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 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

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