CS224n: Deep learning for NLP

Assignment 2.

2) a)

stack	buffer	new dependency	transition
[Root]	[I, parsed, this, sentense, correctly]		Initial Config.
[Root, I]	[Parsed, this, sentence, correctly]		SHIFT
[Root, I, Parsed]	[this, sentence, correctly]		SHIFT
[Root, Parsed]	[this, sentence, correctly]	Parsed → I	Left-Arc
[Root, Parsed, this]	[sentence, correctly]		SHIFT
[Root, Parsed, this, sentence]	[correctly]		SHIFT
[Root, Parsed, sentence]	[correctly]	Sentence → This	Left_Arc
[Root, Parsed]	[correctly]	Parsed → Sentence	Right_Arc
[Root, Parsed, Correctly]	~~∅~~ []		Shift
[Root, Parsed]	~~∅~~ []	Parsed → Correctly	Right-Arc
[Root]	[]	Root → Parsed	Right_Arc

a)(i) ~~Keeping~~ the momentum accumulated throughout the descent of the gradient can help us continue update our parameters in case
- the "real" gradient becomes null
- When we get stuck at a local minimum for instance
- It could help against vanishing gradient

g) (ii) dividind by $\sqrt{V}$ has the effect of increasing the final gradient of previously low values :
ie: very low values of gradient now have slightly bigger values :
$\Rightarrow$ this may help combat vanishing gradient

3) $P(x^{(t+1)} = v_j \mid x^{(t)} \dots x^{(1)}) = \hat{y}_j(t)$

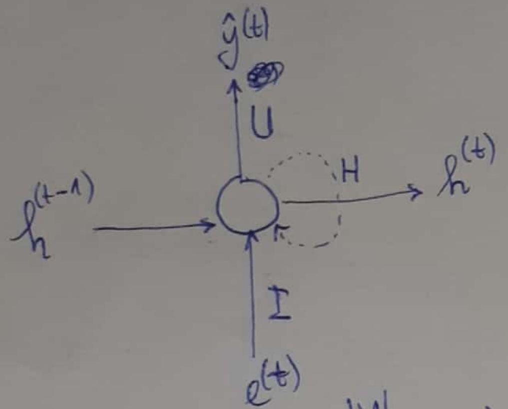
$e^{(t)} = x^{(t)} \cdot L$

$h^{(t)} = \text{sigmoid}(h^{(t-1)} H + e^{(t)} \cdot I + b_1) = \text{sigmoid}(\theta^{(t)})$

$\hat{y}(t) = \text{softmax}(h^{(t)} U + b_2) = \text{softmax}(z^{(t)})$.

$h^{(0)} = h_0 \in \mathbb{R}^{D_h}$; $L \in \mathbb{R}^{|V| \times d}$; $H \in \mathbb{R}^{D_h \times D_h}$; $I \in \mathbb{R}^{d \times D_h}$; $U \in \mathbb{R}^{D_h \times |V|}$;

$b_2 \in \mathbb{R}^{|V|}$



$$J^{(t)}(\theta) = CE(y^{(t)}, \hat{y}^{(t)}) = -\sum_{j=1}^{|V|} y_j^{(t)} \cdot \log(\hat{y}_j^{(t)})$$

$$J_{seq}(\theta) = \frac{1}{T} \sum_{t=1}^{T} J^{(t)}(\theta)$$

(2) $$PP^{(t)}(y^{(t)}, \hat{y}^{(t)}) = \frac{1}{\bar{P}(x_{pred}^{(t+1)} = x^{(t+1)} \mid x^{(t)} \dots x^{(1)})} = \frac{1}{\sum_{j=1}^{|V|} y_j^{(t)} \cdot \hat{y}_j^{(t)}}$$

(perplexity)

$$J_{seq} = -\frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{|V|} y_j^{(t)} \log(\hat{y}_j(t))$$

*) $J^{(t)}(\theta) = -\sum_{j=1}^{|V|} y_j^{(t)} \log(\hat{y}_j^{(t)}) = -\log(\hat{y}_k^{(t)})$ | where k is the $y_k = 1$

$= \log\left(\frac{1}{\hat{y}_k^{(t)}}\right) = \log\left(\frac{1}{\sum_{j=1}^{|V|} y_j^{(t)} \hat{y}_j^{(t)}}\right) = \log(PP^{(t)})$

$$\boxed{J^{(t)} = \log(PP^{(t)}) \iff PP^{(t)} = 2^{J^{(t)}}}$$

(a) * let show that:

$\boxed{T = |V|}$ minimizing $J(\theta) = \frac{1}{T}\sum_{t=1}^{T} J^{(t)} \Leftrightarrow$ minimizing $PP(\theta) = \left[\prod_{t=1}^{T}(pp^{(t)})\right]^{\frac{1}{T}}$

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T} J^{(t)}_{(\theta)} = \frac{1}{T}\sum_{t=1}^{T} \log(pp^{(t)}_{(\theta)})$$ $\log(a \cdot b) = \log(a) + \log(b)$

$$= \frac{1}{T}\log\left[\prod_{t=1}^{T} pp^{(t)}_{(\theta)}\right]$$ $\log a^b = b \log a$

$$= \log\left[\left(\prod_{t=1}^{T} pp^{(t)}_{(\theta)}\right)^{\frac{1}{T}}\right]$$

$$\Rightarrow \underset{\theta}{\operatorname{argmin}}\, J(\theta) = \underset{\theta}{\operatorname{argmin}} \log\left[\left(\prod_{t=1}^{T} pp^{(t)}_{(\theta)}\right)^{\frac{1}{T}}\right]$$ because log is strictly ↗

$$= \underset{\theta}{\operatorname{argmin}}\left[\left(\prod_{t=1}^{T} pp^{(t)}_{(\theta)}\right)^{\frac{1}{T}}\right] \blacksquare$$

* $pp^{(t)}_{(\theta)} = \frac{1}{\sum_{j=1}^{|V|} y_j^{(t)} \hat{y}_j^{(t)}}$; if predictions are random from uniform distribution

$\rightarrow \forall j \quad \hat{y}_j^{(t)} = \frac{1}{|V|}$

$$\Rightarrow pp^{(t)} = \frac{1}{\sum_{j=1}^{|V|} y_j^{(t)} \cdot \frac{1}{|V|}} = \frac{1}{y_k^{(t)} \frac{1}{|V|}}$$ where k is the correct class

$$\boxed{pp^{(t)} = \frac{1}{\frac{1}{|V|}} = |V|}$$

in the case $|V| = 10000$ $\quad J^{(t)} = \log(pp^{(t)}) = \log(10000)$

$$\boxed{J^{(t)} \cong 9.210}$$

Scanned by CamScanner

3) b) let $z^{(t)} = h^{(t)} U + b_2$

$\theta^{(t)} = h^{(t-1)} H + e^{(t)} I + b_1$

$$\frac{\partial J^{(t)}}{\partial z^{(t)}} = \hat{y}^{(t)} - y^{(t)} = \delta_1^{(t)}$$

$$\frac{\partial J^{(t)}}{\partial \theta^{(t)}} = \frac{\partial J^{(t)}}{\partial z^{(t)}} \cdot \frac{\partial z^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial \theta^{(t)}} = \delta_1^{(t)} \cdot U^T \odot \sigma(\theta^{(t)}) \odot (1 - \sigma(\theta^{(t)})) = \delta_2^{(t)}$$

$$\frac{\partial J^{(t)}}{\partial b_2} = \frac{\partial J^{(t)}}{\partial z^{(t)}} \cdot \frac{\partial z^{(t)}}{\partial b_2} = \delta_1^{(t)} \frac{\partial}{\partial b_2}(h^{(t)} \cdot U + b_2) = \delta_1$$

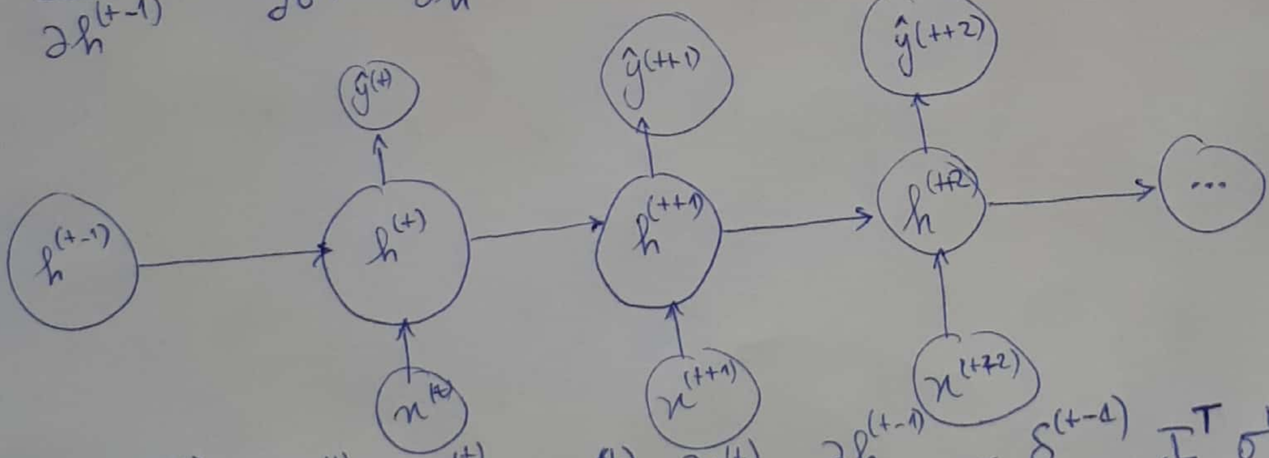
$$\frac{\partial J^{(t)}}{\partial L_{x^{(t)}}} = \frac{\partial J^{(t)}}{\partial e^{(t)}} = \frac{\partial J^{(t)}}{\partial \theta^{(t)}} \cdot \frac{\partial \theta^{(t)}}{\partial e^{(t)}} = \delta_2^{(t)} \cdot I^T$$

$$\left.\frac{\partial J^{(t)}}{\partial I}\right|_{(t)} = \frac{\partial J^{(t)}}{\partial \theta^{(t)}} \cdot \left.\frac{\partial \theta^{(t)}}{\partial I}\right|_{(t)} = \delta_2^{(t)} \cdot (e^{(t)})^T$$

$$\left.\frac{\partial J^{(t)}}{\partial H}\right|_{(t)} = \frac{\partial J^{(t)}}{\partial \theta^{(t)}} \cdot \left.\frac{\partial \theta^{(t)}}{\partial H}\right|_{(t)} = \delta_2^{(t)} \cdot (h^{(t-1)})^T$$

$$\frac{\partial J^{(t)}}{\partial h^{(t-1)}} = \frac{\partial J^{(t)}}{\partial \theta^{(t)}} \cdot \frac{\partial \theta^{(t)}}{\partial h^{(t-1)}} = \delta_2^{(t)} \cdot H^T = \delta^{(t-1)}$$

c)



$$\frac{\partial J^{(t)}}{\partial L_{x^{(t-1)}}} = \frac{\partial J^{(t)}}{\partial e^{(t-1)}} = \frac{\partial J^{(t)}}{\partial \theta^{(t)}} \cdot \frac{\partial \theta^{(t)}}{\partial e^{(t-1)}} = \delta_2^{(t)} \cdot \frac{\partial \theta^{(t)}}{\partial h^{(t-1)}} \cdot \frac{\partial h^{(t-1)}}{\partial e^{(t-1)}} = \delta^{(t-1)} \cdot I^T \cdot \sigma'(\theta^{(t-}$$

$$\left.\frac{\partial J^{(t)}}{\partial I}\right|_{t-1} = \sum_{k=t-1}^{t} \frac{\partial J^{(t)}}{\partial h^{(k)}} \cdot \frac{\partial h^{(k)}}{\partial I} = \underbrace{\frac{\partial J^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial I}}_{\frac{\partial J^{(t)}}{\partial I}} + \underbrace{\frac{\partial J^{(t)}}{\partial h^{(t-1)}} \cdot \frac{\partial h^{(t-1)}}{\partial I}}_{\delta^{(t-1)} \odot \sigma'(\theta^{(t-1)}) e^{(t-1}}$$

$$\Rightarrow \boxed{\left.\frac{\partial J^{(t)}}{\partial I}\right|_{t-1} = \delta_2^{(t)} (e^{(t)})^T + \delta^{(t-1)} \odot \sigma'(\theta^{(t-1)}) e^{(t-1)T}}$$

Scanned by CamScanner

$$\left.\frac{\partial J^{(t)}}{\partial H}\right|_{(t-1)} = \sum_{k=t-1}^{t} \frac{\partial J^{(t)}}{\partial h^{(k)}} \cdot \frac{\partial h^{(k)}}{\partial H}$$

$$= \frac{\partial J^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial H} + \frac{\partial J^{(t)}}{\partial h^{(t-1)}} \cdot \frac{\partial h^{(t-1)}}{\partial H}$$

$$= \frac{\partial J^{(t)}}{\partial H} + \delta^{(t-1)} \odot \sigma'(\theta^{(t-1)}) \cdot (h^{(t-2)})^T$$

$$\boxed{\left.\frac{\partial J^{(t)}}{\partial H}\right|_{(t-1)} = \delta_2^{(t)} (h^{(t-1)})^T + \delta^{(t-1)} \odot \sigma'(\theta^{(t-1)}) \cdot (h^{(t-2)})^T}$$

d) ~~$\#\text{operations}(\delta^{(t)} | h^{(t-1)}) = O(op(\hat{y}^{(t)})) = O(op(h^{(t)}) \times |V| \times D_h^2 + |V|)$~~

~~$\#op(h^{(t)}) = O(\#op(h^{(t-1)}) \times D_h^3 \times \quad +$~~