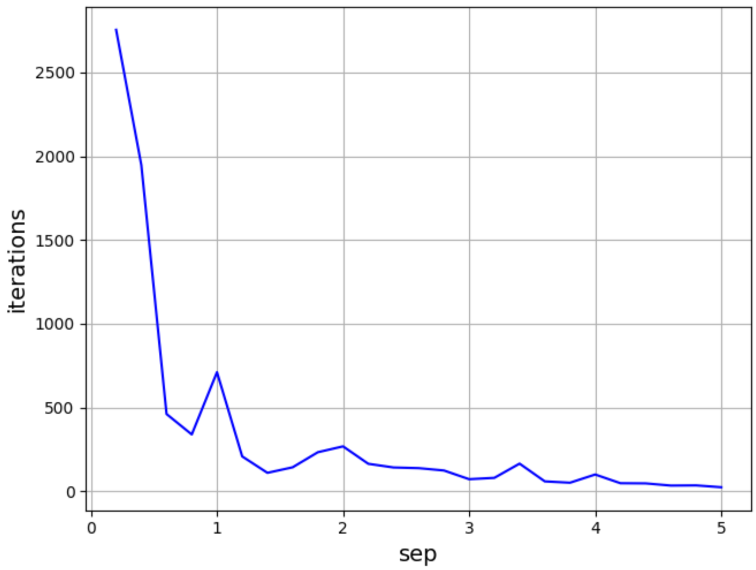
Problem 3.1

a) $w = [-866, -19.31, 73.02]$. PLA is shown in slate gray.



b) $w = [-1.19182, -0.00816, 0.07856]$. Linear regression is shown in purple.
We can see that linear regression can be used for classification, like PLA. Both PLA and Linear regression separated all the data points. Also, w_{lin} could be a good approximation for the perceptron model.

Problem 3.2



With *sep* in the range of $\{0.2, 0.4, ..., 5\}$, we generate 2,000 examples and run the PLA starting with $\mathbf{w} = \mathbf{0}$. We can see that the number of iterations quickly converges to a certain level. This is in agreement with what we had shown from **Problem 1.3** (e) in homework 2, and that is

$$t \leq \frac{R^2||\mathbf{w}^*||^2}{\rho^2}.$$

Here, ρ increases as *sep* increases, making the upper bound smaller. If *sep* gets bigger, it means that the distance between the double-semi-circle increases, which allows for less iterations to take place before PLA converges.

Problem 3.8
Given $E_{out}(h) = \mathbb{E}[h(x) - y)^2]$, we have

$$\begin{aligned}
E_{out}(h) &= \mathbb{E}[h(x) - y)^2] \\
&= \mathbb{E}[(h(x) - y - h^*(x) + h^*(x))^2] \\
&= \mathbb{E}[((h(x) - h^*(x)) + ((h^*(x) - y))^2] \\
&= \mathbb{E}[(h(x) - h^*(x))^2 + (h^*(x) - y)^2 + 2(h(x) - h^*(x))(h^*(x) - y)] \\
&= \mathbb{E}[(h(x) - h^*(x))^2] + \mathbb{E}[(h^*(x) - y)^2] + 2\mathbb{E}[(h(x) - h^*(x))(h^*(x) - y)]
\end{aligned}$$

Looking at the last term $2\mathbb{E}[(h(x) - h^*(x))(h^*(x) - y)]$, we can see that

$$2\mathbb{E}[(h(x) - h^*(x))(h^*(x) - y)] = 2 \times \mathbb{E}[(h(x) - h^*(x))] \times \mathbb{E}[(h^*(x) - y)|x]$$

The last term $\mathbb{E}[(h^*(x) - y)|x]$ can be rewritten as, and using $h^*(x) = \mathbb{E}[y|x]$ given by the problem

$$\mathbb{E}[(h^*(x) - y)|x] = \mathbb{E}[h^*(x)|x] - \mathbb{E}[y|x] = h^*(x) - h^*(x) = 0$$

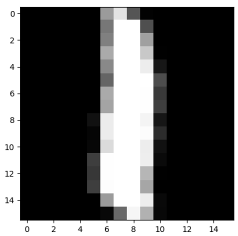
So, we find that

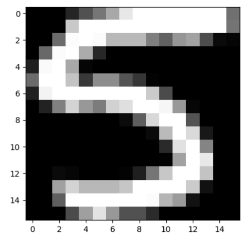
$$E_{out}(h) = \mathbb{E}[\underbrace{(h(x) - h^*(x))^2}_{\text{non-negative}}] + \mathbb{E}[\underbrace{(\mathbf{h}^*(\mathbf{x}) - y)^2}_{\text{non-negative}}]$$

From this, we can see that out of all the hypotheses, $h^*(x)$ is the one that will minimize $E_{out}(h)$.
Given $y = h^*(x) + \epsilon(x)$,

$$\begin{aligned}
y &= h^*(x) + \epsilon(x) \\
&\Downarrow \\
\mathbb{E}[y] &= \mathbb{E}[h^*(x)] + \mathbb{E}[\epsilon(x)] \\
\mathbb{E}[y|x] &= \mathbb{E}[h^*(x)|x] + \mathbb{E}[\epsilon(x)|x] \\
h^*(x) &= h^*(x) + \mathbb{E}[\epsilon(x)|x] \\
0 &= \mathbb{E}[\epsilon(x)]
\end{aligned}$$

Handwritten Digits a)





b)Let $image[i][j]$ define the intensity of the pixel location in an image size of (16×16). Then the average intensity is given as

$$intensity\_val = \frac{1}{256} \sum_{i=0}^{15} \sum_{j=0}^{15} image[i][j]$$

Vertical symmetry is defined as

$$symmetry\_val = \frac{1}{256} \sum_{i=0}^{7} \sum_{j=0}^{15} |\, image[i][j] - image[15 - i][j] \,|$$

c)