A2 (Extra Credit)

● Graded

Student

Sugi Venugeethan

View or edit group

Total Points

5 / 5 pts

Question 1

Extra Credit Challenge 1

2.5 / 2.5 pts

✔ **+ 0.5 pts** **[PART 2]**

$$= -\boldsymbol{u}_o + \frac{1}{\sum_{w' \in \text{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)} \times \sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c) \boldsymbol{u}_w$$

✔ **+ 0.5 pts** **[PART 1]** - Let $\hat{\boldsymbol{y}}$ be the column vector of the softmax prediction of words, and $\boldsymbol{y}$ be the one-hot label which is also a column vector. Then:

$$J_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -\log(\hat{y}_o)$$
$$= -\log\left(\frac{\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}\right)$$
$$= -\left(\log\left(\exp(\boldsymbol{u}_o^\top \boldsymbol{v}_c)\right) - \log\left(\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)\right)\right)$$
$$= -\boldsymbol{u}_o^\top \boldsymbol{v}_c + \log\left(\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)\right)$$

Therefore:

$$\frac{\partial J}{\partial \boldsymbol{v}_c} = -\boldsymbol{u}_o + \frac{\partial}{\partial \boldsymbol{v}_c}\left(\log\left(\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)\right)\right)$$
$$= -\boldsymbol{u}_o + \frac{1}{\sum_{w' \in \text{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)} \times \frac{\partial}{\partial \boldsymbol{v}_c}\left(\sum_{w \in \text{Vocab}} \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)\right)(\text{chain rule})$$

✔ **+ 1 pt** **[PART 3]**

$$= -\boldsymbol{u}_o + \sum_{w \in \text{Vocab}} \left(\frac{\exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)}{\sum_{w' \in \text{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)}\right) \boldsymbol{u}_w (\text{rearrange})$$
$$= -\boldsymbol{u}_o + \sum_{w \in \text{Vocab}} P(O = w \mid C = c) \boldsymbol{u}_w$$

✔ **+ 0.5 pts** **[PART 4]**

$$= -\boldsymbol{u}_o + \sum_{w \in \text{Vocab}} \hat{\boldsymbol{y}}_w \boldsymbol{u}_w$$

Given that $\boldsymbol{y}$ is a 1-hot vector with a 1 at word o, this can be rewritten as:

$$\frac{\partial J}{\partial \boldsymbol{v}_c} = \boldsymbol{U}(\hat{\boldsymbol{y}} - \boldsymbol{y})$$

or equivalently,

$$\frac{\partial J}{\partial \boldsymbol{v}_c} = -\boldsymbol{u}_o + \sum_{w=1}^{V} \hat{y}_w \boldsymbol{u}_w$$

Question 2

Extra Credit Challenge 2

2.5 / 2.5 pts

✔ **+ 0.5 pts** **[PART 1]** - We have:

$$J_{\text{naive-softmax}}(\boldsymbol{v}_c, o, \boldsymbol{U}) = -\boldsymbol{u}_o^\top \boldsymbol{v}_c + \log\left(\sum_{w' \in \text{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)\right)$$

For the first case ($w = o$) we find:

$$\frac{\partial J}{\partial \boldsymbol{u}_w} = -\boldsymbol{v}_c + \frac{\partial}{\partial \boldsymbol{u}_w}\left(\log\left(\sum_{w' \in \text{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)\right)\right)$$
$$= -\boldsymbol{v}_c + \frac{1}{\sum_{w' \in \text{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)} \times \frac{\partial}{\partial \boldsymbol{u}_w}\left(\sum_{w' \in \text{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)\right)(\text{chain rule})$$
$$= -\boldsymbol{v}_c + \frac{1}{\sum_{w' \in \text{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)} \times \exp(\boldsymbol{u}_w^\top \boldsymbol{v}_c)\boldsymbol{v}_c$$

✔ **+ 1.5 pts** **[PART 2]**

$$= -\boldsymbol{v}_c + \hat{y}_w \boldsymbol{v}_c(\text{by definition of } \hat{y}_w)$$
$$= (\hat{y}_w - 1)\boldsymbol{v}_c$$

✔ **+ 0.5 pts** **[PART 3]** - For the second case ($w \neq o$) we find:

$$\frac{\partial J}{\partial \boldsymbol{u}_w} = \frac{\partial}{\partial \boldsymbol{u}_w}\left(\log\left(\sum_{w' \in \text{Vocab}} \exp(\boldsymbol{u}_{w'}^\top \boldsymbol{v}_c)\right)\right)$$
$$= \hat{y}_w \boldsymbol{v}_c(\text{same derivation as before})$$

Given that $\boldsymbol{y}$ is a 1-hot vector with a 1 at word o, this can be rewritten as:

$$\frac{\partial J}{\partial \boldsymbol{U}} = \boldsymbol{v}_c(\hat{\boldsymbol{y}} - \boldsymbol{y})^\top$$

or equivalently:

$$\frac{\partial J}{\partial \boldsymbol{u}_w} =$$

1. $(\hat{y}_w - 1)\boldsymbol{v}_c \text{if } w = o$
2. $\hat{y}_w \boldsymbol{v}_c \text{otherwise}$

Question assigned to the following page: 1

This handout includes space for every question that requires a written response. Please feel free to use it to handwrite your solutions (legibly, please). If you choose to typeset your solutions, the `README.md` for this assignment includes instructions to regenerate this handout with your typeset LaTeX solutions.

1.a $J_{naive\text{-}softmax}(v_c, o, U) = -\log(\hat{y}_o)$

$$= -\log\left(\frac{\exp(u_o^T v_c)}{\sum_{w \in Vocab} \exp(u_w^T v_c)}\right)$$

$$= -\left(\log(\exp(u_o^T v_c)) - \log\left(\sum_{w \in Vocab} \exp(u_w^T v_c)\right)\right)$$

$$= \underbrace{-u_o^T v_c}_{①} + \underbrace{\log\left(\sum_{w \in vocab} \exp(u_w^T v_c)\right)}_{②}$$

Solution:

$$\frac{\partial}{\partial v_c}\left(-u_o^T v_c\right) \quad ①$$

$$= -\frac{\partial}{\partial v_c}\left(u_o^T v_c\right)$$

$$= -u_o$$

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

$$= \frac{1}{\sum_{w \in vocab} \exp(u_w^T v_c)} \cdot \underbrace{\frac{\partial}{\partial v_c} \sum_{w \in vocab} \exp(u_w^T v_c)}_{③}$$

Question assigned to the following page: 1

This handout includes space for every question that requires a written response. Please feel free to use it to handwrite your solutions (legibly, please). If you choose to typeset your solutions, the README.md for this assignment includes instructions to regenerate this handout with your typeset LaTeX solutions.

1.a Solving (3)

$$\sum_{w \in vocab} \frac{\partial}{\partial v_c} \exp(u_w^T v_c)$$

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

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

Substituting (3) derivative above to (2), we get

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

Bringing back derivative of (1), we get

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

expectation

$$\therefore = -u_o + \sum_{w=1}^{V} \hat{y}_w u_w$$

Question assigned to the following page: 2

1.b

$$J_{naive\text{-}softmax}(v_c, o, U) = -u_o^T v_c + \log\left(\sum_{w' \in vocab} \exp(u_{w'}^T v_c)\right)$$

① $$\frac{\partial}{\partial u_w}\left(-u_o^T v_c\right) = -\frac{\partial u_o^T}{\partial u_w} v_c$$

$$= -y_w v_c$$

② $$\frac{\partial}{\partial u_w} \log\left(\sum_{w' \in vocab} \exp(u_{w'}^T v_c)\right)$$

$$= \frac{1}{\sum_{w' \in vocab} \exp(u_{w'}^T v_c)} \cdot \underbrace{\frac{\partial}{\partial u_w}\left(\sum_{w' \in vocab} \exp(u_{w'}^T v_c)\right)}_{③}$$

Solving ③

$$= \sum_{w' \in vocab} \exp(u_{w'}^T v_c) \frac{\partial}{\partial u_w} \exp(u_w^T v_c)$$

$$= \sum_{w' \in vocab} \exp(u_w^T v_c)\, v_c$$

Substituting above ③ derivative to ②, we get

$$= \boxed{\frac{\sum_{w' \in vocab} \exp(u_w^T v_c)\, v_c}{\sum_{w' \in vocab} \exp(u_w^T v_c)}} \Rightarrow \hat{y}_w v_c$$

Question assigned to the following page: 2

1.b

Bringing back derivative from (1)

we get

$$-y_w v_c + \hat{y}_w v_c$$

$$/// \ell r \; (\hat{y}_w - y_w)^T v_c$$

where $y_w = \begin{cases} 1 & \text{if } w = 0 \\ 0 & \text{otherwise} \end{cases}$