Problem 2 (Neural Net Optimization)

In this problem, we will take a closer look at how gradients are calculated for backprop with a simple multi-layer perceptron (MLP). The MLP will consist of a first fully connected layer with a sigmoid activation, followed by a one-dimensional, second fully connected layer with a sigmoid activation to get a prediction for a binary classification problem. Assume bias has not been merged. Let:

- $\mathbf{W}_1$ be the weights of the first layer, $\mathbf{b}_1$ be the bias of the first layer.
- $\mathbf{W}_2$ be the weights of the second layer, $\mathbf{b}_2$ be the bias of the second layer.

The described architecture can be written mathematically as:

$$\hat{y} = \sigma(\mathbf{W}_2 \left[\sigma \left(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1\right)\right] + \mathbf{b}_2)$$

where $\hat{y}$ is a scalar output of the net when passing in the single datapoint $\mathbf{x}$ (represented as a column vector), the additions are element-wise additions, and the sigmoid is an element-wise sigmoid.

1. Let:
   - N be the number of datapoints we have
   - M be the dimensionality of the data
   - H be the size of the hidden dimension of the first layer. Here, hidden dimension is used to describe the dimension of the resulting value after going through the layer. Based on the problem description, the hidden dimension of the second layer is 1.

   Write out the dimensionality of each of the parameters, and of the intermediate variables:

   $$\mathbf{a}_1 = \mathbf{W}_1 \mathbf{x} + \mathbf{b}_1, \qquad \mathbf{z}_1 = \sigma(\mathbf{a}_1)$$
   $$a_2 = \mathbf{W}_2 \mathbf{z}_1 + \mathbf{b}_2, \qquad \hat{y} = z_2 = \sigma(a_2)$$

   and make sure they work with the mathematical operations described above.

2. We will derive the gradients for each of the parameters. The gradients can be used in gradient descent to find weights that improve our model's performance. For this question, assume there is only one datapoint $\mathbf{x}$, and that our loss is $L = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}))$. For all questions, the chain rule will be useful.
   (a) Find $\frac{\partial L}{\partial b_2}$.
   (b) Find $\frac{\partial L}{\partial W_2^h}$, where W_2^h represents the hth element of $\mathbf{W}_2$.
   (c) Find $\frac{\partial L}{\partial b_1^h}$, where b_1^h represents the hth element of $\mathbf{b}_1$. (*Hint: Note that only the hth element of $\mathbf{a}_1$ and $\mathbf{z}_1$ depend on b_1^h - this should help you with how to use the chain rule.)
   (d) Find $\frac{\partial L}{\partial W_1^{h,m}}$, where $W_1^{h,m}$ represents the element in row h, column m in $\mathbf{W}_1$.

Solution:

1. We know that $\mathbf{x}$ is a M-dimensional column vector. Our first hidden layer is a H-dimensional column vector. $\mathbf{b}_1$ is also a H-dimensional column vector. Therefore $\mathbf{W}_1$ must be $H \times M$ dimensional matrix such that $\mathbf{W}_1\mathbf{x} + \mathbf{b}_1$ evaluates to a H-dimensional column vector.

   - Therefore, for $\mathbf{a}_1 = \mathbf{W}_1\mathbf{x} + \mathbf{b}_1$, we must have that $\mathbf{a}_1$ is a H-dimensional column vector.
   - Because our application of the sigmoid function is pointwise, such that each element has the sigmoid function applied individually, then we know that, for $\mathbf{z}_1 = \sigma(\mathbf{a}_1)$, $\mathbf{z}_1$ must also be a H-dimensional column vector.

   We can rewrite our MLP architecture using the variables defined above:

   $$\hat{y} = \sigma(\mathbf{W}_2\mathbf{z}_1 + b_2)$$

   Because our second layer only has a single dimension, our second hidden layer is a scalar quantity. Therefore b_2 is also a scalar. Therefore, $\mathbf{W}_2$ must be a $1 \times H$ dimensional matrix (or a H-dimensional row vector) such that $\mathbf{W}_2\mathbf{z}_1 + b_2$ evaluates to a scalar quantity.

   - Therefore, for $a_2 = \mathbf{W}_2\mathbf{z}_1 + b_2$, we must have that a_2 is a scalar.
   - Because our application of the sigmoid function is pointwise, we know that, for $\hat{y} = z_2 = \sigma(a_2)$, z_2 must be a scalar and therefore $\hat{y}$ is also a scalar.

2. The loss function is given as

   $$L = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}))$$

   Recall that our MLP architecture is given mathematically by the equation

   $$\hat{y} = \sigma(\mathbf{W}_2[\sigma(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1)] + b_2)$$

   We can write this in terms of our intermediary variables as:

   $$\hat{y} = \sigma(a_2)$$

   (a) Applying the chain rule allows us to write:

   $$\frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_2} \frac{\partial a_2}{\partial b_2}$$

   By definition, we have $a_2 = \mathbf{W}_2\mathbf{z}_1 + b_2$ and $\frac{\partial a_2}{\partial b_2} = 1$.
   In the CS 181 textbook from pg 42 equation (3.14), the derivative of the logistic sigmoid function can be expressed as $\frac{\partial(a_2)}{\partial a_2} = \sigma(a_2)(1 - \sigma(a_2))$. Therefore

   $$\frac{\partial \hat{y}}{\partial a_2} \frac{\partial a_2}{\partial b_2} = \sigma(a_2)(1 - \sigma(a_2))$$

   Consider $\frac{\partial L}{\partial \hat{y}}$:

   $$\frac{\partial L}{\partial \hat{y}} = \frac{d}{d\hat{y}}[-(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}))]$$

   $$\frac{\partial L}{\partial \hat{y}} = -[\frac{d}{d\hat{y}}(y \log(\hat{y})) + \frac{d}{d\hat{y}}((1 - y) \log(1 - \hat{y}))]$$

   $$\frac{\partial L}{\partial \hat{y}} = -[\frac{y}{\hat{y}} - \frac{1 - y}{1 - \hat{y}}] = \frac{1 - y}{1 - \hat{y}} - \frac{y}{\hat{y}}$$

   $$\frac{\partial L}{\partial \hat{y}} = \frac{\hat{y} - y}{\hat{y}(1 - \hat{y})}$$

Therefore, combining our results gives

$$\frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_2} \frac{\partial a_2}{\partial b_2}$$

$$\frac{\partial L}{\partial b_2} = \frac{\hat{y} - y}{\hat{y}(1-\hat{y})} \big(\sigma(a_2)(1-\sigma(a_2))\big)$$

Recall that $\hat{y} = \sigma(a_2)$, so we may substitute and rewrite:

$$\frac{\partial L}{\partial b_2} = \frac{\hat{y} - y}{\hat{y}(1-\hat{y})} \big(\hat{y}(1-\hat{y})\big)$$

$$\boxed{\frac{\partial L}{\partial b_2} = \hat{y} - y}$$

or equivalently

$$\boxed{\frac{\partial L}{\partial b_2} = z_2 - y}$$

(b) Applying the chain rule allows us to write:

$$\frac{\partial L}{\partial W_2^h} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_2} \frac{\partial a_2}{\partial W_2^h}$$

Recall that we had already found that

$$\frac{\partial \hat{y}}{\partial a_2} = \sigma(a_2)(1-\sigma(a_2)) = \hat{y}(1-\hat{y})$$

and that

$$\frac{\partial L}{\partial \hat{y}} = \frac{\hat{y} - y}{\hat{y}(1-\hat{y})}$$

Therefore, we know that

$$\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_2} = \frac{\hat{y} - y}{\hat{y}(1-\hat{y})} \big(\hat{y}(1-\hat{y})\big) = \hat{y} - y$$

Now consider $\frac{\partial a_2}{\partial W_2^h}$ where $a_2 = \mathbf{W}_2 \mathbf{z}_1 + b_2$:

$$\frac{\partial a_2}{\partial W_2^h} = \frac{d}{dW_2^h} [\mathbf{W}_2 \mathbf{z}_1 + b_2]$$

Let z_1^h be the hth element of $\mathbf{z}_1$ such that the derivative evaluates to:

$$\frac{\partial a_2}{\partial W_2^h} = z_1^h$$

Therefore, combining our results gives

$$\frac{\partial L}{\partial W_2^h} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_2} \frac{\partial a_2}{\partial W_2^h}$$

$$\boxed{\frac{\partial L}{\partial W_2^h} = z_1^h (\hat{y} - y)}$$

or equivalently

$$\boxed{\frac{\partial L}{\partial W_2^h} = z_1^h (z_2 - y)}$$

(c) Applying the chain rule allows us to write:

$$\frac{\partial L}{\partial b_1^h} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_2} \frac{\partial a_2}{\partial b_1^h}$$

Consider applying the chain rule on the third term $\frac{\partial a_2}{\partial b_1^h}$.

$$\frac{\partial a_2}{\partial b_1^h} = \frac{\partial a_2}{\partial \mathbf{z}_1} \frac{\partial \mathbf{z}_1}{\partial \mathbf{a}_1} \frac{\partial \mathbf{a}_1}{\partial b_1^h}$$

Let a_1^h and z_1^h be the hth element of $\mathbf{a}_1$ and $\mathbf{z}_1$ respectively. Using the hint, since only the hth element of $\mathbf{a}_1$ and $\mathbf{z}_1$ depend on b_1^h, we can write this partial as (because all other terms will not contribute to the partial derivative):

$$\frac{\partial a_2}{\partial b_1^h} = \frac{\partial a_2}{\partial z_1^h} \frac{\partial z_1^h}{\partial a_1^h} \frac{\partial a_1^h}{\partial b_1^h}$$

Therefore, our desired partial derivative, factored by the chain rule, is:

$$\frac{\partial L}{\partial b_1^h} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_2} \frac{\partial a_2}{\partial z_1^h} \frac{\partial z_1^h}{\partial a_1^h} \frac{\partial a_1^h}{\partial b_1^h}$$

Recall that we had previously found that $\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_2} = \hat{y} - y$. Now, consider $\frac{\partial a_2}{\partial z_1^h}$ where $a_2 = \mathbf{W}_2 \mathbf{z}_1 + \mathbf{b}_2$:

$$\frac{\partial a_2}{\partial z_1^h} = \frac{d}{dz_1^h}[\mathbf{W}_2 \mathbf{z}_1 + \mathbf{b}_2] = W_2^h$$

Next, consider $\frac{\partial z_1^h}{\partial a_1^h}$ where $\mathbf{z}_1 = \sigma(\mathbf{a}_1)$. Since we use the point wise application of the sigmoid function, we also have that $z_1^h = \sigma(a_1^h)$. Recall that the derivative of the sigmoid function is: $\sigma' = \sigma(1 - \sigma)$. Therefore we have that:

$$\frac{\partial z_1^h}{\partial a_1^h} = \frac{d}{da_1^h}[\sigma(a_1^h)] = \sigma(a_1^h)(1 - \sigma(a_1^h))$$

or equivalently

$$\frac{\partial z_1^h}{\partial a_1^h} = z_1^h(1 - z_1^h)$$

Finally, consider $\frac{\partial a_1^h}{\partial b_1^h}$ where $\mathbf{a}_1 = \mathbf{W}_1 \mathbf{x} + \mathbf{b}_1$. Let $\mathbf{W}_1^h$ be the hth M-dimensional vector of $\mathbf{W}_1$ (corresponding to the hth row of $\mathbf{W}_1$) such that we can write

$$a_1^h = \mathbf{W}_1^h \mathbf{x} + b_1^h$$

Therefore, $\frac{\partial z_1^h}{\partial a_1^h} = 1$. Combining our results gives

$$\frac{\partial L}{\partial b_1^h} = (\hat{y} - y)(W_2^h)(z_1^h(1 - z_1^h))(1)$$

$$\boxed{\frac{\partial L}{\partial b_1^h} = W_2^h z_1^h (1 - z_1^h)(\hat{y} - y)}$$

or equivalently:

$$\boxed{\frac{\partial L}{\partial b_1^h} = W_2^h z_1^h (1 - z_1^h)(z_2 - y)}$$

(d) Applying the chain rule allows us to write:

$$\frac{\partial L}{\partial W_1^{h,m}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_2} \frac{\partial a_2}{\partial W_1^{h,m}}$$

Consider the third term, which can be expanded as:

$$\frac{\partial a_2}{\partial W_1^{h,m}} = \frac{\partial a_2}{\partial z_1^h} \frac{\partial z_1^h}{\partial a_1^h} \frac{\partial a_1^h}{\partial W_1^{h,m}}$$

Therefore our desired partial derivative, factored by the chain rule, is:

$$\frac{\partial L}{\partial W_1^{h,m}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_2} \frac{\partial a_2}{\partial z_1^h} \frac{\partial z_1^h}{\partial a_1^h} \frac{\partial a_1^h}{\partial W_1^{h,m}}$$

Let x^m be the mth element of $\mathbf{x}$. Recall that earlier we found that

$$a_1^h = \mathbf{W}_1^h \mathbf{x} + b_1^h$$

or equivalently:

$$a_1^h = \big(\sum_{m=1}^{M} W_1^{h,m} x^m \big) + b_1^h$$

Therefore we have that

$$\frac{\partial a_1^h}{\partial W_1^{h,m}} = x^m$$

Recall that we had found previously that $\frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial a_2} = \hat{y} - y$ and that $\frac{\partial a_2}{\partial z_1^h} \frac{\partial z_1^h}{\partial a_1^h} = (W_2^h)(z_1^h(1 - z_1^h))$. Combining our results gives

$$\frac{\partial L}{\partial W_1^{h,m}} = (\hat{y} - y)(W_2^h)(z_1^h(1 - z_1^h))(x^m)$$

$$\boxed{\frac{\partial L}{\partial W_1^{h,m}} = W_2^h x^m z_1^h (1 - z_1^h)(\hat{y} - y)}$$

or equivalently:

$$\boxed{\frac{\partial L}{\partial W_1^{h,m}} = W_2^h x^m z_1^h (1 - z_1^h)(z_2 - y)}$$