Question 1 Optimization.

(a) Stochastic Gradient Descent (SGD).

1.1.1 Minimum Norm Solution. Show that SGD solution is identical to the minimum norm solution w^* obtained by gradient descent, i.e., $\hat{w} = w^*$.

The unique minimum norm solution w^*, obtained by gradient descent, is

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

We just need to show that the SGD converged solution $\hat{w}$ is identical to w^*, i.e.,

$$\hat{w} = X^T(XX^T)^{-1}t$$

As we can see,

$$\begin{aligned}
\mathcal{L} &= \frac{1}{n}||X\hat{w} - t||_2^2 \\
\mathcal{L}_i(x_i, w) &= ||\hat{w}^T x_i - t_i||^2 \\
\frac{\partial \mathcal{L}_i}{\partial \hat{w}} &= 2(\hat{w}^T x_i - t_i)x_i \\
\hat{w}_{t+1} &\leftarrow \hat{w}_t - 2\eta({\hat{w}_t}^T x_i - t_i)x_i
\end{aligned}$$

Since x_i is $d \times 1$, x_i can be written as $x_i = X^T 1$, where X^T is $d \times n$ and 1 is $n \times 1$.
Given i, 1^T will become [0 ... 1 ... 0], where the i_{th} entry is 1.
Thus, we have

$$\hat{w}_{t+1} \leftarrow \hat{w}_t - 2\eta X^T 1({\hat{w}_t}^T x_i - t_i)$$

Then, we can use the same proof method applied on last homework to show that if $w_0 = 0$, we have

$$\hat{w} \propto X^T d$$

where d is $n \times 1$ vector.
To show SGD from zero initialization finds a minimum norm unique minimizer:

$$\begin{aligned}
X\hat{w} &= t \\
XX^T d &= t \\
d &= (XX^T)^{-1}t \\
\hat{w} &= X^T(XX^T)^{-1}t
\end{aligned}$$

(b)

Question 2 .

(a)

(b)

Question 2 continued on next page. . .

Question 3 .

(a)

(b)