1)由于L随时间累积，我们有

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

对V而言

$$\frac{\partial L}{\partial V} = \sum_{t=1}^{n} \frac{\partial L^{(t)}}{\partial O^{(t)}} \frac{\partial O^{(t)}}{\partial V}$$

再由梯度下降

$$V = V - lr * \sum_{t=1}^{n} \frac{\partial L^{(t)}}{\partial O^{(t)}} \frac{\partial O^{(t)}}{\partial V}$$

对于U, V，我们只需要推导到第三次更新($t = 3$)。唯一与上面不同的是更新过程中的$h^{(t)}$也是函数，需要用链式法则再次展开。于是

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

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

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

在由L的公式，我们有：

$$U = U - lr * \sum_{i=1}^{j} \frac{\partial L^{(i)}}{\partial U}, j = 1, 2, 3$$

W的公式完全一样。

2)

$$\sigma'(x) = \sigma(x)(1 - \sigma(x)) \leq \frac{1}{4} < 1$$

从而在训练深层RNN的时候会出现梯度消失问题

$$\tan'(x) = \frac{1}{1 + x^2} \in (0, 1]$$

同上，也会产生梯度消失问题

$$Relu'(x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0 \end{cases}$$

当取值为0时，会有梯度消失问题，当取值为1则不会产生问题。考虑到可以调整其突变的点，一般来说使用RELU函数不会出现梯度消失或爆炸的问题。