ADVANCED METHODS IN NLP
EXERCISE #2 SOLUTION

Uri Avron [uriavron@gmail.com] [308046994]
Ofri Kleinfeld [ofrik@mail.tau.ac.il] [302893680]
Ido Calman [calman.ido@gmail.com] [308353499]

April 22, 2018

Question 1

(b) Our method for grid searching works in the following manner: We first define our grid resolution and halting condition to be 0.1 and 1 respectively. Then, we run through all combinations of $\lambda_1, \lambda_2, \lambda_3$ within the resolution intervals and afterwards divide our resolution by 2. When perplexity difference between searches reaches less than halting condition, we stop the grid search. Our results for resolution = 0.1 are:

$\lambda_2 \backslash \lambda_1$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	189.6	104.2	87.5	79.0	74.2	71.7	71.1	72.7	77.4	90.5
0.1	115.7	82.6	71.6	65.4	61.6	59.4	58.5	59.1	62.1	
0.2	94.7	72.7	64.3	59.5	56.6	54.9	54.6	56.3		
0.3	82.7	66.0	59.4	55.6	53.3	52.4	53.3			
0.4	74.6	61.3	55.9	52.9	51.4	51.8				
0.5	68.9	57.9	53.5	51.2	51.0					
0.6	64.6	55.3	51.9	51.1						
0.7	61.6	53.8	51.9							
0.8	59.6	53.7								
0.9	59.3									

When narrowing down to resolution = 0.05 our run meets the halting condition at: $\lambda_1 = 0.35, \lambda_2 = 0.5, \lambda_3 = 0.15$ with **perplexity = 50.86**

Question 2

(a)

$$\begin{aligned}
\text{softmax}(\theta)_i &= \frac{\exp(\theta_i)}{\sum_j \exp(\theta_j)} \\
\text{CE}(y, \hat{y}) &= -\sum y_i \log \hat{y}_i \underbrace{=}_{\text{y is one-hot vector}} -1 \cdot \log(\hat{y}_i) \\
&= -\log\left(\frac{\exp(\theta_i)}{\sum_j \exp(\theta_j)}\right) \\
&= -\theta_i + \log\left(\sum_j \exp(\theta_j)\right) \\
\left(\frac{\partial \text{CE}(y, \hat{y})}{\partial \theta}\right)_j &= \frac{\partial \text{CE}(y, \hat{y})}{\partial \theta_j}
\end{aligned}$$

Scalar-wise:

$$\begin{aligned}
\frac{\partial \text{CE}(y, \hat{y})}{\partial \theta_{j \neq i}} &= \frac{1 \cdot \exp(\theta_j)}{\sum_k \exp(\theta_k)} \\
\frac{\partial \text{CE}(y, \hat{y})}{\partial \theta_i} &= -1 + \frac{1 \cdot \exp(\theta_i)}{\sum_k \exp(\theta_k)}
\end{aligned}$$

Therefore, by deriving vector-wise we get:

$$\frac{\partial \text{CE}(y, \hat{y})}{\partial \theta} = \hat{y} - y$$

(b)

$$\begin{aligned}
J &= \text{CE}(y, \hat{y}) \\
\theta &= hW_2 + b_2 \\
h &= \sigma(xW_1 + b_1) \\
\frac{\partial J}{\partial \theta} &\underbrace{=}_{\text{section a}} \hat{y} - y \\
\frac{\partial \theta}{\partial h} &= W_2 \\
\frac{\partial h}{\partial x} &= \sigma'(xW_1 + b_1) \cdot W_1
\end{aligned}$$

Now, by applying the chain rule we get:

$$\begin{aligned} \frac{\partial J}{\partial x} &= \frac{\partial J}{\partial \theta} \cdot \frac{\partial \theta}{\partial h} \cdot \frac{\partial h}{\partial x} \\ &= (\hat{y} - y) W_2^T \sigma'(xW_1 + b_1) \cdot W_1^T \end{aligned}$$

where $\sigma'(xW_1 + b_1) = \sigma(xW_1 + b_1) \circ ((1 - \sigma(xW_1 + b_1))$. Note that the multiplication should be performed element-wise, since $xW_1 + b_1$ is a vector.

(d) Results:
#params: 104550
#train examples: 1118296
dev perplexity : **112.967665327**