Homework3

XiongJiao

1)

Lr: the learning rate

The loss:

$$L = \sum_{i=1}^{t} L^{(i)}$$

$$\mathrm{V} = V - lr \bullet \frac{\partial L}{\partial V}$$

$$\frac{\partial L^{(t)}}{\partial V} = \frac{\partial L^{(t)}}{\partial o^{(t)}} \bullet \frac{\partial o^{(t)}}{\partial V}$$

$$\frac{\partial L}{\partial V} = \sum_{i=1}^{t} \frac{\partial L^{(t)}}{\partial o^{(t)}} \bullet \frac{\partial o^{(t)}}{\partial V}$$

So, $V = V - lr \bullet \sum_{i=1}^{t} \frac{\partial L^{(i)}}{\partial o^{(t)}} \bullet \frac{\partial o^{(t)}}{\partial V}$

$$\frac{\partial L^{(1)}}{\partial U} = \frac{\partial L^{(1)}}{\partial o^{(1)}} \bullet \frac{\partial o^{(1)}}{\partial h^{(1)}} \bullet \frac{\partial h^{(1)}}{\partial U}$$

$$\frac{\partial L^{(2)}}{\partial U} = \frac{\partial L^{(2)}}{\partial o^{(2)}} \bullet \frac{\partial o^{(2)}}{\partial h^{(2)}} \bullet \frac{\partial h^{(2)}}{\partial U} + \frac{\partial L^{(2)}}{\partial o^{(2)}} \bullet \frac{\partial o^{(2)}}{\partial h^{(2)}} \bullet \frac{\partial h^{(2)}}{\partial h^{(1)}} \bullet \frac{\partial h^{(1)}}{\partial U}$$

$$\frac{\partial L^{(3)}}{\partial U} = \frac{\partial L^{(3)}}{\partial o^{(3)}} \bullet \frac{\partial o^{(3)}}{\partial h^{(3)}} \bullet \frac{\partial h^{(3)}}{\partial U} + \frac{\partial L^{(3)}}{\partial o^{(3)}} \bullet \frac{\partial o^{(3)}}{\partial h^{(3)}} \bullet \frac{\partial h^{(3)}}{\partial h^{(2)}} \bullet \frac{\partial h^{(2)}}{\partial U} + \frac{\partial L^{(3)}}{\partial o^{(3)}} \bullet \frac{\partial o^{(3)}}{\partial h^{(3)}} \bullet \frac{\partial h^{(3)}}{\partial h^{(2)}} \bullet \frac{\partial h^{(2)}}{\partial h^{(1)}} \bullet \frac{\partial h^{(1)}}{\partial U}$$

$$\frac{\partial L}{\partial U} = \sum_{i=1}^{3} \frac{\partial L^{(i)}}{\partial U}$$

So,

$$U = U - lr \bullet \frac{\partial L}{\partial U}$$

$$\frac{\partial L^{(1)}}{\partial W} = \frac{\partial L^{(1)}}{\partial o^{(1)}} \bullet \frac{\partial o^{(1)}}{\partial h^{(1)}} \bullet \frac{\partial h^{(1)}}{\partial W}$$

$$\frac{\partial L^{(2)}}{\partial W} = \frac{\partial L^{(2)}}{\partial o^{(2)}} \bullet \frac{\partial o^{(2)}}{\partial h^{(2)}} \bullet \frac{\partial h^{(2)}}{\partial W} + \frac{\partial L^{(2)}}{\partial o^{(2)}} \bullet \frac{\partial o^{(2)}}{\partial h^{(2)}} \bullet \frac{\partial h^{(2)}}{\partial h^{(1)}} \bullet \frac{\partial h^{(1)}}{\partial W}$$

$$\cancel{\frac{\partial L^{(3)}}{\partial W} = \frac{\partial L^{(3)}}{\partial o^{(3)}} \bullet \frac{\partial o^{(3)}}{\partial h^{(3)}} \bullet \frac{\partial h^{(3)}}{\partial W} + \frac{\partial L^{(3)}}{\partial o^{(3)}} \bullet \frac{\partial o^{(3)}}{\partial h^{(3)}} \bullet \frac{\partial h^{(3)}}{\partial h^{(2)}} \bullet \frac{\partial h^{(2)}}{\partial W} + \frac{\partial L^{(3)}}{\partial o^{(3)}} \bullet \frac{\partial o^{(3)}}{\partial h^{(3)}} \bullet \frac{\partial h^{(3)}}{\partial h^{(2)}} \bullet \frac{\partial h^{(2)}}{\partial h^{(1)}} \bullet \frac{\partial h^{(1)}}{\partial W}}$$

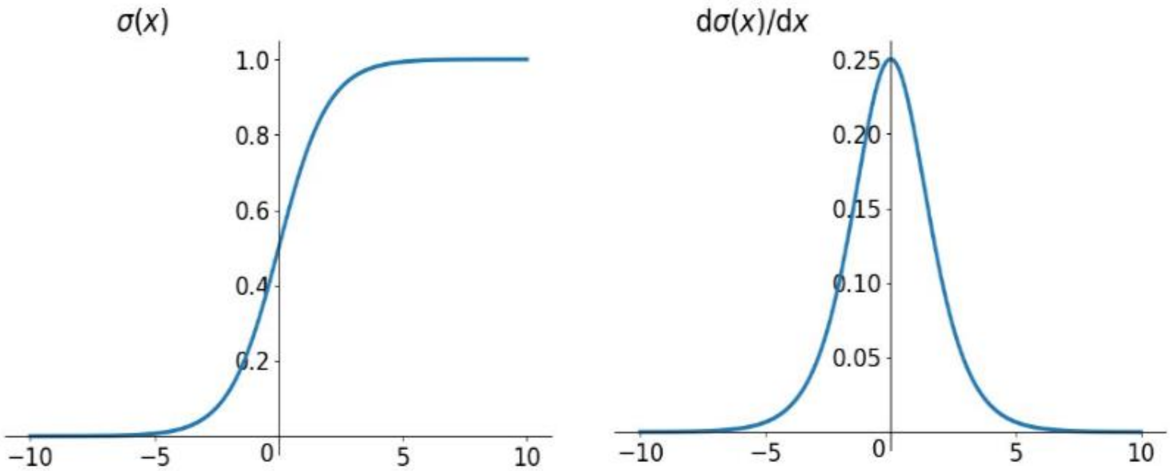
$$\frac{\partial L}{\partial W} = \sum_{i=1}^{3} \frac{\partial L^{(i)}}{\partial W}$$

So,

$$W = W - lr \bullet \frac{\partial L}{\partial W}$$

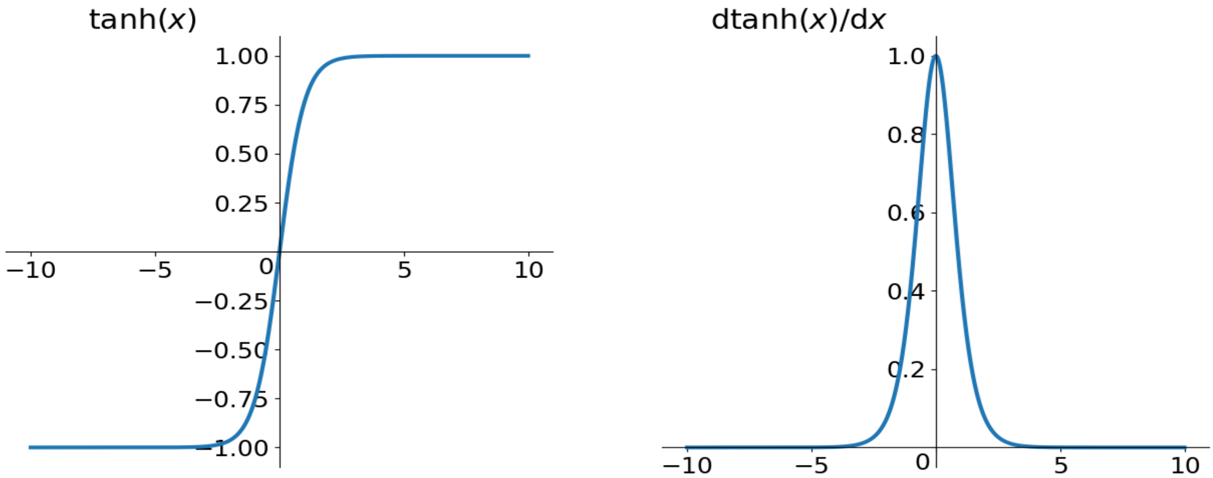
2)

(a)Sigmoid



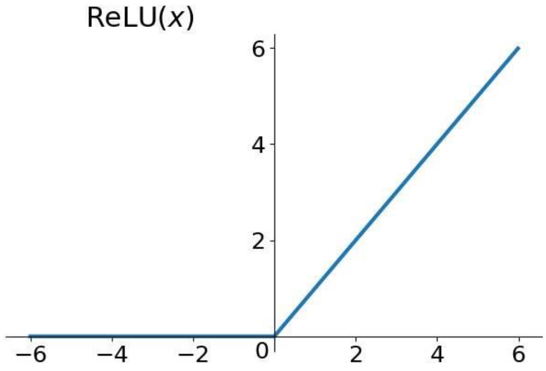
从上图中可以看出，sigmoid 函数的导数取值范围为(0,0.025]，反向传播时每到一个层，梯度变化都会至少缩小四倍，传到神经网络前部很容易造成梯度消失。同时，sigmoid 函数的输出不是中心对称，均大于 0，称为偏移现象，这就导致后一层的神经元会将上一层输出的非 0 均值的信号也学习到作为此层的输入，易学习到噪声。

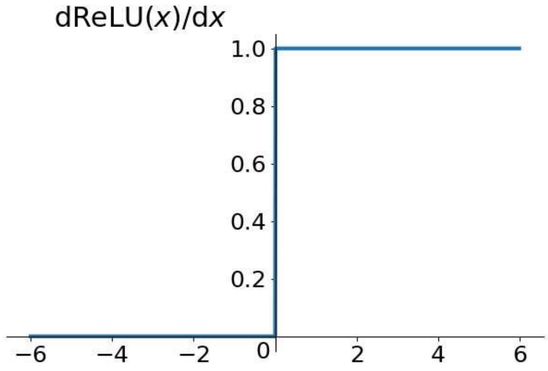
（b）tanh



从图中可以看出，tanh 函数的输出关于零点中心对称，网络收敛性更好，同时，tanh 函数的导数范围为(0,1]，反向传播每经过一层，梯度也会消失，但变化速度较 sigmoid 函数更慢。

（c）Relu





从图中可知，relu 函数的导数左侧为 0，右侧为 1，在一定程度上避免了梯度消失的问题，但是与激活函数相乘的另一个因子在反向传播中呈现增长的趋势，则恒为 1 的导数容易引起梯度爆炸，而恒为 0 的导数有可能把神经元学死，设置合适的步长可有效避免这个问题的发生。