Problem 1.

(a)
$$\sigma'(x) = \frac{d}{dx}\frac{1}{1+e^{-x}}$$
$$= \frac{d}{dx}(1+e^{-x})^{-1}$$
$$= -(1+e^{-x})^{-2}\cdot\frac{d}{dx}(1+e^{-x})$$
$$= -(1+e^{-x})^{-2}\cdot\left(0+\frac{d}{dx}[e^{-x}]\right)$$
$$= -(1+e^{-x})^{-2}\cdot(e^{-x}\cdot -1)$$
$$= (1+e^{-x})^{-2}\cdot e^{-x}$$
$$= \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}}\cdot\frac{e^{-x}}{1+e^{-x}}$$
$$= \sigma(x)\cdot\frac{e^{-x}+1-1}{1+e^{-x}}$$
$$= \sigma(x)\cdot\left(1-\frac{1}{1+e^{-x}}\right)$$
$$= \sigma(x)(1-\sigma(x))$$

(b)
$$(\tanh(x))' = \frac{(e^{x}-e^{-x})'(e^{x}+e^{-x})-(e^{x}-e^{-x})(e^{x}+e^{-x})'}{(e^{x}+e^{-x})^2}$$
$$= \frac{(e^{x}+e^{-x})(e^{x}+e^{-x})-(e^{x}-e^{-x})(e^{x}-e^{-x})}{(e^{x}+e^{-x})^2}$$
$$= \frac{(e^{x}+e^{-x})^2-(e^{x}-e^{-x})^2}{(e^{x}+e^{-x})^2}$$
$$= 1-\left(\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}\right)^2 = 1-\tanh^2(x)$$

Problem 2

(1) Assume $m=1$, skip-gram on average will get more training samples.

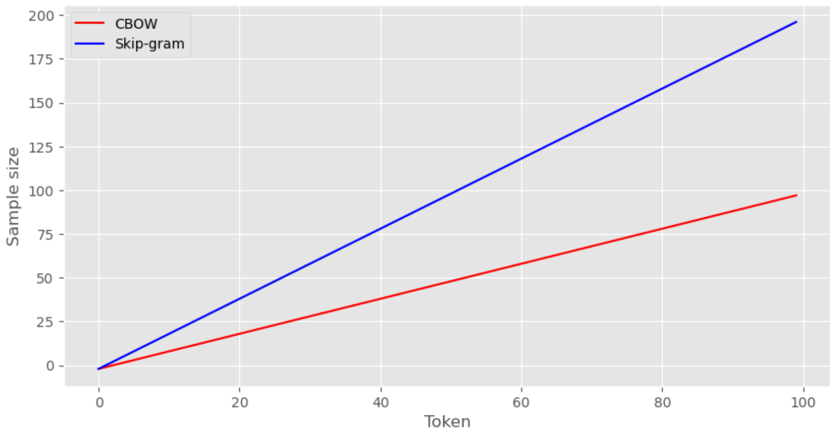
(2) Token = 7, CBOW = 5 Skip-Gram = 12

Token = 8, CBOW = 6 Skip-Gram = 14

Token = 11, CBOW = 9 Skip-Gram = 20

For CBOW method with sentence of n tokens, there are $(n-2)$ samples

For skip-gram method with sentence of n tokens, there are $(2n-2)$ samples



Problem 3

$$\frac{dl}{dw'} = -\frac{1}{a_2} \cdot \frac{da_2}{dw'}$$

$$= -\frac{1}{a_2} \cdot \left(\sigma(z_1) - \sigma(z_1)^2\right) \cdot \frac{dz_1}{dw_1}$$

$$= -\frac{1}{a_2} \cdot \left(\sigma(z_1) - \sigma(z_1)^2\right) \cdot W^2 \cdot \frac{da_1}{dw'}$$

$$= -\frac{1}{a_2} \cdot \left(\sigma(z_1) - \sigma(z_1)^2\right) \cdot W^2 \cdot \left(\sigma(z_0) - \sigma(z_0)^2\right) \cdot \frac{dz_0}{dw'}$$

$$= -\frac{1}{a_2} \cdot \left(\sigma(z_1) - \sigma(z_1)^2\right) \cdot W^2 \cdot \left(\sigma(z_0) - \sigma(z_0)^2\right) \cdot a_0$$

$$\frac{dl}{dw^2} = -\frac{1}{a_2}\cdot\frac{da_2}{dw^2}$$

$$= -\frac{1}{a_2}\cdot(\sigma(z_1) - \sigma(z_1)^2)\cdot\frac{dz_1}{dw^2}$$

$$= -\frac{1}{a_2}\cdot(\sigma(z_1) - \sigma(z_1)^2)\cdot a_1$$

$$\frac{dl}{db^1} = -\frac{1}{a_2}\cdot\frac{da_2}{db^1}$$

$$= -\frac{1}{a_2}\cdot(\sigma(z_1) - \sigma(z_1)^2)\cdot\frac{dz_1}{db^1}$$

$$= -\frac{1}{a_2}\cdot(\sigma(z_1) - \sigma(z_1)^2)\cdot w^2\cdot\frac{da_1}{db^1}$$

$$= -\frac{1}{a_2}\cdot(\sigma(z_1) - \sigma(z_1)^2)\cdot w^2\cdot(\sigma(z_0) - \sigma(z_0)^2)\cdot\frac{dz_0}{db^1}$$

$$= -\frac{1}{a_2}\cdot(\sigma(z_1) - \sigma(z_1)^2)\cdot w^2\cdot(\sigma(z_0) - \sigma(z_0)^2)$$

$$\frac{dl}{db^2} = -\frac{1}{a_2}\cdot\frac{da_2}{db^2}$$

$$= -\frac{1}{a_2}(\sigma(z_1) - \sigma(z_1)^2)\cdot\frac{dz_1}{db^2}$$

$$= -\frac{1}{a_2}(\sigma(z_1) - \sigma(z_1)^2)$$

$$\frac{da_1}{dz_0} = (1 - \sigma(z_0))\sigma(z_0)$$

If z_0 is very large, $\sigma(z_0)$ is close to 1, which means $\frac{da_1}{dz_0}$ will be close to 0. The grediant will be 0. The w_1 and w_2 can't learning because the grediant is close to 0.