1.

(i)

$$\hat{\theta} = \arg\min - \sum_{n=1}^{N} \{ y_n \log h_\theta(x_n) + (1-y_n)\log(1-h_\theta(x_n)) \}.$$

$$J(\theta) = \sum_{n=1}^{N} \{ y_n \log h_\theta(x_n) + (1-y_n)\log[1-h_\theta(x_n)] \}$$

$$= \left(\sum_{n=1}^{N} y_n x_n\right)^T \theta - \sum_{n=1}^{N} \log(1+e^{\theta^T x_n})$$

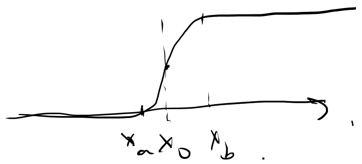
$$\nabla_\theta J(\theta) = \left(\sum_{n=1}^{N} y_n x_n\right) - \sum_{n=1}^{N} \frac{x_n \cdot e^{\theta^T x_n}}{1+e^{\theta^T x_n}}$$

$$= \sum_{n=1}^{N} x_n \cdot \left(y_n - \frac{e^{\theta^T x_n}}{1+e^{\theta^T x_n}}\right)$$

$$= \sum_{n=1}^{N} x_n [y_n - h_\theta(x)] = 0.$$

If the two classes are linearly seperatble, the optimal θ should switch from $y_n = 0$ to $y_n = 1$ instantly to satisfy the gradient being 0.

Assume x_0 is the switching point,



$$h_\theta(x_a) = \frac{1}{1+e^{-\theta^T(x_a - x_0)}}$$

$$h_\theta(x_b) = \frac{1}{1+e^{-\theta^T(x_b - x_0)}} \qquad \lim_{\substack{x_b \to x_0 \\ x_a \to x_0}} \frac{h_\theta(x_b) - h_\theta(x_a)}{x_b - x_a} \to \infty.$$

As the steepness of the curve goes to infinity, $\|\theta\| \to \infty$.

Hence $\|W\| \to \infty$ and $\|W_0\| \to \infty$.

ii) As shown in i).

$$\nabla_\theta J(\theta) = \sum_{n=1}^{N} x_n [y_n - h_\theta(x_n)] < 0,$$

Since the steepness of $h_\theta(x_n)$ only tends to ∞, but never reaches it, the gradient descent would never converges unless it reaches ∞, which is impossible.

iii) If we set a maximum steepness on the logistic curve, the gradient descent will stop once it it reaches there.

Another way to counter this problem is to add regularization term in the loss function.

iv) We haven't met non-convergence issue with other linear classifiers, because if the data are linearly seperable, then we can invert the matrix and perform good'ol linear regression.

2. $J(\theta) = -\frac{1}{N} \sum_{n=1}^{N} \left\{ y_n \log h_\theta(x_n) + (1-y_n) \log(1-h_\theta(x_n)) \right\}$

where $h_\theta(x) = \dfrac{1}{1+\exp(-\theta^T x)}$. Show $J(\theta)$ convex by

Its Hessian, & show it's PSD.

$$J(\theta) = -\frac{1}{N} \sum_{n=1}^{N} \Big\{ \underbrace{y_n \log\left(\frac{h_\theta(x_n)}{1-h_\theta(x_n)}\right)}_{①} + \underbrace{\log(1-h_\theta(x_n))}_{②} \Big\}.$$

①: $y_n \log\left(\dfrac{h_\theta(x_n)}{1-h_\theta(x_n)}\right)$

$$= y_n \log\left(\frac{\frac{1}{1+e^{-\theta^T x_n}}}{\frac{e^{-\theta^T x}}{1+e^{-\theta^T x}}}\right) = y_n \log(e^{\theta^T x_n}) = \underbrace{y_n \theta^T x_n}_{\text{linear}},$$

thus the first term is convex.

② $\nabla_\theta[-\log(1-h_\theta(x))] = -\nabla_\theta\left[\log\left(1-\frac{1}{1+e^{-\theta^T x}}\right)\right]$

$$= -\nabla_\theta\left[\log \frac{e^{-\theta^T x}}{1-e^{-\theta^T x}}\right] = h_\theta(x)\, x,$$

$$\nabla_\theta^2[-\log(1-h_\theta(x))] = h_\theta(x)[1-h_\theta(x)]\, x x^T$$

For any $v \in \mathbb{R}^d$,

$$v^T \nabla_\theta^2 [-\log(1-h_\theta(x))] v = v^T [h_\theta(x)(1-h_\theta(x)) x x^T] v = h_\theta(x)(1-h_\theta(x)) \|v^T x\|^2 \geq 0$$

thus the Hessian is PSD., so $-\log(1-h_\theta(x))$ is convex in θ.