BME 147 HW 2

Andrea Borotta

1) a. $\tilde{\mathcal{L}}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - (x^{(i)} + \delta^{(i)})^T\theta\right)^2$

$= \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - x^{(i)T}\theta + \delta^{(i)T}\theta\right)^2$

$= \frac{1}{N}\sum_{i=1}^{N}\left[\left(y^{(i)} - x^{(i)T}\theta\right) + \delta^{(i)T}\theta\right]^2$

$= \frac{1}{N}\sum_{i=1}^{N}\left[\left(y^{(i)} - x^{(i)T}\theta\right)^2 + 2\delta^{(i)T}\theta\left(y^{(i)} - x^{(i)T}\theta\right) + \theta^T\delta^{(i)}\delta^{(i)T}\theta\right]$

Then, by the linearity of the expectation function,

$E[\tilde{\mathcal{L}}(\theta)] = \frac{1}{N}\sum_{i=1}^{N}\left(E\left[\left(y^{(i)} - x^{(i)T}\theta\right)^2\right] + E\left[2\delta^{(i)T}\theta\left(y^{(i)} - x^{(i)T}\theta\right)\right] + E\left[\theta^T\delta^{(i)}\delta^{(i)T}\theta\right]\right)$

Since θ, $y^{(i)}$, $x^{(i)}$ are independent of $\delta^{(i)}$,

$= \frac{1}{N}\sum_{i=1}^{N}\left(\left(y^{(i)} - x^{(i)T}\theta\right)^2 + 2E[\delta^{(i)T}]\theta\left(y^{(i)} - x^{(i)T}\theta\right) + \theta^T E[\delta^{(i)}\delta^{(i)T}]\theta\right)$

For a zero centered Gaussian distribution, $E[\delta^{(i)}] = 0$, and we know $E[\delta^{(i)}\delta^{(i)T}] = \sigma^2 I$, so

$= \mathcal{L}(\theta) + \frac{1}{N}\sum_{i=1}^{N}\left(2 \cdot 0 \cdot \theta\left(y^{(i)} - x^{(i)T}\theta\right) + \theta^T(\sigma^2 I)\theta\right)$

$= \mathcal{L}(\theta) + \sigma^2\theta^T\theta$

$= \mathcal{L}(\theta) + \sigma^2\|\theta\|^2$

b. The addition of noise would regularize by attempting to also minimize the magnitude of θ.

c. If $\sigma \to 0$, $\tilde{\mathcal{L}}(\theta) = \mathcal{L}(\theta) + \sigma^2\|\theta\|^2 \to \mathcal{L}(\theta)$, so there would be no regularization effect

d. If $\sigma \to \infty$, $\tilde{\mathcal{L}}(\theta) = \mathcal{L}(\theta) + \sigma^2\|\theta\|^2 \to \infty$, so it would be impossible to minimize the cost function

3) We have that

$$\text{softmax}_i(x) = \frac{e^{\tilde{w}_i^T \tilde{x}}}{\sum_{j=1}^{c} e^{\tilde{w}_j^T \tilde{x}}}$$

for $\tilde{x} = \begin{bmatrix} \hat{x} \\ 1 \end{bmatrix}$, $\tilde{w} = \begin{bmatrix} w \\ b \end{bmatrix}$.

Then, let $\mathcal{L}_i$ be the gradient of the softmax for the i-th observation in the data set. Since $\mathcal{L}$ is the negative log likelihood, it is additive. Then

$$\mathcal{L}_i(\tilde{x}^{(i)}) = -\log \text{softmax}_{y^{(i)}}(\tilde{x}^{(i)})$$

$$= \log \sum_{j=1}^{c} e^{\tilde{w}_j^T \tilde{x}^{(i)}} - \log e^{\tilde{w}_{y^{(i)}}^T \tilde{x}^{(i)}}$$

$$= \log \sum_{j=1}^{c} e^{\tilde{w}_j^T \tilde{x}^{(i)}} - \tilde{w}_{y^{(i)}}^T \tilde{x}^{(i)}$$

Then

$$\frac{\partial \mathcal{L}_i}{\partial w_k} = \frac{1}{\sum_{j=1}^{c} e^{\tilde{w}_j^T \tilde{x}}} e^{\tilde{w}_k^T \tilde{x}^{(i)}} \tilde{x}^{(i)} - \frac{\partial}{\partial w_k}\left(\tilde{w}_{y^{(i)}}^T \tilde{x}^{(i)}\right)$$

$$= \text{softmax}_k(\tilde{x}^{(i)})\, \tilde{x}^{(i)} - \mathbb{I}_{\{y^{(i)}=k\}} \tilde{x}^{(i)}$$

$$= \left(\text{softmax}_k(\tilde{x}^{(i)}) - \mathbb{I}_{\{y^{(i)}=k\}}\right) \tilde{x}^{(i)}$$

And

$$\frac{\partial \mathcal{L}}{\partial w_k} = \sum_{i=1}^{n} \frac{\partial \mathcal{L}_i}{\partial w_k}$$

4) $$\mathcal{L}(w, b) = \frac{1}{k} \sum_{i=1}^{k} \text{hinge}_{y^{(i)}}(x^{(i)})$$

$$= \frac{1}{k} \sum_{i=1}^{k} \max\left(0,\ 1 - y^{(i)}(w^T x^{(i)} + b)\right)$$

$$\frac{\partial \mathcal{L}}{\partial w} = \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}_{\{y^{(i)}(w^T x^{(i)} + b) < 1\}} \left(\frac{\partial}{\partial w} \left(1 - y^{(i)}(w^T x^{(i)} + b)\right)\right)$$

$$= \frac{1}{k} \sum \mathbb{1}_{\{y^{(i)}(w^T x^{(i)} + b) < 1\}} \left(-y^{(i)} x^{(i)}\right)$$

$$= \frac{1}{k} X^T \begin{bmatrix} \mathbb{1}_{\{y^{(1)}(w^T x^{(1)} + b)} & \cdots & 0 \\ 0 & & \mathbb{1}_{\{y^{(k)}(w^T x^{(k)} + b)} \end{bmatrix} Y$$

Where $X = \begin{bmatrix} \text{—} & x^{(1)} & \text{—} \\ & \vdots & \\ \text{—} & x^{(k)} & \text{—} \end{bmatrix}$ and $Y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(k)} \end{bmatrix}$

Also,

$$\frac{\partial \mathcal{L}}{\partial b} = \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}_{\{y^{(i)}(w^T x^{(i)} + b) < 1\}} \left(\frac{\partial}{\partial b} \left(1 - y^{(i)}(w^T x^{(i)} + b)\right)\right)$$

$$= \frac{1}{k} \sum_{i=1}^{k} \mathbb{1}_{\{y^{(i)}(w^T x^{(i)} + b) < 1\}} \left(-y^{(i)}\right)$$

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