Question 1

First prove that
If $x > \theta z$, then optimal w > 0. If $x < \theta z$, then optimal w < 0. [Optimal w refers to the value such that C is minimized]

When $x > \theta z$, the minimal value C can get is greater than 0 if w < 0.
$f(x) = \frac{1}{1+\exp(-x)}$ is monolithic increasing. When w < 0, $wx + b < w\theta z + b$, $\frac{1}{1+\exp(-wx-b)} < \frac{1}{1+\exp(-w\theta z-b)}$. C is always greater than 0. When w > 0, $wx + b > w\theta z + b$, $\frac{1}{1+\exp(-wx-b)} > \frac{1}{1+\exp(-w\theta z-b)}$. C is less than 0. Therefore, if $x > \theta z$, then optimal w is greater than 0. The second half can be proved similarly.

Scenario 1: x > $\theta_0 z$

After step 2: w will be assigned a value w_1 such that $\frac{dC}{dw} = 0$ or clipped 1 if $w_1$is outside the range [-1, 1]. (w is greater than 0)

Scenario 1.1 z > 0
$\frac{dC}{d\theta} > 0$ therefore, after step 3 θ will increase until $\theta z = \beta$.
After running step 2 again, w will be less than 0 since $\theta z = \beta > x$.
$\frac{dC}{d\theta} < 0$, θ will decrease until $\theta z = \alpha$.
Then w will be assigned a positive value and θ will increase until $\theta z = \beta$ again. The program will never converge.

Scenario 1.1 z < 0
$\frac{dC}{d\theta} < 0$ therefore, after step 3, θ will decrease. $\alpha \leq \theta z \leq \beta$, thus $\frac{\beta}{z} \leq \theta \leq \frac{\alpha}{z}$. θ will decrease until $\theta = \frac{\beta}{z}$, which is equivalent to $\theta z = \beta$. This is the same as Scenario 1.1.

Scenario 2: x < $\theta_0 z$
This is the same as scenario 1 and can be proved in the same manner.

Therefore, the algorithm does not converge.

Question 2

In order to approximate the Lipschitz function, the weights of the neural network are constrained to some compact space. That is w is clipped after each iteration. This means neural network cannot approximate all the Lipschitz function, and the w obtained from line 3 to 8 may not be correct. The network may not be able to approximate the function that produces max $W(P_r, P_\theta)$.

Training time is long as well.