

## Homework\_3

Lv\_Xinpeng

### Part\_1

1) Assume learning rate is lr, then :

$$V = V - lr \cdot \frac{\partial L}{\partial V}$$

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

For RNN, since we have a loss function at each position of the sequence,

the final loss L is :  $L = \sum_{i=1}^t L^{(i)}$  , so:

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

The solution of the partial guide of U involves historical data.:

$$W = W - lr \cdot \frac{\partial L^{(3)}}{\partial W}$$

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

$$W = W - lr \cdot \left( \frac{\partial L^{(3)}}{\partial o^{(3)}} \cdot \frac{\partial o^{(3)}}{\partial h^{(3)}} \cdot \frac{\partial h^{(3)}}{\partial W} + \frac{\partial L^{(3)}}{\partial o^{(3)}} \cdot \frac{\partial o^{(3)}}{\partial h^{(3)}} \cdot \frac{\partial h^{(3)}}{\partial h^{(2)}} \cdot \frac{\partial h^{(2)}}{\partial W} + \frac{\partial L^{(3)}}{\partial o^{(3)}} \cdot \frac{\partial o^{(3)}}{\partial h^{(3)}} \cdot \frac{\partial h^{(3)}}{\partial h^{(2)}} \cdot \frac{\partial h^{(2)}}{\partial h^{(1)}} \cdot \frac{\partial h^{(1)}}{\partial W} \right)$$

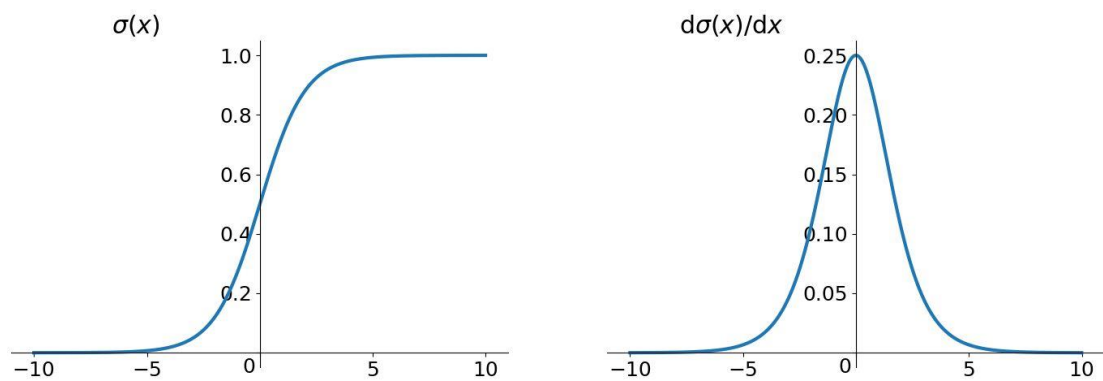
In the same reason:

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

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

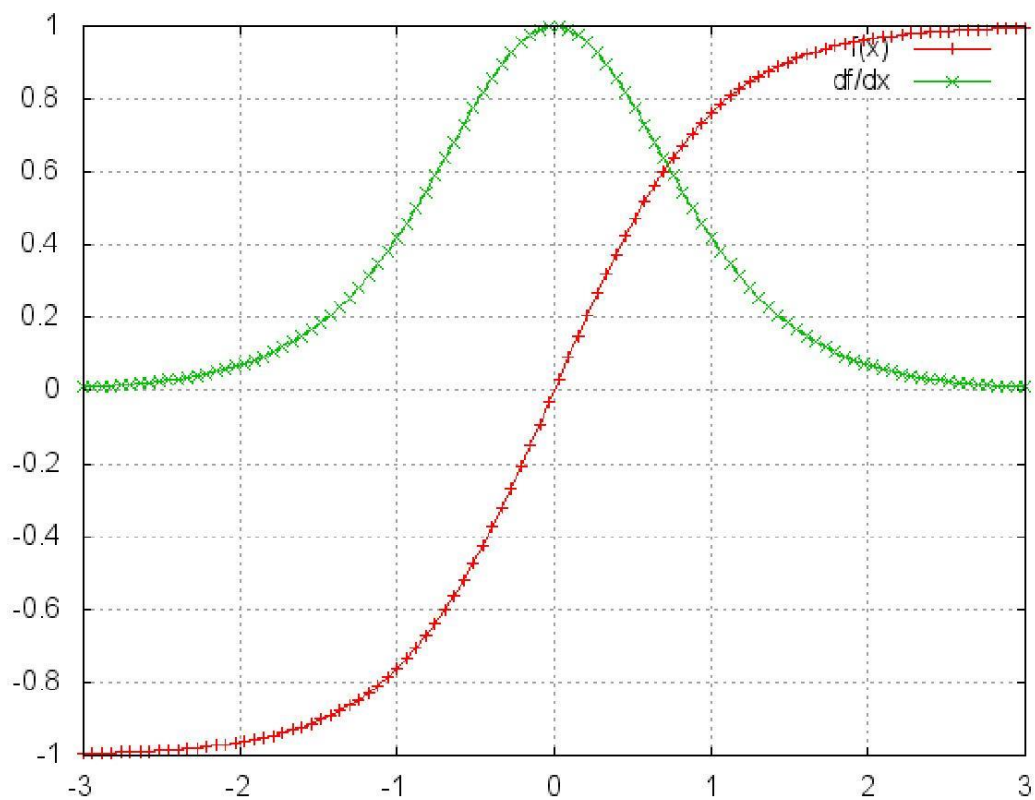
$$U = U - lr \cdot \left( \frac{\partial L^{(3)}}{\partial o^{(3)}} \cdot \frac{\partial o^{(3)}}{\partial h^{(3)}} \cdot \frac{\partial h^{(3)}}{\partial U} + \frac{\partial L^{(3)}}{\partial o^{(3)}} \cdot \frac{\partial o^{(3)}}{\partial h^{(3)}} \cdot \frac{\partial h^{(3)}}{\partial h^{(2)}} \cdot \frac{\partial h^{(2)}}{\partial U} + \frac{\partial L^{(3)}}{\partial o^{(3)}} \cdot \frac{\partial o^{(3)}}{\partial h^{(3)}} \cdot \frac{\partial h^{(3)}}{\partial h^{(2)}} \cdot \frac{\partial h^{(2)}}{\partial h^{(1)}} \cdot \frac{\partial h^{(1)}}{\partial U} \right)$$

For the sigmoid activation-function, its function image and derivative image are as follows:



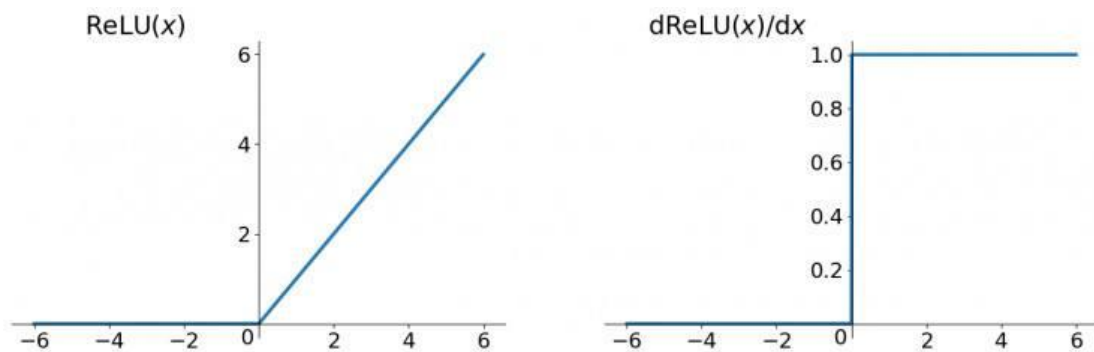
In the partial derivatives of W and U, there is a process in which the activation-function continuously seeks and multiplied. So, from the picture above, the derivative range of the sigmoid function is (0,0.25). It will cause the gradient-disappearance.

And for the tanh activation-function:



From the picture above, the derivative range of the tanh function is (0,1). Just like sigmoid, it will cause the gradient-disappearance.

For the ReLU activation-function :



The left derivative of the ReLU function is 0, and the right derivative is always 1. While the derivative of constant 1 is easy to cause gradient-explosion. Also the 0 derivative makes the gradient disappeared.