

Question 1

Let $z_1=w_1*x+b_1$, $z_2=w*h_1+b$, $z_3=w*h_2+b$, $z_4=w*h_3+b$, $z_5=w*h_4+b$. We can then generate the formula as below:

For the linear chain:
According to the chain rule:

$$\frac{dy}{dw_1} = \frac{dy}{dz_5} \frac{dz_5}{dh_4} \frac{dh_4}{dh_3} \frac{dh_3}{dh_2} \frac{dh_2}{dh_1} \frac{dh_1}{dw_1} = \sigma'(z_5) \frac{dz_5}{dh_4} \sigma'(z_4) \frac{dz_4}{dh_3} \sigma'(z_3) \frac{dz_3}{dh_2} \sigma'(z_2) \frac{dz_2}{dh_1} \sigma'(z_1) \frac{dz_1}{dw_1}$$

$$\frac{dy}{dw_1} = w^4 \sigma'(z_5) \sigma'(z_4) \sigma'(z_3) \sigma'(z_2) \sigma'(z_1) \frac{dz_2}{dh_1} \sigma'(z_1) \frac{dz_1}{dw_1}$$

$$= w^4 x \sigma'(z_5) \sigma'(z_4) \sigma'(z_3) \sigma'(z_2) \sigma'(z_1)$$

For the network with the short circuit connections:

$$h_3^* = \text{sigma}(wh_2th_1+b) = \sigma(z_3th_1) = \sigma(w\sigma(wh_1+b)+h_1+b)$$

$$y^* = \text{sigma}(wh_4th_3^*+b) = \sigma(z_5+h_3^*) = \sigma(w\sigma(wh_3^*+b)+b+h_3^*)$$

So, the derivative of $\frac{dy^*}{dw_1}$ & $\frac{dy^*}{db_1}$ can be simplified as:

$$\frac{dy^*}{dw_1} = \frac{dy^*}{dh_3^*} \frac{dh_3^*}{dh_1} \frac{dh_1}{dw_1} = \sigma'(z_5+h_3^*) [w^2 \sigma'(z_4)+1] \sigma'(z_3th_1) (z_3th_1)$$

$$= w^4 x \sigma'(z_5+h_3^*) + w^2 x \sigma'(z_5+h_3^*) \sigma'(z_4) \sigma'(z_3th_1) \sigma'(z_2) \sigma'(z_1) +$$

$$w^2 x \sigma'(z_5+h_3^*) \sigma'(z_4) \sigma'(z_3th_1) \sigma'(z_1) +$$

$$w^2 x \sigma'(z_5+h_3^*) \sigma'(z_3th_1) \sigma'(z_2) \sigma'(z_1) + x \sigma'(z_5+h_3^*) \sigma'(z_3th_1) \sigma'(z_1)$$

$$\frac{dy^*}{db_1} = \sigma'(z_5+h_3^*) [w^2 \sigma'(z_4)+1] \sigma'(z_3th_1) [w^2 \sigma'(z_2)+1] \sigma'(z_1)$$

In dy^*/dw_1 & dy^*/db_1 , we can find some additional element comparing with dy/dw_1 & dy/db_1 . So, we can conclude that there are $|dy^*/dw_1| > |dy/dw_1|$ and $|dy^*/db_1| > |dy/db_1|$.

Question 2

1. Let the initial value of θ as θ_0 . Then the points θ will go are $\theta_0+0.3$, $\theta_0+0.6$, $\theta_0+0.9$, $\theta_0+1.2$, and then it will come back to $\theta_0+0.9$, and converge.
2. Now we let $\theta_0=-1$. If we keep on implementing our Adam optimizer without stopping, we will find a maximum the θ will reach, which is exactly equalling to the value of h .

By implementing the Adam optimization by code, we generated a series of value of θ (as below), and we can conclude that the maximum is 0.4101, meaning the maximum of h is also the same as 0.4101.

```
[-1,  
-0.7,  
-0.399999999999999786,  
-0.099999999999999776,  
0.200000000000000367,  
0.35348343141804006,  
0.41018429512997384,  
0.39851277161401233,  
0.3362158088702168,  
0.23511522423117243,  
0.10347681809094891]
```

Question 3

