CS 1678/2078 HW Backprop

Abstract

In this assignment you will be computing the gradients of the weights of multi-layered neural network by hand. This serves as a precursor to part B of HW2 where you will be implementing backprop to train a multi-layered neural network. To submit this assignment, upload a `.pdf` to Gradescope containing your responses to the questions below. You are required to use LaTeX for your write up.

1 Partial Derivatives With a Single Sample (34 points)

Consider a neural network with two hidden layers and a linear output layer. The input to the network is a vector of length four, the first hidden layer has three hidden units, the second layer has two, and the last layer as a single unit. Each hidden layer uses the ReLU activation function.

For a single input x and target value $y \in \mathbb{R}$, the loss function for the network is

$$l(\theta) = \frac{1}{2}\left(f(x,\theta) - y\right)^2,$$

with each layer computing,

$$h^i = f^i(h^{i-1}, W^i) = \sigma\left(h^{i-1}{W^i}^\top\right),$$

where $h^i \in \mathbb{R}^{1\times n_i}$ and $W^i \in \mathbb{R}^{n_i \times n_{i-1}}$. Note that we dropped the dataset D in notation for the loss function $l_D(\theta)$. This just makes notation simpler for the assignment. Let the partial derivative of the loss with respect to $f(x,\theta)$ be δ, e.g.

$$\delta = \frac{\partial l(\theta)}{\partial f(x,\theta)} = f(x,\theta) - y$$

1. What is the partial derivative of $l(\theta)$ with respect to the weight $W^3_{1,1}$?

$$\begin{aligned}
\frac{\partial l(\theta)}{\partial W^3_{1,1}} &= \frac{\partial h^3}{\partial W^3_{1,1}} \frac{\partial l(\theta)}{\partial h^3} \\
&= \frac{\partial\left(h^2{W^3}^\top\right)}{\partial W^3_{1,1}}\delta \\
&= \delta\frac{\partial\left(h^2_{1,1}W^3_{1,1} + h^2_{1,2}W^3_{1,2}\right)}{\partial W^3_{1,1}} \\
&= \delta h^2_{1,1}
\end{aligned}$$

2. What is the partial derivative of $l(\theta)$ with respect to the weight $W^3_{1,2}$?

$$\begin{aligned}
\frac{\partial l(\theta)}{\partial W^3_{1,2}} &= \frac{\partial f^3(h^2, W^3)}{\partial W^3_{1,2}} \frac{\partial l(\theta)}{\partial f(x,\theta)} \\
&= \delta h^2_{1,2}
\end{aligned}$$

3. What are the partial derivatives of $l(\theta)$ with respect to W^3.

$$\begin{aligned}\frac{\partial l(\theta)}{\partial W^3} &= \begin{bmatrix}\frac{\partial l(\theta)}{\partial W^3_{1,1}} & \frac{\partial l(\theta)}{\partial W^3_{1,2}}\end{bmatrix} \\ &= \delta \begin{bmatrix}h^2_{1,1} & h^2_{1,2}\end{bmatrix} = \delta h^2\end{aligned}$$

4. What are the partial derivatives of $l(\theta)$ with respect to $h^2_{1,1}$.

$$\begin{aligned}\frac{\partial l(\theta)}{\partial h^2_{1,1}} &= \frac{\partial f^3(h^2, W^3)}{\partial h^2_{1,1}} \frac{\partial l(\theta)}{\partial f(x,\theta)} \\ &= \frac{\partial \left(h^2 {W^3}^\top\right)}{\partial h^2_{1,1}} \delta \\ &= \delta \frac{\partial \left(h^2_{1,1} W^3_{1,1} + h^2_{1,2} W^3_{1,2}\right)}{\partial h^2_{1,1}} \\ &= \delta W^3_{1,1}\end{aligned}$$

5. What are the partial derivatives of $l(\theta)$ with respect to h^2.

$$\begin{aligned}\frac{\partial l(\theta)}{\partial h^2} &= \begin{bmatrix}\frac{\partial l(\theta)}{\partial h^2_{1,1}} & \frac{\partial l(\theta)}{\partial h^2_{1,2}}\end{bmatrix} \\ &= \delta \begin{bmatrix}W^3_{1,1} & W^3_{1,2}\end{bmatrix} = \delta W^3\end{aligned}$$

6. What is the derivative for the ReLU activation function $\sigma(x) = \max(x, 0)$? You can use the notation that $x > y$ evaluates to 1 if true and 0 if false.

$$\frac{d\sigma(x)}{dx} = x \geq 0$$

ReLU is not differentiable at $x = 0$, but in practice we use the subderivative (which is the answer above) `https://en.wikipedia.org/wiki/Subderivative`. You could use $x > 0$ or $x \geq 0$ because both are valid subderivatives as $x = 0$.

7. What are the partial derivatives with respect to $W^2_{1,j}$ for $h^2_{1,1} = f^2_1(h^1, W^2)$? You may use $z^i = h^{i-1} {W^i}^\top$ and $z^i_{1,1} = h^{i-1} {W^i_{1,\cdot}}^\top$ to simplify your answer.

$$
\begin{aligned}
\frac{\partial h^2_{1,1}}{\partial W^2_{1,j}} &= \frac{\partial \sigma(z^2_{1,1})}{\partial W^2_{1,j}} \\
&= \frac{\partial z^2_{1,1}}{\partial W^2_{1,j}} \frac{\partial \sigma(z^2_{1,1})}{\partial z^2_{1,1}} \\
&= \frac{\partial z^2_{1,1}}{\partial W^2_{1,j}} (z^2_{1,1} \geq 0) \\
&= \frac{\partial (h^1_{1,1} W^2_{1,1} + h^1_{1,2} W^2_{1,2} + h^1_{1,3} W^2_{1,3})}{\partial W^2_{1,j}} (z^2_{1,1} \geq 0) \\
&= \frac{\partial h^1_{1,j} W^2_{1,j}}{\partial W^2_{1,j}} (z^2_{1,1} \geq 0) \\
&= h^1_{1,j} (z^2_{1,1} \geq 0) \\
&= (z^2_{1,1} \geq 0) h^1_{1,j} \text{ flipping sides for simpler connections in part 2}
\end{aligned}
$$

8. What are the partial derivatives with respect to $W^2_{2,j}$ for $h^2_{1,1} = f^2_1(h^1, W^2)$?

$$
\begin{aligned}
\frac{\partial h^2_{1,1}}{\partial W^2_{2,j}} &= \frac{\partial (h^1_{1,1} W^2_{1,1} + h^1_{1,2} W^2_{1,2} + h^1_{1,3} W^2_{1,3})}{\partial W^2_{1,j}} (z^2_{1,1} \geq 0) \\
&= 0 (z^2_{1,1} \geq 0) = 0
\end{aligned}
$$

9. What are the partial derivatives with respect to W^2 for $h^2_{1,1} = f^2_1(h^1, W^2)$?

$$
\frac{\partial h^2_{1,1}}{\partial W^2} = \begin{bmatrix} \frac{\partial h^2_{1,1}}{\partial W^2_{1,1}} & \frac{\partial h^2_{1,1}}{\partial W^2_{1,2}} & \frac{\partial h^2_{1,1}}{\partial W^2_{1,3}} \\ \frac{\partial h^2_{1,1}}{\partial W^2_{2,1}} & \frac{\partial h^2_{1,1}}{\partial W^2_{2,2}} & \frac{\partial h^2_{1,1}}{\partial W^2_{2,3}} \end{bmatrix} = (z^2_{1,1} \geq 0) \begin{bmatrix} h^1_{1,1} & h^1_{1,2} & h^1_{1,3} \\ 0 & 0 & 0 \end{bmatrix} = (z^2_{1,1} \geq 0) \begin{bmatrix} h^1_{1,\cdot} \\ 0 \end{bmatrix}
$$

10. What are the partial derivatives of $l(\theta)$ with respect to $W^2_{i,j}$? Note that using scalar notation we express h^3 as

$$
h^3_{1,1} = \sum_{q=1}^{n_2} h^2_{1,q} W^3_{1,q} = \sum_{q=1}^{n_2} \sigma \left(\sum_{r=1}^{n_1} h^1_{1,r} W^2_{q,r} \right) W^3_{1,q}.
$$

You can use this expression as a starting point for the derivative if you are not comfortable with linear algebra.

$$
\begin{aligned}
\frac{\partial l(\theta)}{W^2_{i,j}} &= \frac{\partial l(\theta)}{\partial h^3_{1,1}} \frac{\partial h^3_{1,1}}{\partial W^2_{i,j}} \\
&= \delta \frac{\partial}{\partial W^2_{i,j}} \sum_{q=1}^{n_2} h^2_{1,q} W^3_{1,q} \\
&= \delta \sum_{q=1}^{n_2} \frac{\partial h^2_{1,q} W^3_{1,q}}{\partial W^2_{i,j}} \\
&= \delta \sum_{q=1}^{n_2} \frac{\partial h^2_{1,q} W^3_{1,q}}{\partial h^2_{1,q}} \frac{\partial h^2_{1,q}}{\partial W^2_{i,j}} \\
&= \delta \sum_{q=1}^{n_2} W^3_{1,q} \qquad \underbrace{\frac{\partial h^2_{1,q}}{\partial W^2_{i,j}}}_{=0 \text{ if } q \neq i, \text{ see \#10}} \\
&= \delta W^3_{1,i} \frac{\partial h^2_{1,i}}{\partial W^2_{i,j}} \\
&= \delta W^3_{1,i} \left(z^2_{1,i} \geq 0\right) h^1_{1,j}
\end{aligned}
$$

11. What are the partial derivatives with respect to W^2 for $l(\theta)$?

$$
\begin{aligned}
\frac{\partial l(\theta)}{\partial W^2} &= \begin{bmatrix} \frac{\partial l(\theta)}{\partial W^2_{1,1}} & \frac{\partial l(\theta)}{\partial W^2_{1,2}} & \frac{\partial l(\theta)}{\partial W^2_{1,3}} \\ \frac{\partial l(\theta)}{\partial W^2_{2,1}} & \frac{\partial l(\theta)}{\partial W^2_{2,2}} & \frac{\partial l(\theta)}{\partial W^2_{2,3}} \end{bmatrix} \\
&= \delta \begin{bmatrix} W^3_{1,1}\left(z^2_{1,1} \geq 0\right) h^1_{1,1} & W^3_{1,1}\left(z^2_{1,1} \geq 0\right) h^1_{1,2} & W^3_{1,1}\left(z^2_{1,1} \geq 0\right) h^1_{1,3} \\ W^3_{1,2}\left(z^2_{1,2} \geq 0\right) h^1_{1,1} & W^3_{1,2}\left(z^2_{1,2} \geq 0\right) h^1_{1,2} & W^3_{1,2}\left(z^2_{1,2} \geq 0\right) h^1_{1,3} \end{bmatrix} \\
&= \delta \begin{bmatrix} W^3_{1,1}(z^2_{1,1} \geq 0) \\ W^3_{1,2}(z^2_{1,2} \geq 0) \end{bmatrix} \begin{bmatrix} h^1_{1,1} & h^1_{1,2} & h^1_{1,3} \end{bmatrix} \\
&= \delta \left(W^3 \odot (z^2 \geq 0)\right)^\top h^1
\end{aligned}
$$

12. What are the partial derivatives of $h^2_{1,1}$ with respect to $h^1_{1,j}$?

$$
\begin{aligned}
\frac{\partial h^2_{1,1}}{\partial h^1_{1,j}} = \frac{\partial \sigma(z^2_{1,1})}{\partial h^1_{1,j}} &= \frac{\partial \sigma(z^2_{1,1})}{\partial z^2_{1,1}} \frac{\partial z^2_{1,1}}{\partial h^1_{1,j}} \\
&= (z^2_{1,1} \geq 0) \frac{\partial \left(h^1_{1,1} W^2_{1,1} + h^1_{1,2} W^2_{1,2} + h^1_{1,3} W^2_{1,3}\right)}{\partial h^1_{1,j}} \\
&= (z^2_{1,1} \geq 0) \frac{\partial h^1_{1,j} W^2_{1,j}}{\partial h^1_{1,j}} \\
&= (z^2_{1,1} \geq 0) W^2_{1,j}
\end{aligned}
$$

13. What are the partial derivatives of $h^2_{1,i}$ with respect to h^1?

$$
\frac{\partial h^2_{1,i}}{\partial h^1} = \begin{bmatrix} (z^2_{1,i} \geq 0) W^2_{i,1} & (z^2_{1,i} \geq 0) W^2_{i,2} & (z^2_{1,i} \geq 0) W^2_{i,3} \end{bmatrix}
$$

14. What are the partial derivatives of $l(\theta)$ with respect to $h^1_{1,j}$?

$$\begin{aligned}
\frac{\partial l(\theta)}{\partial h^1_{1,j}} &= \frac{\partial l(\theta)}{\partial h^3_{1,1}} \frac{\partial h^3_{1,1}}{\partial h^1_{1,j}} \\
&= \delta \frac{\partial}{\partial h^1_{1,j}} \sum_{q=1}^{n_2} h^2_{1,q} W^3_{1,q} \\
&= \delta \sum_{q=1}^{n_2} \frac{\partial h^2_{1,q} W^3_{1,q}}{\partial h^1_{1,j}} \\
&= \delta \sum_{q=1}^{n_2} \frac{\partial h^2_{1,q} W^3_{1,q}}{\partial h^2_{1,q}} \frac{\partial h^2_{1,q}}{\partial h^1_{1,j}} \\
&= \delta \sum_{q=1}^{n_2} W^3_{1,q} \frac{\partial h^2_{1,q}}{\partial h^1_{1,j}} \\
&= \delta \sum_{q=1}^{n_2} W^3_{1,q} (z^2_{1,q} \geq 0) W^2_{q,j} \\
&= \delta \begin{bmatrix} W^3_{1,1}(z^2_{1,1} \geq 0) & W^3_{1,1}(z^2_{1,1} \geq 0) \end{bmatrix} \begin{bmatrix} W^2_{1,j} \\ W^2_{2,j} \end{bmatrix} \\
&= \delta \left(W^3 \odot (z^2 \geq 0) \right) W^2_{\cdot,j}
\end{aligned}$$

15. What are the partial derivatives of $l(\theta)$ with respect to h^1?

Using matrix expression from previous answer:

$$\begin{aligned}
\frac{\partial l(\theta)}{\partial h^1} &= \begin{bmatrix} \frac{\partial l(\theta)}{\partial h^1_{1,1}} & \frac{\partial l(\theta)}{\partial h^1_{1,2}} & \frac{\partial l(\theta)}{\partial h^1_{1,3}} \end{bmatrix} \\
&= \begin{bmatrix} \delta \left(W^3 \odot (z^2 \geq 0) \right) W^2_{\cdot,1} & \delta \left(W^3 \odot (z^2 \geq 0) \right) W^2_{\cdot,2} & \delta \left(W^3 \odot (z^2 \geq 0) \right) W^2_{\cdot,3} \end{bmatrix} \\
&= \delta \left(W^3 \odot (z^2 \geq 0) \right) \begin{bmatrix} W^2_{\cdot,1} & W^2_{\cdot,2} & W^2_{\cdot,3} \end{bmatrix} \\
&= \delta \left(W^3 \odot (z^2 \geq 0) \right) W^2
\end{aligned}$$

Scalar version:

$$\frac{\partial l(\theta)}{\partial h^1} = \delta \begin{bmatrix} \sum_{q=1}^{n_2} W^3_{1,q}(z^2_{1,q} \geq 0) W^2_{q,1} & \sum_{q=1}^{n_2} W^3_{1,q}(z^2_{1,q} \geq 0) W^2_{q,2} & \sum_{q=1}^{n_2} W^3_{1,q}(z^2_{1,2} \geq 0) W^2_{q,3} \end{bmatrix}$$

16. What is the partial derivative of $l(\theta)$ with respect to $W^1_{i,j}$? Notice that $h^1_{1,i}$ is the only term of h^1 that has dependence on $W^1_{i,j}$. We have also already derived the partial derivative $\frac{\partial h^i_{1,j}}{\partial W^i_{j,k}}$ when $i = 2$, so we can reuse that result here.

$$\begin{aligned}
\frac{\partial l(\theta)}{\partial W^1_{i,j}} &= \frac{\partial l(\theta)}{\partial h^1_{1,i}} \frac{\partial h^1_{1,i}}{\partial W^1_{i,j}} \\
&= \delta \sum_{q=1}^{n_2} W^3_{1,q} (z^2_{1,q} \geq 0) W^2_{q,i} \frac{\partial h^1_{1,i}}{\partial W^1_{i,j}} \\
&= \delta \sum_{q=1}^{n_2} W^3_{1,q} (z^2_{1,q} \geq 0) W^2_{q,i} h^0_{1,j} \\
&= \delta \left(W^3 \odot (z^2 \geq 0) \right) W^2_{\cdot,i} h^0_{1,j}
\end{aligned}$$

17. What are the partial derivatives of $l(\theta)$ with respect to W^1? For conciseness you may use leave your answer in terms of $\frac{\partial l(\theta)}{\partial h^1_{1,j}}$. For further ease of notation, you can write these partial derivatives as $\partial_{h^1_{1,j}} l(\theta)$.

$$
\begin{aligned}
\frac{\partial l(\theta)}{\partial W^1} &= \begin{bmatrix} \frac{\partial l(\theta)}{\partial W^1_{1,1}} & \frac{\partial l(\theta)}{\partial W^1_{1,2}} & \frac{\partial l(\theta)}{\partial W^1_{1,3}} & \frac{\partial l(\theta)}{\partial W^1_{1,4}} \\ \frac{\partial l(\theta)}{\partial W^1_{2,1}} & \frac{\partial l(\theta)}{\partial W^1_{2,2}} & \frac{\partial l(\theta)}{\partial W^1_{2,3}} & \frac{\partial l(\theta)}{\partial W^1_{2,4}} \\ \frac{\partial l(\theta)}{\partial W^1_{3,1}} & \frac{\partial l(\theta)}{\partial W^1_{3,2}} & \frac{\partial l(\theta)}{\partial W^1_{3,3}} & \frac{\partial l(\theta)}{\partial W^1_{3,4}} \end{bmatrix} \\
&= \begin{bmatrix} \partial_{h^1_{1,1}} l(\theta) \frac{\partial h^1_{1,1}}{\partial W^1_{1,1}} & \partial_{h^1_{1,1}} l(\theta) \frac{\partial h^1_{1,1}}{\partial W^1_{1,2}} & \partial_{h^1_{1,1}} l(\theta) \frac{\partial h^1_{1,1}}{\partial W^1_{1,3}} & \partial_{h^1_{1,1}} l(\theta) \frac{\partial h^1_{1,1}}{\partial W^1_{1,4}} \\ \partial_{h^1_{1,2}} l(\theta) \frac{\partial h^1_{1,2}}{\partial W^1_{2,1}} & \partial_{h^1_{1,2}} l(\theta) \frac{\partial h^1_{1,2}}{\partial W^1_{2,2}} & \partial_{h^1_{1,2}} l(\theta) \frac{\partial h^1_{1,2}}{\partial W^1_{2,3}} & \partial_{h^1_{1,2}} l(\theta) \frac{\partial h^1_{1,2}}{\partial W^1_{2,4}} \\ \partial_{h^1_{1,3}} l(\theta) \frac{\partial h^1_{1,3}}{\partial W^1_{3,1}} & \partial_{h^1_{1,3}} l(\theta) \frac{\partial h^1_{1,3}}{\partial W^1_{3,2}} & \partial_{h^1_{1,3}} l(\theta) \frac{\partial h^1_{1,3}}{\partial W^1_{3,3}} & \partial_{h^1_{1,3}} l(\theta) \frac{\partial h^1_{1,3}}{\partial W^1_{3,4}} \end{bmatrix} \\
&= \begin{bmatrix} \partial_{h^1_{1,1}} l(\theta)(z^1_{1,1} \geq 0)h^0_{1,1} & \partial_{h^1_{1,1}} l(\theta)(z^1_{1,1} \geq 0)h^0_{1,2} & \partial_{h^1_{1,1}} l(\theta)(z^1_{1,1} \geq 0)h^0_{1,3} & \partial_{h^1_{1,1}} l(\theta)(z^1_{1,1} \geq 0)h^0_{1,4} \\ \partial_{h^1_{1,2}} l(\theta)(z^1_{1,2} \geq 0)h^0_{1,1} & \partial_{h^1_{1,2}} l(\theta)(z^1_{1,2} \geq 0)h^0_{1,2} & \partial_{h^1_{1,2}} l(\theta)(z^1_{1,2} \geq 0)h^0_{1,3} & \partial_{h^1_{1,2}} l(\theta)(z^1_{1,2} \geq 0)h^0_{1,4} \\ \partial_{h^1_{1,3}} l(\theta)(z^1_{1,3} \geq 0)h^0_{1,1} & \partial_{h^1_{1,3}} l(\theta)(z^1_{1,3} \geq 0)h^0_{1,2} & \partial_{h^1_{1,3}} l(\theta)(z^1_{1,3} \geq 0)h^0_{1,3} & \partial_{h^1_{1,3}} l(\theta)(z^1_{1,3} \geq 0)h^0_{1,4} \end{bmatrix} \\
&= \begin{bmatrix} \partial_{h^1_{1,1}} l(\theta)(z^1_{1,1} \geq 0) \\ \partial_{h^1_{1,2}} l(\theta)(z^1_{1,2} \geq 0) \\ \partial_{h^1_{1,2}} l(\theta)(z^1_{1,3} \geq 0) \end{bmatrix} \begin{bmatrix} h^0_{1,1} & h^0_{1,2} & h^0_{1,3} & h^0_{1,4} \end{bmatrix} \\
&= \left(\partial_{h^1} l(\theta) \odot (z^1 \geq 0)\right)^\top h^1
\end{aligned}
$$

2 Partial Derivatives for a Batch of Data (16 points)

Instead of computing derivatives for a single data point at a time, it is faster to compute a derivatives for a mini-batch of m data points. First consider a mini-batch size of $m = 2$, e.g., $x \in \mathbb{R}^{2\times4}$, $y \in \mathbb{R}^{2\times1}$, $h^1 \in \mathbb{R}^{2\times3}$, $h^2 \in \mathbb{R}^{2\times2}$, $h^3 \in \mathbb{R}^{2\times1}$. Let

$$l_k(\theta) = \frac{1}{2}\left(h^3_{k,1} - y_{k,1}\right)^2.$$

The loss function is now

$$l(\theta) = \frac{1}{m}\sum_{k=1}^{m} l_k(\theta) = \frac{1}{2}\frac{1}{m}\sum_{k=1}^{m}\left(h^3_{k,1} - y_{k,1}\right)^2.$$

1. What is the partial derivative of $l(\theta)$ with respect to $h^3 = f(x, \theta)$? Express your final answer using vector notation.

$$
\begin{aligned}
\delta = \frac{\partial l(\theta)}{\partial h^3} &= \begin{bmatrix} \frac{\partial l(\theta)}{\partial h^3_{1,1}} \\ \frac{\partial l(\theta)}{\partial h^3_{2,1}} \end{bmatrix} \\
&= \begin{bmatrix} (h^3_{1,1} - y_1)/m \\ (h^3_{1,1} - y_1)/m \end{bmatrix} = \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix}
\end{aligned}
$$

2. What is the partial derivative of $l(\theta)$ with respect to $W^3_{1,1}$?

$$
\begin{aligned}
\frac{\partial l(\theta)}{\partial W^3_{1,1}} &= \frac{\partial}{\partial W^3_{1,1}} \frac{1}{2}\frac{1}{m}\sum_{i=1}^{m} \left(h^3_{i,1} - y_{i,1}\right)^2 = \frac{1}{2}\frac{1}{m}\sum_{i=1}^{m} \frac{\partial}{\partial W^3_{1,1}} \left(h^3_{i,1} - y_{i,1}\right)^2 \\
&= \frac{1}{2}\frac{1}{m}\sum_{i=1}^{m} \frac{\partial \left(h^3_{i,1} - y_{i,1}\right)^2}{\partial h^3_{i,1}} \frac{\partial h^3_{i,1}}{\partial W^3_{1,1}} \\
&= \frac{1}{m}\sum_{i=1}^{m} \left(h^3_{i,1} - y_{i,1}\right) \frac{\partial h^3_{i,1}}{\partial W^3_{1,1}} \\
&= \sum_{i=1}^{m} \delta_i \frac{\partial h^3_{i,1}}{\partial W^3_{1,1}} \\
&= \sum_{i=1}^{m} \delta_i h^2_{i,1} \\
&= \begin{bmatrix} \delta_1 & \delta_2 \end{bmatrix} \begin{bmatrix} h^2_{1,1} \\ h^2_{2,1} \end{bmatrix} \\
&= \delta^\top h^2_{\cdot,1}
\end{aligned}
$$

We can also get here a little more directly by realizing this answer is just the average of the partial derivatives from each data point. Let $l_i(\theta) = \frac{1}{2}(h^3_{i,1} - y_{i,1})^2$, thus $l(\theta) = \frac{1}{m}\sum_{i=1}^{m} l_i(\theta)$.

$$
\frac{\partial l(\theta)}{\partial W^3_{1,1}} = \frac{1}{m}\sum_{i=1}^{m} \frac{\partial}{\partial W^3_{1,1}} l_i(\theta) = \frac{1}{m}\sum_{i=1}^{m} (h^3_{i,1} - y_{i,1}) h^2_{i,1} = \sum_{i=1}^{m} \delta_i h^2_{i,1}
$$

We can apply this principle to compute all partial derivatives with respect to each weight below.

3. What are the partial derivatives of $l(\theta)$ with respect to W^3? Express the final answer using vector notation.

$$
\begin{aligned}
\frac{\partial l(\theta)}{\partial W^3} &= \begin{bmatrix} \frac{\partial l(\theta)}{\partial W^3_{1,1}} & \frac{\partial l(\theta)}{\partial W^3_{1,2}} \end{bmatrix} \\
&= \begin{bmatrix} \delta^\top h^2_{\cdot,1} & \delta^\top h^2_{\cdot,2} \end{bmatrix} = \delta^\top \begin{bmatrix} h^2_{\cdot,1} & h^2_{\cdot,2} \end{bmatrix} = \delta^\top h^2
\end{aligned}
$$

4. What are the partial derivatives $l(\theta)$ with respect to $h^2_{\cdot,1}$? Express the final answer using vector notation.

   To make the answer simple, we first show the partial derivatives for $h^2_{1,1}$.

$$
\begin{aligned}
\frac{\partial l(\theta)}{\partial h^2_{1,1}} &= \frac{1}{2}\frac{1}{m}\sum_{i=1}^{m} \frac{\partial}{\partial h^2_{1,1}} \left(h^3_{i,1} - y_{i,1}\right)^2 \\
&= \frac{1}{2}\frac{1}{m}\frac{\partial}{\partial h^2_{1,1}} \left(h^3_{1,1} - y_{1,1}\right)^2 \\
&= \delta_1 \frac{\partial h^3_{1,1}}{\partial h^2_{1,1}} \\
&= \delta_1 W^3_{1,1}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l(\theta)}{\partial h^2_{\cdot,1}} &= \begin{bmatrix} \frac{\partial l(\theta)}{\partial h^2_{1,1}} \\ \frac{\partial l(\theta)}{\partial h^2_{2,1}} \end{bmatrix} \\
&= \begin{bmatrix} \delta_1 W^3_{1,1} \\ \delta_2 W^3_{1,1} \end{bmatrix} \\
&= \delta W^3_{1,1}
\end{aligned}
$$

5. What are the partial derivatives $l(\theta)$ with respect to h^2? Express the final answer using vector notation.

$$\begin{aligned}
\frac{\partial l(\theta)}{\partial h^2} &= \begin{bmatrix} \frac{\partial l(\theta)}{\partial h^2_{1,1}} & \frac{\partial l(\theta)}{\partial h^2_{1,2}} \\ \frac{\partial l(\theta)}{\partial h^2_{2,1}} & \frac{\partial l(\theta)}{\partial h^2_{2,2}} \end{bmatrix} \\
&= \begin{bmatrix} \delta_1 W^3_{1,1} & \delta_1 W^3_{1,2} \\ \delta_2 W^3_{1,1} & \delta_2 W^3_{1,2} \end{bmatrix} \\
&= \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} \begin{bmatrix} W^3_{1,1} & W^3_{1,1} \end{bmatrix} \\
&= \delta W^3
\end{aligned}$$

6. What is the partial derivative of $l(\theta)$ with respect to $W^2_{i,j}$?

$$\begin{aligned}
\frac{\partial l(\theta)}{\partial W^2_{i,j}} &= \frac{\partial}{\partial W^2_{i,j}} \frac{1}{2} \frac{1}{m} \sum_{k=1}^{m} \left(h^3_{k,1} - y_{k,1}\right)^2 \\
&= \frac{1}{2} \frac{1}{m} \sum_{k=1}^{m} \frac{\partial \left(h^3_{k,1} - y_{k,1}\right)^2}{\partial h^3_{k,1}} \frac{\partial h^3_{k,1}}{\partial W^2_{i,j}} \\
&= \sum_{k=1}^{m} \delta_k \frac{\partial h^3_{k,1}}{\partial W^2_{i,j}} \quad \text{now plug in answer from part 1} \\
&= \sum_{k=1}^{m} \delta_k W^3_{1,i} (z^2_{k,i} \geq 0) h^1_{k,j} \\
&= \sum_{k=1}^{m} \partial_{h^2_{k,i}} l(\theta) (z^2_{k,i} \geq 0) h^1_{k,j} \\
&= \begin{bmatrix} \partial_{h^2_{1,i}} l(\theta)(z^2_{1,i} \geq 0) & \partial_{h^2_{2,i}} l(\theta)(z^2_{2,i} \geq 0) \end{bmatrix} \begin{bmatrix} h^1_{1,j} \\ h^1_{2,j} \end{bmatrix} \\
&= \left(\partial_{h^2_{:,i}} l(\theta) \odot (z^2_{:,i} \geq 0)\right)^\top h^1_{:,j}
\end{aligned}$$

7. What are the partial derivatives of $l(\theta)$ with respect to W^2? Express your answer using vector notation.

$$\begin{aligned}
\frac{\partial l(\theta)}{\partial W^2} &= \begin{bmatrix} \frac{\partial l(\theta)}{\partial W^2_{1,1}} & \frac{\partial l(\theta)}{\partial W^2_{1,2}} & \frac{\partial l(\theta)}{\partial W^2_{1,3}} \\ \frac{\partial l(\theta)}{\partial W^2_{2,1}} & \frac{\partial l(\theta)}{\partial W^2_{2,2}} & \frac{\partial l(\theta)}{\partial W^2_{2,3}} \end{bmatrix} \\
&= \begin{bmatrix} \left(\partial_{h^2_{:,1}} l(\theta) \odot (z^2_{:,1} \geq 0)\right)^\top h^1_{:,1} & \left(\partial_{h^2_{:,1}} l(\theta) \odot (z^2_{:,1} \geq 0)\right)^\top h^1_{:,2} & \left(\partial_{h^2_{:,1}} l(\theta) \odot (z^2_{:,1} \geq 0)\right)^\top h^1_{:,3} \\ \left(\partial_{h^2_{:,2}} l(\theta) \odot (z^2_{:,2} \geq 0)\right)^\top h^1_{:,1} & \left(\partial_{h^2_{:,2}} l(\theta) \odot (z^2_{:,2} \geq 0)\right)^\top h^1_{:,2} & \left(\partial_{h^2_{:,2}} l(\theta) \odot (z^2_{:,2} \geq 0)\right)^\top h^1_{:,3} \end{bmatrix} \\
&= \begin{bmatrix} \left(\partial_{h^2_{:,1}} l(\theta) \odot (z^2_{:,1} \geq 0)\right)^\top \\ \left(\partial_{h^2_{:,2}} l(\theta) \odot (z^2_{:,1} \geq 0)\right)^\top \end{bmatrix} \begin{bmatrix} h^1_{:,1} & h^1_{:,2} & h^1_{:,2} \end{bmatrix} \\
&= \left(\partial_{h^2} l(\theta) \odot (z^2 \geq 0)\right)^\top h^1 \\
&= \left(\delta W^3 \odot (z^2 \geq 0)\right)^\top h^1
\end{aligned}$$

8. What are the partial derivatives of $l(\theta)$ with respect to h^1? Express your answer using vector notation. You can use $\partial_{h^2} l(\theta) = \frac{\partial l(\theta)}{\partial h^2}$ to simplify your answer.

Starting with derivative with respect to $h^1_{1,1}$.

$$
\begin{aligned}
\frac{\partial l(\theta)}{\partial h^1_{1,1}} &= \sum_{k=1}^{m} \delta_k \frac{\partial h^3_{k,1}}{\partial h^1_{1,1}} \\
&= \delta_1 \frac{\partial h^3_{1,1}}{\partial h^1_{1,1}} \\
&= \left(\delta_1 W^3 \odot (z^2_{1,\cdot} \geq 0)\right) W^2_{\cdot,1} \\
&= \left(\partial_{h^2_{1,\cdot}} l(\theta) \odot (z^2_{1,\cdot} \geq 0)\right) W^2_{\cdot,1}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial l(\theta)}{\partial h^1} &= \begin{bmatrix} \frac{\partial l(\theta)}{\partial h^1_{1,1}} & \frac{\partial l(\theta)}{\partial h^1_{1,2}} & \frac{\partial l(\theta)}{\partial h^1_{1,3}} \\ \frac{\partial l(\theta)}{\partial h^1_{2,1}} & \frac{\partial l(\theta)}{\partial h^1_{2,2}} & \frac{\partial l(\theta)}{\partial h^1_{2,3}} \end{bmatrix} \\
&= \begin{bmatrix} \left(\partial_{h^2_{1,\cdot}} l(\theta) \odot (z^2_{1,\cdot} \geq 0)\right) W^2_{\cdot,1} & \left(\partial_{h^2_{1,\cdot}} l(\theta) \odot (z^2_{1,\cdot} \geq 0)\right) W^2_{\cdot,2} & \left(\partial_{h^2_{1,\cdot}} l(\theta) \odot (z^2_{1,\cdot} \geq 0)\right) W^2_{\cdot,3} \\ \left(\partial_{h^2_{2,\cdot}} l(\theta) \odot (z^2_{2,\cdot} \geq 0)\right) W^2_{\cdot,1} & \left(\partial_{h^2_{2,\cdot}} l(\theta) \odot (z^2_{2,\cdot} \geq 0)\right) W^2_{\cdot,2} & \left(\partial_{h^2_{2,\cdot}} l(\theta) \odot (z^2_{2,\cdot} \geq 0)\right) W^2_{\cdot,3} \end{bmatrix} \\
&= \begin{bmatrix} \left(\partial_{h^2_{1,\cdot}} l(\theta) \odot (z^2_{1,\cdot} \geq 0)\right) \\ \left(\partial_{h^2_{2,\cdot}} l(\theta) \odot (z^2_{2,\cdot} \geq 0)\right) \end{bmatrix} \begin{bmatrix} W^2_{\cdot,1} & W^2_{\cdot,2} & W^2_{\cdot,3} \end{bmatrix} \\
&= \left(\partial_{h^2} l(\theta) \odot (z^2 \geq 0)\right) W^2
\end{aligned}
$$