1. (a) (i) $\delta_x^{(1)} = \nabla_{\theta_x} l_x(x; W^{(2)}, W_x^{(1)})$

$$= \nabla_{\theta_x}\left(\frac{1}{2}\|xW^{(2)}W_x^{(1)} - x\|_2^2\right)$$

=> In terms of θ_x => $\nabla_{\theta_x}\left(\frac{1}{2}\|\theta_x - x\|_2^2\right)$, $\theta_x = hW_x^{(1)}$, $h = xW^{(2)}$

=> $\frac{1}{2}\nabla_{\theta_x}[(\theta_x - x)^T(\theta_x - x)] = \frac{1}{2}\nabla_{\theta_x}[(\theta_x^T - x^T)(\theta_x - x)]$

$= \frac{1}{2}\nabla_{\theta_x}[\theta_x^T\theta_x - \theta_x^T x - x^T\theta_x + \cancel{x^T x}^{\,0}$

$= \frac{1}{2}[2\theta_x - \theta_x - x^T] \boxed{= \frac{1}{2}(\theta_x - x^T) \in \mathbb{R}^{d\times 1}}$

(ii) $\nabla_{W_y^{(1)}} l(x, y; W^{(2)}, W_x^{(1)}, W_y^{(1)}) = \nabla_{W_y^{(1)}}(l_y(\text{softmax}(xW^{(2)}W_y^{(1)}), y)) + 0$

↳ since l_x does not have $W_y^{(1)}$ $\nabla l_x = 0$

$\nabla_y^{(1)} l_y = \nabla_{W_y^{(1)}} \theta_y \; \nabla_{\theta_y} l_y = h\delta_y^{(1)} = \boxed{xW^{(2)}(\text{softmax}(xW^{(2)}W_y^{(1)}) - y) \in \mathbb{R}^{p\times m}}$

(iii) $\nabla_h \; l_y(\text{softmax}(hW_y^{(1)}), y) + \beta l_x(hW_x^{(1)}, x)$

$= (\nabla_h \theta_y)(\text{softmax}(hW_y^{(1)}) - y) + \beta(\nabla_h \theta_x)(1/2\,(\theta_x - x^T))$

$$\boxed{= W_y^{(1)}(\text{softmax}(hW_y^{(1)}) - y) + \beta/2\; W_x^{(1)}(hW_x^{(1)} - x^T)) \in \mathbb{R}^{p\times 1}}$$

(iv)

$\nabla_{W^{(2)}} l = \nabla_{W^{(2)}} h \cdot \nabla_h(l)$

$$\boxed{= x(W_y^{(1)}(\text{softmax}(hW_y^{(1)}) - y) + \beta/2\; W_x^{(1)}(hW_x^{(1)} - x^T)) \in \mathbb{R}^{p\times m}}$$

b) When $\beta = 0$, $\hat{y} = \text{softmax}(xW^{(2)}W_y^{(1)})$ is equivalent to $\hat{y} = \sigma(xW)$

which ends up being a linear classifier.

c)

d)