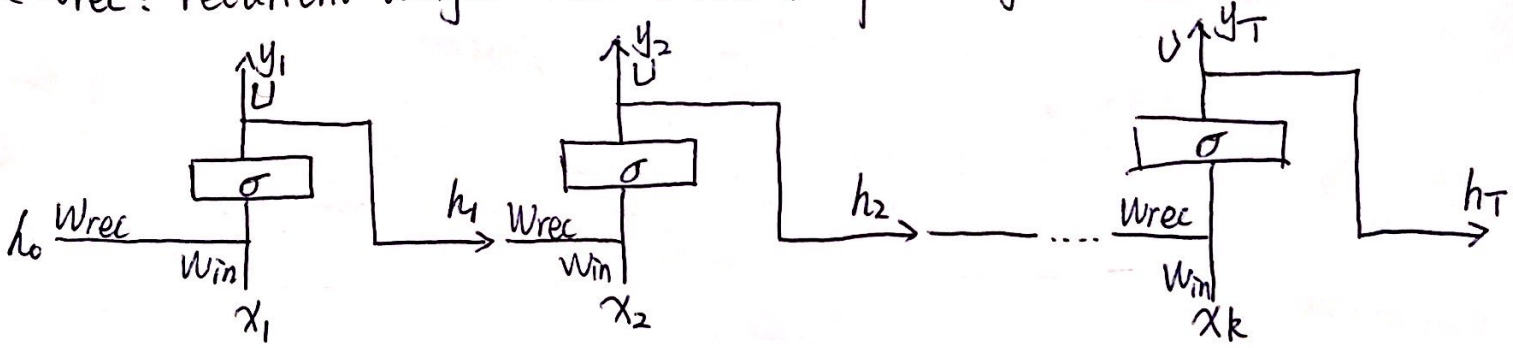
Recurrent Neural Network (RNN)

RNN基本结构

(Wrec: recurrent weight matrix, Win = input weight matrix)



$$h_t = \sigma(W_{rec} h_{t-1} + W_{in} x_t + b_t)$$

Back-prop ~~in~~ through time (BPTT). let $W = [W_{rec}, W_{in}]$

$$\frac{\partial E}{\partial W} = \sum_{t=1}^{T} \frac{\partial E_t}{\partial W}$$ (t: 第t个 timestep)

$$\frac{\partial E_k}{\partial W} = \frac{\partial E_k}{\partial y_k} \cdot \frac{\partial y_k}{\partial h_k} \left(\prod_{t=2}^{k} \frac{\partial h_t}{\partial h_{t-1}} \right) \cdot \frac{\partial h_1}{\partial W}$$

其中: $\frac{\partial h_t}{\partial h_{t-1}} = diag(\sigma'(W_{rec} h_{t-1} + W_{in} x_t + b_t)) W_{rec}$

so: $\prod_{t=2}^{k} \frac{\partial h_t}{\partial h_{t-1}} = \prod_{t=2}^{k} \left(diag(\sigma'(W_{rec} h_{t-1} + W_{in} x_t + b_t)) \cdot W_{rec} \right)$ ①

σ为sigmoid时，σ'最大为$\frac{1}{4}$；σ为tanh或ReLU，最大为1。当网络较深时，即K较大时，①式在连乘下会变得极小导致梯度消失；而当Wrec过大时，①又会导致梯度爆炸.

RNN权向传播解释（BPTT：Back-prop through time）

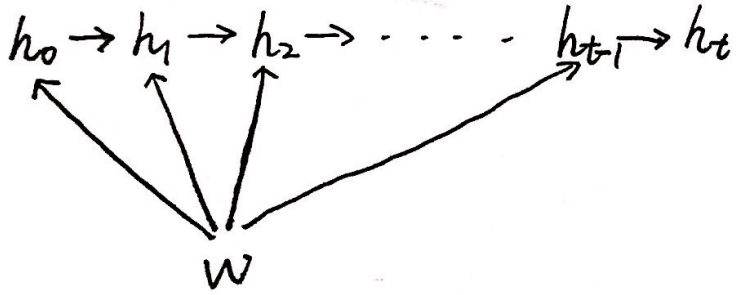
① error是每个timestep的error总和

$$\frac{\partial E}{\partial W}=\sum_{t=1}^{T}\frac{\partial E_t}{\partial W}$$

$$h_t=\sigma(W^{hh}h_{t-1}+W^{hx}x_t)$$

$$\hat{y}_t=\text{softmax}(W^{S}h_t)$$

② 偏导数是与W有关的路径的偏导数之和



每个timestep的 $\frac{\partial E_t}{\partial W}=\frac{\partial E_t}{\partial y_t}\frac{\partial y_t}{\partial h_t}\left(\sum_{k=1}^{t-1}\frac{\partial h_t}{\partial h_k}\right)\frac{\partial h_k}{\partial W}$

其中：$\frac{\partial h_t}{\partial h_k}=\prod_{j=k+1}^{t}\frac{\partial h_j}{\partial h_{j-1}}=\prod_{j=k+1}^{t}W^{hh}\times diag(f'(j_{j-1}))$

↳雅克比矩阵（Jacobian matrix）

diag指对角线为$f'(j_{j-1})$的向量，其余为0的矩阵

③ $\frac{\partial E}{\partial W}=\sum_{t=1}^{T}\sum_{k=1}^{t-1}\frac{\partial E_t}{\partial y_t}\frac{\partial y_t}{\partial h_t}\left(\prod_{j=k+1}^{t}\frac{\partial h_j}{\partial h_{j-1}}\right)\frac{\partial h_k}{\partial W}$

序列越长，对一个序列进行BPTT，每次时刻都有连乘项，刚开始的几层影响不大，越往前传播，W连乘的级数急剧上升。W矩阵初始化不当（小于1或大于1），都会使得梯度消失或梯度爆炸出现。

发生梯度消失，距离当前t timestep越远的单词对预测timestep t+1的单词的能力会减弱（lose long-term dependency）

（可理解为越远的单词通过权重矩阵W^{hh}和W^{hx}形成非线性变换影响h_{t+1}，梯度消失，表示W^{hh}和W^{hx}学不到这个关系。）