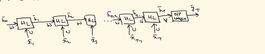
Backpropagation through time (BPTT)

Simple hidden layer RNN

Supervised Pattern Classification
- Training outputs

$x = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_T)$



Training data set:

$$\mathcal{D} = \{ X_n, \bar{t}_n \}_{n=1}^{N}$$

$$X_n = (\bar{x}_{n1}, \bar{x}_{n2}, \ldots, \bar{x}_{nT})$$

$\bar{t}_n$: 1-of-K representation of class label for X_n

Loss fn:

$$\bar{t}_n^T = \frac{1}{K} \sum_{k=1}^{K} (\bar{y}_{nk} - \bar{t}_{nk})^2$$

$$\Delta \bar{y}_{nk} = \eta \, \delta_{nk}^{y} \, h_{nk}$$

$$\Delta w_{jm}$$

1 and 1st are biases in hidden & output of NL

$$\Delta w_{jk}$$

$$\Delta w_{jm}^{h} = \eta \, \delta_{nj}^{h} \, h_{n,m}$$

$$\delta_{nj}^{h} = \left(\sum_{k=1}^{K} v_{kj} \, \delta_{nk}^{y} \right) \cdot \frac{\Delta \phi(a_{nj}^{h})}{\Delta a_{nj}^{h}}$$

$$a_{nj}^{h} = \sum_{k=1}^{I} w_{jk} \, h_{n,k} + \sum_{l=1}^{J} w_{jl} \, x_{nl} + b_{nj}$$

$$\Delta w_{jm}^{hh} = \eta \, \delta_{nj}^{h} \cdot h_{n,k}$$

$$\delta_{nj}^{h} = \left(\sum_{k=1}^{K} v_{kj} \, \delta_{nk}^{y} \right) \frac{\Delta \phi(a_{nj}^{h})}{\Delta a_{nj}^{h}}$$

$$\Delta w_{jm}^{h}$$

$$\Delta w_{km} = \sum_{t=1}^{T} \Delta w_{km}^{h}$$

$$\Delta w_{jm} = \eta \sum_{t=1}^{T} \delta_{nj}^{h} \, h_{n,k}$$

$$\Delta w_{jk} = \eta \sum_{t=1}^{T} \delta_{nj}^{h} \, x_{jk}$$

$$\Delta w_{jk}^{h} = \eta \, \delta_{nj}^{h} \, x_{nk}$$

Minimizing outermost outputs
- Energy conformity

$\bar{h}_0 = \bar{\varepsilon}$ ($\bar{\varepsilon}$: representation of interest)

Training data set

$$\mathcal{D} = \{ \bar{x}_n, \bar{y}_n \}_{n=1}^{N}$$

$\bar{x}_n$: Representation of all things in the training data set

$\bar{y}_n = (\bar{y}_{n1}, \bar{y}_{n2}, \ldots, \bar{y}_{nK})$

$$\bar{E}_n = \frac{1}{K} \sum_{k=1}^{K} (\bar{y}_{nk} - \bar{x}_{nk})^2$$

$$\Delta w_{km}^{y} = \eta \, \delta_{nk}^{y} \, h_{nk}$$

$$\Delta w_{km}^{y} = -\eta \frac{\partial \bar{E}_n}{\partial w_{km}}$$

$$= -\eta \frac{\partial}{\partial w_{km}} \frac{1}{K} \sum_{k=1}^{K} (\bar{y}_{nk} - \bar{x}_{nk})^2$$

$$= \eta \sum_{k=1}^{K} (\bar{y}_{nk} - \bar{x}_{nk}) \cdot \frac{\partial \bar{y}_{nk}}{\partial w_{km}}$$

$$= \eta \sum_{k=1}^{K} \underbrace{(\bar{y}_{nk} - \bar{x}_{nk}) \cdot \frac{\Delta \phi(a_{nk}^{y})}{\Delta a_{nk}^{y}}}_{\delta_{nk}^{y}} \cdot \frac{\partial a_{nk}^{y}}{\partial w_{km}}$$

$$= \eta \sum_{k=1}^{K} \delta_{nk}^{y} \cdot \frac{\partial}{\partial w_{km}} \left[\sum_{j=1}^{J} v_{kj} \, h_{n,j} \right]$$

$$h_{nj} = \sum_{k=1}^{I} w_{jk} \, x_{nk} + b_{nj}$$

Where do w_{km} come from?

$$= \eta \sum_{k=1}^{K} v_{km} \, \delta_{nk}^{y} \cdot \underbrace{\frac{\partial}{\partial w_{km}} (h_{nj})}$$

$$\left\{ \begin{array}{l} \frac{\Delta \phi(a_{nj}^{h})}{\Delta a_{nj}^{h}} \cdot \underbrace{\frac{\partial}{\partial w_{km}} (a_{nj}^{h})} \\ \frac{\partial}{\partial w_{km}} \left[\sum_{j=1}^{J} v_{kj} \, h_{n,j} + \sum_{k=1}^{K} v_{kj} \, x_{nk} + b_{j}^{y} \right] \end{array} \right.$$

$$h_{n,m} + v_{km} \cdot \frac{\partial h_{n,m}}{\partial w_{km}} + \sum_{k=1}^{K} v_{kj} \frac{\partial x_{nk}}{\partial w_{km}}$$

$$= \eta \sum_{k=1}^{K} v_{km} \, \delta_{nk}^{y} \cdot \frac{\partial h_{nj}}{\partial w_{km}}$$

$$\Delta w_{km}^{h} = \eta \sum_{k=1}^{K} v_{km} \, \delta_{nk}^{y} \left[h_{n,m} + v_{km} \cdot \frac{\partial h_{n,m}}{\partial w_{km}} + \sum_{k=1}^{K} v_{km} \frac{\partial h_{n,m}}{\partial w_{km}} \right]$$

$$\Delta w_{km}^{h} = \sum_{t=1}^{T} \Delta w_{km}^{h}$$

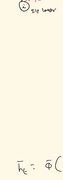
$$\Delta w_{jk}^{h} = -\eta \frac{\partial \bar{E}_n}{\partial w_{jk}}$$

$$= -\eta \frac{\partial}{\partial w_{jk}} \left[\frac{1}{K} \sum_{k=1}^{K} (\bar{y}_{nk} - \bar{x}_{nk})^2 \right]$$

$$= \eta \sum_{k=1}^{K} (\bar{y}_{nk} - \bar{x}_{nk}) \cdot \frac{\partial \bar{y}_{nk}}{\partial w_{jk}}$$

$$\Delta w_{jk}^{h} = \eta \sum_{k=1}^{K} v_{km} \, \delta_{nk}^{y} \cdot \frac{\partial h_{nj}}{\partial w_{jk}}$$

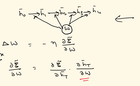
$$\Delta w_{jk}^{h} = \eta \sum_{k=1}^{K} v_{km} \, \delta_{nk}^{y} \cdot \frac{\partial h_{nj}}{\partial w_{jk}} \left[h_{n,k} + w_{jk} \cdot \frac{\partial h_{n,k}}{\partial w_{jk}} + \sum_{k=1}^{K} v_{km} \frac{\partial h_{n,k}}{\partial w_{jk}} \right]$$



Vanishing and Exploding Gradients Problem

Loops & sequences in Nets
- Time & Energy

$T = 4$



$$\Delta w = -\eta \frac{\partial \bar{E}}{\partial w}$$

Training outputs case

$$\frac{\partial \bar{E}}{\partial w} = \frac{\partial \bar{E}}{\partial \bar{h}_T} \cdot \frac{\partial \bar{h}_T}{\partial w}$$

$$\bar{h}_t = \Phi\left(W \bar{h}_{t-1} + U \bar{x}_t + \bar{b}_h \right)$$

$$\frac{\partial \bar{h}_T}{\partial w} = \underbrace{\frac{\partial^+ \bar{h}_T}{\partial w}}_{\text{Explicit dependency}} + \underbrace{\frac{\partial \bar{h}_T}{\partial w} \cdot \frac{\partial \bar{h}_T}{\partial w}}_{\text{implicit dependency}}$$

$$\frac{\partial \bar{h}_T}{\partial w} = \frac{\partial^+ \bar{h}_T}{\partial w} + \frac{\partial \bar{h}_T}{\partial \bar{h}_T} \cdot \frac{\partial \bar{h}_T}{\partial w}$$

$$\frac{\partial \bar{h}_T}{\partial w} = \frac{\partial^+ \bar{h}_T}{\partial w} + \frac{\partial \bar{h}_T}{\partial \bar{h}_T} \cdot \frac{\partial \bar{h}_T}{\partial w}$$

$$\frac{\partial \bar{h}_T}{\partial w} = \frac{\partial^+ \bar{h}_T}{\partial w} + \frac{\partial \bar{h}_T}{\partial \bar{h}_T} \cdot \underbrace{\frac{\partial \bar{h}_T}{\partial \bar{h}_T}}_{=0} = \frac{\partial^+ \bar{h}_T}{\partial w}$$

$$\frac{\partial \bar{h}_T}{\partial w} = \frac{\partial^+ \bar{h}_T}{\partial w} + \frac{\partial \bar{h}_T}{\partial \bar{h}_T} \cdot \frac{\partial \bar{h}_T}{\partial w}$$

$$= \frac{\partial^+ \bar{h}_T}{\partial w} + \frac{\partial \bar{h}_T}{\partial \bar{h}_T} \cdot \frac{\partial^+ \bar{h}_T}{\partial w} + \frac{\partial \bar{h}_T}{\partial \bar{h}_T} \cdot \frac{\partial \bar{h}_T}{\partial \bar{h}_T} \cdot \frac{\partial^+ \bar{h}_T}{\partial w}$$

$$+ \underbrace{\frac{\partial \bar{h}_T}{\partial \bar{h}_T} \cdot \frac{\partial \bar{h}_T}{\partial \bar{h}_T} \cdot \frac{\partial \bar{h}_T}{\partial \bar{h}_T} \cdot \frac{\partial \bar{h}_T}{\partial w}}$$

$$= \frac{\partial \bar{h}_T}{\partial \bar{h}_T} \cdot \frac{\partial \bar{h}_T}{\partial w} + \frac{\partial \bar{h}_T}{\partial \bar{h}_T} \cdot \frac{\partial^+ \bar{h}_T}{\partial w} + \frac{\partial \bar{h}_T}{\partial \bar{h}_T} \cdot \frac{\partial \bar{h}_T}{\partial \bar{h}_T} \cdot \frac{\partial^+ \bar{h}_T}{\partial w}$$

$$= \sum_{t=1}^{4} \frac{\partial \bar{h}_T}{\partial \bar{h}_t} \cdot \frac{\partial^+ \bar{h}_t}{\partial w}$$

$$\frac{\partial \bar{E}_T}{\partial w} = \sum_{t=1}^{T} \underbrace{\frac{\partial \bar{E}_T}{\partial \bar{h}_t} \cdot \frac{\partial^+ \bar{h}_t}{\partial w}}$$

$$\frac{\partial \bar{h}_T}{\partial \bar{h}_k} = \prod_{t=k+1}^{T} \frac{\partial \bar{h}_t}{\partial \bar{h}_{t-1}}$$

$$\frac{\partial \bar{h}_t}{\partial \bar{h}_{t-1}} = \frac{\Delta \phi(a_t^h)}{\Delta a_t^h} \cdot \frac{\partial \bar{a}_t^h}{\partial \bar{h}_{t-1}}$$

$$\boxed{\begin{array}{l} \bar{h}_t = \Phi\left(W \bar{h}_{t-1} + U \bar{x}_t + \bar{b}_h \right) \\ = \Phi(\bar{a}_t^h) \end{array}}$$

$$\frac{\partial \bar{h}_t}{\partial \bar{h}_{t-1}} = \frac{\Delta \phi(\bar{a}_t^h)}{\Delta \bar{a}_t^h} \cdot W$$

$$\frac{\Delta \phi(\bar{a}_t^h)}{\Delta \bar{a}_t^h} = \left[\frac{\partial \phi(\bar{a}_i^h)}{\partial a_j^h} \right]_{J \times J} \qquad i, j = 1, 2, \ldots, J$$

$$= \begin{bmatrix} \phi'(\bar{a}_1^h) & 0 & \cdots & 0 \\ 0 & \phi'(\bar{a}_2^h) & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \phi'(\bar{a}_J^h) \end{bmatrix}$$

$$= \operatorname{diag}\left(\phi'(\bar{a}_t^h) \right)$$

$$\frac{\partial \bar{h}_t}{\partial \bar{h}_{t-1}} = \frac{\Delta \phi(\bar{a}_t^h)}{\Delta \bar{a}_t^h} \cdot \frac{\partial \bar{a}_t^h}{\partial \bar{h}_{t-1}}$$

$$\frac{\partial \bar{h}_t}{\partial \bar{h}_{t-1}} = \operatorname{diag}\left(\phi'(\bar{a}_t^h) \right) \cdot W$$

Frobenius norm as a matrix:
- Square root of sum of squares of elements of a matrix

$$\left\| \frac{\partial \bar{h}_t}{\partial \bar{h}_{t-1}} \right\|_F = \left\| \operatorname{diag}\left(\phi'(\bar{a}_t^h) \right) \cdot W \right\|_F$$

Cauchy-Schwarz inequality

$$|\bar{a}^T \bar{b}| \le \|\bar{a}\| \, \|\bar{b}\|$$

$$\|a \cdot b\| \le \|a\| \, \|b\|$$

$$\left\| \frac{\partial \bar{h}_t}{\partial \bar{h}_{t-1}} \right\|_F \le \left\| \operatorname{diag}\left(\phi'(\bar{a}_t^h) \right) \right\|_F \left\| W \right\|_F$$

$$\|W\|_F \le \lambda$$

$$\left\| \operatorname{diag}\left(\phi'(\bar{a}_t^h) \right) \right\|_F$$

$$\frac{\Delta \phi(a_j^h)}{\Delta a_j^h} \le \frac{\beta}{4} \text{ , logistic sigmoid} \qquad \frac{\Delta \phi(a_j)}{\Delta a_j} = \beta \cdot \phi(a_j)(1 - \phi(a_j))$$

$$\le \beta \text{ , tanh} \qquad \le \frac{\beta}{4}$$

$$\le \beta \text{ , tanh sigmoid}$$

$$\frac{\Delta \phi(a_j^h)}{\Delta a_j^h} \le \beta \cdot \gamma \qquad \frac{\Delta \phi(a_j)}{\Delta a_j} = \beta \cdot \underbrace{\phi(a_j)(1 - \phi(a_j))}_{\le \beta}$$

$$\gamma = \frac{1}{4} \text{ , for logistic sigmoid}$$

$$= 1 \text{ , for tanh}$$

$$\sqrt{J \beta \gamma \lambda} = \sqrt{J} \beta \gamma \lambda$$

$$\left\| \operatorname{diag}\left(\phi'(\bar{a}_t^h) \right) \right\|_F \le \sqrt{J} \beta \cdot \gamma$$

$$\left\| \frac{\partial \bar{h}_t}{\partial \bar{h}_{t-1}} \right\|_F \le \sqrt{J} \beta \cdot \gamma \cdot \lambda$$

$$\left\| \frac{\partial \bar{h}_T}{\partial \bar{h}_k} \right\|_F \le \prod_{m=k+1}^{T} \underbrace{\left\| \frac{\partial \bar{h}_m}{\partial \bar{h}_{m-1}} \right\|_F}_{T-k}$$

$$\le \underbrace{\left(\sqrt{J} \beta \cdot \gamma \cdot \lambda \right)^{T-k}}$$

$$\sqrt{J} \beta \cdot \gamma \cdot \lambda < 1.0 \qquad \left\| \frac{\partial \bar{h}_T}{\partial \bar{h}_k} \right\|_F \text{ vanishes}$$

$$\sqrt{J} \beta \cdot \gamma \cdot \lambda > 1.0 \qquad \left\| \frac{\partial \bar{h}_T}{\partial \bar{h}_k} \right\|_F \text{ explodes}$$