Homework_3

Lv_Xinpeng

Part_1

1) Assume learning rate is lr, then :

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

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

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：

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

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

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

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

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

$$W^{(4)} = W^{(3)} - lr \bullet (\frac{\partial L^{(3)}}{\partial o^{(3)}} \bullet \frac{\partial o^{(3)}}{\partial h^{(3)}} \bullet \frac{\partial h^{(3)}}{\partial W^{(3)}} + \frac{\partial L^{(3)}}{\partial o^{(3)}} \bullet \frac{\partial o^{(3)}}{\partial h^{(3)}} \bullet \frac{\partial h^{(3)}}{\partial h^{(2)}} \bullet \frac{\partial h^{(2)}}{\partial W^{(2)}} + \frac{\partial L^{(3)}}{\partial o^{(3)}} \bullet \frac{\partial o^{(3)}}{\partial h^{(3)}} \bullet \frac{\partial h^{(3)}}{\partial h^{(2)}} \bullet \frac{\partial h^{(2)}}{\partial h^{(1)}} \bullet \frac{\partial h^{(1)}}{\partial W^{(1)}})$$

In the same reason：

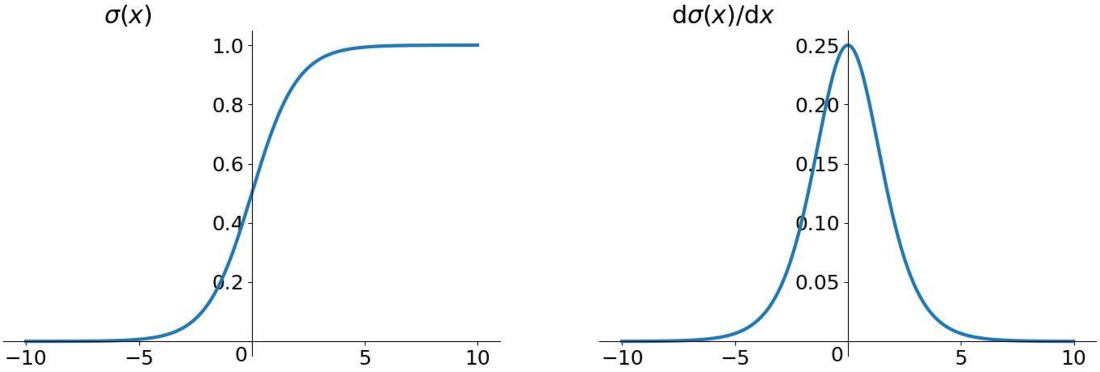
$$U = U - lr \bullet \frac{\partial L}{\partial U}$$

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

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

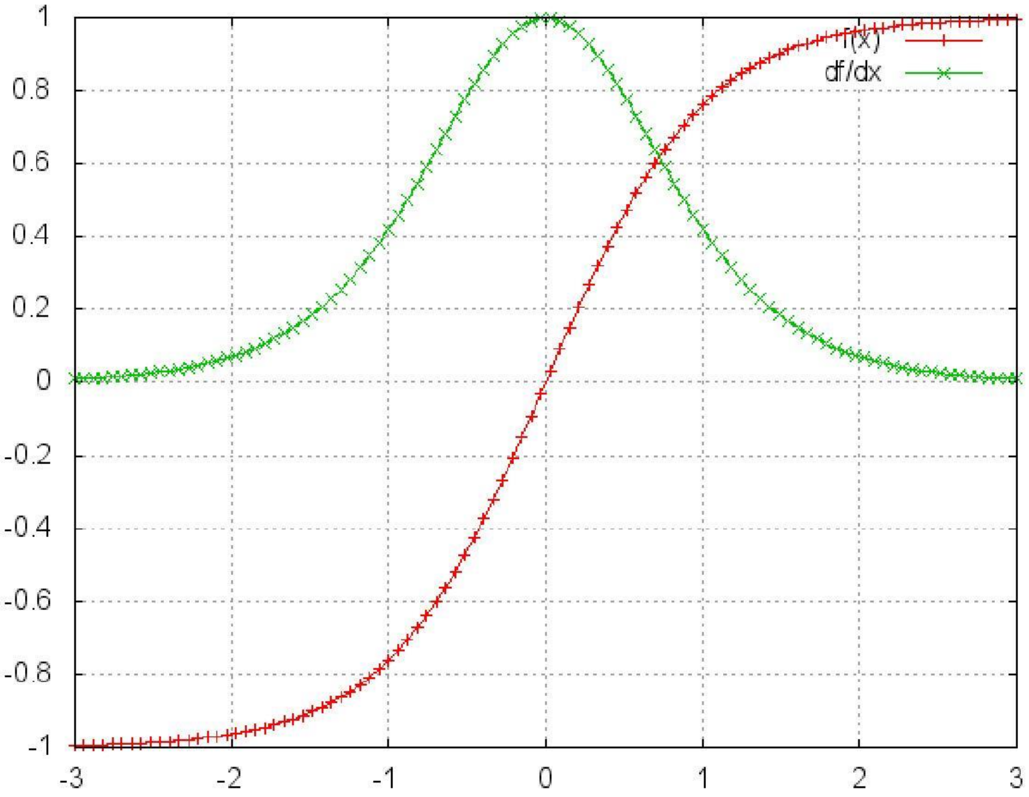
$$U^{(4)} = U^{(3)} - lr \bullet (\frac{\partial L^{(3)}}{\partial o^{(3)}} \bullet \frac{\partial o^{(3)}}{\partial h^{(3)}} \bullet \frac{\partial h^{(3)}}{\partial U^{(3)}} + \frac{\partial L^{(3)}}{\partial o^{(3)}} \bullet \frac{\partial o^{(3)}}{\partial h^{(3)}} \bullet \frac{\partial h^{(3)}}{\partial h^{(2)}} \bullet \frac{\partial h^{(2)}}{\partial U^{(2)}} + \frac{\partial L^{(3)}}{\partial o^{(3)}} \bullet \frac{\partial o^{(3)}}{\partial h^{(3)}} \bullet \frac{\partial h^{(3)}}{\partial h^{(2)}} \bullet \frac{\partial h^{(2)}}{\partial h^{(1)}} \bullet \frac{\partial h^{(1)}}{\partial U^{(1)}})$$

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



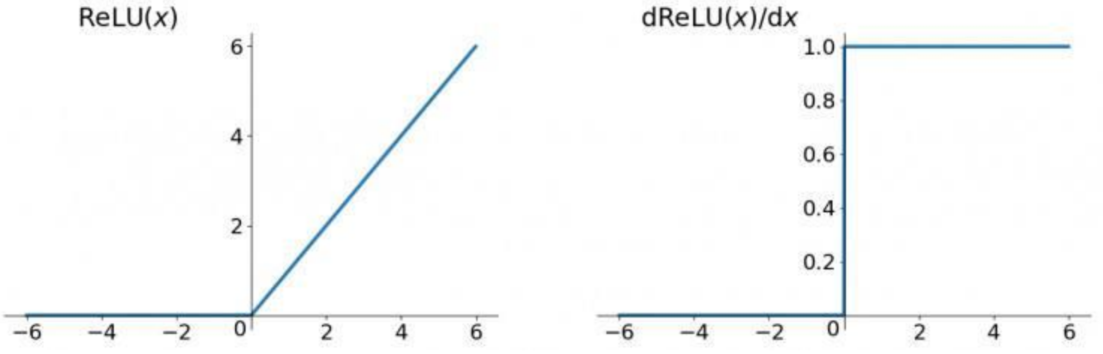
In the partial derivatives of W and U, there is a process in which the activation-function continuously seeks and multiplied. So, from the picture above, the derivative range of the sigmoid function is (0,0.25). It will cause the gradient-disappearance.

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.