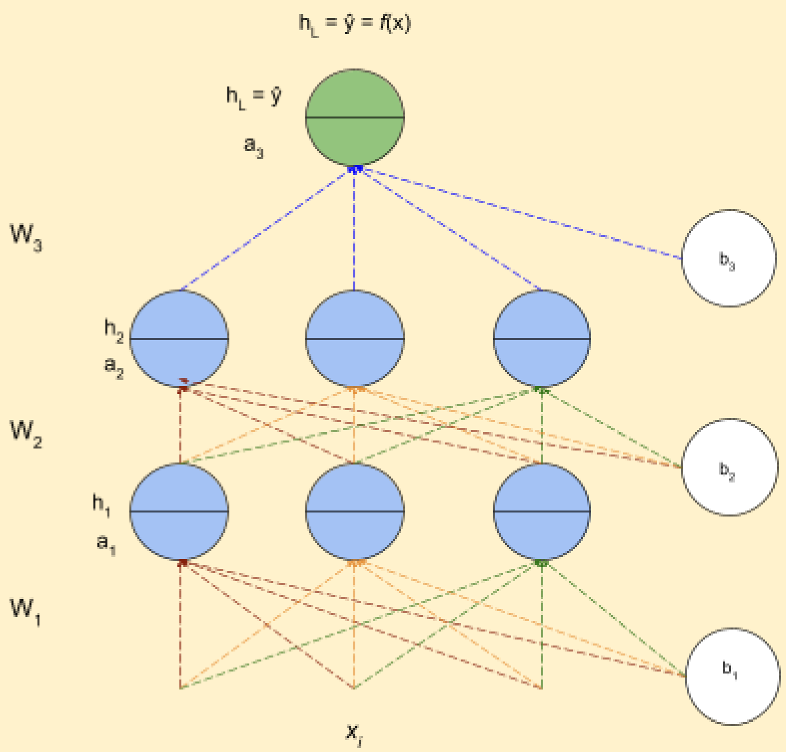
Backpropagation (Light Math)

Setting the context

Can we use the same learning algorithm as before?

1. Here is the learning algorithm as discussed in the previous chapter, the no-math version
2. Consider the Neural Network with the following configuration



3. The algorithm
   a. **Initialise:** w_{111}, w_{112}, ... w_{313}, b_1, b_2, b_3 randomly
   b. **Iterate over data**
      i. Compute $\hat{y}$
      ii. Compute L(w,b) Cross-entropy loss function
      iii. $w_{111} = w_{111} - \eta \Delta w_{111}$
      iv. $w_{112} = w_{112} - \eta \Delta w_{112}$
      ...
      v. $w_{313} = w_{111} - \eta \Delta w_{313}$
      vi. $b_i = b_i + \eta \Delta b_i$
      vii. Pytorch/Tensorflow have functions to compute $\frac{\delta l}{\delta w}$ *and* $\frac{\delta l}{\delta b}$
   c. **Till satisfied**
      i. Number of epochs is reached (ie 1000 passes/epochs)
      ii. Continue till Loss < ε (some defined value)

PadhAI: Backpropagation - the light math version

One Fourth Labs

---

4. In this section, we will be looking at the light-math version, where we will be computing the derivatives
5. Derivatives for all layers from 1 to L

$$\begin{bmatrix} \frac{\partial \mathscr{L}(\theta)}{\partial W_{111}} & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial W_{11n}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{211}} & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial W_{21n}} & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial W_{L,11}} & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial W_{L,1k}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{L,1k}} & \frac{\partial \mathscr{L}(\theta)}{\partial b_{11}} & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial b_{L1}} \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{121}} & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial W_{12n}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{221}} & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial W_{22n}} & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial W_{L,21}} & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial W_{L,2k}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{L,2k}} & \frac{\partial \mathscr{L}(\theta)}{\partial b_{12}} & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial b_{L2}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial \mathscr{L}(\theta)}{\partial W_{1n1}} & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial W_{1nn}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{2n1}} & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial W_{2nn}} & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial W_{L,n1}} & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial W_{L,nk}} & \frac{\partial \mathscr{L}(\theta)}{\partial W_{L,nk}} & \frac{\partial \mathscr{L}(\theta)}{\partial b_{1n}} & \cdots & \frac{\partial \mathscr{L}(\theta)}{\partial b_{Lk}} \end{bmatrix}$$

Layer 1 Layer 2 Output Layer nxk weights Bias terms

6. Once we know the gradients, we can use them in the Gradient Descent algorithm to compute the weights of the network

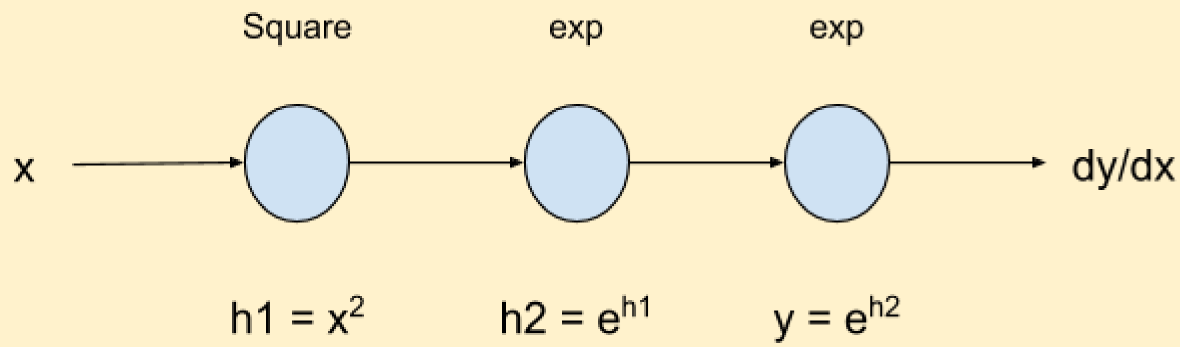
PadhAI: Backpropagation - the light math version

One Fourth Labs

Revisiting Basic Calculus

Let's do a quick recap of some basic calculus concepts

1. Here are some examples of simple derivatives
   a. $\frac{de^x}{dx} = e^x$
   b. $\frac{dx^2}{dx} = 2x$
   c. $\frac{d(\frac{1}{x})}{dx} =- \frac{1}{x^2}$
2. Now, let's look at a slightly more complicated derivative
   a. a $\frac{de^{x^2}}{dx}$
   b. Here, we break it into two parts
      i. h = x^2
      ii. y = $e^{(term)}$
   c. Therefore, $\frac{de^{x^2}}{dx} = \frac{dy}{dh}\frac{dh}{dx}$
   d. $\frac{dh}{dx} = 2x$
   e. $\frac{dy}{dh} = e^h$
   f. $\frac{de^{x^2}}{dx} = \frac{dy}{dh}\frac{dh}{dx} = (e^h).(2x) = (e^{x^2}).(2x) = 2xe^{x^2}$
   g. Here, the output is a composite function of the input. This process of breaking the equation into parts and solving them sequentially is known as **Chain Rule**
   h. Consider another example $\frac{de^{e^{x^2}}}{dx}$
   i. Here is the flow diagram of chain rule applied to the above equation

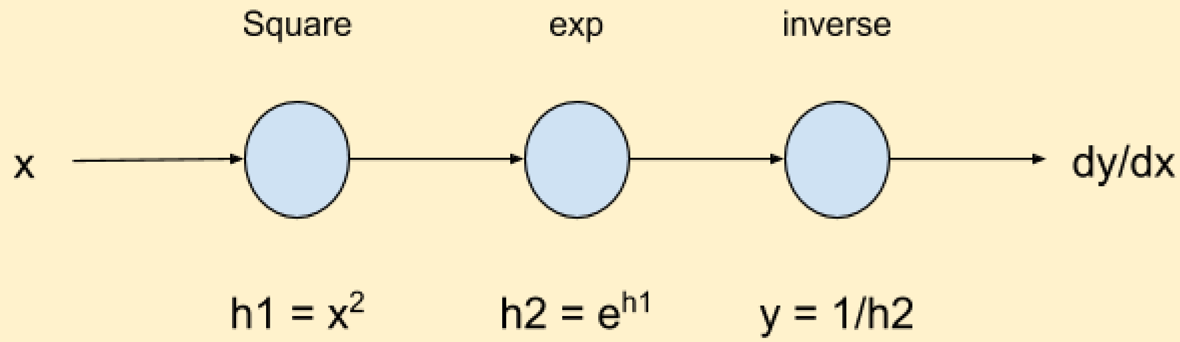


   j. $\frac{de^{e^{x^2}}}{dx} = \frac{dy}{dh2}\frac{dh2}{dh1}\frac{dh1}{dx} = (e^{h2}).(e^{h1}).(2x) = (e^{e^{x^2}}).(e^{x^2}).(2x) = 2xe^{e^{x^2}}e^{x^2}$
   k. Another example $\frac{d(\frac{1}{e^{x^2}})}{dx}$

PadhAI: Backpropagation - the light math version

One Fourth Labs

l. Flow diagram of chain rule

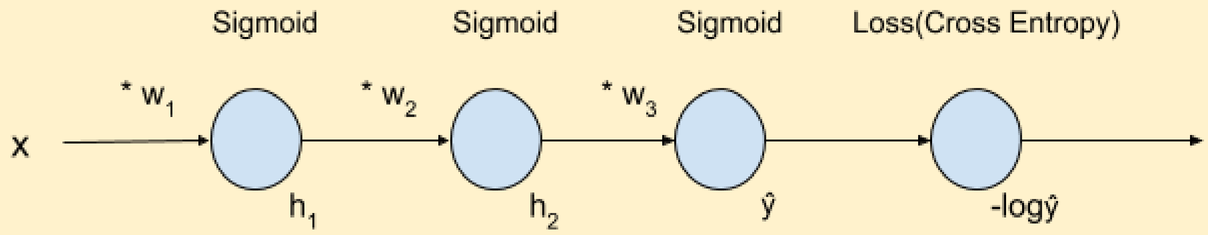


m. $\frac{d(\frac{1}{e^{x^2}})}{dx} = \frac{dy}{dh2}\frac{dh2}{dh1}\frac{dh1}{dx} = (\frac{-1}{(h2)^2}).(e^{h1}).(2x) = (\frac{-1}{(e^{x^2})^2}).(e^{x^2}).(2x) = (\frac{-1}{(e^{x^2})^2}).2xe^{x^2}$

Why do we care about the chain rule of derivatives

Importance of chain rule in Deep Learning

1. Let us look at a sample chain rule flow of a shallow neural network



2. Here, the output ŷ is a composite dependent on input x and all of the parameters w
3. $Loss\,function:\ L = f(x,\ w_1,\ w_2,\ w_3)$
4. Now, for the gradient, we want the derivative of the loss function with respect to the various weights $\frac{\partial L}{\partial w_i}$
5. If we want the derivative w.r.t w_2 then we do the following $\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial h_2} \frac{\partial h_2}{\partial w_2}$
6. Here, computation happens from input layer to the output layer ie forward propagation
7. Derivative calculation happens backwards from the output layer to the input, ie back propagation

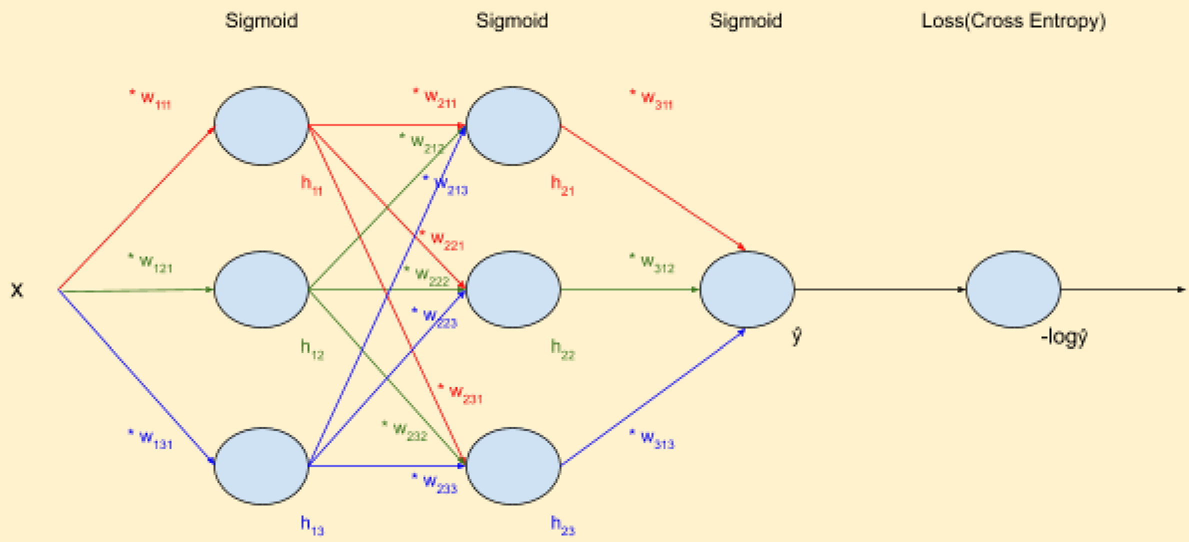
PadhAI: Backpropagation - the light math version

One Fourth Labs

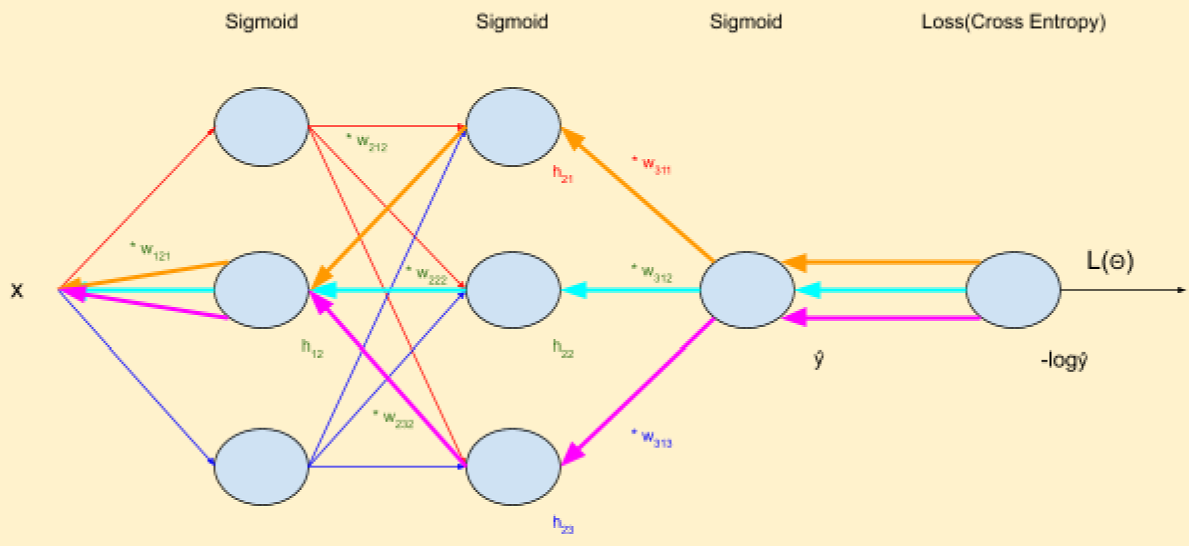
Applying chain rule across multiple paths

Importance of chain rule in deep learning

1. Let us look at a more complex neural network



2. In the shallow Neural Network from the previous example, we apply the chain rule along a straight path. However, in a more practical Neural Network as shown above, the chain rule needs to be applied across multiple parallel paths in order to find a particular gradient
3. For example, to calculate $\frac{\partial L}{\partial w_{121}}$ we need to operate along 3 different paths



4. Summing up the derivatives across the three paths (cyan, orange and pink) will give us the required derivative $\frac{\partial L}{\partial w_{121}}$
5. This scales across as many paths as there are in the neural network.
6. Here, these are not regular derivatives but partial derivatives.

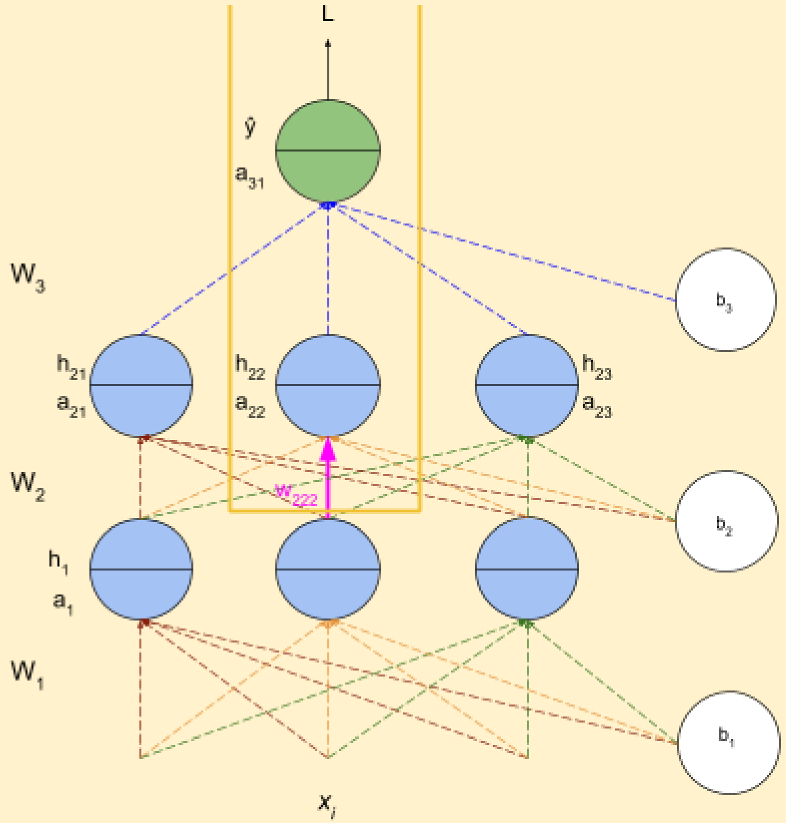
PadhAI: Backpropagation - the light math version

One Fourth Labs

Applying chain rule in a neural network

How many derivatives do we need to compute and how do we compute them?

1. Let's focus on the highlighted weight(w_{222}) of the following neural network



2. To learn this weight, we have to compute the partial derivative w.r.t loss function a

$$(w_{222})_{t+1} = (w_{222})_t - \eta * (\frac{\partial L}{\partial w_{222}})$$

3. We can calculate $\frac{\partial L}{\partial w_{222}}$ as follows

   a. $\frac{\partial L}{\partial w_{222}} = (\frac{\partial L}{\partial a_{22}}).(\frac{\partial a_{22}}{\partial w_{222}})$

   b. $\frac{\partial L}{\partial w_{222}} = (\frac{\partial L}{\partial h_{22}}).(\frac{\partial h_{22}}{\partial a_{22}}).(\frac{\partial a_{22}}{\partial w_{222}})$

   c. $\frac{\partial L}{\partial w_{222}} = (\frac{\partial L}{\partial a_{31}}).(\frac{\partial a_{31}}{\partial h_{22}}).(\frac{\partial h_{22}}{\partial a_{22}}).(\frac{\partial a_{22}}{\partial w_{222}})$

   d. $\frac{\partial L}{\partial w_{222}} = (\frac{\partial L}{\partial \hat{y}}).(\frac{\partial \hat{y}}{\partial a_{31}}).(\frac{\partial a_{31}}{\partial h_{22}}).(\frac{\partial h_{22}}{\partial a_{22}}).(\frac{\partial a_{22}}{\partial w_{222}})$

4. Thus, by breaking the partial derivative into all the subdivisions along that path and multiplying it, we will get the required solution.

PadhAI: Backpropagation - the light math version

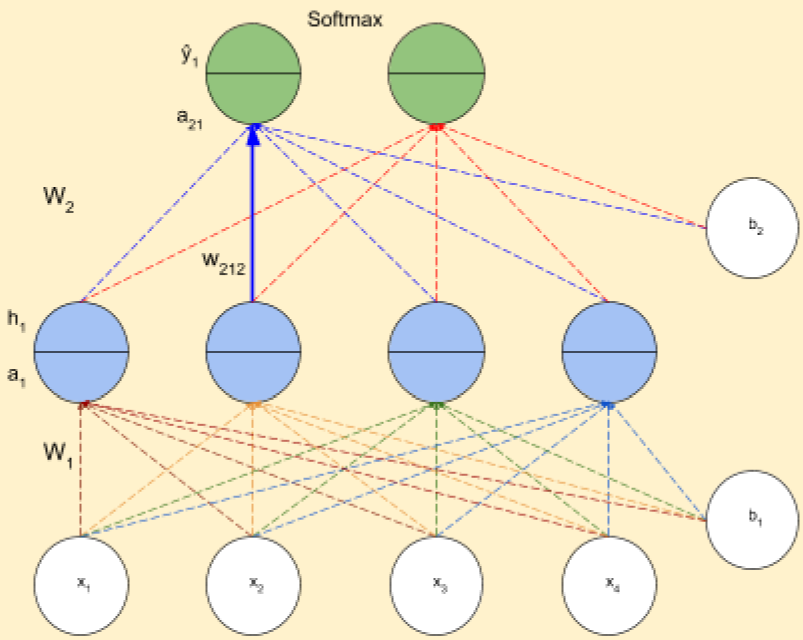
One Fourth Labs

Partial Derivatives with respect to a

Part 1

How do we compute partial derivatives

1. The following neural network will be used to demonstrate the calculations



2. Here are the parameters of the network
   a. $b = [\,0.5 \;\; 0.3\,]$
   b.

$$W_1 = \begin{bmatrix} 0.1 & 0.3 & 0.8 & -0.4 \\ -0.3 & -0.2 & 0.5 & 0.5 \\ -0.3 & 0 & 0.5 & 0.4 \\ 0.2 & 0.5 & -0.9 & 0.7 \end{bmatrix}$$

   c.

$$W_2 = \begin{bmatrix} 0.5 & 0.8 & 0.2 & 0.4 \\ 0.5 & 0.2 & 0.3 & -0.5 \end{bmatrix}$$

   d. $x = [\,2 \;\; 5 \;\; 3 \;\; 3\,]$ true distribution $y = [\,1 \;\; 0\,]$

---

3. Now, we want to find the partial derivative w.r.t w_{212} as highlighted in the figure $\frac{\partial L}{\partial w_{212}}$
4. $\frac{\partial L}{\partial w_{212}} = (\frac{\partial L}{\partial a_{21}}).(\frac{\partial a_{21}}{\partial w_{212}}) = (\frac{\partial L}{\partial \hat{y}_1}).(\frac{\partial \hat{y}_1}{\partial a_{21}}).(\frac{\partial a_{21}}{\partial w_{212}})$
5. We will solve the above equation sequentially
   a. Consider square error loss function L
   b. $\frac{\partial L}{\partial \hat{y}_1} = \Sigma^2{}_{i=1}(y_i - \hat{y}_i)^2$
      i. $\frac{\partial L}{\partial \hat{y}_1} = \frac{\partial}{\partial y_1}[(y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2]$
      ii. Here, the y_2 terms get cancelled, leaving $\frac{\partial}{\partial y_1}[(y_1 - \hat{y}_1)^2] = -2(y_1 - \hat{y}_1)$

PadhAI: Backpropagation - the light math version

One Fourth Labs

Partial Derivatives with respect to a

Part 2

How do we compute partial derivatives

1. Let us continue calculating the partial derivative of L w.r.t w_{212}
2. Solving the equation sequentially
    a. Let's look at the second partial derivative $\frac{\partial \hat{y}_1}{\partial a_{21}}$
        i. Here, $\hat{y}_1 = (\frac{e^{a_{21}}}{e^{a_{21}} + e^{a_{22}}})$, this is the softmax applied on a_{21}
        ii. To make it easier to compute, multiply both numerator and denominator by $e^{-a_{21}}$
        iii. $\hat{y}_1 = (\frac{e^{-a_{21}}}{e^{-a_{21}}})(\frac{e^{a_{21}}}{e^{a_{21}} + e^{a_{22}}}) = \frac{1}{1 + e^{-(a_{21} - a_{22})}}$
        iv. $\frac{\partial \hat{y}_1}{\partial a_{21}} = \frac{\partial}{\partial a_{21}}(\frac{1}{1 + e^{-(a_{21} - a_{22})}})$
        v. $\frac{\partial \hat{y}_1}{\partial a_{21}} = (\frac{-1}{(1 + e^{-(a_{21} - a_{22})})^2}).(1).(e^{-(a_{21} - a_{22})}).(-1) = (\frac{1}{1 + e^{-(a_{21} - a_{22})}}).(\frac{e^{-(a_{21} - a_{22})}}{1 + e^{-(a_{21} - a_{22})}})$
        vi. Rewriting the terms $\frac{\partial \hat{y}_1}{\partial a_{21}} = \hat{y}_1(1 - \hat{y}_1)$

PadhAI: Backpropagation - the light math version

One Fourth Labs

Partial Derivatives with respect to a

Part 3

How do we compute partial derivatives?

1. Solving the equation sequentially
   a. Let's look at the third partial derivative $\frac{\partial a_{21}}{\partial w_{212}}$
      i. Here $a_{21} = w_{211}h_{11} + w_{212}h_{12} + w_{213}h_{13} + w_{214}h_{14}$
      ii. $\frac{\partial a_{21}}{\partial w_{212}} = h_{12}$, as all other terms cancel out.
2. Consider the following output values
   a. $a_1 = W_1 * x + b_1 = [2.9\ \ 1.4\ \ 2.1\ \ 2.3]$
   b. $h_1 = sigmoid(a_1) = [0.95\ \ 0.80\ \ 0.89\ \ 0.91]$
   c. $a_2 = W_2 * h_1 + b_2 = [1.66\ \ 0.45]$
   d. $\hat{y} = softmax(a_2) = [0.77\ \ 0.23]$
   e. Squared error loss $L(\Theta) = (1 - 0.77)^2 + (1 - 0.23)^2 = 0.1058$
3. Substituting these values in our formulae
   a. $\frac{\partial L}{\partial \hat{y}_1} = -2(y_1 - \hat{y}_1) = -0.46$
   b. $\frac{\partial \hat{y}_1}{\partial a_{21}} = \hat{y}_1(1 - \hat{y}_1) = 0.1771$
   c. $\frac{\partial a_{21}}{\partial w_{212}} = h_{12} = 0.8$
   d. $\frac{\partial L}{\partial w_{212}} = (-2(y_1 - \hat{y}_1)) * (\hat{y}_1(1 - \hat{y}_1)) * (h_{12}) = (-0.46) * (0.1771) * (0.8) = -0.065$
4. Now we can calculate the updated value of w_{212}
5. $w_{212} = w_{212} - \eta(\frac{\partial L}{\partial w_{212}})$
   a. $w_{212} = 0.8 - (1) * (-0.065)$
   b. $w_{212} = 0.865$
6. We can repeat this process for each weight.

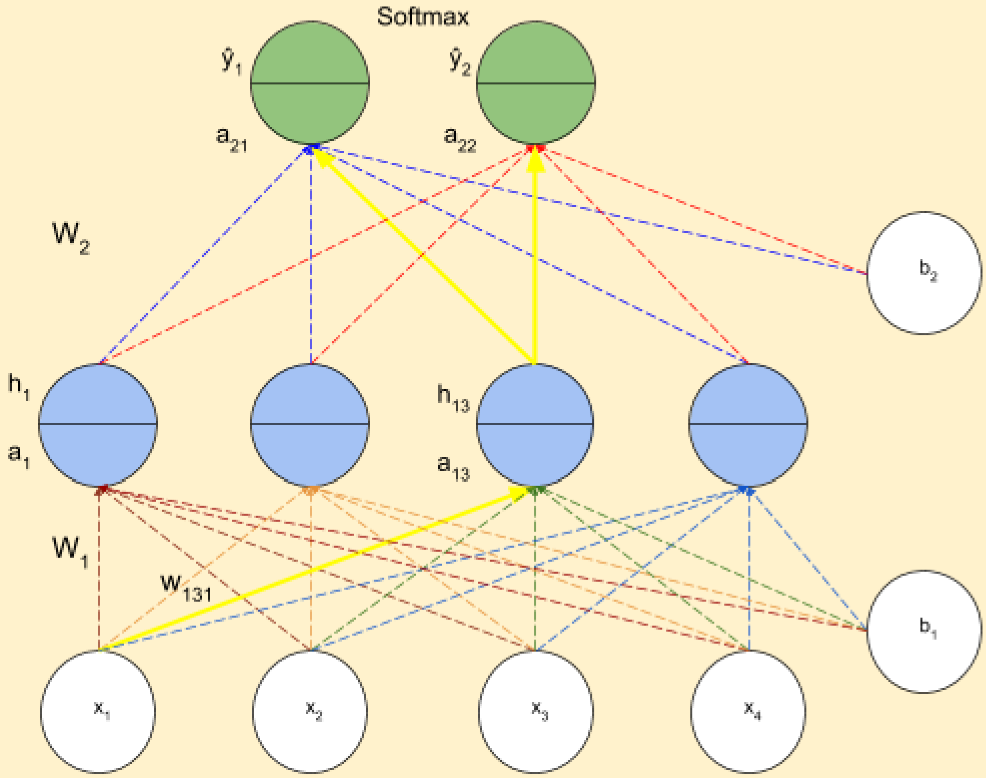
PadhAI: Backpropagation - the light math version

One Fourth Labs

Multiple Paths

Can we see one more example?

1. Let's look at a different weight from the previous example, which would require multiple paths to perform the calculations



2. Here are the parameters of the network
   a. $b = [\,0 \;\; 0\,]$
   b.

$$W_1 = \begin{bmatrix} 0.1 & 0.3 & 0.8 & -0.4 \\ -0.3 & -0.2 & 0.5 & 0.5 \\ -0.3 & 0 & 0.5 & 0.4 \\ 0.2 & 0.5 & -0.9 & 0.7 \end{bmatrix}$$

c.

$$W_2 = \begin{bmatrix} 0.5 & 0.8 & 0.2 & 0.4 \\ 0.5 & 0.2 & 0.3 & -0.5 \end{bmatrix}$$

d. x = [2 5 3 3] true distribution y = [1 0]

3. Now, we want to find the partial derivative w.r.t w_{212} as highlighted in the figure $\frac{\partial L}{\partial w_{212}}$
4. $\frac{\partial L}{\partial w_{131}} = (\frac{\partial L}{\partial a_{13}}).(\frac{\partial a_{13}}{\partial w_{131}}) = (\frac{\partial L}{\partial h_{13}}).(\frac{\partial h_{13}}{\partial a_{13}}).(\frac{\partial a_{13}}{\partial w_{131}}) = (\frac{\partial L}{\partial a_{21}}.\frac{\partial a_{21}}{\partial h_{13}} + \frac{\partial L}{\partial a_{22}}.\frac{\partial a_{22}}{\partial h_{13}}).(\frac{\partial h_{13}}{\partial a_{13}}).(\frac{\partial a_{13}}{\partial w_{131}})$ a
5. The final split is $\frac{\partial L}{\partial w_{131}} = (\frac{\partial L}{\partial \hat{y}_1}.\frac{\partial \hat{y}_1}{\partial a_{21}}.\frac{\partial a_{21}}{\partial h_{13}} + \frac{\partial L}{\partial \hat{y}_2}.\frac{\partial \hat{y}_2}{\partial a_{22}}.\frac{\partial a_{22}}{\partial h_{13}}).(\frac{\partial h_{13}}{\partial a_{13}}).(\frac{\partial a_{13}}{\partial w_{131}})$
6. Let us sequentially solve both splits

$\frac{\partial L}{\partial \hat{y}_1} = -2(y_1 - \hat{y}_1) = -0.46$	$\frac{\partial L}{\partial \hat{y}_2} = -2(y_2 - \hat{y}_2) = 0.46$
$\frac{\partial \hat{y}_1}{\partial a_{21}} = \hat{y}_1(1 - \hat{y}_1) = 0.1771$	$\frac{\partial \hat{y}_2}{\partial a_{22}} = \hat{y}_2(1 - \hat{y}_2) = 0.1771$
$\frac{\partial a_{21}}{\partial h_{13}} = w_{213} = 0.2$	$\frac{\partial a_{22}}{\partial h_{13}} = w_{223} = 0.3$
$\frac{\partial h_{13}}{\partial a_{13}} = h_{13} * (1 - h_{13}) = 0.0979$	$\frac{\partial h_{13}}{\partial a_{13}} = h_{13} * (1 - h_{13}) = 0.0979$
$\frac{\partial a_{13}}{\partial w_{131}} = x_1 = 2$	$\frac{\partial a_{13}}{\partial w_{131}} = x_1 = 2$
Path1: $(-0.46 * 0.1771 * 0.2 * 0.0979 * 2) = -0.003190$	Path1: $(0.46 * 0.1771 * 0.3 * 0.0979 * 2) = 0.004785$
Sum of the paths is $\frac{\partial L}{\partial w_{131}} = 0.001595$	

7. Now we can calculate the updated value of w_{212}
8. $w_{131} = w_{131} - \eta(\frac{\partial L}{\partial w_{131}})$
   a. $w_{131} = -0.3 - (1) * (0.001595)$
   b. $w_{131} = -0.301595$
9. We can repeat this process for each weight

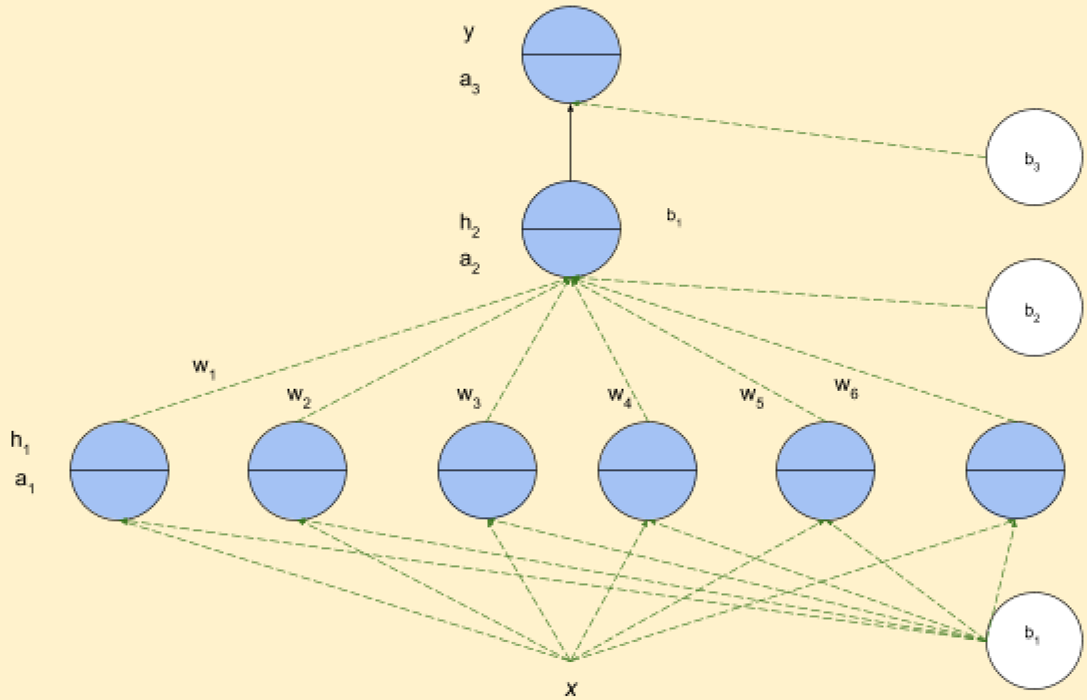
PadhAI: Backpropagation - the light math version

One Fourth Labs

Takeaways and what next?

What have we learned so far and what more do we need to learn?

1. For calculating w_{132}, the paths are almost identical to w_{131} except for the last step
2. This is applicable for the weights w_{133} and w_{134} as well
3. In the next slot (math-heavy), we will learn how to re-use a lot of the computations when calculating new weights.



4. No matter how complex the function, we can always compute the derivative w.r.t any variable using the chain rule.
5. We can reuse a lot of work by starting backwards and computing simpler elements in the chain