Written by Tianyi Luo

Solution for Homework Assignment 4

1. Let $\mathbf{x}$ denote the vector of points and $\mathbf{y}$ the vector of labels. Note the $\mathbf{x}^2 = \mathbf{x}^\top \mathbf{x} = \sum_i x_i^2 = \|\mathbf{x}\|_2^2$.

$$
\begin{aligned}
\mathbb{E}\left[L(w_i)\right] &= \sum_i \frac{x_i^2}{\mathbf{x}^2} \sum_j (\underbrace{\frac{y_i}{x_i}}_{w_i^*} x_j - y_j)^2 \\
&= \frac{1}{\mathbf{x}^2} \sum_i \sum_j (y_i x_j - y_j x_i)^2 \\
&= \frac{1}{\mathbf{x}^2} \sum_i \sum_j (y_i^2 x_j^2 - 2 y_i x_j y_j x_i + y_j^2 x_i^2) \\
&= \frac{1}{\mathbf{x}^2} \sum_i \sum_j (2 y_i^2 x_j^2 - 2 y_i x_j y_j x_i) \\
&= 2\mathbf{y}^2 - \frac{2}{\mathbf{x}^2} \sum_i \sum_j y_i y_j x_i x_j \\
L(w^*) &= (\mathbf{x} \underbrace{\frac{\mathbf{x}^\top \mathbf{y}}{\mathbf{x}^2}}_{w^*} - \mathbf{y})^2 \\
&= (\frac{\mathbf{x}\mathbf{x}^\top}{\mathbf{x}^2}\mathbf{y} - \mathbf{y})^\top (\frac{\mathbf{x}\mathbf{x}^\top}{\mathbf{x}^2}\mathbf{y} - \mathbf{y}) \\
&= \mathbf{y}^\top \frac{\mathbf{x}\cancel{\mathbf{x}^\top}}{\mathbf{x}^2} \frac{\cancel{\mathbf{x}}\mathbf{x}^\top}{\cancel{\mathbf{x}^2}} \mathbf{y} - 2\mathbf{y}^\top \frac{\mathbf{x}\mathbf{x}^\top}{\mathbf{x}^2}\mathbf{y} + \mathbf{y}^2 \\
&= \mathbf{y}^2 - \mathbf{y}^\top \frac{\mathbf{x}\mathbf{x}^\top}{\mathbf{x}^2}\mathbf{y} \\
&= \mathbf{y}^2 - \frac{1}{\mathbf{x}^2} \sum_i \sum_j y_i y_j x_i x_j
\end{aligned}
$$

2. Let us define the notations:

$$
\begin{aligned}
t_f &: \textit{True label.} \\
y_f &: \textit{Prediction result.} \\
E_n &: \textit{Loss function of the nth data point.} \\
a_f &: \textit{Linear summation of the hidden layer output plus the bias.} \\
w_{fj} &: \textit{Weights between the hidden layer and output layer.} \\
b_f &: \textit{Bias item of the hidden layer output.} \\
z_f &: \textit{One final output of output layer after non} - \textit{linear transformation.} \\
a_j &: \textit{Linear summation of the input layer output plus the bias.} \\
w_{ji} &: \textit{Weights between the input layer and hidden layer.} \\
b_j &: \textit{Bias item of the input layer output.} \\
z_j &: \textit{The jth node's output of hidden layer after non} - \textit{linear transformation.}
\end{aligned}
$$

Based on the chain rule of derivative and the notations mentioned above, we have:

$$
\begin{aligned}
a_f &= \sum_j w_{fj} z_j + b_f \quad\quad z_f = \Phi(a_f) \\
a_j &= \sum_i w_{ji} x_i + b_j \quad\quad z_j = \Phi(a_j)
\end{aligned}
$$

So we can calculate the derivatives of weights and bias:

$$
\begin{aligned}
(1):\quad & \frac{\partial E_n}{\partial w_{fj}} = \frac{\partial E_n}{\partial a_f} \cdot \frac{\partial a_f}{\partial w_{fj}} \\
& \delta_f = \frac{\partial E_n}{\partial a_f} = \frac{\frac{1}{2}(z_f - t_f)^2}{\partial a_f} = \frac{\frac{1}{2}(z_f - t_f)^2}{\partial z_f} \cdot \frac{\partial z_f}{\partial a_f} = (z_f - t_f) \cdot \Phi^{'}(a_f) = (y_f - t_f) \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{a_f^2}{2}} \\
& (\Phi^{'}(a) = (\int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}} dz)^{'} = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{a^2}{2}} \quad and \quad z_f = y_f) \\
& \frac{\partial a_f}{\partial w_{fj}} = z_j, \quad So \quad \frac{\partial E_n}{\partial w_{fj}} = (y_f - t_f) \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{a_f^2}{2}} \cdot z_j \\
(2):\quad & \frac{\partial E_n}{\partial w_{ji}} = \frac{\partial E_n}{\partial a_j} \cdot \frac{\partial a_j}{\partial w_{ji}} \\
& \delta_j = \frac{\partial E_n}{\partial a_j} = \frac{\partial E_n}{\partial a_f} \cdot \frac{\partial a_f}{\partial a_j} = \delta_f \cdot \frac{\partial a_f}{\partial a_j} = \delta_f \cdot \frac{\partial a_f}{\partial z_j} \cdot \frac{\partial z_j}{\partial a_j} = \delta_f \cdot w_{fj} \cdot \sqrt{2\pi} \cdot e^{-\frac{a_j^2}{2}} \\
& (\Phi^{'}(a) = (\int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}} dz)^{'} = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{a^2}{2}}) \\
& \frac{\partial a_j}{\partial w_{ji}} = x_i, \quad So \quad \frac{\partial E_n}{\partial w_{ji}} = \delta_f \cdot w_{fj} \cdot \sqrt{2\pi} \cdot e^{-\frac{a_j^2}{2}} \cdot x_i \\
(3):\quad & \frac{\partial E_n}{\partial b_f} = \frac{\partial E_n}{\partial a_f} \cdot \frac{\partial a_f}{\partial b_f} = \delta_f = (y_f - t_f) \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{a_f^2}{2}} \\
(4):\quad & \frac{\partial E_n}{\partial b_j} = \frac{\partial E_n}{\partial a_j} \cdot \frac{\partial a_j}{\partial b_j} = \delta_j = \delta_f \cdot w_{fj} \cdot \sqrt{2\pi} \cdot e^{-\frac{a_j^2}{2}}
\end{aligned}
$$

3. The definition of max function is as following:

$$
h(a) = max(0, a) = \begin{cases} a \text{ if } a \geq 0 \\ 0 \text{ if } a < 0 \end{cases}
$$

And the integral function of max function is as following:

$$
\begin{aligned}
H(a) = \int max(0, a) &= \begin{cases} \frac{1}{2}a^2 \text{ if } a \geq 0 \\ 0 \text{ if } a < 0 \end{cases} \\
&= \frac{1}{2} max(0, a) a
\end{aligned}
$$

Based on the 16/23 of "Bregman divergences for constructing regularizers and losses" in the lecture 5, the formula of matching loss is as following:

$$
\begin{aligned}
& \triangle_H(w \cdot x, h^{-1}) = H(w \cdot x) - H(h^{-1}(y)) - (w \cdot x - h^{-1}(y))y \\
& Since\ we\ have\ h^{-1}(y) = y,\ so\ we\ have: \\
& \triangle_H(w \cdot x, y) = H(w \cdot x) - H(y) - (w \cdot x - y)y \\
& So\ we\ have:\ Mathing\ loss = \frac{1}{2} max(0, w \cdot x) w \cdot x - \frac{1}{2} max(0, y) y - (w \cdot x - y)y
\end{aligned}
$$