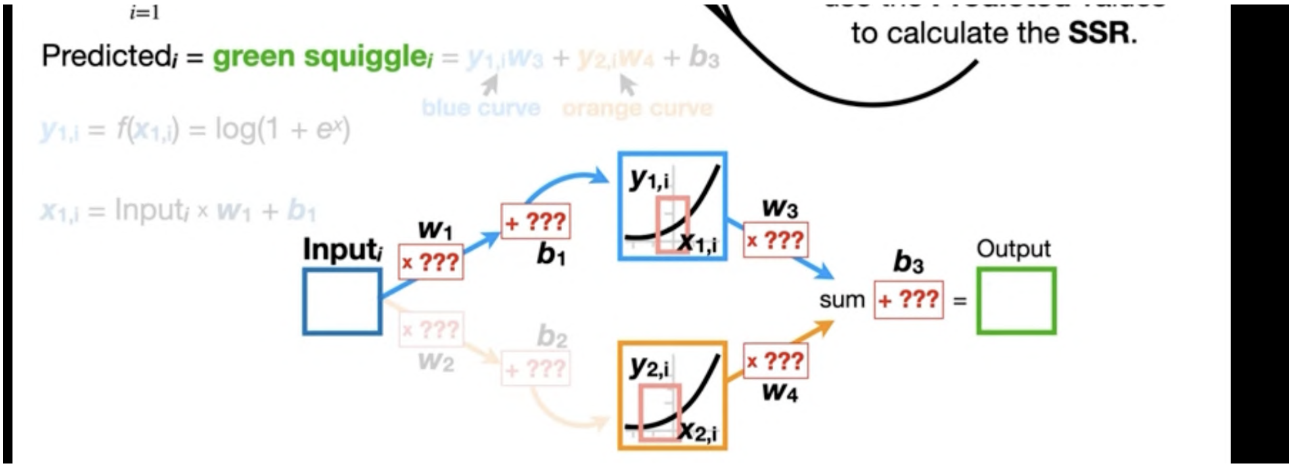
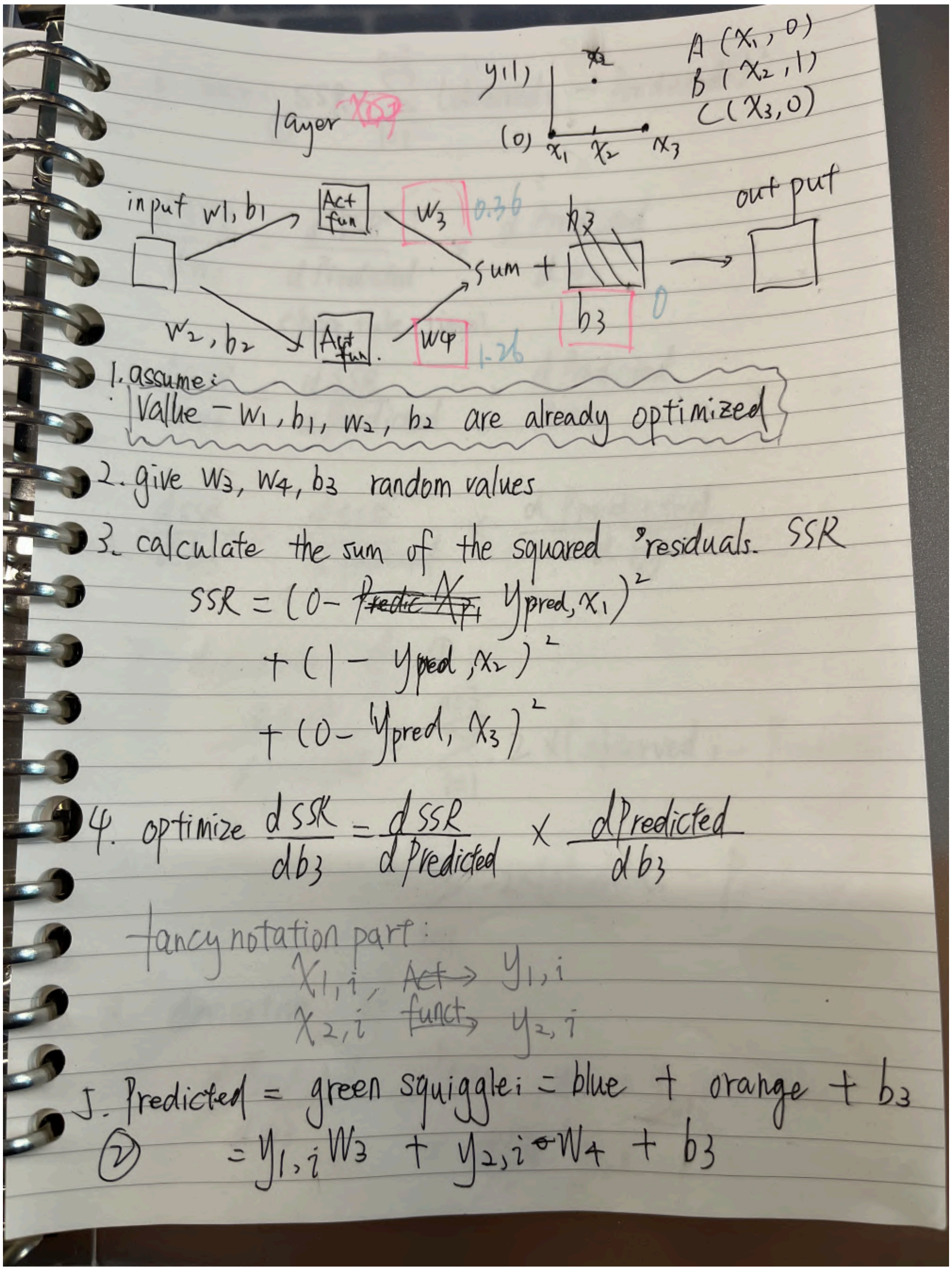
[Backpropagation Details Pt. 2: Going bonkers with The Chain Rule](#)





layer

$y(1)$, (0), x_1, x_2, x_3

$A(x_1, 0)$
$B(x_2, 1)$
$C(x_3, 0)$

input w_1, b_1 Act fun w_3 0.36 b_3 output
sum +
w_2, b_2 Act fun w_4 1.26 b_3 0

1. assume:
Value — w_1, b_1, w_2, b_2 are already optimized

2. give W_3, W_4, b_3 random values

3. calculate the sum of the squared residuals. SSR

$$SSR = (0 - \cancel{\text{Predic } X_1}\ y_{pred, x_1})^2 + (1 - y_{pred, x_2})^2 + (0 - y_{pred, x_3})^2$$

4. optimize $\frac{dSSR}{db_3} = \frac{dSSR}{dPredicted} \times \frac{dPredicted}{db_3}$

fancy notation part:
$x_{1,i} \xrightarrow{\text{Act}} y_{1,i}$
$x_{2,i} \xrightarrow{\text{funct}} y_{2,i}$

5. Predicted = green squiggle$_i$ = blue + orange + b_3
② $= y_{1,i}W_3 + y_{2,i}W_4 + b_3$

6. since: $SSR = \sum_{i=1}^{n=3} (\text{observed}_i - \text{Predicted}_i)^2$ ①

$$\frac{dSSR}{dw_3} = \frac{dSSR}{d\,\text{Predicted}} \times \frac{d\,\text{Predicted}}{dw_3}$$

chain rule time!

$$\frac{dSSR}{dw_4} = \frac{dSSR}{d\,\text{Predicted}} \times \frac{d\,\text{Predicted}}{dw_4}$$

$$\frac{dSSR}{db_3} = \frac{dSSR}{d\,\text{predicted}} \times \frac{d\,\text{Predicted}}{db_3}$$

7. derivative of ①:

$$\frac{dSSR}{d\,\text{predicted}} = \sum_{i=1}^{n=3} 2 \times (\text{observed}_i - \text{Predicted}_i) \times (-1)$$

$$= \sum_{i=1}^{n=3} -2 \times (\text{observed}_i - \text{Predicted}_i)$$

8. derivative of

$$\frac{d\,\text{Predicted}}{dw_3} = (y_{1,i} w_3 + y_{2,i} w_4 + b_3)'$$

$$= y_{1,i}$$

$$\frac{d\ \text{Predicted}}{d w_4} = y_{2,i}$$

$$\frac{d\,\text{Predicted}}{b_3} = 1$$

9. chain rule time:

$$\frac{dSSR}{dw_3} = \sum_{i=1}^{n=3} -2(\text{observed}_i - \text{Predicted}_i) \times y_{1,i}$$

$$\frac{dSSR}{dw_2} = \sum_{i=1}^{n=3} -2(\text{observed}_i - \text{Predicted}_i) \times y_{2,i}$$

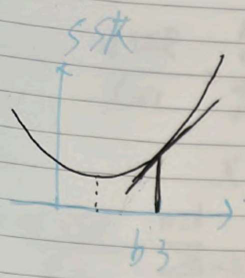
$$\frac{dSSR}{db_3} = \sum_{i=1}^{n=3} -2(\text{observed}_i - \text{Predicted}_i) \times 1$$

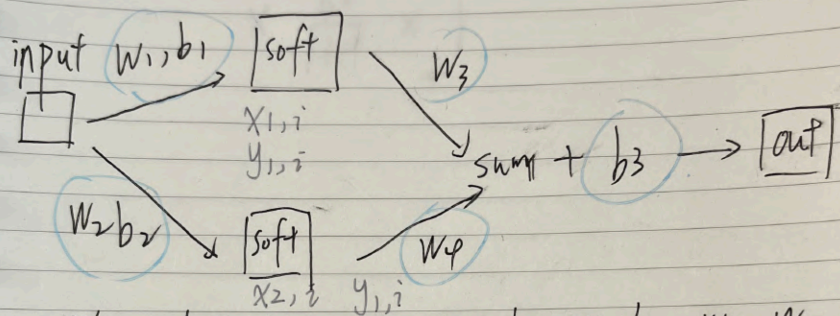
10. Stepsize =

$$SS = \frac{dSSR}{dw_3} \times \text{learning rate}$$

$$\text{New } w_3 = 0.36 - \text{Stepsize}$$

11. repeat the process until the Predictions no longer improve very much, or reach a maximum number of steps





1. Now Let's optimize $w_1, b_1, w_2, b_2, w_3, w_4, b_3$

2. $$\frac{d\,SSR}{dw_1} = \overset{①}{\frac{d\,SSR}{d\,Predicted}} \times \overset{②}{\frac{d\,Predicted}{dy_1}} \times \overset{③}{\frac{dy_{1,i}}{dx_{1,i}}} \times \overset{④}{\frac{dx_{1,i}}{dw_1}}$$

① $SSR = \sum_{i=1}^{n=3} (observed_i - Predicted_i)^2 \Rightarrow \sum_{i=1}^{n=3} -2(observed_i - Predicted_i)$

② $Predicted_i = y_{1,i} w_3 + y_{2,i} w + b_3 \Rightarrow w_3$

③ $y_{1,i} = f(x_{1,i}) = \log(1 + e^x) \Rightarrow \frac{e^x}{1+e^x}$

④ $x_{1,i} = input_i \times w_1 + b_1 \Rightarrow input_i$

3. $$\frac{dSSR}{dw_1} = \sum_{i=1}^{n=3} -2(observed_i - Predicted_i) \times w_3 \times \frac{e^x}{1+e^x} \times Input_i$$

$$\frac{dSSR}{db_1} = \sum_{i=1}^{n=3} -2(observed_i - Predicted_i) \times w_3 \times \frac{e^x}{1+e^x} \times 1$$