1. Exercise 3.4

(a) we know $y = w^{*T}x + \epsilon$ and $H = X(X^TX)^{-}1X^T$ from (3.6), and we know $\hat{y} = Hy$ by definition, we want to prove $\hat{y} = Xw^* + H\epsilon$

$$\begin{aligned}\hat{y} &= H(w^*X + \epsilon) \\ &= X(X^TX)^{-1}X^T(w^*X + \epsilon) \\ &= X(X^TX)^{-1}X^Tw^*X + X(X^TX)^{-1}X^T\epsilon \\ &= w^*X + H\epsilon\end{aligned}$$

(b) for $\hat{y} - y$, we have

$$\begin{aligned}\hat{y} - y &= w^*X + H\epsilon - (w^* + \epsilon) \\ &= H\epsilon - \epsilon \\ &= \epsilon(H - I)\end{aligned}$$

where I denotes the identity matrix

(c) let $E_{in}(w) = \frac{1}{N}||\hat{y} - y||^2$

$$\begin{aligned}E_{in}(w) &= \frac{1}{N}||\epsilon(H - I)||^2 \\ &= \frac{1}{N}(\epsilon(H - I))^T(\epsilon(H - I)) \\ &= \frac{1}{N}\epsilon^T(H - I)^T\epsilon(H - I)\end{aligned}$$

We know $H - I$ is symmetric, so $(H - I)^T = (H - I)$

$$\begin{aligned}E_{in}(w) &= \frac{1}{N}\epsilon^T\epsilon(H - I)^2 \\ &= \frac{1}{N}\epsilon^T\epsilon(I - H)^2\end{aligned}$$

(d) We know

$$\begin{aligned}E_D[E_{in}(w_{lin})] &= E_D[\frac{1}{N}(\epsilon^T\epsilon(I - H))] \\ &= \frac{1}{N}(E_D[\epsilon^T\epsilon] - E_D[\epsilon^T\epsilon H])\end{aligned}$$

Given that ϵ is a noise term with zero mean and σ^2 variance. The variance of each noise component ϵ is σ^2, so

$$\begin{aligned} E_D[E_{in}(w_{lin})] &= \frac{1}{N}(N\sigma^2 - E_D[\epsilon^T \epsilon H]) \\ &= \sigma^2 - \frac{1}{N} E_D[\epsilon^T \epsilon H] \end{aligned}$$

Now we can calculate

$$\begin{aligned} E_D[\epsilon^T \epsilon H] &= E_D[\sum_{i=1}^{N} \epsilon_i^2 H] \\ &= H \sum_{i=1}^{N} E_D[\epsilon_i^2] \end{aligned}$$

By the problem, we know that each component of ϵ is a random variable with zero mean and variance σ^2, so this means that $E_D[\epsilon_i] = 0$ and $E_D[\epsilon_i^2] = \sigma^2$ for all i.

$$\begin{aligned} E_D[\epsilon^T \epsilon H] &= H \sum_{i=1}^{N} \sigma^2 \\ &= HN\sigma^2 \end{aligned}$$

We continue the problem by substituting the result into our original equation

$$\begin{aligned} E_D[E_{in}(w_{lin})] = \sigma^2 - \frac{1}{N} E_D[\epsilon^T \epsilon H] &= \sigma^2 - \frac{1}{N} HN\sigma^2 \\ &= \sigma^2 - H\sigma^2 \end{aligned}$$

Then, we can calculate for the $trace(H)$

$$\begin{aligned} trace(H) &= trace(X(X^T X)^{-1} X^T) \\ &= trace((X^T X)^{-1}(X^T X)) \end{aligned}$$

Given that $X^T X$ is a square matrix of size $(d + 1)$, and it's inverse $(X^T X)^{-1}$ is also present, then we have

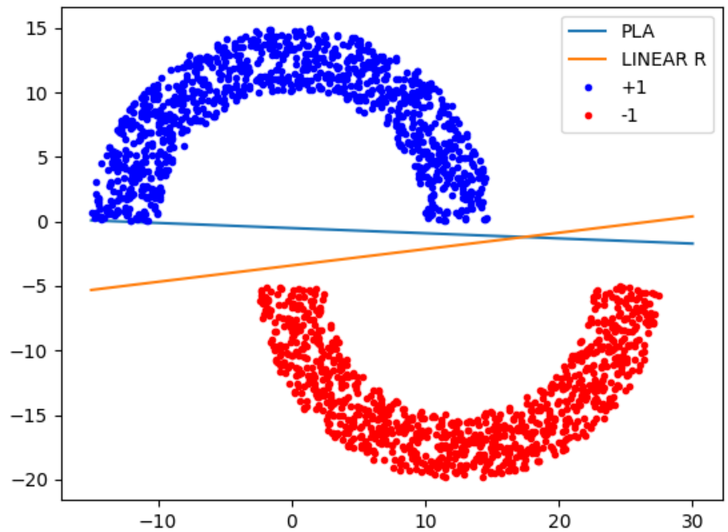
$$\begin{aligned} trace((X^T X)^{-1}(X^T X)) &= d + 1 \\ trace(H) &= \frac{d+1}{N} \end{aligned}$$

Then, we have proved that $E_D[E_{in}(w_{lin})] = \sigma^2(1 - \frac{d+1}{N})$ ■

(e) to do

$$E_{D,\epsilon'}[E_{test}(w_{lin})] = E_{D,\epsilon'}[\frac{1}{N}||Xw - y'||^2]$$

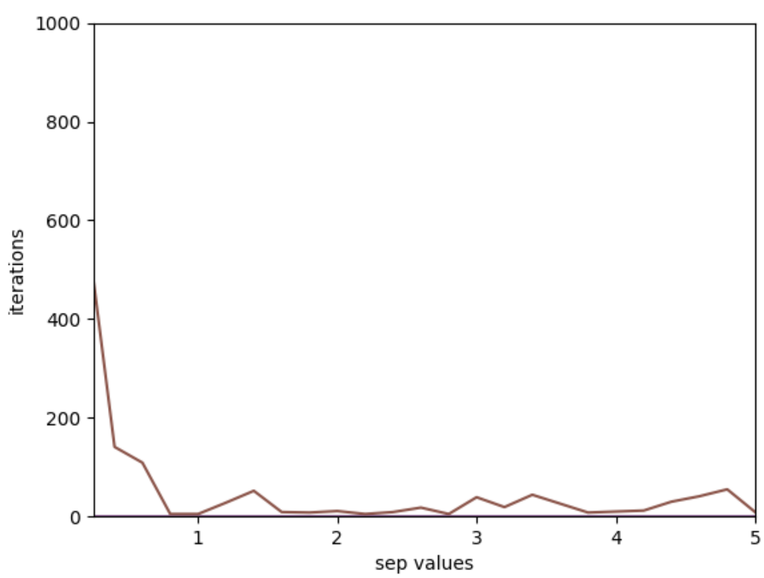
2. Problem 3.1



(a)

(b) Both PLA and linear regression found ways to separate this data, however, one could say that the linear regression algorithm found a better way to separate the data as the PLA appears to be closer to the top part of the semicircle, barely missing on misclassifying one of the +1 points. With this, one can predict that linear regression will have a lower E_{out} than the PLA, however, this isn't guaranteed.

3. Problem 3.2



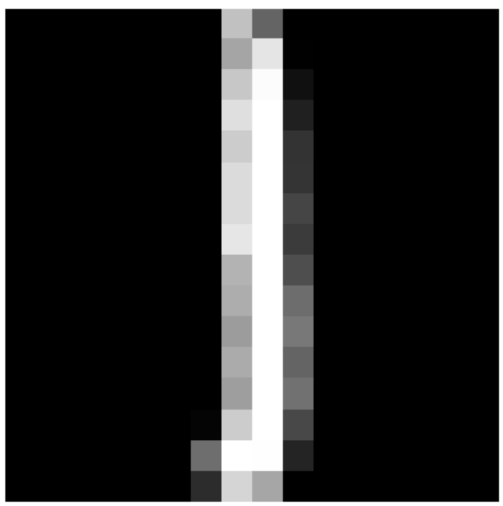
When *sep* is small, we require a large number of iterations to find the line of best fit. When the separation between the data sets is larger, iterations decrease significantly. Intuitively, we can give this reasoning due to the number of possible hypothesis when you have a larger gap between the two datasets. When *sep* is high, we can imagine a large gap between the two datasets, and any line inside of that gap can fit. However, when *sep* is small, we need the perfect line to fit in between the dataset, which limits how many lines are possible, thus causing PLA to run more iterations.

Mathematically, we proved that

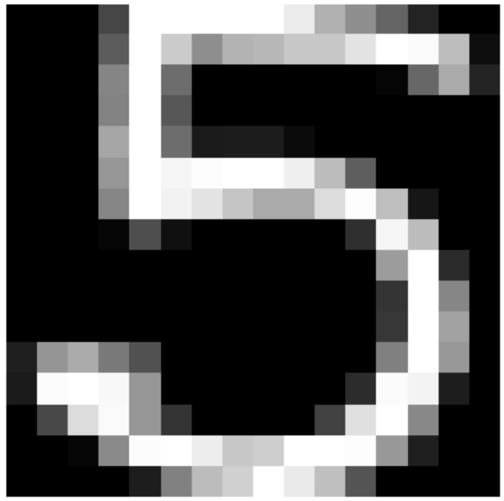
$$t \leq \frac{R^2||w^*||^2}{p^2}$$

Our sep values increase with our p, so as *sep* grows larger, p grows larger, creating a smaller bound for our iterations t.

4. Problem 3.8



(a)



(b)