1(a) $\tilde{L}(\theta) = \frac{1}{N}\|y^{(i)} - (x^{(i)} + \delta^{(i)})^T\theta\|_2^2$

$= \frac{1}{N}(Y-(X+\delta)^T\theta)^T(Y-(X+\delta)^T\theta)$.

$= \frac{1}{N}(Y^T - \theta^T(X+\delta))(Y-(X+\delta)^T\theta)$

$= \frac{1}{N}[Y^TY - Y^T(X+\delta)^T\theta - \theta^T(X+\delta)Y +$

$\theta^T(X+\delta)(X+\delta)^T\theta]$

$= \frac{1}{N}[Y^TY - 2Y^T(X+\delta)^T\theta + \theta^T(X+\delta)(X+\delta)^T\theta]$

$= \frac{1}{N}[Y^TY - 2Y^T(X^T+\delta^T)\theta + \theta^T(XX^T+X\delta^T+\delta\delta^T)\theta]$

since $L(\theta) = \frac{1}{N}(Y^TY - 2Y^TX^T\theta + \theta^TXX^T\theta)$

so $\tilde{L}(\theta) = L(\theta) - \frac{1}{N}[2Y^T\delta^T\theta - \theta^T(2X\delta^T + \delta\delta^T)\theta]$

$E[\frac{1}{N}(2Y^T\delta^T\theta - \theta^T(2X\delta^T+\delta\delta^T)\theta]$

$= -E[\frac{1}{N}(\theta^T\delta\delta^T\theta)] = -\frac{1}{N}\theta^T\sigma^2I\theta = -\frac{1}{N}\theta\sigma^2\theta$

so $E_{\delta\sim N}[\tilde{L}(\theta)] = L(\theta) + \frac{1}{N}\theta^T\sigma^2\theta$

$= L(\theta) + \frac{1}{N}\sigma^2\|\theta\|_2^2$

1(b) Under expectation, the loss function will be increased by $\frac{1}{N}\sigma^2\|\theta\|_2^2$ because of the noise.

1(c) $\sigma \to 0$, $E[\tilde{L}(\theta)] = L(\theta)$, the model will overfit the data.

1(d) $\sigma \to \infty$, $\tilde{L}(\theta) \approx \sigma^2\|\theta\|_2^2$, the model will underfit the data.

3. Assuming the samples $(x^{(1)}, y^{(1)}), \cdots (x^{(m)}, y^{(m)})$ are iid,

$$p(x^{(1)}, \cdots x^{(m)}, y^{(1)}, \cdots y^{(m)} | \theta) = \prod_{i=1}^{m} p(x^{(i)}, y^{(i)} | \theta)$$

$$= \prod_{i=1}^{m} (p(x^{(i)}|\theta)) p(y^{(i)}|x^{(i)}, \theta)$$

$$\operatorname{argmax}_{\theta} \prod_{i=1}^{m} p(x^{(i)}|\theta) p(y^{(i)}|x^{(i)}, \theta)$$

$$= \operatorname{argmax}_{\theta} \prod_{i=1}^{m} p(y^{(i)}|x^{(i)}, \theta)$$

$$= \operatorname{argmax}_{\theta} \sum_{i=1}^{m} \log \operatorname{softmax}_{y^{(i)}}(x^{(i)})$$

$$= \operatorname{argmax}_{\theta} \sum_{i=1}^{m} \log \left[\frac{e^{a_i(x^{(i)})}}{\sum_{j=1}^{c} e^{a_j(x^{(i)})}} \right]$$

$$= \operatorname{argmax}_{\theta} \frac{1}{m} \sum_{i=1}^{m} \left[a_{y^{(i)}}(x^{(i)}) - \log \sum_{j=1}^{c} e^{a_j(x^{(i)})} \right]$$

$$= \operatorname{argmin}_{\theta} \frac{1}{m} \sum_{i=1}^{m} \left[\log \sum_{j=1}^{c} e^{a_j(x^{(i)})} - a_{y^{(i)}}(x^{(i)}) \right]$$

So $L(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left[\log \sum_{j=1}^{c} e^{a_j(x^{(i)})} - a_{y^{(i)}}(x^{(i)}) \right]$

Expand $a_i(x) = w_i^T x + b_i$ as below:

$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left[\log\sum_{j=1}^{c} e^{w_j^T x^{(i)} + b_j} - \left(w_{y^{(i)}}^T x^{(i)} + b_{y^{(i)}}\right)\right]$$

Since $L(\theta)$ is linear, we can calculate two terms seperately,

For the first term, $\nabla_{w_i} L_1 = \text{softmax}_j(x^{(i)}) \cdot x^{(i)}$

For the second term, $\nabla_{w_i} L_2 = \begin{cases} -x^{(i)} & \text{if } y^{(i)} = i \\ 0 & \text{o.w.} \end{cases}$

so $\nabla_{w_i} L(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(\text{softmax}(x^{(i)}) \cdot x^{(i)} - x^{(i)} \cdot 1_{\{y^{(i)}=i\}}\right)$

Similarly, $\nabla_{b_i} L_1 = \text{softmax}_j(x^{(i)})$

$$\nabla_{b_2} L_2 = \begin{cases} -1 & \text{if } y^{(i)} = i \\ 0 & \text{o.w.} \end{cases}$$

so $\nabla_{b_i} L(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(\text{softmax}(x^{(i)}) - 1_{\{y^{(i)}=i\}}\right)$

Since we are doing argmin to replace argmax,

$$\therefore \nabla_{w_i} L(\theta) = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}\left(1_{\{y^{(i)}=i\}} - \text{softmax}(x^{(i)})\right)$$

$$\nabla_{b_i} L(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(1_{\{y^{(i)}=i\}} - \text{softmax}(x^{(i)})\right).$$

4. $\text{hinge}_{y^{(i)}}(x^{(i)}) = \begin{cases} 0 & y^{(i)}(w^T x^{(i)} + b) > 1 \\ 1 - y^{(i)}(w^T x^{(i)} + b) & \text{o.w.} \end{cases}$

If $y^{(i)}(w^T x^{(i)} + b) > 1$, $\nabla_w \text{hinge}_{y^{(i)}}(x^{(i)}) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^d$

else, then $\nabla_w \text{hinge}_{y^{(i)}}(x^{(i)}) = -y^{(i)} x^{(i)}$

$$\nabla_w \text{hinge}_{y^{(i)}}(x^{(i)}) = \mathbb{I}_{\{y^{(i)}(w^T x^{(i)} + b) < 1\}} \odot (-y^{(i)} x^{(i)})$$

so, $\nabla_w L = \frac{1}{k} \sum_{i=1}^{k} \mathbb{I}_{\{y^{(i)}(w^T x^{(i)} + b) < 1\}} \odot (-y^{(i)} x^{(i)})$

If $y^{(i)}(w^T x^{(i)} + b) > 1$, $\nabla_b \text{hinge}_{y^{(i)}}(x^{(i)}) = 0$

else, $\nabla_b \text{hinge}_{y^{(i)}}(x^{(i)}) = -y^{(i)}$

so, $\nabla_b L = \frac{1}{k} \sum_{i=1}^{k} \mathbb{I}_{\{y^{(i)}(w^T x^{(i)} + b) < 1\}} \odot (-y^{(i)})$