CS589 Homework 5 Solutions

Proofs: Task 1, Task 2

Task 1

Show that when $C = 2$, $y_p \in \{1, -1\}$, multi-class Softmax cost reduces to two-class Softmax cost. Multi-class Softmax is given in the equation below (from *Machine Learning Refined*, Eq. 7.23):

$$g(\mathbf{w}_0, ..., \mathbf{w}_{C-1}) = \frac{1}{P}\sum_{p=1}^{P}\left[\log\left(\sum_{j=0}^{C-1} e^{\dot{\mathbf{x}}_p^{\intercal}\mathbf{w}_j}\right) - \dot{\mathbf{x}}_p^{\intercal}\mathbf{w}_{y_p}\right] \tag{1}$$

Two-class Softmax is given as below (*Machine Learning Refined*, Eq. 6.25):

$$g(\mathbf{w}) = \frac{1}{P}\sum_{p=1}^{P}\log\left(1 + e^{-y_p\dot{\mathbf{x}}_p^{\intercal}\mathbf{w}}\right) \tag{2}$$

Solution. To disambiguate, we denote with $g^{(C)}$ the multi-class Softmax cost function and with $\breve{g}$ the two-class Softmax cost function. Let $C = 2$ with $y_p \in \{1, -1\}$, and define a class label assignment function $f : \{-1, 1\} \to \{0, 1\}$ such that $f(-1) = 0$, $f(1) = 1$. Then we have weights $\mathbf{w}_{f(y_p)} = \mathbf{w}_0$ for $y_p = -1$, $\mathbf{w}_{f(y_p)} = \mathbf{w}_1$ for $y_p = 1$.

Consider the following form of the multi-class Softmax cost (*Machine Learning Refined*, Eq. 7.24):

$$g^{(C)}(\mathbf{w}_0, ..., \mathbf{w}_{C-1}) = \frac{1}{P}\sum_{p=1}^{P}\log\left(1 + \sum_{\substack{j=0\\ j\neq y_p}}^{C-1} e^{\dot{\mathbf{x}}_p^{\intercal}(\mathbf{w}_j - \mathbf{w}_{y_p})}\right). \tag{3}$$

We substitute $C = 2$ and use the class label assignment function f to assign $\mathbf{w}_{y_p}$ from our labels $y_p \in \{-1, 1\}$:

$$g^{(2)}(\mathbf{w}_0, \mathbf{w}_1) = \frac{1}{P}\sum_{p=1}^{P}\log\left(1 + \sum_{\substack{j=0\\ j\neq f(y_p)}}^{2-1} e^{\dot{\mathbf{x}}_p^{\intercal}(\mathbf{w}_j - \mathbf{w}_{f(y_p)})}\right)$$

and expand the summation over values of j to get:

$$g^{(2)}(\mathbf{w}_0, \mathbf{w}_1) = \frac{1}{P}\sum_{p=1}^{P}\log\left(1 + \mathbb{1}\Big[f(y_p) \neq 0\Big]e^{\dot{\mathbf{x}}_p^{\intercal}(\mathbf{w}_0 - \mathbf{w}_{f(y_p)})} + \mathbb{1}\Big[f(y_p) \neq 1\Big]e^{\dot{\mathbf{x}}_p^{\intercal}(\mathbf{w}_1 - \mathbf{w}_{f(y_p)})}\right).$$

Now use the assignment of $\mathbf{w}_{f(y_p)}$ from y_p to simplify and define a step function on y_p:

$$
\begin{aligned}
g^{(2)}(\mathbf{w}_0, \mathbf{w}_1) &= \frac{1}{P}\sum_{p=1}^{P} \log\left(1 + \mathbb{1}\Big[y_p \neq -1\Big] e^{\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_0 - \mathbf{w}_{f(y_p)})} + \mathbb{1}\Big[y_p \neq 1\Big] e^{\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_1 - \mathbf{w}_{f(y_p)})}\right) \\
&= \frac{1}{P}\sum_{p=1}^{P} \begin{cases} \log\left(1 + e^{\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_0 - \mathbf{w}_1)}\right) & y_p = 1 \\ \log\left(1 + e^{\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_1 - \mathbf{w}_0)}\right) & y_p = -1. \end{cases}
\end{aligned}
$$

Then rearrange case expressions to present in similar terms of $\mathbf{w}_0$, $\mathbf{w}_1$:

$$
\begin{aligned}
g^{(2)}(\mathbf{w}_0, \mathbf{w}_1) &= \frac{1}{P}\sum_{p=1}^{P} \begin{cases} \log\left(1 + e^{\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_0 - \mathbf{w}_1)}\right) & y_p = 1 \\ \log\left(1 + e^{\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_1 - \mathbf{w}_0)}\right) & y_p = -1 \end{cases} \\
&= \frac{1}{P}\sum_{p=1}^{P} \begin{cases} \log\left(1 + e^{1\cdot\dot{\mathbf{x}}_p^\intercal(-1\cdot-1(\mathbf{w}_0 - \mathbf{w}_1))}\right) & y_p = 1 \\ \log\left(1 + e^{-1\cdot-1\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_1 - \mathbf{w}_0)}\right) & y_p = -1 \end{cases} \\
&= \frac{1}{P}\sum_{p=1}^{P} \begin{cases} \log\left(1 + e^{1\cdot\dot{\mathbf{x}}_p^\intercal(-1(\mathbf{w}_1 - \mathbf{w}_0))}\right) & y_p = 1 \\ \log\left(1 + e^{-1\cdot-1\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_1 - \mathbf{w}_0)}\right) & y_p = -1 \end{cases} \\
&= \frac{1}{P}\sum_{p=1}^{P} \begin{cases} \log\left(1 + e^{-1\cdot1\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_1 - \mathbf{w}_0)}\right) & y_p = 1 \\ \log\left(1 + e^{-1\cdot-1\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_1 - \mathbf{w}_0)}\right) & y_p = -1 \end{cases}
\end{aligned}
$$

and push y_p into the function to obtain equivalent summand terms and collapse the function inline:

$$
\begin{aligned}
g^{(2)}(\mathbf{w}_0, \mathbf{w}_1) &= \frac{1}{P}\sum_{p=1}^{P} \begin{cases} \log\left(1 + e^{-1\cdot(1)\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_1 - \mathbf{w}_0)}\right) & y_p = 1 \\ \log\left(1 + e^{-1\cdot(-1)\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_1 - \mathbf{w}_0)}\right) & y_p = -1 \end{cases} \\
&= \frac{1}{P}\sum_{p=1}^{P} \log\left(1 + e^{-y_p\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_1 - \mathbf{w}_0)}\right).
\end{aligned}
$$

Now let $\mathbf{w}_s = \mathbf{w}_1 - \mathbf{w}_0$. Then we have:

$$
g^{(2)}(\mathbf{w}_0, \mathbf{w}_1) = \frac{1}{P}\sum_{p=1}^{P} \log\left(1 + e^{-y_p\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_1 - \mathbf{w}_0)}\right) = \frac{1}{P}\sum_{p=1}^{P} \log\left(1 + e^{-y_p\dot{\mathbf{x}}_p^\intercal\mathbf{w}_s}\right) = \breve{g}(\mathbf{w}_s). \tag{4}
$$

So for $C = 2$, $y_p \in \{-1, 1\}$, and $\mathbf{w}_s = \mathbf{w}_1 - \mathbf{w}_0$, we have shown $g^{(2)}(\mathbf{w}_0, \mathbf{w}_1) = \breve{g}(\mathbf{w}_s)$.

■

Task 2

Show that when $C = 2$, $y_p \in \{0, 1\}$, multi-class Softmax cost is equivalent to the two-class Cross Entropy cost.

The multi-class Softmax cost is given in Equation 1. The two-class Cross Entropy cost is given below (from *Machine Learning Refined*, Eq. 6.12):

$$g(\mathbf{w}) = -\frac{1}{P}\sum_{p=1}^{P} y_p \log\left(\sigma(\dot{\mathbf{x}}_p^\intercal \mathbf{w})\right) + (1 - y_p)\log\left(1 - \sigma(\dot{\mathbf{x}}_p^\intercal \mathbf{w})\right) \tag{5}$$

Solution. To disambiguate cost functions, we denote with $g^{(C)}$ the multi-class Softmax cost and with $g^{\diamondsuit}$ the two-class Cross Entropy cost. Let $C = 2$ with $y_p \in \{0, 1\}$.

Consider the following form of the multi-class Softmax cost (*Machine Learning Refined*, Eq. 7.28):

$$g^{(C)}(\mathbf{w}_0, ..., \mathbf{w}_{C-1}) = -\frac{1}{P}\sum_{p=1}^{P} \log\left(\frac{e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_{y_p}}}{\sum_{j=0}^{C-1} e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_j}}\right). \tag{6}$$

Substituting $C = 2$, we have:

$$g^{(2)}(\mathbf{w}_0, \mathbf{w}_1) = -\frac{1}{P}\sum_{p=1}^{P} \log\left(\frac{e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_{y_p}}}{e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_0} + e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_1}}\right)$$

and use the assignment of $\mathbf{w}_{y_p}$ from y_p to define a step function with cases on y_p:

$$\begin{aligned}
g^{(2)}(\mathbf{w}_0, \mathbf{w}_1) &= -\frac{1}{P}\sum_{p=1}^{P} \begin{cases} \log\left(\frac{e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_0}}{e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_0} + e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_1}}\right) & y_p = 0 \\ \log\left(\frac{e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_1}}{e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_0} + e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_1}}\right) & y_p = 1 \end{cases} \\
&= -\frac{1}{P}\sum_{p=1}^{P} \begin{cases} \log\left(\frac{(e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_0}) \cdot 1}{(e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_0})\left(1 + \frac{e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_1}}{e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_0}}\right)}\right) & y_p = 0 \\ \log\left(\frac{(e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_1}) \cdot 1}{(e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_1})\left(1 + \frac{e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_0}}{e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_1}}\right)}\right) & y_p = 1 \end{cases} \\
&= -\frac{1}{P}\sum_{p=1}^{P} \begin{cases} \log\left(\frac{1}{1 + \frac{e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_1}}{e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_0}}}\right) & y_p = 0 \\ \log\left(\frac{1}{1 + \frac{e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_0}}{e^{\dot{\mathbf{x}}_p^\intercal \mathbf{w}_1}}}\right) & y_p = 1 \end{cases} \\
&= -\frac{1}{P}\sum_{p=1}^{P} \begin{cases} \log\left(\frac{1}{1 + e^{\dot{\mathbf{x}}_p^\intercal (\mathbf{w}_1 - \mathbf{w}_0)}}\right) & y_p = 0 \\ \log\left(\frac{1}{1 + e^{\dot{\mathbf{x}}_p^\intercal (\mathbf{w}_0 - \mathbf{w}_1)}}\right) & y_p = 1. \end{cases}
\end{aligned}$$

Recall the definition of the sigmoid function, $\sigma(x) = \frac{1}{1+e^{-x}}$. Then we obtain:

$$
\begin{aligned}
g^{(2)}(\mathbf{w}_0, \mathbf{w}_1) &= -\frac{1}{P}\sum_{p=1}^{P} \begin{cases} \log\left(\frac{1}{1+e^{\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_1-\mathbf{w}_0)}}\right) & y_p = 0 \\ \log\left(\frac{1}{1+e^{\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_0-\mathbf{w}_1)}}\right) & y_p = 1 \end{cases} \\
&= -\frac{1}{P}\sum_{p=1}^{P} \begin{cases} \log\left(\sigma(-\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_1-\mathbf{w}_0))\right) & y_p = 0 \\ \log\left(\sigma(-\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_0-\mathbf{w}_1))\right) & y_p = 1 \end{cases} \\
&= -\frac{1}{P}\sum_{p=1}^{P} \begin{cases} \log\left(\sigma(\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_0-\mathbf{w}_1))\right) & y_p = 0 \\ \log\left(\sigma(\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_1-\mathbf{w}_0))\right) & y_p = 1. \end{cases}
\end{aligned}
$$

Let $\mathbf{w}_h = \mathbf{w}_1 - \mathbf{w}_0$. Then simplifying and applying the sigmoid identity $\sigma(-x) = 1 - \sigma(x)$, we have:

$$
\begin{aligned}
g^{(2)}(\mathbf{w}_0, \mathbf{w}_1) &= -\frac{1}{P}\sum_{p=1}^{P} \begin{cases} \log\left(\sigma(-\dot{\mathbf{x}}_p^\intercal\mathbf{w}_h)\right) & y_p = 0 \\ \log\left(\sigma(\dot{\mathbf{x}}_p^\intercal\mathbf{w}_h\right) & y_p = 1 \end{cases} \\
&= -\frac{1}{P}\sum_{p=1}^{P} \begin{cases} \log\left(1-\sigma(\dot{\mathbf{x}}_p^\intercal\mathbf{w}_h)\right) & y_p = 0 \\ \log\left(\sigma(\dot{\mathbf{x}}_p^\intercal\mathbf{w}_h)\right) & y_p = 1. \end{cases}
\end{aligned}
$$

Using the values of y_p, we express the step function inline:

$$
\begin{aligned}
g^{(2)}(\mathbf{w}_0, \mathbf{w}_1) &= -\frac{1}{P}\sum_{p=1}^{P} \mathbb{1}\Big[y_p = 1\Big]\Big(\log\left(\sigma(\dot{\mathbf{x}}_p^\intercal\mathbf{w}_h)\right)\Big) + \mathbb{1}\Big[y_p = 0\Big]\Big(\log\left(1-\sigma(\dot{\mathbf{x}}_p^\intercal\mathbf{w}_h)\right)\Big) \\
&= -\frac{1}{P}\sum_{p=1}^{P} y_p\Big(\log\left(\sigma(\dot{\mathbf{x}}_p^\intercal\mathbf{w}_h)\right)\Big) + (1-y_p)\Big(\log\left(1-\sigma(\dot{\mathbf{x}}_p^\intercal\mathbf{w}_h)\right)\Big) \\
&= -\frac{1}{P}\sum_{p=1}^{P} y_p \log\left(\sigma(\dot{\mathbf{x}}_p^\intercal\mathbf{w}_h)\right) + (1-y_p)\log\left(1-\sigma(\dot{\mathbf{x}}_p^\intercal\mathbf{w}_h)\right).
\end{aligned}
$$

So we have shown:

$$
\begin{aligned}
g^{(2)}(\mathbf{w}_0, \mathbf{w}_1) &= -\frac{1}{P}\sum_{p=1}^{P} y_p \log(\sigma(\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_1-\mathbf{w}_0))) + (1-y_p)\log(1-\sigma(\dot{\mathbf{x}}_p^\intercal(\mathbf{w}_1-\mathbf{w}_0))) \\
&= -\frac{1}{P}\sum_{p=1}^{P} y_p \log(\sigma(\dot{\mathbf{x}}_p^\intercal\mathbf{w}_h)) + (1-y_p)\log(1-\sigma(\dot{\mathbf{x}}_p^\intercal\mathbf{w}_h)) \\
&= g^{\diamondsuit}(\mathbf{w}_h).
\end{aligned}
$$

Then for $C = 2$, $y_p \in \{0, 1\}$, and $\mathbf{w}_h = \mathbf{w}_1 - \mathbf{w}_0$, we have shown $g^{(2)}(\mathbf{w}_0, \mathbf{w}_1) = g^{\diamondsuit}(\mathbf{w}_h)$.

$\blacksquare$