1. (Math) Recall that when ℓ_1 regularization is used, and assuming that the Hessian matrix is diagonal with positive entries, the objective function can be approximated by

$$\hat{L}_R(\theta) := \sum_{i=1}^{d}\left[\frac{1}{2}H_{ii}(\theta_i - \theta_i^*)^2 + \alpha|\theta_i|\right]$$

Solve this in the close form expression: show that the optimal solution θ_R^* for the objective $\hat{L}_R(\theta)$ is that as shown on slide 26 of "Deep learning lecture 3: Regularization I".

Hint: note that $\alpha/H_{ii} > 0$.

Proof: $\hat{L}_R(\theta_R^*) = \sum_{i=1}^{d}\left[\frac{1}{2}H_{ii}(\theta_{Ri} - \theta_{Ri}^*)^2 + \alpha|\theta_{Ri}|\right] = \hat{L}(\theta_R^*) + \alpha\|\theta_R^*\|_1$

$$(\theta_R^*)_i = \frac{dL_R}{d\theta} = \frac{d}{d\theta}\left[\frac{1}{2}H_{ii}(\theta_i - \theta_i^*)^2\right] + \frac{d}{d\theta}\left[\alpha|\theta_i|\right]$$

$$= \theta_i^* - \frac{\alpha}{H_{ii}}$$

Since θ_R^* is optimal, $\theta_R^* = \theta^* - \frac{\alpha}{H_{ii}} = 0$, $\theta_R^* = \theta^* = \frac{\alpha}{H_{ii}}$

$\hat{L}(\theta) = L(\theta^*) + \frac{1}{2}(\theta-\theta^*)^T H(\theta-\theta^*)$

$\because$ H matrix is positive and diagonal

$$\begin{bmatrix} +\frac{d^2}{dx^2} & 0 \\ 0 & +\frac{d^2}{dy^2} \end{bmatrix}$$

$\therefore$ $(\theta-\theta^*)^T H(\theta-\theta^*) = H(\theta-\theta^*)^2$

$\hat{L}(\theta) = L(\theta^*) + \frac{1}{2}H(\theta-\theta^*)^2$

By drop the constant $L(\theta^*)$,

$\hat{L}(\theta) = \underline{\frac{1}{2}H(\theta-\theta^*)^2}$

$\sum\alpha|\theta_i| = \alpha\sum|\theta_i| = \underline{\alpha\|\theta\|_1}$

$\hat{L}_R(\theta_R^*) = \sum_{i=1}^{d}\left[\frac{1}{2}H_{ii}(\theta_{Ri} - \theta_{Ri}^*)^2 + \alpha|\theta_{Ri}|\right]$

$= \hat{L}(\theta_R^*) + \alpha\|\theta_R^*\|_1$

= Slide 26

2. (Math) Consider a three layer network:

$$h^1 = \sigma(W^1 x), \qquad h^2 = \sigma(W^2 h^1), \qquad f(x) = \langle w^3, h^2 \rangle.$$

See Figure 1 for an illustration. Compute $\frac{\partial f}{\partial W^1_{i,j}}$

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

$$\sigma'(x) = \frac{d}{dx}\sigma(x) = \sigma(x)(1-\sigma(x))$$

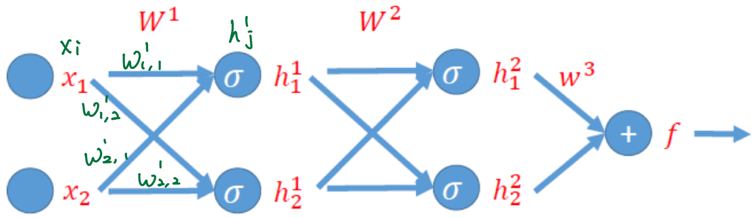

Figure 1: An illustration of the three layer network

$$\frac{\partial f}{\partial w^1_{ij}} = \frac{\partial f}{\partial h^2_1} \cdot \frac{\partial h^2_1}{\partial net^2_1} \cdot \frac{\partial net^2_1}{\partial h^1_j} \cdot \frac{\partial h^1_j}{\partial net_j} \cdot \frac{\partial net^1_j}{\partial w^1_{ij}}$$

$$+ \frac{\partial f}{\partial h^2_2} \cdot \frac{\partial h^2_2}{\partial net^2_2} \cdot \frac{\partial net^2_2}{\partial h^1_j} \cdot \frac{\partial h^1_j}{\partial net_j} \cdot \frac{\partial net^1_j}{\partial w^1_{ij}}$$

$$= w^3 \times \sigma'(net^2_1) \times w^2 \times \sigma'(net^1_j) \times X_i$$

$$+ w^3 \times \sigma'(net^2_2) \times w^2 \times \sigma'(net^1_j) \times X_i$$

$$= \left[\sigma'(net^2_1) + \sigma'(net^2_2)\right] \times w^3 w^2 \sigma'(net^1_j) X_i$$

$$\boxed{= \left[h^2_1(1-h^2_1) + h^2_2(1-h^2_2)\right] \cdot w^3 w^2 \cdot h^1_j(1-h^1_j) \cdot X_i}$$

$$\sigma'(net^2_1) = \sigma(net^2_1)(1-\sigma(net^2_1)) = h^2_1(1-h^2_1)$$

$$\sigma'(net^2_2) = h^2_2(1-h^2_2)$$

$$\sigma'(net^1_j) = h^1_j(1-h^1_j)$$