1. Let $z = b + W_1 X_1 + W_2 X_2$ and $L = \frac{1}{1}\sum_{i=1}^{1}(y - h(x_1, x_2))^2 = \left(y' - h(x_1', x_2')\right)^2$

$$\frac{\partial L}{\partial b} = -2\left(y - \sigma(z)\right)\sigma(z)\left(1 - \sigma(z)\right)$$

$$= \left(y - \sigma(b + W_1 X_1 + W_2 X_2)\right)^2$$

$$\frac{\partial L}{\partial W_1} = -2\left(y - \sigma(z)\right)\sigma(z)\left(1 - \sigma(z)\right)X_1$$

$$\frac{\partial L}{\partial W_2} = -2\left(y - \sigma(z)\right)\sigma(z)\left(1 - \sigma(z)\right)X_2 .$$

$(X_1, X_2, y) = (1, 2, 3)$, $(b^\circ, W_1^\circ, W_2^\circ) = (4, 5, 6)$.

$$\Rightarrow z = 4 + 5 \cdot 1 + 6 \cdot 2 = 21 .$$

By SGD: $\theta' = \theta^\circ - \alpha \nabla L(\theta^\circ)$

$$= \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} - \alpha \begin{bmatrix} -2(3 - \sigma(21))\sigma(21)(1 - \sigma(21)) \\ -2(3 - \sigma(21))\sigma(21)(1 - \sigma(21)) \cdot 1 \\ -2(3 - \sigma(21))\sigma(21)(1 - \sigma(21)) \cdot 2 \end{bmatrix}$$

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

$$\frac{d}{dx}\sigma(x) = \frac{-(-e^{-x})}{(1 + e^{-x})^2} = \frac{e^{-x}}{(1 + e^{-x})^2} = \frac{1}{1 + e^{x}} \cdot \frac{e^{-x}}{1 + e^{-x}}$$

$$= \sigma(x) \cdot (1 - \sigma(x))$$

$$\frac{d^2}{dx^2}\sigma(x) = \frac{d}{dx}\left(\sigma(x)(1 - \sigma(x))\right) = \frac{d}{dx}[\sigma(x)] \cdot (1 - \sigma(x)) + \sigma(x) \cdot \frac{d}{dx}(1 - \sigma(x))$$

$$= \sigma(x)(1 - \sigma(x)) \cdot (1 - \sigma(x)) + \sigma(x) \cdot \left(- \sigma(x)(1 - \sigma(x))\right)$$

$$= \sigma(x)(1 - \sigma(x))^2 - \sigma(x)^2(1 - \sigma(x))$$

$$= \sigma(x)(1 - \sigma(x))(1 - 2\sigma(x))$$

$$\frac{d^3}{dx^3}\sigma(x) = \frac{d}{dx}\left(\sigma(x)(1 - \sigma(x))(1 - 2\sigma(x))\right) = \sigma(x)(1 - \sigma(x))^2(1 - 2\sigma(x)) + \sigma(x)\left(- \sigma(x)(1 - \sigma(x))\right)(1 - 2\sigma(x))$$

$$+ \sigma(x)(1 - \sigma(x))\left(-2\sigma(x)(1 - \sigma(x))\right)$$

$$= \sigma(x)(1 - \sigma(x))^2(1 - 2\sigma(x)) - \sigma(x)^2(1 - \sigma(x))(1 - 2\sigma(x))$$

$$- 2\sigma(x)^2(1 - \sigma(x))^2$$

$$= \sigma(x)(1 - \sigma(x))\left[(1 - \sigma(x))(1 - 2\sigma(x)) - \sigma(x)(1 - 2\sigma(x)) - 2\sigma(x)(1 - \sigma(x))\right]$$

$$= \sigma(x)(1-\sigma(x))\left[1-\sigma(x)+2\sigma(x)^2 - \sigma(x)+2\sigma(x)^2 - 2\sigma(x)+2\sigma(x)^2\right]$$

$$= \sigma(x)(1-\sigma(x))\left(1-6\sigma(x)+6\sigma(x)^2\right) \quad *$$

(b) $\tanh(x) = \dfrac{e^x - e^{-x}}{e^x + e^{-x}} = \dfrac{1-e^{-2x}}{1+e^{-2x}} = \dfrac{2-(1+e^{-2x})}{1+e^{-2x}} = \dfrac{2}{1+e^{-2x}} - 1 = 2\sigma(2x) - 1 \quad *$

3. ① How can we effectively choose the learning rate α?

② Are there any methods for automatically adjusting the learning rate?

③ What can be done when gradient descent stuck in a local minimum?