Q 2

a) For the bias term :-
we can find the gradient of the log loss as $y = 1$ as all the training examples are +ve.

$\therefore$ log loss $= -[(y \log \hat{y}) - (1-y) \log(1-\hat{y}))]$

$\therefore \nabla_b f_{loss} = \nabla b [-(1) \log \hat{y} - 0 \cdot \log(1-\hat{y}))$

$\therefore \nabla_b f_{loss} = \nabla b [-\log \hat{y}]$

$= -\nabla_b \log \sigma(x^T w + b)$

$= -\dfrac{\sigma(x^T w + b)(1 - \sigma(x^T w + b))}{\sigma(x^T w + b)}$

$\therefore \nabla_b f_{loss} = ((\sigma(x^T w + b) - 1)$

1) Now, if we assume $w = 0$

$\nabla_b f_{loss} = \sigma(b) - 1$

The bias update is

$b_{new} = b_{old} - [\text{learning rate} \times (\sigma(b) - 1)]$

$b_{new} = b_{old} - \alpha(\sigma(b) - 1)$

As $0 < \sigma(b) < 1 \quad \therefore \sigma(b) - 1 < 0$; let $\alpha(b) - 1 = a$

$b_{new} = b_{old} + \alpha \times a$

If we assume b convergence as 10000 Using the above equation we won't be able to provide an upper bound for iterations.

2) If $w \neq 0$

$\nabla_b f_{loss} = \sigma(x^T w + b) - 1$

The bias update is:

$b_{new} = b_{old} - ([\sigma(x^T w + b) - 1] \times \alpha)$

Now here we can see that the convergence of bias will depend on the values of the training example as well.

$\therefore$ The convergence of the bias will depend and cannot be guaranted.