Gradient Matrices Calculations for a1.mlp.3.layers.py code for Assignment 1

Eric Leifer, James Troendle; November 15, 2021

We compute the gradient matrices for the three backpropagation steps for the **a1.mlp.3.layers.py** code by Hui Xue. In the **a1.mlp.3.layers.py** code, these are computed on lines 134-146 as $dz3, dW3, db3, da2, dz2, dW2, db2, da1, dz1, dW1, db1$.

Let B = batch size. For $i = 0, 1, \ldots, B-1$, each input observation $X^{[i]}$ is a $1 \times n_0$ matrix where $n_0 = 28 \times 28 = 284$ corresponding to the 284 pixels for digit in the MNIST data. There are $K = 10$ output neurons corresponding to the 10 one hot labels which gives the nine 0's and one 1 representation of the actual digit between 0 and 9. There are two hidden layers. There are $n_1 = 200$ neurons in the first hidden layer and $n_2 = 100$ neurons in the second hidden layer.

The loss function is the cross-entropy loss averaged over the observations in the batch. Let y be the $B \times 10$ matrix of one hot labels and $y^{(i)}$ be the i^{th} row of y for $i = 0, 1, \ldots, B-1$. Let $\hat{y}$ be the $B \times 10$ estimated probabilities of each of the hot labels and $\hat{y}_i^{(i)}$ be the rows of $\hat{y}$. Let θ be the weights and biases which we want to optimize with respect to the loss function. The loss function is

$$L(y, \hat{y}; \theta) = \frac{1}{B} \sum_{i=0}^{B-1} L_{CE}(y^{(i)}, \hat{y}^{(i)}) \tag{1}$$

where in (1) we compactly write the 10×1 vector $\hat{y}^{(i)} = (\hat{y}_0^{(i)}, \ldots, \hat{y}_9^{(i)}) = \hat{y}^{(i)}(X^{(i)}, \theta)$ and

$$L_{CE}(y^{(i)}, \hat{y}^{(i)}; \theta) = -\sum_{j=0}^{K-1} y_j^{(i)} \log(\hat{y}_j^{(i)}) \tag{2}$$

Note 1. Below, we will often suppress the dependence of L_{CE} on the observation $i = 0, 1, \ldots, B-1$. However, it will be understood that L_{CE} is a function of a particular observation i. Sometimes to emphasize that dependence, we will write $L_{CE}^{[i]}$.

First backpropagation step going from the output layer to the second hidden layer:

We compute the gradient matrices $dz3$, $dW3$, $db3$ on lines 134-136 of **a1.mlp.3.layers.py**. For observation $i = 0, 1, \ldots, B-1$,

$$z^{[3,i]} = a^{[2,i]} W^{[3]} + b^{[3]} \tag{3}$$

$$\hat{y}^{(i)} = a^{[3,i]} = \text{softmax}(z^{[3,i]}) \tag{4}$$

where $z^{[3,i]}$ is a $1 \times K$ matrix, $a^{[2,i]}$ is a $1 \times n_2$ matrix, $W^{[3]}$ is a $n_2 \times K$ matrix, $b^{[3,i]}$ is a $1 \times K$ matrix, and $\hat{y}^{(i)} = a^{[3,i]}$ is a $1 \times K$ matrix. From (3)-(4), we have the mappings

$$z^{[3,i]} \longrightarrow L_{CE}\big[y, \hat{y}(z^{[3,i]})\big] \quad \text{which maps} \quad \mathbb{R}^K \longrightarrow \mathbb{R}$$

$$W^{[3]} \longrightarrow z^{[3,i]} \quad \text{which maps} \quad \mathbb{R}^{n_2 \times K} \longrightarrow \mathbb{R}^K$$

$$b^{[3]} \longrightarrow z^{[3,i]} \quad \text{which maps} \quad \mathbb{R}^K \longrightarrow \mathbb{R}^K$$

We compute $\dfrac{\partial L_{CE}}{\partial W^{[3]}}$ and $\dfrac{\partial L_{CE}}{\partial b^{[3]}}$ using the tensor methodology described in Johnson (2017). $\dfrac{\partial L_{CE}}{\partial W^{[3]}}$ is a $1 \times (n_2 \times K)$ tensor, although we consider it to be a $n_2 \times K$ matrix. Similarly $\dfrac{\partial L_{CE}}{\partial b^{[3]}}$ is a $1 \times K$ matrix.

By the chain rule for tensors described in Johnson,

$$\frac{\partial L_{CE}}{\partial W^{[3]}} = \frac{\partial L_{CE}}{\partial z^{[3,i]}} \cdot \frac{\partial z^{[3,i]}}{\partial W^{[3]}}, \qquad \frac{\partial L_{CE}}{\partial b^{[3]}} = \frac{\partial L_{CE}}{\partial z^{[3,i]}} \cdot \frac{\partial z^{[3,i]}}{\partial b^{[3]}} \tag{5}$$

In (5), $\dfrac{\partial L_{CE}}{\partial z^{[3,i]}}$ is a $1 \times K$ matrix, $\dfrac{\partial z^{[3,i]}}{\partial W^{[3]}}$ is a $K \times (n_2 \times K)$ tensor, $\dfrac{\partial L_{CE}}{\partial b^{[3]}}$ is a $1 \times K$ matrix, and $\dfrac{\partial z^{[3,i]}}{\partial b^{[3]}}$ is a $K \times K$ matrix. From Johnson (2017), the matrix entries of (5) are for $r = 0, 1, \ldots, n_2 - 1$ and $j = 0, 1, \ldots, K - 1$:

$$\left(\frac{\partial L_{CE}}{\partial W^{[3]}}\right)_{(r,j)} = \sum_{t=0}^{K-1} \left(\frac{\partial L_{CE}}{\partial z^{[3,i]}}\right)_{1,t} \left(\frac{\partial z^{[3,i]}}{\partial W^{[3]}}\right)_{t,(r,j)}, \quad \left(\frac{\partial L_{CE}}{\partial b^{[3]}}\right)_{(1,j)} = \sum_{t=0}^{K-1} \left(\frac{\partial L_{CE}}{\partial z^{[3,i]}}\right)_{1,t} \left(\frac{\partial z^{[3,i]}}{\partial b^{[3]}}\right)_{(t,j)} \tag{6}$$

To compute (6), we use the following lemmas.

Lemma 1. ***Derivative of softmax***

Let $z = (z_0, z_2, \ldots, z_{K-1})$ *and* $softmax_j(z) = \sigma_j(z) = \dfrac{e^{z_j}}{\sum_{t=0}^{K-1} e^{z_t}}$. *Then*

$$\frac{d\sigma_j(z)}{dz_j} = \sigma_j(z)\big[1 - \sigma_j(z)\big] \tag{7}$$

Proof.

$$\frac{d\sigma_j(z)}{dz_j} = \frac{e^{z_j} \sum e^{z_t} - e^{2z_j}}{\left(\sum e^{z_t}\right)^2} = \sigma_j(z) - \big(\sigma_j(z)\big)^2 = \sigma_j(z)\big[1 - \sigma_j(z)\big] \qquad \text{QED} \tag{8}$$

Lemma 2. *For* $i = 0, 1, \ldots, B - 1$ *and* $j = 0, 1, \ldots, K - 1$,

$$\left(\frac{\partial L_{CE}}{\partial z^{[3,i]}}\right)_{1,j} = \hat{y}_j^{(i)} - y_j^{(i)} \tag{9}$$

Proof. By definition of L_{CE} in (2) and writing $z_j = z_j^{[3,i]}$, $\hat{y} = \hat{y}^{(i)}$, and $y = y^{(i)}$,

$$\frac{\partial L_{CE}(y, \hat{y}(z))}{\partial z_j} = -\sum_{t=0}^{K-1} \frac{y_t}{\hat{y}_t(z)} \cdot \frac{d}{dz_j}\big[\hat{y}_t(z)\big] \tag{10}$$

To evaluate $\dfrac{d}{dz_j}\big[\hat{y}_t(z)\big]$ in (10), we consider two situations: $t = j$ and $t \neq j$. First suppose $t = j$. Since $\hat{y}_j(z) = \text{softmax}_j(z)$, it follows from Lemma 1 that

$$\frac{d}{dz_j}\big[\hat{y}_j(z)\big] = \hat{y}_j(z)\big[1 - \hat{y}_j(z)\big] \tag{11}$$

Next suppose $t \neq j$. Then

$$\frac{d}{dz_j}\big[\hat{y}_t(z)\big] = e^{z_t} \cdot \frac{d}{dz_j}\Big[\Big(\sum_{u=0}^{K-1} e^{z_u}\Big)^{-1}\Big] = -e^{z_t} \cdot \frac{e^{z_j}}{\Big(\sum_{u=0}^{K-1} e^{z_u}\Big)^2} = -\hat{y}_t(z) \cdot \hat{y}_j(z). \tag{12}$$

We have from (10) and the sum of the one-hot-labels $\sum_{t=0}^{K-1} y_t = 1$ that

$$\begin{aligned}
\frac{\partial L_{CE}(y, \hat{y})}{\partial z_j} &= -\sum_{t=0}^{K-1} \frac{y_t}{\hat{y}_t}\Big[\hat{y}_j(1-\hat{y}_j)I(t=j) - \hat{y}_t\hat{y}_j I(t \neq j)\Big] \\
&= -\sum_{t=0}^{K-1} \Big[y_j(1-\hat{y}_j)I(t=j) - y_t\hat{y}_j I(t \neq j)\Big] \\
&= -y_j + y_j\hat{y}_j + \sum_{t=0}^{K-1} y_t\hat{y}_j I(t \neq j) = -y_j + \hat{y}_j \sum_{t=0}^{K-1} y_t = -y_j + \hat{y}_j \cdot 1 = \hat{y}_j - y_j \qquad \text{QED}
\end{aligned} \tag{13}$$

Before giving the next lemma, we establish some notation. For $j = 0, 1, ..., K-1$, let $z_j^{[3,i]}$ denote the j^{th} entry of $z^{[3,i]}$ and $a_j^{[2,i]}$ the j^{th} entry of $a_j^{[2,i]}$. Let $w_{rj}^{[3]}$ denote the (r, j) entry of $W^{[3]}$, and $b_j^{[3]}$ the j^{th} entry of $b^{[3]}$.

Lemma 3. *For $i = 0, 1, \ldots, B-1$; $t, j = 0, 1, \ldots, K-1$; and $r = 0, 1, \ldots, n_2 - 1$,*

$$\left(\frac{\partial z^{[3,i]}}{\partial W^{[3]}}\right)_{t,(r,j)} = I(t=j) \cdot a_r^{[2,i]}, \quad \left(\frac{\partial z^{[3,i]}}{\partial b^{[3]}}\right)_{(t,j)} = I(t=j) \tag{14}$$

where $I(t = j)$ is the indicator function which equals 1 when $t = j$ and equals 0 when $t \neq j$.

Proof. From (3), for $j = 0, 1, \ldots, K-1$, the j^{th} entry of $z^{[3,i]}$ is

$$z_j^{[3,i]} = \sum_{q=0}^{n_2-1} a_q^{[2,i]} w_{qj}^{[3]} + b_j^{[3]} \tag{15}$$

Thus,

$$\left(\frac{\partial z^{[3,i]}}{\partial W^{[3]}}\right)_{t,(r,j)} = \frac{\partial}{\partial w_{rj}^{[3]}}\Big[z_t^{[3,i]}\Big] = \frac{\partial}{\partial w_{rj}^{[3]}}\left[\sum_{q=0}^{n_2-1} a_q^{[2,i]} w_{qt}^{[3]} + b_t^{[3]}\right] = I(t=j) \cdot a_r^{[2,i]} \tag{16}$$

and

$$\left(\frac{\partial z^{[3,i]}}{\partial W b[3]}\right)_{t,j)} = \frac{\partial}{\partial b_j^{[3]}}\Big[z_t^{[3,i]}\Big] = \frac{\partial}{\partial b_j^{[3]}}\left[\sum_{q=0}^{n_2-1} a_q^{[2,i]} w_{qt}^{[3]} + b_t^{[3]}\right] = I(t=j) \qquad \text{QED} \tag{17}$$

Lemma 4. *For $i = 0, 1, \ldots, B-1$; $t, j = 0, 1, \ldots, K-1$; and $r = 0, 1, \ldots, n_2 - 1$,*

$$\left(\frac{\partial L_{CE}}{\partial W^{[3]}}\right)_{(r,j)} = \left(\hat{y}_j^{(i)} - y_j^{(i)}\right) \cdot a_r^{[2,i]}, \quad \left(\frac{\partial L_{CE}}{\partial b^{[3]}}\right)_{(1,j)} = \hat{y}_j^{(i)} - y_j^{(i)} \tag{18}$$

Proof. This follows from (6) and Lemmas 2 and 3. QED.

Note 2. In lines 134-136 of Hui's **a1.mlp.3.layers.py** program:

1. By (1), dz_3 is a $B \times K$ matrix with (i, j) entry given by (9).

2. By (1) and (18), dW_3 is a $n_2 \times K$ matrix with (r, j) entry

$$B \cdot \left(\frac{\partial L}{\partial W^{[3]}}\right)_{r,j} = \sum_{i=0}^{B-1} \left(\frac{\partial L_{CE}^{[i]}}{\partial W^{[3]}}\right)_{(r,j)} = \sum_{i=0}^{B-1} \left(\hat{y}_j^{(i)} - y_j^{(i)}\right) \cdot a_r^{[2,i]} \tag{19}$$

3. By (1) and (18), $db3$ is a $1 \times K$ matrix with $(1, j)$ entry

$$B \cdot \left(\frac{\partial L}{\partial Wb[3]}\right)_{r,j} = \sum_{i=0}^{B-1} \left(\frac{\partial L_{CE}^{[i]}}{\partial b^{[3]}}\right)_{(1,j)} = \sum_{i=0}^{B-1} \left(\hat{y}_j^{(i)} - y_j^{(i)}\right) \tag{20}$$

Second backpropagation step going from the second to the first hidden layer: We compute the gradient matrices $da2$, $dz2$, $dW2$, $db2$ on lines 138-141 of **a1.mlp.3.layers.py**. For observation $i = 0, 1, \ldots, B-1$, We have

$$z^{[2,i]} = a^{[1,i]} W^{[2]} + b^{[2]} \tag{21}$$

$$a^{[2,i]} = \text{sigmoid}(z^{[2,i]}) \tag{22}$$

where $z^{[2,i]}$ is a $1 \times n_2$ matrix, $a^{[1,i]}$ is a $1 \times n_1$ matrix, $W^{[2]}$ is a $n_1 \times n_2$ matrix, $b^{[2]}$ is a $1 \times n_2$ matrix, and $a^{[2,i]}$ is a $1 \times n_2$ matrix. For $q = 0, 1, \ldots, n_1 - 1$; $r = 0, 1, \ldots, n_2 - 1$, let $w_{qr}^{[2]}$ denote the (q, r) entry of $W^{[2]}$. We have the mappings

$$z^{[2,i]} \longrightarrow a^{[2,i]} \quad \text{which maps} \quad \mathbb{R}^{n_2} \longrightarrow \mathbb{R}^{n_2} \tag{23}$$

$$W^{[2]} \longrightarrow z^{[2,i]} \quad \text{which maps} \quad \mathbb{R}^{n_1 \times n_2} \longrightarrow \mathbb{R}^{n_2} \tag{24}$$

$$b^{[2]} \longrightarrow z^{[2,i]} \quad \text{which maps} \quad \mathbb{R}^{n_2} \longrightarrow \mathbb{R}^{n_2} \tag{25}$$

Applying the chain rule,

$$\frac{\partial L_{CE}}{\partial W^{[2]}} = \frac{\partial L_{CE}}{\partial z^{[3,i]}} \cdot \frac{\partial z^{[3,i]}}{\partial a^{[2,i]}} \cdot \frac{\partial a^{[2,i]}}{\partial z^{[2,i]}} \cdot \frac{\partial z^{[2,i]}}{\partial W^{[2]}}, \qquad \frac{\partial L_{CE}}{\partial b^{[2]}} = \frac{\partial L_{CE}}{\partial z^{[3,i]}} \cdot \frac{\partial z^{[3,i]}}{\partial a^{[2,i]}} \cdot \frac{\partial a^{[2,i]}}{\partial z^{[2,i]}} \cdot \frac{\partial z^{[2,i]}}{\partial b^{[2]}} \tag{26}$$

In (26), $\dfrac{\partial L_{CE}}{\partial W^{[2]}}$ is a $n_1 \times n_2$ matrix, $\dfrac{\partial L_{CE}}{\partial z^{[3,i]}}$ is a $1 \times K$ matrix, $\dfrac{\partial z^{[3,i]}}{\partial a^{[2,i]}}$ is a $K \times n_2$ matrix, $\dfrac{\partial a^{[2,i]}}{\partial z^{[2,i]}}$ is an $n_2 \times n_2$ matrix, $\dfrac{\partial z^{[2,i]}}{\partial W^{[2]}}$ is a $n_2 \times (n_1 \times n_2)$ tensor, $\dfrac{\partial L_{CE}}{\partial b^{[2]}}$ is a $1 \times n_2$ matrix, and $\dfrac{\partial z^{[2,i]}}{\partial b^{[2]}}$ is a $n_2 \times n_2$ matrix.

We next compute the entries of the $1 \times n_2$ matrix $\dfrac{\partial L_{CE}}{\partial a^{[2,i]}}$. We first see from (15), for $j = 0, 1, \ldots, K-1$ and $r = 0, 1, \ldots, n_2 - 1$,

$$\left(\frac{\partial z^{[3,i]}}{\partial a^{[2,i]}}\right)_{(j,r)} = w^{[3]}_{r,j} \tag{27}$$

Thus, for $r = 0, 1, \ldots, n_2 - 1$,

$$\left(\frac{\partial L_{CE}}{\partial a^{[2,i]}}\right)_{(1,r)} = \sum_{j=0}^{K-1} \left(\frac{\partial L_{CE}}{\partial z^{[3,i]}}\right)_{(1,j)} \left(\frac{\partial z^{[3,i]}}{\partial a^{[2,i]}}\right)_{(j,r)} = \sum_{j=0}^{K-1} \left(\hat{y}^{(i)}_j - y^{(i)}_j\right) \cdot w^{[3]}_{r,j} \tag{28}$$

We next compute the entries of the $1 \times n_2$ matrix $\dfrac{\partial L_{CE}}{\partial z^{[2,i]}}$. To do this, we first compute $\dfrac{\partial a^{[2,i]}}{\partial z^{[2,i]}}$. For this, we state the following lemma whose proof is straightforward.

Lemma 5. *Let* $\sigma(x) = sigmoid(x) = \dfrac{1}{1+e^{-x}}$. *Then*

$$\frac{d}{dx}\sigma(x) = \sigma(x) \cdot \big[1 - \sigma(x)\big]$$

By Lemma 5, for $i = 0, 1, \ldots, B-1$; $r, q = 0, 1, \ldots, n_2 - 1$,

$$\left(\frac{\partial a^{[2,i]}}{\partial z^{[2,i]}}\right)_{r,q} = I(r = q) \cdot a^{[2,i]}_r \left(1 - a^{[2,i]}_r\right) \tag{29}$$

so

$$\left(\frac{\partial L_{CE}}{\partial z^{[2,i]}}\right)_{1,r} = \sum_{t=0}^{n_2-1} \left(\frac{\partial L_{CE}}{\partial a^{[2,i]}}\right)_{1,t} \left(\frac{\partial a^{[2,i]}}{\partial z^{[2,i]}}\right)_{t,r} = \left(\frac{\partial L_{CE}}{\partial a^{[2,i]}}\right)_{1,r} a^{[2,i]}_r \left(1 - a^{[2,i]}_r\right) \tag{30}$$

Finally, we compute the entries of the $n_1 \times n_2$ matrix $\dfrac{\partial L_{CE}}{\partial W^{[2]}}$ and the $1 \times n_2$ matrix $\dfrac{\partial L_{CE}}{\partial b^{[2]}}$. To do this, we first compute $\dfrac{\partial z^{[2,i]}}{\partial W^{[2]}}$ and $\dfrac{\partial z^{[2,i]}}{\partial b^{[2]}}$. From (21), for $i = 0, 1, \ldots, B-1$,

$$z^{[2,i]}_r = \sum_{q=0}^{n_1-1} a^{[1,i]}_q w^{[2]}_{qr} + b^{[2]}_r \tag{31}$$

Consequently,

$$\left(\frac{\partial z^{[2,i]}}{\partial W^{[2]}}\right)_{t,(q,r)} = \frac{\partial}{\partial w^{[2]}_{qr}} \left[z^{[2,i]}_t\right] = I(t = r) \cdot a^{[1,i]}_q, \quad \left(\frac{\partial z^{[2,i]}}{\partial b^{[2]}}\right)_{t,r} = \frac{\partial}{\partial b^{[2]}_r} \left[z^{[2,i]}_t\right] = I(t = r) \tag{32}$$

Thus,

$$\left(\frac{\partial L_{CE}}{\partial W^{[2]}}\right)_{q,r} = \sum_{t=0}^{n_2-1} \left(\frac{\partial L_{CE}}{\partial z^{[2,i]}}\right)_{1,t} \left(\frac{\partial z^{[2,i]}}{\partial W^{[2]}}\right)_{t,(q,r)} = \left(\frac{\partial L_{CE}}{\partial z^{[2,i]}}\right)_{1,r} a^{[1,i]}_q \tag{33}$$

and

$$\left(\frac{\partial L_{CE}}{\partial b^{[2]}}\right)_{1,r} = \sum_{t=0}^{n_2-1}\left(\frac{\partial L_{CE}}{\partial z^{[2,i]}}\right)_{1,t}\left(\frac{\partial z^{[2,i]}}{\partial b^{[2]}}\right)_{t,r} = \left(\frac{\partial L_{CE}}{\partial z^{[2,i]}}\right)_{1,r} \tag{34}$$

Note 3. In lines 138-141 of Hui's **a1.mlp.3.layers.py** program:

1. By (1), da_2 is a $B \times n_2$ matrix with (i, r) entry given by (28).

2. By (1), dz_2 is a $B \times n_2$ matrix with (i, r) entry given by (30).

3. By (1) and (33), dW_2 is a $n_1 \times n_2$ matrix with (q, r) entry

$$B \cdot \left(\frac{\partial L}{\partial W^{[2]}}\right)_{r,j} = \sum_{i=0}^{B-1}\left(\frac{\partial L_{CE}^{[i]}}{\partial W^{[2]}}\right)_{r,j} = \sum_{i=0}^{B-1}\left(\frac{\partial L_{CE}}{\partial z^{[2,i]}}\right)_{1,r} a_q^{[1,i]} \tag{35}$$

4. By (1) and (34), $db2$ is a $1 \times n_2$ matrix with $(1, r)$ entry

$$B \cdot \left(\frac{\partial L}{\partial b^{[2]}}\right)_{1,r} = \sum_{i=0}^{B-1}\left(\frac{\partial L_{CE}^{[i]}}{\partial b^{[2]}}\right)_{1,r} = \sum_{i=0}^{B-1}\left(\frac{\partial L_{CE}}{\partial z^{[2,i]}}\right)_{1,r} \tag{36}$$

Third backpropagation step going from the first hidden layer to the input layer: We compute the gradient matrices $da1$, $dz1$, $dW1$, $db1$ on lines 143-146 of **a1.mlp.3.layers.py**. The computations are very similar to the second backpropagation step. For $i = 0, 1, \ldots, B-1$,

$$z^{[1,i]} = X^{[i]}W^{[1]} + b^{[1]} \tag{37}$$

$$a^{[1,i]} = \text{sigmoid}(z^{[1]}) \tag{38}$$

where $z^{[1,i]}$ is a $1 \times n_1$ matrix, $X^{[i]}$ is a $1 \times n_0$ matrix of input data, $W^{[1]}$ is a $n_0 \times n_1$ matrix, $b^{[1]}$ is a $1 \times n_1$ matrix, and $a^{[1,i]}$ is a $B \times n_1$ matrix. For $q = 0, 1, \ldots, n_0 - 1$; $r = 0, 1, \ldots, n_1 - 1$, let $x_q^{[i]}$ denote the (i, q) entry of X and $w_{qr}^{[1]}$ denote the (q, r) entry of $W^{[1]}$. We have the mappings

$$z^{[1,i]} \longrightarrow a^{[1,i]} \quad \text{which maps} \quad \mathbb{R}^{n_1} \longrightarrow \mathbb{R}^{n_1} \tag{39}$$

$$W^{[1]} \longrightarrow z^{[1,i]} \quad \text{which maps} \quad \mathbb{R}^{n_0 \times n_1} \longrightarrow \mathbb{R}^{n_1} \tag{40}$$

$$b^{[1]} \longrightarrow z^{[1,i]} \quad \text{which maps} \quad \mathbb{R}^{n_1} \longrightarrow \mathbb{R}^{n_1} \tag{41}$$

Applying the chain rule,

$$\frac{\partial L_{CE}}{\partial W^{[1]}} = \frac{\partial L_{CE}}{\partial z^{[2,i]}} \cdot \frac{\partial z^{[2,i]}}{\partial a^{[1,i]}} \cdot \frac{\partial a^{[1,i]}}{\partial z^{[1,i]}} \cdot \frac{\partial z^{[1,i]}}{\partial W^{[1]}}, \qquad \frac{\partial L_{CE}}{\partial b^{[1]}} = \frac{\partial L_{CE}}{\partial z^{[2,i]}} \cdot \frac{\partial z^{[2,i]}}{\partial a^{[1,i]}} \cdot \frac{\partial a^{[1,i]}}{\partial z^{[1,i]}} \cdot \frac{\partial z^{[1,i]}}{\partial b^{[1]}} \tag{42}$$

In (42), $\dfrac{\partial L_{CE}}{\partial W^{[1]}}$ is a $n_0 \times n_1$ matrix, $\dfrac{\partial L_{CE}}{\partial z^{[2,i]}}$ is a $1 \times n_2$ matrix, $\dfrac{\partial z^{[2,i]}}{\partial a^{[1,i]}}$ is a $n_2 \times n_1$ matrix, $\dfrac{\partial a^{[1,i]}}{\partial z^{[1,i]}}$ is an $n_1 \times n_1$ matrix, $\dfrac{\partial z^{[1,i]}}{\partial W^{[1]}}$ is a $n_1 \times (n_0 \times n_1)$ tensor, $\dfrac{\partial L_{CE}}{\partial b^{[1]}}$ is a $1 \times n_1$ matrix, and $\dfrac{\partial z^{[1,i]}}{\partial b^{[1]}}$ is a $n_1 \times n_1$ matrix.

We next compute the entries of the $1 \times n_1$ matrix $\dfrac{\partial L_{CE}}{\partial a^{[1,i]}}$. We first see from (31), for $j = 0, 1, \ldots, n_2 - 1$ and $r = 0, 1, \ldots, n_1 - 1$,

$$\left(\frac{\partial z^{[2,i]}}{\partial a^{[1,i]}}\right)_{(j,r)} = w^{[1]}_{r,j} \tag{43}$$

Thus, for $r = 0, 1, \ldots, n_1 - 1$,

$$\left(\frac{\partial L_{CE}}{\partial a^{[1,i]}}\right)_{(1,r)} = \sum_{j=0}^{n_2-1} \left(\frac{\partial L_{CE}}{\partial z^{[2,i]}}\right)_{(1,j)} \left(\frac{\partial z^{[2,i]}}{\partial a^{[1,i]}}\right)_{(j,r)} = \sum_{j=0}^{n_2-1} \left(\frac{\partial L_{CE}}{\partial z^{[2,i]}}\right)_{(1,j)} \cdot w^{[1]}_{r,j} \tag{44}$$

We next compute the entries of the $1 \times n_2$ matrix $\dfrac{\partial L_{CE}}{\partial z^{[1,i]}}$. To do this, we first compute $\dfrac{\partial a^{[1,i]}}{\partial z^{[1,i]}}$. By Lemma 5, for $i = 0, 1, \ldots, B-1$; $r, q = 0, 1, \ldots, n_1 - 1$,

$$\left(\frac{\partial a^{[1,i]}}{\partial z^{[1,i]}}\right)_{r,q} = I(r = q) \cdot a^{[1,i]}_r \left(1 - a^{[1,i]}_r\right) \tag{45}$$

so

$$\left(\frac{\partial L_{CE}}{\partial z^{[1,i]}}\right)_{1,r} = \sum_{t=0}^{n_1-1} \left(\frac{\partial L_{CE}}{\partial a^{[1,i]}}\right)_{1,t} \left(\frac{\partial a^{[1,i]}}{\partial z^{[1,i]}}\right)_{t,r} = \left(\frac{\partial L_{CE}}{\partial a^{[1,i]}}\right)_{1,r} a^{[1,i]}_r \left(1 - a^{[1,i]}_r\right) \tag{46}$$

Finally, we compute the entries of the $n_0 \times n_1$ matrix $\dfrac{\partial L_{CE}}{\partial W^{[1]}}$ and the $1 \times n_1$ matrix $\dfrac{\partial L_{CE}}{\partial b^{[1]}}$. To do this, we first compute $\dfrac{\partial z^{[1,i]}}{\partial W^{[1]}}$ and $\dfrac{\partial z^{[1,i]}}{\partial b^{[1]}}$. From (37), for $i = 0, 1, \ldots, B-1$,

$$z^{[1,i]}_r = \sum_{q=0}^{n_0-1} x^{[i]}_q w^{[1]}_{qr} + b^{[1]}_r \tag{47}$$

Consequently,

$$\left(\frac{\partial z^{[1,i]}}{\partial W^{[1]}}\right)_{t,(q,r)} = \frac{\partial}{\partial w^{[1]}_{qr}} \left[z^{[1,i]}_t\right] = I(t = r) \cdot x^{[i]}_q, \quad \left(\frac{\partial z^{[1,i]}}{\partial b^{[1]}}\right)_{t,r} = \frac{\partial}{\partial b^{[1]}_r} \left[z^{[1,i]}_t\right] = I(t = r) \tag{48}$$

Thus,

$$\left(\frac{\partial L_{CE}}{\partial W^{[1]}}\right)_{q,r} = \sum_{t=0}^{n_1-1} \left(\frac{\partial L_{CE}}{\partial z^{[1,i]}}\right)_{1,t} \left(\frac{\partial z^{[1,i]}}{\partial W^{[1]}}\right)_{t,(q,r)} = \left(\frac{\partial L_{CE}}{\partial z^{[1,i]}}\right)_{1,r} x^{[i]}_q \tag{49}$$

and

$$\left(\frac{\partial L_{CE}}{\partial b^{[1]}}\right)_{r} = \sum_{t=0}^{n_1-1} \left(\frac{\partial L_{CE}}{\partial z^{[1,i]}}\right)_{1,t} \left(\frac{\partial z^{[1,i]}}{\partial b^{[1]}}\right)_{t,r} = \left(\frac{\partial L_{CE}}{\partial z^{[1,i]}}\right)_{1,r} \tag{50}$$

Note 4. In lines 143-146 of Hui's **a1.mlp.3.layers.py** program:

1. By (1), da_1 is a $B \times n_1$ matrix with (i, r) entry given by (44).

2. By (1), dz_1 is a $B \times n_1$ matrix with (i, r) entry given by (46).

3. By (1) and (49), dW_1 is a $n_0 \times n_1$ matrix with (q, r) entry

$$B \cdot \left(\frac{\partial L}{\partial W^{[1]}}\right)_{r,j} = \sum_{i=0}^{B-1} \left(\frac{\partial L_{CE}^{[i]}}{\partial W^{[1]}}\right)_{r,j} = \sum_{i=0}^{B-1} \left(\frac{\partial L_{CE}}{\partial z^{[1,i]}}\right)_{1,r} x_q^{[i]} \tag{51}$$

4. By (1) and (50), $db1$ is a $1 \times n_1$ matrix with $(1, r)$ entry

$$B \cdot \left(\frac{\partial L}{\partial b^{[1]}}\right)_{1,r} = \sum_{i=0}^{B-1} \left(\frac{\partial L_{CE}^{[i]}}{\partial b^{[2]}}\right)_{1,r} = \sum_{i=0}^{B-1} \left(\frac{\partial L_{CE}}{\partial z^{[1,i]}}\right)_{1,r} \tag{52}$$

Note 5. Lines 156-159 introduce L^2 regularization. We justify line 157 with lines 158-159 being similar. This corresponds to part (d) from Problem 5 in Assignment 1 where the loss function with L^2 regularization is defined as

$$L^\lambda \equiv L + \lambda\Big(||W^{[1]}||_2^2 + ||W^{[2]}||_2^2 + ||W^{[3]}||_2^2\Big) \tag{53}$$

For $r = 0, 1, \ldots, n_0 - 1$ and $j = 0, 1, \ldots, n_1 - 1$,

$$\left(\frac{\partial L^\lambda}{\partial W^{[1]}}\right)_{r,j} = \left(\frac{\partial L}{\partial W^{[1]}}\right)_{r,j} + \lambda \frac{\partial}{\partial w_{rj}^{[1]}}\left[\sum_{q=0}^{n_0-1}\sum_{k=0}^{n_1-1} \left(w_{qk}^{[1]}\right)^2\right] = \left(\frac{\partial L}{\partial W^{[1]}}\right)_{r,j} + 2\lambda w_{rj}^{[1]} \tag{54}$$

Thus,

$$\frac{\partial L^\lambda}{\partial W^{[1]}} = \frac{\partial L}{\partial W^{[1]}} + 2\lambda W^{[1]} \tag{55}$$

References

1. Johnson J. Derivatives, Backpropagation, and Vectorization. 2017.