Deep Learning - Spring 2019
Homework 1

Daniel Rivera Ruiz
Department of Computer Science
New York University
drr342@nyu.edu

1 Backprop

1.1 Warm-up

The chain rule is at the heart of backpropagation. Assume you are given input $\boldsymbol{x}$ and output $\boldsymbol{y}$, both in $\mathbb{R}^2$, and the error backpropagated to the output is $\frac{\partial L}{\partial \boldsymbol{y}}$. In particular, let

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}, \tag{1}$$

where $\boldsymbol{W} \in \mathbb{R}^{2\times 2}$ and $\boldsymbol{x}, \boldsymbol{b} \in \mathbb{R}^2$. Give an expression for $\frac{\partial L}{\partial \boldsymbol{W}}$ and $\frac{\partial L}{\partial \boldsymbol{b}}$ in terms of $\frac{\partial L}{\partial \boldsymbol{y}}$ and $\boldsymbol{x}$ using the chain rule.

Solution
If we let

$$\begin{aligned}
\boldsymbol{y} &= \begin{bmatrix} y_1 & y_2 \end{bmatrix}^\top \\
\boldsymbol{x} &= \begin{bmatrix} x_1 & x_2 \end{bmatrix}^\top \\
\boldsymbol{b} &= \begin{bmatrix} b_1 & b_2 \end{bmatrix}^\top \\
\boldsymbol{W} &= \begin{bmatrix} W_{1,1} & W_{1,2} \\ W_{2,1} & W_{2,2} \end{bmatrix}
\end{aligned}$$

We can rewrite equation 1 as follows:

$$\begin{aligned}
y_1 &= W_{1,1}x_1 + W_{1,2}x_2 + b_1 \\
y_2 &= W_{2,1}x_1 + W_{2,2}x_2 + b_2
\end{aligned}$$

Obtaining the partial derivatives of these equations and using chain rule it is easy to see that

$$\begin{aligned}
\frac{\partial L}{\partial W_{i,j}} &= \frac{\partial L}{\partial y_i} \cdot \frac{\partial y_i}{\partial W_{i,j}} \\
&= \frac{\partial L}{\partial y_i} \cdot x_j \\
\frac{\partial L}{\partial b_i} &= \frac{\partial L}{\partial y_i} \cdot \frac{\partial y_i}{\partial b_i} \\
&= \frac{\partial L}{\partial y_i}
\end{aligned}$$

Finally, we can write the vector expressions for these derivatives:

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{W}} &= \begin{bmatrix} \frac{\partial L}{\partial W_{1,1}} & \frac{\partial L}{\partial W_{1,2}} \\ \frac{\partial L}{\partial W_{2,1}} & \frac{\partial L}{\partial W_{2,2}} \end{bmatrix} = \begin{bmatrix} \frac{\partial L}{\partial y_1} \cdot x_1 & \frac{\partial L}{\partial y_1} \cdot x_2 \\ \frac{\partial L}{\partial y_2} \cdot x_1 & \frac{\partial L}{\partial y_2} \cdot x_2 \end{bmatrix} \\
&= \frac{\partial L}{\partial \boldsymbol{y}} \cdot \boldsymbol{x}^\top \\
\frac{\partial L}{\partial \boldsymbol{b}} &= \begin{bmatrix} \frac{\partial L}{\partial b_1} \\ \frac{\partial L}{\partial b_2} \end{bmatrix} = \begin{bmatrix} \frac{\partial L}{\partial y_1} \\ \frac{\partial L}{\partial y_2} \end{bmatrix} \\
&= \frac{\partial L}{\partial \boldsymbol{y}}
\end{aligned}$$

1.2 Softmax

Multinomial logistic regression is a generalization of logistic regression into multiple classes. The softmax expression which indicates the probability of the j-th class is as follows:

$$y_j = \frac{\exp(\beta x_j)}{\sum_{k=1}^{n} \exp(\beta x_k)}$$

What is the expression for $\frac{\partial y_j}{\partial x_i}$? (Hint: Answer differs when $i = j$ and $i \neq j$).

Solution

I. $i \neq j$

$$\begin{aligned}
\frac{\partial y_j}{\partial x_i} &= \frac{\partial}{\partial x_i} \left[\frac{\exp(\beta x_j)}{\sum_{k=1}^{n} \exp(\beta x_k)} \right] \\
&= \exp(\beta x_j) \left[(-1) \left(\sum_{k=1}^{n} \exp(\beta x_k) \right)^{-2} \right] \frac{\partial}{\partial x_i} \left[\sum_{k=1}^{n} \exp(\beta x_k) \right] \\
&= -\frac{\exp(\beta x_j)}{\left[\sum_{k=1}^{n} \exp(\beta x_k)\right]^2} \left[\beta \exp(\beta x_i)\right] \\
&= -\beta \cdot \frac{\exp(\beta x_j)}{\sum_{k=1}^{n} \exp(\beta x_k)} \cdot \frac{\exp(\beta x_i)}{\sum_{k=1}^{n} \exp(\beta x_k)} \\
&= -\beta y_j y_i
\end{aligned}$$

II. $i = j$

$$\begin{aligned}
\frac{\partial y_j}{\partial x_j} &= \frac{\partial}{\partial x_j} \left[\frac{\exp(\beta x_j)}{\sum_{k=1}^{n} \exp(\beta x_k)} \right] \\
&= \exp(\beta x_j) \cdot \frac{\partial}{\partial x_j} \left[\frac{1}{\sum_{k=1}^{n} \exp(\beta x_k)} \right] + \frac{1}{\sum_{k=1}^{n} \exp(\beta x_k)} \cdot \frac{\partial}{\partial x_j} \left[\exp(\beta x_j)\right] \\
&= -\beta y_j^2 + \frac{\beta \exp(\beta x_j)}{\sum_{k=1}^{n} \exp(\beta x_k)} \\
&= -\beta y_j^2 + \beta y_j \\
&= \beta y_j (1 - y_j)
\end{aligned}$$