Practical No. 2

Jan Dziuba

March 19, 2023

1 Part 1

a)

$$-\sum_{w\in\text{Vocab}} \mathbf{y}_w \log(\hat{\mathbf{y}}_w) = -((\sum_{w\in\text{Vocab},w\neq o} \mathbf{y}_w \log(\hat{\mathbf{y}}_w)) + \mathbf{y_o} \log(\hat{\mathbf{y}}_o))$$

$$= -(\sum_{w\in\text{Vocab},w\ neqo} 0 \cdot \log(\hat{y}_w)) - 1 \cdot \log(\hat{y}_o) = -\log(\hat{y}_o)$$

b)

$$\frac{\partial}{\partial \mathbf{v}_c} J_{\text{naive-softmax}(\mathbf{v}_c,o,\mathbf{U})} = -\frac{\partial}{\partial \mathbf{v}_c} \log \frac{\exp(\mathbf{u}_o^T \mathbf{v}_c)}{\sum_{w\in\text{Vocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c)}$$

$$= -\frac{\partial}{\partial \mathbf{v}_c}(\mathbf{u}_o^T \mathbf{v}_c - \log\left(\sum_{w\in\text{Vocab}} \exp(\mathbf{u}_w^T \mathbf{v}_c)\right))$$

$$= -(\mathbf{u}_o - \frac{\sum_{w_1\in\text{Vocab}} \exp(\mathbf{u}_{w_1}^T \mathbf{v}_c)\mathbf{u}_{w_1}}{\sum_{w_2\in\text{Vocab}} \exp(\mathbf{u}_{w_2}^T \mathbf{v}_c)})$$

$$= -(\mathbf{u}_o - (\sum_{w_1\in\text{Vocab}} P(O = w_1|C = c)\mathbf{u}_{w_1}))$$

$$= -(\mathbf{u}_o - \sum_{w_1\in\text{Vocab}} \hat{y}_{w_1}\mathbf{u}_{w_1}) = -\mathbf{u}_o + \mathbf{U}\hat{\mathbf{y}} = \mathbf{U}(\hat{\mathbf{y}} - \mathbf{y})$$

c)

$$\frac{\partial}{\partial \mathbf{u}_w} J_{\text{naive-softmax}(\mathbf{v}_c,o,\mathbf{U})} = -\frac{\partial}{\partial \mathbf{u}_w} \log \frac{\exp(\mathbf{u}_o^T \mathbf{v}_c)}{\sum_{w_1\in\text{Vocab}} \exp(\mathbf{u}_{w_1}^T \mathbf{v}_c)}$$

$$= -\frac{\partial}{\partial \mathbf{u}_w}(\mathbf{u}_o^T \mathbf{v}_c - \log\left(\sum_{w_1\in\text{Vocab}} \exp(\mathbf{u}_{w_1}^T \mathbf{v}_c)\right))$$

If $w \neq o$

$$= -(0 - \frac{\exp\big(\mathbf{u}_w^T \mathbf{v}_c\big)\mathbf{v}_c}{\sum_{w_1 \in \text{Vocab}} \exp\big(\mathbf{u}_{w_1}^T \mathbf{v}_c\big)}) = P(O = w|C = c)\mathbf{v}_c = \hat{y}_w \mathbf{v}_c$$

else if $w = o$

$$= -(\mathbf{v}_c - \frac{\exp\big(\mathbf{u}_w^T \mathbf{v}_c\big)\mathbf{v}_c}{\sum_{w_1 \in \text{Vocab}} \exp\big(\mathbf{u}_{w_1}^T \mathbf{v}_c\big)}) = -\mathbf{v}_c + P(O = w|C = c)\mathbf{v}_c = (\hat{y}_w - 1)\mathbf{v}_c$$

d)

$$\frac{\partial}{\partial x}\sigma(x) = \frac{\partial}{\partial x}\frac{1}{1+e^{-x}} = \frac{\partial}{\partial x}(1+e^{-x})^{-1}$$

$$= -1 \cdot (1+e^{-x})^{-2} \cdot (-e^{-x}) = \frac{e^{-x}}{(1+e^{-x})^2}$$

$$= \frac{1}{(1+e^{-x})} \cdot \frac{e^{-x}+1-1}{(1+e^{-x})} = \frac{1}{(1+e^{-x})} \cdot (1 - \frac{1}{(1+e^{-x})})$$

$$= \sigma(x) \cdot (1 - \sigma(x))$$

e)

1) $\partial \mathbf{v}_c$

$$\frac{\partial}{\partial \mathbf{v}_c} J_{neg-sample}(\mathbf{v}_c, o, \mathbf{U}) = -\frac{\partial}{\partial \mathbf{v}_c} \log\big(\sigma(\mathbf{u}_o^T \cdot \mathbf{v}_c)\big) - \frac{\partial}{\partial \mathbf{v}_c} \sum_{k=1}^{K} \log\big(\sigma(-\mathbf{u}_k^T \cdot \mathbf{v}_c)\big)$$

$$= -\frac{\sigma(\mathbf{u}_o^T \cdot \mathbf{v}_c) \cdot (1 - \sigma(\mathbf{u}_o^T \cdot \mathbf{v}_c))}{\sigma(\mathbf{u}_o^T \cdot \mathbf{v}_c)} \cdot \mathbf{u}_o - (\sum_{k=1}^{K} \frac{\sigma(-\mathbf{u}_k^T \cdot \mathbf{v}_c) \cdot (1 - \sigma(-\mathbf{u}_k^T \cdot \mathbf{v}_c))}{\sigma(-\mathbf{u}_k^T \cdot \mathbf{v}_c)} \cdot (-\mathbf{u}_k))$$

$$= (\sigma(\mathbf{u}_o^T \cdot \mathbf{v}_c) - 1) \cdot \mathbf{u}_o + \sum_{k=1}^{K} (1 - \sigma(-\mathbf{u}_k^T \cdot \mathbf{v}_c)) \cdot \mathbf{u}_k$$

$$= (\sigma(\mathbf{u}_o^T \cdot \mathbf{v}_c) - 1) \cdot \mathbf{u}_o + \sum_{k=1}^{K} \sigma(\mathbf{u}_k^T \cdot \mathbf{v}_c) \cdot \mathbf{u}_k$$

2) $\partial \mathbf{u}_o$

$$\frac{\partial}{\partial \mathbf{u}_o} J_{neg-sample}(\mathbf{v}_c, o, \mathbf{U}) = -\frac{\partial}{\partial \mathbf{u}_o} \log\big(\sigma(\mathbf{u}_o^T \cdot \mathbf{v}_c)\big) - \frac{\partial}{\partial \mathbf{u}_o} \sum_{k=1}^{K} \log\big(\sigma(-\mathbf{u}_k^T \cdot \mathbf{v}_c)\big)$$

$$= (\sigma(\mathbf{u}_o^T \cdot \mathbf{v}_c) - 1)\mathbf{v}_c,$$

3) $\partial \mathbf{u}_k$

$$\frac{\partial}{\partial \mathbf{u}_k} J_{neg-sample}(\mathbf{v}_c, o, \mathbf{U}) = -\frac{\partial}{\partial \mathbf{u}_k} \log\big(\sigma(\mathbf{u}_o^T \cdot \mathbf{v}_c)\big) - \frac{\partial}{\partial \mathbf{u}_k} \sum_{k_1=1}^{K} \log\big(\sigma(-\mathbf{u}_{k_1}^T \cdot \mathbf{v}_c)\big)$$

$$= (1 - \sigma(-\mathbf{u}_k^T \mathbf{v}_c))\mathbf{v}_c = \sigma(\mathbf{u}_k^T \mathbf{v}_c)\mathbf{v}_c$$

This loss function is more efficient to compute, because we don't need to compute softmax over the whole dataset.

f)

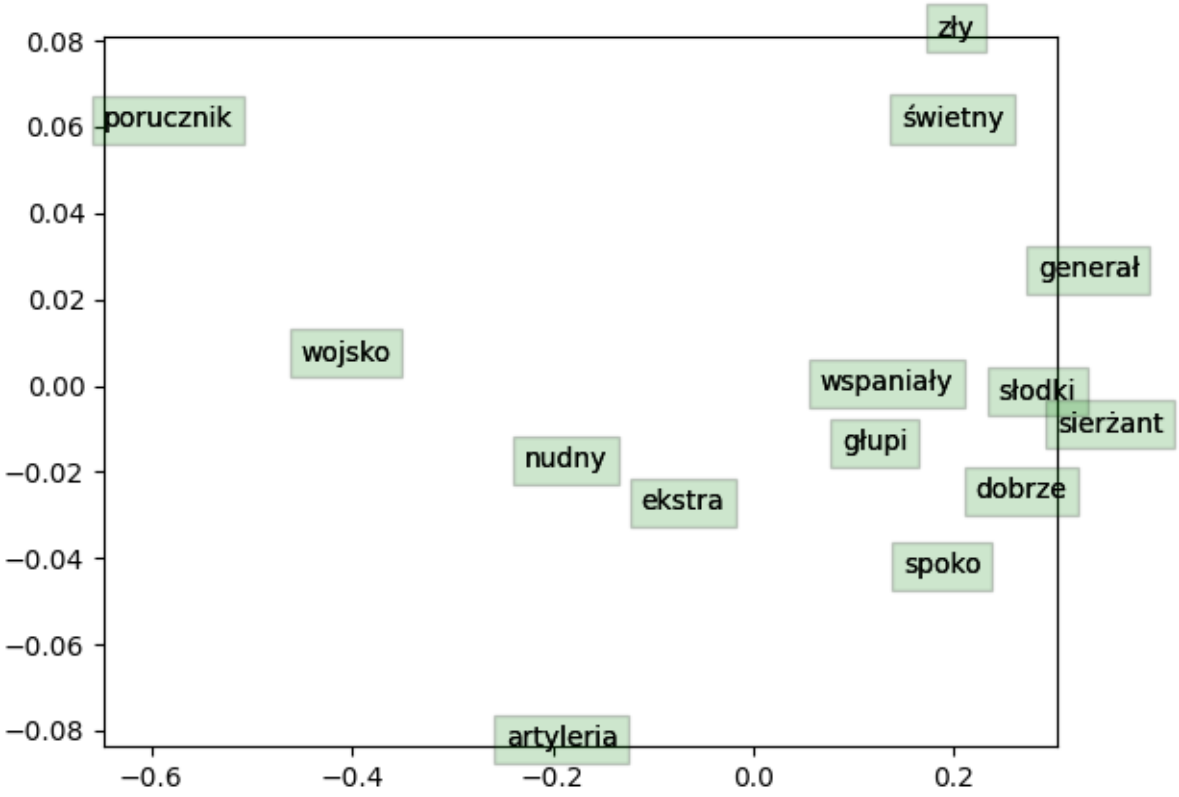
$$\frac{\partial J_{\text{skip-gram}}}{\partial \mathbf{U}}(\mathbf{v_c}, w_{t-m}, \ldots, w_{t+m}, \mathbf{U}) = \sum_{-m \le j \le m, j \neq 0} \frac{\partial J_{\text{skip-gram}}}{\partial \mathbf{U}}(\mathbf{v_c}, w_{t+j}, \mathbf{U})$$

$$\frac{\partial J_{\text{skip-gram}}}{\partial \mathbf{v_c}}(\mathbf{v_c}, w_{t-m}, \ldots, w_{t+m}, \mathbf{U}) = \sum_{-m \le j \le m, j \neq 0} \frac{\partial J_{\text{skip-gram}}}{\partial \mathbf{v_c}}(\mathbf{v_c}, w_{t+j}, \mathbf{U})$$

$$\frac{\partial J_{\text{skip-gram}}}{\partial \mathbf{v_w}}(\mathbf{v_c}, w_{t-m}, \ldots, w_{t+m}, \mathbf{U}) = \sum_{-m \le j \le m, j \neq 0} \frac{\partial J_{skip-gram}}{\partial \mathbf{v_w}}(\mathbf{v_c}, w_{t+j}, \mathbf{U})$$

2 Part 2

c)



Words on the plot are either military related, or describing an opinion. I would expect to see 2 distinct clusters separating those 2 groups. That is not the case. All patterns i can see on the plot have a counterexample. For example sierżant is quite close to generał, but very far from porucznik. Distribution of words seems not random, but it is hard to tell with such a small sample.

3 Part 3

a)

Explain how using m stops the updates from varying as much and why this low variance may be helpful to learning, overall.

By adding the momentum term, the updates to the parameters in each iteration become smoother and less erratic. The update direction is less likely to be influenced by noise in the

data. The low variance in updates helps the learning algorithm to get a better sense of the direction in which the parameters need to be updated.

b)

Since Adam divides the update by a rolling average of the magnitudes of the gradients, which of the model parameters will get larger updates? Why might this help with by learning?

Adam reduces learning rate of parameters with big gradients and increases learning rate of parameters with small gradients. This normalization helps parameters learn at a similar pace and prevents the gradient from fading prematurely.

c)

Why L2 regularization can be helpful in low-data setting?

In low-data settings, machine learning models are more prone to overfitting, as there are fewer examples available to learn from. By encouraging the model to use smaller parameter values, L2 regularization can help to prevent the model from memorizing the training data. L2 reduces the variance in the predictions and making it more robust to small changes in the input data.

The authors of an improved version of Adam (called AdamW optimizer) argue that the equivalence between L2 regularization and weight decay from the point b), true for SGD, does not hold for adaptive schemes. This leads to situation where L2 regularization is not effective in Adam. What is their proposal for improving the Adam update scheme?

The main idea of the paper is to apply the weight decay factor in the last part of the update, so that it does not influence the learning rate.

d)

Why should we apply dropout during training but not during evaluation?

During evaluation, dropout should not be applied because we want the full predictive power of the model to be available. Dropout introduces randomness and variability during training, which helps to improve generalization performance, but this same variability can be problematic during evaluation because it can lead to inconsistent and unpredictable model outputs.

How dropout can be seen from the perspective of Bayesian theory.

From the perspective of Bayesian theory, dropout can be seen as a way of approximating an ensemble of neural networks with a single network. Dropout can be seen as a form

of model averaging, where during training, the dropout masks can be viewed as random variables that sample from a distribution of possible network architectures.