e Σe- e·e

输入：$a_t^{(in)} = \sigma(W^{(in)} x_t + b^{(in)})$

隐层：$h_t = \sigma(U a_t^{(in)} + V h_{t-1} + b_h)$

$o_t = \sigma(W h_t + b_o)$

输出：$h_t^{(out)} = W^{(out)} o_t + b^{(out)}$

$\hat{r}_t = Softmax(h_t^{(out)})$

Loss：$L = \sum_{t=1}^{Tr} L_t = \sum_{t=1}^{Tr} Loss(r_t, \hat{r}_t)$

使用交叉熵：$L = \sum_{t=1}^{Tr} -r_t^T \log \hat{r}_t$

$$\frac{\partial \sigma(z)}{\partial z} = -\frac{-e^{-z}}{(1+e^{-z})^2} = \frac{e^{-z}}{(1+e^{-z})^2} = \sigma(z)\cdot(1-\sigma(z))$$

对于SoftMax：令之前为$(X_1, \dots, X_C)$ 则 $Softmax(X_i) = \frac{e^{X_i}}{\sum_{j=1}^{C} e^{X_j}}$

$$\frac{\partial Softmax(X_i)}{\partial X_j} = -\frac{e^{X_j}}{(\sum_{j=1}^{C} e^{X_j})^2} \cdot e^{X_i} \quad (i \neq j)$$

$$\frac{\partial Softmax(X_i)}{\partial X_j} = \frac{e^{X_i}\cdot(\sum_{j=1}^{C} e^{X_j} - e^{X_j})}{(\sum_{j=1}^{C} e^{X_j})^2} \quad (i = j)$$

$$\therefore \frac{\partial L}{\partial h_{t,i}^{(out)}} = \sum_{t=1}^{Tr}\sum_{a=1}^{C} \frac{\partial L_t}{\partial \hat{r}_{t,a}} \cdot \frac{\partial \hat{r}_{t,a}}{\partial h_{t,i}^{(out)}}$$

$$\frac{\partial L_t}{\partial \hat{r}_{t,a}} = -\frac{r_t}{\hat{r}_t}$$

$\frac{\partial r_{ta}}{\partial h_{t,i}^{(out)}}$ 如上式

$$\therefore 上式 = \sum_{t=1}^{Tr}\left[\frac{r_{t,i}}{\hat{r}_{t,i}}\hat{r}_{t,i}(1-\hat{r}_{t,i}) - \sum_{i\neq j}\frac{r_{t,j}}{\hat{r}_{t,j}}\hat{r}_{t,i}\hat{r}_{t,j}\right]$$

$$= -\sum_{t=1}^{Tr}\left[r_{t,i}(1-\hat{r}_{t,i}) - \sum_{i\neq j} r_{t,j}\cdot\hat{r}_{t,i}\right]$$

$$= -\sum_{t=1}^{Tr}\left[r_{t,i} - \hat{r}_{t,i}\sum_{j}^{C} r_{t,j}\right] = \sum_{t=1}^{Tr}(\hat{r}_{t,i} - r_{t,i})$$

根据 $h_t^{(out)} = W^{(out)} o_t + b^{(out)}$

$$\frac{\partial L}{\partial b_i^{(out)}} = \frac{\partial L}{\partial h_t^{(out)}} \cdot \frac{\partial h_t^{(out)}}{\partial b_i^{(out)}} = \sum_{t=1}^{Tr} (\hat{r}_t - r_t)$$

$$\frac{\partial L}{\partial W^{(out)}} = \frac{\partial L}{\partial h_t^{(out)}} \cdot \frac{\partial h_t^{(out)}}{\partial W^{(out)}} = \sum_{t=1}^{Tr} (\hat{r}_t - r_t) \, o_t^T$$

$$\frac{\partial L}{\partial o_t} = \frac{\partial L}{\partial h_t^{(out)}} \cdot \frac{\partial h_t^{(out)}}{\partial o_t} = \sum_{t=1}^{Tr} (\hat{r}_t - r_t) \cdot W^{T}_{(out)}$$

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial o_t} \cdot \frac{\partial o_t}{\partial W} = \sum_{t=1}^{Tr} (\hat{r}_t - r_t) W^{T}_{(out)} \cdot o_t(1-o_t) \cdot h_t^T$$

$$\frac{\partial L}{\partial b_o} = \sum_{t=1}^{Tr} (\hat{r}_t - r_t) W^{T}_{(out)} o_t(1-o_t)$$

L对 h_t 的梯度包含 h_{t+1} 和 o_t 即当 $t = Tr$ 时只有 o_t

故 $t \neq Tr$ 时：$\frac{\partial L_t}{\partial h_t} = \left(\frac{\partial o_t}{\partial h_t}\right)^T \frac{\partial L_t}{\partial o_t} + \left(\frac{\partial h_{t+1}}{\partial h_t}\right)^T \frac{\partial L_t}{\partial h_{t+1}}$

$= o_t(1-o_t) W^T (\hat{r}_t - r_t) \cdot W^{T}_{(out)} + h_{t+1}(1-h_{t+1}) \cdot V^T \cdot \frac{\partial L_t}{\partial h_{t+1}}$

$t = Tr$：上式化为前半部分

$$\therefore \frac{\partial L}{\partial U} = \sum_{t=1}^{Tr} h_t \cdot (1-h_t) \frac{\partial L_t}{\partial h_t} \cdot a_t^{(in)T} \quad \left(\frac{\partial L_t}{\partial h_t} \text{ 又 D 上}\right)$$

$$\frac{\partial L}{\partial V} = \sum_{t=1}^{Tr} h_t (1-h_t) \frac{\partial L_t}{\partial h_t} \cdot h_{t-1}^T$$

$$\frac{\partial L}{\partial b_h} = \sum_{t=1}^{Tr} h_t(1-h_t) \frac{\partial L_t}{\partial h_t}$$

$$\frac{\partial L}{\partial a_t^{(in)}} = \sum_{t=1}^{Tr} h_t(1-h_t) \frac{\partial L_t}{\partial h_t} U^T$$

$$\frac{\partial L}{\partial W^{(in)}} = \frac{\partial L}{\partial a_t^{(in)}} \cdot \frac{\partial a_t^{(in)}}{\partial W^{(in)}}$$

$$= \sum_{t=1}^{Tr} h_t(1-h_t) \frac{\partial L_t}{\partial h_t} U^T \cdot x_t^T$$

$$\frac{\partial L}{\partial b^{(in)}} = \sum_{t=1}^{Tr} h_t(1-h_t) \frac{\partial L_t}{\partial h_t} U^T$$