CS577: Assignment 2

Jane Downer
Department of Computer Science
Illinois Institute of Technology

October 6, 2022

Artificial Neurons

1. $weights = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.3 \end{bmatrix}$

   $vector = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

   $output = 0.1 + (0.2)(1) + (0.3)(1) = 0.6$

2. Values of g(x) greater than 0 on one side of the decision boundary and less than 0 on the other side. When g(x) equals 0, the output falls on the decision boundary.
3. In the case of a linear discriminator, θ_0 is proportional to the negative distance of the decision boundary to the origin, whereas θ_1 and θ_2 define the (un-normalized) normal vector to the decision boundary.
4. $g(x_1, x_2) = 1 + 2x_1 + 3x_2$
   $\theta_0 = 1, \theta_1 = 2, \theta_3 = 3$
   $n = \langle \theta_1, \theta_2 \rangle = \langle 2,3 \rangle$

   $$\hat{n} = \frac{\langle \theta_1, \theta_2 \rangle}{\sqrt{\theta_0^2 + {\theta_1}^2 + {\theta_2}^2}} = \frac{\langle 2,3 \rangle}{\sqrt{1^2 + 2^2 + 3^2}} = \left\langle \frac{1\sqrt{6}}{3}, \frac{1\sqrt{6}}{3} \right\rangle$$

   $$d = -\frac{\theta_0}{\sqrt{\theta_0^2 + {\theta_1}^2 + {\theta_2}^2}} = -\frac{1\sqrt{6}}{6}$$

5. When writing a discriminant function as $g(x) = \theta^T x$, using the bias trick we set x equal to $[1, x_1, \dots, x_n]$, giving it the same dimensions as θ. This means that when we take the dot product of θ and x, the bias term is added to the sum without being scaled by the input vector.
6. Step function: $h_\theta(x) = \begin{cases} 1 & if\ \theta^T x > 0 \\ 0 & otherwise \end{cases}$

   Logistic (sigmoid) function: $h_\theta(x) = \frac{1}{1+\exp(-\theta^T x)}$

   The logistic function outputs a probability of belonging to a class rather than making a binary classification. For example, if a discriminant function outputs a value that is 0.0001, the step function would output 1, whereas the logistic function would output $\frac{1}{1+\exp(-0.0001)} = 0.500025$ – reflecting that the observation has close to a 50/50 chance of belonging to either category.
7. Start with this activation function:

   $$h_\theta(x) = \begin{cases} c_1 & if\ \dfrac{P(y=1|x)}{P(y=0|x)} > 1 \\ c_0 & if\ \dfrac{P(y=0|x)}{P(y=0|x)} < 1 \end{cases}$$

   Taking the log of the likelihood ratio is a monotonic transformation, and converts values over 1 to positive values and values under 1 to negative values. Therefore,

$$h_\theta(x) = \begin{cases} c_1 & if \;\; \log\left(\frac{P(y=1|x)}{P(y=0|x)}\right) > 0 \\ c_0 & if \;\; \log\left(\frac{P(y=1|x)}{P(y=0|x)}\right) < 0 \end{cases}$$

This can be modeled in terms of a linear function: $\log\left(\frac{P(y=1|x)}{P(y=0|x)}\right) = \theta^T x$

Therefore:

$$\begin{aligned}
\frac{P(y=1|x)}{P(y=0|x)} &= \exp\,(\theta^T x) \\
P(y=1|x) &= P(y=0|x) * \exp\,(\theta^T x) \\
P(y=1|x) &= \big(1 - P(y=1|x)\big) * \exp(\theta^T x) \\
\frac{P(y=1|x)}{1-P(y=1|x)} &= \exp(\theta^T x) \\
P(y=1|x) *(1+\exp(\theta^T x)) &= \exp(\theta^T x) \\
P(y=1|x) &= \frac{\exp(\theta^T x)}{1+\exp(\theta^T x)} \\
P(y=1|x) &= \frac{1}{1+\exp\,(-\theta^T x)} \\
P(y=1|x) &= sigmoid(x) \\
h_\theta(x) &= \begin{cases} c_1 & if \;\; sigmoid(x) > 0.5 \\ c_0 & \qquad otherwise \end{cases}
\end{aligned}$$

8.

$$\begin{aligned}
sigmoid(a) &= \frac{1}{1+\exp\,(-a)} \\
\frac{d}{da} sigmoid(a) &= \big(sigmoid(a)\big) * \big(1 - sigmoid(a)\big) \\
\\
\frac{d}{dx}\log(a) &= \frac{1}{a} * a' \\
\\
\frac{d}{da}\log\big(sigmoid(a)\big) &= \frac{1}{sigmoid(a)} * \frac{d}{da} sigmoid(a) \\
\frac{d}{da}\log\big(sigmoid(a)\big) &= \frac{1}{sigmoid(a)} * \big(sigmoid(a)\big) * \big(1 - sigmoid(a)\big) \\
\frac{d}{da}\log\big(sigmoid(a)\big) &= 1 - sigmoid(a)
\end{aligned}$$

9. In gradient descent, the direction of the update for the $(i+1)^{th}$ guess is given by the negative of the previous gradient, $\nabla J(\theta_i)$. The update is controlled by the learning rate, η. The formula that ties these together is:

$$\theta_{i+1} = \theta_i - \eta \cdot \nabla J(\theta_i)$$

10. The stop condition for gradient descent is that we should stop making updates to θ when the resulting change in loss is deemed sufficiently small. The condition is determined by the change in loss rather the change in θ, because our objective is to minimize loss, and we don't know that a small change in θ also means that the change in loss was similarly small. (i.e., we do not know how sensitive the loss function is to changes in θ).
11. When the learning rate is too small, it will take a very long time to minimize the loss. When the learning rate is too big, the gradient descent process may make updates that skip over the optimal solution and never converge.
12. Empirical error loss is a count of the incorrect classifications. This cannot be used in gradient descent because this loss function is piecewise-constant. As a result, every value of θ will produce a local minimum for the loss function (gradient of 0), so gradient descent will never make any updates / will be unable to minimize the loss.
13. Binary cross-entropy loss is the negative log likelihood for a given value of θ.

$$\begin{aligned}
L(\theta) &= \textstyle\prod_{i=1}^{m} P(y=1|x_i)^{y_i} \cdot P(y=0|x_i)^{1-y_i} \\
\\
l(\theta) &= -\log\big(L(\theta)\big) \\
&= -\log(\textstyle\prod_{i=1}^{m} P(y=1|x_i)^{y_i} \cdot P(y=0|x_i)^{1-y_i})
\end{aligned}$$

$$
\begin{aligned}
&= -\sum_{i=1}^{m}\Big(y_i \cdot \log\big(P(y=1|x_i)\big) + (1-y_i)\cdot \log\big(P(y=0|x_i)\big)\Big)\\
&= -\sum_{i=1}^{m}\Big(y_i \cdot \log\big(h_\theta(x)\big) + (1-y_i)\cdot \log\Big(\big(1-h_\theta(x)\big)\Big)\Big)\\
&= -\sum_{i=1}^{m}\big(y_i \cdot \log(\hat{y}_i) + (1-y_i)\cdot \log\big((1-\hat{y}_i)\big)\big) \quad \leftarrow \text{binary cross-entropy}
\end{aligned}
$$

14. Assuming a sigmoid activation function:

$$
\begin{aligned}
h_\theta(x) &= \sigma(x) \qquad \leftarrow \text{sigmoid function}\\
\frac{d}{d\theta}\log\big(h_\theta(x)\big) &= \frac{d}{dx}\log\big(\sigma(\theta^T x)\big)\cdot \frac{d}{d\theta}x\\
&= \big(1-\sigma(\theta^T x)\big)\cdot x\\
\frac{d}{d\theta}h_\theta(x) &= \frac{d}{dx}\sigma(\theta^T x)\cdot \frac{d}{d\theta}(\theta^T x)\\
&= \sigma(\theta^T x)\cdot\big(1-\sigma(\theta^T x)\big)\cdot x
\end{aligned}
$$

<u>Gradient</u>

$$
\begin{aligned}
\frac{d}{d\theta}l(\theta) &= \frac{d}{d\theta}\Big(-\sum_{i=1}^{m}\big(y_i\cdot\log(\hat{y}_i) + (1-y_i)\cdot\log\big((1-\hat{y}_i)\big)\big)\Big)\\
&= \frac{d}{d\theta}\Big(-\sum_{i=1}^{m}\Big(y_i\cdot\log\big(h_\theta(x_i)\big) + (1-y_i)\cdot\log\Big(\big(1-h_\theta(x_i)\big)\Big)\Big)\Big)\\
&= -\sum_{i=1}^{m}\left(\frac{d}{d\theta}\Big(y_i\cdot\log\big(h_\theta(x_i)\big)\Big)\right) - \sum_{i=1}^{m}\left(\frac{d}{d\theta}\Big((1-y_i)\cdot\log\Big(\big(1-h_\theta(x_i)\big)\Big)\Big)\right)\\
&= -\sum_{i=1}^{m}\left(\frac{d}{d\theta}\Big(y_i\cdot\log\big(\sigma(\theta^T x_i)\big)\Big)\right) - \sum_{i=1}^{m}\left(\frac{d}{d\theta}\Big((1-y_i)\cdot\log\Big(\big(1-\sigma(\theta^T x_i)\big)\Big)\Big)\right)\\
&= -\sum_{i=1}^{m}y_i\cdot\big(1-\sigma(\theta^T x_i)\big)\cdot x_i - \sum_{i=1}^{m}(1-y_i)\cdot\frac{1}{1-\sigma(\theta^T x_i)}\cdot\frac{d}{d\theta}\big(1-\sigma(\theta^T x_i)\big)\\
&= -\sum_{i=1}^{m}y_i\cdot\big(1-\sigma(\theta^T x_i)\big)\cdot x_i - \sum_{i=1}^{m}(1-y_i)\cdot\frac{1}{1-\sigma(\theta^T x_i)}\cdot\big(-\sigma(\theta^T x_i)\big)\cdot\big(1-\sigma(\theta^T x_i)\big)\cdot x_i\\
&= -\sum_{i=1}^{m}y_i\cdot\big(1-\sigma(\theta^T x_i)\big)\cdot x_i - \sum_{i=1}^{m}(1-y_i)(-\sigma(\theta^T x_i)\cdot x_i)\\
&= -\sum_{i=1}^{m}y_i\cdot\big(1-\sigma(\theta^T x_i)\big)\cdot x_i + \sum_{i=1}^{m}(1-y_i)\cdot\sigma(\theta^T x_i)\cdot x_i\\
&= -\sum_{i=1}^{m}x_i\cdot\big(y_i - y_i\cdot\sigma(\theta^T x_i) - \sigma_i(\theta^T x_i) + y_i\cdot\sigma(\theta^T x_i)\big)\\
&= -\sum_{i=1}^{m}x_i\cdot\big(y_i - \sigma(\theta^T x_i)\big)
\end{aligned}
$$

<u>Update Rule</u>

$$
\begin{aligned}
\theta_{j+1} &= \theta_j - \eta\cdot\Big(-\sum_{i=1}^{m}\big(y_i - \sigma(\theta^T x_i)\big)\cdot x_i\Big)\\
\theta_{j+1} &= \theta_j - \eta\cdot\Big(\sum_{i=1}^{m}(\sigma(\theta^T x_i) - y_i)\cdot x_i\Big)\\
\theta_{j+1} &= \theta_j - \eta\cdot\left(\sum_{i=1}^{m}\left(\frac{1}{1+\exp(-\theta_i^T x_i)} - y_i\right)\cdot x_i\right)
\end{aligned}
$$

The value of each update to θ equals the previous value of θ minus the learning rate times the gradient of the loss function. In this case, the gradient is equal to the summed differences between $sigmoid(\theta^T x)$ and y, times x, for every observation i through m in the batch. Put simply, the amount of change is the negative of the input times the summed prediction errors, scaled by a learning rate. In any iteration i through m where an error is not made, then that iteration does not contribute to the parameter update.

15. One against all others: Each of the k categories has its own discriminant model that acts as a binary classifier, treating the given category as one class and grouping all other observations into a second class. The combination of the k discriminant functions jointly divide up the input space into regions (theoretically)

belonging to each class. For a given observation, choose the class whose decision boundary has the greatest signed distance from that observation (where the sign comes from the normal vector).
One against each other: create a binary discriminant model between each unique pair of classes – this results in $k(k+1)/2$ discriminant functions – a much larger number of models than with the "one against all others" method. This will do a better job of separating the classes, but the other method is usually easier.

16. The template matching interpretation means we think of the rows of the parameter matrix as templates for each of the k categories. If the j^{th} value in the output vector is larger than the others, then we can think of this observation as matching the template for the j^{th} category, defined by the j^{th} row of Θ.
17. If we were to use the sigmoid activation function for multi-class classification, the output would be k values that sum to more than 1, which doesn't make sense if we are looking for probability. The softmax activation function outputs k probabilities that sum to 1, representing, for each observation, the probabilities of belonging to each of the k classes.
18.

$$softmax(z_j) = \frac{\exp(z_j)}{\sum_{i=1}^{k} \exp(z_i)}$$

$$\begin{aligned}
\frac{\partial}{\partial z_i} softmax(z_j) &= \frac{\partial}{\partial z_i}\left(\exp(z_j) \cdot \left(\sum_{i=1}^{k} \exp(z_i)\right)^{-1}\right) \\
&= \delta_{ij} \cdot \left(\exp(z_j) \cdot \left(\sum_{i=1}^{k} \exp(z_i)\right)^{-1}\right) + \exp(z_j) \cdot (-1) \cdot \left(\sum_{i=1}^{k} \exp(z_i)\right)^{-2} \cdot \exp(z_i) \\
&= \left(\exp(z_j) \cdot \left(\sum_{i=1}^{k} \exp(z_i)\right)^{-1}\right) \cdot \left(\delta_{ij} - \left(\sum_{i=1}^{k} \exp(z_i)\right)^{-1} \cdot \exp(z_i)\right) \\
&= softmax(z_j) \cdot \left(\delta_{ij} - softmax(z_i)\right)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \theta_i} softmax(z_j) &= \frac{\partial}{\partial z_i} softmax(z_j) \cdot \frac{\partial}{\theta_i} z_j \\
&= softmax(\theta_j^T x) \cdot \left(\delta_{ij} - softmax(\theta_i^T x)\right) \cdot \frac{\partial}{\theta_i} \theta_j^T x \\
&= softmax(\theta_j^T x) \cdot \left(\delta_{ij} - softmax(\theta_i^T x)\right) \cdot x \cdot \delta_{ij} \\
&= softmax(\theta_j^T x) \cdot \left(\delta_{ij} - softmax(\theta_i^T x)\right) \cdot x
\end{aligned}$$

$$\begin{aligned}
\log\left(softmax(z_j)\right) &= \log\left(\frac{exp(z_j)}{\sum_{i=1}^{k} exp(z_i)}\right) \\
&= \log(\exp(z_j)) - \log\left(\sum_{i=1}^{k} exp(z_i)\right) \\
&= z_j - \log\left(\sum_{i=1}^{k} exp(z_i)\right)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \theta_i} \log\left(softmax(z_j)\right) &= \frac{1}{softmax(z_j)} \cdot \frac{\partial}{\partial \theta_i} softmax(z_j) \\
&= \frac{1}{softmax(z_j)} \cdot softmax(\theta_j^T x) \cdot \left(\delta_{ij} - softmax(\theta_i^T x)\right) \cdot x \\
&= \frac{1}{softmax(\theta_j^T x)} \cdot softmax(\theta_j^T x) \cdot \left(\delta_{ij} - softmax(\theta_i^T x)\right) \cdot x \\
&= \left(\delta_{ij} - softmax(\theta_i^T x)\right) \cdot x
\end{aligned}$$

19.

Likelihood: for every observation i through m, multiply the probability of belonging to classes j through k.

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^{m} \prod_{j=1}^{k} P(y_i = j|x)^{\mathbf{1}(y_i=j)} \qquad \leftarrow \quad \mathbf{1} = \text{indicator function} \\
l(\theta) &= -\log\left(\prod_{i=1}^{m} \prod_{j=1}^{k} P(y_i = j|x)^{\mathbf{1}(y_i=j)}\right) \\
&= -\sum_{i=1}^{m} \sum_{j=1}^{k} \log\left(P(y_i = j|x)^{\mathbf{1}(y_i=j)}\right) \\
&= -\sum_{i=1}^{m} \sum_{j=1}^{k} \mathbf{1}(y_i = j) \cdot \log(P(y_i = j|x)) \\
&= -\sum_{i=1}^{m} \sum_{j=1}^{k} \mathbf{1}(y_i = j) \cdot \log\left(h_{\theta_j}(x_i)\right) \qquad \leftarrow \text{categorical cross-entropy}
\end{aligned}$$

20.

Gradient:

$$\begin{aligned}
\frac{\partial}{\partial\theta_j} l(\theta) &= -\frac{\partial}{\partial\theta_j}\sum_{i=1}^{m}\sum_{j=1}^{k}\mathbf{1}(y_i = j)\cdot\log\left(h_{\theta_j}(x_i)\right) \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{k}\mathbf{1}(y_i = j)\cdot\frac{\partial}{\partial\theta_j}\log\left(h_{\theta_j}(x_i)\right) \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{k}\mathbf{1}(y_i = j)\cdot\frac{1}{h_{\theta_j}(x_i)}\cdot\frac{\partial}{\partial\theta_j}h_{\theta_j}(x_i) \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{k}\mathbf{1}(y_i = j)\cdot\frac{1}{softmax\left(\theta_j^T x\right)}\cdot\frac{\partial}{\partial\theta_j}softmax\left(\theta_j^T x\right) \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{k}\mathbf{1}(y_i = j)\cdot\frac{1}{softmax\left(\theta_j^T x\right)}\cdot softmax\left(\theta_j^T x\right)\cdot\left(1 - softmax\left(\theta_j^T x\right)\right)\cdot x_i \\
&= -\sum_{i=1}^{m}\sum_{j=1}^{k}\left(\mathbf{1}(y_i = j) - softmax\left(\theta_j^T x\right)\right)\cdot x_i \\
&= \sum_{i=1}^{m}\left(softmax\left(\theta_j^T x\right) - \mathbf{1}(y_i = j)\right)\cdot x_i
\end{aligned}$$

Update rule:

$$\begin{aligned}
\theta_{j+1} &= \theta_j - \eta\cdot\frac{\partial}{\partial\theta_j}l(\theta) \\
&= \theta_j - \eta\cdot\sum_{i=1}^{m}\left(softmax\left(\theta_j^T x\right) - \mathbf{1}(y_i = j)\right)\cdot x_i
\end{aligned}$$

For each observation in the batch, the parameter θ_{j+1} is equal to the previous guess, θ_j, minus the step size times the gradient of the loss function. When the loss function is categorical cross-entropy, the gradient is equal to the summed values of an observation x_i times the distance of $softmax\left(\theta_j^T x\right)$ to the truth (0 if $y_i = j$, 1 otherwise). In other words, for each observation, the closer our prediction to the ground truth, the smaller the contribution of that observation to the update.

Neural Networks

1. The number of units in a layer indicates the size of the output. So in a two-layer network where there are fewer units in the first layer than the second, mapping values from the first layer onto the units in the second layer will increase the dimensionality. This is beneficial in the case of non-linear classification where there may be more complicated patterns to learn.
2. Conversely to the above explanation, if there are fewer units in the second layer, then when the outputs from the first layer are mapped onto the units in the second, the dimensionality will decrease. This is beneficial when we want to simplify the data – for example, in the case of regression, where we may want to take many input values and output a single value.
3.

$$\begin{aligned}
v_j &\leftarrow v_j - \eta\cdot\sum_{i=1}^{m}\left(\hat{y}_j{}^{(i)} - \mathbf{1}(y^{(i)} = j)\right)\cdot z^{(i)} && \text{update equation for output layer} \\
w_j &\leftarrow w_j - \eta\cdot\sum_{i=1}^{m}\left(\hat{z}_j{}^{(i)} - \mathbf{1}(z^{(i)} = j)\right)\cdot x^{(i)} && \text{update equation for hidden layer}
\end{aligned}$$

   We cannot directly use the update equations for the hidden layer because the terms $\hat{z}_j{}^{(i)}$ and $z^{(i)}$ are unknown.
4. For single output regression:

$$\begin{aligned}
\hat{y} &= v^T z && \text{linear activation in final layer} \\
z_j &= sigmoid(w_j^T x) && \text{sigmoid activation hidden layer} \\
E(\theta) &= \frac{1}{2}\sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right)^2
\end{aligned}$$

Gradient w.r.t. **outer layer parameters** (v):

$$\begin{aligned}\frac{\partial E}{\partial v} &= \frac{\partial E}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial v} \\ &= 2 \cdot \frac{1}{2} \cdot \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right) \cdot z^{(i)} \\ &= \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right) \cdot z^{(i)}\end{aligned}$$

Gradient w.r.t. **hidden layer parameters** (w):

$$\frac{\partial E}{\partial w} = \left[\frac{\partial E}{\partial w_1}, \dots, \frac{\partial E}{\partial w_k}\right]$$

where:

$$\begin{aligned}\frac{\partial E}{\partial w_j} &= \frac{\partial E}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z_j} \cdot \frac{\partial z_j}{\partial w_j} \\ &= 2 \cdot \frac{1}{2} \cdot \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right) \cdot v_j \cdot z^{(i)} \cdot \left(1 - z_j^{(i)}\right) \cdot x^{(i)} \\ &= \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right) \cdot v_j \cdot z^{(i)} \cdot \left(1 - z_j^{(i)}\right) \cdot x^{(i)}\end{aligned}$$

5. For multiple output regression:

$$\begin{aligned}\hat{y} &= [\hat{y}_1, \dots, \hat{y}_k] \\ \hat{y}_j &= v_j^T z_j && \text{linear activation in final layer} \\ z_j &= sigmoid(w_j^T x) && \text{sigmoid activation in hidden layer} \\ E(\theta_j) &= \sum_{j=1}^{k} \frac{1}{2} \sum_{i=1}^{m} \left(\hat{y}_j{}^{(i)} - y_j^{(i)}\right)^2\end{aligned}$$

Gradient w.r.t. **outer layer parameters** (v):

$$\begin{aligned}\frac{\partial E}{\partial v_j} &= \frac{\partial E}{\partial \hat{y}_j} \cdot \frac{\partial \hat{y}_j}{\partial v_j} \\ &= 2 \cdot \frac{1}{2} \cdot \sum_{i=1}^{m} \left(\hat{y}_j{}^{(i)} - y_j^{(i)}\right) \cdot z_j^{(i)} \\ &= \sum_{i=1}^{m} \left(\hat{y}_j{}^{(i)} - y_j^{(i)}\right) \cdot z_j^{(i)}\end{aligned}$$

Gradient w.r.t. **hidden layer parameters** (w):

$$\begin{aligned}\frac{\partial E}{\partial w_j} &= \sum_{j=1}^{k} \frac{\partial E}{\partial \hat{y}_j} \cdot \frac{\partial \hat{y}_j}{\partial z_j} \cdot \frac{\partial z_j}{\partial w_j} \\ &= \sum_{j=1}^{k} 2 \cdot \frac{1}{2} \cdot \sum_{i=1}^{m} \left(\hat{y}_j{}^{(i)} - y_j^{(i)}\right) \cdot v_j \cdot z^{(i)} \cdot \left(1 - z_j^{(i)}\right) \cdot x^{(i)} \\ &= \sum_{j=1}^{k} \sum_{i=1}^{m} \left(\hat{y}_j{}^{(i)} - y_j^{(i)}\right) \cdot v_j \cdot z^{(i)} \cdot \left(1 - z_j^{(i)}\right) \cdot x^{(i)}\end{aligned}$$

6. For binary classification:

$$\begin{aligned}\hat{y} &= sigmoid\left(v^T z^{(i)}\right) && \text{sigmoid activation in final layer} \\ z_j &= sigmoid(w_j^T x) && \text{sigmoid activation in hidden layer} \\ E(\theta) &= -\left(\sum_{i=1}^{m} y^{(i)} \cdot \log\left(\hat{y}^{(i)}\right) + \left(1 - y^{(i)}\right) \cdot \log\left(1 - \hat{y}^{(i)}\right)\right)\end{aligned}$$

Gradient w.r.t. **outer layer parameters** (v):

$$\begin{aligned}\frac{\partial E}{\partial v} &= \frac{\partial E}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z_j} \\ &= \sum_{i=1}^{m} \left(\frac{\hat{y}^{(i)} - y^{(i)}}{\hat{y}^{(i)} \cdot \left(1 - \hat{y}^{(i)}\right)}\right) \cdot \hat{y}^{(i)} \cdot \left(1 - \hat{y}^{(i)}\right) \cdot z^{(i)} \\ &= \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right) \cdot z^{(i)}\end{aligned}$$

Gradient w.r.t. **hidden layer parameters** (w):

$$\begin{aligned}\frac{\partial E}{\partial w_j} &= \frac{\partial E}{\partial \hat{y}_j}\cdot\frac{\partial \hat{y}_j}{\partial z_j}\cdot\frac{\partial z_j}{\partial w_j}\\ &= \sum_{i=1}^{m}\left(\frac{\hat{y}^{(i)}-y^{(i)}}{\hat{y}^{(i)}\cdot\left(1-\hat{y}^{(i)}\right)}\right)\cdot\hat{y}\cdot(1-\hat{y})\cdot v_j\cdot\left(z_j^{(i)}\right)*\left(1-z_j^{(i)}\right)\cdot x_j^{(i)}\\ &= \sum_{i=1}^{m}\left(\hat{y}^{(i)}-y^{(i)}\right)\cdot v_j\cdot\left(z_j^{(i)}\right)*\left(1-z_j^{(i)}\right)\cdot x_j^{(i)}\end{aligned}$$

7. For multi-class classification:

$$\begin{aligned}\hat{y} &= \frac{\exp\left(v^T z^{(i)}\right)}{\sum_{j=1}^{k}\exp\left(v^T z_j^{(i)}\right)} && \text{softmax activation in final layer}\\ z_j &= sigmoid\left(w_j^T x\right) && \text{sigmoid activation in hidden layer}\\ E(\theta) &= -\sum_{i=1}^{m}\sum_{j=1}^{k}\mathbf{1}\left(y^{(i)}=j\right)\cdot\log\left(\hat{y}_j^{(i)}\right)\end{aligned}$$

Gradient w.r.t. **outer layer parameters** (v):

$$\begin{aligned}\frac{\partial E}{\partial v_j} &= \frac{\partial E}{\partial \hat{y}}\cdot\frac{\partial \hat{y}}{\partial v_j}\\ &= -\sum_{i=1}^{m}\mathbf{1}\left(y^{(i)}=j\right)\cdot\frac{1}{\hat{y}_j^{(i)}}\cdot\sum_{i=1}^{m}\hat{y}_j^{(i)}\cdot\left(\delta_{ij}-\hat{y}_j^{(i)}\right)\cdot z_j^{(i)}\\ &= \sum_{i=1}^{m}\left(\hat{y}_j^{(i)}-\mathbf{1}\left(y^{(i)}=j\right)\right)\cdot z_j^{(i)}\end{aligned}$$

Gradient w.r.t. **hidden layer parameters** (w):

$$\begin{aligned}\frac{\partial E}{\partial w_j} &= \sum_{j=1}^{k}\frac{\partial E}{\partial \hat{y}_j}\cdot\frac{\partial \hat{y}_j}{\partial z_j}\cdot\frac{\partial z_j}{\partial w_j}\\ &= \sum_{j=1}^{k}\sum_{i=1}^{m}-\mathbf{1}\left(y^{(i)}=j\right)\cdot\frac{1}{\hat{y}_j^{(i)}}\cdot\hat{y}_j\cdot\left(\delta_{ij}-\hat{y}_j\right)\cdot v_j^{(i)}\cdot z_j\cdot\left(1-z_j\right)\cdot x_j\\ &= \sum_{j=1}^{k}\sum_{i=1}^{m}\left(\hat{y}_j^{(i)}-\mathbf{1}\left(y^{(i)}=j\right)\right)\cdot v_j\cdot z_j^{(i)}\cdot\left(1-z_j^{(i)}\right)\cdot x_j^{(i)}\end{aligned}$$

8. Weights should be initialized to small, random values – the randomness is necessary, otherwise all nodes would learn identically.

Computation Graphs

1. When there are many layers, it is difficult to compute the gradient manually. Computation graphs provide a more intuitive way of calculating the gradient. During the forward pass, each input is pushed through the network and, we compute the output using the current estimate of the parameters. During the backward pass, we compute the gradient with respect to each parameter in the network. Each node needs to be able to compute the output given the inputs and the function it represents, store these values, use these values to find the gradient of the loss with respect to its parameters, and then store the learned weights for the next round of updates.
2.

$$\begin{aligned}z &= f_1(W_1, x) && \leftarrow \text{hidden layer output}\\ \hat{y} &= f_2(W_2, z)\\ &= f_2\left(W_2, f_1(W_1, x)\right) && \leftarrow \text{final layer output}\\ \hat{y}_i &= f_2\left(W_2, f_1(W_1, x_i)\right)\\ loss &= L(\hat{y}, y)\\ &= L\left(f_2\left(W_2, f_1(W_1, x_i)\right), y\right)\end{aligned}$$

Generally:

$$\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial \hat{y}}\cdot\frac{\partial \hat{y}}{\partial z}\cdot\frac{\partial z}{\partial W_1}$$

$$= \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial f_2(W_2,z)}{\partial z} \cdot \frac{\partial f_1(W_1,x)}{\partial W_1}$$

$$\begin{aligned} \frac{\partial L}{\partial W_2} &= \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial W_2} \\ &= \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial f_2(W_2,z)}{\partial W_2} \end{aligned}$$

When the loss function is $L_2(\,)$:

$$L = (\hat{y} - y)^2$$

$$\frac{\partial L}{\partial \hat{y}} = 2 \cdot (\hat{y} - y)$$

$$\frac{\partial L}{\partial W_1} = 2 \cdot (\hat{y} - y) \cdot \frac{\partial f_2(W_2,z)}{\partial z} \cdot \frac{\partial f_1(W_1,x)}{\partial W_1}$$

$$\frac{\partial L}{\partial W_2} = 2 \cdot (\hat{y} - y) \cdot \frac{\partial f_2(W_2,z)}{\partial W_2}$$

When the loss function is cross-entropy:

$$L = \; -\sum_{i=1}^{m} (y_i \cdot \log(\hat{y}_i)) + (1 - y_i) \cdot \log((1 - \hat{y}_i)))$$

$$\begin{aligned} \frac{\partial L}{\partial \hat{y}} &= -\left(\frac{y}{\hat{y}} + \frac{1-y}{(1-\hat{y})} \cdot (-1)\right) \\ &= -\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}} \end{aligned}$$

$$\frac{\partial L}{\partial W_1} = \left(-\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}}\right) \cdot \frac{\partial f_2(W_2,z)}{\partial z} \cdot \frac{\partial f_1(W_1,x)}{\partial W_1}$$

$$\frac{\partial L}{\partial W_2} = \left(-\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}}\right) \cdot \frac{\partial f_2(W_2,z)}{\partial W_2}$$

3.

When the loss function is cross-entropy:

$$\begin{aligned} z &= f_1(W_1, x) && \leftarrow \text{hidden layer output} \\ \hat{y} &= f_2(W_2, z) \\ &= f_2\big(W_2, f_1(W_1, x)\big) && \leftarrow \text{final layer output} \\ \hat{y}_i &= f_2\big(W_2, f_1(W_1, x_i)\big) \end{aligned}$$

$$\frac{\partial \hat{y}}{\partial W_1} = \frac{\partial f_2}{\partial W_1} \qquad \frac{\partial \hat{y}}{\partial z} = \frac{\partial f_2}{\partial z} \qquad \frac{\partial z}{\partial W_1} = \frac{\partial f_1}{\partial W_1} \qquad \frac{\partial z}{\partial x} = \frac{\partial f_1}{\partial x}$$

Generally:

$$\begin{aligned} \frac{\partial L}{\partial W_1} &= \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial z}{\partial W_1} \\ &= \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial f_2(W_2,z)}{\partial z} \cdot \frac{\partial f_1(W_1,x)}{\partial W_1} \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial W_2} &= \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial W_2} \cdot \frac{\partial z}{\partial W_1} \\ &= \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial f_2(W_2,z)}{\partial W_2} \end{aligned}$$

When the loss function is $L_2(\,)$:

$$L = (\hat{y} - y)^2$$

$$\frac{\partial L}{\partial \hat{y}} = 2 \cdot (\hat{y} - y)$$

$$\frac{\partial L}{\partial W_2} = 2 \cdot (\hat{y} - y) \cdot \frac{\partial \hat{y}}{\partial f_2} \cdot \frac{\partial f_2}{\partial W_2}$$

$$\frac{\partial L}{\partial W_1} = 2 \cdot (\hat{y} - y) \cdot \frac{\partial \hat{y}}{\partial f_2} \cdot \frac{\partial f_2}{\partial f_1} \cdot \frac{\partial f_1}{\partial W_1}$$

When the loss function is cross-entropy:

$$L = -\sum_{i=1}^{m}\left(y_i \cdot \log(\hat{y}_i)\right) + (1 - y_i) \cdot \log\big((1 - \hat{y}_i)\big)\Big)$$

$$\begin{aligned}\frac{\partial L}{\partial \hat{y}} &= -\left(\frac{y}{\hat{y}} + \frac{1-y)}{(1-\hat{y})} \cdot (-1)\right) \\ &= -\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}}\end{aligned}$$

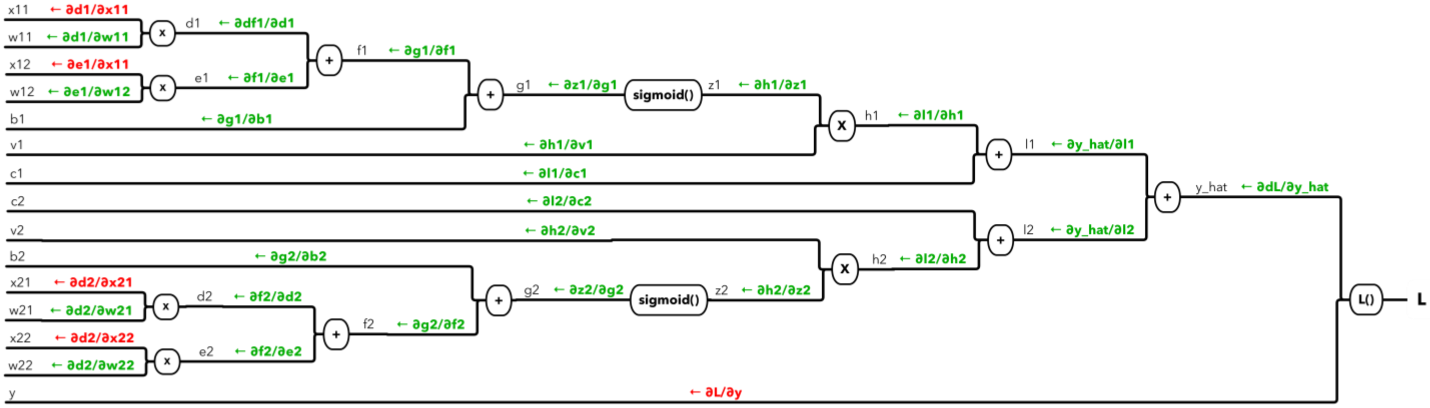
$$\frac{\partial L}{\partial W_2} = \left(-\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}}\right) \cdot \frac{\partial \hat{y}}{\partial f_2} \cdot \frac{\partial f_2}{\partial W_2}$$

$$\frac{\partial L}{\partial W_1} = \left(-\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}}\right) \cdot \frac{\partial \hat{y}}{\partial f_2} \cdot \frac{\partial f_2}{\partial f_1} \cdot \frac{\partial f_1}{\partial W_1}$$

4.

Gradients needed: $\frac{\partial L}{\partial \hat{y}}, \frac{\partial \hat{y}}{\partial l_1}, \frac{\partial \hat{y}}{\partial l_2}, \frac{\partial l_1}{\partial c_1}, \frac{\partial l_2}{\partial c_2}, \frac{\partial l_1}{\partial h_1}, \frac{\partial l_2}{\partial h_2}, \frac{\partial h_1}{\partial v_1}, \frac{\partial h_2}{\partial v_2}, \frac{\partial h_1}{\partial z_1}, \frac{\partial h_2}{\partial z_2}, \frac{\partial z_1}{\partial g_1}, \frac{\partial z_2}{\partial g_2}, \frac{\partial g_1}{\partial b_1}, \frac{\partial g_2}{\partial b_2}, \frac{\partial g_1}{\partial f_1}, \frac{\partial g_2}{\partial f_2}, \frac{\partial f_1}{\partial d_1}, \frac{\partial f_2}{\partial d_2}, \frac{\partial f_1}{\partial e_1}, \frac{\partial f_2}{\partial e_2}, \frac{\partial d_1}{\partial w_{11}}, \frac{\partial d_2}{\partial w_{21}}, \frac{\partial e_1}{\partial w_{12}}, \frac{\partial e_2}{\partial w_{22}}$

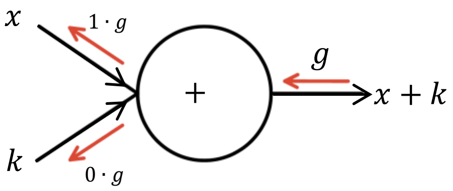
Gradients not needed: $\frac{\partial d_1}{\partial x_{11}}, \frac{\partial d_2}{\partial x_{21}}, \frac{\partial e_1}{\partial x_{12}}, \frac{\partial e_2}{\partial x_{22}}, \frac{\partial L}{\partial y}$



$$\begin{aligned}
\frac{\partial L}{\partial w_{11}} &= \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial l_1} \cdot \frac{\partial l_1}{\partial h_1} \cdot \frac{\partial h_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial g_1} \cdot \frac{\partial g_1}{\partial f_1} \cdot \frac{\partial f_1}{\partial d_1} \cdot \frac{\partial d_1}{\partial w_{11}} \\
\frac{\partial L}{\partial w_{12}} &= \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial l_1} \cdot \frac{\partial l_1}{\partial h_1} \cdot \frac{\partial h_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial g_1} \cdot \frac{\partial g_1}{\partial f_1} \cdot \frac{\partial f_1}{\partial e_1} \cdot \frac{\partial e_1}{\partial w_{12}} \\
\frac{\partial L}{\partial b_1} &= \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial l_1} \cdot \frac{\partial l_1}{\partial h_1} \cdot \frac{\partial h_1}{\partial z_1} \cdot \frac{\partial z_1}{\partial g_1} \cdot \frac{\partial g_1}{\partial b_1}
\end{aligned}$$

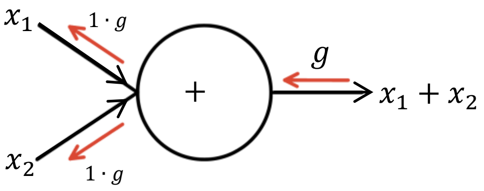
$$\begin{aligned}
\frac{\partial L}{\partial w_{21}} &= \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial l_2} \cdot \frac{\partial l_2}{\partial h_2} \cdot \frac{\partial h_2}{\partial z_2} \cdot \frac{\partial z_2}{\partial g_2} \cdot \frac{\partial g_2}{\partial f_2} \cdot \frac{\partial f_2}{\partial d_2} \cdot \frac{\partial d_2}{\partial w_{21}} \\
\frac{\partial L}{\partial w_{22}} &= \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial l_2} \cdot \frac{\partial l_2}{\partial h_2} \cdot \frac{\partial h_2}{\partial z_2} \cdot \frac{\partial z_2}{\partial g_2} \cdot \frac{\partial g_2}{\partial f_2} \cdot \frac{\partial f_2}{\partial e_2} \cdot \frac{\partial e_2}{\partial w_{22}} \\
\frac{\partial L}{\partial b_2} &= \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial l_2} \cdot \frac{\partial l_2}{\partial h_2} \cdot \frac{\partial h_2}{\partial z_2} \cdot \frac{\partial z_2}{\partial g_2} \cdot \frac{\partial g_2}{\partial b_2}
\end{aligned}$$

5.



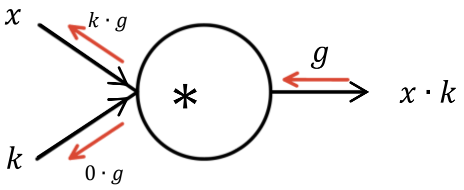
$$\frac{\partial}{\partial x}(x + k) = 1$$

$$\frac{\partial}{\partial k}(x + k) = 0$$



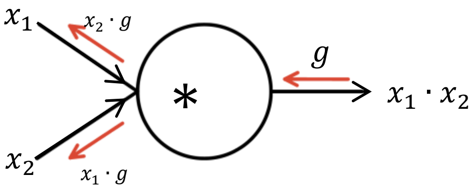
$$\frac{\partial}{\partial x_1}(x_1 + x_2) = 1$$

$$\frac{\partial}{\partial x_2}(x_1 + x_2) = 1$$



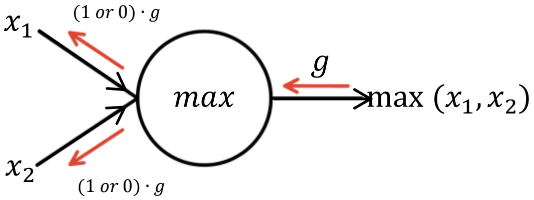
$$\frac{\partial}{\partial x}(x \cdot k) = k$$

$$\frac{\partial}{\partial k}(x \cdot k) = 0$$



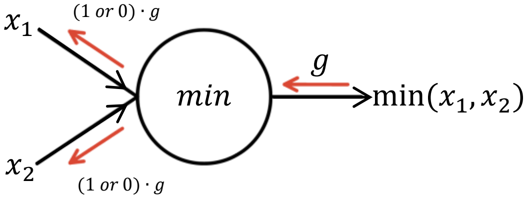
$$\frac{\partial}{\partial x_1}(x_1 \cdot x_2) = x_2$$

$$\frac{\partial}{\partial x_2}(x_1 \cdot x_2) = x_1$$



$$\frac{\partial}{\partial x_1}\max(x_1, x_2) = \begin{Bmatrix} 1 & if\ x_1 \geq x_2 \\ 0 & otherwise \end{Bmatrix}$$

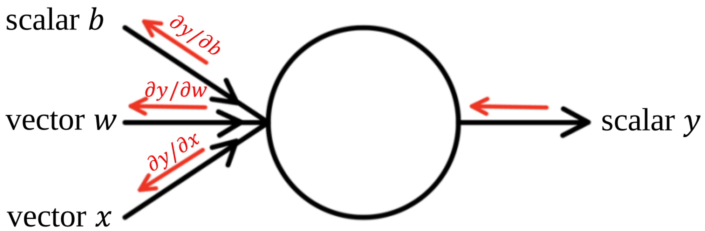
$$\frac{\partial}{\partial x_2}\max(x_1, x_2) = \begin{Bmatrix} 0 & if\ x_1 \geq x_2 \\ 1 & otherwise \end{Bmatrix}$$



$$\frac{\partial}{\partial x_1}\min(x_1, x_2) = \begin{Bmatrix} 1 & if\ x_1 \leq x_2 \\ 0 & otherwise \end{Bmatrix}$$

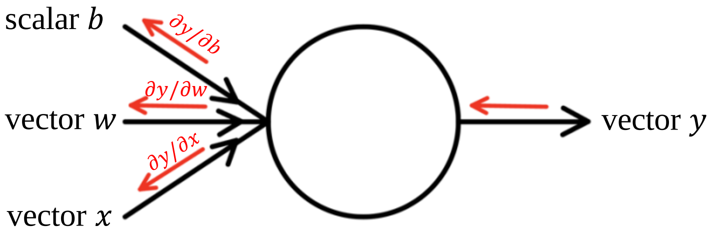
$$\frac{\partial}{\partial x_2}\min(x_1, x_2) = \begin{Bmatrix} 0 & if\ x_1 \leq x_2 \\ 1 & otherwise \end{Bmatrix}$$

6.



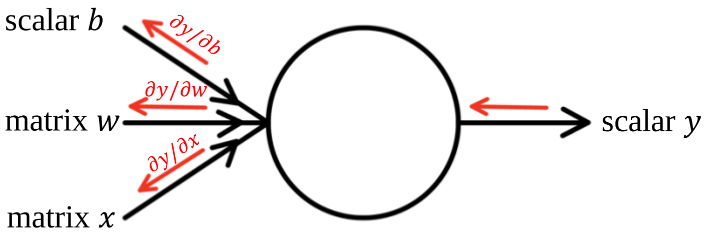
$\partial y/\partial b$: scalar
$\partial y/\partial w$: vector (rank 1 tensor)
$\partial y/\partial x$: vector (rank 1 tensor)

7.



$\partial y/\partial b$: vector (rank 1 tensor)
$\partial y/\partial w$: rank 2 tensor
$\partial y/\partial x$: rank 2 tensor

8.



$\partial y/\partial b$: scalar
$\partial y/\partial w$: rank 2 tensor
$\partial y/\partial x$: rank 2 tensor

9. Given vector-valued functions F and G:

$$(F \circ G)(x) = F\big(G(x)\big)$$

$$\frac{dF}{dx} = \frac{dF}{dG} \cdot \frac{dG}{dx} = F'(G) \cdot G'(x)$$

10. Renamed variables to avoid confusion:

$$G(a,b) = \begin{bmatrix} a - 5b \\ a \cdot b \\ a - b \end{bmatrix} \qquad F(c,d,e) = \begin{bmatrix} 3cd \\ d - e \end{bmatrix}$$

$$(F \circ G)(a,b) = \begin{bmatrix} 3 \cdot (a - 5b) \cdot ab \\ ab - (a - b) \end{bmatrix}$$

$$\frac{dF}{d(a)} = \frac{dF}{dG} \cdot \frac{dG}{da} = F'(G) \cdot G'(a)$$
$$\frac{dF}{d(b)} = \frac{dF}{dG} \cdot \frac{dG}{db} = F'(G) \cdot G'(b)$$

$$\frac{dF}{dG} = \begin{bmatrix} dF_1/\,dG_1 & dF_1/dG_2 & dF_1/dG_3 \\ dF_2/dG_1 & dF_2/dG_2 & dF_2/dG_3 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{d(3\cdot(a-5b)\cdot ab)}{d(a-5b)} & \frac{d(3\cdot(a-5b)\cdot ab)}{d(ab)} & \frac{d\left(3\cdot\left((a-b)-4b\right)\cdot ab\right)}{d(a-b)} \\ \frac{d(ab-(a-b))}{d(a-5b)} & \frac{d\left(ab-(a-b)\right)}{d(ab)} & \frac{d\left(ab-(a-b)\right)}{d(a-b)} \end{bmatrix}$$
$$= \begin{bmatrix} 3ab & 3(a-5b) & 3ab \\ 0 & 1 & -1 \end{bmatrix}$$

$$\frac{dG}{da} = \begin{bmatrix} 1 \\ b \\ 1 \end{bmatrix}, \frac{dG}{db} = \begin{bmatrix} -5 \\ a \\ -1 \end{bmatrix}$$

$$\frac{dF}{da} = \begin{bmatrix} 3ab & 3(a-5b) & 3ab \\ 0 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ b \\ 1 \end{bmatrix} = \begin{bmatrix} 3ab + 3ab - 15b^2 + 3ab \\ b-1 \end{bmatrix} = \begin{bmatrix} 9ab - 15b^2 \\ b-1 \end{bmatrix}$$

$$\frac{dF}{db} = \begin{bmatrix} 3ab & 3(a-5b) & 3ab \\ 0 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} -5 \\ a \\ -1 \end{bmatrix} = \begin{bmatrix} -15ab + 3a^2 - 15ab - 3ab \\ a+1 \end{bmatrix} = \begin{bmatrix} -33ab + 3a^2 \\ a+1 \end{bmatrix}$$

11. The order of the nodes is important because it affects the order of operations in the forward pass, as well as the gradients needed for the chain rule in the backward pass.
12.

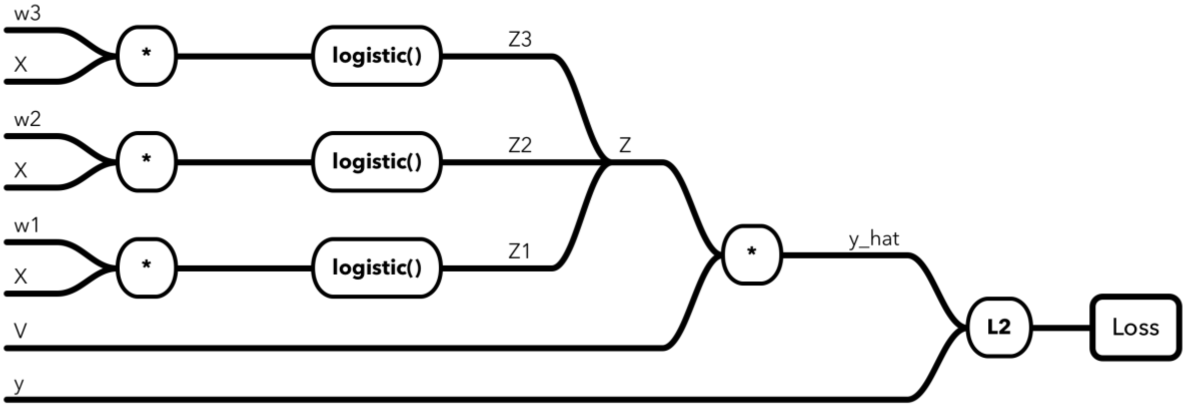
$$X = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 3 \\ 1 & 2 & 2 \end{bmatrix}$$ (prepended column of 1s – bias trick)

$$w_1^T = [0.01 \quad 0.02 \quad 0.03]$$
$$w_2^T = [0.03 \quad 0.01 \quad 0.02]$$
$$w_3^T = [0.02 \quad 0.03 \quad 0.01]$$

$$v = \begin{bmatrix} 0.01 \\ 0.02 \\ 0.03 \\ 0.04 \end{bmatrix}$$

$$y = \begin{bmatrix} 8 \\ 11 \\ 10 \end{bmatrix}$$

Part (a)



Part (b)

For X^*w_1, X^*w_2, and X^*w_3 – prepend rows of X with 1 so we can use the bias trick – call this modified matrix X^*.

$$\begin{aligned}
X^*W_1 &= \begin{bmatrix} 1 & \leftarrow X_1^T \rightarrow \\ 1 & \leftarrow X_2^T \rightarrow \\ 1 & \leftarrow X_3^T \rightarrow \end{bmatrix} \begin{bmatrix} W_{11} \\ W_{12} \\ W_{13} \end{bmatrix} \\
&= \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 3 \\ 1 & 2 & 2 \end{bmatrix} \cdot \begin{bmatrix} 0.01 \\ 0.02 \\ 0.03 \end{bmatrix} \\
&= \begin{bmatrix} 1 \ast 0.01 + 1 \ast 0.02 + 2 \ast 0.03 \\ 1 \ast 0.01 + 1 \ast 0.02 + 3 \ast 0.03 \\ 1 \ast 0.01 + 2 \ast 0.02 + 3 \ast 0.03 \end{bmatrix} \\
&= \begin{bmatrix} 0.09 \\ 0.12 \\ 0.11 \end{bmatrix}
\end{aligned}$$

$$\begin{aligned}
X^*W_2 &= \begin{bmatrix} 1 & \leftarrow X_1^T \rightarrow \\ 1 & \leftarrow X_2^T \rightarrow \\ 1 & \leftarrow X_3^T \rightarrow \end{bmatrix} \begin{bmatrix} W_{21} \\ W_{22} \\ W_{23} \end{bmatrix} \\
&= \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 3 \\ 1 & 2 & 2 \end{bmatrix} \cdot \begin{bmatrix} 0.03 \\ 0.01 \\ 0.02 \end{bmatrix} \\
&= \begin{bmatrix} 1 \ast 0.03 + 1 \ast 0.01 + 2 \ast 0.02 \\ 1 \ast 0.03 + 1 \ast 0.01 + 3 \ast 0.02 \\ 1 \ast 0.03 + 2 \ast 0.01 + 3 \ast 0.02 \end{bmatrix} \\
&= \begin{bmatrix} 0.08 \\ 0.10 \\ 0.09 \end{bmatrix}
\end{aligned}$$

$$\begin{aligned}
X^*W_3 &= \begin{bmatrix} 1 & \leftarrow X_1^T \rightarrow \\ 1 & \leftarrow X_2^T \rightarrow \\ 1 & \leftarrow X_3^T \rightarrow \end{bmatrix} \begin{bmatrix} w_{31} \\ w_{32} \\ w_{33} \end{bmatrix} \\
&= \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 3 \\ 1 & 2 & 2 \end{bmatrix} \cdot \begin{bmatrix} 0.02 \\ 0.03 \\ 0.01 \end{bmatrix} \\
&= \begin{bmatrix} 1 \ast 0.02 + 1 \ast 0.03 + 2 \ast 0.01 \\ 1 \ast 0.02 + 1 \ast 0.03 + 3 \ast 0.01 \\ 1 \ast 0.02 + 2 \ast 0.03 + 3 \ast 0.01 \end{bmatrix} \\
&= \begin{bmatrix} 0.07 \\ 0.08 \\ 0.10 \end{bmatrix}
\end{aligned}$$

$$Z_1 = logistic(X^* W_1) = \begin{bmatrix} logistic(0.09) \\ logistic(0.12) \\ logistic(0.11) \end{bmatrix} = \begin{bmatrix} 0.52248 \\ 0.52996 \\ 0.52747 \end{bmatrix}$$

$$Z_2 = logistic(X^* W_2) = \begin{bmatrix} logistic(0.08) \\ logistic(0.10) \\ logistic(0.09) \end{bmatrix} = \begin{bmatrix} 0.51999 \\ 0.52498 \\ 0.52248 \end{bmatrix}$$

$$Z_3 = logistic(X^* W_3) = \begin{bmatrix} logistic(0.10) \\ logistic(0.08) \\ logistic(0.10) \end{bmatrix} = \begin{bmatrix} 0.51749 \\ 0.51999 \\ 0.52498 \end{bmatrix}$$

$$Z = \begin{bmatrix} Z_1^T \\ Z_2^T \\ Z_3^T \end{bmatrix} = \begin{bmatrix} 0.52248 & 0.52996 & 0.52747 \\ 0.51998 & 0.52497 & 0.52248 \\ 0.52497 & 0.51998 & 0.52497 \end{bmatrix}$$

For calculation of $\hat{y}$ – prepend 1 to rows in Z terms so we can use the bias trick – call this modified matrix Z^*.

$$\begin{aligned} \hat{y} &= Z^* v \\ &= \begin{bmatrix} 1 & \leftarrow Z_1^T \rightarrow \\ 1 & \leftarrow Z_2^T \rightarrow \\ 1 & \leftarrow Z_3^T \rightarrow \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0.52248 & 0.52996 & 0.52747 \\ 1 & 0.51999 & 0.52498 & 0.52248 \\ 1 & 0.51749 & 0.51999 & 0.52498 \end{bmatrix} \cdot \begin{bmatrix} 0.01 \\ 0.02 \\ 0.03 \\ 0.04 \end{bmatrix} \\ &= \begin{bmatrix} 1 * 0.01 + 0.52248 * 0.02 + 0.52996 * 0.03 + 0.52747 * 0.04 \\ 1 * 0.01 + 0.51999 * 0.02 + 0.52498 * 0.03 + 0.52248 * 0.04 \\ 1 * 0.01 + 0.51749 * 0.02 + 0.51999 * 0.03 + 0.52498 * 0.04 \end{bmatrix} \\ &= \begin{bmatrix} 0.0574 \\ 0.0570 \\ 0.0569 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} L &= L_2(\hat{y}, y) \\ &= \begin{bmatrix} L_2(\hat{y}_1, y_1) \\ L_2(\hat{y}_2, y_2) \\ L_2(\hat{y}_1, y_3) \end{bmatrix} \\ &= \begin{bmatrix} (8 - 0.0574)^2 \\ (11 - 0.0570)^2 \\ (10 - 0.0571)^2 \end{bmatrix} \\ &= \begin{bmatrix} 63.0841 \\ 119.7482 \\ 98.9643 \end{bmatrix} \end{aligned}$$

Part (c) and Part (d)

(Apologies here – I got down to the last minute and these parts are mixed)

$$\begin{aligned} L &= L_2(\hat{y}, y) \\ &= \begin{bmatrix} L_2(\hat{y}_1, y_1) \\ L_2(\hat{y}_2, y_2) \\ L_2(\hat{y}_1, y_3) \end{bmatrix} \\ &= \begin{bmatrix} (8 - 0.0574)^2 \\ (11 - 0.0570)^2 \\ (10 - 0.0571)^2 \end{bmatrix} \end{aligned}$$

$$\frac{\partial L}{\partial \hat{y}} = \begin{bmatrix} \frac{\partial}{\partial \hat{y}_1}(\hat{y}_1 - y_1)^2 \\ \frac{\partial}{\partial \hat{y}_2}(\hat{y}_2 - y_2)^2 \\ \frac{\partial}{\partial \hat{y}_3}(\hat{y}_3 - y_3)^2 \end{bmatrix}$$

$$= \begin{bmatrix} 2 \cdot (\hat{y}_1 - y_1) \\ 2 \cdot (\hat{y}_2 - y_2) \\ 2 \cdot (\hat{y}_3 - y_3) \end{bmatrix}$$

$$= \begin{bmatrix} 15.885 \\ 21.886 \\ 19.886 \end{bmatrix}$$

$$\frac{\partial \hat{y}}{\partial Z^*} = v$$

$$= \begin{bmatrix} 0.01 \\ 0.02 \\ 0.03 \\ 0.04 \end{bmatrix}$$

$$\frac{\partial Z^*}{\partial (WX^{*T})} = \frac{\partial \left(logistic\left(WX^{*T}\right)\right)}{\partial WX^{*T}}$$

$$= \begin{bmatrix} WX^{*T}_{11}(1 - WX^{*T}_{11}) & WX^{*T}_{12}(1 - WX^{*T}_{12}) & WX^{*T}_{13}(1 - WX^{*T}_{13}) \\ WX^{*T}_{21}(1 - WX^{*T}_{21}) & WX^{*T}_{22}(1 - WX^{*T}_{22}) & WX^{*T}_{23}(1 - WX^{*T}_{23}) \\ WX^{*T}_{31}(1 - WX^{*T}_{31}) & WX^{*T}_{32}(1 - WX^{*T}_{32}) & WX^{*T}_{33}(1 - WX^{*T}_{33}) \end{bmatrix}$$

$$= \begin{bmatrix} 0.2495 & 0.2491 & 0.2492 \\ 0.2496 & 0.2494 & 0.2495 \\ 0.2497 & 0.2496 & 0.2494 \end{bmatrix}$$

$$\frac{\partial \left(WX^{*T}\right)}{\partial W} = X^*$$

$$= \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 3 \\ 1 & 2 & 2 \end{bmatrix}$$

$$\frac{\partial \hat{y}}{\partial v} = Z^*$$

$$= \begin{bmatrix} 1 & 0.52248 & 0.52996 & 0.52747 \\ 1 & 0.51999 & 0.52498 & 0.52248 \\ 1 & 0.51749 & 0.51999 & 0.52498 \end{bmatrix}$$

$$\frac{\partial L}{\partial v} = Z^{*T} \cdot \frac{\partial L}{\partial \hat{y}}$$

$$= \begin{bmatrix} 1 & 1 & 1 \\ 0.52248 & 0.51999 & 0.51749 \\ 0.52996 & 0.52498 & 0.51999 \\ 0.52747 & 0.52248 & 0.52498 \end{bmatrix} \cdot \begin{bmatrix} 15.885 \\ 21.886 \\ 19.886 \end{bmatrix}$$

$$= \begin{bmatrix} 1 * 15.885 + 1 * 21.886 + 1 * 19.886 \\ 0.52248 * 15.885 + 0.51999 * 21.886 + 0.51749 * 19.886 \\ 0.52996 * 15.885 + 0.52498 * 21.886 + 0.51999 * 19.886 \\ 0.52747 * 15.885 + 0.52248 * 21.886 + 0.52498 * 19.886 \end{bmatrix}$$

$$= \begin{bmatrix} 57.657 \\ 29.971 \\ 30.249 \\ 30.254 \end{bmatrix}$$

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial Z^*} \cdot \frac{\partial Z^*}{\partial (WX^{*T})} \cdot \frac{\partial \left(WX^{*T}\right)}{\partial W}$$

$$= \begin{bmatrix} 2\cdot(\hat{y}_1 - y_1) \\ 2\cdot(\hat{y}_2 - y_2) \\ 2\cdot(\hat{y}_3 - y_3) \end{bmatrix} \cdot v^T \cdot \begin{bmatrix} WX{^*}^T_{11}(1 - WX{^*}^T_{11}) & WX{^*}^T_{12}(1 - WX{^*}^T_{12}) & WX{^*}^T_{13}(1 - WX{^*}^T_{13}) \\ WX{^*}^T_{21}(1 - WX{^*}^T_{21}) & WX{^*}^T_{22}(1 - WX{^*}^T_{22}) & WX{^*}^T_{23}(1 - WX{^*}^T_{23}) \\ WX{^*}^T_{31}(1 - WX{^*}^T_{31}) & WX{^*}^T_{32}(1 - WX{^*}^T_{32}) & WX{^*}^T_{33}(1 - WX{^*}^T_{33}) \end{bmatrix} \cdot X^*$$

$$= \begin{bmatrix} 15.885 \\ 21.886 \\ 19.886 \end{bmatrix} \cdot \begin{bmatrix} 0.01 & 0.02 & 0.03 & 0.04 \end{bmatrix} \cdot \begin{bmatrix} 0.2495 & 0.2491 & 0.2492 \\ 0.2496 & 0.2494 & 0.2495 \\ 0.2497 & 0.2496 & 0.2494 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 3 \\ 1 & 2 & 2 \end{bmatrix}$$

$$\frac{\partial \hat{y}}{\partial Z_1} = \frac{\partial}{\partial Z_1} \begin{bmatrix} 1 & \leftarrow Z_1^T \rightarrow \\ 1 & \leftarrow Z_2^T \rightarrow \\ 1 & \leftarrow Z_3^T \rightarrow \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} 0.02 \\ 0.03 \\ 0.04 \end{bmatrix}$$

$$\frac{\partial Z_1}{\partial W_1} = \frac{\partial}{\partial W_1}\left(logistic(\ X_1^{*T} \cdot W_1\ ) \right)$$

$$= \frac{\partial}{\partial W_1}\left(\begin{bmatrix} logistic({X_{11}^*}^T \cdot W_1) \\ logistic({X_{12}^*}^T \cdot W_1) \\ logistic({X_{13}^*}^T \cdot W_1) \end{bmatrix} \right)$$

$$= \begin{bmatrix} Z_{11} \cdot (1 - Z_{11}) \cdot {X_{11}^*}^T \\ Z_{12} \cdot (1 - Z_{12}) \cdot {X_{12}^*}^T \\ Z_{13} \cdot (1 - Z_{13}) \cdot {X_{13}^*}^T \end{bmatrix} = \begin{bmatrix} 0.2495 \cdot 1 \\ 0.2491 \cdot 1 \\ 0.2492 \cdot 2 \end{bmatrix} = \begin{bmatrix} 0.2495 \\ 0.2491 \\ 0.4984 \end{bmatrix}$$

Extending the above logic:

$$\frac{\partial \hat{y}}{\partial Z_2} = \begin{bmatrix} v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} 0.02 \\ 0.03 \\ 0.04 \end{bmatrix}$$

$$\frac{\partial \hat{y}}{\partial Z_3} = \begin{bmatrix} v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} 0.02 \\ 0.03 \\ 0.04 \end{bmatrix}$$

$$\frac{\partial Z_2}{\partial W_2} = \begin{bmatrix} Z_{21} \cdot (1 - Z_{21}) \cdot {X_{21}^*}^T \\ Z_{22} \cdot (1 - Z_{22}) \cdot {X_{22}^*}^T \\ Z_{23} \cdot (1 - Z_{23}) \cdot {X_{23}^*}^T \end{bmatrix} \begin{bmatrix} 0.2496 \cdot 1 \\ 0.2494 \cdot 1 \\ 0.2495 \cdot 3 \end{bmatrix} = \begin{bmatrix} 0.2496 \\ 0.2495 \\ 0.7485 \end{bmatrix}$$

$$\frac{\partial Z_3}{\partial W_3} = \begin{bmatrix} Z_{31} \cdot (1 - Z_{31}) \cdot {X_{31}^*}^T \\ Z_{32} \cdot (1 - Z_{32}) \cdot {X_{32}^*}^T \\ Z_{33} \cdot (1 - Z_{33}) \cdot {X_{33}^*}^T \end{bmatrix} \begin{bmatrix} 0.2494 \cdot 1 \\ 0.2496 \cdot 2 \\ 0.2494 \cdot 2 \end{bmatrix} = \begin{bmatrix} 0.2494 \\ 0.4992 \\ 0.4988 \end{bmatrix}$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_1}$$

$$= 57.6568 \cdot dot\left(\begin{bmatrix} 0.02 \\ 0.03 \\ 0.04 \end{bmatrix} \cdot \begin{bmatrix} 0.2495 \\ 0.2491 \\ 0.4984 \end{bmatrix} \right)$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z_2} \cdot \frac{\partial z_2}{\partial w_2}$$

$$= 57.6568 \cdot dot\left(\begin{bmatrix} 0.02 \\ 0.03 \\ 0.04 \end{bmatrix} \cdot \begin{bmatrix} 0.2496 \\ 0.2495 \\ 0.7485 \end{bmatrix} \right)$$

$$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z_3} \cdot \frac{\partial z_3}{\partial w_3}$$

$$= 57.6568 \cdot dot\left(\begin{bmatrix} 0.02 \\ 0.03 \\ 0.04 \end{bmatrix} \cdot \begin{bmatrix} 0.2494 \\ 0.4992 \\ 0.4988 \end{bmatrix} \right)$$

$$\frac{\partial L}{\partial v} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial v}$$

$$= 57.6568 \cdot \begin{bmatrix} 1 & 0.52248 & 0.52996 & 0.52747 \\ 1 & 0.51998 & 0.52497 & 0.52248 \\ 1 & 0.52497 & 0.51998 & 0.52497 \end{bmatrix}$$

13.

a. $f(x,y) = (2x+3y)^2$

$$\nabla f(x,y) = \begin{bmatrix} \partial f/\partial x \\ \partial f/\partial y \end{bmatrix} = \begin{bmatrix} 2(2x+3y)\cdot 2 \\ 2(2x+3y)\cdot 3 \end{bmatrix} = \begin{bmatrix} 8x+12y \\ 12x+18y \end{bmatrix}$$

b. $F(x,y) = \begin{bmatrix} x^2+2y \\ 3x+4y^2 \end{bmatrix}$

$$\nabla DF(1,2) = \begin{bmatrix} \partial f_1/\partial x & \partial f_1/\partial y \\ \partial f_2/\partial x & \partial f_2/\partial y \end{bmatrix} = \begin{bmatrix} 2x & 2 \\ 3 & 8y \end{bmatrix}$$

c. $G(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$

Without chain rule:

$$F \circ G = \begin{bmatrix} (x)^2 + 2(x^2) \\ 3(x) + 4(x^2)^2 \end{bmatrix} = \begin{bmatrix} 3x^2 \\ 3x+4x^4 \end{bmatrix}$$

$$Jacobian = \begin{bmatrix} \frac{d}{dx} 3x^2 \\ \frac{d}{dx}(3x+4x^4) \end{bmatrix} = \begin{bmatrix} 6x \\ 3+16x^3 \end{bmatrix}$$

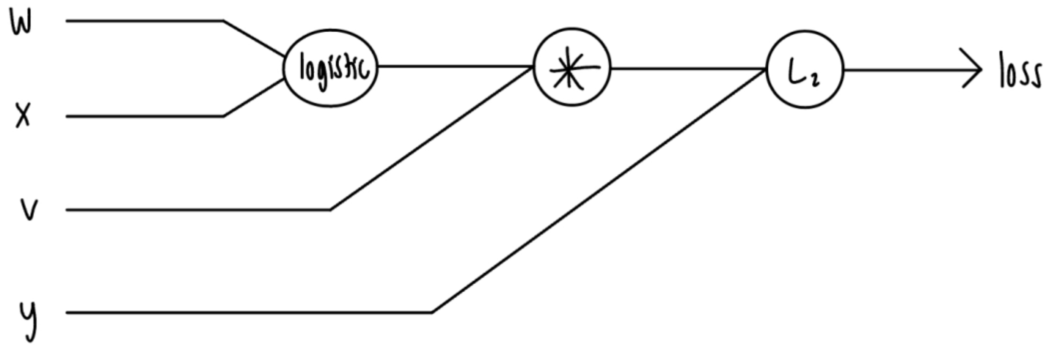
With chain rule:

$$DF\big(G(x)\big) = DF \circ G = \begin{bmatrix} 2(x) & 2 \\ 3 & 8(x^2) \end{bmatrix} = \begin{bmatrix} 2x & 2 \\ 3 & 8x^2 \end{bmatrix}$$

$$DG(x) = \begin{bmatrix} \frac{d}{dx}(x) \\ \frac{d}{dx}(x^2) \end{bmatrix} = \begin{bmatrix} 1 \\ 2x \end{bmatrix}$$

$$Jacobian = DF\big(DG(x)\big) = \tfrac{DF}{DG} \cdot \tfrac{DG}{Dx} = \begin{bmatrix} 2x & 2 \\ 3 & 8x^2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 2x \end{bmatrix} = \begin{bmatrix} 4x+2x \\ 3+16x^3 \end{bmatrix} = \begin{bmatrix} 6x \\ 3+16x^3 \end{bmatrix}$$

d.



Forward pass:

$$WX^{*T} = \begin{bmatrix} 0.01 & 0.02 & 0.03 \\ 0.03 & 0.01 & 0.02 \\ 0.02 & 0.03 & 0.01 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 2 & 3 & 2 \end{bmatrix} = \begin{bmatrix} 0.09 & 0.12 & 0.11 \\ 0.08 & 0.10 & 0.09 \\ 0.07 & 0.08 & 0.10 \end{bmatrix}$$

$$Z = logistic\left(WX^{*T}\right) = \begin{bmatrix} 0.522 & 0.530 & 0.527 \\ 0.520 & 0.525 & 0.522 \\ 0.525 & 0.520 & 0.525 \end{bmatrix}$$

$$\hat{y} = Z^{*}v = \begin{bmatrix} 1 & 0.522 & 0.530 & 0.527 \\ 1 & 0.520 & 0.525 & 0.522 \\ 1 & 0.525 & 0.520 & 0.525 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix} = \begin{bmatrix} 0.0574 \\ 0.0570 \\ 0.0571 \end{bmatrix}$$

$$\hat{y} = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 = (8 - 0.0574)^2 + (11 - 0.0570)^2 + (10 - 0.0571)^2 = 287.694$$

Backward pass:

$$\frac{\partial L}{\partial W} = \frac{DL}{D\hat{y}} \cdot \frac{D\hat{y}}{DZ} \cdot \frac{DZ}{DW}$$

$$L = f(\hat{y}, y) = L_2(\hat{y}, y)$$

$$\frac{DL}{D\hat{y}} \quad \rightarrow \quad scalar = \frac{\partial L_2(\hat{y}, y)}{\partial \hat{y}}$$

$$\hat{y} = f(Z^{*}, v) = Z^{*} \cdot v = v_1 + Z \cdot \begin{bmatrix} v2 \\ v3 \\ v4 \end{bmatrix}$$

$$\frac{D\hat{y}}{DZ} \quad \rightarrow \quad Jacobian = \begin{bmatrix} \partial\left(v_1 + Z \cdot \begin{bmatrix} v2 \\ v3 \\ v4 \end{bmatrix}\right) / (\ \partial Z_1^{*}\ ) \\ \partial\left(v_1 + Z \cdot \begin{bmatrix} v2 \\ v3 \\ v4 \end{bmatrix}\right) / (\ \partial Z_2^{*}\ ) \\ \partial\left(v_1 + Z \cdot \begin{bmatrix} v2 \\ v3 \\ v4 \end{bmatrix}\right) / (\ \partial Z_3^{*}\ ) \\ \partial\left(v_1 + Z \cdot \begin{bmatrix} v2 \\ v3 \\ v4 \end{bmatrix}\right) / (\ \partial Z_4^{*}\ ) \end{bmatrix} = \begin{bmatrix} v_2 \\ v_3 \\ v_4 \end{bmatrix}$$

$$Z^{*} = f(X^{*}, W) = logistic\left(WX^{*T}\right)$$

$$\frac{DZ^{*}}{DW} \quad \rightarrow \quad Jacobian =$$

$$\begin{bmatrix} \frac{\partial}{\partial W_1} logistic\left(WX^{*T}\right) \\ \frac{\partial}{\partial W_2} logistic\left(WX^{*T}\right) \\ \frac{\partial}{\partial W_3} logistic\left(WX^{*T}\right) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial WX^T} logistic\left(WX^{*T}\right) \cdot \frac{\partial WX^T}{\partial W} \cdot \frac{\partial W}{\partial W_1} = top\ row\ of\ WXt(1 - WXt) * XT \\ \frac{\partial}{\partial W_2} logistic\left(WX^{*T}\right) \\ \frac{\partial}{\partial W_3} logistic\left(WX^{*T}\right) \end{bmatrix}$$

$$\frac{\partial L}{\partial W} = \frac{DL}{D\hat{y}} \cdot \frac{D\hat{y}}{DZ^{*}} \cdot \frac{DZ^{*}}{DW}$$

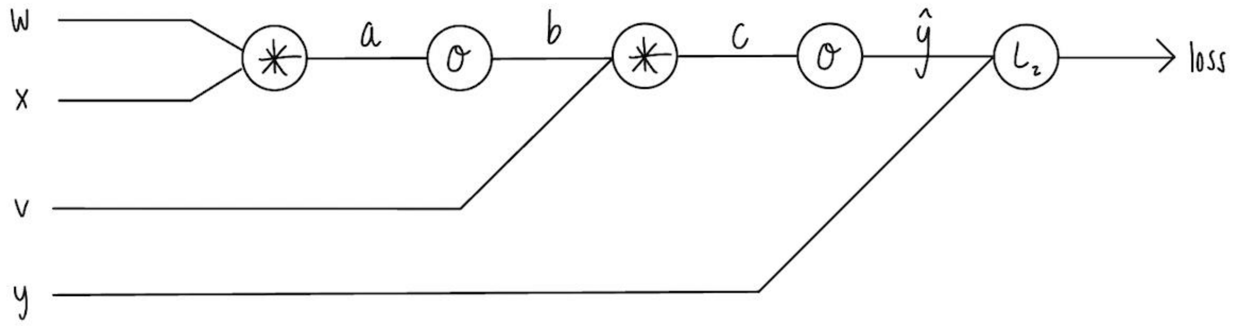
$$= \sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2 \cdot \begin{bmatrix} v_2 \\ v_3 \\ v_4 \end{bmatrix} \cdot \begin{bmatrix} \begin{bmatrix} Z_{11}(1-Z_{11}) \cdot X_1^* \\ Z_{12}(1-Z_{12}) \cdot X_1^* \\ Z_{13}(1-Z_{13}) \cdot X_1^* \end{bmatrix} \\ \begin{bmatrix} Z_{21}(1-Z_{21}) \cdot X_2^* \\ Z_{22}(1-Z_{22}) \cdot X_2^* \\ Z_{33}(1-Z_{23}) \cdot X_2^* \end{bmatrix} \\ \begin{bmatrix} Z_{31}(1-Z_{31}) \cdot X_3^* \\ Z_{32}(1-Z_{32}) \cdot X_3^* \\ Z_{33}(1-Z_{33}) \cdot X_3^* \end{bmatrix} \end{bmatrix}$$

$$= 57.6568 \cdot \begin{bmatrix} 0.02 \\ 0.03 \\ 0.04 \end{bmatrix} \cdot \begin{bmatrix} \begin{bmatrix} 0.2495 \\ 0.2491 \\ 0.4984 \end{bmatrix} \\ \begin{bmatrix} 0.2496 \\ 0.2404 \\ 0.7485 \end{bmatrix} \\ \begin{bmatrix} 0.2494 \\ 0.4992 \\ 0.4988 \end{bmatrix} \end{bmatrix}$$

$$= 57.6568 \cdot \begin{bmatrix} 0.06 & 0.06 & 0.12 \\ 0.02 & 0.03 & 0.05 \\ 0.03 & 0.04 & 0.07 \end{bmatrix}$$

$$= \begin{bmatrix} 3.46 & 3.46 & 6.92 \\ 1.15 & 1.73 & 2.88 \\ 1.73 & 2.31 & 4.04 \end{bmatrix}$$

14.



$$L = L_2(\hat{y} - y)$$

$$\frac{\partial L}{\partial \hat{y}} = 2 \cdot \frac{1}{2}(\hat{y} - y) = (\hat{y} - y)$$

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

$$\frac{\partial c}{\partial b} = V$$

$$\frac{\partial c}{\partial V} = b$$

$$\frac{\partial b}{\partial a} = a \cdot (1 - a)$$

$$\frac{\partial a}{\partial W} = X$$

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial c} \cdot \frac{\partial c}{\partial b} \cdot \frac{\partial b}{\partial a} \cdot \frac{\partial a}{\partial W} = (\hat{y} - y) \cdot c \cdot (1 - c) \cdot V \cdot a \cdot (1 - a)$$

$$\frac{\partial L}{\partial V} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial c} \cdot \frac{\partial c}{\partial V} = (\hat{y} - y) \cdot c \cdot (1 - c) \cdot b$$