

Assignment #2

Peiyu Liu

1 [15 points] Understanding word2vec.

(a) According to notations defined in writeup:

$$\mathbf{J}_{\text{naive-softmax}}(\mathbf{v}_c, o, \mathbf{U}) = -\log P(O = o|C = c) = -\log \hat{y}_o = -\sum_{w \in \text{Vocab}} y_w \log \hat{y}_w.$$

Note that

1. scalar $\hat{y}_o = P(O = o|C = c) = \frac{\exp(\mathbf{u}_o^T \mathbf{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c)}$, and
2. \mathbf{y} is a one-hot vector where only $y_o = 1$. □

(b) Using chain rule, we obtain

$$\begin{aligned} \frac{\partial \mathbf{J}}{\partial \mathbf{v}_c} &= -\frac{\partial}{\partial \mathbf{v}_c} \log P(O = o|C = c) \\ &= -\frac{1}{P(O = o|C = c)} \cdot \frac{\partial P(O = o|C = c)}{\partial \mathbf{v}_c} \\ &= -\frac{\sum_{w \in \text{Vocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c)}{\exp(\mathbf{u}_o^T \mathbf{v}_c)} \\ &\quad \cdot \frac{\exp(\mathbf{u}_o^T \mathbf{v}_c) [\mathbf{u}_o \sum_{w \in \text{Vocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c) - \sum_{w \in \text{Vocab}} \mathbf{u}_w \exp(\mathbf{u}_w^T \mathbf{v}_c)]}{[\sum_{w \in \text{Vocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c)]^2} \\ &= \frac{\sum_{w \in \text{Vocab}} \mathbf{u}_w \exp(\mathbf{u}_w^T \mathbf{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c)} - \mathbf{u}_o. \end{aligned}$$

Since $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{|\text{Vocab}|}]$, we have

$$\frac{\sum_{w \in \text{Vocab}} \mathbf{u}_w \exp(\mathbf{u}_w^T \mathbf{v}_c)}{\sum_{w \in \text{Vocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c)} = \mathbf{U} \hat{\mathbf{y}},$$

and

$$\mathbf{u}_o = \mathbf{U} \mathbf{y}$$

Therefore, the derivative is $\mathbf{U}(\hat{\mathbf{y}} - \mathbf{y})$, where $\mathbf{U} \in \mathbb{R}^{d \times |\text{Vocab}|}$, and $\hat{\mathbf{y}}, \mathbf{y} \in \mathbb{R}^{|\text{Vocab}|}$.

The derivative equals to zero when

1. $\hat{\mathbf{y}} = \mathbf{y}$ (impossible due to positive nature of softmax), or
2. $\hat{\mathbf{y}} - \mathbf{y} \in \ker(\mathbf{U})$ (possible since d is usually much less than $|\text{Vocab}|$).

Assume vector \mathbf{v}_c is randomly initialized and \mathbf{U} remains fixed, then gradient descent updates \mathbf{v}_c by pulling it towards the true context vector $\mathbf{U} \mathbf{y}$ and repelling it from the (false) predicted distribution $\mathbf{U} \hat{\mathbf{y}}$.