CS 224n
Assignment #2

Andrew Hinh
9/3/24

1 a) $J_{naive\text{-}softmax}(v_c, o, U) = -\log P(O=o \mid C=c)$

$= -\sum_{w \in vocab} y_w \log(\hat{y}_w) = -\log(\hat{y}_o)$

b) $\hat{y} = P(O \mid C=c)$, so eqn 1 = eqn 3 ■

b) i) $J_{naive\text{-}softmax}(v_c, o, U) = -\log \frac{\exp(u_o^T v_c)}{\sum_{w \in vocab} \exp(u_w^T v_c)}$

$\frac{dJ}{dv_c} = -\frac{1}{\hat{y}_o} \cdot \frac{\partial \hat{y}_o}{\partial v_c}$,

$\frac{\partial \hat{y}_o}{\partial v_c} = \frac{\partial}{\partial v_c}\left(\frac{\exp(u_o^T v_c)}{\sum_{w \in vocab} \exp(u_w^T v_c)}\right)$

$= \left(\sum_{w \in vocab} \exp(u_w^T v_c) \cdot \exp(u_o^T v_c) \cdot u_o\right)$

$- \left(\exp(u_o^T v_c) \cdot \sum_{w \in vocab} \exp(u_w^T v_c) \cdot u_w\right)$

$\left(\sum_{w \in vocab} \exp(u_w^T v_c)\right)^2$

$= \frac{\exp(u_o^T v_c)}{\sum_{w \in vocab} \exp(u_w^T v_c)} (u_o - u_w)$ = softmax

$= \hat{y}_o u_o - \sum_{w \in vocab} \hat{y}_w u_w$

$\frac{dJ}{dv_c} = -\frac{1}{\hat{y}_o}\left(\hat{y}_o u_o - \sum_{w \in vocab} \hat{y}_w u_w\right)$

$= -u_o + \sum_{w \in vocab} \hat{y}_w u_w$

$= U(\hat{y} - y)$ ■

ii) $\frac{dJ}{dv_c} = 0$ when $\boxed{y = \hat{y}}$ ✓

iii) when $\frac{dJ}{dv_c}$ subtracted from word vec v_c)

it makes v_c closer to u_o (aka the correct class) ✓

c) When magnitude important, info lost vs when dir. " not lost, ✓

d) $W = 0$: $\frac{dJ}{du_o} = -\frac{1}{\hat{y}_o} \cdot \frac{d\hat{y}_o}{du_o}$

$$\frac{2\hat{y}_o}{2u_o} = \frac{2}{2u_o}\left(\frac{\exp(u_o^T v_c)}{\sum_{w \in Vocab}(u_w^T v_c)}\right)$$

$$= \frac{2}{2u_o}\left(\frac{\exp(u_o^T v_c)}{u_o^T v_c}\right)$$

$$= (u_o^T v_c) \cdot \exp(u_o^T v_c) \cdot v_c - \exp(u_o^T v_c) \cdot v_c$$

$$\frac{(u_o^T v_c)^2}{}$$

$$= \frac{\exp(u_o^T v_c) v_c ((u_o^T v_c) - 1)}{(u_o^T v_c)^2}$$

$$= \hat{y}_o v_c \cdot ((u_o^T v_c) - 1),$$

$$\frac{dJ}{du_o} = \boxed{-v_c(\hat{y} - 1)},$$

$w \neq o$: $\frac{\partial J_p}{\partial u_w} = \frac{\partial}{\partial u_w}\left(\frac{\exp(u_o^T v_c)}{\sum_{w \in vocab} \exp(u_w^T v_c)}\right)$

$= \frac{\exp(u_o^T v_c) \cdot 0 \cdot \sum_{w \in vocab} - \exp(u_o^T v_c) \cdot \exp(u_w^T v_c) \cdot v_c}{\left(\sum_{w \in vocab} \exp(u_w^T v_c)\right)^2}$

$= -\hat{y} \cdot v_c$

$\frac{\partial J}{\partial u_w} = \frac{1}{y_o} \cdot -\hat{y} \cdot v_c$

$= \boxed{y_w \cdot v_c}$ ✱

e) $\frac{dJ}{dU} = \left[\frac{\partial J(v_c, o, U)}{\partial u_1}\right]\left[\cdots\right]\left[\frac{\partial J(v_c, o, U)}{\partial u_{|vocab|}}\right]$

$= \left[y_1 \cdot v_c \quad \right] \cdots \left[y_w \cdot v_c\right]$ ✱

2. a) i) Updates can only affect gradient by small amount → less var → less noise when training ✱

ii) adaptive lr → bigger updates when grad mag ↓, smaller lr ↑, move faster in right dir & slower in wrong dir. ✱

b) i) $E_{p_{drop}}[\gamma_d \circ h]_i = h_i$

$= \gamma E[d_i] \cdot h_i = h_i$

$\Rightarrow \gamma = \frac{1}{E[d_i]} = \frac{1}{0 \cdot p_{drop} + 1 \cdot (1 - p_{drop})} = \boxed{\frac{1}{1 - p_{drop}}}$ ✱

ii) Only during training b/c thrown away prof ✱

3) a.) (3 given), n = 8 words

Stack	Buffer	New dep.	Trans.
[ROOT, presented, my]	[findings, at, the, NLP, conference]		SHIFT
[ROOT, presented, my, findings]	[at, the, NLP, conf.]		SHIFT
[ROOT, pres, findings]	"	findings → my	LEFT-ARC
[ROOT, pres.]	"	pres. → findings	RIGHT-ARC
[ROOT, pres., at]	[the, NLP, conf.]		SHIFT
[ROOT, pres., at, the]	[NLP, conf.]		SHIFT
[ROOT, pres., at, the, NLP]	[conf.]		SHIFT
[ROOT, pres., at, the, NLP, conf.]			SHIFT
[ROOT, pres., the, NLP, conf.]		conf. → at	LEFT
[ROOT, pres., NLP, conf.]		conf. → the	LEFT
[ROOT, pres., conf.]		conf. → NLP	LEFT
[ROOT, pres.]		pres. → conf.	RIGHT
[ROOT]		pres. → ROOT	RIGHT

b) $n = 8 \rightarrow 2n$ steps $= 8(2) = 16$ steps bc n steps to → stack, n steps to remove.

c) Code

d) code $\overbrace{(3\times2)\ (2\times4)\ (4)}^{3\times4}$ 4×2

e) i) $\frac{\partial h_i}{\partial x_j} = (xW + b_1) > 0 \times \overbrace{W.T}$

ii) $\frac{\partial CE(y,\hat{y})}{\partial l_i} = \sum_j \frac{\partial CE(y,\hat{y})}{\partial \hat{y}_j} \cdot \frac{\partial \hat{y}_j}{\partial l_i} \quad (j = c)$

$= \frac{\partial CE(y,\hat{y})}{\partial \hat{y}_c} \cdot \frac{\partial \hat{y}_c}{\partial l_i}$

$= \frac{y_j}{\hat{y}_c} \cdot \frac{\partial}{\partial l_i} \frac{e^{l_i}}{\sum_{j=1}^{3} e^{l_j}}$

$= \frac{y^i}{\hat{y}_c} \cdot e^{l_i - l_j} \cdot 1 = \boxed{\frac{e^{l_i - l_j}(y_i)}{\hat{y}_c}}$

iii) code

f) i) error: :: verb error

incorr. dep.: sitting → ,

corr. dep.: sitting → blocked

ii) " mod error

elements → most

crucial → most

iii) " prep. phrase attach error

decided → decision

reasons → decision

iv) " coord error

one → and

Quebec → and

g) POS tags can give signal as they often follow certain words