Assignment 1 writeup

Thomas Lu

1 Problem 1

(a) We have

$$\begin{aligned}
\text{softmax}(\mathbf{x}+c)_i &= \frac{e^{x_i+c}}{\sum_j e^{x_j+c}} \\
&= \frac{e^c(e^{x_i})}{e^c \sum_j e^{x_j}} \\
&= \frac{e^c}{e^c} \frac{e^{x_i}}{\sum_j e^{x_j}} \\
&= \frac{e^{x_i}}{\sum_j e^{x_j}} \\
&= \text{softmax}(\mathbf{x})_i,
\end{aligned}$$

so $\text{softmax}(\mathbf{x}+c) = \text{softmax}(\mathbf{x})$, as desired.

2 Problem 2

(a) We have

$$\begin{aligned}
\sigma(x) &= \frac{1}{1+e^{-x}} \\
\sigma'(x) &= \frac{d}{dx}(1+e^{-x})^{-1} \\
&= -(1+e^{-x})^{-2} \frac{d}{dx}\left(1+e^{-x}\right) \\
&= -\left(\frac{1}{1+e^{-x}}\right)^2 \left(-e^{-x}\right) \\
&= \left(\frac{1}{1+e^{-x}}\right)\left(1-\frac{1}{1+e^{-x}}\right) \\
&= \left(\sigma(x)\right)\left(1-\sigma(x)\right).
\end{aligned}$$

(b) We have $\hat{y} = \text{softmax}(\theta)$. Suppose that y is one-hot with $y_k = 1$. Then

$$\begin{aligned}\frac{\partial}{\partial \theta_j} CE(y, \hat{y}) &= \frac{\partial}{\partial \theta_j} \sum_i -y_i \log \frac{e^{\theta_i}}{\sum_i e^{\theta_i}} \\ &= -\frac{\partial}{\partial \theta_j} \log \frac{e^{\theta_k}}{\sum_i e^{\theta_i}} \\ &= \frac{\partial}{\partial \theta_j} \left(\log \left(\sum_i e^{\theta_i} \right) - \log e^{\theta_k} \right)\end{aligned}$$

.

We now split into two cases. If $j \neq k$, then

$$\begin{aligned}\frac{\partial}{\partial \theta_j} CE(y, \hat{y}) &= \frac{\partial}{\partial \theta_j} \left(\log \left(\sum_i e^{\theta_i} \right) - \log e^{\theta_k} \right) \\ &= \frac{1}{\sum_i e^{\theta_i}} (e^{\theta_j}) \\ &= \text{softmax}(\theta)_j.\end{aligned}$$

If $j = k$, then

$$\begin{aligned}\frac{\partial}{\partial \theta_j} CE(y, \hat{y}) &= \frac{\partial}{\partial \theta_k} \left(\log \left(\sum_i e^{\theta_i} \right) - \log e^{\theta_k} \right) \\ &= \frac{1}{\sum_i e^{\theta_i}} (e^{\theta_k}) - \frac{\partial}{\partial \theta_k} \theta_k \\ &= \text{softmax}(\theta)_k - 1.\end{aligned}$$

Thus

$$\vec{\nabla}_\theta CE(y, \hat{y}) = \text{softmax}(\theta) - y.$$

(c) We begin with a couple of theorems:

Theorem: If $f_1, f_2, \ldots, f_n, g$ are differentiable, and

$$y = g(f_1(x), f_2(x), \ldots, f_n(x)),$$

then

$$\frac{dy}{dx} = \sum_{i=1}^{n} \frac{\partial y}{\partial f_i} \frac{df_i}{dx}.$$

Theorem: If x and y are row vectors and z is a scalar, and f and g are differentiable with $y = f(x)$ and $z = g(y)$, then

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} \frac{\partial y}{\partial x},$$

where

$$\left(\frac{\partial y}{\partial x} \right)_{ij} = \frac{\partial y_i}{\partial x_j}.$$

We now compute $\partial J / \partial x$. Letting $z_1 = xW_1 + b_1$ and $z_2 = hW_2 + b_2$, we have

$$\frac{\partial J}{\partial x} = \frac{\partial J}{\partial z_2} \frac{\partial z_2}{\partial h} \frac{\partial h}{\partial z_1} \frac{\partial z_1}{\partial x}.$$

We evaluate the partials on the RHS of the above equation in sequence. We have:

$$\begin{aligned}
\frac{\partial J}{\partial z_2} &= \text{softmax}(z_2) - y \\
\frac{\partial z_2}{\partial h} &= W_2^T \\
\frac{\partial h}{\partial z_1} &= \text{diag}(\sigma(z_1))\text{diag}(1 - \sigma(z_1)) \\
\frac{\partial z_1}{\partial x} &= W_1^T \\
\Rightarrow \frac{\partial J}{\partial x} &= (\text{softmax}(z_2) - y)W_2^T\text{diag}(\sigma(z_1))\text{diag}(1 - \sigma(z_1))W_1^T \\
&= (\text{softmax}(z_2) - y)W_2^T W_1^T \circ \sigma(z_1) \circ (1 - \sigma(z_1))
\end{aligned}$$

where $\text{diag}(v)$ denotes the diagonal matrix D with $D_i i = v_i$ and 1 denotes a vector of ones where appropriate.

(d) There are four groups of parameters:

- b_1, a bias vector with H entries,
- b_2, a bias vector with D_y entries,
- W_1, a $D_x \times H$ weight matrix, and
- W_2, a $H \times D_y$ weight matrix.

Thus the total number of parameters is

$$H + D_y + HD_x + HD_y = D_y(H + 1) + H(D_x + 1).$$

3 Problem 3

(a) Let V be our vocabulary, and for $o, c \in V$, let v_c denote the center word (column) vector for c and u_o denote the outer word (column) vector for o. Let U denote the matrix

$$\begin{bmatrix} — & u_{w_1}^T — \\ — & u_{w_2}^T — \\ & \vdots \\ — & u_{w_N}^T — \end{bmatrix},$$

where $V = \{w_1, w_2, \dots, w_N\}$. Let

$$\hat{y}_i = p(w_i|c) = \frac{\exp(u_{w_i}^T v_c)}{\sum_{w \in V} \exp(u_w^T v_c)}$$

and $\hat{y} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_N]$. Supposing that y is one-hot with the one at o, we have

$$\begin{aligned}
\frac{\partial J}{\partial v_c} &= \frac{\partial}{\partial v_c} CE(y, \hat{y}) \\
&= -\frac{\partial}{\partial v_c} \sum_{i=1}^{N} y_i \log \hat{y}_i \\
&= -\frac{\partial}{\partial v_c} \log \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp\left(u_w^T v_c\right)} \\
&= -\frac{\partial}{\partial v_c} \left(\log \left(\exp(u_o^T v_c) \right) - \log \left(\sum_{w \in V} \exp\left(u_w^T v_c\right) \right) \right) \\
&= -u_o + \frac{1}{\sum_{w \in V} \exp(u_w^T v_c)} \sum_{w \in V} \frac{\partial}{\partial v_c} \exp\left(u_w^T v_c\right) \\
&= -u_o + \frac{1}{\sum_{w \in V} \exp(u_w^T v_c)} \sum_{w \in V} u_w \exp\left(u_w^T v_c\right) \\
&= -u_o + \sum_{i=1}^{N} u_{w_i} \hat{y}_i \\
&= U^T \left(\hat{y} - y\right).
\end{aligned}$$

(b) We have, using the same variable conventions as in (a),

$$\begin{aligned}
\frac{\partial J}{\partial u_{w_i}} &= -\frac{\partial}{\partial u_{w_i}} \log \frac{\exp(u_o^T v_c)}{\sum_{w \in V} \exp\left(u_w^T v_c\right)} \\
&= -\frac{\partial}{\partial u_{w_i}} \left(\log \left(\exp(u_o^T v_c) \right) - \log \left(\sum_{w \in V} \exp\left(u_w^T v_c\right) \right) \right) \\
&= -v_c(1_{w_i = o}) + \frac{1}{\sum_{w \in V} \exp(u_w^T v_c)} \frac{\partial}{\partial u_{w_i}} \left(\sum_{w \in V} \exp\left(u_w^T v_c\right) \right) \\
&= -v_c(1_{w_i = o}) + \left(\frac{\exp\left(u_{w_i}^T v_c\right)}{\sum_{w \in V} \exp(u_w^T v_c)} \right) v_c \\
&= -v_c y_i + \hat{y}_i v_c \\
\Rightarrow \frac{\partial J}{\partial U} &= (\hat{y} - y) v_c^T.
\end{aligned}$$

(c) Note first that

$$\frac{d}{dx} \log(\sigma(x)) = \frac{1}{\sigma(x)} \sigma'(x) = \frac{1}{\sigma(x)} \left(\sigma(x)(1 - \sigma(x))\right) = 1 - \sigma(x).$$

For $w \notin \{o, 1, 2, \dots, K\}$, it is clear that $\partial J / \partial u_w = 0$. Now

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

$$\frac{\partial J}{\partial u_i} = -\frac{\partial}{\partial u_i} \log \sigma(-u_i^T v_c) = v_c(1 - \sigma(-u_i^T v_c)),$$

where $i \in \{1, 2, \ldots, K\}$. Similarly, we can compute

$$\frac{\partial J}{\partial v_c} = u_o(\sigma(u_o^T v_c) - 1) + \sum_{k=1}^{K} u_k(1 - \sigma(-u_k^T v_c)).$$

This is much faster to compute because we only need to perform K/V (proportionally) as many dot products.

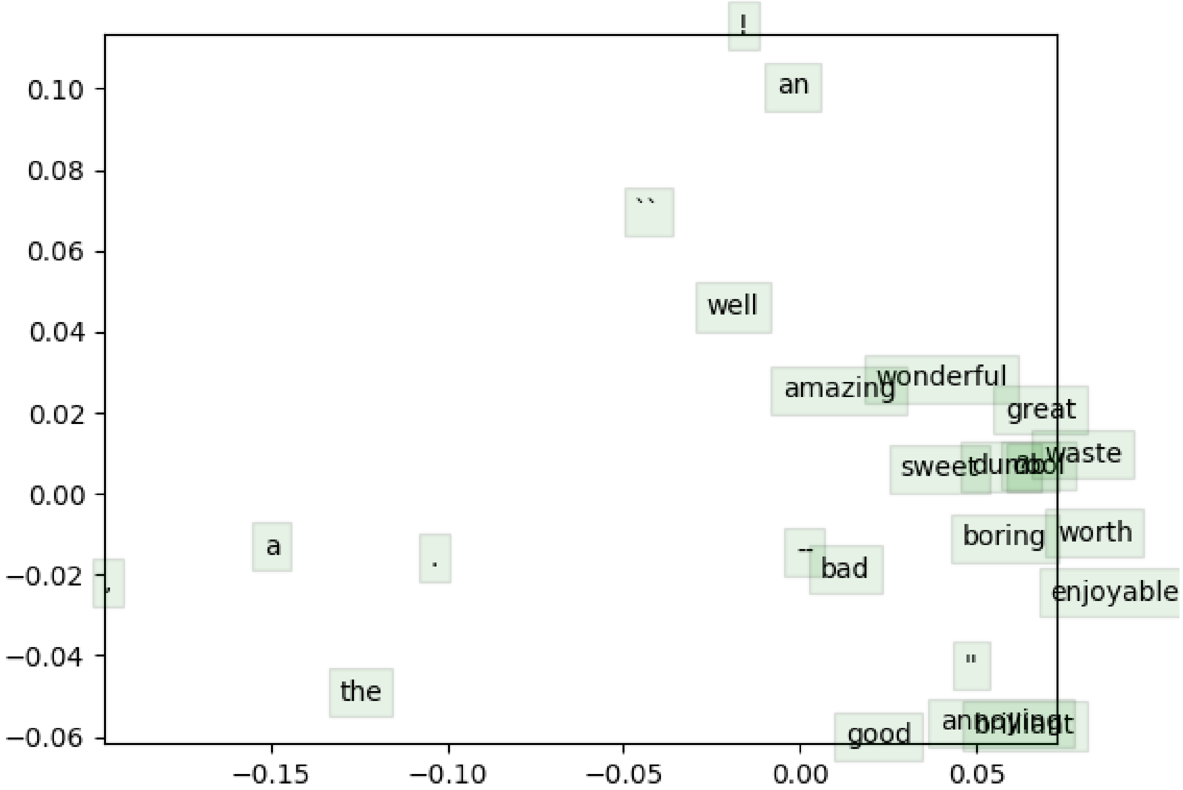
(d) For the skip-gram model, we have

$$\begin{aligned}
\frac{\partial J_{\text{skipgram}}(w_{t-m\ldots t+m})}{\partial v_c} &= \sum_{\substack{t-m \leq j \leq t+m \\ j \neq 0}} \frac{\partial F(w_j, v_c)}{\partial v_c} \\
\frac{\partial J_{\text{skipgram}}(w_{t-m\ldots t+m})}{\partial v_j} &= 0,\ j \neq c \\
\frac{\partial J_{\text{skipgram}}(w_{t-m\ldots t+m})}{\partial U} &= \sum_{\substack{t-m \leq j \leq t+m \\ j \neq 0}} \frac{\partial F(w_j, v_c)}{\partial U}
\end{aligned}$$

For CBOW, we have

$$\begin{aligned}
\frac{\partial J_{\text{CBOW}}(w_{t-m\ldots t+m})}{\partial v_c} &= \frac{\partial F(w_t, \hat{v})}{\partial \hat{v}} \frac{\partial \hat{v}}{\partial v_c} = \frac{\partial F(w_t, \hat{v})}{\partial \hat{v}} \\
\frac{\partial J_{\text{CBOW}}(w_{t-m\ldots t+m})}{\partial U} &= \frac{\partial F(w_t, \hat{v})}{\partial U} \\
\frac{\partial J_{\text{CBOW}}(w_{t-m\ldots t+m})}{\partial v_j} &= 0,\ j \neq c.
\end{aligned}$$

(g)



The plot is a 2-D projection of the word vectors trained from our corpus. We see that a number of generic quality-describing adjectives are clustered together (amazing, wonderful, great) along with some less general but still quite versatile ones (sweet, dumb, boring, enjoyable). Interestingly, “waste” is also clustered pretty closely to these words; it is often used in a similar context (“x was a waste of time” vs. “x was boring”) although it is a noun rather than an adjective.