$$\sinh(x) = \frac{e^x - e^{-x}}{2} \qquad \cosh(x) = \frac{e^x + e^{-x}}{2} \qquad \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$\tanh'(x) = 1 - \tanh^2(x)$$

$$S_1 = \tanh(UX_1 + WS_0 + b_1) \qquad O_1 = VS_1 + b_2$$

$$S_2 = \tanh(UX_2 + WS_1 + b_1) \qquad O_2 = VS_2 + b_2$$

$$S_3 = \tanh(UX_3 + WS_2 + b_1) \qquad O_3 = VS_3 + b_3$$

$$L_3 = \frac{1}{2}(Y_3 - O_3)^2$$

$$\frac{\partial L_3}{\partial U} = \frac{\partial L_3}{\partial O_3} \cdot \frac{\partial O_3}{\partial S_3} \cdot \frac{\partial S_3}{\partial U}$$

$$= (O_3 - Y_3) \cdot V \cdot \frac{\partial S_3}{\partial U}$$

$$\frac{\partial S_3}{\partial U} = \cdots\cdots$$

令 $\theta_3 = UX_3 + WS_2 + b_1 \qquad S_3 = \tanh(\theta_3)$

① $$\boxed{\frac{\partial S_3}{\partial U}} = \frac{\partial S_3}{\partial \theta_3} \cdot \frac{\partial \theta_3}{\partial U}$$

$$= \frac{\partial S_3}{\partial \theta_3} \cdot \left(\frac{\partial \theta_3}{\partial U} + \underline{\frac{\partial \theta_3}{\partial S_2} \cdot \frac{\partial S_2}{\partial U}} \right)$$

$$= \frac{\partial \tanh(\theta_3)}{\partial \theta_3} x_3 + \frac{\partial \tanh(\theta_3)}{\partial \theta_3} \cdot \boxed{\frac{\partial S_2}{\partial U}} \cdot W$$

② $$\boxed{\frac{\partial S_2}{\partial U}} = \frac{\partial \tanh(\theta_2)}{\partial \theta_2} x_2 + \frac{\partial \tanh(\theta_2)}{\partial \theta_2} \cdot \boxed{\frac{\partial S_1}{\partial U}} \cdot W$$

③ $$\boxed{\frac{\partial S_1}{\partial U}} = \frac{\partial \tanh(\theta_1)}{\partial \theta_1} x_1$$

④

$$\frac{\partial S_3}{\partial U} = \frac{\partial \tanh(\theta_3)}{\partial \theta_3} x_3 + \frac{\partial \tanh(\theta_3)}{\partial \theta_3} \cdot W \Big(\frac{\partial \tanh(\theta_2)}{\partial \theta_2} x_2 + \frac{\partial \tanh(\theta_2)}{\partial \theta_2} \cdot W \Big(\frac{\partial \tanh(\theta_1)}{\partial \theta_1} x_1 \Big)\Big)$$

$$= \frac{\partial \tanh(\theta_3)}{\partial \theta_3} x_3 + \frac{\partial \tanh(\theta_3)}{\partial \theta_3} \cdot \frac{\partial \tanh(\theta_2)}{\partial \theta_2} \cdot x_2 W$$

$$+ \frac{\partial \tanh(\theta_3)}{\partial \theta_3} \cdot \frac{\partial \tanh(\theta_2)}{\partial \theta_2} \cdot \frac{\partial \tanh(\theta_1)}{\partial \theta_1} x_1 W^2$$

$$\Rightarrow \sum_{K=1}^{3} \frac{\partial L_3}{\partial U} = \frac{\partial L_3}{\partial O_3} \cdot \frac{\partial O_3}{\partial S_3} \cdot$$

$$= \sum_{K=1}^{3} \frac{\partial L_3}{\partial O_3} \cdot \frac{\partial O_3}{\partial S_3} \cdot \Big(\prod_{j=4-K}^{3} \frac{\partial \tanh(\theta_j)}{\partial \theta_j} \Big) x_K W^{K-1}$$

$\tanh'(x) = 1 - \tanh^2(x)$

$\tanh(x) \in [-1, 1]$ $\tanh'(x) \in [0, 1]$

假设 $t=20$

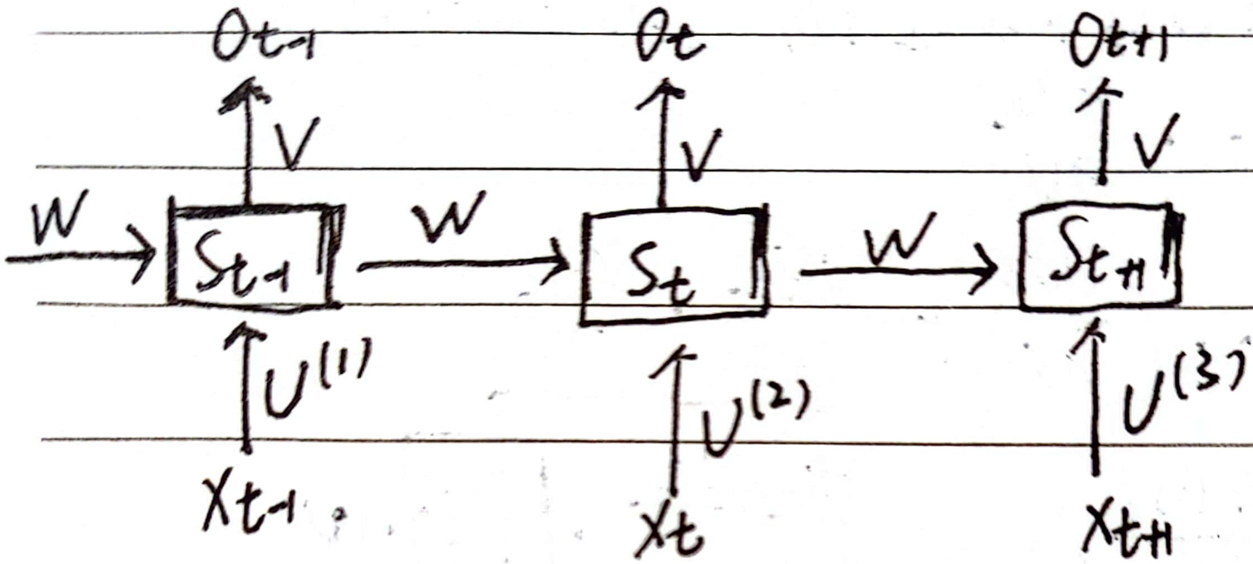
$$\frac{\partial L_{20}}{\partial U} = \frac{\partial L_{20}}{\partial O_{20}} \cdot \frac{\partial O_{20}}{\partial S_{20}} \cdot \Big(\frac{\partial \tanh(\theta_{20})}{\partial \theta_{20}} \cdot x_{20}$$

$$+ \frac{\partial \tanh(\theta_{20})}{\partial \theta_{20}} \cdot \frac{\partial \tanh(\theta_{19})}{\partial \theta_{19}} x_{19} W$$

$$+ \frac{\partial \tanh(\theta_{20})}{\partial \theta_{20}} \cdot \frac{\partial \tanh(\theta_{19})}{\partial \theta_{19}} \cdot \frac{\partial \tanh(\theta_{18})}{\partial \theta_{18}} \cdot x_{18} \cdot W^2$$

$+ \cdots$

$+ \cdots 0 + 0 + \cdots \Big)$



$t=3$, $\frac{\partial L_3}{\partial U} = \frac{\partial L_3}{\partial U^{(3)}} + \frac{\partial L_3}{\partial U^{(2)}} + \frac{\partial L_3}{\partial U^{(1)}}$ [从前向传播]

1> $\frac{\partial L_3}{\partial U^{(3)}} = \frac{\partial L_3}{\partial O_3} \cdot \frac{\partial O_3}{\partial S_3} \cdot \frac{\partial S_3}{\partial U^{(3)}}$

2> $\frac{\partial L_3}{\partial U^{(2)}} = \frac{\partial L_3}{\partial O_3} \cdot \frac{\partial O_3}{\partial S_3} \cdot \underline{\frac{\partial S_3}{\partial S_2}} \cdot \frac{\partial S_2}{\partial U^{(2)}}$

3> $\frac{\partial L_3}{\partial U^{(1)}} = \frac{\partial L_3}{\partial O_3} \cdot \frac{\partial O_3}{\partial S_3} \cdot \underline{\frac{\partial S_3}{\partial S_2} \cdot \frac{\partial S_2}{\partial S_1}} \cdot \frac{\partial S_1}{\partial U^{(1)}}$

$$\frac{\partial L_3}{\partial V}=\frac{\partial L_3}{\partial V^{(3)}}+\frac{\partial L_3}{\partial V^{(2)}}+\frac{\partial L_3}{\partial V^{(1)}}$$

$$=\frac{\partial L_3}{\partial O_3}\cdot\frac{\partial O_3}{\partial S_3}\cdot\frac{\partial S_3}{\partial V^{(3)}}$$

$$+\frac{\partial L_3}{\partial O_3}\cdot\frac{\partial O_3}{\partial S_3}\cdot\frac{\partial S_3}{\partial S_2}\cdot\frac{\partial S_2}{\partial V^{(2)}}$$

$$+\frac{\partial L_3}{\partial O_3}\cdot\frac{\partial O_3}{\partial S_3}\cdot\frac{\partial S_3}{\partial S_2}\cdot\frac{\partial S_2}{\partial S_1}\cdot\frac{\partial S_1}{\partial V^{(1)}}$$

$$=\frac{\partial L_3}{\partial O_3}\cdot\frac{\partial O_3}{\partial S_3}\cdot\frac{\partial S_3}{\partial V^{(3)}}$$

$$+\frac{\partial L_3}{\partial O_3}\cdot\frac{\partial O_3}{\partial S_3}\cdot\left(\begin{array}{l}\dfrac{\partial S_3}{\partial S_2}\cdot\dfrac{\partial S_2}{\partial V^{(2)}}\\ +\dfrac{\partial S_3}{\partial S_2}\cdot\dfrac{\partial S_2}{\partial S_1}\cdot\dfrac{\partial S_1}{\partial V^{(1)}}\end{array}\right)$$

$$=\frac{\partial L_3}{\partial O_3}\cdot\frac{\partial O_3}{\partial S_3}\cdot\frac{\partial S_3}{\partial V^{(3)}}$$

$$+\frac{\partial L_3}{\partial O_3}\cdot\frac{\partial O_3}{\partial S_3}\cdot\left(\begin{array}{l}\dfrac{\partial S_2}{\partial S_1}\cdot\dfrac{\partial S_3}{\partial S_2}\cdot\dfrac{\partial S_1}{\partial V^{(1)}}\\ +\dfrac{\partial S_3}{\partial S_2}\cdot\dfrac{\partial S_2}{\partial V^{(2)}}\end{array}\right)\begin{array}{l}k=1\\ \\ k=2\end{array}$$

$$=\frac{\partial L_3}{\partial O_3}\cdot\frac{\partial O_3}{\partial S_3}\cdot\frac{\partial S_3}{\partial V^{(3)}}$$

$$+\sum_{k=1}^{2}\cdot\frac{\partial L_3}{\partial O_3}\cdot\frac{\partial O_3}{\partial S_3}\cdot\left(\prod_{j=k+1}^{3}\frac{\partial S_j}{\partial S_{j-1}}\right)\frac{\partial S_k}{\partial V^{(k)}}$$

k↑, j↑, 连乘项越少; k↓, j↓ 连乘项越多

在任意t时刻下，有

$$\frac{\partial L_t}{\partial U}=\frac{\partial L_t}{\partial U^{(t)}}+\cdots+\frac{\partial L_t}{\partial U^{(3)}}+\frac{\partial L_t}{\partial U^{(2)}}+\frac{\partial L_t}{\partial U^{(1)}}$$

$$=\frac{\partial L_t}{\partial O_t}\cdot\frac{\partial O_t}{\partial S_t}\cdot\frac{\partial S_t}{\partial U^{(t)}}$$

$$+\sum_{k=1}^{t-1}\frac{\partial L_t}{\partial O_t}\cdot\frac{\partial O_t}{\partial S_t}\left(\prod_{j=k+1}^{t}\frac{\partial S_j}{\partial S_{j-1}}\right)\frac{\partial S_k}{\partial U^{(k)}}$$

$$\frac{\partial S_j}{\partial S_{j-1}}=\tanh'(\theta_j)W$$

$$=[1-\tanh^2(\theta_j)]W$$

$K\downarrow$，$j\downarrow$ 连乘的项越多

$W\uparrow$ 梯度爆炸

$W\downarrow$ 梯度消失

前面的消息无法传到当前时刻，失去序列性.

$t=20$

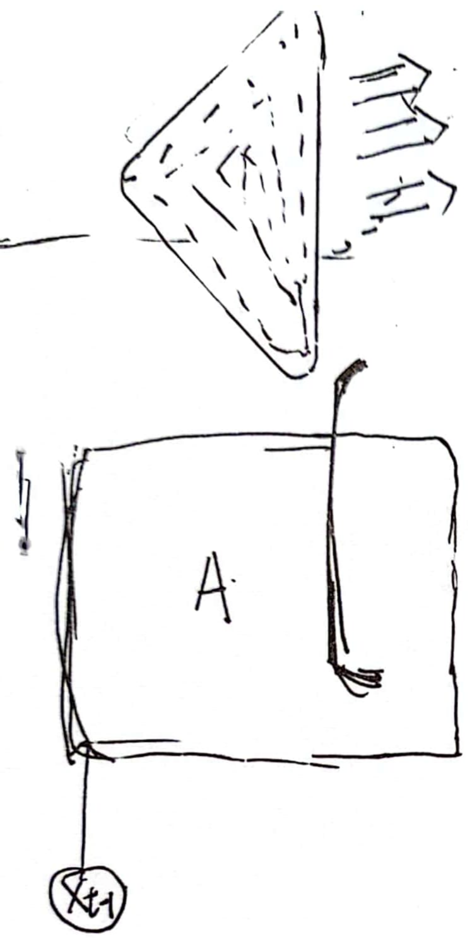
$$\frac{\partial L_{20}}{\partial U}=\frac{\partial L_{20}}{\partial O_{20}}\cdot\frac{\partial O_{20}}{\partial S_{20}}\cdot\frac{\partial S_{20}}{\partial U^{(20)}}$$

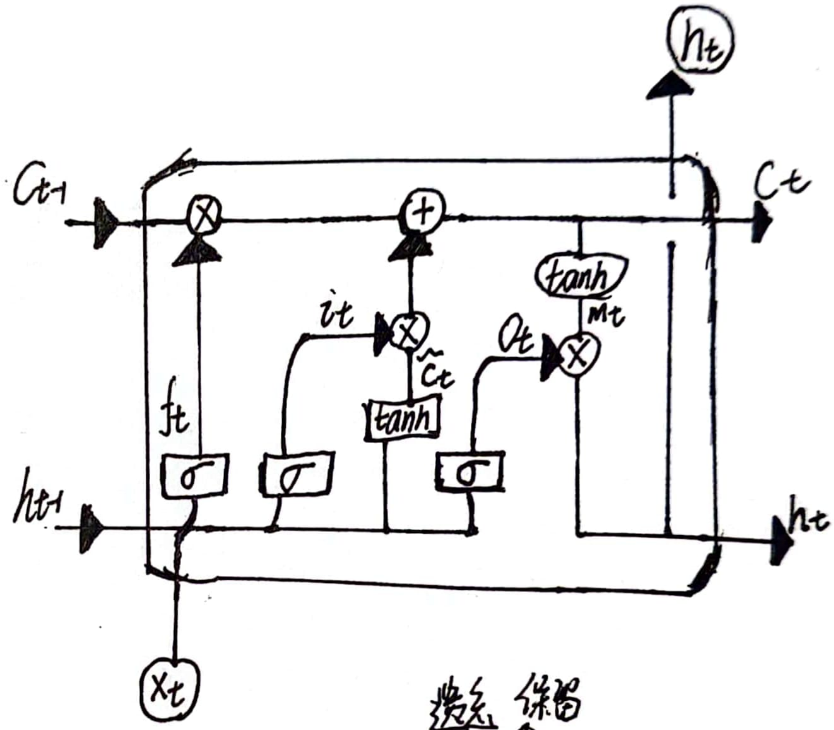
$$+\frac{\partial L_{20}}{\partial O_{20}}\cdot\frac{\partial O_{20}}{\partial S_{20}}\cdot\left(\frac{\partial S_{20}}{\partial S_{19}}\cdot\frac{\partial S_{19}}{\partial U^{(19)}}+\frac{\partial S_{20}}{\partial S_{19}}\cdot\frac{\partial S_{19}}{\partial S_{18}}\cdot\frac{\partial S_{18}}{\partial U^{(18)}}+\cdots\right)$$

$+\cdots$

$+\cdots$

$+\cdots$





$f_t = \sigma(W$
$g_t = tan$

eg: 场景：期末考试周

A时间步：考高等数学

下一时间步：考线性代数

X_t：复习线性代数

h_t：考试得分

C_t：新的记忆

C_{t-1}：旧的记忆，考完高数的记忆

h_{t-1}：考完高数当下的状态

目标：h_t 尽可能好 [有监督学习]

遗忘 保留

σ：[0,1] → [0 1 … 0 1 1]

f_t：考完高数遗忘与线代无关的知识 [遗忘门]

~~i_t~~

$\tilde{C}_t$：tanh：相当于复习线代生成新的记忆.

i_t：生成的新的记忆并不是都有用，通过 i_t σ 进行更新 [更新门]

C_t：旧的记忆 + 数学的记忆 = 新记忆

$$\underbrace{f_t * C_{t-1}}_{\text{数学基础能力}} + \underbrace{i_t * \tilde{C}_t}_{\text{线代考点内容}}$$

$t=3$

$$\frac{\partial L_3}{\partial W_{xf}} = \frac{\partial L_3}{\partial W_{xf}^{(3)}} + \frac{\partial L_3}{\partial W_{xf}^{(2)}} + \frac{\partial L_3}{\partial W_{xf}^{(1)}}$$

① $$\frac{\partial L_3}{\partial W_{xf}^{(3)}} = \frac{\partial L_3}{\partial y_3}\cdot\frac{\partial y_3}{\partial h_3}\cdot\frac{\partial h_3}{\partial m_3}\cdot\frac{\partial m_3}{\partial c_3}\cdot\frac{\partial c_3}{\partial f_3}\cdot\frac{\partial f_3}{\partial W_{xf}^{(3)}}$$

$$= \frac{\partial L_3}{\partial y_3}\cdot\frac{\partial y_3}{\partial h_3}\cdot\frac{\partial h_3}{\partial c_3}\cdot\frac{\partial c_3}{\partial f_3}\cdot\frac{\partial f_3}{\partial W_{xf}^{(3)}}$$

② $$\frac{\partial L_3}{\partial W_{xf}^{(2)}} = \frac{\partial L_3}{\partial y_3}\cdot\frac{\partial y_3}{\partial h_3}\cdot\left(\begin{aligned}&\frac{\partial h_3}{\partial o_3}\cdot\frac{\partial o_3}{\partial h_2}\cdot\frac{\partial h_2}{\partial c_2}\cdot\frac{\partial c_2}{\partial f_2}\cdot\frac{\partial f_2}{\partial W_{xf}^{(2)}}\\ &+\frac{\partial h_3}{\partial c_3}\cdot\frac{\partial c_3}{\partial c_2}\cdot\frac{\partial c_2}{\partial f_2}\cdot\frac{\partial f_2}{\partial W_{xf}^{(2)}}\\ &+\frac{\partial h_3}{\partial c_3}\cdot\frac{\partial c_3}{\partial f_3}\cdot\frac{\partial f_3}{\partial h_2}\cdot\frac{\partial h_2}{\partial c_2}\cdot\frac{\partial c_2}{\partial f_2}\cdot\frac{\partial f_2}{\partial W_{xf}^{(2)}}\\ &+\frac{\partial h_3}{\partial c_3}\cdot\frac{\partial c_3}{\partial i_3}\cdot\frac{\partial i_3}{\partial h_2}\cdot\frac{\partial h_2}{\partial c_2}\cdot\frac{\partial c_2}{\partial f_2}\cdot\frac{\partial f_2}{\partial W_{xf}^{(2)}}\\ &+\frac{\partial h_3}{\partial c_3}\cdot\frac{\partial c_3}{\partial g_3}\cdot\frac{\partial g_3}{\partial h_2}\cdot\frac{\partial h_2}{\partial c_2}\cdot\frac{\partial c_2}{\partial f_2}\cdot\frac{\partial f_2}{\partial W_{xf}^{(2)}}\end{aligned}\right)$$

$$= \frac{\partial L_3}{\partial y_3}\cdot\frac{\partial y_3}{\partial h_3}\left(\begin{aligned}&\frac{\partial h_3}{\partial o_3}\cdot\frac{\partial o_3}{\partial h_2}\cdot\frac{\partial h_2}{\partial c_2}\\ &+\frac{\partial h_3}{\partial c_3}\left(\begin{aligned}&\frac{\partial c_3}{\partial c_2}\\ &+\frac{\partial c_3}{\partial f_3}\cdot\frac{\partial f_3}{\partial h_2}\cdot\frac{\partial h_2}{\partial c_2}\\ &+\frac{\partial c_3}{\partial i_3}\cdot\frac{\partial i_3}{\partial h_2}\cdot\frac{\partial h_2}{\partial c_2}\\ &+\frac{\partial c_3}{\partial g_3}\cdot\frac{\partial g_3}{\partial h_2}\cdot\frac{\partial h_2}{\partial c_2}\end{aligned}\right\}\frac{\partial c_3}{\partial c_2}\end{aligned}\right)\frac{\partial c_2}{\partial f_2}\cdot\frac{\partial f_2}{\partial W_{xf}^{(2)}}$$

③ (1) $t_3 \to t_2$.

$$y_3 \to h_3 \to \begin{cases} o_3 \to h_2 \\ c_3 \to f_3 \to h_2 \\ c_3 \to i_3 \to h_2 \\ c_3 \to c_2 \\ c_3 \to g_3 \to h_2 \end{cases}$$

(2) $t_2 \to t_1$

$$\left\{\begin{array}{l} \left. h_2 \to \begin{cases} o_2 \to h_1 \\ c_2 \to f_2 \to h_1 \\ c_2 \to i_2 \to h_1 \\ c_2 \to g_2 \to h_1 \end{cases} \right\} \to c_1 \to f_1 \to W_{xf}^{(1)} \\ \\ \left. c_2 \to \begin{cases} c_1 \\ f_2 \to h_2 \to c_1 \\ i_2 \to h_2 \to c_1 \\ g_2 \to h_2 \to c_1 \end{cases} \right\} \to f_1 \to W_{xf}^{(1)} \end{array}\right. = \left\{\begin{array}{l} h_2 \to \left\{\begin{array}{l} o_2 \to h_1 \\ c_2 \to c_1 \end{array}\right\} \to f_1 \\ \\ c_2 \to c_1 \to f_1 \end{array}\right.$$

$$= \frac{\partial L_3}{\partial y_3} \cdot \frac{\partial y_3}{\partial h_3} \cdot \left(\begin{array}{l} \frac{\partial h_3}{\partial o_3} \cdot \frac{\partial o_3}{\partial h_2} \cdot \frac{\partial h_2}{\partial c_2} \\ \frac{\partial h_3}{\partial c_3} \cdot \frac{\partial c_3}{\partial c_2} \end{array} \right) \frac{\partial c_2}{\partial f_2} \cdot \frac{\partial f_2}{\partial W_{xf}^{(2)}}$$

---

(3) $t_3 \to t_1$

$$y_3 \to h_3 \to \left(\begin{array}{l}
o_3 \to h_2 \to \begin{pmatrix} o_2 \to h_1 \\ c_2 \to c_1 \end{pmatrix} \to f_1 \to W_{xf}^{(1)} \\
c_3 \to f_3 \to h_2 \to \begin{pmatrix} o_2 \to h_1 \\ c_2 \to c_1 \end{pmatrix} \to f_1 \to W_{xf}^{(1)} \\
c_3 \to i_3 \to h_2 \to \begin{pmatrix} o_2 \to h_1 \\ c_2 \to c_1 \end{pmatrix} \to f_1 \to W_{xf}^{(1)} \\
c_3 \to g_3 \to h_2 \to \begin{pmatrix} o_2 \to h_1 \\ c_2 \to c_1 \end{pmatrix} \to f_1 \to W_{xf}^{(1)} \\
c_3 \to c_2 \to c_1 \to f_1 \to W_{xf}^{(1)}
\end{array}\right)$$

$$= y_3 \to h_3 \left(\begin{array}{l}
\begin{pmatrix} o_3 \to h_2 \\ c_3 \to f_3 \to h_2 \\ c_3 \to i_3 \to h_2 \\ c_3 \to g_3 \to h_2 \end{pmatrix} \to o_2 \to h_1 \to c_1 \to f_1 \to W_{xf}^{(1)} \\
\\
\begin{pmatrix} o_3 \to h_2 \\ c_3 \\ c_3 \to f_3 \to h_2 \\ c_3 \to i_3 \to h_2 \\ c_3 \to g_3 \to h_2 \end{pmatrix} \to c_2 \to c_1 \to f_1 \to W_{xf}^{(1)}
\end{array}\right.$$

$$= y_3 \to h_3 \to \left|\begin{array}{l}
\begin{pmatrix} o_3 \to h_2 \\ c_3 \to f_3 \to h_2 \\ c_3 \to i_3 \to h_2 \\ c_3 \to g_3 \to h_2 \end{pmatrix} \to o_2 \to h_1 \to c_1 \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad \downarrow \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad f_1 \to W_{xf}^{(1)} \\
o_3 \to h_2 \\
\qquad\quad \downarrow \\
c_3 \to c_2 \to c_1 \nearrow
\end{array}\right.$$

的前面知识结合起来。

$$\frac{\partial L_3}{\partial W_{xf}^{(1)}} = \frac{\partial L_3}{\partial y_3}\cdot\frac{\partial y_3}{\partial h_3}\cdot\left(\left(\begin{array}{l}\frac{\partial h_3}{\partial o_3}\cdot\frac{\partial o_3}{\partial h_2}\\ \frac{\partial h_3}{\partial c_3}\cdot\frac{\partial c_3}{\partial f_3}\cdot\frac{\partial f_3}{\partial h_2}\\ \frac{\partial h_3}{\partial c_3}\cdot\frac{\partial c_3}{\partial i_3}\cdot\frac{\partial i_3}{\partial h_2}\\ +\frac{\partial h_3}{\partial c_3}\cdot\frac{\partial c_3}{\partial g_3}\cdot\frac{\partial g_3}{\partial h_2}\end{array}\right)\cdot\frac{\partial h_2}{\partial o_2}\cdot\frac{\partial o_2}{\partial h_1}\cdot\frac{\partial h_1}{\partial c_1} + \left(\begin{array}{l}\frac{\partial h_3}{\partial o_3}\cdot\frac{\partial o_3}{\partial h_2}\cdot\frac{\partial h_2}{\partial c_2}\cdot\frac{\partial c_2}{\partial c_1}\\ \frac{\partial h_3}{\partial c_3}\cdot\frac{\partial c_3}{\partial c_2}\cdot\frac{\partial c_2}{\partial c_1}\end{array}\right)\right)\cdot\frac{\partial c_1}{\partial f_1}\cdot\frac{\partial f_1}{\partial W_{xf}^{(1)}}$$

$\cancel{E_t \quad C_{t-1} \odot f_t + g_t \odot i_t}$

$\cancel{\frac{\partial C_t}{\partial C_{t-1}} = \frac{\partial C_t}{\partial C_{t-1}} + \frac{\partial C_t}{\partial f_t}}$

$$\frac{\partial C_t}{\partial C_{t-1}} = \frac{\partial C_t}{\partial C_{t-1}} \quad : \quad f_t$$

$$+ \frac{\partial C_t}{\partial g_t} \cdot \frac{\partial g_t}{\partial h_{t-1}} \cdot \frac{\partial h_{t-1}}{\partial m_{t-1}} \cdot \frac{\partial m_{t-1}}{\partial C_{t-1}}$$

$$+ \frac{\partial C_t}{\partial i_t} \cdot \frac{\partial i_t}{\partial h_{t-1}} \cdot \frac{\partial h_{t-1}}{\partial m_{t-1}} \cdot \frac{\partial m_{t-1}}{\partial C_{t-1}}$$

$$+ \frac{\partial C_t}{\partial f_t} \cdot \frac{\partial f_t}{\partial h_{t-1}} \cdot \frac{\partial h_{t-1}}{\partial m_{t-1}} \cdot \frac{\partial m_{t-1}}{\partial C_{t-1}}$$

$$= f_t$$

$$+ i_t \cdot \tanh'(\theta) \cdot O_{t-1} \cdot \tanh'(C_{t-1}) \cdot W_{hg}$$

$$+ g_t \cdot \sigma'(\theta) \cdot O_{t-1} \cdot \tanh'(C_{t-1}) \cdot W_{hi}$$

$$+ C_{t-1} \cdot \sigma'(\theta) \cdot O_{t-1} \cdot \tanh'(C_{t-1}) \cdot W_{hf}$$

通过调节 W_{hf}, W_{hi}, W_{hg} 可以灵活控制 $\frac{\partial C_t}{\partial C_{t-1}}$ 的值

当要从 n 时刻长期记忆某个东西直到 m 时刻时，该路径上的 $\prod_{t=n}^{m} \frac{\partial C_t}{\partial C_{t-1}} \approx 1 \times 1 \times \cdots \times 1$，从而缓解梯度消失.

LSTM 可以通过调节参数W_{hi}，W_{hf}，W_{hg}灵活控制记忆细胞$\frac{\partial C_t}{\partial C_{t-1}}$，使得其值接近 1。为什么 RNN 中$\frac{\partial S_t}{\partial S_{t-1}}$不可以通过调节参数 U 来控制其值近似为 1?

$\frac{\partial C_t}{\partial C_{t-1}}$是多个 W 的线性相加的综合结果，其中某个参数 W 很大或者很小，可以由其它的 W 进行协调，不会影响最后的结果。RNN 中$\frac{\partial S_t}{\partial S_{t-1}}$只由一个参数 U 控制，就会导致梯度爆炸或梯度消失。[梯度正常+梯度消失=梯度正常]。