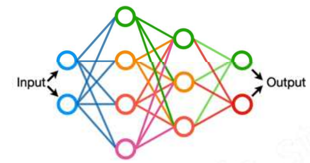
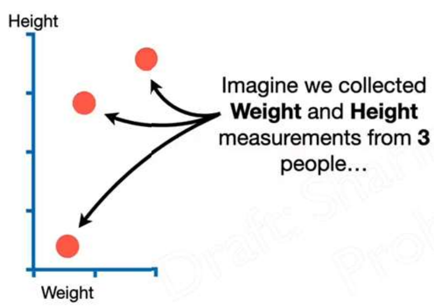
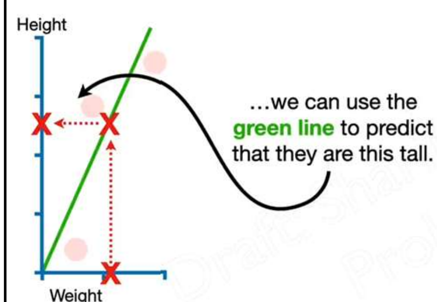
# Neural Networks…

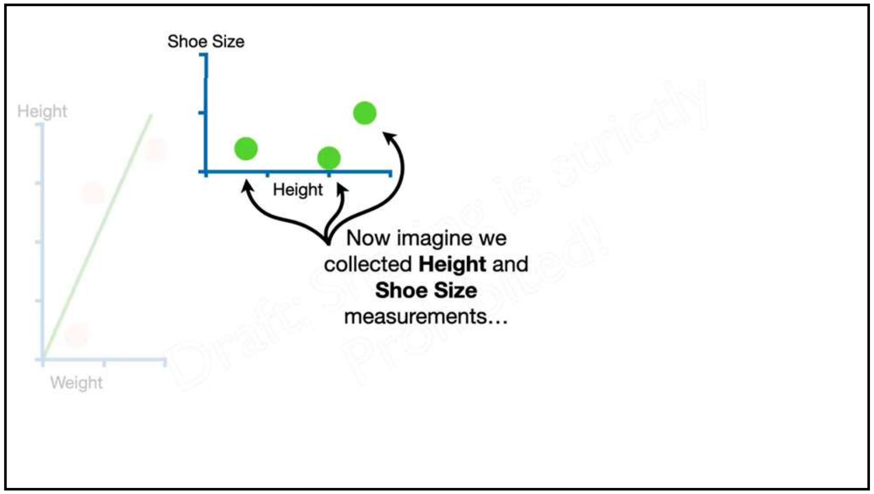
**BACKPROPAGATION DETAILS**

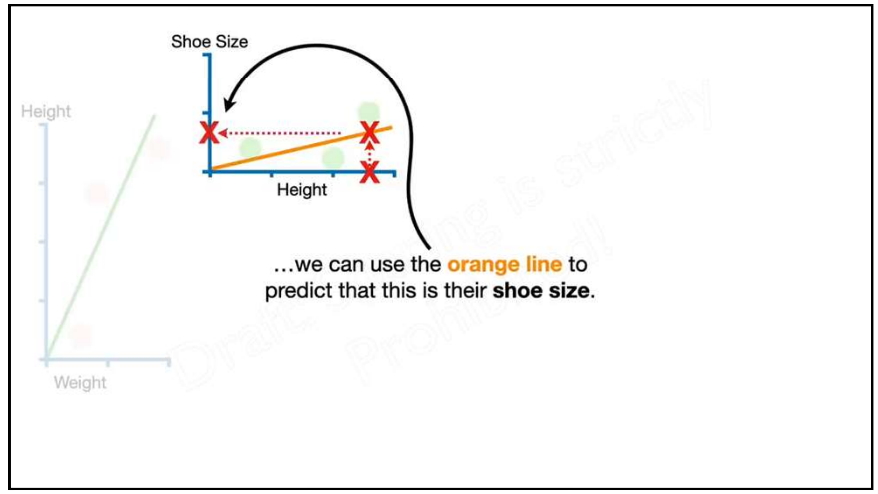


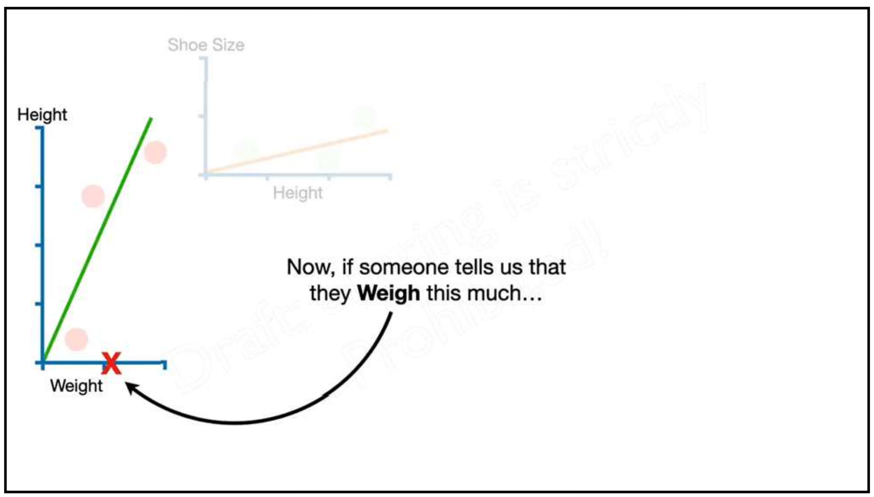

# The Chain Rule

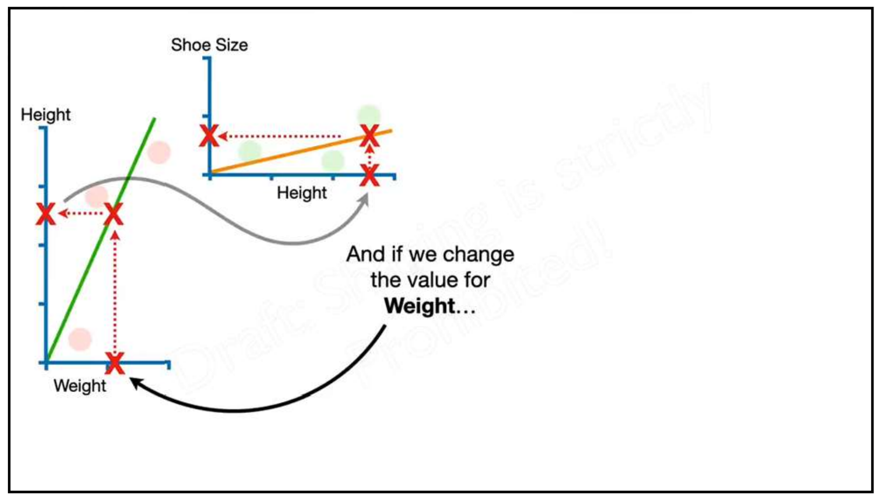


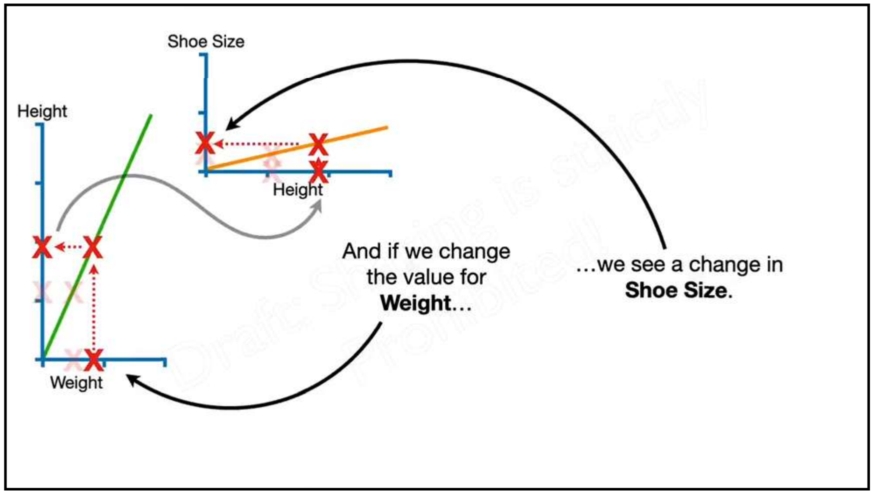



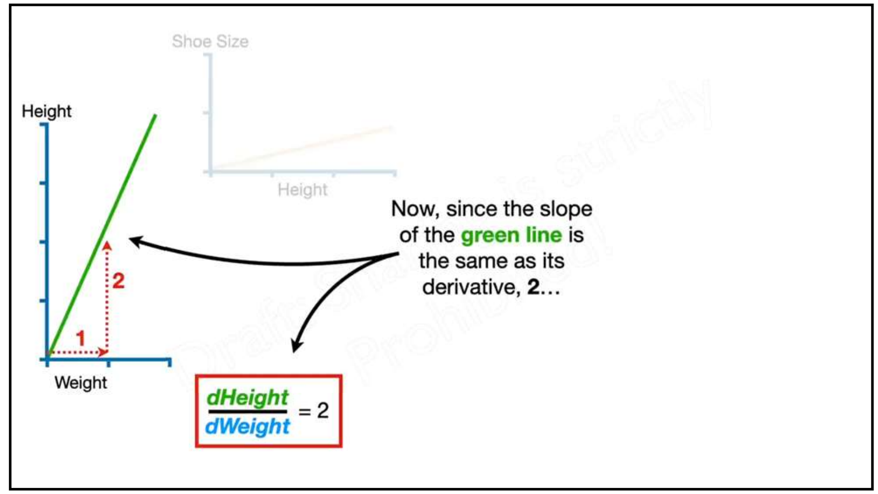

Shoe Size
Height
Height
Weight
Now imagine we collected **Height** and **Shoe Size** measurements…

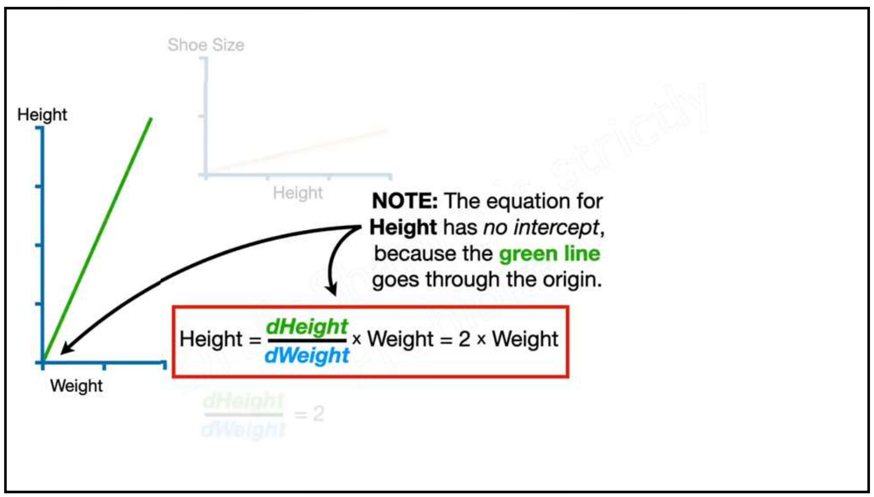

Shoe Size
Height
Height
Weight
…we can use the orange line to predict that this is their **shoe size**.

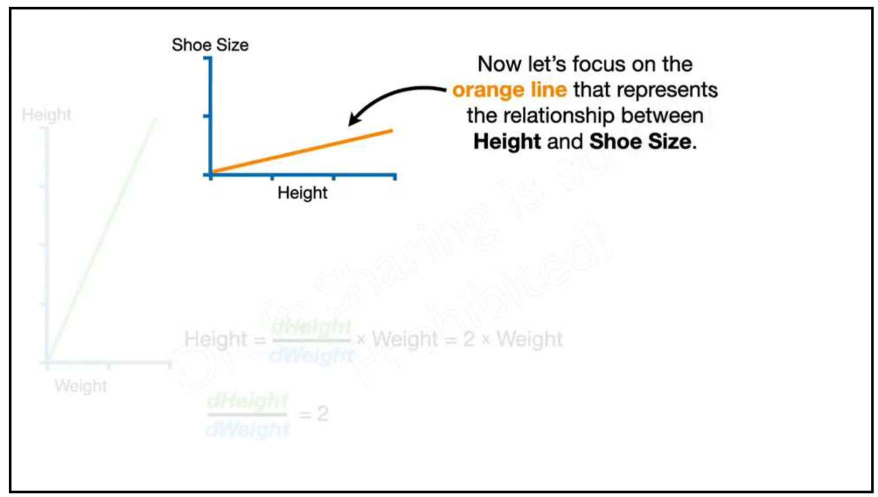

Shoe Size
Height
Height
Weight
Now, if someone tells us that they **Weigh** this much…

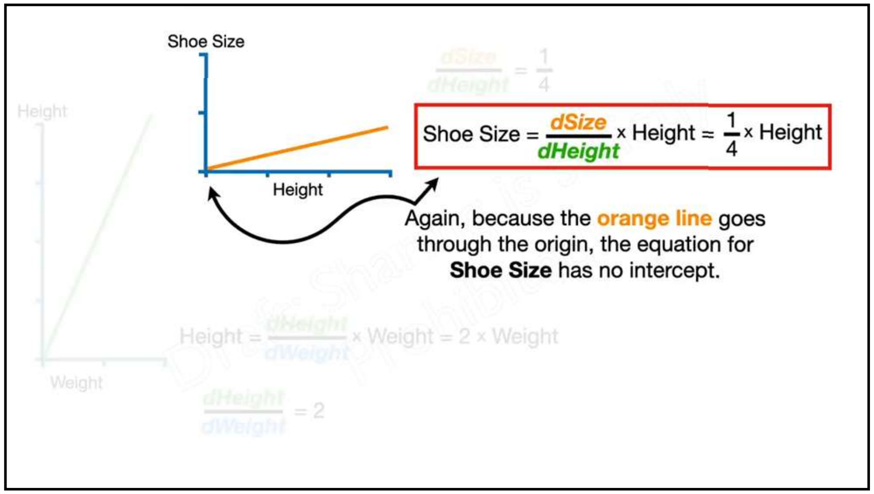

Shoe Size
Height
Height
Weight
And if we change the value for **Weight**…


Shoe Size
Height
Height
Weight
And if we change the value for **Weight**…
…we see a change in **Shoe Size**.

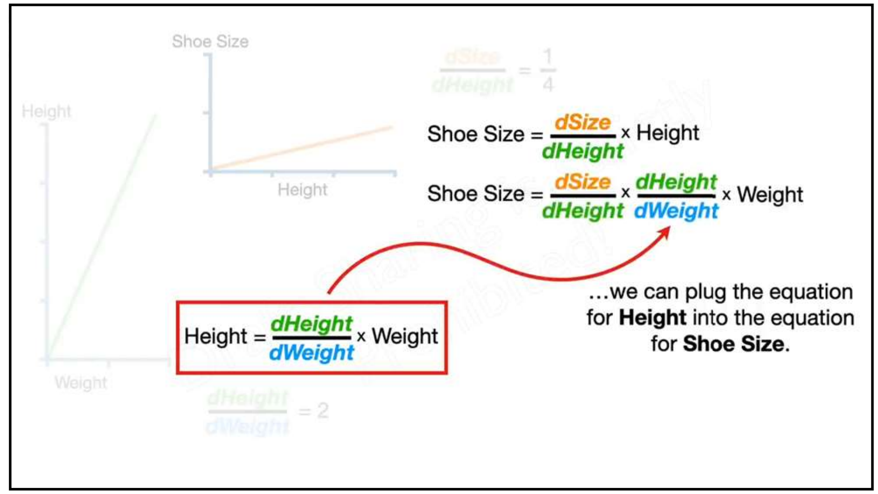

Shoe Size
Height
Height
Weight
2
1
Now, since the slope of the **green line** is the same as its derivative, **2**…
$\frac{dHeight}{dWeight} = 2$


Shoe Size
Height
Height
Weight
**NOTE:** The equation for **Height** has *no intercept*, because the **green line** goes through the origin.
Height = $\frac{dHeight}{dWeight}$ × Weight = 2 × Weight
$\frac{dHeight}{dWeight} = 2$

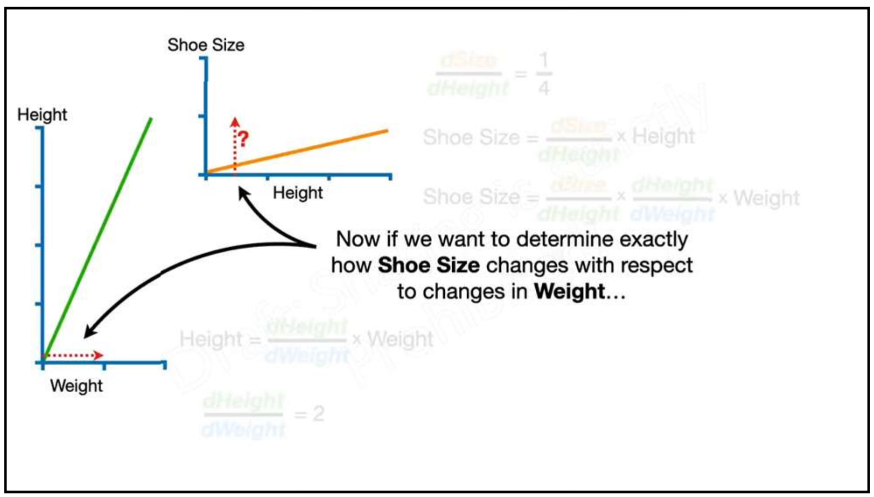

Shoe Size
Height
Now let's focus on the **orange line** that represents the relationship between **Height** and **Shoe Size**.
Height
Weight
Height = $\frac{dHeight}{dWeight}$ × Weight = 2 × Weight
$\frac{dHeight}{dWeight} = 2$



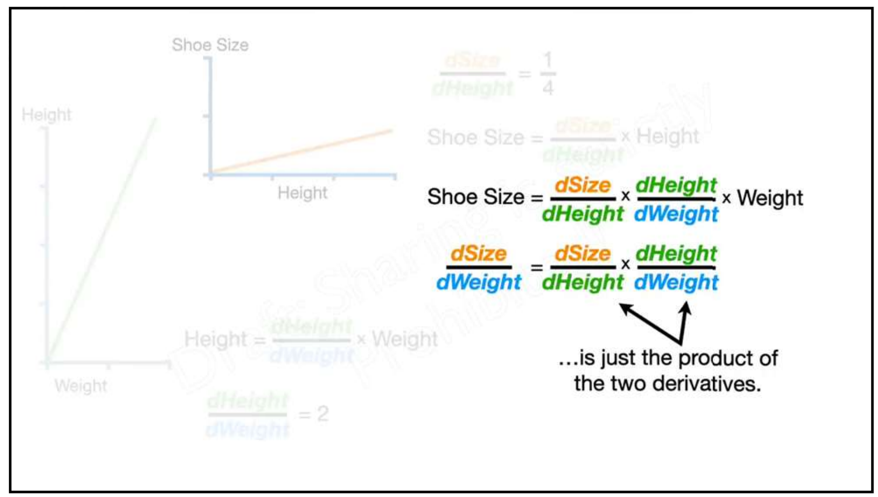
Shoe Size
Height
$\text{Shoe Size} = \frac{dSize}{dHeight} \times \text{Height} = \frac{1}{4} \times \text{Height}$
Again, because the orange line goes through the origin, the equation for Shoe Size has no intercept.

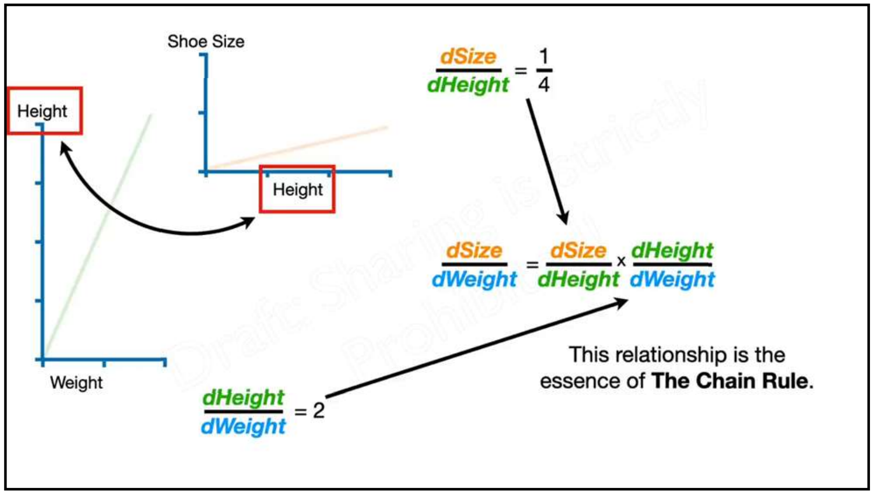
$\text{Shoe Size} = \frac{dSize}{dHeight} \times \text{Height}$
$\text{Shoe Size} = \frac{dSize}{dHeight} \times \frac{dHeight}{dWeight} \times \text{Weight}$
$\text{Height} = \frac{dHeight}{dWeight} \times \text{Weight}$
…we can plug the equation for Height into the equation for Shoe Size.

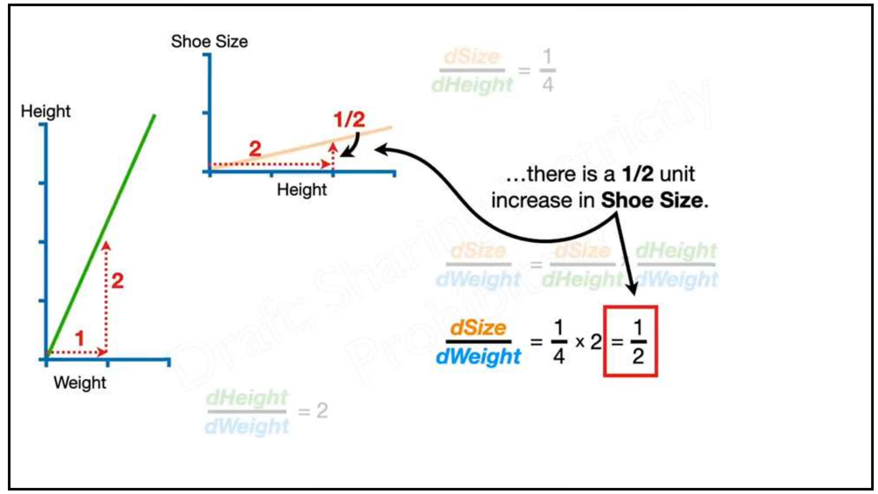



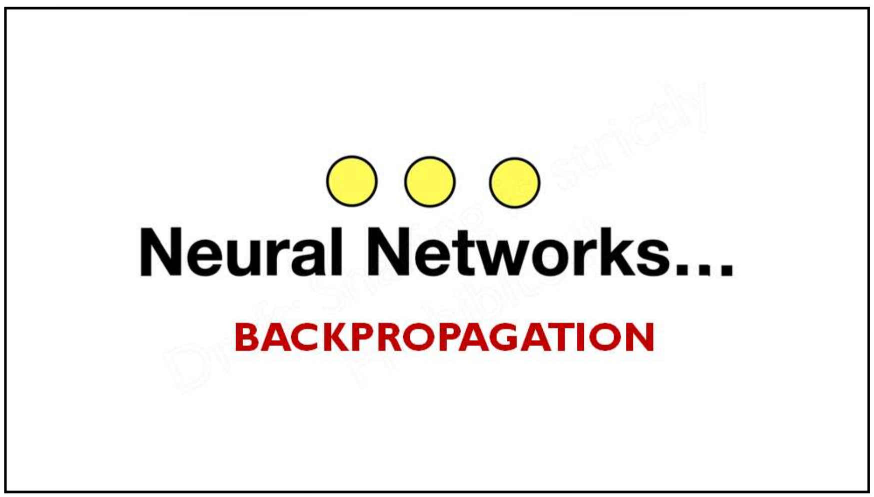
Shoe Size
Height
Height
Weight
?
Now if we want to determine exactly how Shoe Size changes with respect to changes in Weight…

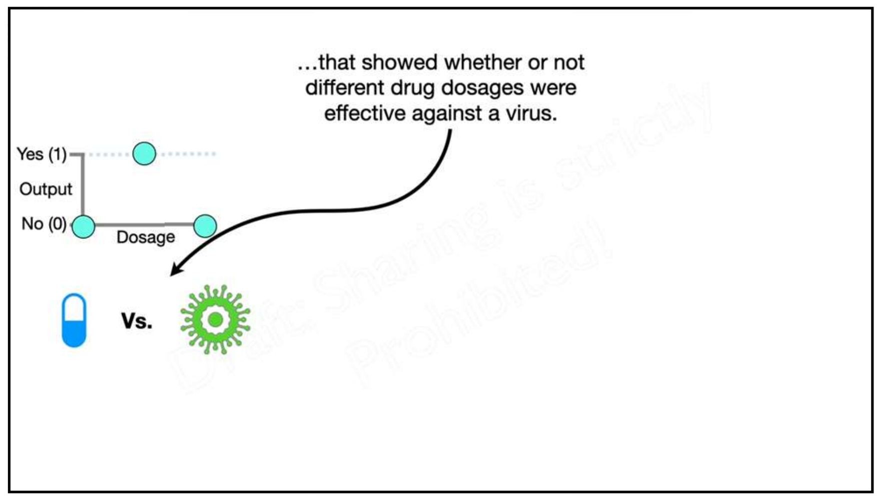



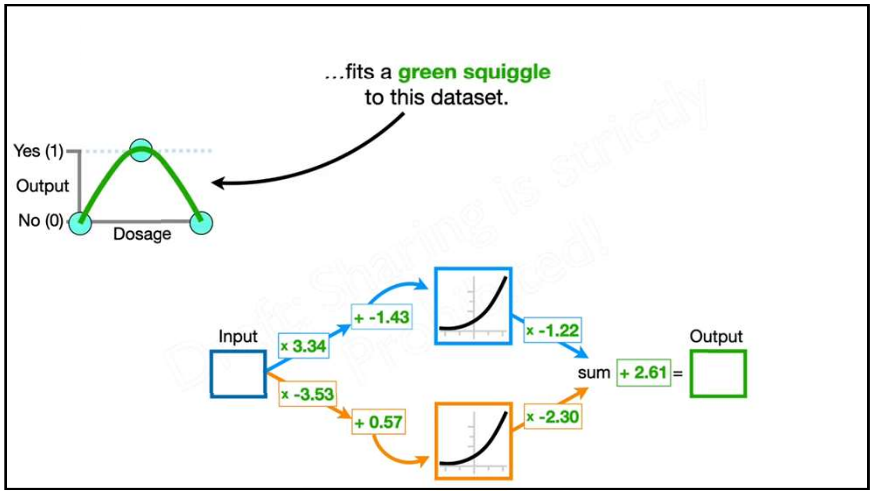
$\text{Shoe Size} = \frac{dSize}{dHeight} \times \frac{dHeight}{dWeight} \times \text{Weight}$
$\frac{dSize}{dWeight} = \frac{dSize}{dHeight} \times \frac{dHeight}{dWeight}$
…is just the product of the two derivatives.

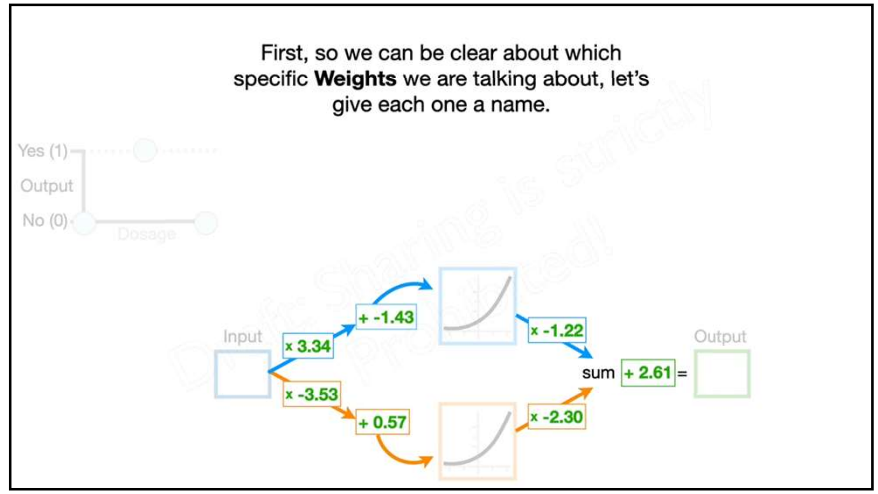

Shoe Size
Height
Height
Weight
dSize / dHeight = 1/4
dSize / dWeight = dSize / dHeight x dHeight / dWeight
dHeight / dWeight = 2
This relationship is the essence of **The Chain Rule**.

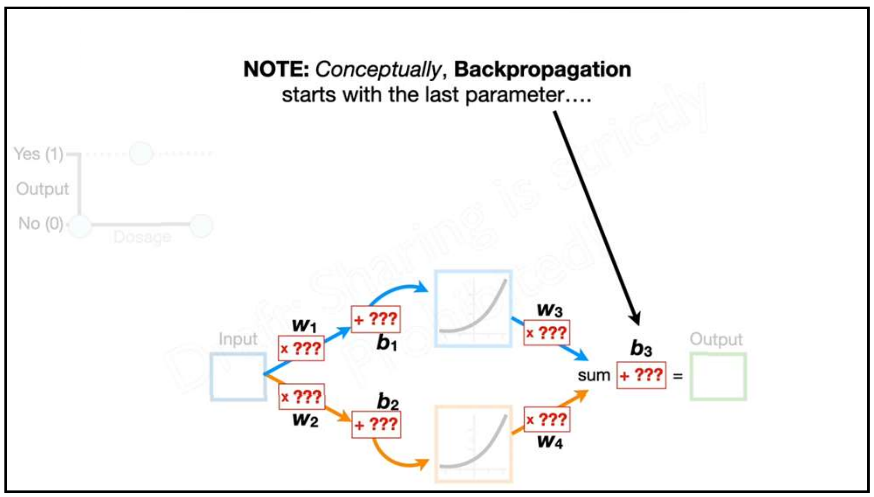

Shoe Size
Height
Height
Weight
1/2
2
2
1
dSize / dHeight = 1/4
…there is a **1/2** unit increase in **Shoe Size**.
dSize / dWeight = dSize / dHeight x dHeight / dWeight
dSize / dWeight = 1/4 x 2 = 1/2
dHeight / dWeight = 2

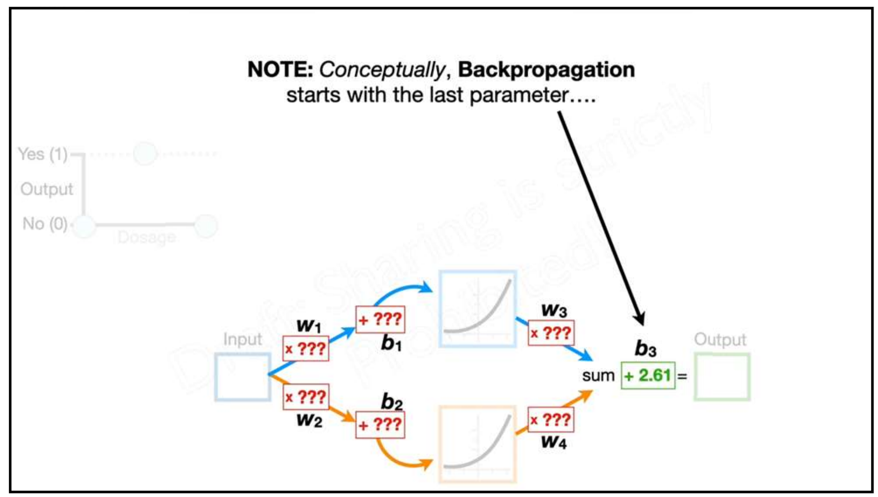

Neural Networks…
BACKPROPAGATION


…that showed whether or not different drug dosages were effective against a virus.
Yes (1)
Output
No (0)
Dosage
Vs.

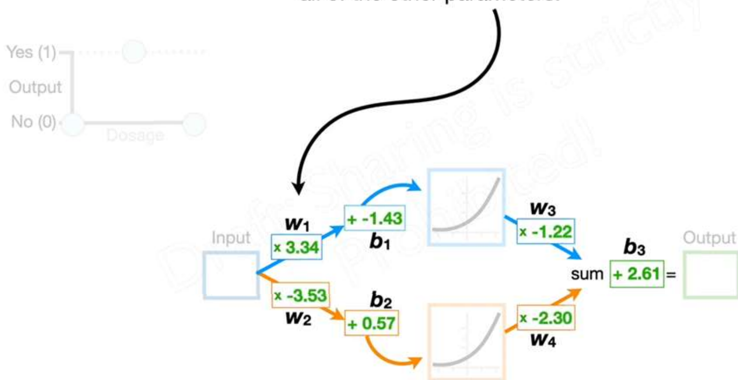

…fits a green squiggle
to this dataset.
Yes (1)
Output
No (0)
Dosage
Input
x 3.34
+ -1.43
x -1.22
x -3.53
+ 0.57
x -2.30
sum
+ 2.61
=
Output


First, so we can be clear about which
specific **Weights** we are talking about, let's
give each one a name.
Yes (1)
Output
No (0)
Input
x 3.34
+ -1.43
x -1.22
x -3.53
+ 0.57
x -2.30
sum
+ 2.61
=
Output


**NOTE:** *Conceptually*, **Backpropagation**
starts with the last parameter….
Yes (1)
Output
No (0)
Input
$w_1$
x ???
+ ???
$b_1$
$w_3$
x ???
$b_3$
sum
+ ???
=
Output
x ???
$w_2$
$b_2$
+ ???
x ???
$w_4$


**NOTE:** *Conceptually*, **Backpropagation**
starts with the last parameter….
Yes (1)
Output
No (0)
Input
$w_1$
x ???
+ ???
$b_1$
$w_3$
x ???
$b_3$
sum
+ 2.61
=
Output
x ???
$w_2$
$b_2$
+ ???
x ???
$w_4$

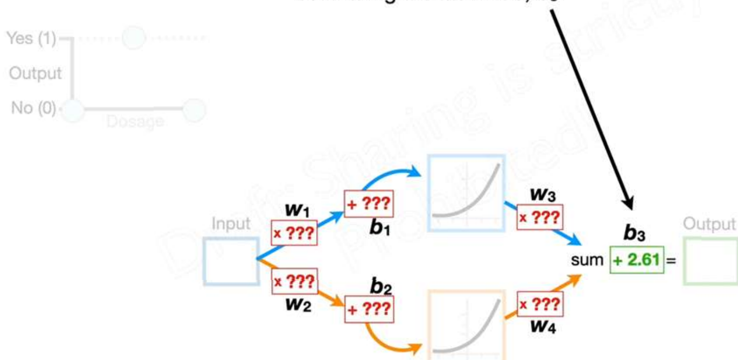
…and works its way backwards to estimate all of the other parameters.



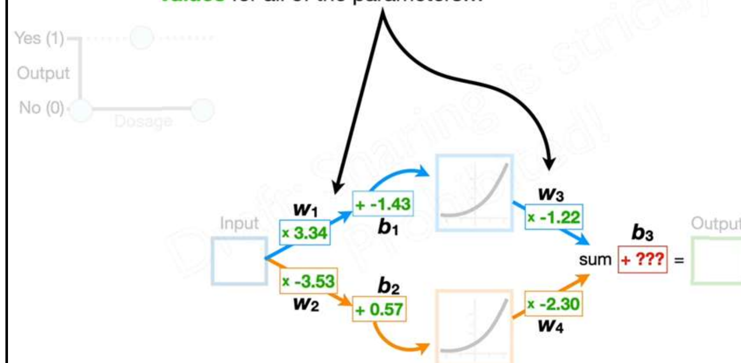

# Neural Networks…

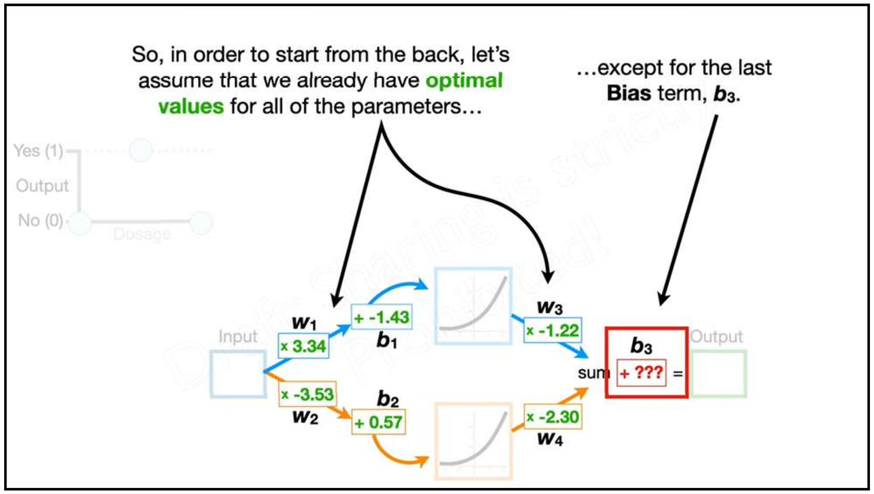
## UPDATING A SINGLE PARAMETER

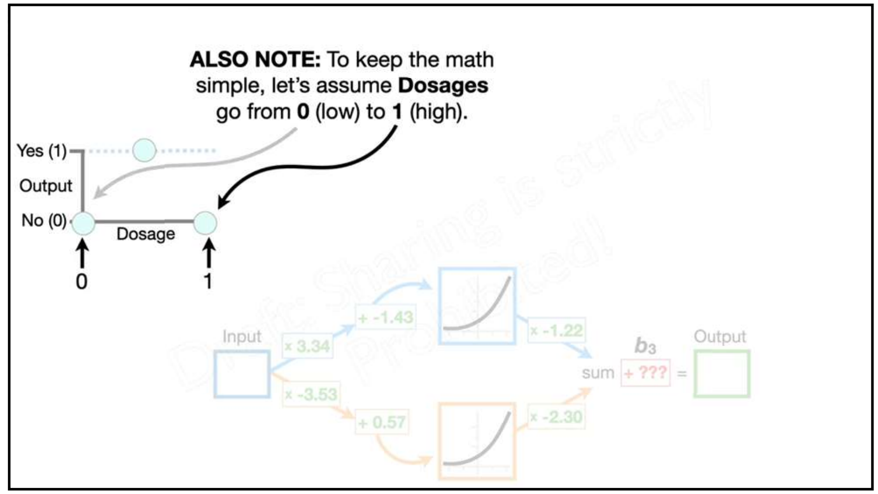
However, we can discuss all of the **Main Ideas** behind **Backpropagation** by just estimating the last **Bias**, $b_3$.

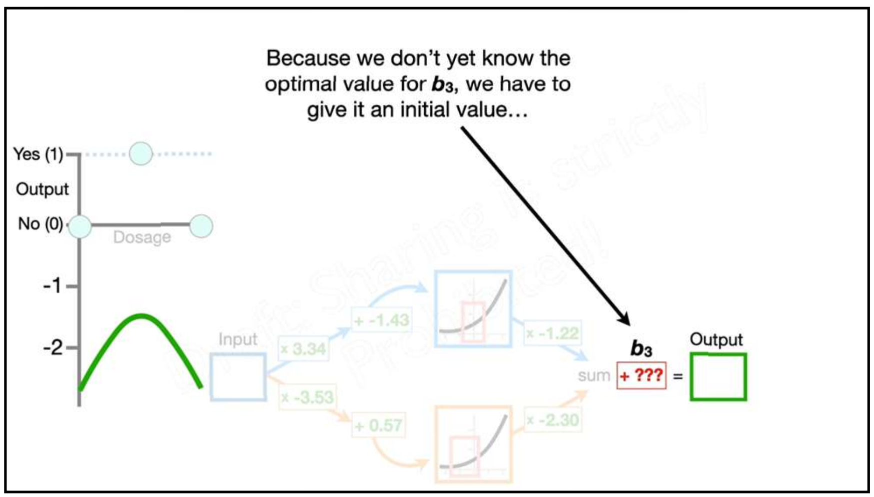


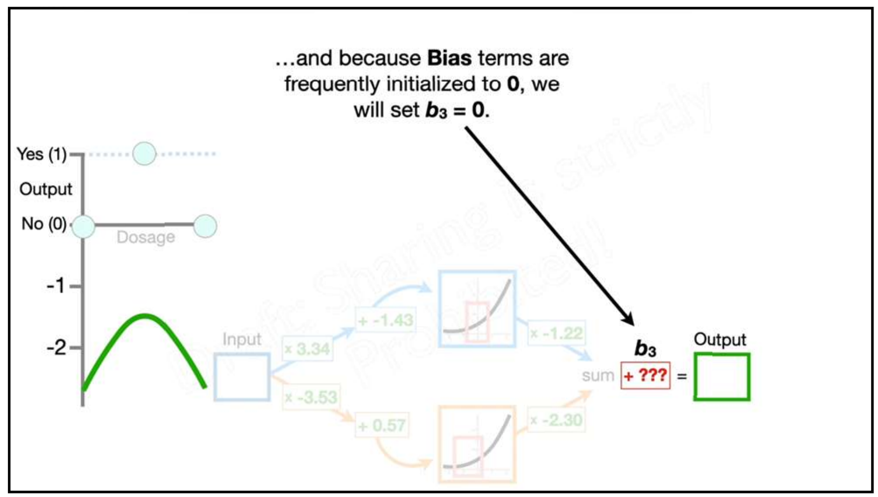

So, in order to start from the back, let's assume that we already have **optimal values** for all of the parameters…

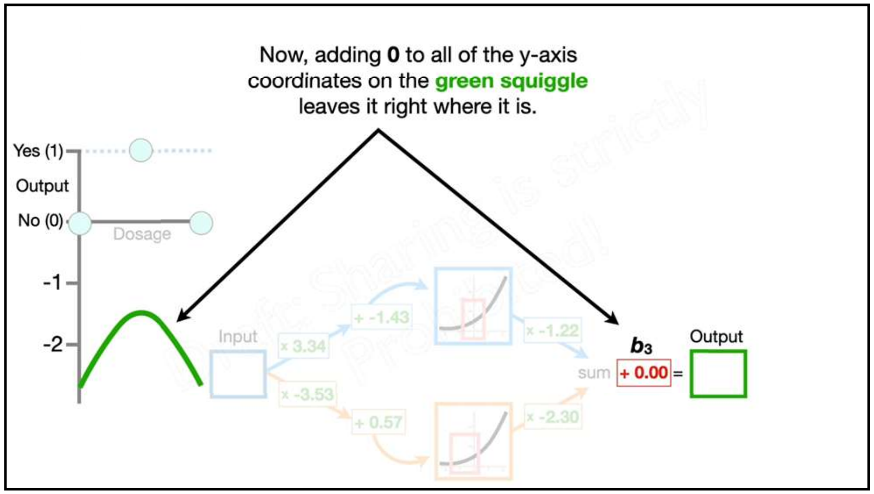


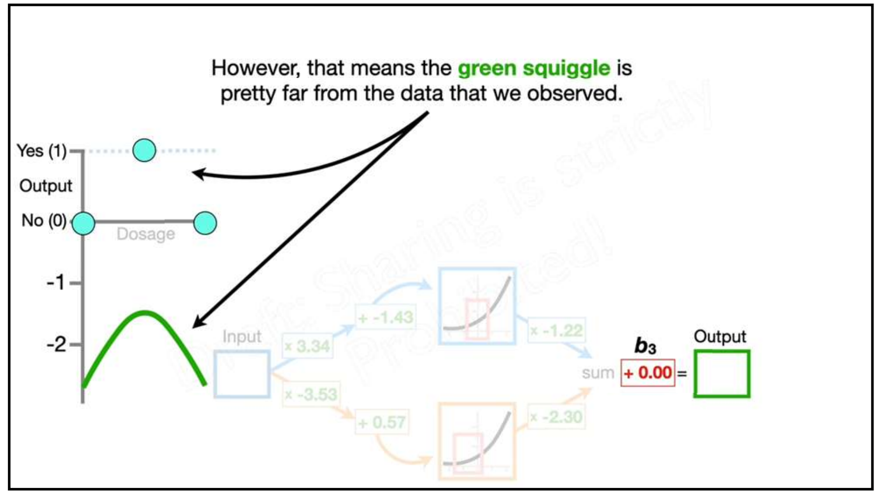

So, in order to start from the back, let's assume that we already have optimal values for all of the parameters…
…except for the last Bias term, $b_3$.
$w_1$
x 3.34
+ -1.43
$b_1$
$w_3$
x -1.22
$b_3$
sum
+ ???
=
x -3.53
$w_2$
$b_2$
+ 0.57
x -2.30
$w_4$
Output

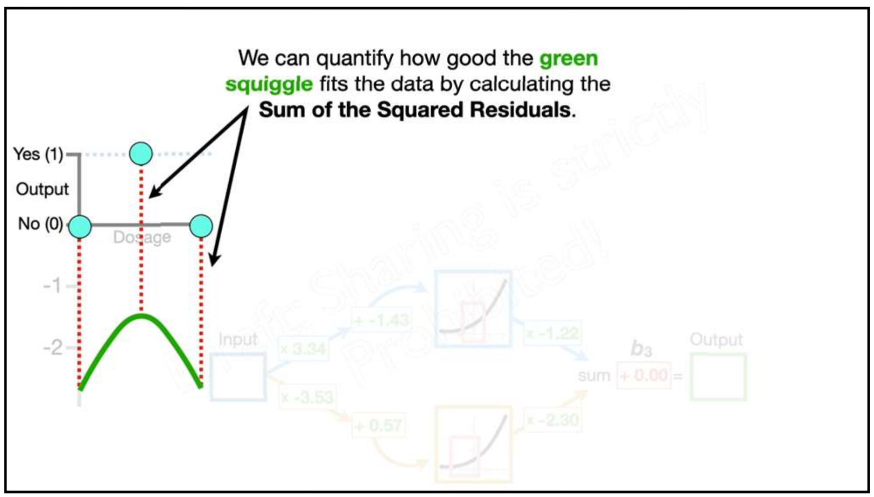

ALSO NOTE: To keep the math simple, let's assume Dosages go from 0 (low) to 1 (high).
Yes (1)
Output
No (0)
Dosage
0
1

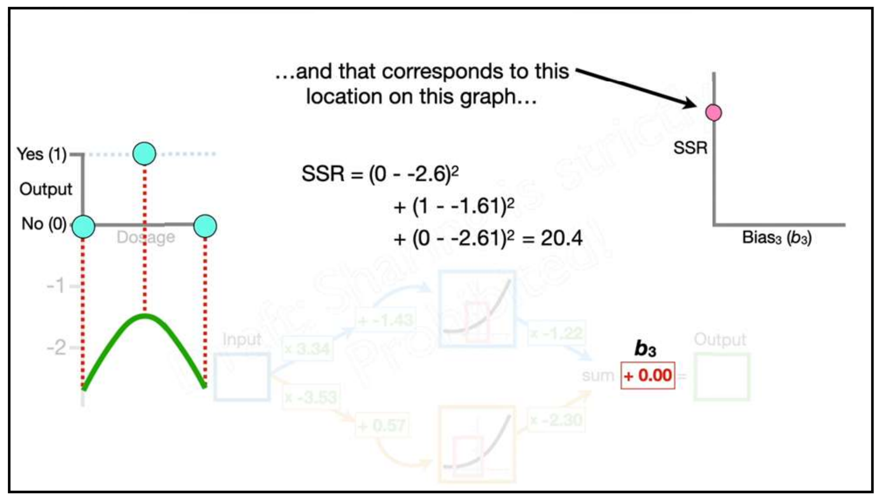

Because we don't yet know the optimal value for $b_3$, we have to give it an initial value…
Yes (1)
Output
No (0)
-1
-2
$b_3$
+ ???
=
Output

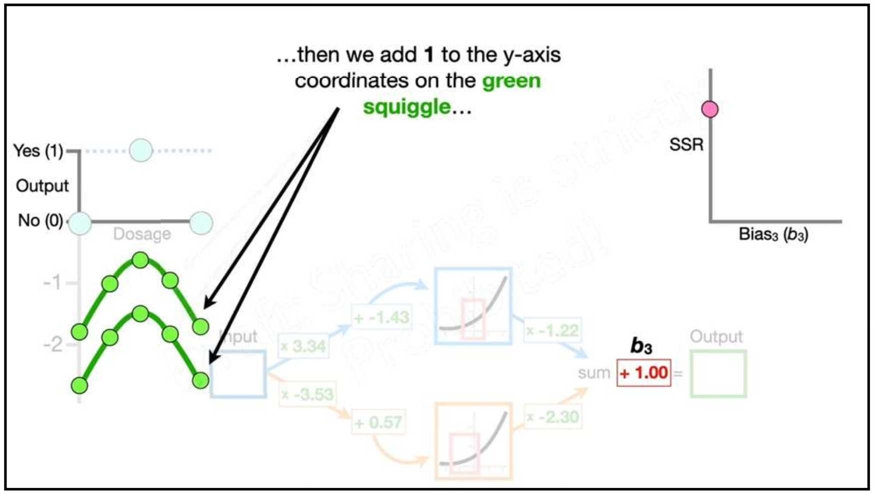

…and because Bias terms are frequently initialized to 0, we will set $b_3$ = 0.
Yes (1)
Output
No (0)
-1
-2
$b_3$
+ ???
=
Output

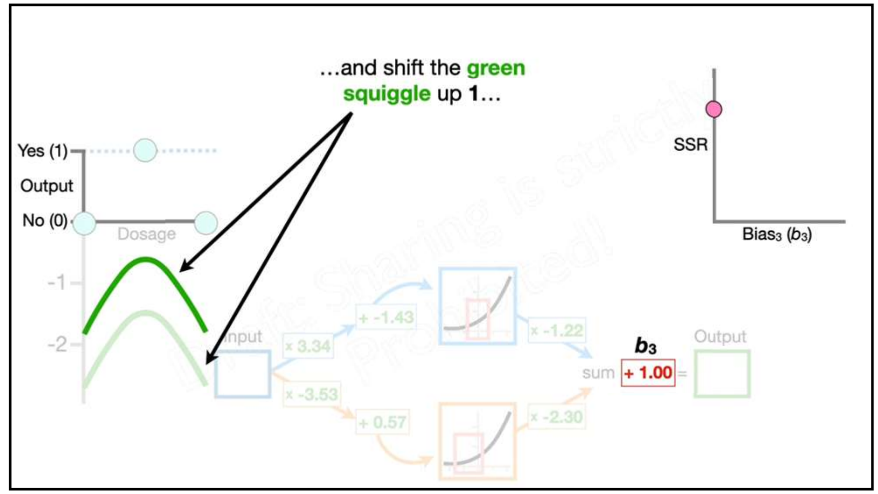

Now, adding **0** to all of the y-axis coordinates on the **green squiggle** leaves it right where it is.
Yes (1)
Output
No (0)
Dosage
-1
-2
Input
$b_3$
sum + 0.00 =
Output

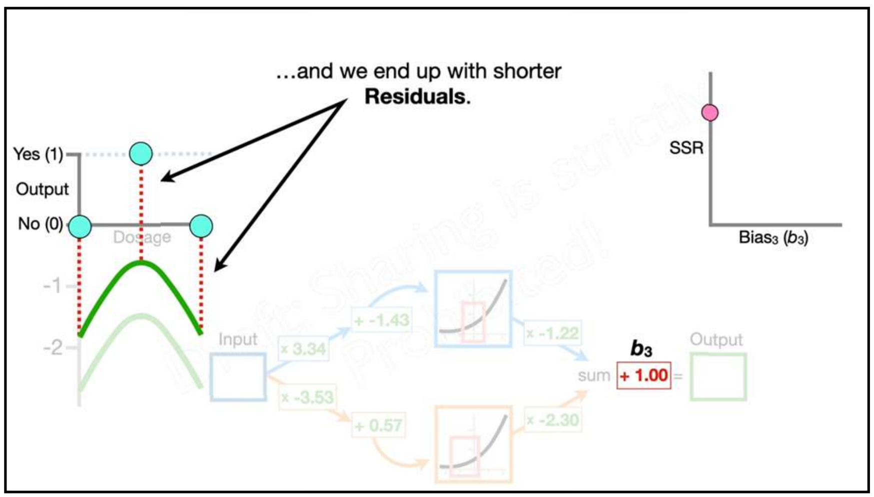

However, that means the **green squiggle** is pretty far from the data that we observed.
Yes (1)
Output
No (0)
Dosage
-1
-2
Input
$b_3$
sum + 0.00 =
Output

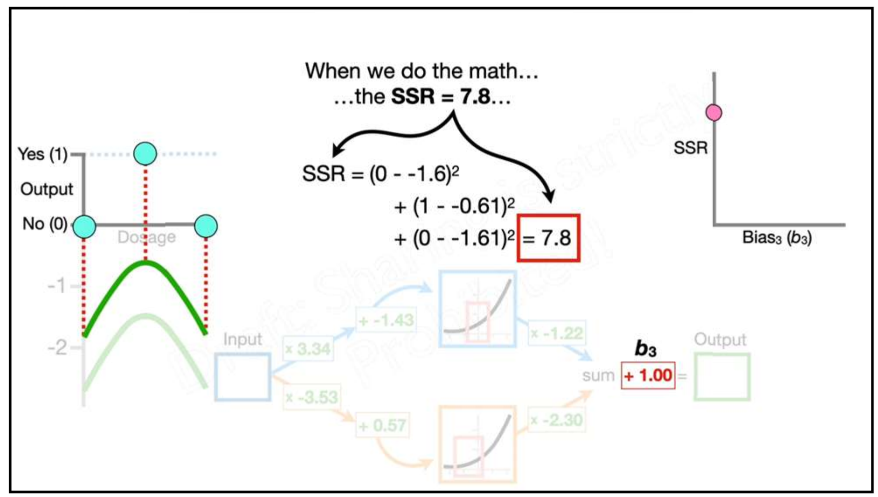

We can quantify how good the **green squiggle** fits the data by calculating the **Sum of the Squared Residuals**.
Yes (1)
Output
No (0)
Dosage
-1
-2

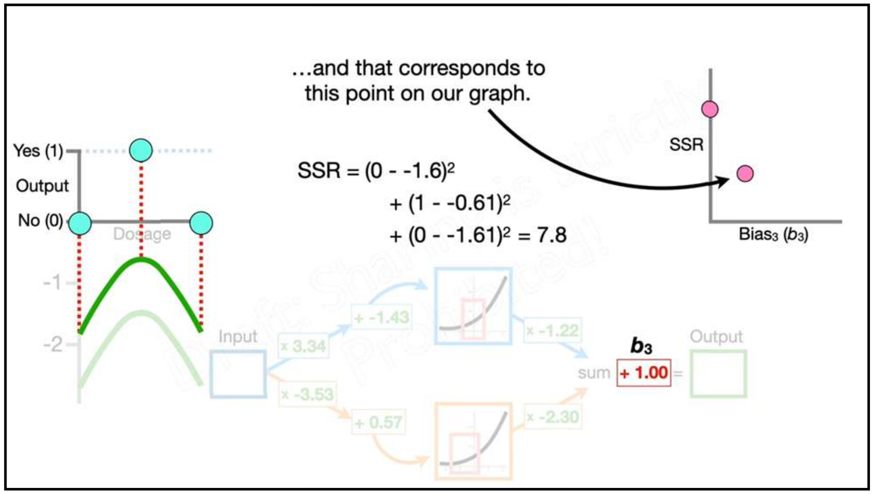

…and that corresponds to this location on this graph…
SSR = $(0 - -2.6)^2$
$+ (1 - -1.61)^2$
$+ (0 - -2.61)^2 = 20.4$
SSR
Bias$_3$ ($b_3$)
Yes (1)
Output
No (0)
Dosage
-1
-2
$b_3$
+ 0.00

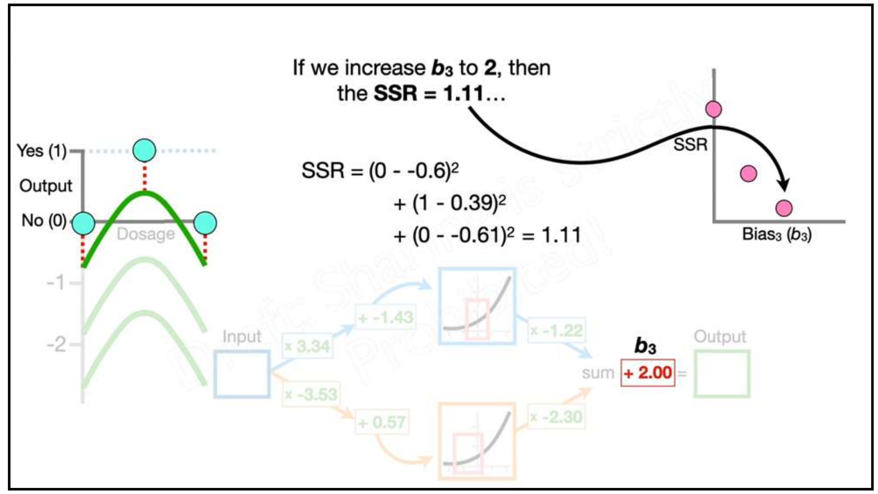

...then we add 1 to the y-axis coordinates on the green squiggle...
Yes (1)
Output
No (0)
Dosage
-1
-2
SSR
Bias3 (b3)
Input
x 3.34
x -3.53
+ -1.43
+ 0.57
x -1.22
x -2.30
b3
sum
+ 1.00
Output

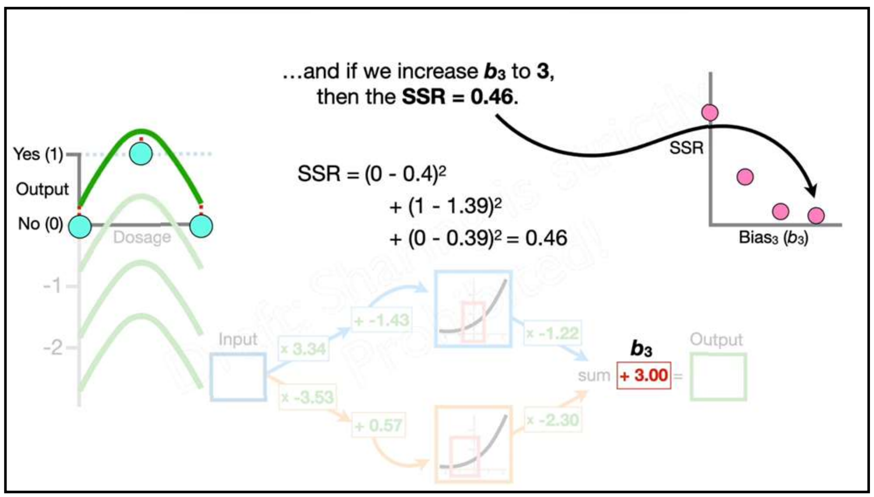

...and shift the green squiggle up 1...
Yes (1)
Output
No (0)
Dosage
-1
-2
SSR
Bias3 (b3)
Input
x 3.34
x -3.53
+ -1.43
+ 0.57
x -1.22
x -2.30
b3
sum
+ 1.00
Output

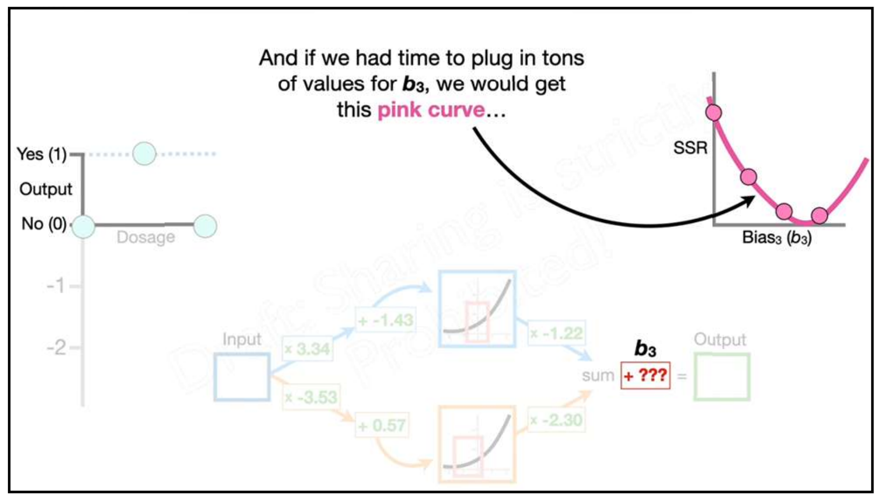

...and we end up with shorter Residuals.
Yes (1)
Output
No (0)
Dosage
-1
-2
SSR
Bias3 (b3)
Input
x 3.34
x -3.53
+ -1.43
+ 0.57
x -1.22
x -2.30
b3
sum
+ 1.00
Output

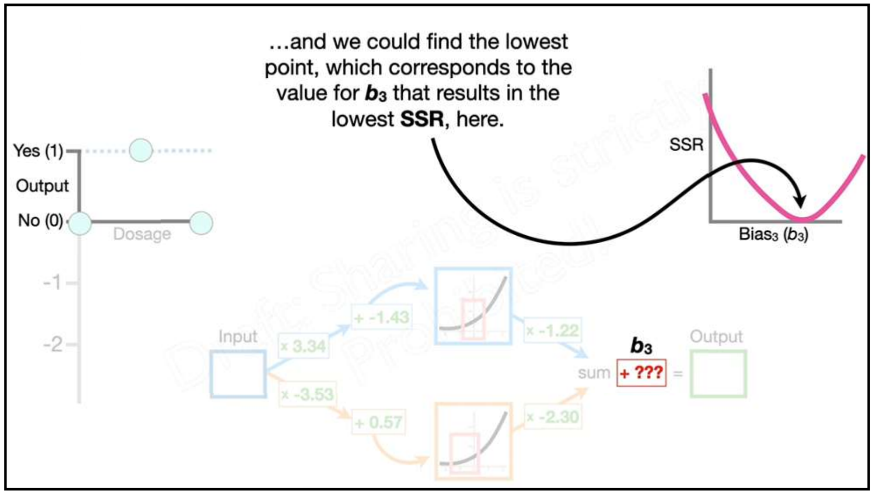

When we do the math...
...the SSR = 7.8...
SSR = (0 - -1.6)2
+ (1 - -0.61)2
+ (0 - -1.61)2 = 7.8
Yes (1)
Output
No (0)
Dosage
-1
-2
SSR
Bias3 (b3)
Input
x 3.34
x -3.53
+ -1.43
+ 0.57
x -1.22
x -2.30
b3
sum
+ 1.00
Output

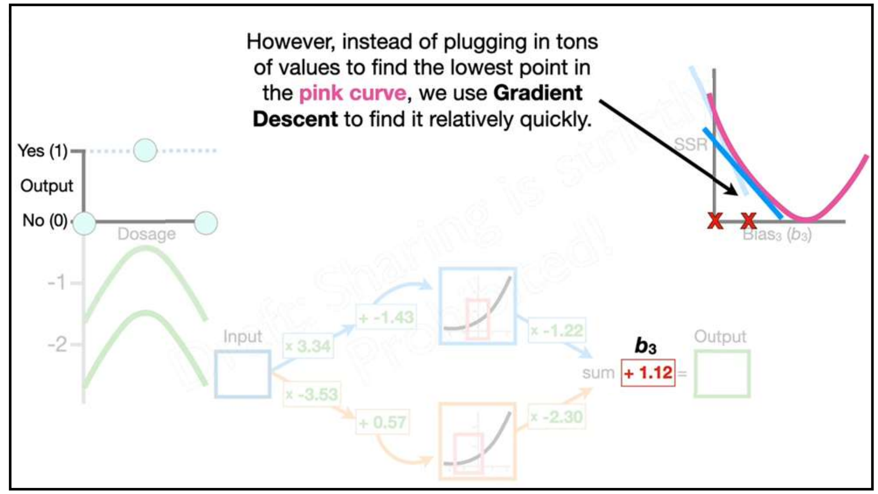

…and that corresponds to
this point on our graph.
SSR = (0 - -1.6)²
+ (1 - -0.61)²
+ (0 - -1.61)² = 7.8
SSR
Bias₃ (b₃)
Yes (1)
Output
No (0)
Dosage
-1
-2
Input
b₃
sum
+ 1.00
Output

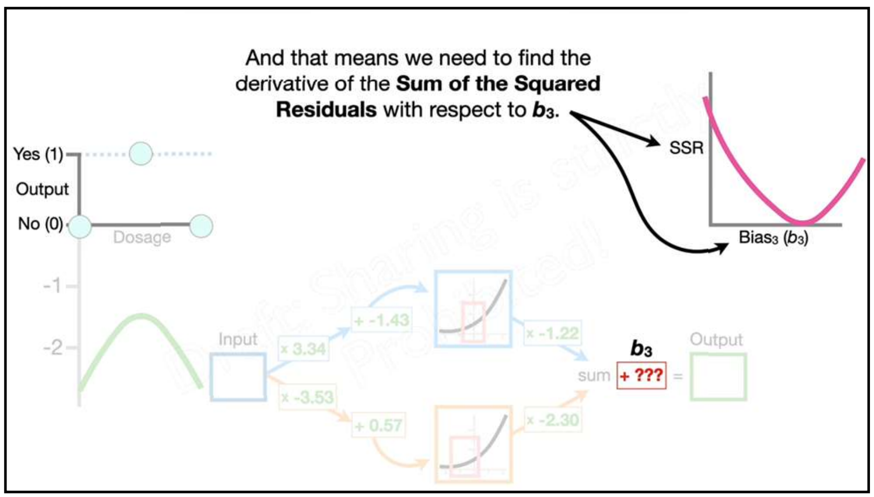

If we increase b₃ to 2, then
the SSR = 1.11…
SSR = (0 - -0.6)²
+ (1 - 0.39)²
+ (0 - -0.61)² = 1.11
SSR
Bias₃ (b₃)
Yes (1)
Output
No (0)
Dosage
-1
-2
Input
b₃
sum
+ 2.00
Output

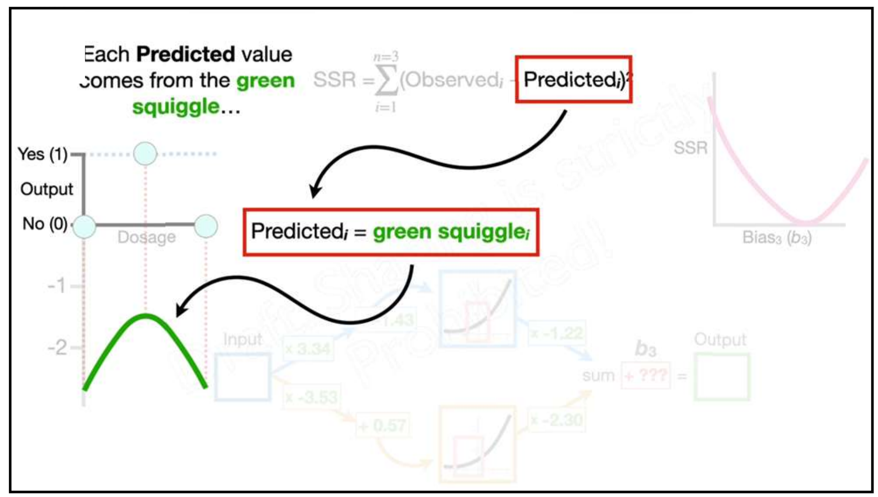

…and if we increase b₃ to 3,
then the SSR = 0.46.
SSR = (0 - 0.4)²
+ (1 - 1.39)²
+ (0 - 0.39)² = 0.46
SSR
Bias₃ (b₃)
Yes (1)
Output
No (0)
Dosage
-1
-2
Input
b₃
sum
+ 3.00
Output

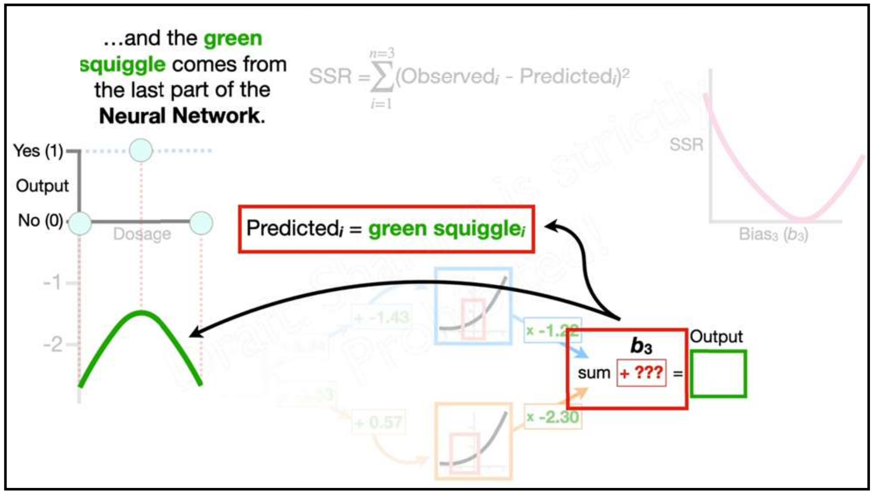

And if we had time to plug in tons
of values for b₃, we would get
this pink curve…
SSR
Bias₃ (b₃)
Yes (1)
Output
No (0)
Dosage
-1
-2
Input
b₃
sum
+ ???
Output


…and we could find the lowest point, which corresponds to the value for b3 that results in the lowest SSR, here.
SSR
Bias3 (b3)
Yes (1)
Output
No (0)
Dosage
-1
-2
Input
x 3.34
x -3.53
+ -1.43
+ 0.57
x -1.22
x -2.30
b3
sum + ??? =
Output

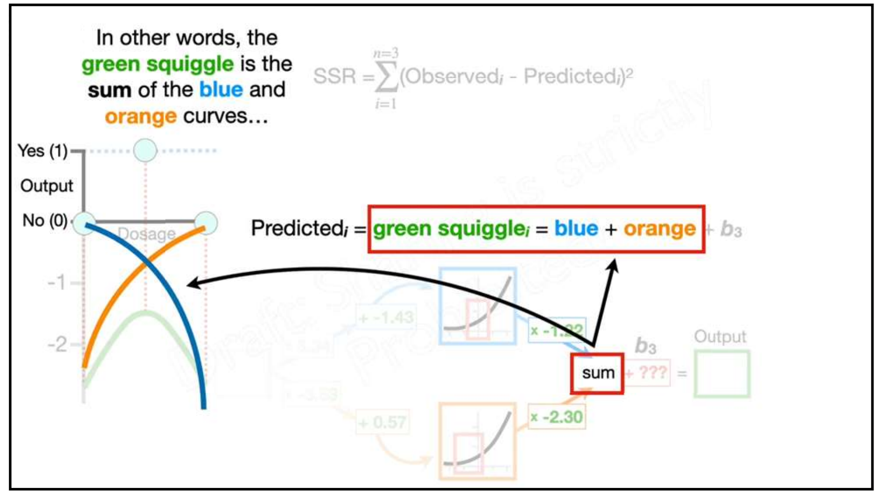

However, instead of plugging in tons of values to find the lowest point in the pink curve, we use Gradient Descent to find it relatively quickly.
SSR
Bias3 (b3)
Yes (1)
Output
No (0)
Dosage
-1
-2
Input
x 3.34
x -3.53
+ -1.43
+ 0.57
x -1.22
x -2.30
b3
sum + 1.12 =
Output


And that means we need to find the derivative of the Sum of the Squared Residuals with respect to b3.
SSR
Bias3 (b3)
Yes (1)
Output
No (0)
Dosage
-1
-2
Input
x 3.34
x -3.53
+ -1.43
+ 0.57
x -1.22
x -2.30
b3
sum + ??? =
Output

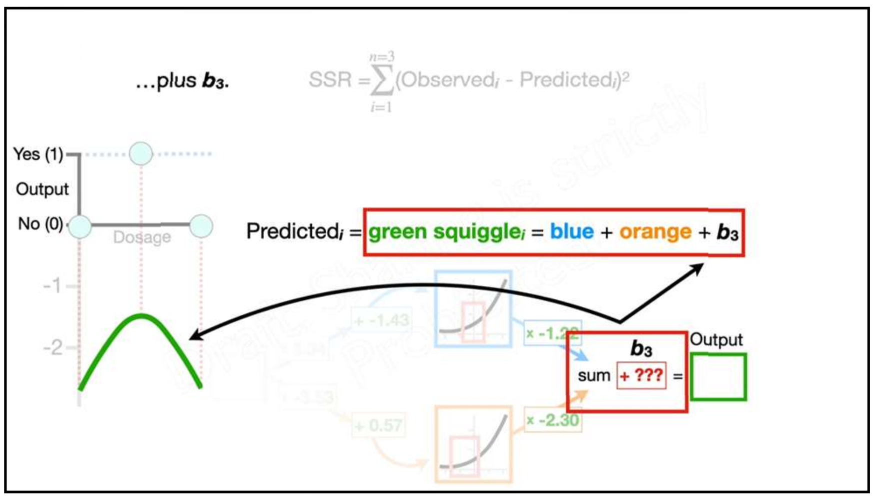

Each Predicted value comes from the green squiggle…
SSR = $\sum_{i=1}^{n=3}$(Observed$_i$ - Predicted$_i$)$^2$
Predicted$_i$ = green squiggle$_i$
SSR
Bias3 (b3)
Yes (1)
Output
No (0)
Dosage
-1
-2



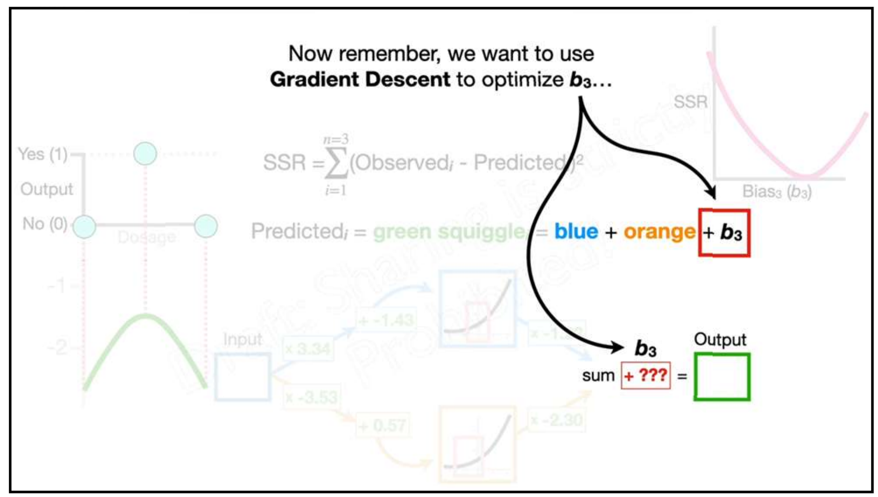
…and the **green squiggle** comes from the last part of the **Neural Network**.
SSR = $\sum_{i=1}^{n=3}$(Observed$_i$ - Predicted$_i$)$^2$
Yes (1)
Output
No (0)
Dosage
Predicted$_i$ = green squiggle$_i$
SSR
Bias$_3$ ($b_3$)
$b_3$
sum + ??? =
Output
x -2.30

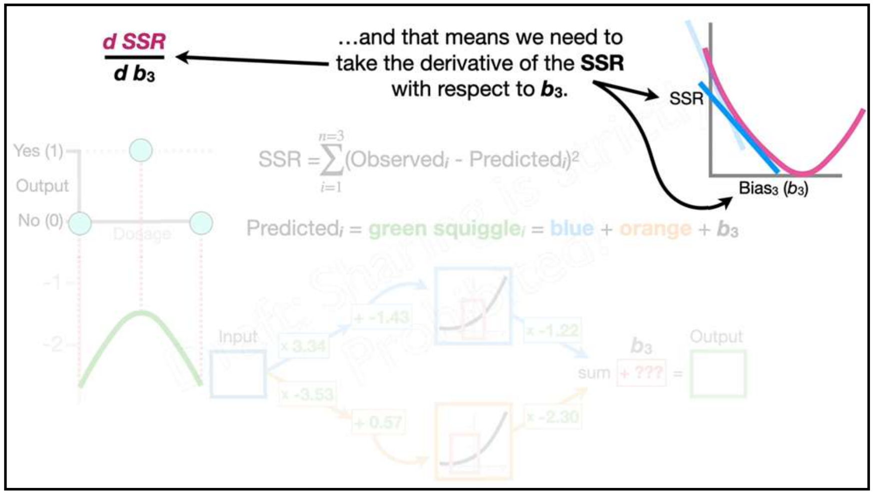
In other words, the **green squiggle** is the **sum** of the **blue** and **orange** curves…
SSR = $\sum_{i=1}^{n=3}$(Observed$_i$ - Predicted$_i$)$^2$
Yes (1)
Output
No (0)
Dosage
Predicted$_i$ = green squiggle$_i$ = blue + orange + $b_3$
sum
x -2.30

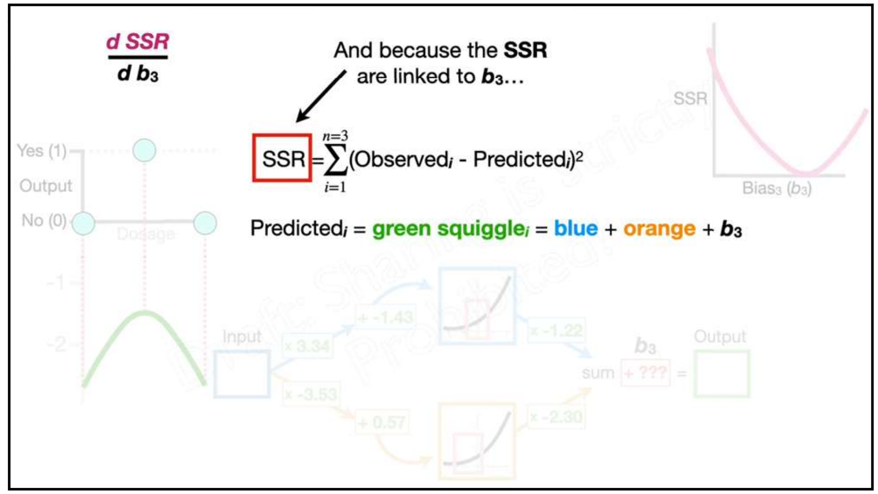
…plus $b_3$.
SSR = $\sum_{i=1}^{n=3}$(Observed$_i$ - Predicted$_i$)$^2$
Yes (1)
Output
No (0)
Dosage
Predicted$_i$ = green squiggle$_i$ = blue + orange + $b_3$
$b_3$
sum + ??? =
Output
x -2.30

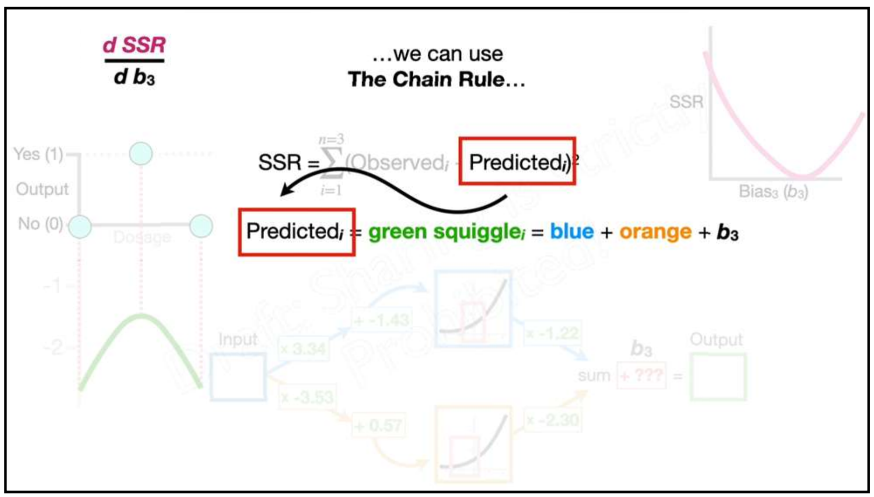
Now remember, we want to use **Gradient Descent** to optimize $b_3$…
SSR
Bias$_3$ ($b_3$)
SSR = $\sum_{i=1}^{n=3}$(Observed$_i$ - Predicted$_i$)$^2$
Predicted$_i$ = green squiggle$_i$ = blue + orange + $b_3$
$b_3$
sum + ??? =
Output



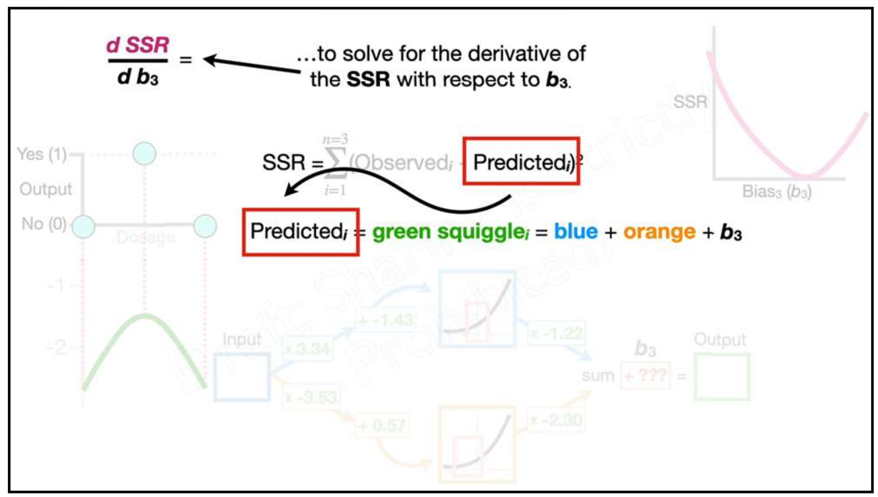
d SSR / d b3
…and that means we need to take the derivative of the SSR with respect to b3.
SSR
Bias3 (b3)
Yes (1)
Output
No (0)
Dosage
SSR = $\sum_{i=1}^{n=3}$(Observedi - Predictedi)2
Predictedi = green squigglei = blue + orange + b3
Input
b3
Output
sum + ??? =

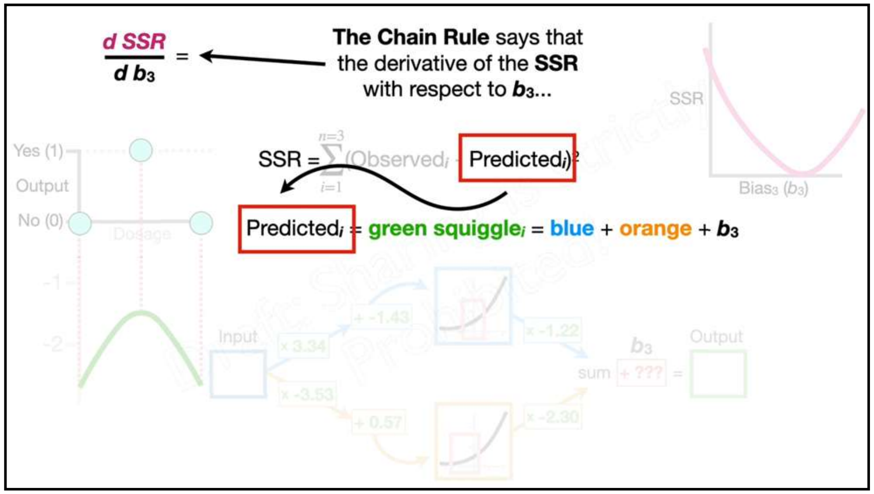
d SSR / d b3
And because the SSR are linked to b3…
SSR = $\sum_{i=1}^{n=3}$(Observedi - Predictedi)2
Predictedi = green squigglei = blue + orange + b3
SSR
Bias3 (b3)
Yes (1)
Output
No (0)
Input
b3
Output
sum + ??? =

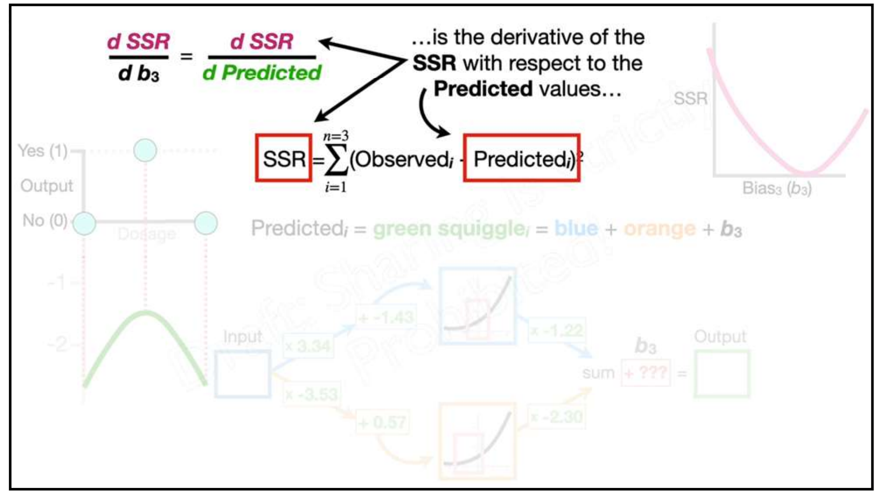



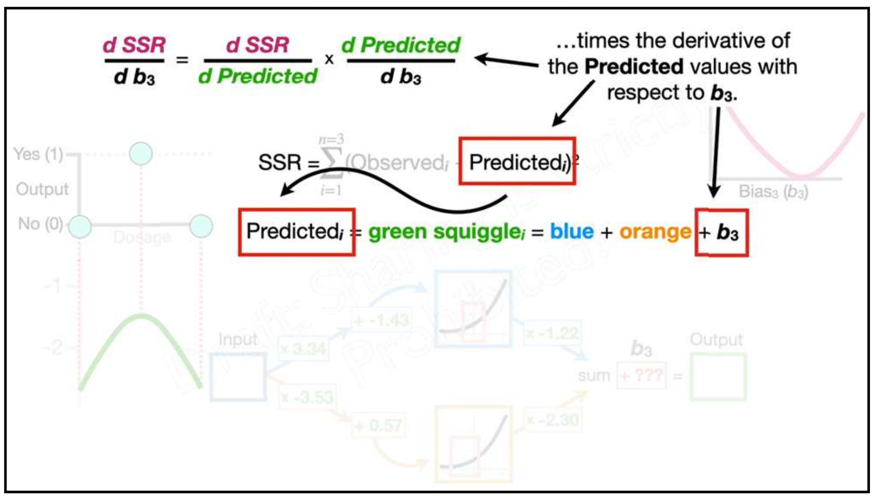
d SSR / d b3
…we can use The Chain Rule…
SSR = $\sum_{i=1}^{n=3}$(Observedi - Predictedi)2
Predictedi = green squigglei = blue + orange + b3
SSR
Bias3 (b3)
Yes (1)
Output
No (0)
Input
b3
Output
sum + ??? =

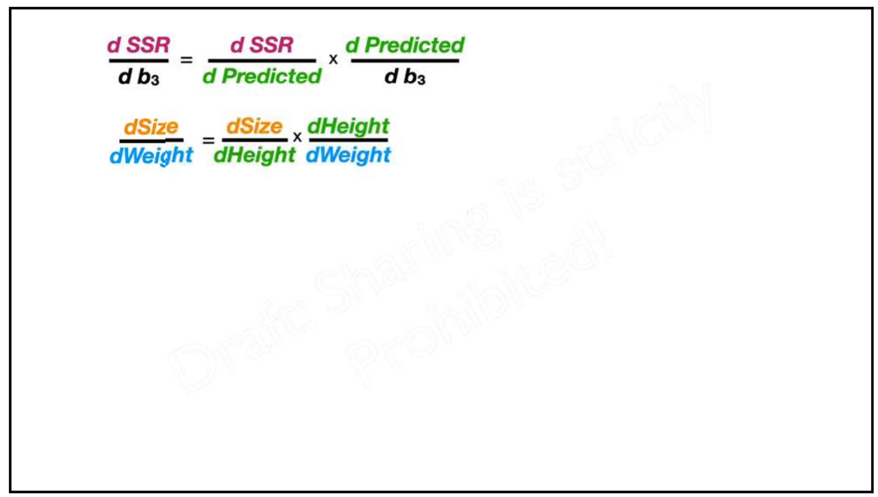



d SSR / d b3 =
…to solve for the derivative of the SSR with respect to b3.
SSR = $\sum_{i=1}^{n=3}$(Observedi - Predictedi)2
Predictedi = green squigglei = blue + orange + b3
SSR
Bias3 (b3)
Yes (1)
Output
No (0)
Input
b3
Output
sum + ??? =


d SSR / d b3 =
The Chain Rule says that the derivative of the SSR with respect to b3...
SSR = Σ (Observedi - Predictedi)²
Predictedi = green squigglei = blue + orange + b3


d SSR / d b3 = d SSR / d Predicted
...is the derivative of the SSR with respect to the Predicted values...
SSR = Σ (Observedi - Predictedi)²


d SSR / d b3 = d SSR / d Predicted x d Predicted / d b3
...times the derivative of the Predicted values with respect to b3.
SSR = Σ (Observedi - Predictedi)²
Predictedi = green squigglei = blue + orange + b3


d SSR / d b3 = d SSR / d Predicted x d Predicted / d b3
dSize / dWeight = dSize / dHeight x dHeight / dWeight

$$\frac{d\ SSR}{d\ b_3} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ b_3}$$

Now, we can solve for the derivative of the **SSR** with respect to the **Predicted** values by first substituting in the equation…

$$\frac{d\ SSR}{d\ Predicted} = \frac{d}{d\ Predicted} \sum_{i=1}^{n=3} (\text{Observed}_i - \text{Predicted}_i)^2$$

$$\boxed{\text{SSR} = \sum_{i=1}^{n=3} (\text{Observed}_i - \text{Predicted}_i)^2} \quad \text{Predicted}_i = \text{green squiggle}_i = \text{blue} + \text{orange} + b_3$$

---

$$\frac{d\ SSR}{d\ b_3} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ b_3}$$

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

So we multiply the derivative of the **SSR** with respect to the **Predicted** values by **1**.

$$\frac{d\ Predicted}{d\ b_3} = \frac{d}{d\ b_3}\ \text{green squiggle} = \frac{d}{d\ b_3} (\text{blue} + \text{orange} + b_3) = 1$$

$$\text{SSR} = \sum_{i=1}^{n=3} (\text{Observed}_i - \text{Predicted}_i)^2 \quad \text{Predicted}_i = \text{green squiggle}_i = \text{blue} + \text{orange} + b_3$$

---

$$\frac{d\ SSR}{d\ b_3} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ b_3}$$

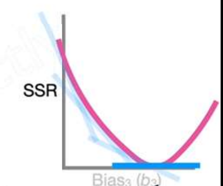
$$= \sum_{i=1}^{n=3} -2 \times (\text{Observed}_i - \text{Predicted}_i) \times 1$$

And at long last, we have the derivative of the **SSR** with respect to ***b*₃**.

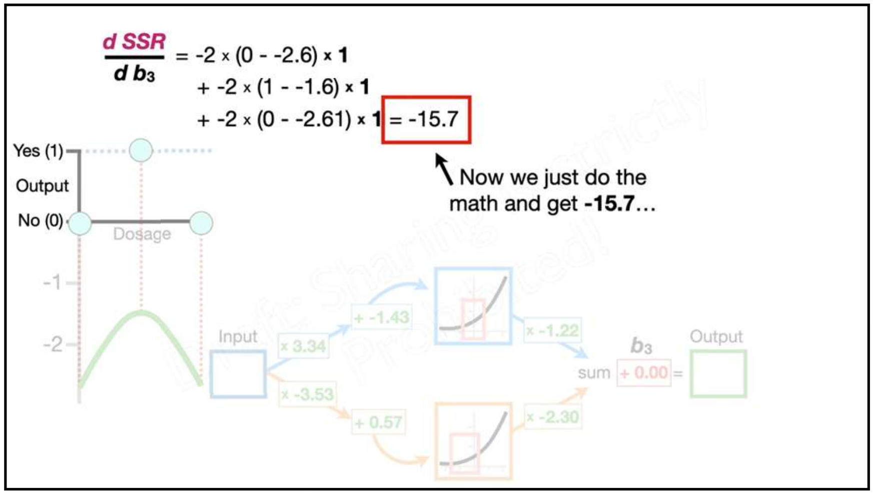
---

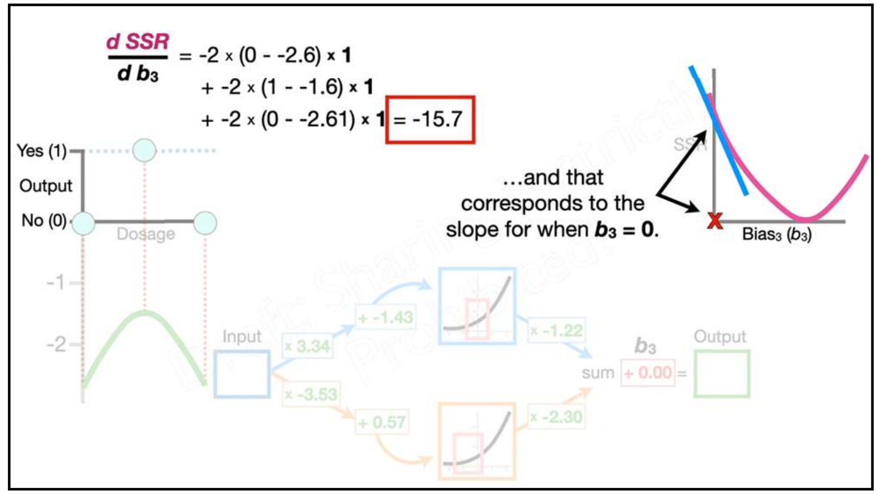
$$\frac{d\ SSR}{d\ b_3} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ b_3}$$

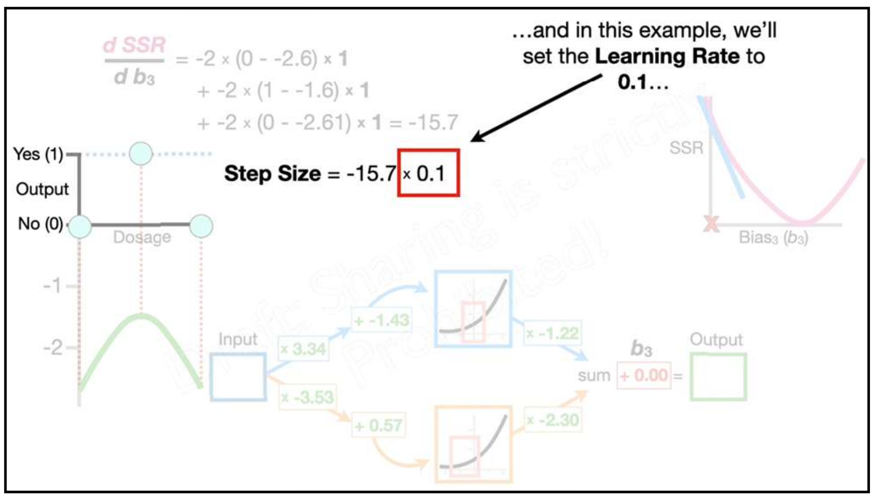
$$\boxed{= \sum_{i=1}^{n=3} -2 \times (\text{Observed}_i - \text{Predicted}_i) \times 1}$$

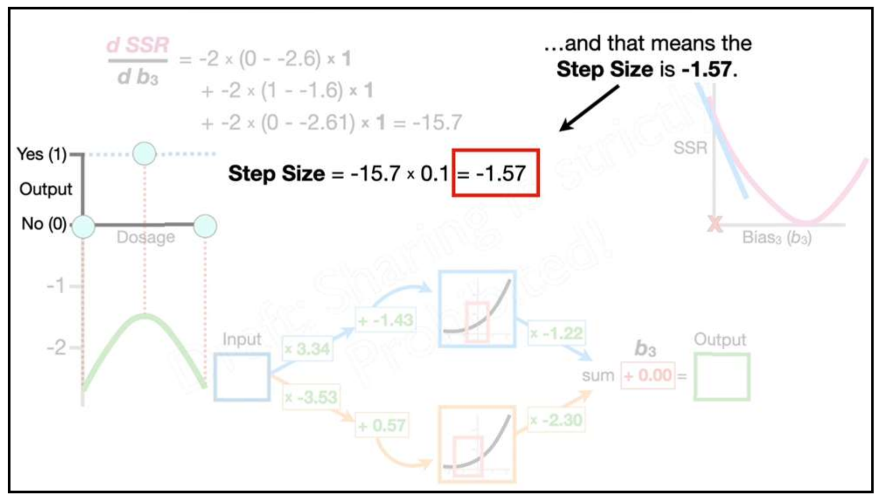


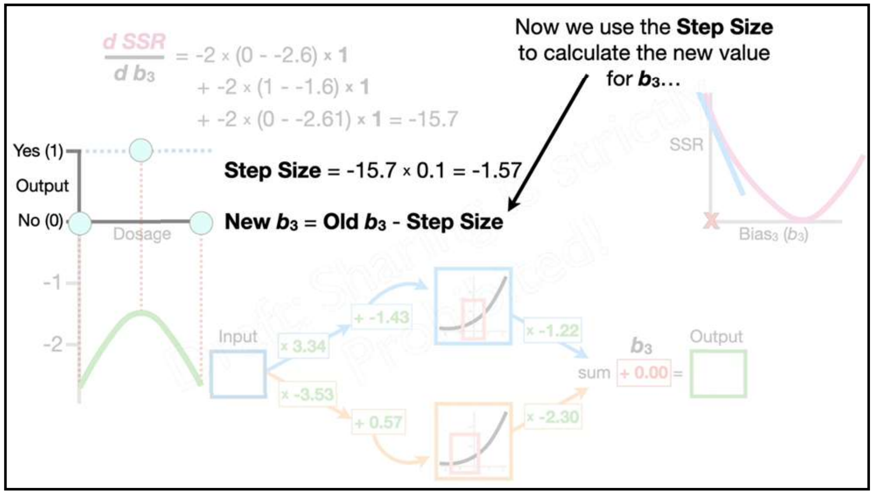

And that means we can plug this derivative into **Gradient Descent** to find the optimal value for $b_3$.


d SSR / d b3 = -2 x (0 - -2.6) x 1
+ -2 x (1 - -1.6) x 1
+ -2 x (0 - -2.61) x 1 = -15.7
Now we just do the math and get -15.7…
Yes (1)
Output
No (0)
Dosage
Input
Output
sum

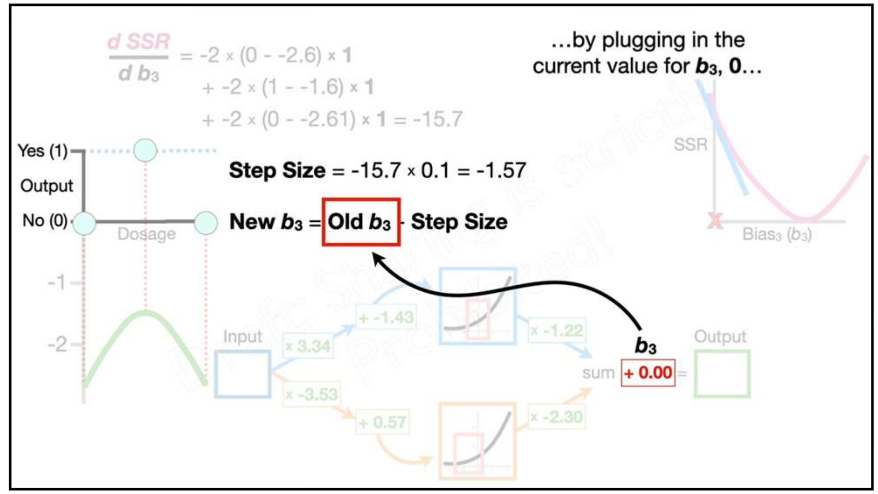

d SSR / d b3 = -2 x (0 - -2.6) x 1
+ -2 x (1 - -1.6) x 1
+ -2 x (0 - -2.61) x 1 = -15.7
…and that corresponds to the slope for when b3 = 0.
SSR
Bias3 (b3)
Yes (1)
Output
No (0)
Dosage


d SSR / d b3 = -2 x (0 - -2.6) x 1
+ -2 x (1 - -1.6) x 1
+ -2 x (0 - -2.61) x 1 = -15.7
…and in this example, we'll set the Learning Rate to 0.1…
Step Size = -15.7 x 0.1
SSR
Bias3 (b3)
Yes (1)
Output
No (0)
Dosage

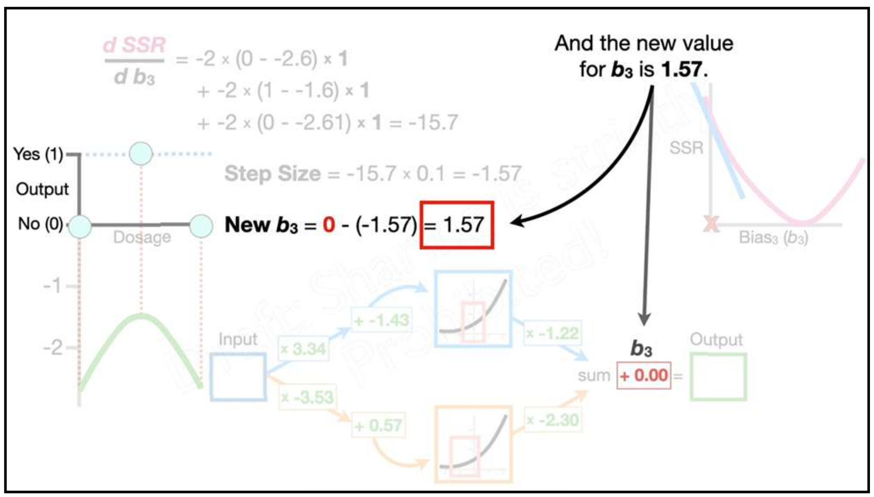

d SSR / d b3 = -2 x (0 - -2.6) x 1
+ -2 x (1 - -1.6) x 1
+ -2 x (0 - -2.61) x 1 = -15.7
…and that means the Step Size is -1.57.
Step Size = -15.7 x 0.1 = -1.57
SSR
Bias3 (b3)
Yes (1)
Output
No (0)
Dosage



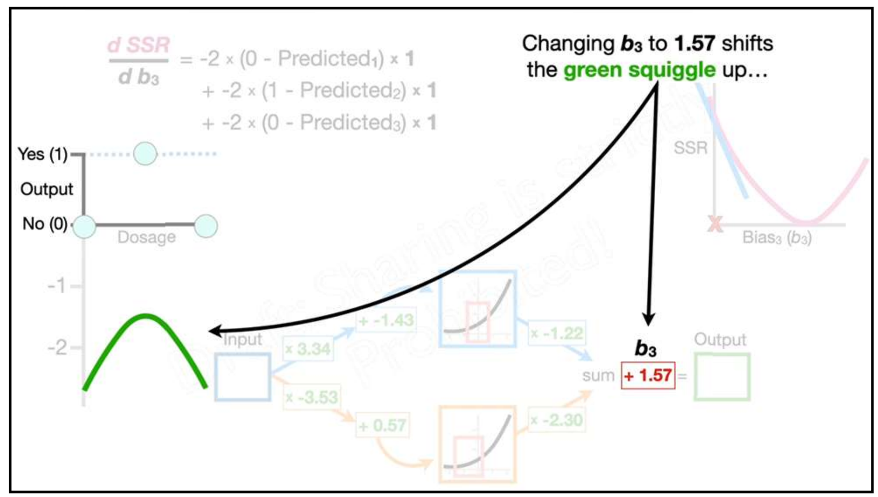
Now we use the **Step Size** to calculate the new value for ***b***$_3$…
$\frac{d\ SSR}{d\ b_3} = -2 \times (0 - -2.6) \times 1 + -2 \times (1 - -1.6) \times 1 + -2 \times (0 - -2.61) \times 1 = -15.7$
**Step Size** = -15.7 × 0.1 = -1.57
**New *b*$_3$ = Old *b*$_3$ - Step Size**
Yes (1)
Output
No (0)
Dosage
SSR
Bias$_3$ ($b_3$)
Input
x 3.34
x -3.53
+ -1.43
+ 0.57
x -1.22
x -2.30
$b_3$
sum + 0.00 =
Output

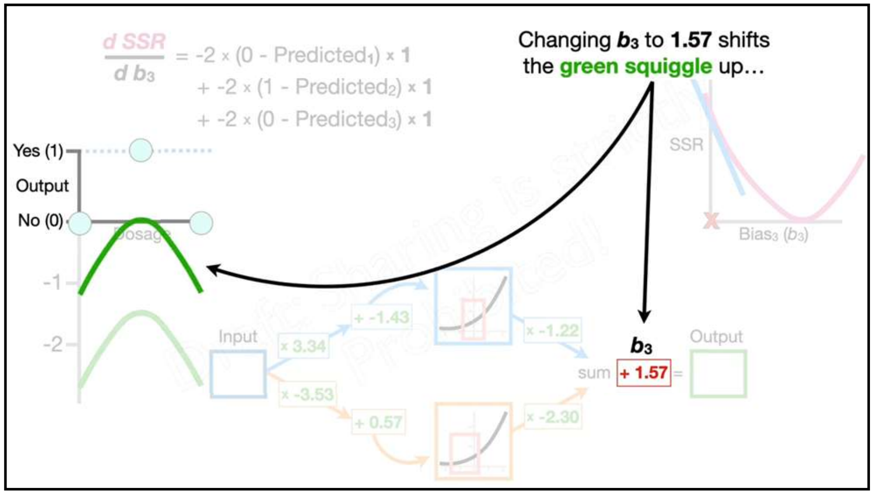
…by plugging in the current value for ***b***$_3$, **0**…
$\frac{d\ SSR}{d\ b_3} = -2 \times (0 - -2.6) \times 1 + -2 \times (1 - -1.6) \times 1 + -2 \times (0 - -2.61) \times 1 = -15.7$
**Step Size** = -15.7 × 0.1 = -1.57
**New *b*$_3$ = Old *b*$_3$ - Step Size**
Yes (1)
Output
No (0)
Dosage
SSR
Bias$_3$ ($b_3$)
Input
$b_3$
sum + 0.00 =
Output

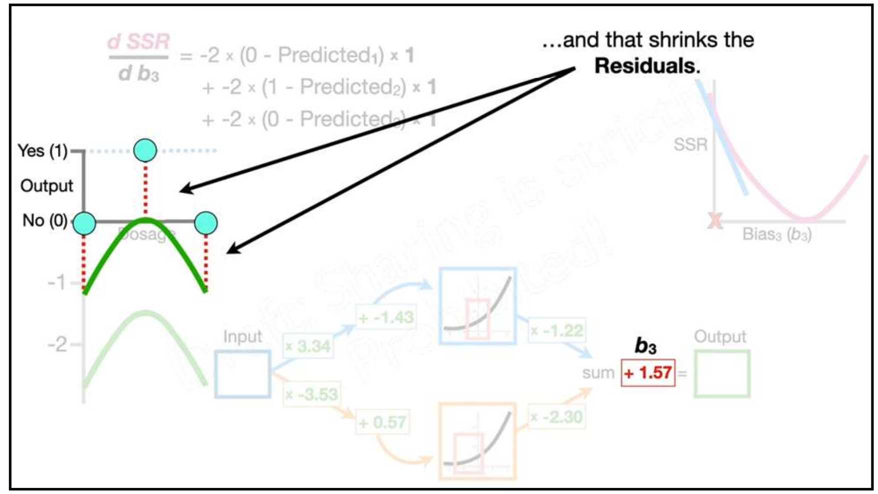



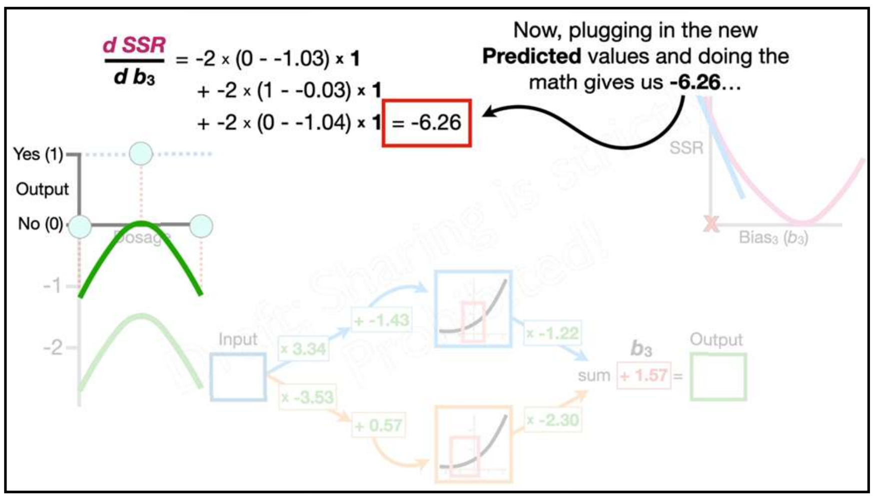
And the new value for ***b***$_3$ is **1.57**.
$\frac{d\ SSR}{d\ b_3} = -2 \times (0 - -2.6) \times 1 + -2 \times (1 - -1.6) \times 1 + -2 \times (0 - -2.61) \times 1 = -15.7$
Step Size = -15.7 × 0.1 = -1.57
**New *b*$_3$** = 0 - (-1.57) = 1.57
Yes (1)
Output
No (0)
Dosage
SSR
Bias$_3$ ($b_3$)
Input
$b_3$
sum + 0.00 =
Output

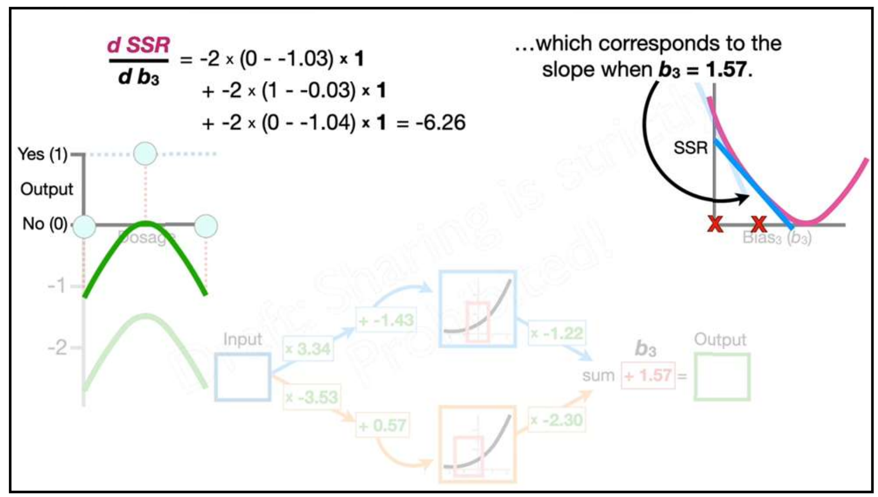



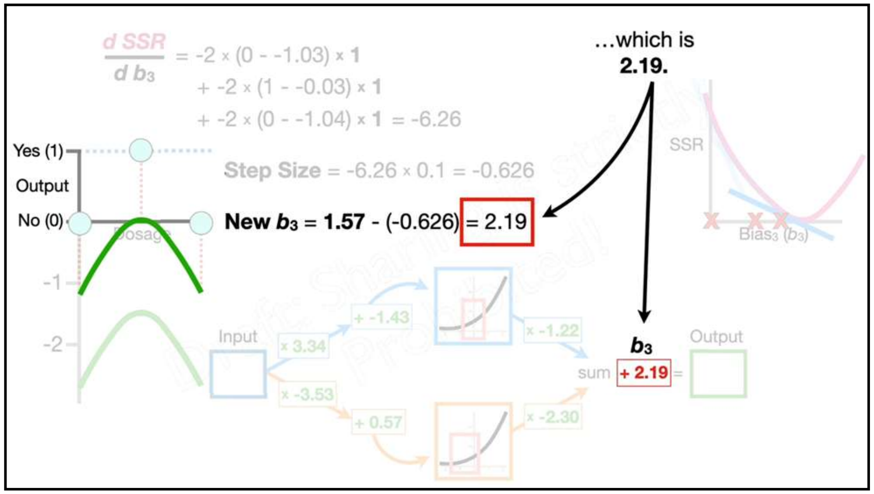
Changing ***b***$_3$ to **1.57** shifts the **green squiggle** up…
$\frac{d\ SSR}{d\ b_3} = -2 \times (0 - \text{Predicted}_1) \times 1 + -2 \times (1 - \text{Predicted}_2) \times 1 + -2 \times (0 - \text{Predicted}_3) \times 1$
Yes (1)
Output
No (0)
Dosage
SSR
Bias$_3$ ($b_3$)
Input
$b_3$
sum + 1.57 =
Output


Changing $b_3$ to 1.57 shifts the green squiggle up…
d SSR / d $b_3$ = -2 x (0 - Predicted$_1$) x 1
+ -2 x (1 - Predicted$_2$) x 1
+ -2 x (0 - Predicted$_3$) x 1
Yes (1)
Output
No (0)
Dosage
-1
-2
Input
x 3.34
x -3.53
+ -1.43
+ 0.57
x -1.22
x -2.30
$b_3$
sum
+ 1.57
Output
SSR
Bias$_3$ ($b_3$)

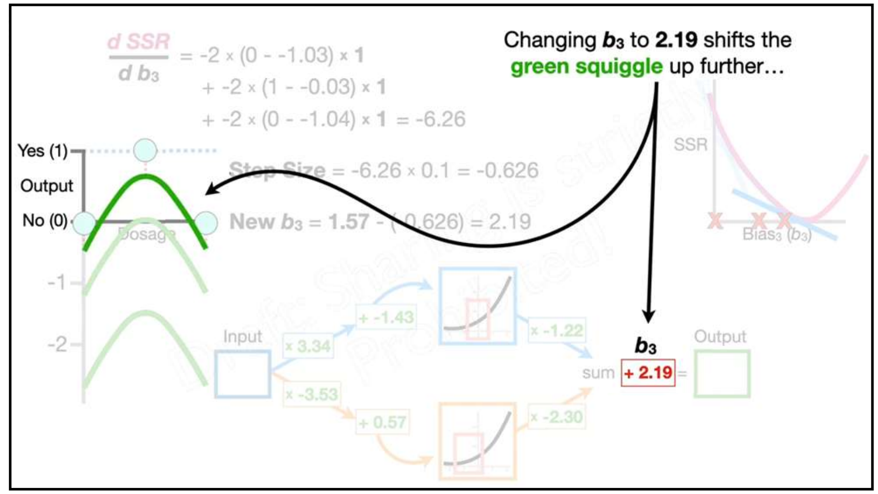

…and that shrinks the Residuals.
d SSR / d $b_3$ = -2 x (0 - Predicted$_1$) x 1
+ -2 x (1 - Predicted$_2$) x 1
+ -2 x (0 - Predicted$_3$) x 1
Yes (1)
Output
No (0)
Dosage
-1
-2
Input
x 3.34
x -3.53
+ -1.43
+ 0.57
x -1.22
x -2.30
$b_3$
sum
+ 1.57
Output
SSR
Bias$_3$ ($b_3$)


Now, plugging in the new Predicted values and doing the math gives us -6.26…
d SSR / d $b_3$ = -2 x (0 - -1.03) x 1
+ -2 x (1 - -0.03) x 1
+ -2 x (0 - -1.04) x 1 = -6.26
Yes (1)
Output
No (0)
Dosage
-1
-2
Input
x 3.34
x -3.53
+ -1.43
+ 0.57
x -1.22
x -2.30
$b_3$
sum
+ 1.57
Output
SSR
Bias$_3$ ($b_3$)

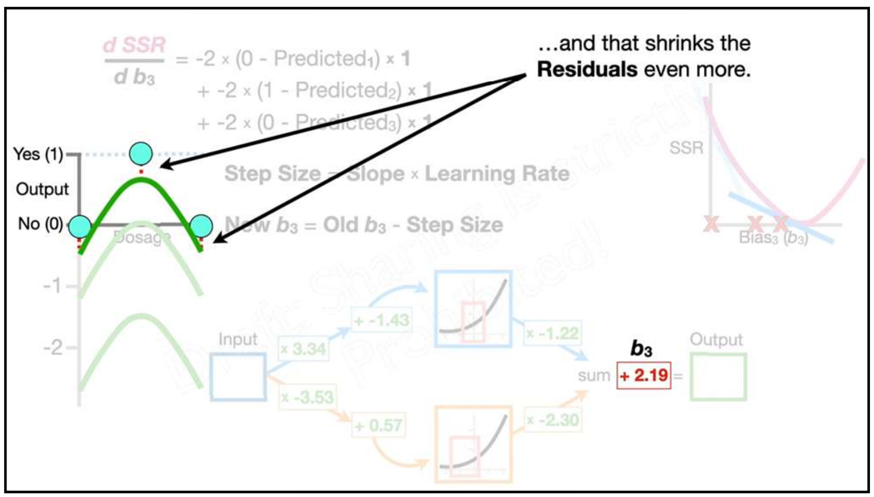

…which corresponds to the slope when $b_3$ = 1.57.
d SSR / d $b_3$ = -2 x (0 - -1.03) x 1
+ -2 x (1 - -0.03) x 1
+ -2 x (0 - -1.04) x 1 = -6.26
Yes (1)
Output
No (0)
Dosage
-1
-2
Input
x 3.34
x -3.53
+ -1.43
+ 0.57
x -1.22
x -2.30
$b_3$
sum
+ 1.57
Output
SSR
Bias$_3$ ($b_3$)



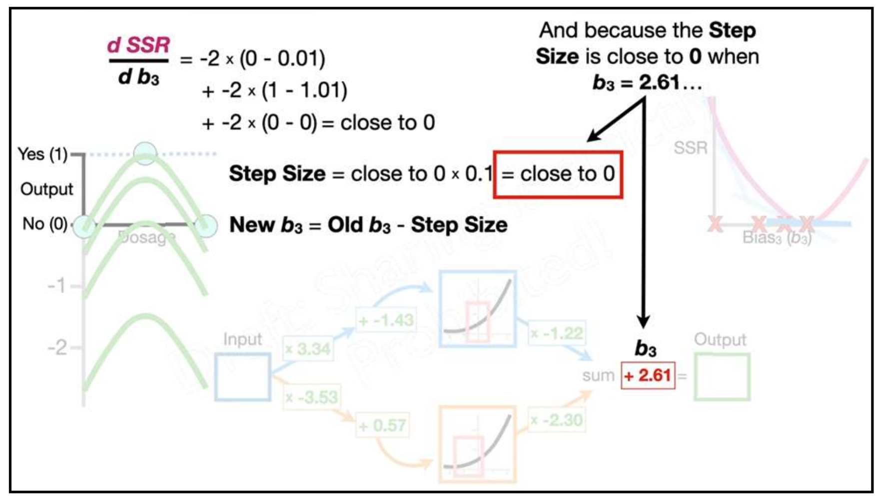
…which is
2.19.
d SSR / d b3 = -2 x (0 - -1.03) x 1
+ -2 x (1 - -0.03) x 1
+ -2 x (0 - -1.04) x 1 = -6.26
Step Size = -6.26 x 0.1 = -0.626
New b3 = 1.57 - (-0.626) = 2.19
Yes (1)
Output
No (0)
Dosage
-1
-2
SSR
Bias3 (b3)
Input
x 3.34
+ -1.43
x -1.22
x -3.53
+ 0.57
x -2.30
b3
sum + 2.19 =
Output

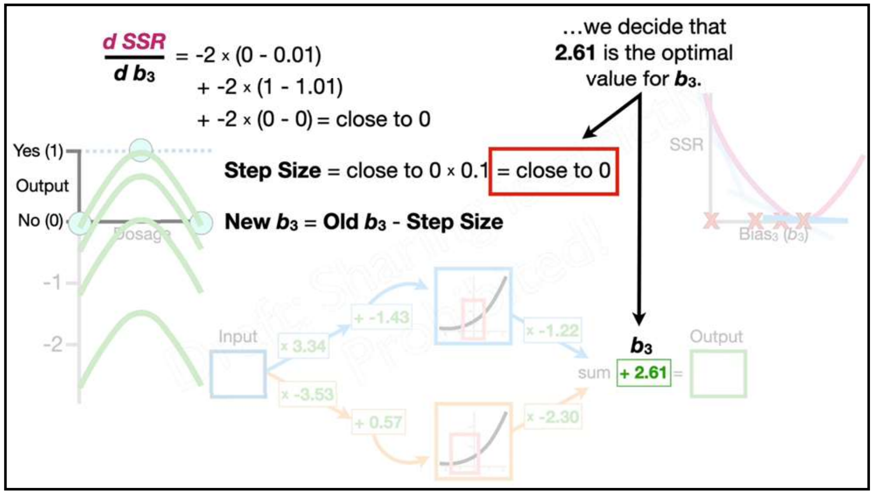
Changing b3 to 2.19 shifts the
green squiggle up further…
d SSR / d b3 = -2 x (0 - -1.03) x 1
+ -2 x (1 - -0.03) x 1
+ -2 x (0 - -1.04) x 1 = -6.26
Step Size = -6.26 x 0.1 = -0.626
New b3 = 1.57 - (-0.626) = 2.19
Yes (1)
Output
No (0)
Dosage
-1
-2
SSR
Bias3 (b3)
Input
x 3.34
+ -1.43
x -1.22
x -3.53
+ 0.57
x -2.30
b3
sum + 2.19 =
Output

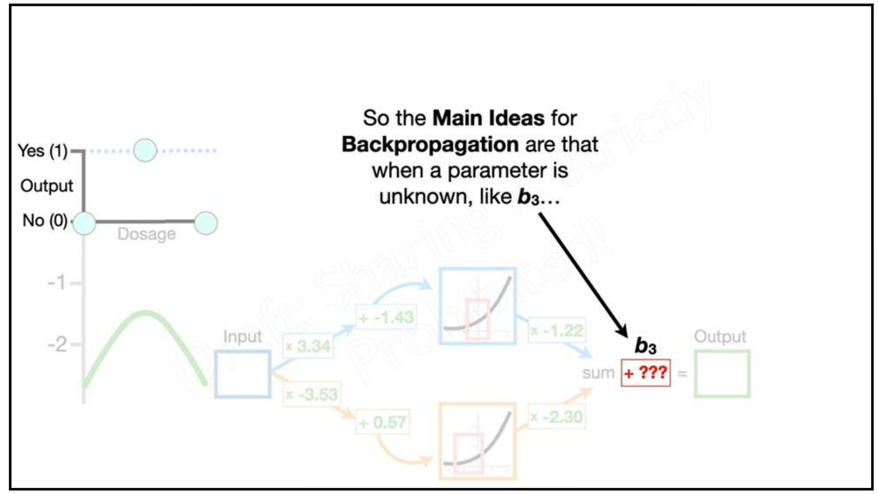



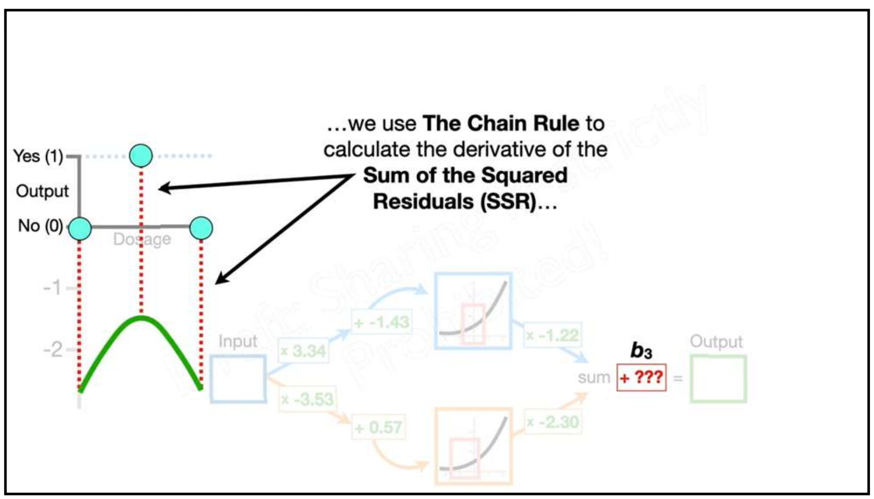
…and that shrinks the
Residuals even more.
d SSR / d b3 = -2 x (0 - Predicted1) x 1
+ -2 x (1 - Predicted2) x 1
+ -2 x (0 - Predicted3) x 1
Step Size = Slope x Learning Rate
New b3 = Old b3 - Step Size
Yes (1)
Output
No (0)
Dosage
-1
-2
SSR
Bias3 (b3)
Input
x 3.34
+ -1.43
x -1.22
x -3.53
+ 0.57
x -2.30
b3
sum + 2.19 =
Output

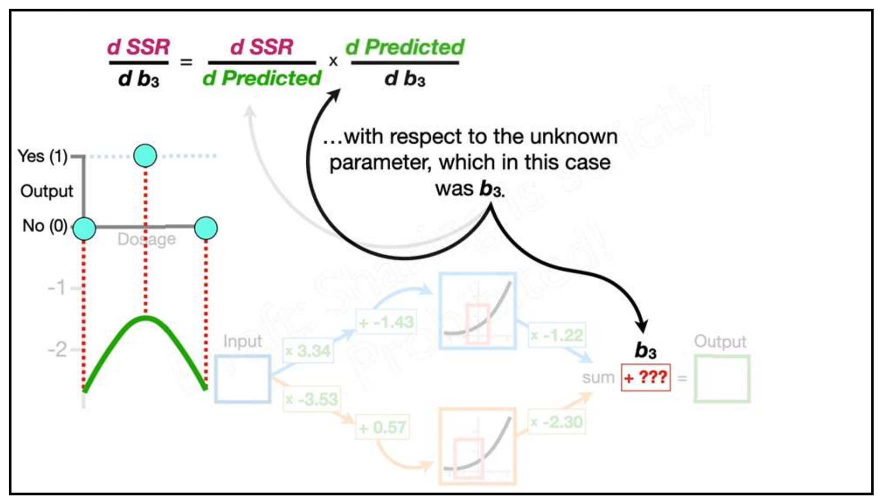



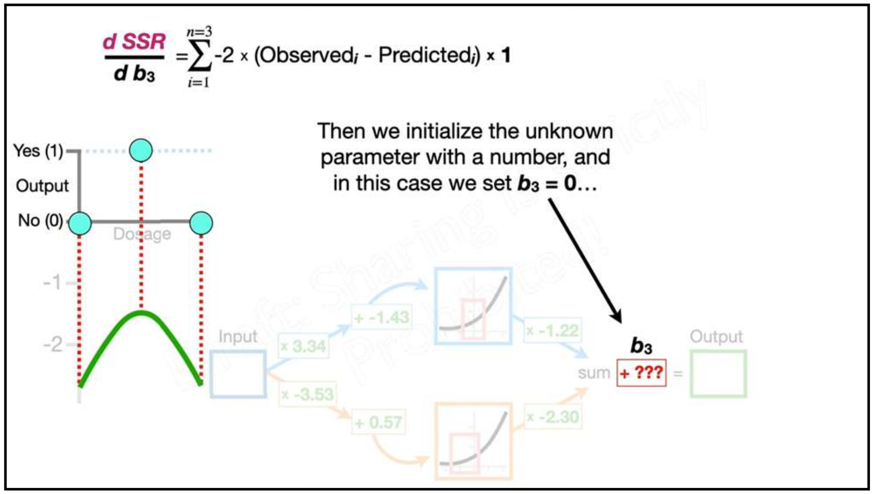
And because the Step
Size is close to 0 when
b3 = 2.61…
d SSR / d b3 = -2 x (0 - 0.01)
+ -2 x (1 - 1.01)
+ -2 x (0 - 0) = close to 0
Step Size = close to 0 x 0.1 = close to 0
New b3 = Old b3 - Step Size
Yes (1)
Output
No (0)
Dosage
-1
-2
SSR
Bias3 (b3)
Input
x 3.34
+ -1.43
x -1.22
x -3.53
+ 0.57
x -2.30
b3
sum + 2.61 =
Output

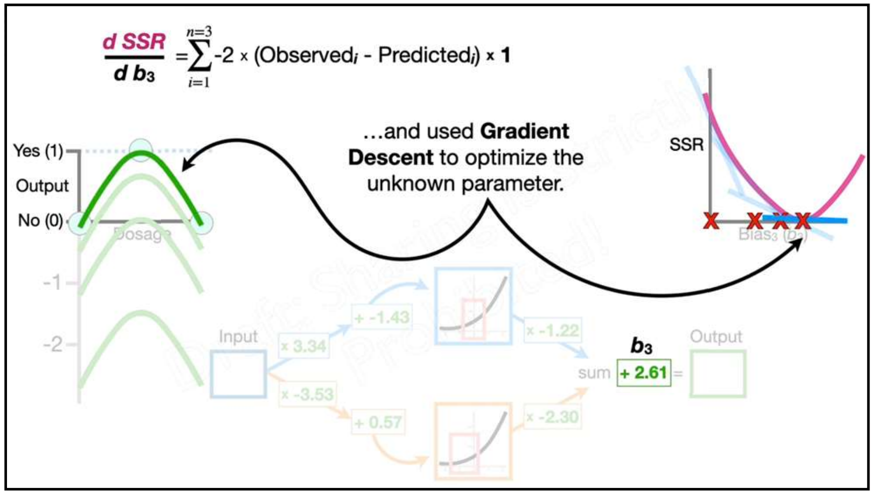

d SSR / d b3 = -2 x (0 - 0.01)
+ -2 x (1 - 1.01)
+ -2 x (0 - 0) = close to 0
...we decide that
2.61 is the optimal
value for b3.
Step Size = close to 0 x 0.1 = close to 0
New b3 = Old b3 - Step Size
Yes (1)
Output
No (0)
Dosage
-1
-2
Input
SSR
Bias3 (b3)
b3
sum
+ 2.61
Output

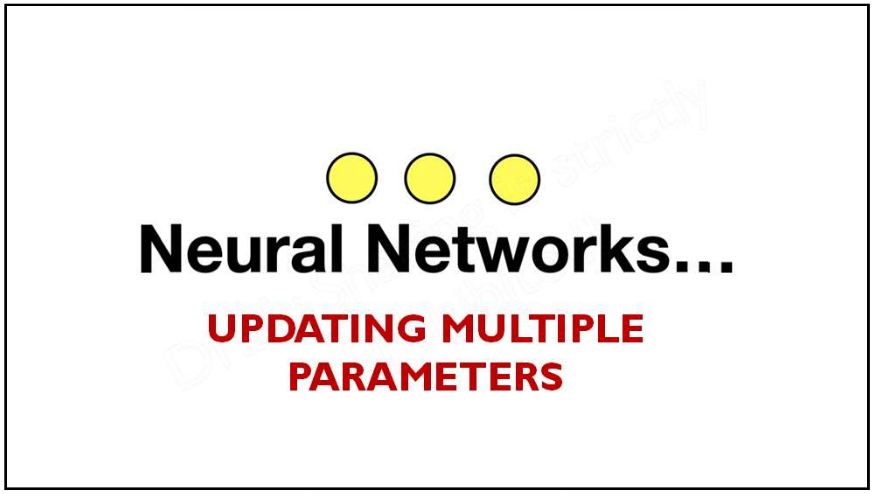

So the Main Ideas for
Backpropagation are that
when a parameter is
unknown, like b3...
Yes (1)
Output
No (0)
Dosage
-1
-2
Input
b3
sum
+ ???
Output

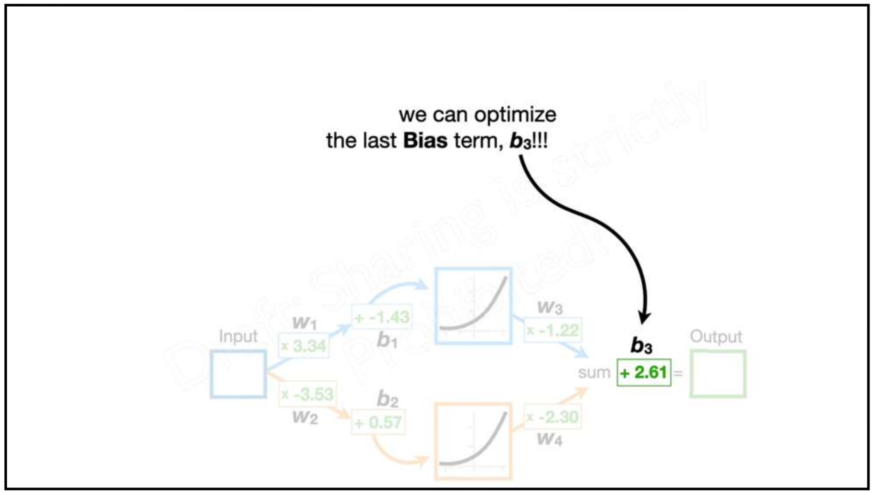

...we use The Chain Rule to
calculate the derivative of the
Sum of the Squared
Residuals (SSR)...
Yes (1)
Output
No (0)
Dosage
-1
-2
Input
b3
sum
+ ???
Output

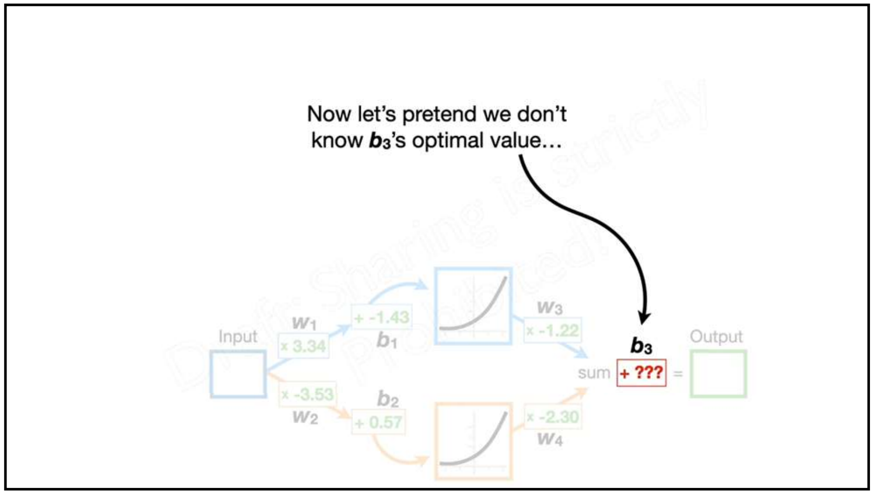

d SSR / d b3 = d SSR / d Predicted x d Predicted / d b3
...with respect to the unknown
parameter, which in this case
was b3.
Yes (1)
Output
No (0)
Dosage
-1
-2
Input
b3
sum
+ ???
Output

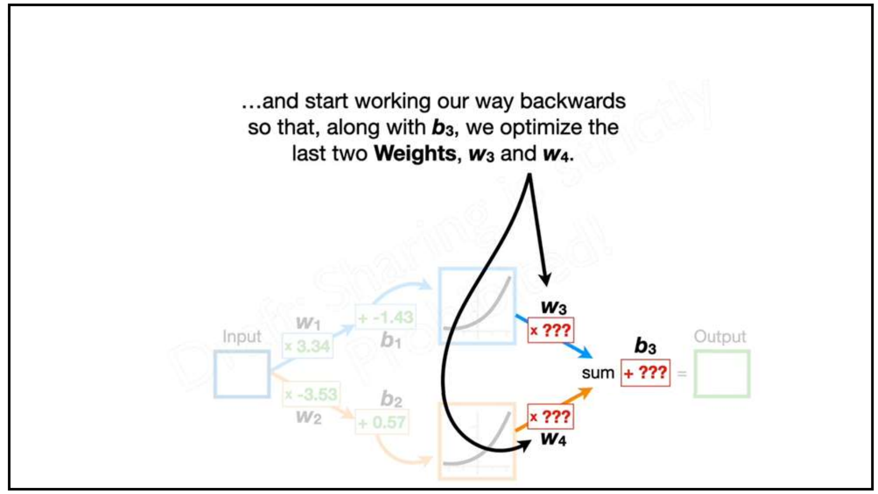

$\frac{d\,SSR}{d\,b_3} = \sum_{i=1}^{n=3} -2 \times (\text{Observed}_i - \text{Predicted}_i) \times 1$
Then we initialize the unknown parameter with a number, and in this case we set $b_3 = 0$…
Yes (1)
Output
No (0)
Dosage
-1
-2
Input
x 3.34
x -3.53
+ -1.43
+ 0.57
x -1.22
x -2.30
$b_3$
sum + ??? =
Output

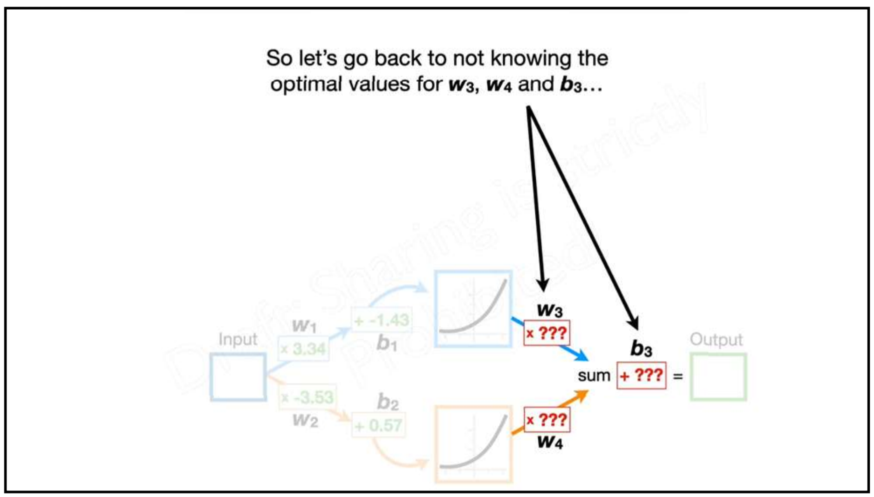

$\frac{d\,SSR}{d\,b_3} = \sum_{i=1}^{n=3} -2 \times (\text{Observed}_i - \text{Predicted}_i) \times 1$
…and used **Gradient Descent** to optimize the unknown parameter.
SSR
Yes (1)
Output
No (0)
-1
-2
Input
x 3.34
x -3.53
+ -1.43
+ 0.57
x -1.22
x -2.30
$b_3$
sum + 2.61 =
Output

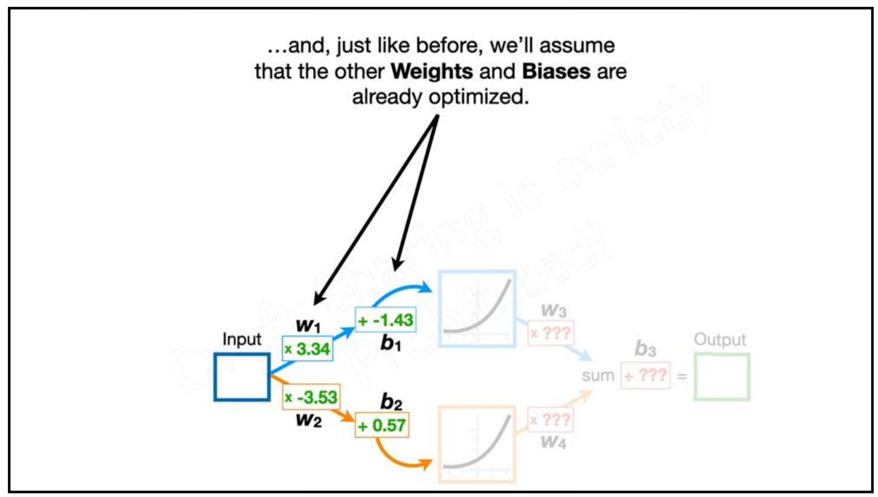

Neural Networks…
UPDATING MULTIPLE PARAMETERS

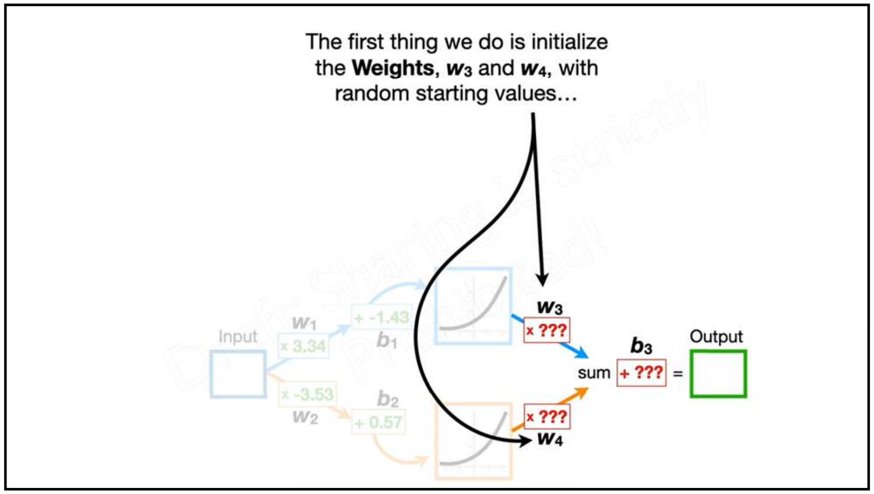

we can optimize the last **Bias** term, $b_3$!!!
Input
$w_1$
x 3.34
x -3.53
$w_2$
+ -1.43
$b_1$
$b_2$
+ 0.57
$w_3$
x -1.22
x -2.30
$w_4$
$b_3$
sum + 2.61 =
Output

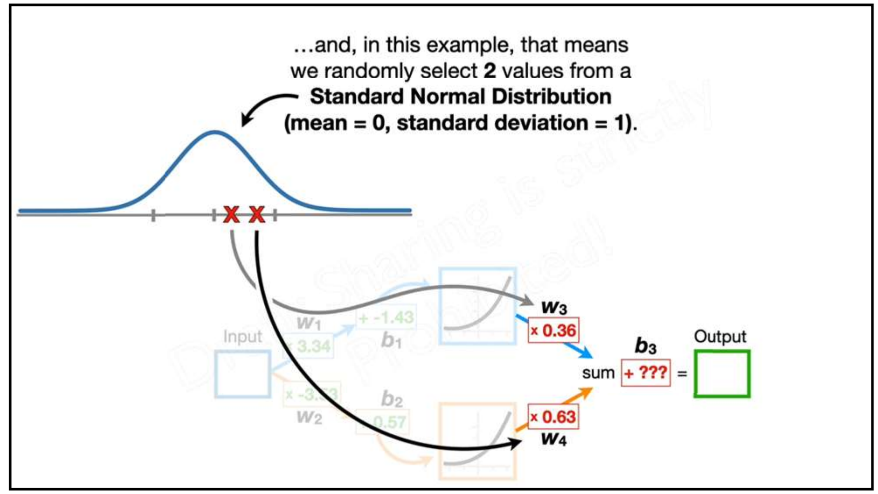

Now let's pretend we don't
know b3's optimal value…
Input
w1
x 3.34
+ -1.43
b1
w3
x -1.22
b3
sum
+ ???
=
Output
x -3.53
w2
b2
+ 0.57
x -2.30
w4

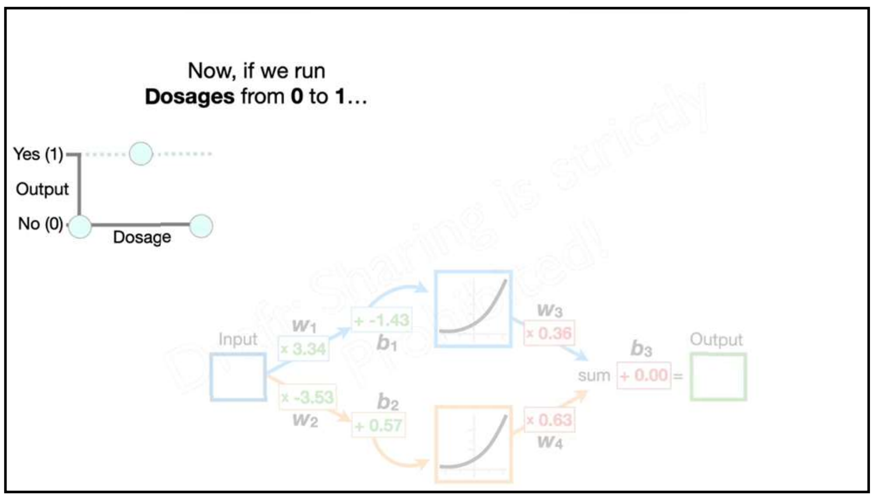

…and start working our way backwards
so that, along with b3, we optimize the
last two Weights, w3 and w4.
Input
w1
x 3.34
+ -1.43
b1
w3
x ???
b3
sum
+ ???
=
Output
x -3.53
w2
b2
+ 0.57
x ???
w4

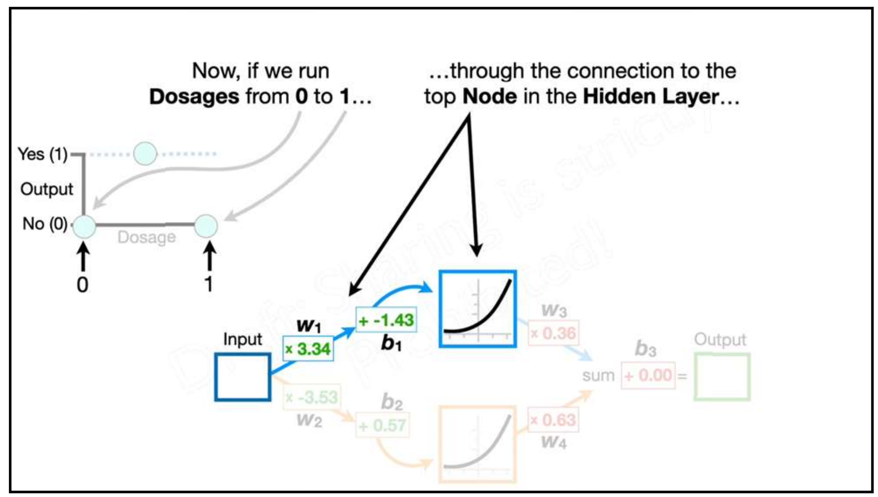

So let's go back to not knowing the
optimal values for w3, w4 and b3…
Input
w1
x 3.34
+ -1.43
b1
w3
x ???
b3
sum
+ ???
=
Output
x -3.53
w2
b2
+ 0.57
x ???
w4

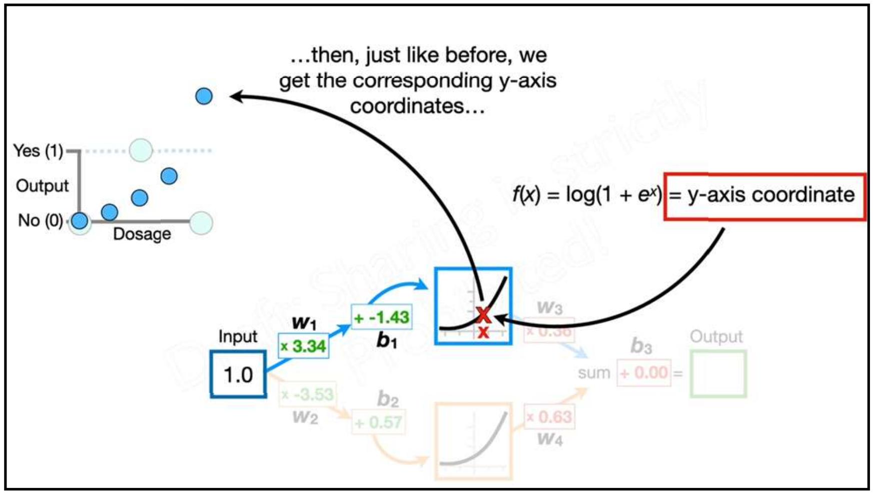

…and, just like before, we'll assume
that the other Weights and Biases are
already optimized.
Input
w1
x 3.34
+ -1.43
b1
w3
x ???
b3
sum
+ ???
=
Output
x -3.53
w2
b2
+ 0.57
x ???
w4

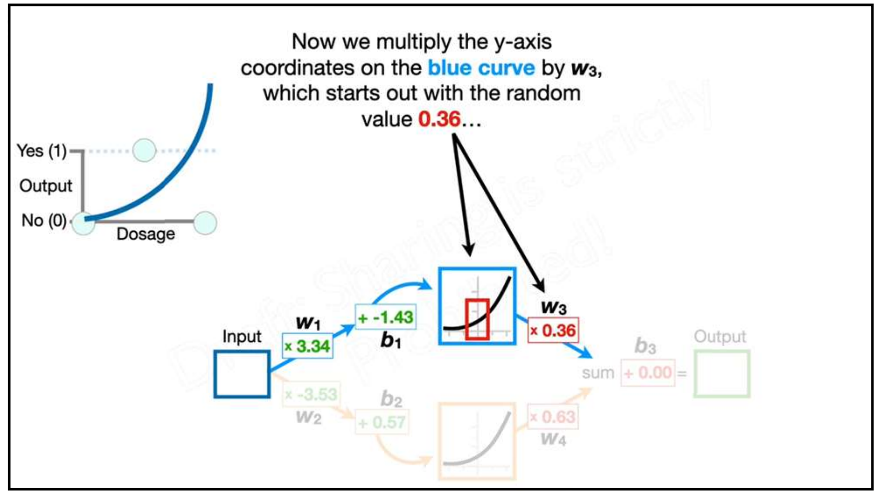

The first thing we do is initialize
the **Weights**, $w_3$ and $w_4$, with
random starting values…
Input
$w_1$
x 3.34
+ -1.43
$b_1$
x -3.53
$w_2$
$b_2$
+ 0.57
$w_3$
x ???
x ???
$w_4$
$b_3$
sum
+ ???
=
Output

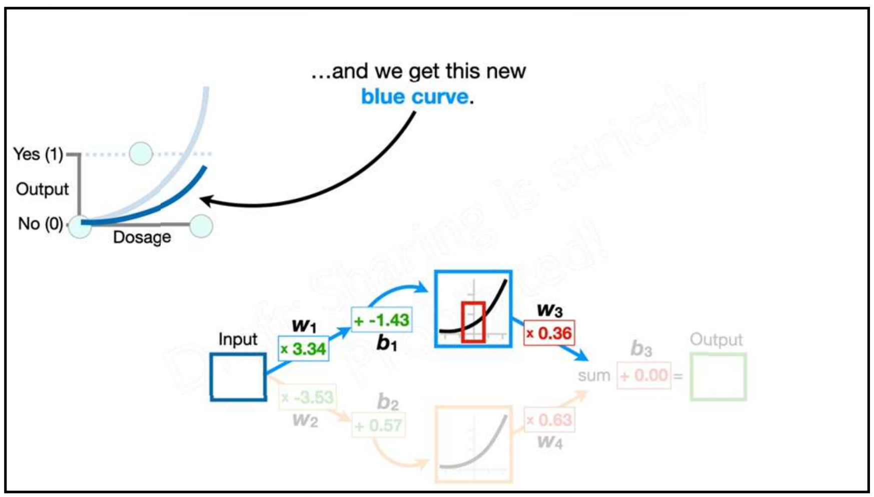

…and, in this example, that means
we randomly select **2** values from a
**Standard Normal Distribution**
**(mean = 0, standard deviation = 1).**
Input
$w_1$
$b_1$
+ -1.43
$w_2$
$b_2$
$w_3$
x 0.36
x 0.63
$w_4$
$b_3$
sum
+ ???
=
Output

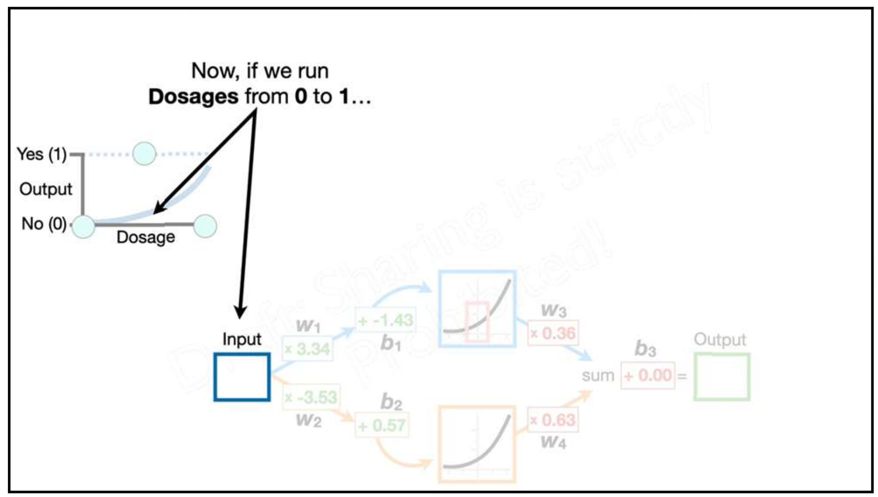

Now, if we run
**Dosages** from **0** to **1**…
Yes (1)
Output
No (0)
Dosage
Input
$w_1$
x 3.34
+ -1.43
$b_1$
x -3.53
$w_2$
$b_2$
+ 0.57
$w_3$
x 0.36
x 0.63
$w_4$
$b_3$
sum
+ 0.00
=
Output

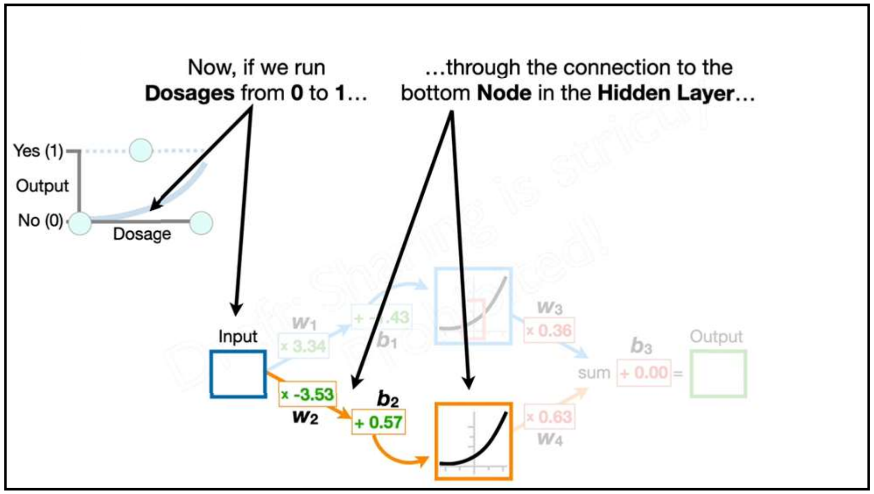

Now, if we run
**Dosages** from **0** to **1**…
…through the connection to the
top **Node** in the **Hidden Layer**…
Yes (1)
Output
No (0)
Dosage
0
1
Input
$w_1$
x 3.34
+ -1.43
$b_1$
x -3.53
$w_2$
$b_2$
+ 0.57
$w_3$
x 0.36
x 0.63
$w_4$
$b_3$
sum
+ 0.00
=
Output

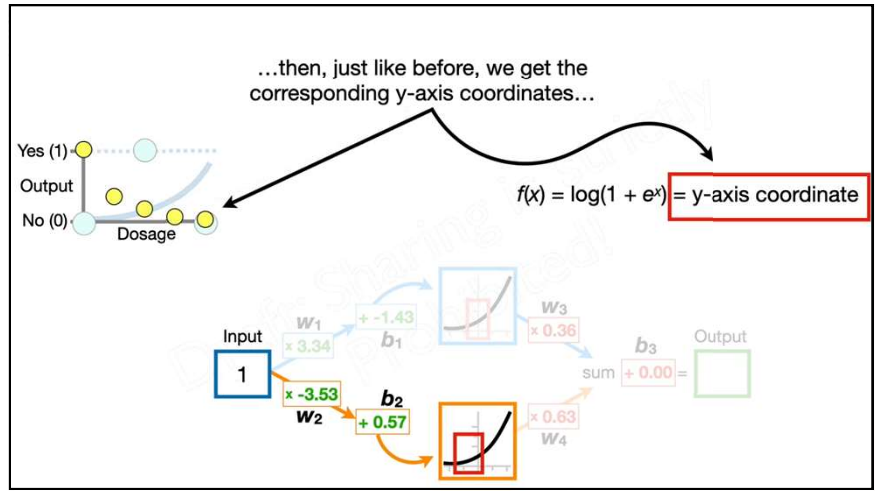

…then, just like before, we get the corresponding y-axis coordinates…
f(x) = log(1 + e^x) = y-axis coordinate
Yes (1)
Output
No (0)
Dosage
Input
1.0
w1
x 3.34
+ -1.43
b1
w3
x 0.36
b3
sum + 0.00 =
Output
x -3.53
w2
b2
+ 0.57
x 0.63
w4

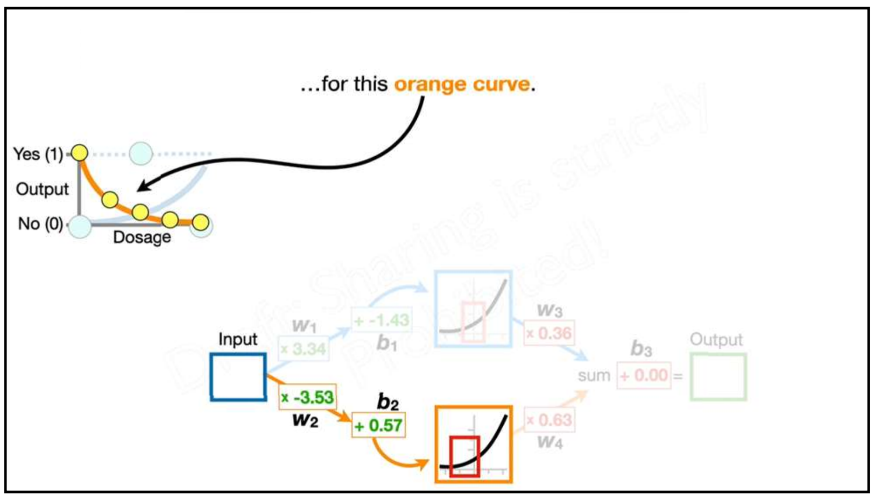

Now we multiply the y-axis coordinates on the blue curve by w3, which starts out with the random value 0.36…
Yes (1)
Output
No (0)
Dosage
Input
w1
x 3.34
+ -1.43
b1
w3
x 0.36
b3
sum + 0.00 =
Output
x -3.53
w2
b2
+ 0.57
x 0.63
w4

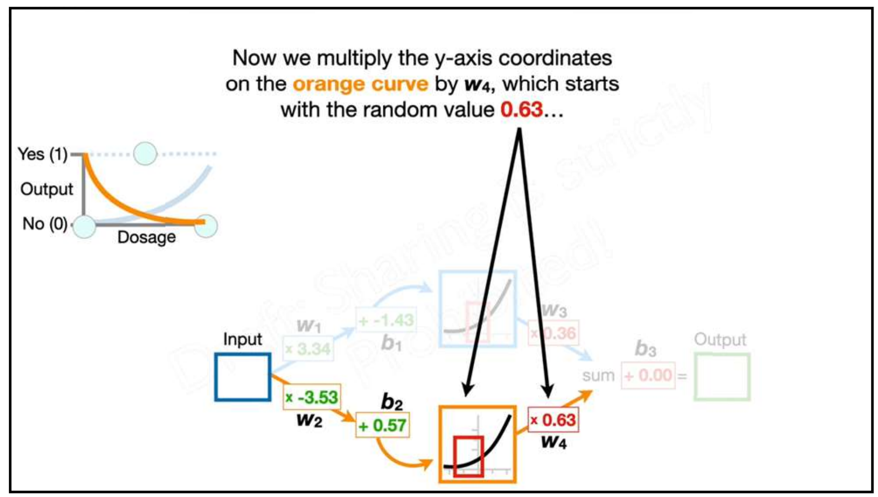

…and we get this new blue curve.
Yes (1)
Output
No (0)
Dosage
Input
w1
x 3.34
+ -1.43
b1
w3
x 0.36
b3
sum + 0.00 =
Output
x -3.53
w2
b2
+ 0.57
x 0.63
w4

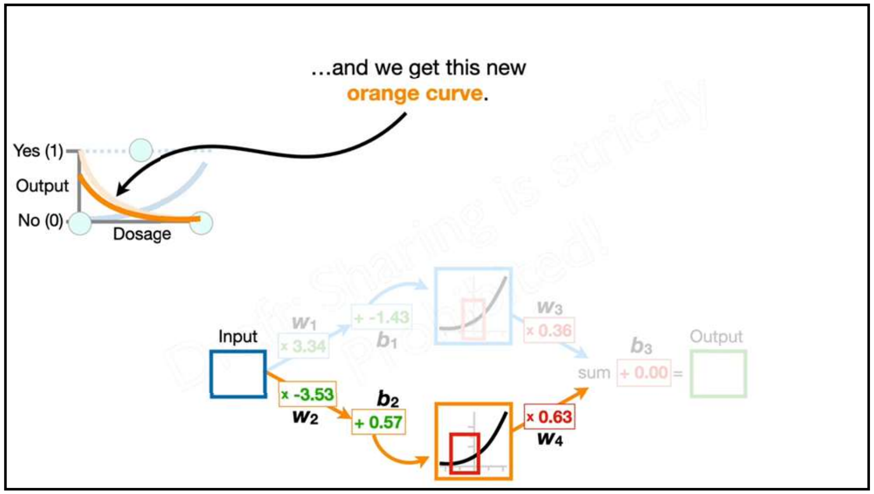

Now, if we run Dosages from 0 to 1…
Yes (1)
Output
No (0)
Dosage
Input
w1
x 3.34
+ -1.43
b1
w3
x 0.36
b3
sum + 0.00 =
Output
x -3.53
w2
b2
+ 0.57
x 0.63
w4

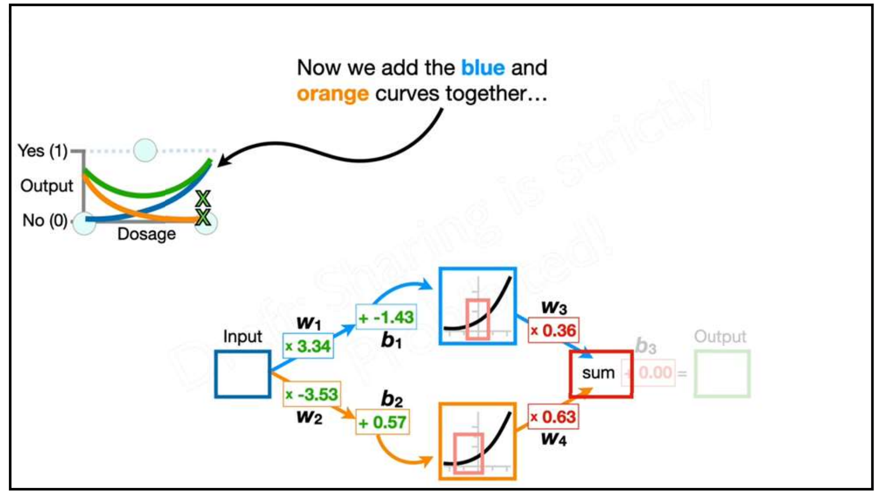

Now, if we run
**Dosages** from **0** to **1**…
…through the connection to the
bottom **Node** in the **Hidden Layer**…
Yes (1)
Output
No (0)
Dosage
Input
x -3.53
w2
b2
+ 0.57

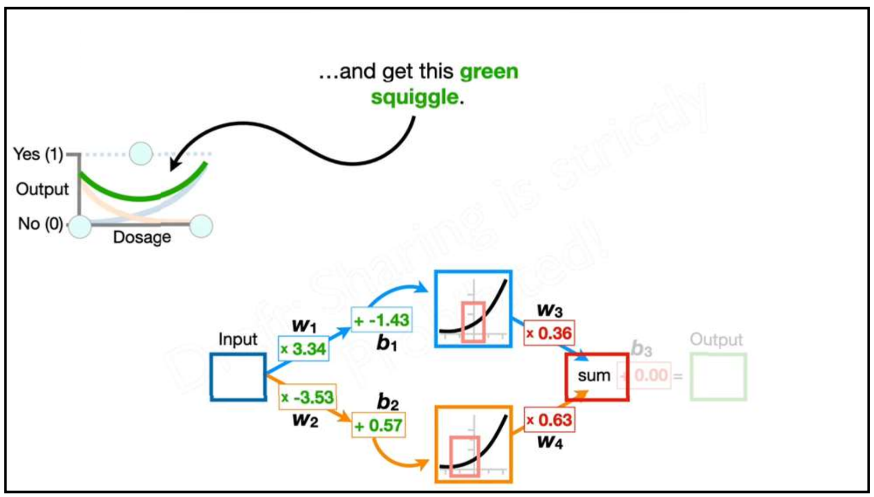

…then, just like before, we get the
corresponding y-axis coordinates…
f(x) = log(1 + e^x) = y-axis coordinate
Yes (1)
Output
No (0)
Dosage
Input
1
x -3.53
w2
b2
+ 0.57

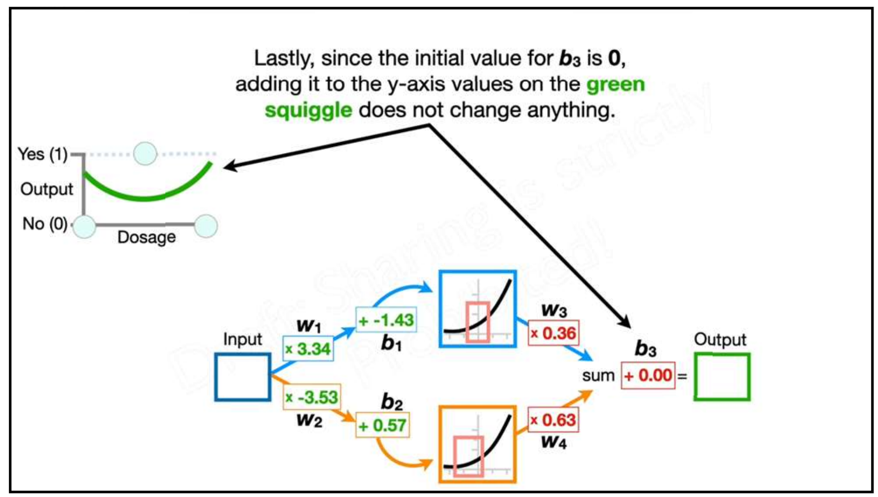

…for this orange curve.
Yes (1)
Output
No (0)
Dosage
Input
x -3.53
w2
b2
+ 0.57

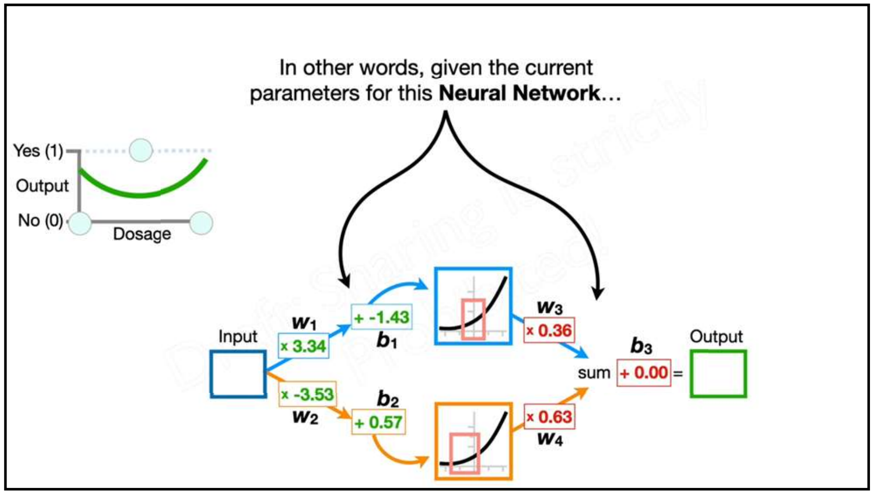

Now we multiply the y-axis coordinates
on the orange curve by w4, which starts
with the random value 0.63…
Yes (1)
Output
No (0)
Dosage
Input
x -3.53
w2
b2
+ 0.57
x 0.63
w4

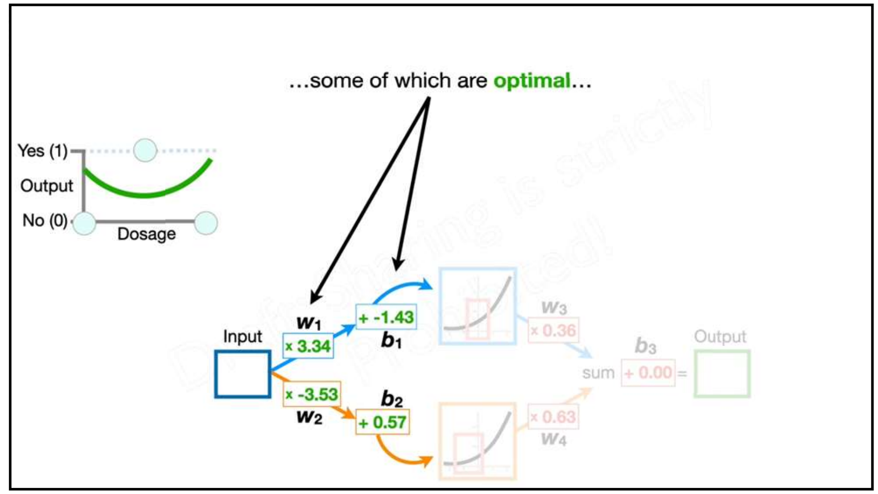

...and we get this new orange curve.
Yes (1)
Output
No (0)
Dosage
Input
$w_1$
x 3.34
+ -1.43
$b_1$
x -3.53
$w_2$
$b_2$
+ 0.57
$w_3$
x 0.36
x 0.63
$w_4$
sum
$b_3$
+ 0.00
Output

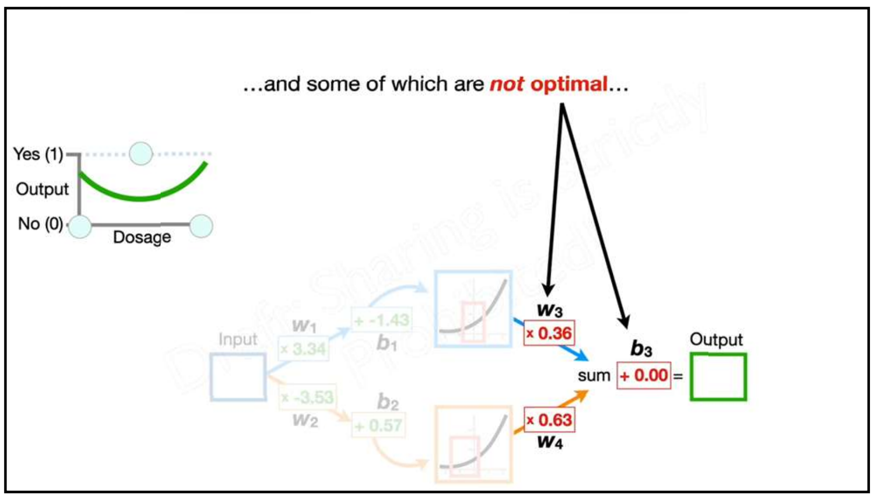

Now we add the blue and orange curves together...
Yes (1)
Output
No (0)
Dosage
Input
$w_1$
x 3.34
+ -1.43
$b_1$
x -3.53
$w_2$
$b_2$
+ 0.57
$w_3$
x 0.36
x 0.63
$w_4$
sum
$b_3$
0.00
Output

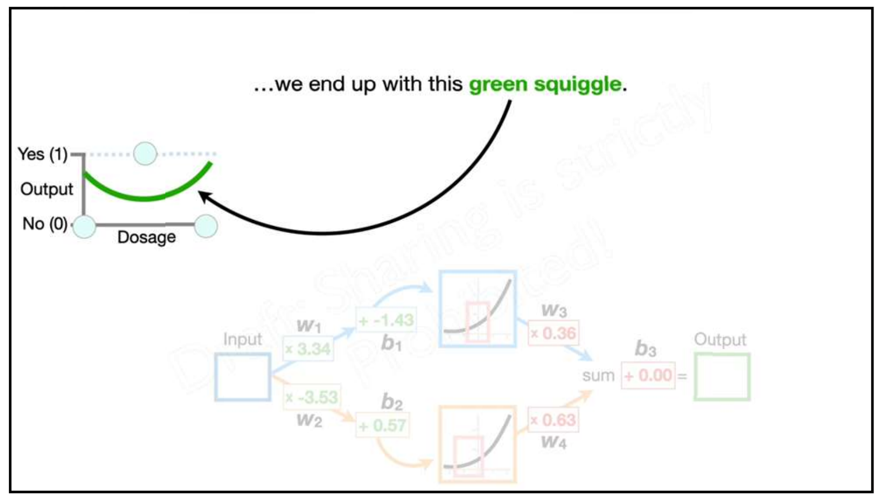

...and get this green squiggle.
Yes (1)
Output
No (0)
Dosage
Input
$w_1$
x 3.34
+ -1.43
$b_1$
x -3.53
$w_2$
$b_2$
+ 0.57
$w_3$
x 0.36
x 0.63
$w_4$
sum
$b_3$
0.00
Output

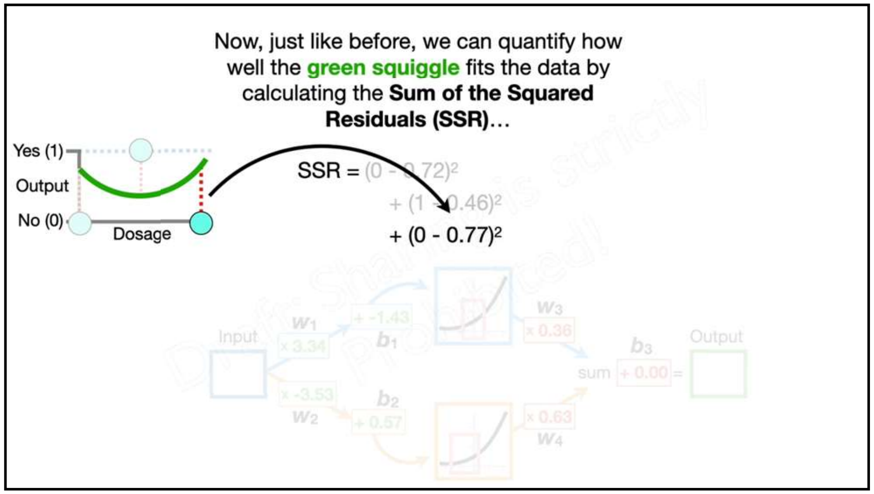

Lastly, since the initial value for $b_3$ is 0, adding it to the y-axis values on the green squiggle does not change anything.
Yes (1)
Output
No (0)
Dosage
Input
$w_1$
x 3.34
+ -1.43
$b_1$
x -3.53
$w_2$
$b_2$
+ 0.57
$w_3$
x 0.36
x 0.63
$w_4$
sum
$b_3$
+ 0.00
Output

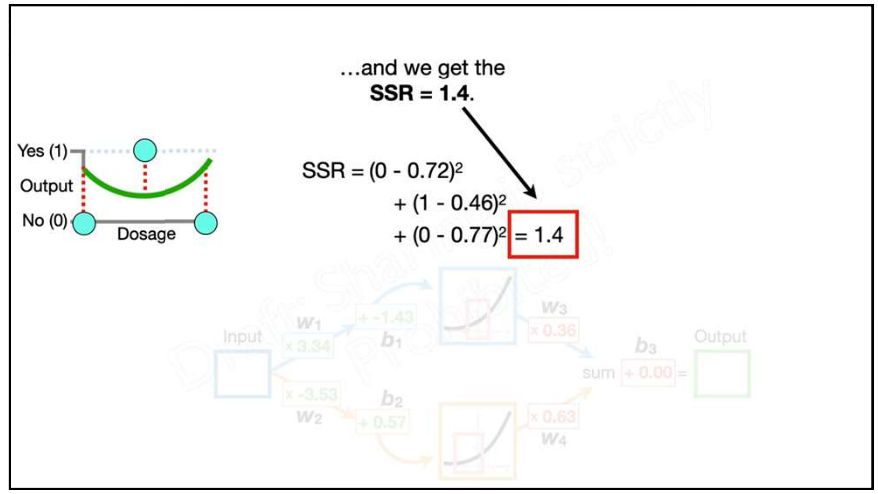

In other words, given the current parameters for this **Neural Network**…
Yes (1)
Output
No (0)
Dosage
Input
$w_1$
x 3.34
+ -1.43
$b_1$
x -3.53
$w_2$
$b_2$
+ 0.57
$w_3$
x 0.36
$b_3$
sum
+ 0.00
=
Output
x 0.63
$w_4$

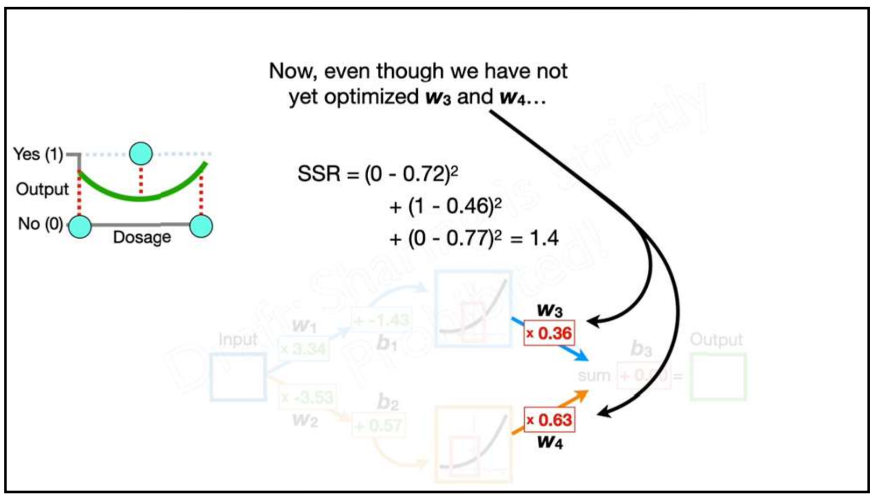

…some of which are **optimal**…

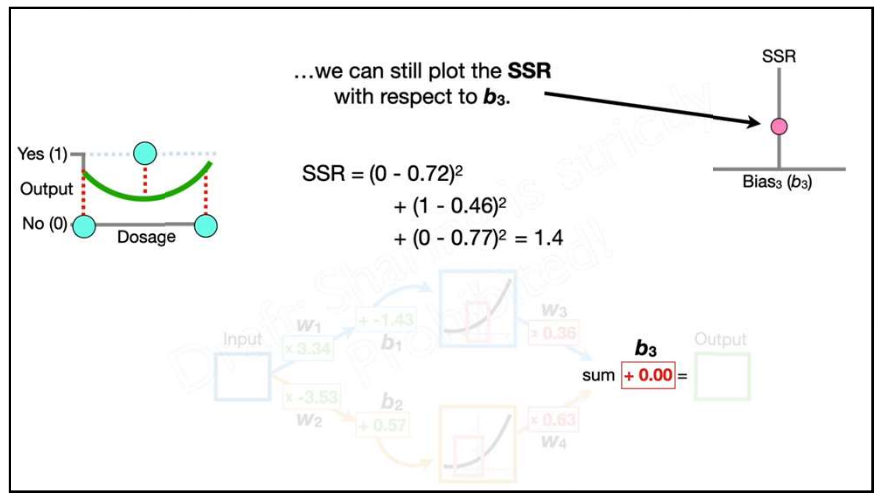

…and some of which are ***not* optimal**…

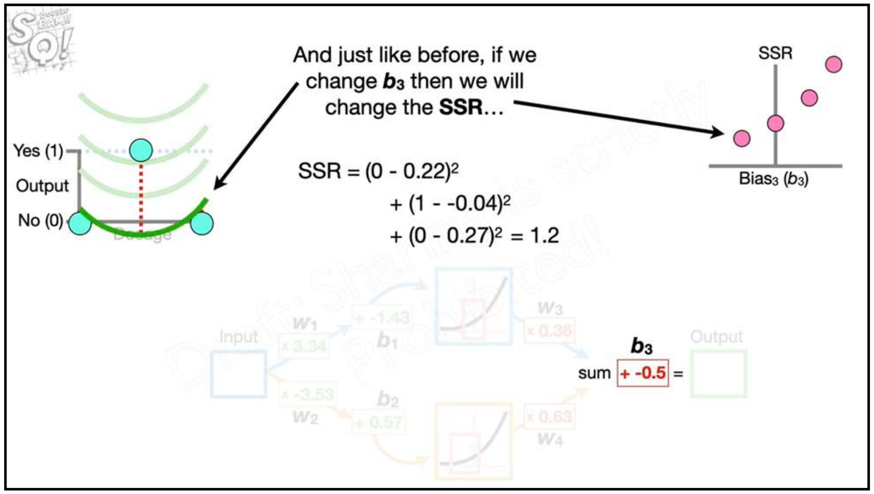

…we end up with this **green squiggle**.

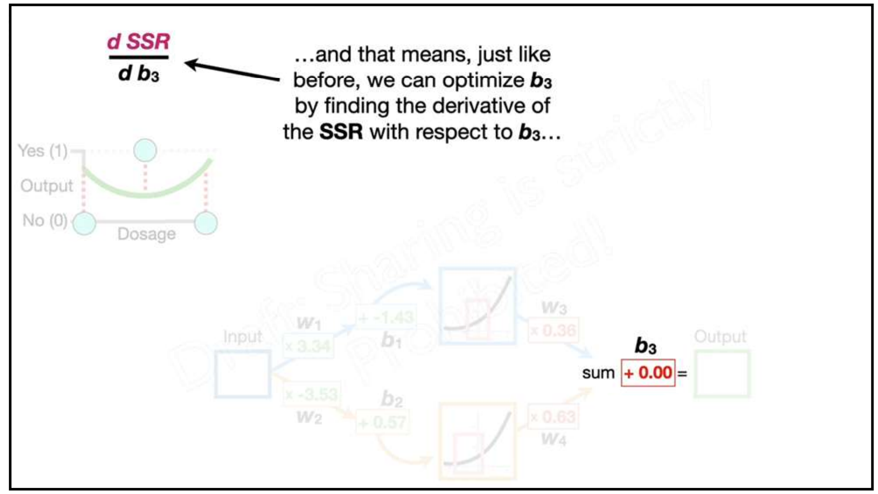

Now, just like before, we can quantify how well the green squiggle fits the data by calculating the Sum of the Squared Residuals (SSR)…
Yes (1)
Output
No (0)
Dosage
SSR = (0 - 0.72)²
+ (1 - 0.46)²
+ (0 - 0.77)²

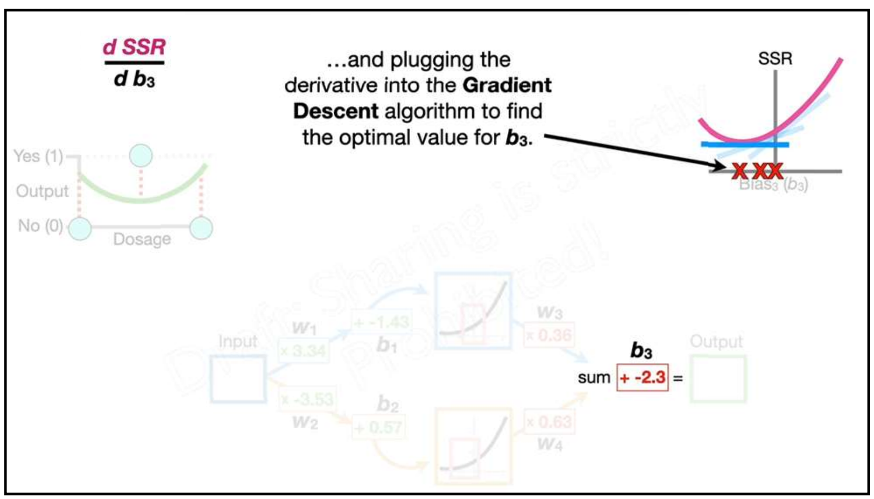

…and we get the SSR = 1.4.
Yes (1)
Output
No (0)
Dosage
SSR = (0 - 0.72)²
+ (1 - 0.46)²
+ (0 - 0.77)² = 1.4

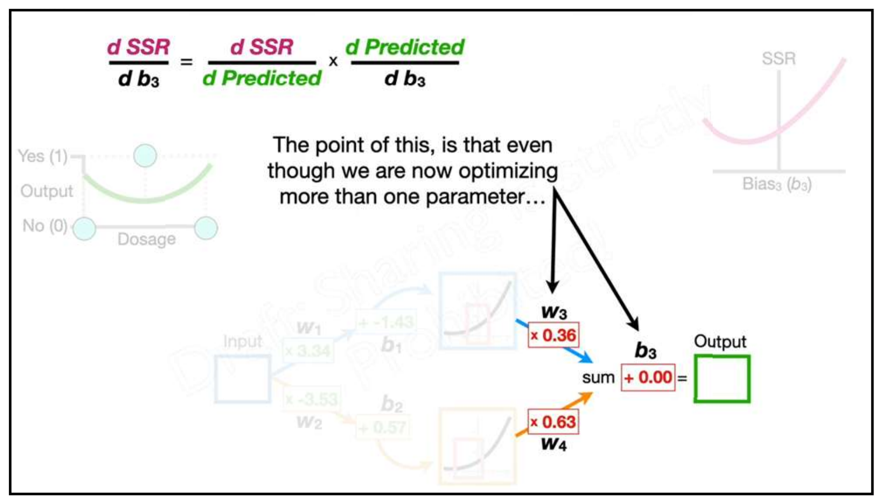

Now, even though we have not yet optimized w3 and w4…
Yes (1)
Output
No (0)
Dosage
SSR = (0 - 0.72)²
+ (1 - 0.46)²
+ (0 - 0.77)² = 1.4
w3
x 0.36
x 0.63
w4

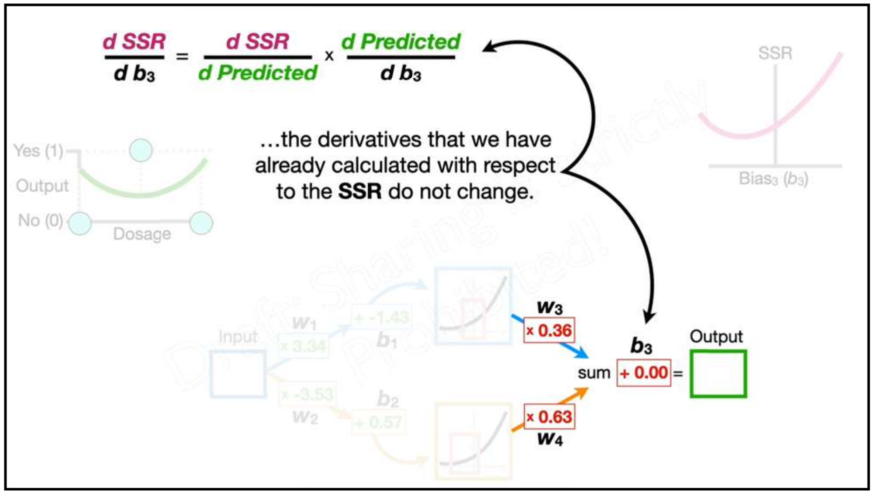

…we can still plot the SSR with respect to b3.
SSR
Bias3 (b3)
Yes (1)
Output
No (0)
Dosage
SSR = (0 - 0.72)²
+ (1 - 0.46)²
+ (0 - 0.77)² = 1.4
b3
sum + 0.00 =

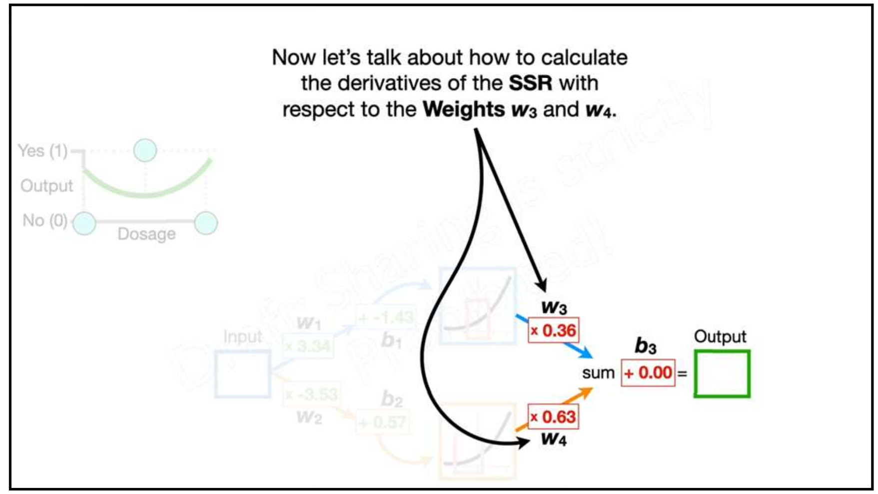

And just like before, if we change $b_3$ then we will change the **SSR**…
SSR = (0 - 0.22)² + (1 - -0.04)² + (0 - 0.27)² = 1.2
SSR
Bias₃ ($b_3$)
Yes (1)
Output
No (0)
Dosage
Input
$w_1$
x 3.34
+ -1.43
$b_1$
x -3.53
$w_2$
$b_2$
+ 0.57
$w_3$
x 0.36
x 0.63
$w_4$
$b_3$
sum + -0.5 =
Output

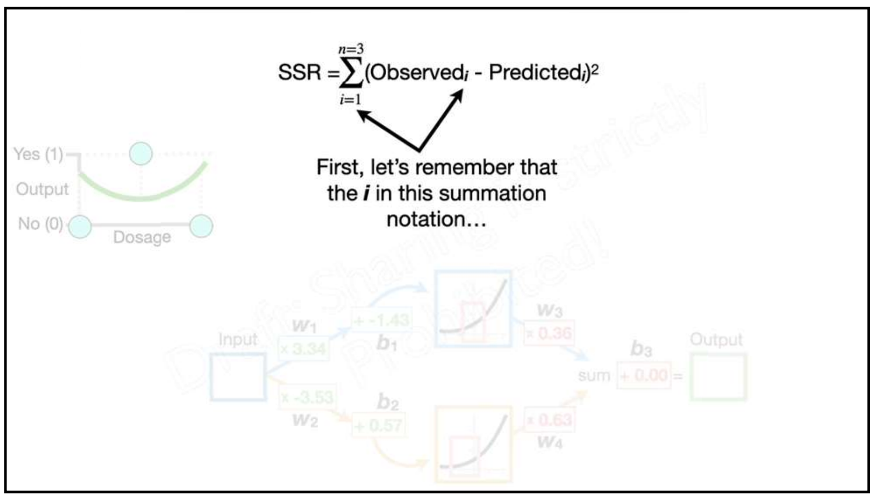

$\frac{d\ SSR}{d\ b_3}$
…and that means, just like before, we can optimize $b_3$ by finding the derivative of the **SSR** with respect to $b_3$…
Yes (1)
Output
No (0)
Dosage
Input
$b_3$
sum + 0.00 =
Output

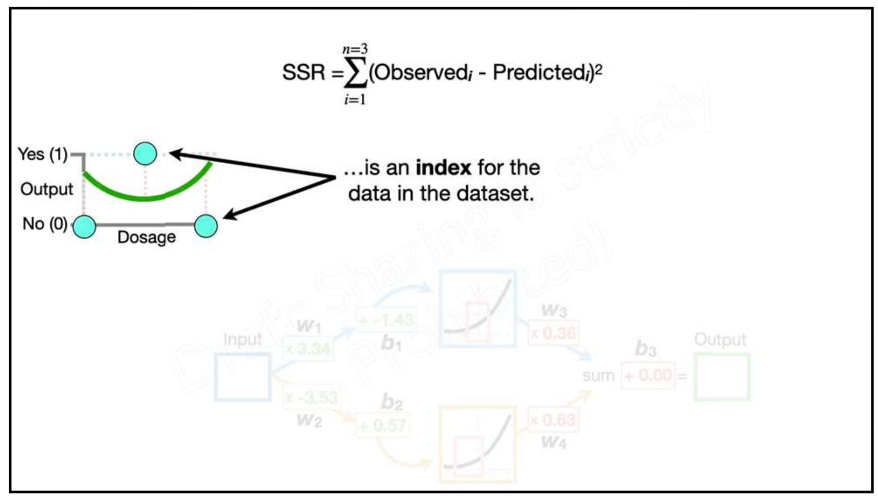

$\frac{d\ SSR}{d\ b_3}$
…and plugging the derivative into the **Gradient Descent** algorithm to find the optimal value for $b_3$.
SSR
Bias₃ ($b_3$)
Yes (1)
Output
No (0)
Dosage
Input
$b_3$
sum + -2.3 =
Output

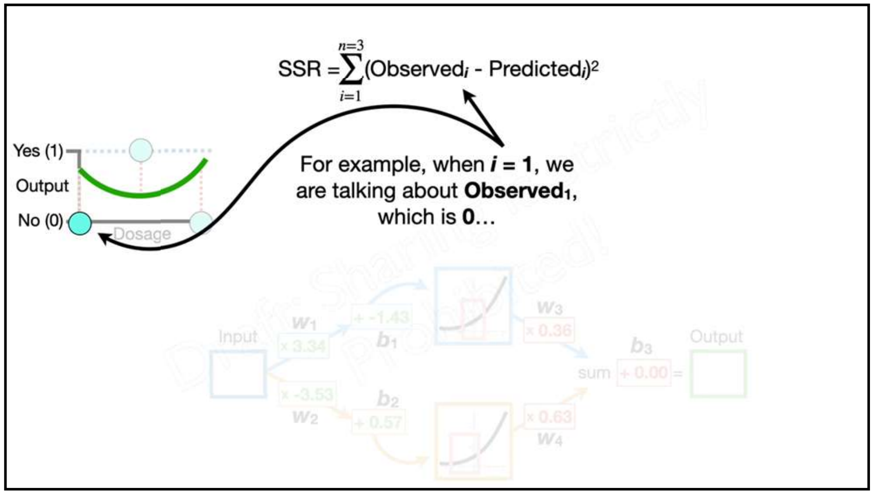

$\frac{d\ SSR}{d\ b_3} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ b_3}$
The point of this, is that even though we are now optimizing more than one parameter…
SSR
Bias₃ ($b_3$)
Yes (1)
Output
No (0)
Dosage
Input
$w_3$
x 0.36
$b_3$
sum + 0.00 =
Output
x 0.63
$w_4$

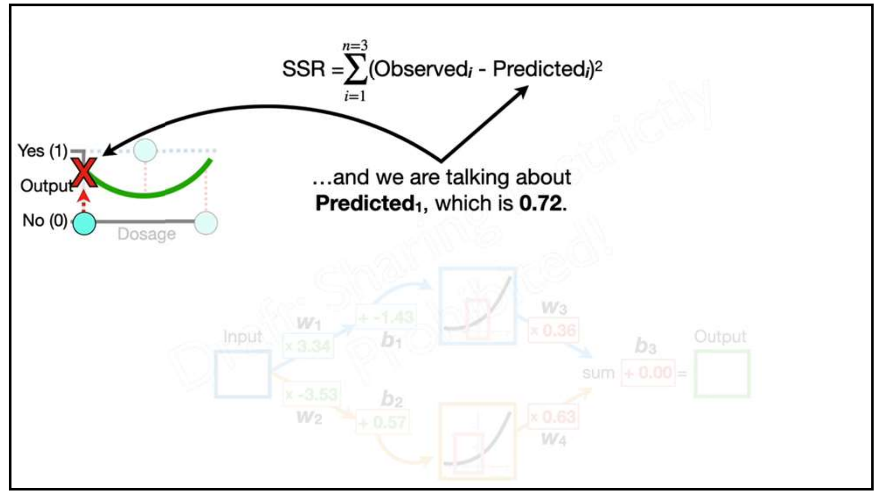

$\frac{d\ SSR}{d\ b_3} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ b_3}$
…the derivatives that we have already calculated with respect to the **SSR** do not change.
$w_3$
× 0.36
$b_3$
sum + 0.00 =
Output
× 0.63
$w_4$

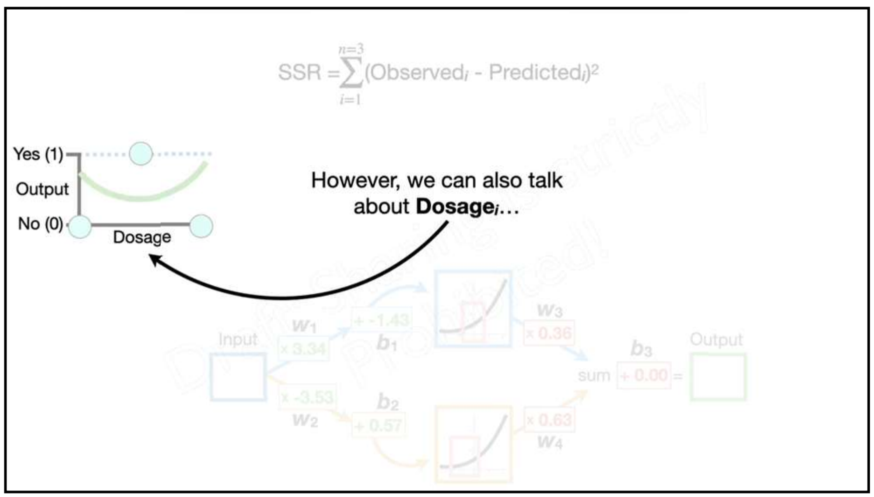

Now let's talk about how to calculate the derivatives of the **SSR** with respect to the **Weights** $w_3$ and $w_4$.
$w_3$
× 0.36
$b_3$
sum + 0.00 =
Output
× 0.63
$w_4$

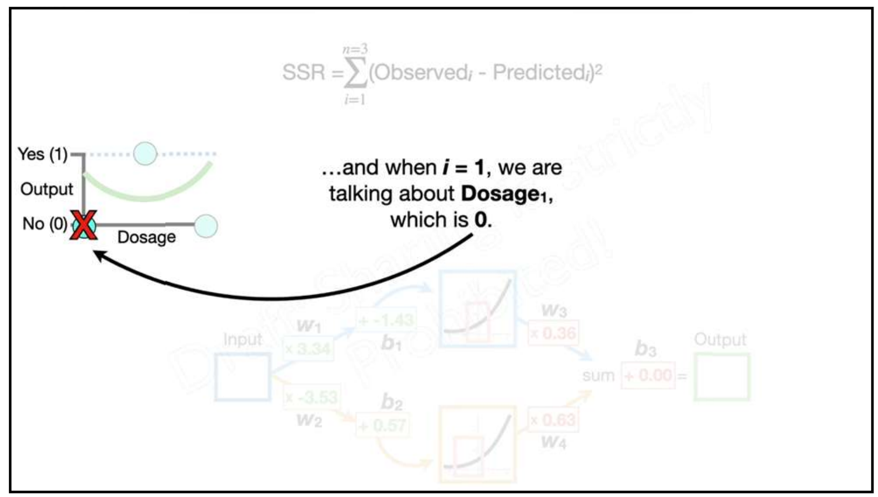

SSR = $\sum_{i=1}^{n=3}(\text{Observed}_i - \text{Predicted}_i)^2$
First, let's remember that the $i$ in this summation notation…

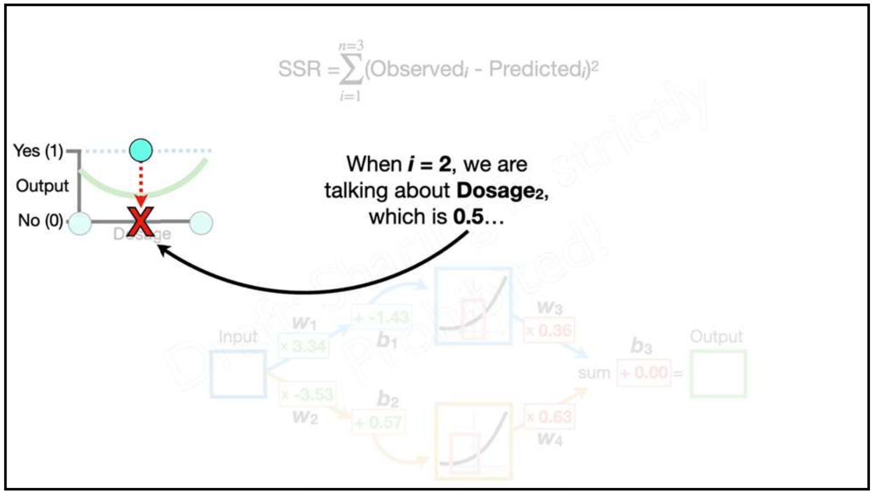

SSR = $\sum_{i=1}^{n=3}(\text{Observed}_i - \text{Predicted}_i)^2$
Yes (1)
Output
No (0)
Dosage
…is an **index** for the data in the dataset.

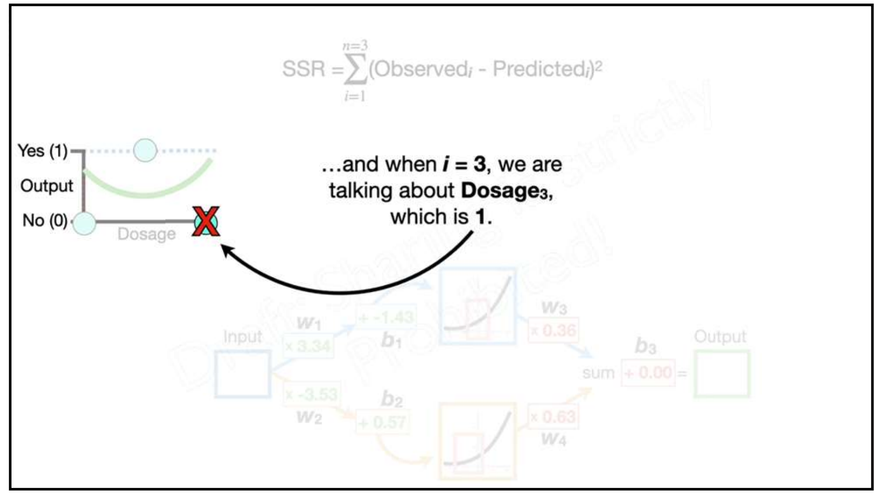

$SSR = \sum_{i=1}^{n=3} (Observed_i - Predicted_i)^2$
Yes (1)
Output
No (0)
Dosage
For example, when ***i* = 1**, we are talking about **Observed$_1$**, which is **0**…

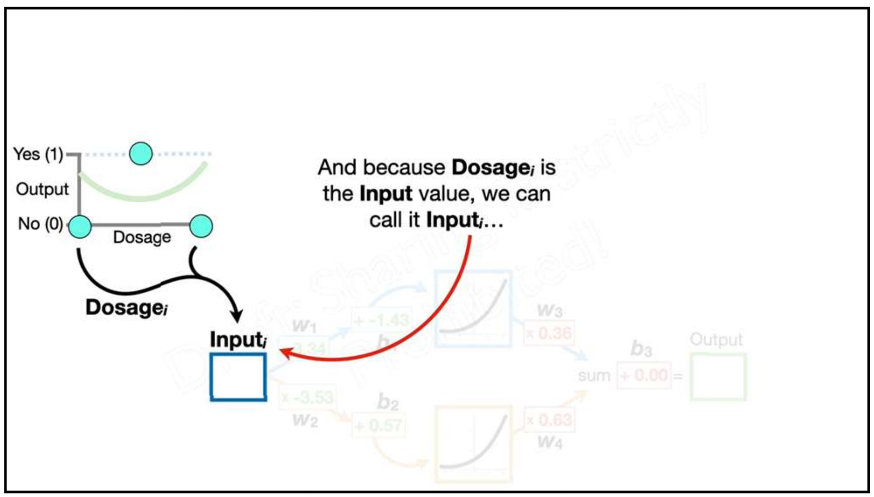

$SSR = \sum_{i=1}^{n=3} (Observed_i - Predicted_i)^2$
Yes (1)
Output
No (0)
Dosage
…and we are talking about **Predicted$_1$**, which is **0.72**.

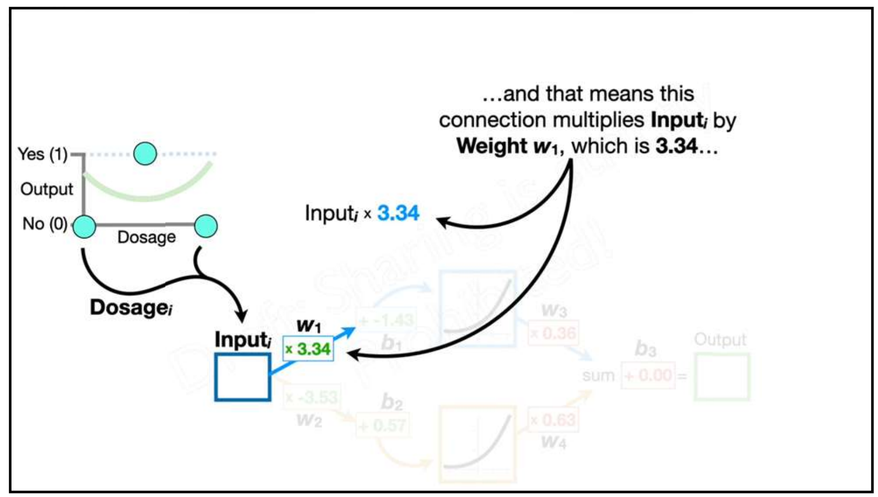

$SSR = \sum_{i=1}^{n=3} (Observed_i - Predicted_i)^2$
Yes (1)
Output
No (0)
Dosage
However, we can also talk about **Dosage*$_i$***…

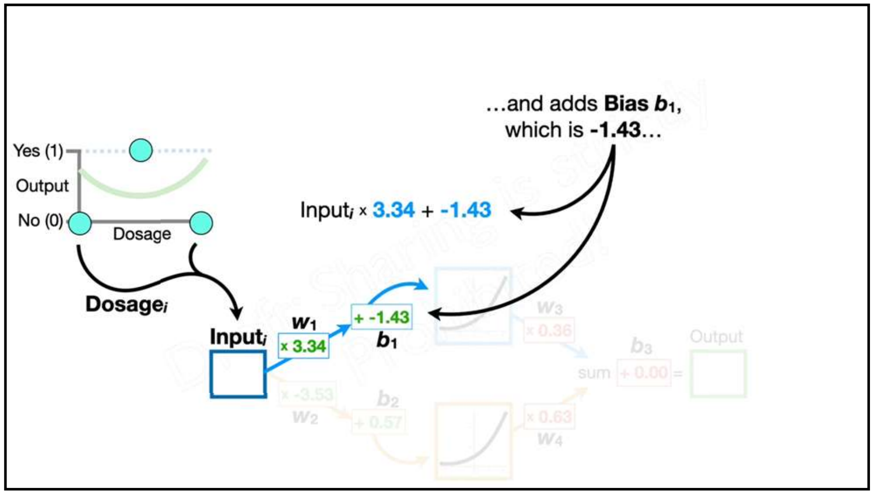

$SSR = \sum_{i=1}^{n=3} (Observed_i - Predicted_i)^2$
Yes (1)
Output
No (0)
Dosage
…and when ***i* = 1**, we are talking about **Dosage$_1$**, which is **0**.

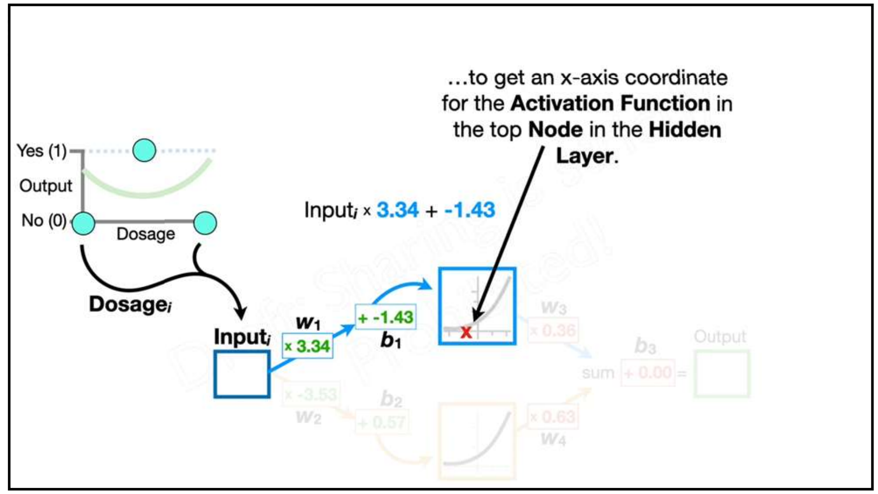

SSR = $\sum_{i=1}^{n=3}$(Observed$_i$ - Predicted$_i$)$^2$
Yes (1)
Output
No (0)
Dosage
When ***i* = 2**, we are talking about **Dosage$_2$**, which is **0.5**…

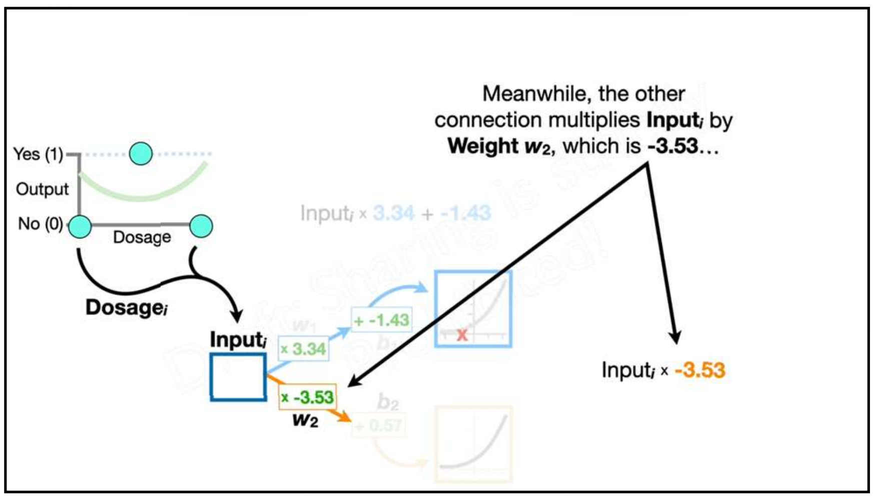

SSR = $\sum_{i=1}^{n=3}$(Observed$_i$ - Predicted$_i$)$^2$
Yes (1)
Output
No (0)
Dosage
…and when ***i* = 3**, we are talking about **Dosage$_3$**, which is **1**.

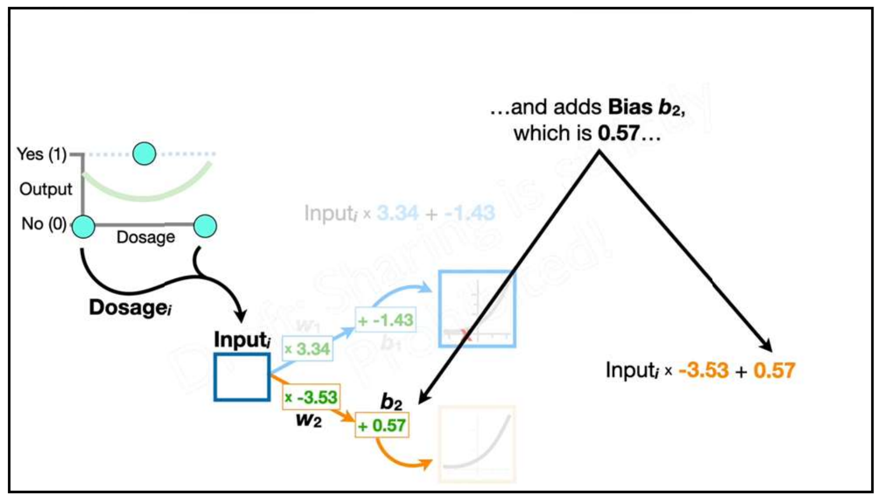

Yes (1)
Output
No (0)
Dosage
**Dosage$_i$**
**Input$_i$**
And because **Dosage$_i$** is the **Input** value, we can call it **Input$_i$**…

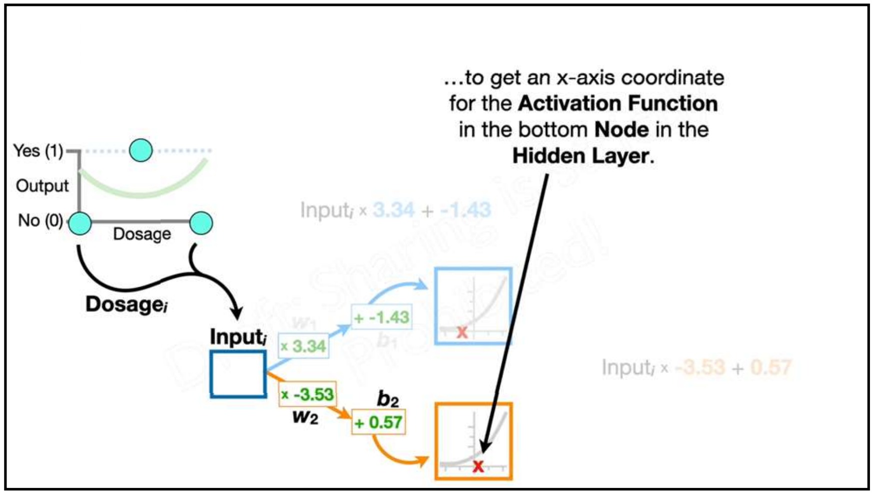

Yes (1)
Output
No (0)
Dosage
**Dosage$_i$**
**Input$_i$**
$w_1$
x 3.34
Input$_i$ × 3.34
…and that means this connection multiplies **Input$_i$** by **Weight $w_1$**, which is **3.34**…

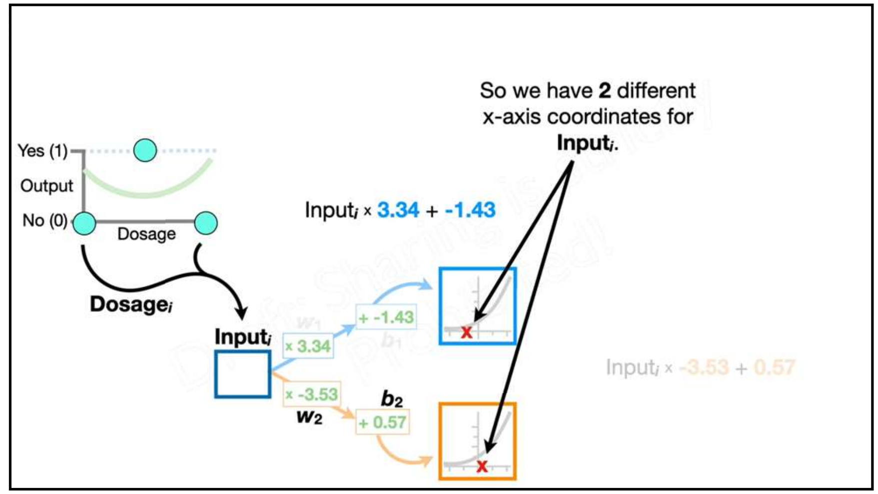

...and adds **Bias $b_1$**, which is **-1.43**...
Input$_i$ × 3.34 + -1.43
Yes (1)
Output
No (0)
Dosage
Dosage$_i$
Input$_i$
$w_1$
x 3.34
+ -1.43
$b_1$

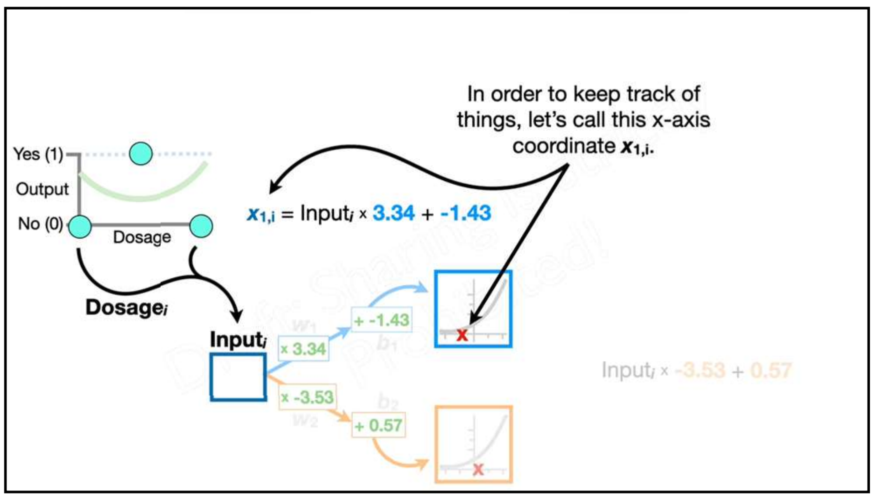

...to get an x-axis coordinate for the **Activation Function** in the top **Node** in the **Hidden Layer**.
Input$_i$ × 3.34 + -1.43
Yes (1)
Output
No (0)
Dosage
Dosage$_i$
Input$_i$
$w_1$
x 3.34
+ -1.43
$b_1$

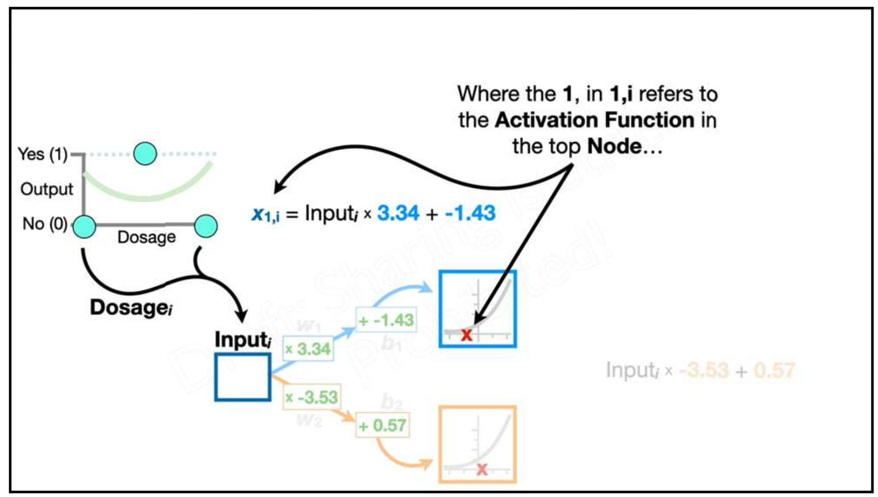

Meanwhile, the other connection multiplies **Input$_i$** by **Weight $w_2$**, which is **-3.53**...
Input$_i$ × -3.53
Yes (1)
Output
No (0)
Dosage
Dosage$_i$
Input$_i$
x -3.53
$w_2$

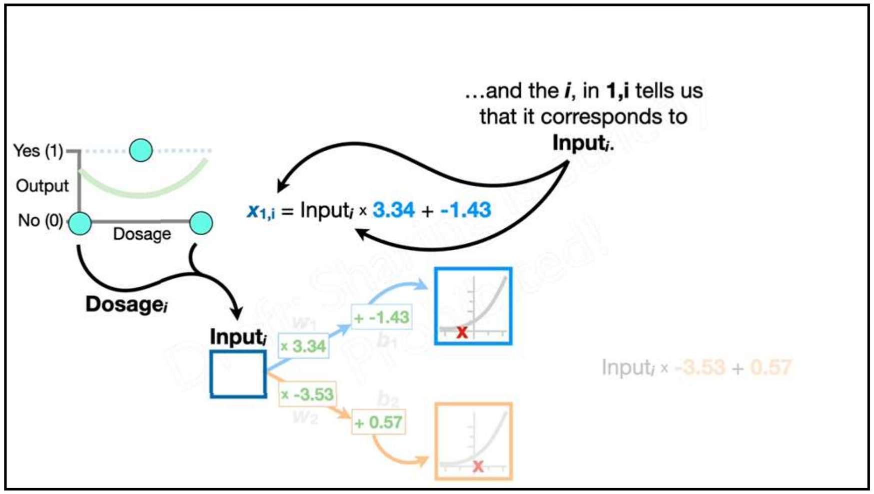

...and adds **Bias $b_2$**, which is **0.57**...
Input$_i$ × -3.53 + 0.57
Yes (1)
Output
No (0)
Dosage
Dosage$_i$
Input$_i$
x -3.53
$w_2$
$b_2$
+ 0.57

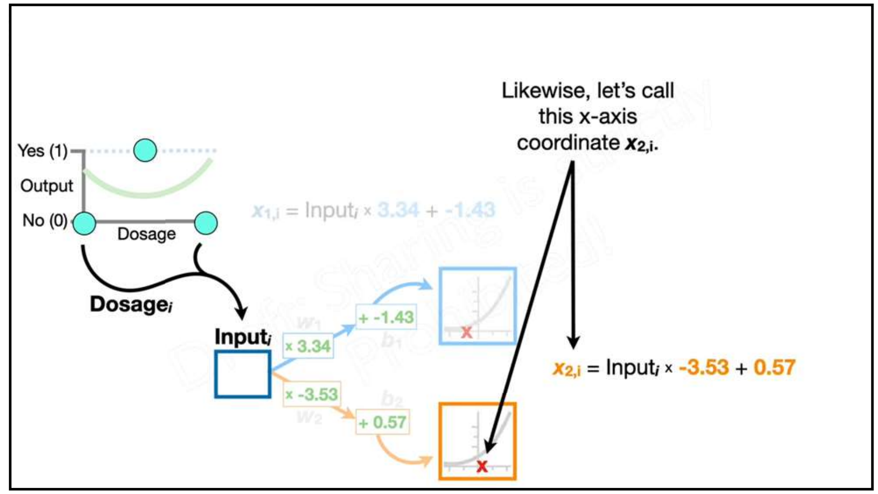

…to get an x-axis coordinate
for the **Activation Function**
in the bottom **Node** in the
**Hidden Layer**.
Yes (1)
Output
No (0)
Dosage
Dosage$_i$
Input$_i$
Input$_i$ × 3.34 + -1.43
Input$_i$ × -3.53 + 0.57
x 3.34
+ -1.43
x -3.53
+ 0.57
$w_1$
$b_1$
$w_2$
$b_2$

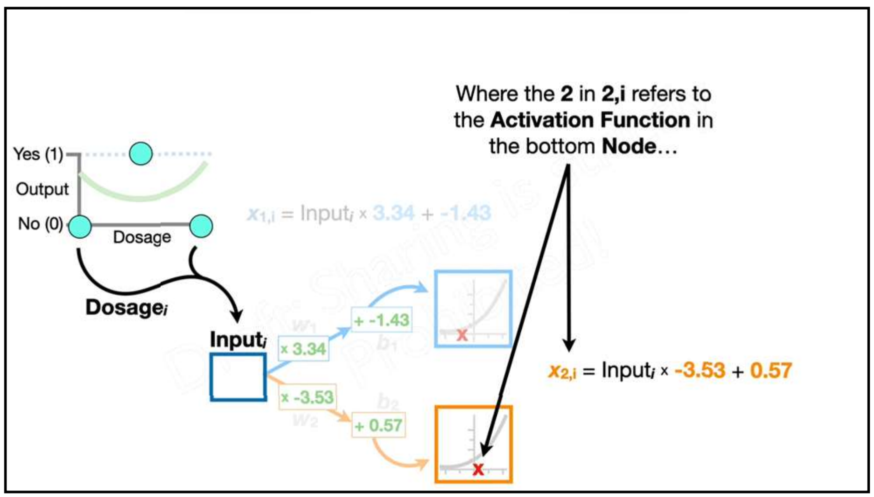

So we have **2** different
x-axis coordinates for
**Input$_i$**.
Yes (1)
Output
No (0)
Dosage
Dosage$_i$
Input$_i$
Input$_i$ × 3.34 + -1.43
Input$_i$ × -3.53 + 0.57
x 3.34
+ -1.43
x -3.53
+ 0.57
$w_1$
$b_1$
$w_2$
$b_2$

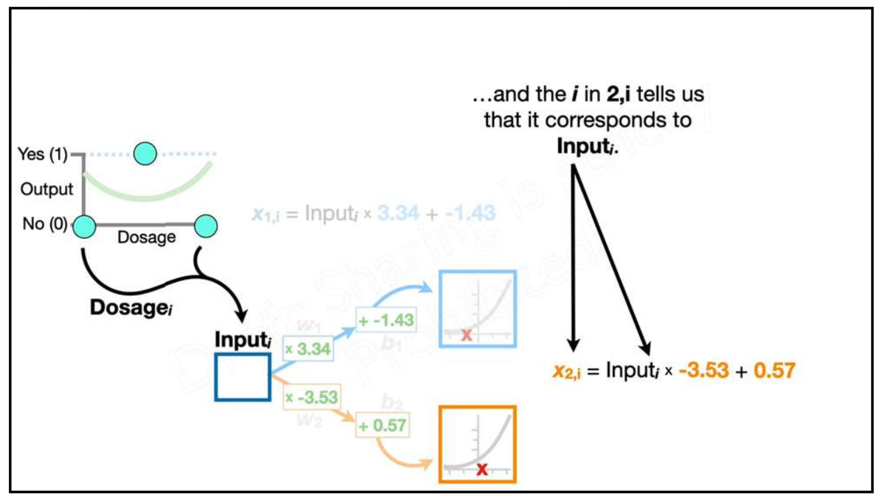

In order to keep track of
things, let's call this x-axis
coordinate $x_{1,i}$.
$x_{1,i}$ = Input$_i$ × 3.34 + -1.43
Yes (1)
Output
No (0)
Dosage
Dosage$_i$
Input$_i$
Input$_i$ × -3.53 + 0.57
x 3.34
+ -1.43
x -3.53
+ 0.57
$w_1$
$b_1$
$w_2$
$b_2$

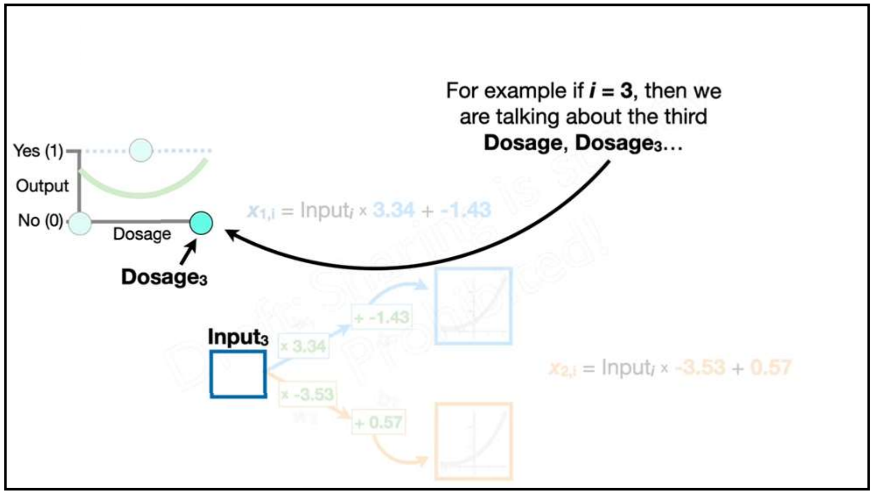

Where the **1**, in **1,i** refers to
the **Activation Function** in
the top **Node**…
$x_{1,i}$ = Input$_i$ × 3.34 + -1.43
Yes (1)
Output
No (0)
Dosage
Dosage$_i$
Input$_i$
Input$_i$ × -3.53 + 0.57
x 3.34
+ -1.43
x -3.53
+ 0.57
$w_1$
$b_1$
$w_2$
$b_2$

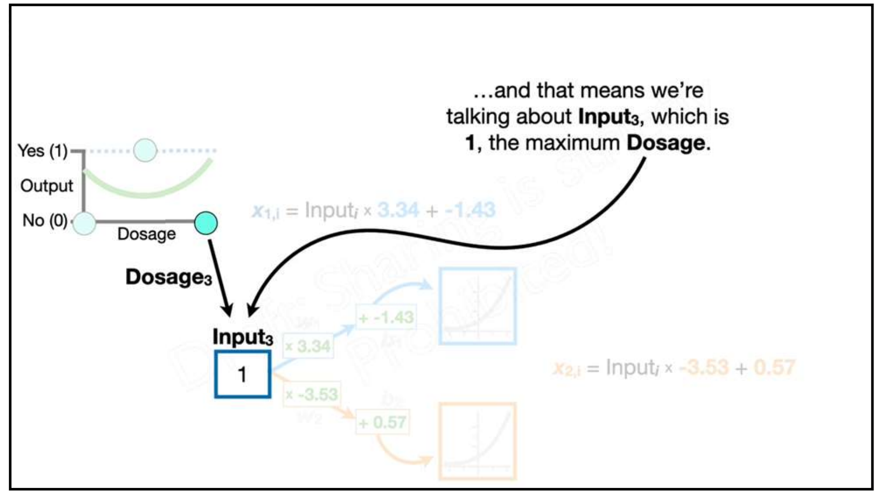

…and the *i*, in **1,i** tells us that it corresponds to **Input*i***.
Yes (1)
Output
No (0)
Dosage
**Dosage*i***
**Input*i***
$x_{1,i}$ = Input$_i$ × 3.34 + -1.43
x 3.34
+ -1.43
x -3.53
+ 0.57
Input$_i$ × -3.53 + 0.57

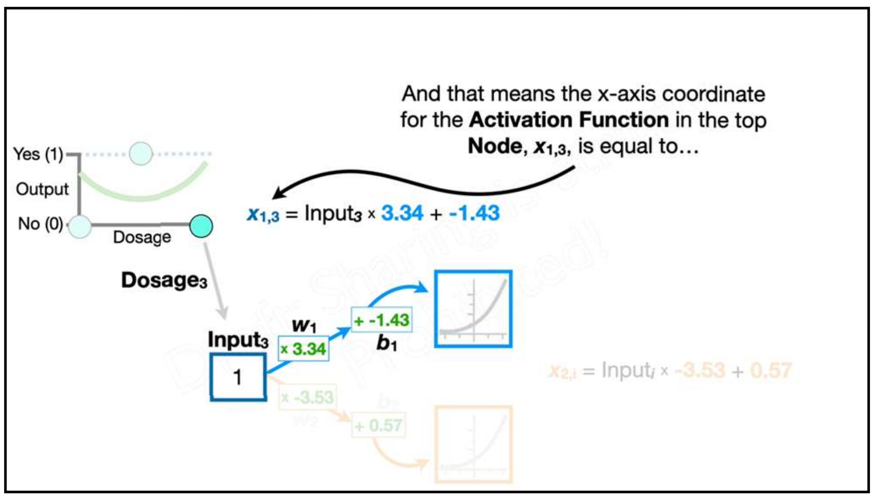

Likewise, let's call this x-axis coordinate $x_{2,i}$.
Yes (1)
Output
No (0)
Dosage
**Dosage*i***
**Input*i***
$x_{1,i}$ = Input$_i$ × 3.34 + -1.43
x 3.34
+ -1.43
x -3.53
+ 0.57
$x_{2,i}$ = Input$_i$ × -3.53 + 0.57

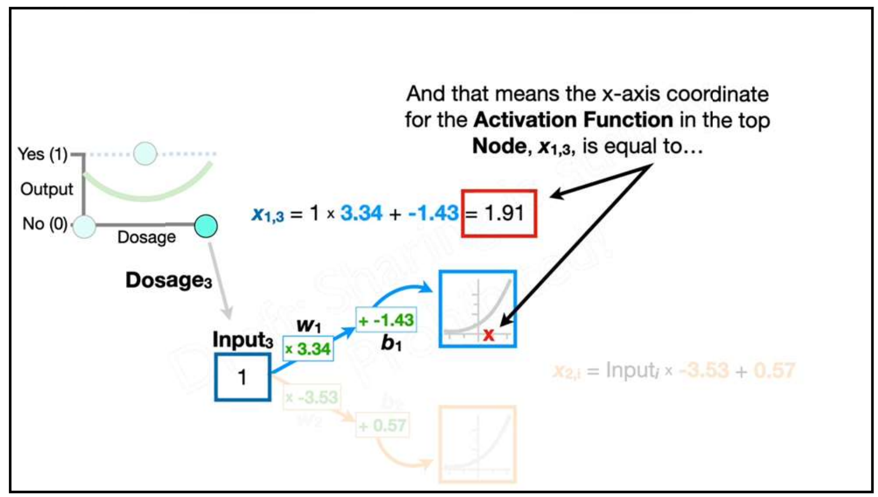

Where the **2** in **2,i** refers to the **Activation Function** in the bottom **Node**…
Yes (1)
Output
No (0)
Dosage
**Dosage*i***
**Input*i***
$x_{1,i}$ = Input$_i$ × 3.34 + -1.43
x 3.34
+ -1.43
x -3.53
+ 0.57
$x_{2,i}$ = Input$_i$ × -3.53 + 0.57

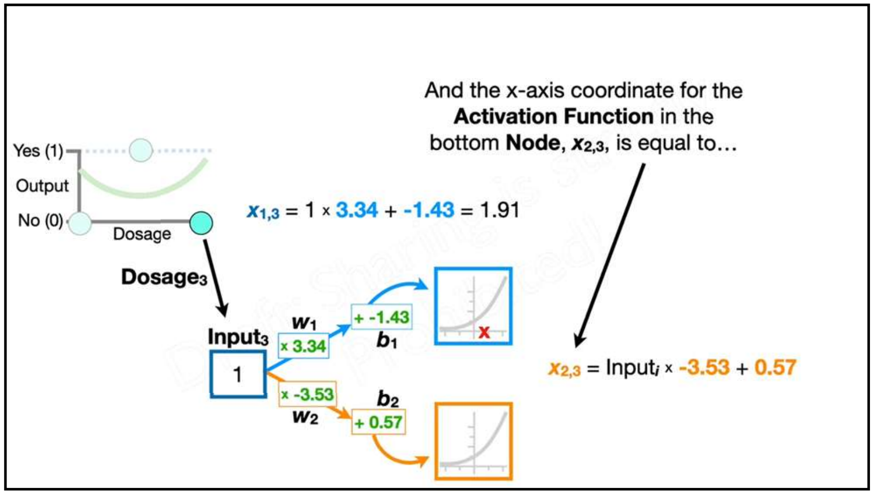

…and the *i* in **2,i** tells us that it corresponds to **Input*i***.
Yes (1)
Output
No (0)
Dosage
**Dosage*i***
**Input*i***
$x_{1,i}$ = Input$_i$ × 3.34 + -1.43
x 3.34
+ -1.43
x -3.53
+ 0.57
$x_{2,i}$ = Input$_i$ × -3.53 + 0.57

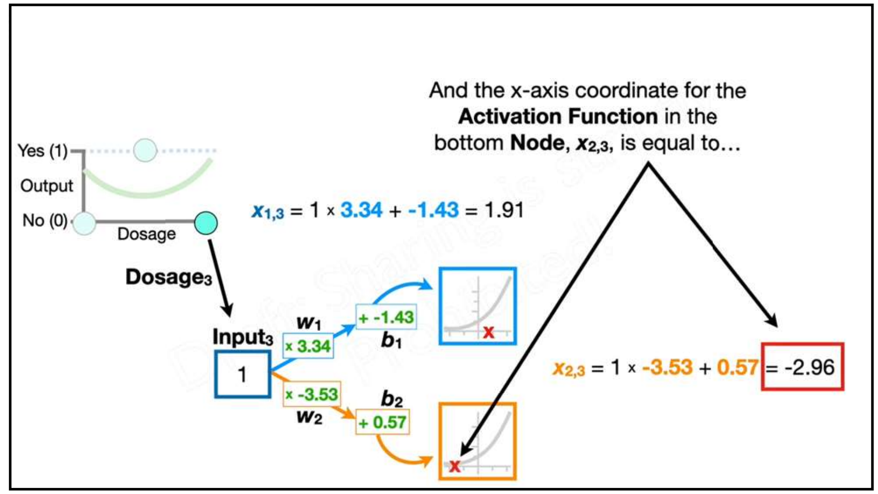

For example if *i* = 3, then we are talking about the third **Dosage**, **Dosage₃**…
Yes (1)
Output
No (0)
Dosage
Dosage3
Input3

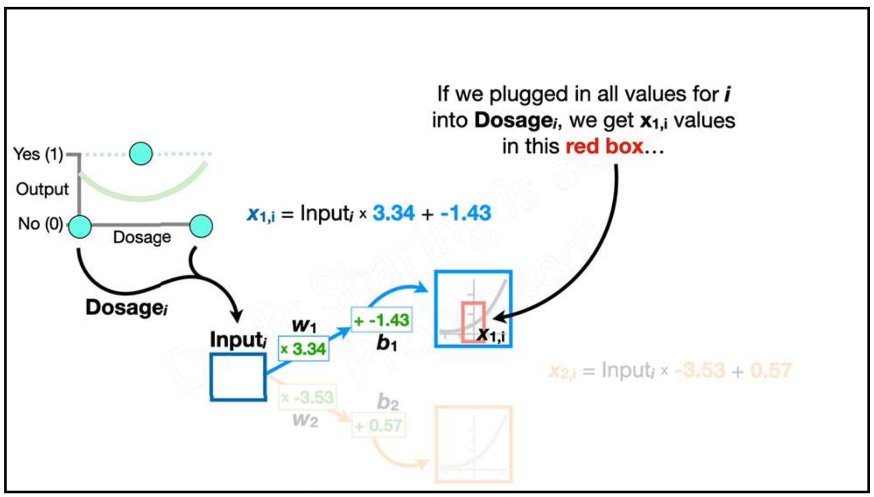

…and that means we're talking about **Input₃**, which is **1**, the maximum **Dosage**.
Yes (1)
Output
No (0)
Dosage
Dosage3
Input3
1

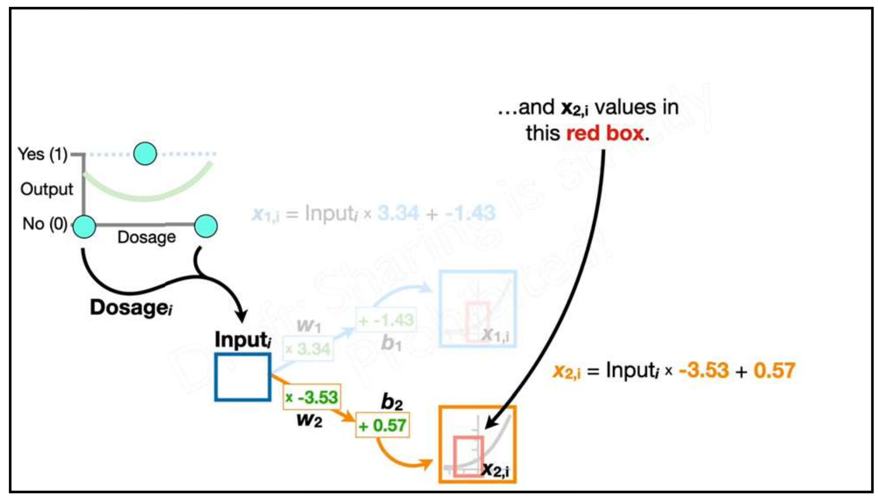

And that means the x-axis coordinate for the **Activation Function** in the top **Node**, $x_{1,3}$, is equal to…
Yes (1)
Output
No (0)
Dosage
$x_{1,3}$ = Input3 × 3.34 + -1.43
Dosage3
Input3
1
w1
x 3.34
+ -1.43
b1

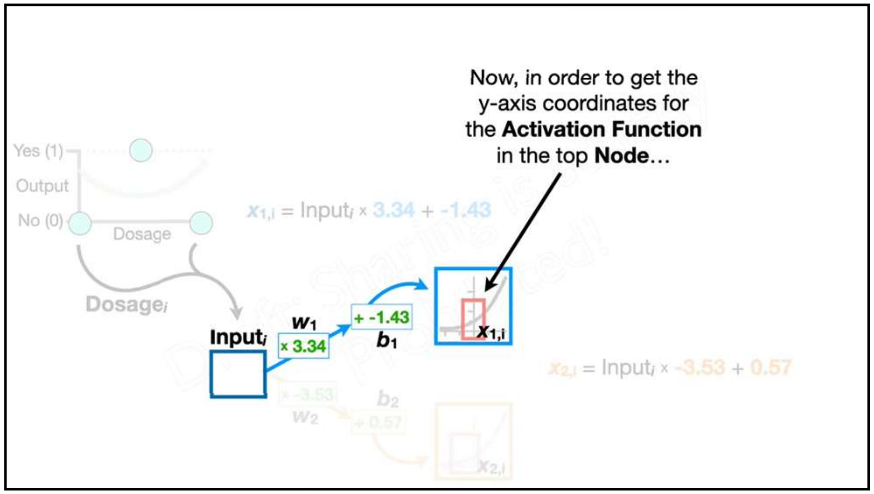

And that means the x-axis coordinate for the **Activation Function** in the top **Node**, $x_{1,3}$, is equal to…
Yes (1)
Output
No (0)
Dosage
$x_{1,3}$ = 1 × 3.34 + -1.43 = 1.91
Dosage3
Input3
1
w1
x 3.34
+ -1.43
b1


And the x-axis coordinate for the **Activation Function** in the bottom **Node**, $x_{2,3}$, is equal to...
Yes (1)
Output
No (0)
Dosage
Dosage$_3$
$x_{1,3} = 1 \times 3.34 + -1.43 = 1.91$
Input$_3$
1
$w_1$
x 3.34
+ -1.43
$b_1$
x -3.53
$w_2$
+ 0.57
$b_2$
$x_{2,3}$ = Input$_i$ × -3.53 + 0.57

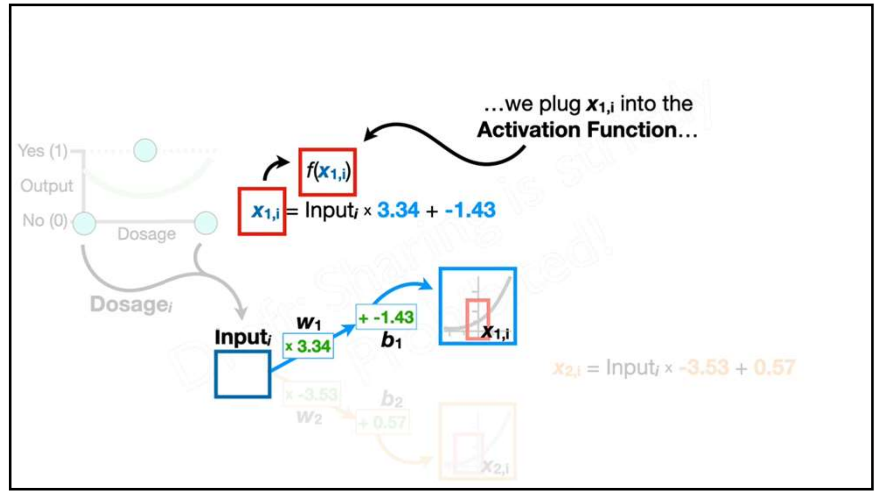

And the x-axis coordinate for the **Activation Function** in the bottom **Node**, $x_{2,3}$, is equal to...
Yes (1)
Output
No (0)
Dosage
Dosage$_3$
$x_{1,3} = 1 \times 3.34 + -1.43 = 1.91$
Input$_3$
1
$w_1$
x 3.34
+ -1.43
$b_1$
x -3.53
$w_2$
+ 0.57
$b_2$
$x_{2,3} = 1 \times -3.53 + 0.57 = -2.96$


If we plugged in all values for ***i*** into **Dosage*$_i$***, we get **$x_{1,i}$** values in this **red box**...
Yes (1)
Output
No (0)
Dosage
Dosage$_i$
$x_{1,i}$ = Input$_i$ × 3.34 + -1.43
Input$_i$
$w_1$
x 3.34
+ -1.43
$b_1$
$x_{1,i}$
$x_{2,i}$ = Input$_i$ × -3.53 + 0.57
x -3.53
$w_2$
+ 0.57
$b_2$

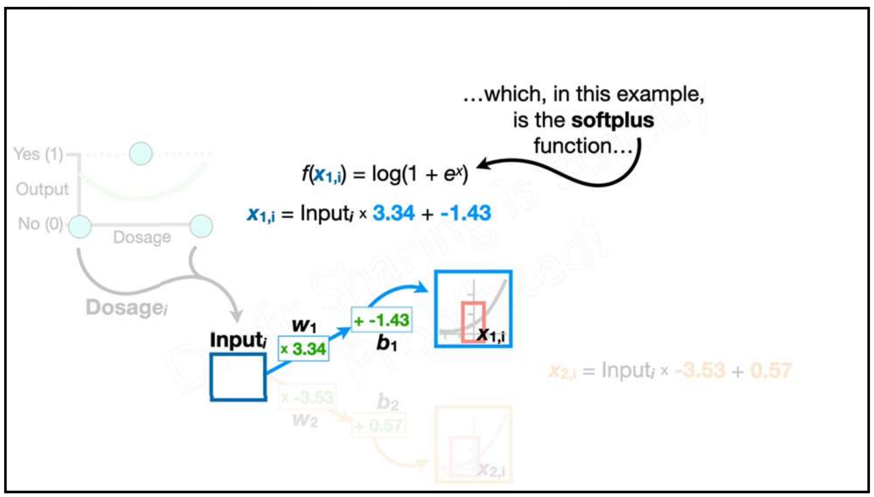

...and **$x_{2,i}$** values in this **red box**.
Yes (1)
Output
No (0)
Dosage
Dosage$_i$
$x_{1,i}$ = Input$_i$ × 3.34 + -1.43
Input$_i$
$w_1$
x 3.34
+ -1.43
$b_1$
$x_{1,i}$
$x_{2,i}$ = Input$_i$ × -3.53 + 0.57
x -3.53
$w_2$
+ 0.57
$b_2$
$x_{2,i}$



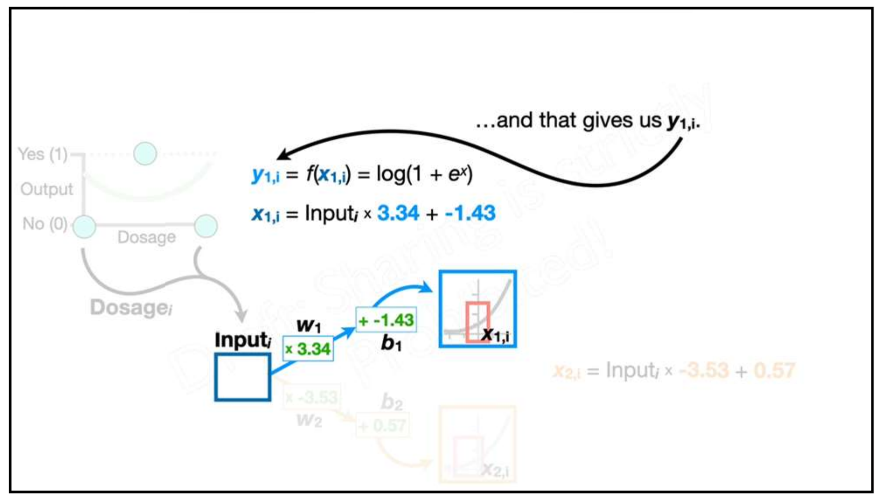
Now, in order to get the y-axis coordinates for the **Activation Function** in the top **Node**…
Yes (1)
Output
No (0)
Dosage
$x_{1,i}$ = Input$_i$ × 3.34 + -1.43
Dosage$_i$
Input$_i$
$w_1$
× 3.34
+ -1.43
$b_1$
$x_{1,i}$
$x_{2,i}$ = Input$_i$ × -3.53 + 0.57
× -3.53
$w_2$
$b_2$
+ 0.57
$x_{2,i}$

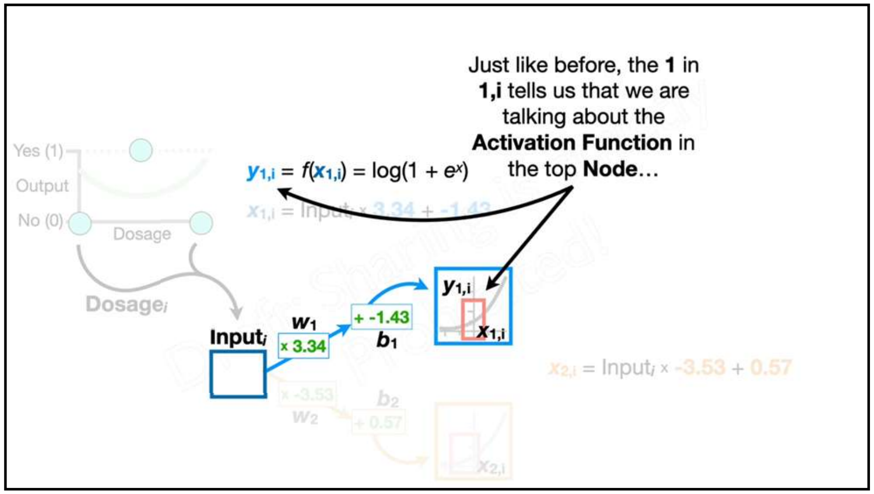
…we plug $x_{1,i}$ into the **Activation Function**…
$f(x_{1,i})$
$x_{1,i}$ = Input$_i$ × 3.34 + -1.43
Yes (1)
Output
No (0)
Dosage
Dosage$_i$
Input$_i$
$w_1$
× 3.34
+ -1.43
$b_1$
$x_{1,i}$
$x_{2,i}$ = Input$_i$ × -3.53 + 0.57
× -3.53
$w_2$
$b_2$
+ 0.57
$x_{2,i}$

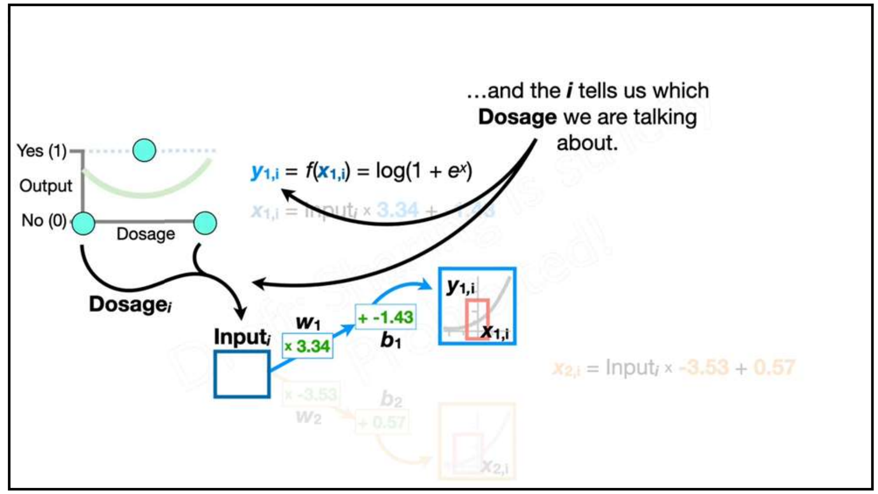



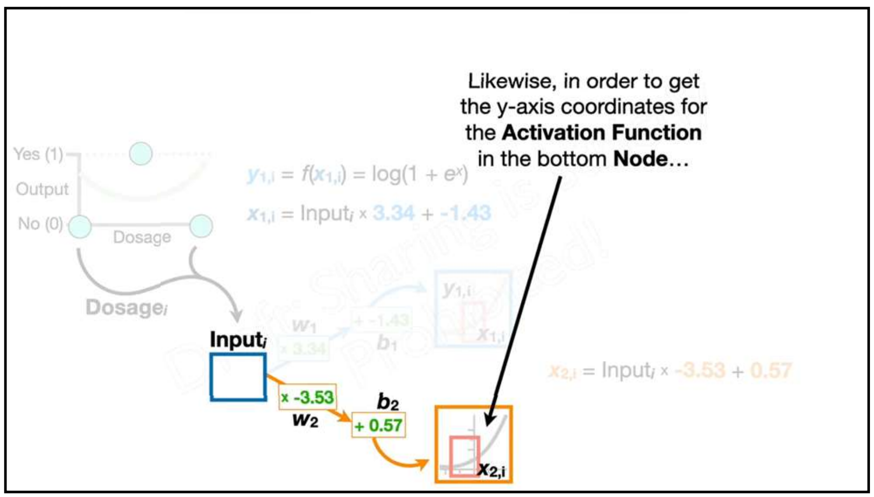
…which, in this example, is the **softplus** function…
$f(x_{1,i}) = \log(1 + e^x)$
$x_{1,i}$ = Input$_i$ × 3.34 + -1.43
Yes (1)
Output
No (0)
Dosage
Dosage$_i$
Input$_i$
$w_1$
× 3.34
+ -1.43
$b_1$
$x_{1,i}$
$x_{2,i}$ = Input$_i$ × -3.53 + 0.57
× -3.53
$w_2$
$b_2$
+ 0.57
$x_{2,i}$

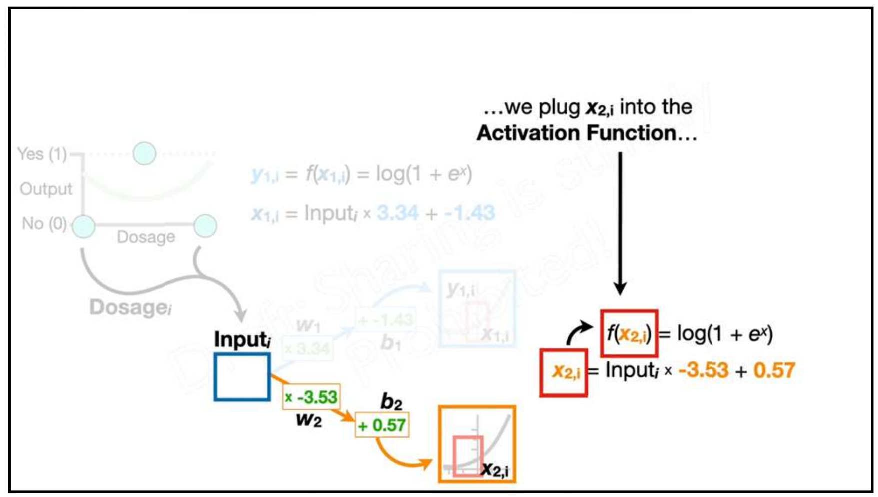



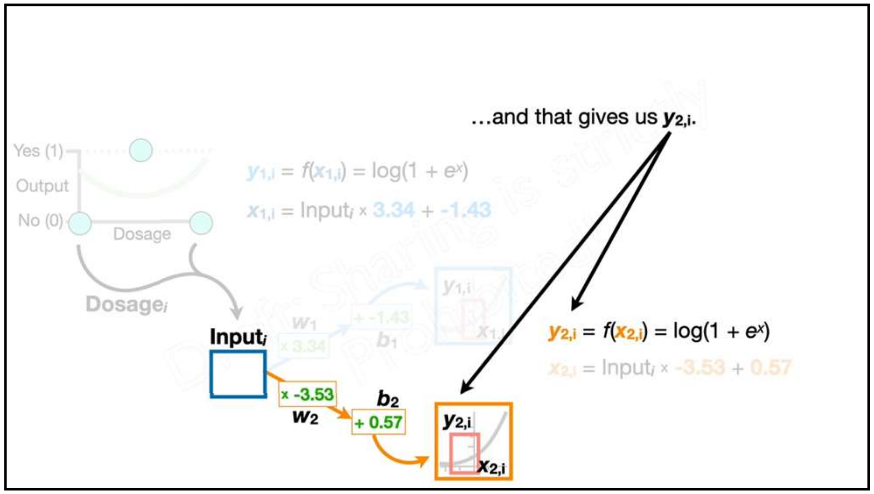
…and that gives us $y_{1,i}$.
$y_{1,i} = f(x_{1,i}) = \log(1 + e^x)$
$x_{1,i}$ = Input$_i$ × 3.34 + -1.43
Yes (1)
Output
No (0)
Dosage
Dosage$_i$
Input$_i$
$w_1$
× 3.34
+ -1.43
$b_1$
$x_{1,i}$
$x_{2,i}$ = Input$_i$ × -3.53 + 0.57
× -3.53
$w_2$
$b_2$
+ 0.57
$x_{2,i}$

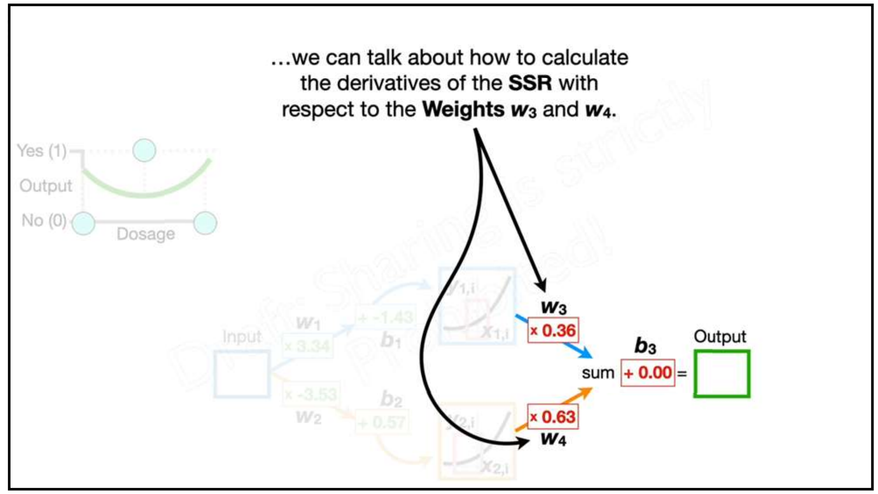

Just like before, the **1** in **1,i** tells us that we are talking about the **Activation Function** in the top **Node**…
$y_{1,i} = f(x_{1,i}) = \log(1 + e^x)$
Yes (1)
Output
No (0)
Dosage
Dosage$_i$
Input$_i$
$w_1$
× 3.34
+ -1.43
$b_1$
$y_{1,i}$
$x_{1,i}$
$x_{2,i}$ = Input$_i$ × -3.53 + 0.57

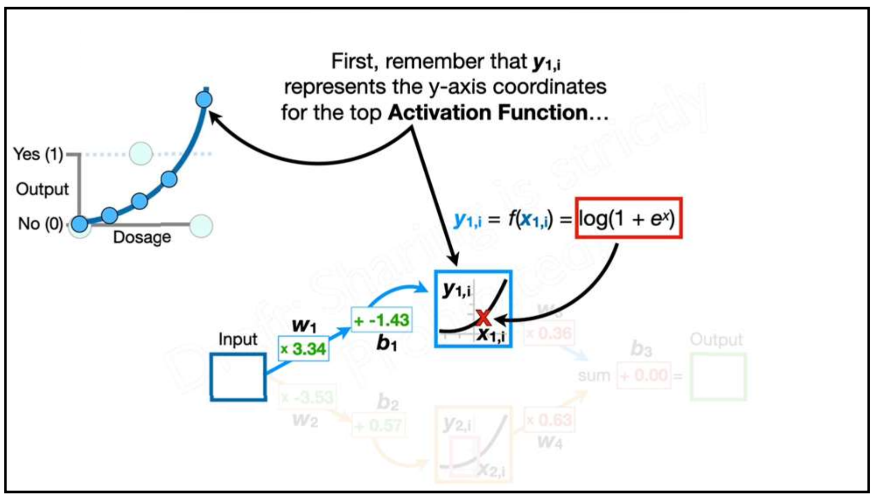

…and the ***i*** tells us which **Dosage** we are talking about.
$y_{1,i} = f(x_{1,i}) = \log(1 + e^x)$
Yes (1)
Output
No (0)
Dosage
Dosage$_i$
Input$_i$
$w_1$
× 3.34
+ -1.43
$b_1$
$y_{1,i}$
$x_{1,i}$
$x_{2,i}$ = Input$_i$ × -3.53 + 0.57

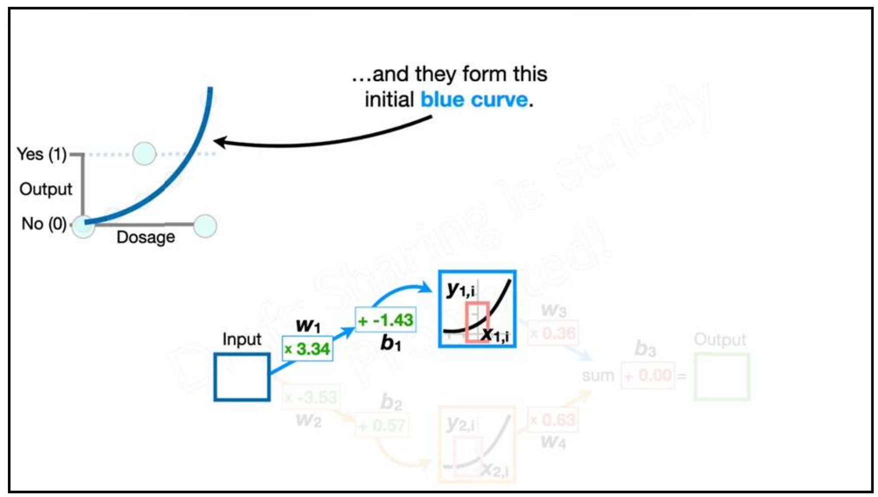

Likewise, in order to get the y-axis coordinates for the **Activation Function** in the bottom **Node**…
$y_{1,i} = f(x_{1,i}) = \log(1 + e^x)$
$x_{1,i}$ = Input$_i$ × 3.34 + -1.43
Dosage$_i$
Input$_i$
× -3.53
$w_2$
$b_2$
+ 0.57
$x_{2,i}$
$x_{2,i}$ = Input$_i$ × -3.53 + 0.57

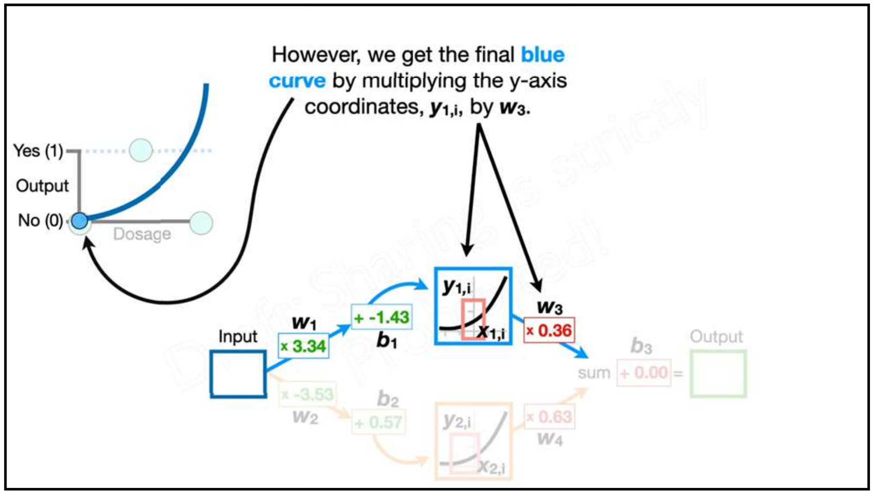

…we plug $x_{2,i}$ into the **Activation Function**…
$f(x_{2,i}) = \log(1 + e^x)$
$x_{2,i}$ = Input$_i$ × -3.53 + 0.57
$y_{1,i} = f(x_{1,i}) = \log(1 + e^x)$
$x_{1,i}$ = Input$_i$ × 3.34 + -1.43
Input$_i$
× -3.53
$w_2$
$b_2$
+ 0.57
$x_{2,i}$

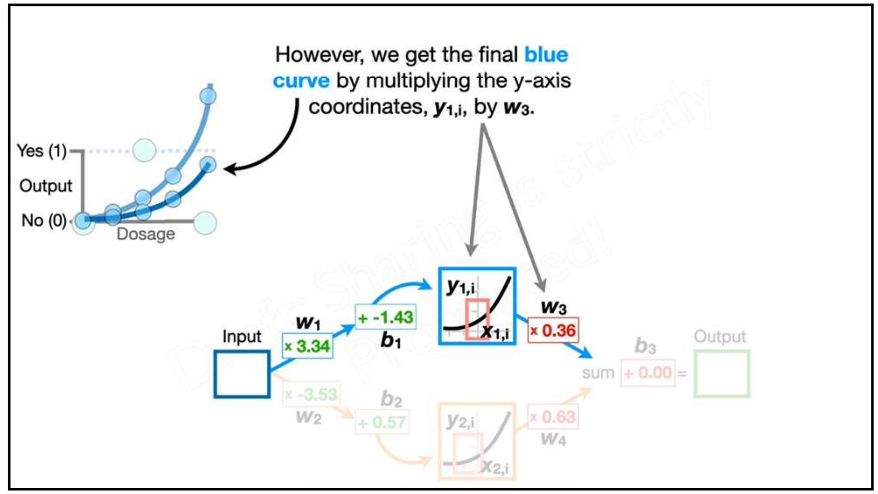

…and that gives us $y_{2,i}$.
$y_{1,i} = f(x_{1,i}) = \log(1 + e^x)$
$x_{1,i} = \text{Input}_i \times 3.34 + -1.43$
$y_{2,i} = f(x_{2,i}) = \log(1 + e^x)$
$x_{2,i} = \text{Input}_i \times -3.53 + 0.57$
Yes (1)
Output
No (0)
Dosage
$\text{Dosage}_i$
$\text{Input}_i$
$w_2$
$b_2$
x -3.53
+ 0.57

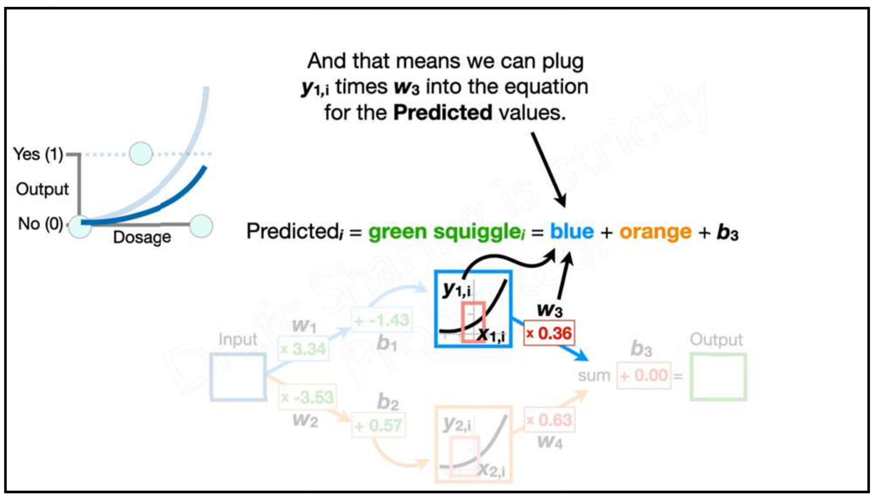

…we can talk about how to calculate the derivatives of the **SSR** with respect to the **Weights** $w_3$ and $w_4$.
$w_3$
x 0.36
$w_4$
x 0.63
$b_3$
sum + 0.00 =
Output

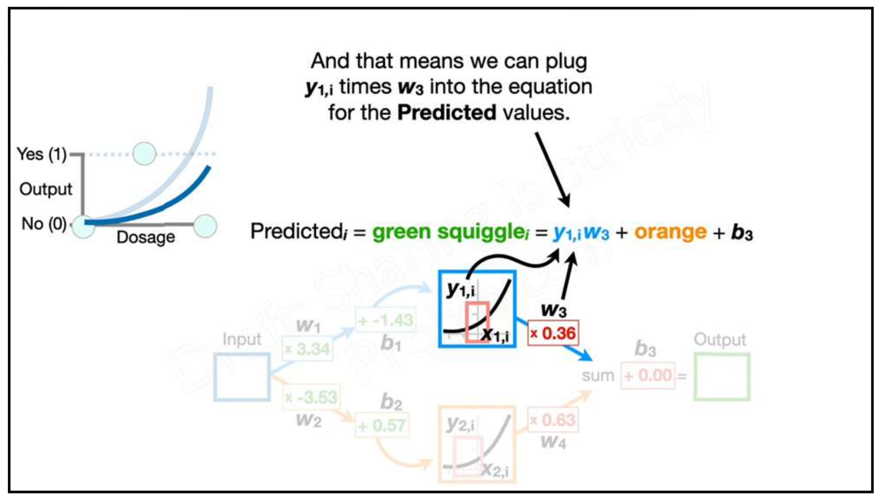

First, remember that $y_{1,i}$ represents the y-axis coordinates for the top **Activation Function**…
$y_{1,i} = f(x_{1,i}) = \log(1 + e^x)$
Yes (1)
Output
No (0)
Dosage
Input
$w_1$
x 3.34
+ -1.43
$b_1$

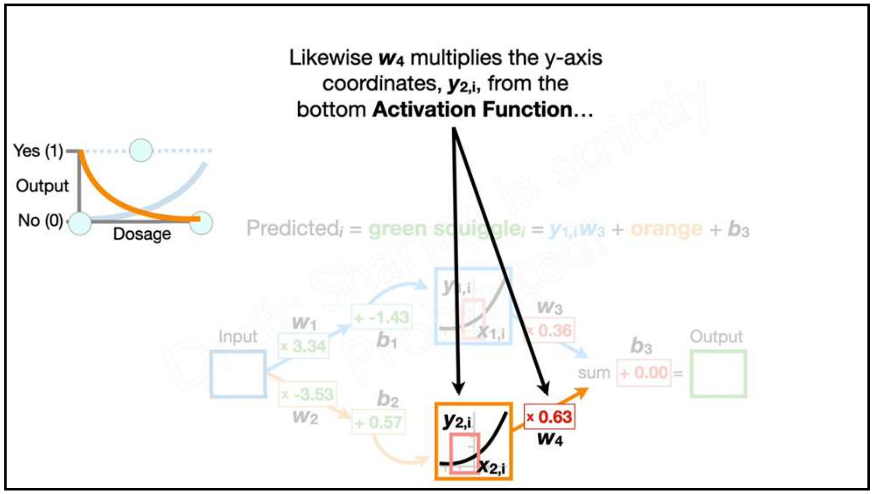

…and they form this initial **blue curve**.
Yes (1)
Output
No (0)
Dosage
Input
$w_1$
x 3.34
+ -1.43
$b_1$

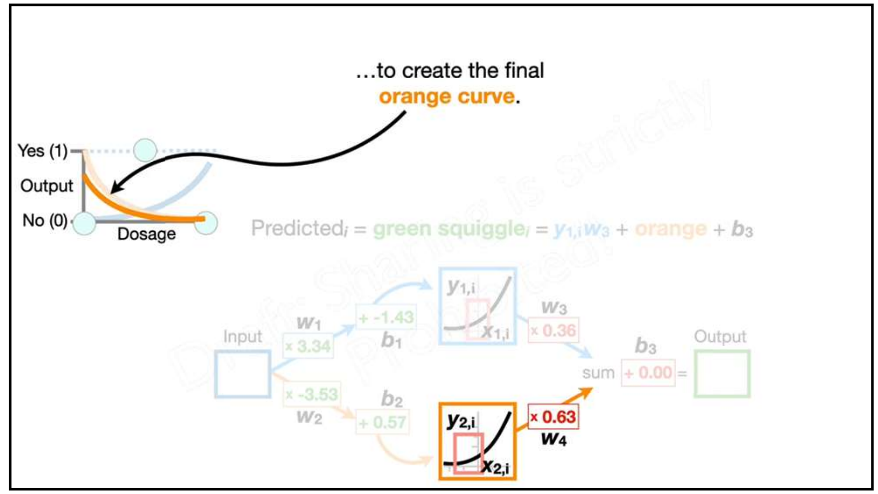

However, we get the final blue curve by multiplying the y-axis coordinates, $y_{1,i}$, by $w_3$.
Yes (1)
Output
No (0)
Dosage
Input
$w_1$
× 3.34
+ -1.43
$b_1$
$y_{1,i}$
$x_{1,i}$
$w_3$
× 0.36
$b_3$
Output
sum
+ 0.00
× -3.53
$w_2$
$b_2$
+ 0.57
$y_{2,i}$
$x_{2,i}$
× 0.63
$w_4$

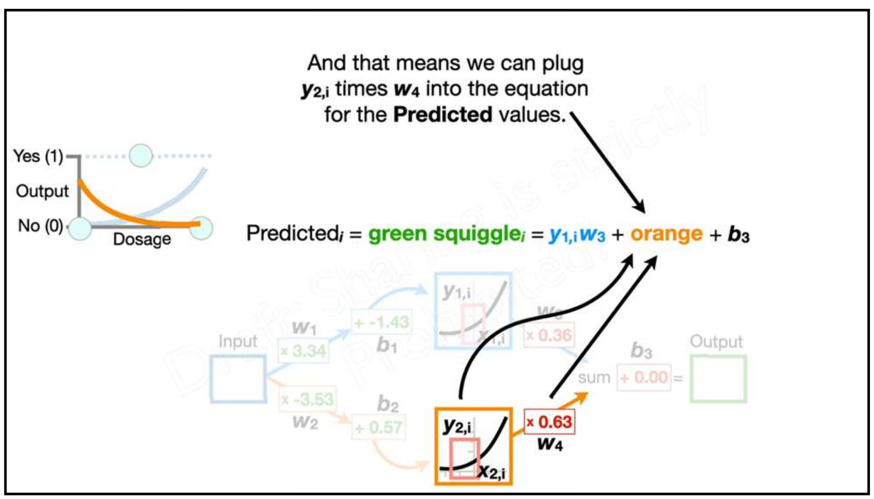

However, we get the final blue curve by multiplying the y-axis coordinates, $y_{1,i}$, by $w_3$.
Yes (1)
Output
No (0)
Dosage
Input
$w_1$
× 3.34
+ -1.43
$b_1$
$y_{1,i}$
$x_{1,i}$
$w_3$
× 0.36
$b_3$
Output
sum
+ 0.00
× -3.53
$w_2$
$b_2$
+ 0.57
$y_{2,i}$
$x_{2,i}$
× 0.63
$w_4$

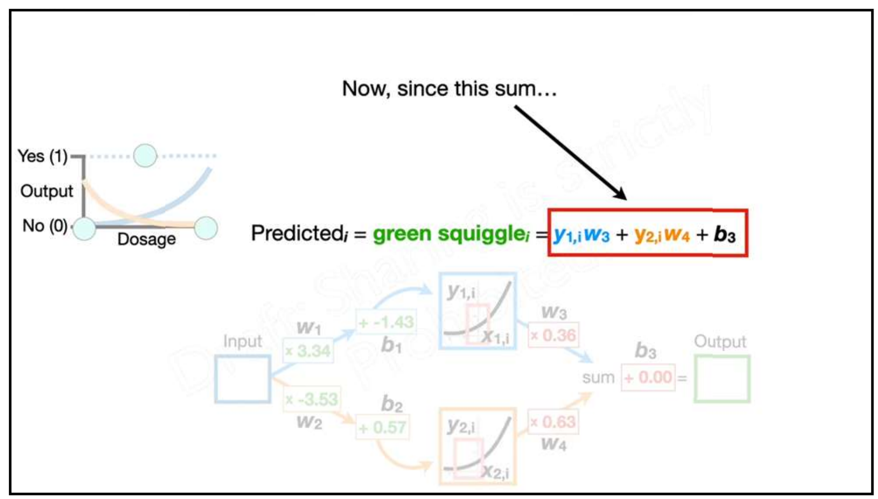

And that means we can plug $y_{1,i}$ times $w_3$ into the equation for the Predicted values.
Yes (1)
Output
No (0)
Dosage
Predicted$_i$ = green squiggle$_i$ = blue + orange + $b_3$
$y_{1,i}$
$x_{1,i}$
$w_3$
× 0.36

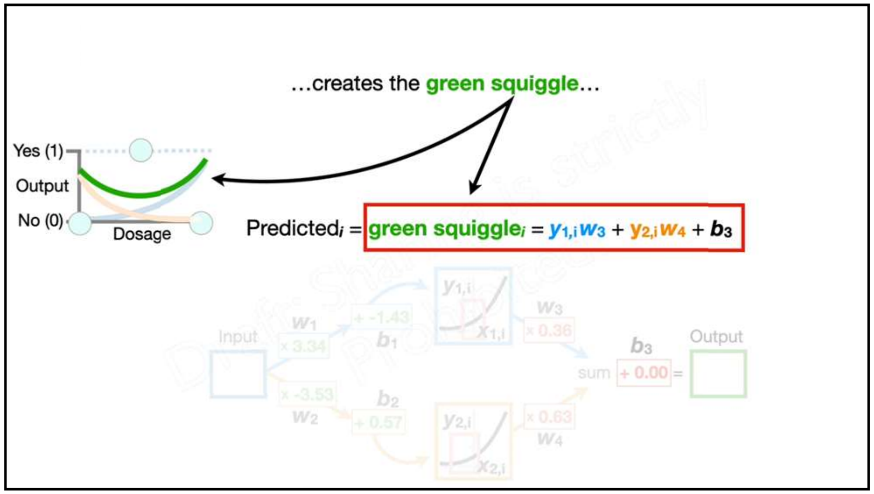

And that means we can plug $y_{1,i}$ times $w_3$ into the equation for the Predicted values.
Yes (1)
Output
No (0)
Dosage
Predicted$_i$ = green squiggle$_i$ = $y_{1,i} w_3$ + orange + $b_3$
$y_{1,i}$
$x_{1,i}$
$w_3$
× 0.36

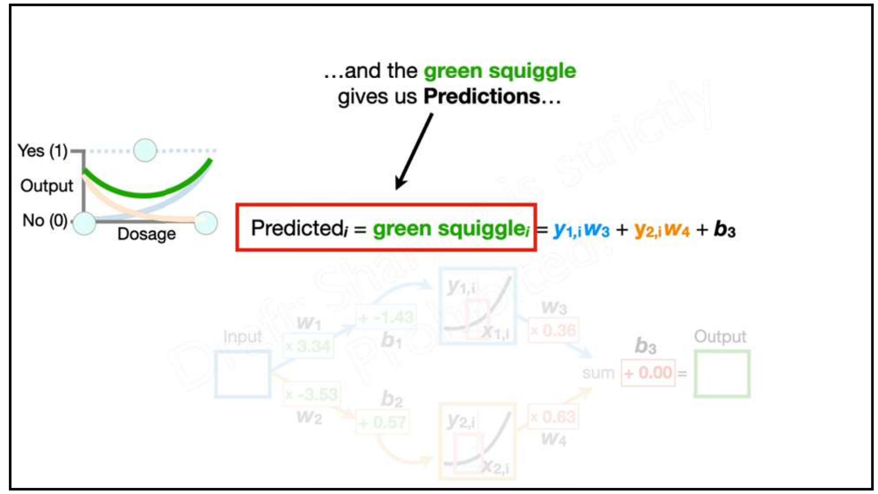

Likewise w4 multiplies the y-axis coordinates, y2,i, from the bottom Activation Function…
Yes (1)
Output
No (0)
Dosage
Predictedi = green squigglei = y1,iw3 + orange + b3
Input
w1
x 3.34
+ -1.43
b1
y1,i
x1,i
w3
x 0.36
b3
sum
+ 0.00
=
Output
x -3.53
w2
b2
+ 0.57
y2,i
x2,i
x 0.63
w4

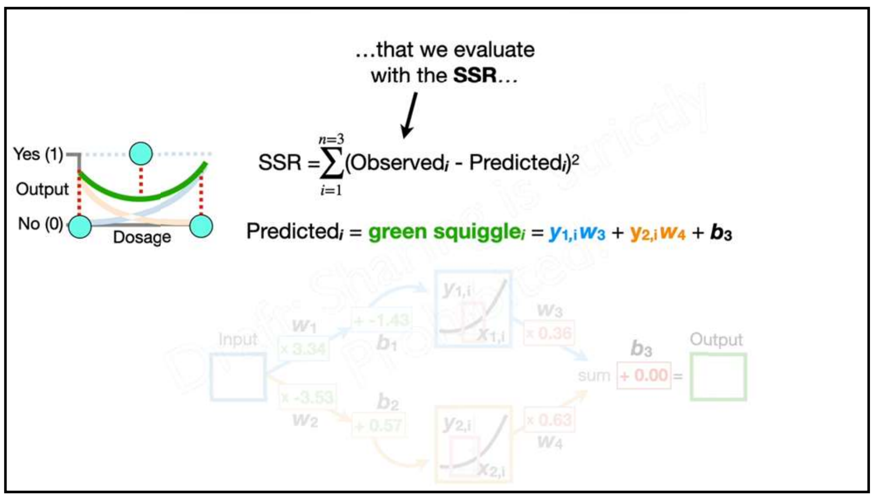

…to create the final orange curve.
Yes (1)
Output
No (0)
Dosage
Predictedi = green squigglei = y1,iw3 + orange + b3
Input
w1
x 3.34
+ -1.43
b1
y1,i
x1,i
w3
x 0.36
b3
sum
+ 0.00
=
Output
x -3.53
w2
b2
+ 0.57
y2,i
x2,i
x 0.63
w4

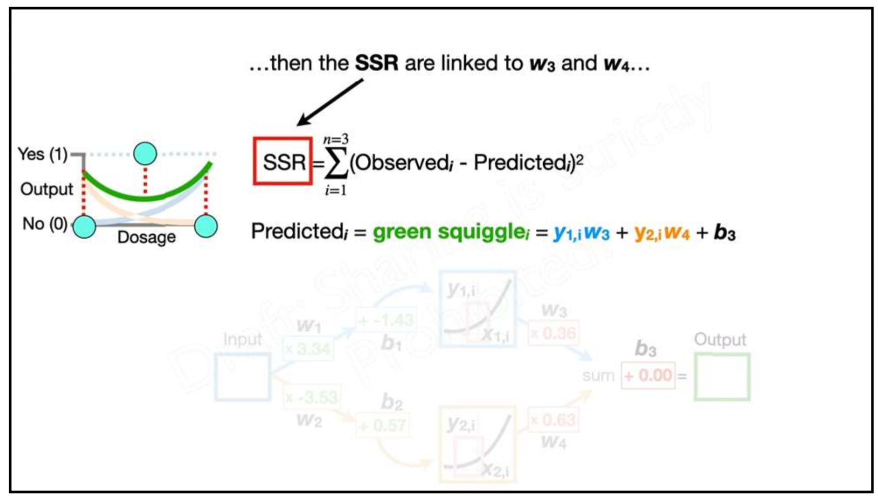

And that means we can plug y2,i times w4 into the equation for the Predicted values.
Yes (1)
Output
No (0)
Dosage
Predictedi = green squigglei = y1,iw3 + orange + b3
Input
w1
x 3.34
+ -1.43
b1
y1,i
x1,i
x 0.36
b3
sum
+ 0.00
=
Output
x -3.53
w2
b2
+ 0.57
y2,i
x2,i
x 0.63
w4

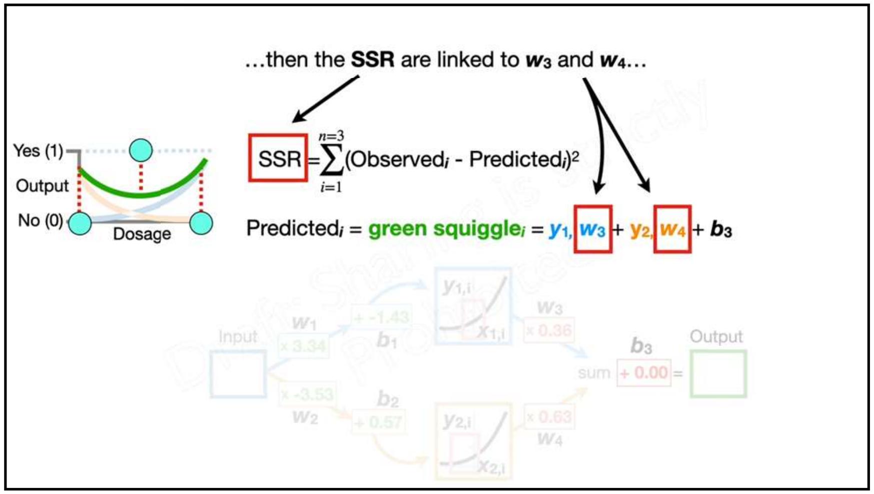

Now, since this sum…
Yes (1)
Output
No (0)
Dosage
Predictedi = green squigglei = y1,iw3 + y2,iw4 + b3
Input
w1
x 3.34
+ -1.43
b1
y1,i
x1,i
w3
x 0.36
b3
sum
+ 0.00
=
Output
x -3.53
w2
b2
+ 0.57
y2,i
x2,i
x 0.63
w4

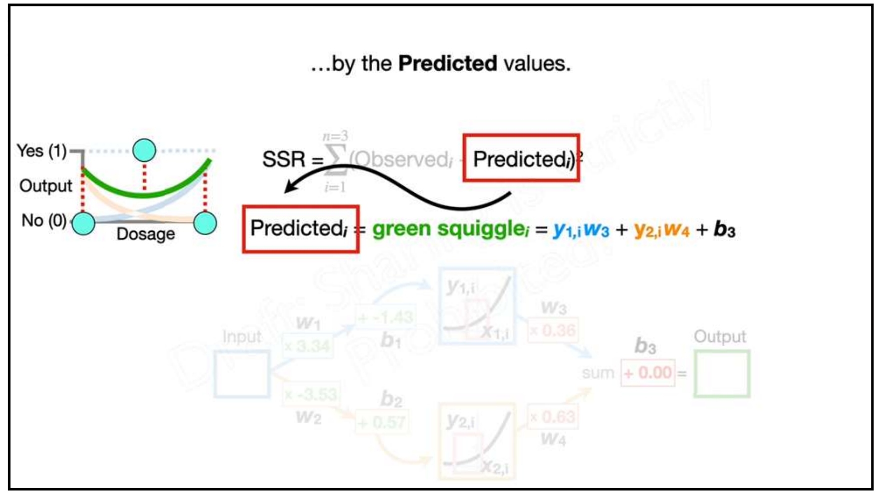

…creates the green squiggle…
Yes (1)
Output
No (0)
Dosage
Predicted$_i$ = green squiggle$_i$ = $y_{1,i}w_3 + y_{2,i}w_4 + b_3$

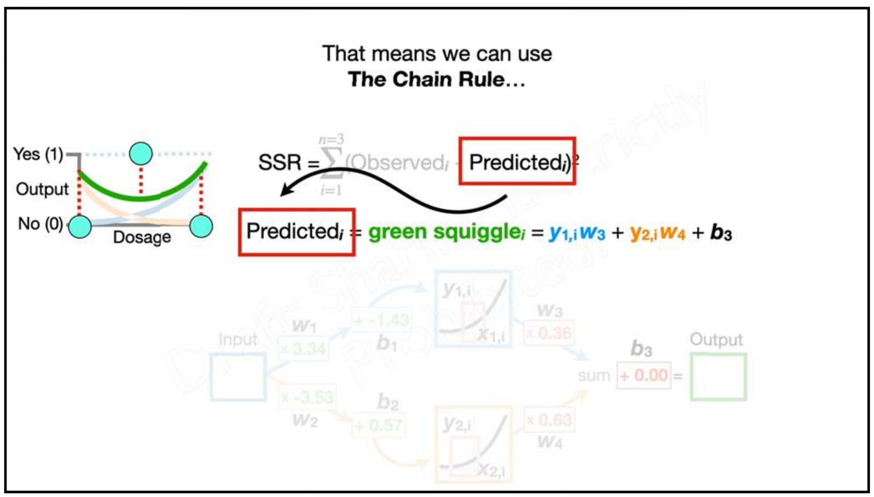

…and the green squiggle gives us Predictions…
Yes (1)
Output
No (0)
Dosage
Predicted$_i$ = green squiggle$_i$ = $y_{1,i}w_3 + y_{2,i}w_4 + b_3$

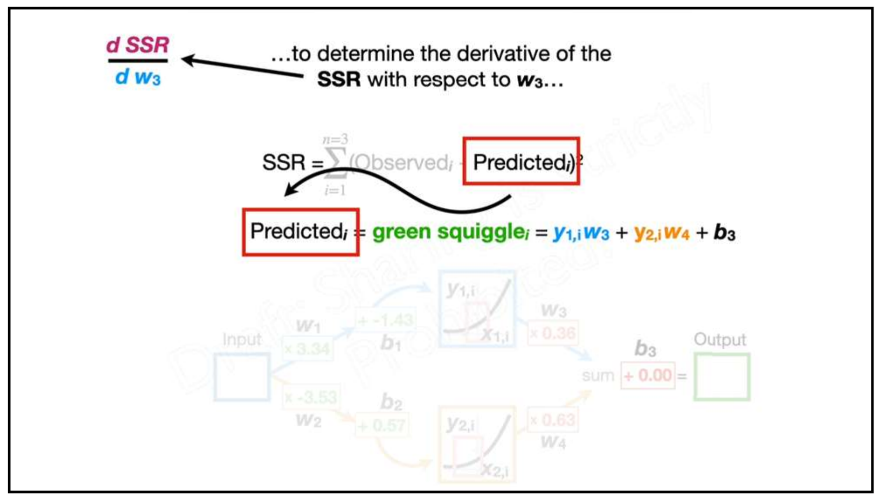

…that we evaluate with the SSR…
Yes (1)
Output
No (0)
Dosage
SSR = $\sum_{i=1}^{n=3}$(Observed$_i$ - Predicted$_i$)$^2$
Predicted$_i$ = green squiggle$_i$ = $y_{1,i}w_3 + y_{2,i}w_4 + b_3$

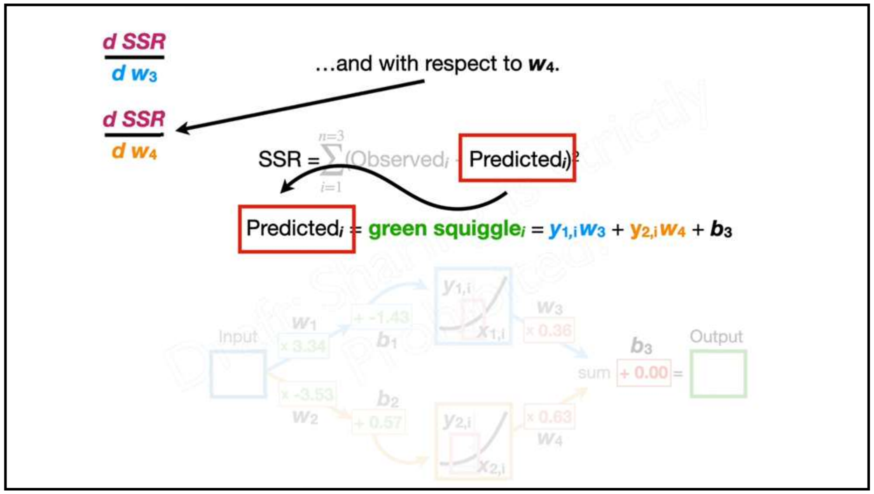

…then the SSR are linked to $w_3$ and $w_4$…
Yes (1)
Output
No (0)
Dosage
SSR = $\sum_{i=1}^{n=3}$(Observed$_i$ - Predicted$_i$)$^2$
Predicted$_i$ = green squiggle$_i$ = $y_{1,i}w_3 + y_{2,i}w_4 + b_3$

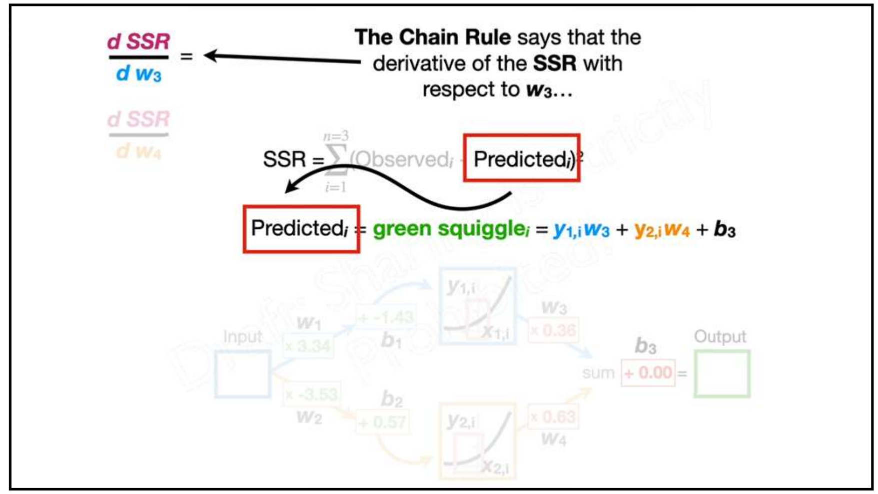

…then the **SSR** are linked to $w_3$ and $w_4$…
Yes (1)
Output
No (0)
Dosage
$SSR = \sum_{i=1}^{n=3}(\text{Observed}_i - \text{Predicted}_i)^2$
$\text{Predicted}_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$

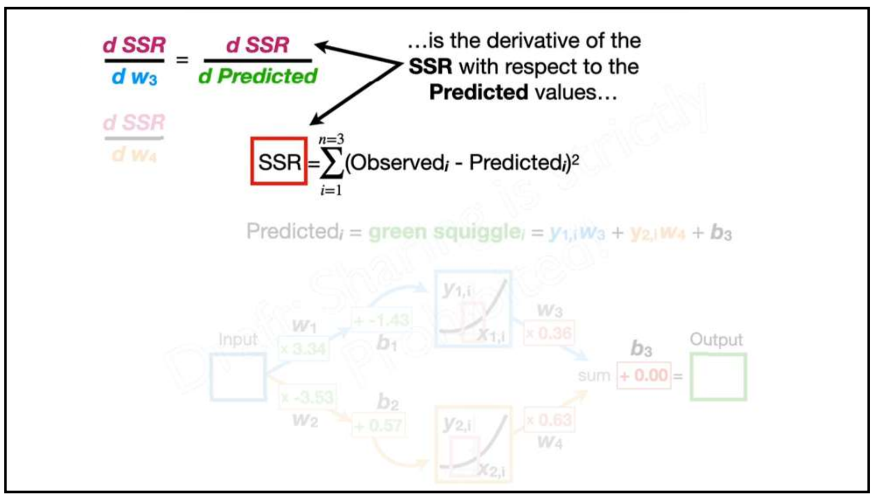

…by the **Predicted** values.
Yes (1)
Output
No (0)
Dosage
$SSR = \sum_{i=1}^{n=3}(\text{Observed}_i - \text{Predicted}_i)^2$
$\text{Predicted}_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$

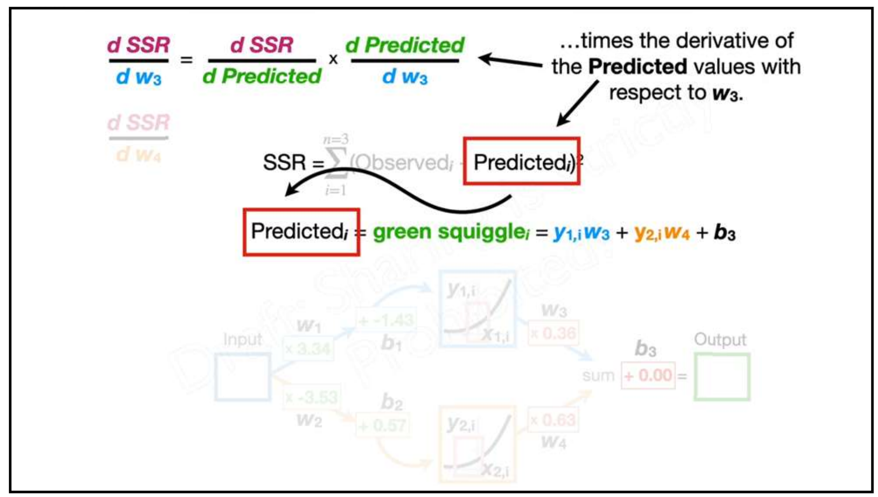

That means we can use
**The Chain Rule**…
Yes (1)
Output
No (0)
Dosage
$SSR = \sum_{i=1}^{n=3}(\text{Observed}_i - \text{Predicted}_i)^2$
$\text{Predicted}_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$


$\frac{d\ SSR}{d\ w_3}$
…to determine the derivative of the
**SSR** with respect to $w_3$…
$SSR = \sum_{i=1}^{n=3}(\text{Observed}_i - \text{Predicted}_i)^2$
$\text{Predicted}_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$


d SSR / d w3
d SSR / d w4
…and with respect to w4.
SSR = Σ_{i=1}^{n=3} (Observed_i - Predicted_i)^2
Predicted_i = green squiggle_i = y1,i w3 + y2,i w4 + b3


d SSR / d w3 =
The Chain Rule says that the derivative of the SSR with respect to w3…
SSR = Σ_{i=1}^{n=3} (Observed_i - Predicted_i)^2
Predicted_i = green squiggle_i = y1,i w3 + y2,i w4 + b3


d SSR / d w3 = d SSR / d Predicted
…is the derivative of the SSR with respect to the Predicted values…
SSR = Σ_{i=1}^{n=3} (Observed_i - Predicted_i)^2


d SSR / d w3 = d SSR / d Predicted x d Predicted / d w3
…times the derivative of the Predicted values with respect to w3.
SSR = Σ_{i=1}^{n=3} (Observed_i - Predicted_i)^2
Predicted_i = green squiggle_i = y1,i w3 + y2,i w4 + b3

$$\frac{d\,SSR}{d\,w_3} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_3}$$

$$\frac{d\,SSR}{d\,w_4} =$$

Likewise, the derivative with respect to ***w*₄**…

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

$$\boxed{Predicted_i} = \textbf{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$$

---

$$\frac{d\,SSR}{d\,w_3} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_3}$$

$$\frac{d\,SSR}{d\,w_4} = \frac{d\,SSR}{d\,Predicted}$$

…is the derivative of the **SSR** with respect to the **Predicted** values…

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

$$Predicted_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$$

---

$$\frac{d\,SSR}{d\,w_3} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_3}$$

$$\frac{d\,SSR}{d\,w_4} = \frac{d\,SSR}{d\,Predicted}$$

…is the derivative of the **SSR** with respect to the **Predicted** values…

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

$$Predicted_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$$

---

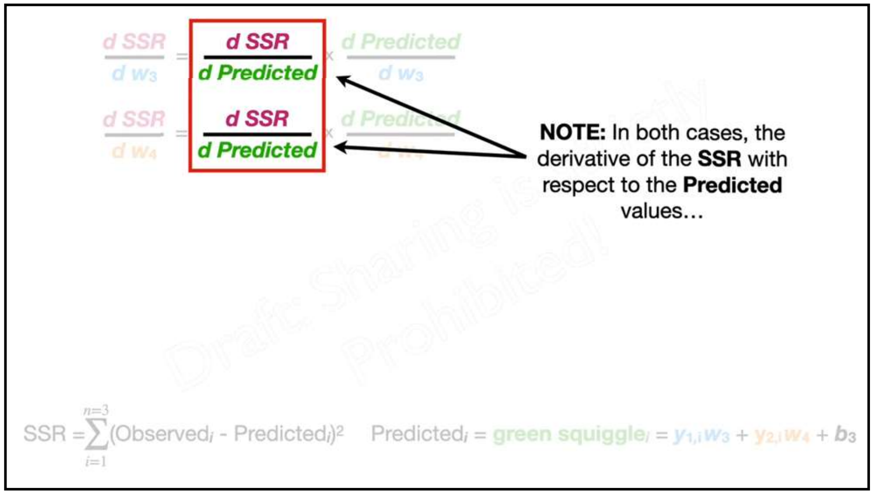
$$\frac{d\,SSR}{d\,w_3} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_3}$$

$$\frac{d\,SSR}{d\,w_4} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_4}$$

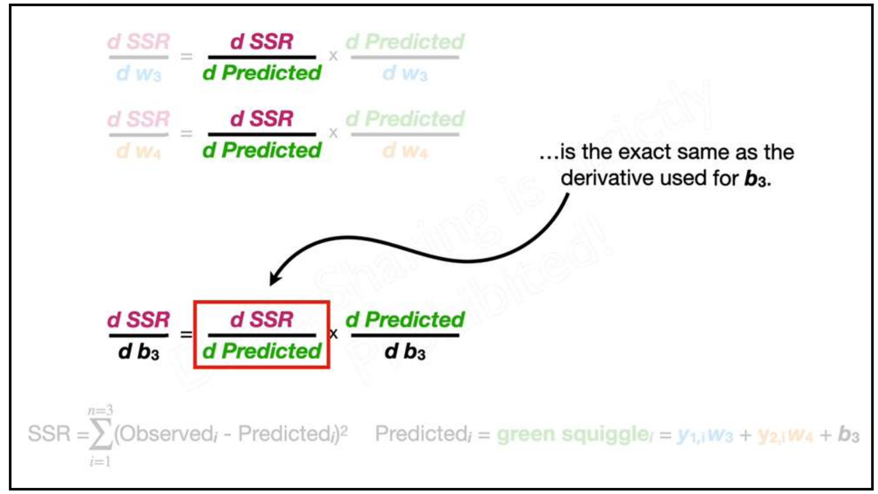
…times the derivative of the **Predicted** values with respect to ***w*₄**.

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

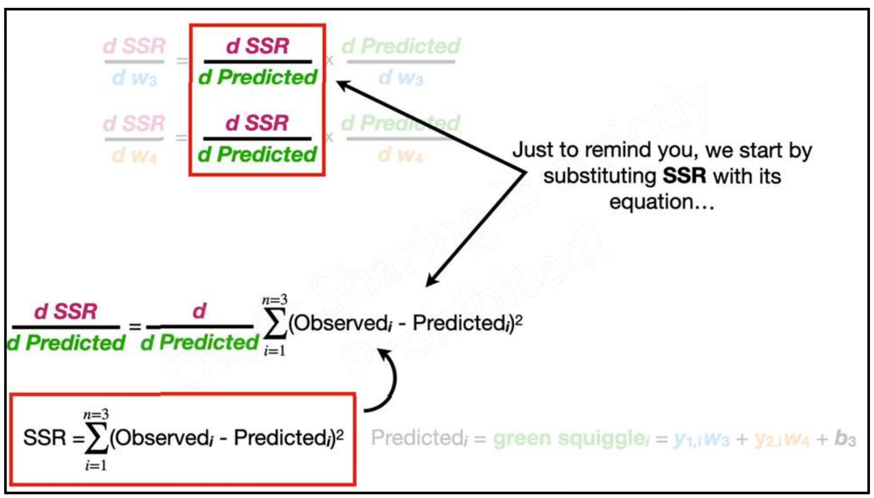
$$\boxed{Predicted_i} = \textbf{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$$



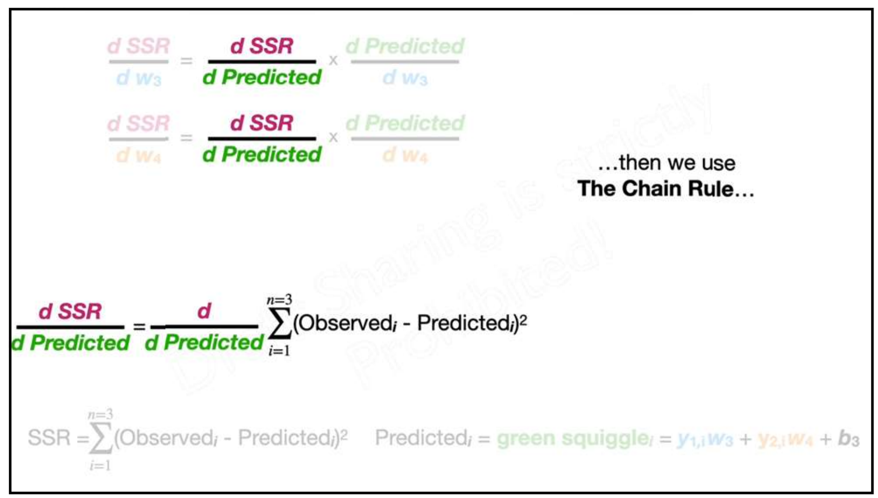
$\frac{d\ SSR}{d\ w_3} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ w_3}$
$\frac{d\ SSR}{d\ w_4} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ w_4}$
**NOTE:** In both cases, the derivative of the **SSR** with respect to the **Predicted** values…
$SSR = \sum_{i=1}^{n=3}(\text{Observed}_i - \text{Predicted}_i)^2$
$\text{Predicted}_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$

$\frac{d\ SSR}{d\ w_3} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ w_3}$
$\frac{d\ SSR}{d\ w_4} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ w_4}$
…is the exact same as the derivative used for $b_3$.
$\frac{d\ SSR}{d\ b_3} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ b_3}$
$SSR = \sum_{i=1}^{n=3}(\text{Observed}_i - \text{Predicted}_i)^2$
$\text{Predicted}_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$

$\frac{d\ SSR}{d\ w_3} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ w_3}$
$\frac{d\ SSR}{d\ w_4} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ w_4}$
Just to remind you, we start by substituting **SSR** with its equation…
$\frac{d\ SSR}{d\ Predicted} = \frac{d}{d\ Predicted}\sum_{i=1}^{n=3}(\text{Observed}_i - \text{Predicted}_i)^2$
$SSR = \sum_{i=1}^{n=3}(\text{Observed}_i - \text{Predicted}_i)^2$
$\text{Predicted}_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$

$\frac{d\ SSR}{d\ w_3} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ w_3}$
$\frac{d\ SSR}{d\ w_4} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ w_4}$
…then we use **The Chain Rule…**
$\frac{d\ SSR}{d\ Predicted} = \frac{d}{d\ Predicted}\sum_{i=1}^{n=3}(\text{Observed}_i - \text{Predicted}_i)^2$
$SSR = \sum_{i=1}^{n=3}(\text{Observed}_i - \text{Predicted}_i)^2$
$\text{Predicted}_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$

$$\frac{d\,SSR}{d\,w_3} = \boxed{\frac{d\,SSR}{d\,Predicted}} \times \frac{d\,Predicted}{d\,w_3}$$

$$\frac{d\,SSR}{d\,w_4} = \boxed{\frac{d\,SSR}{d\,Predicted}} \times \frac{d\,Predicted}{d\,w_4}$$

…and this is the derivative of the **SSR** with respect to the **Predicted** values.

$$\frac{d\,SSR}{d\,Predicted} = \frac{d}{d\,Predicted}\sum_{i=1}^{n=3}(Observed_i - Predicted_i)^2 = \boxed{\sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i)}$$

$SSR = \sum_{i=1}^{n=3}(Observed_i - Predicted_i)^2$ $\quad$ $Predicted_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$



$$\frac{d\,SSR}{d\,w_3} = \boxed{\frac{d\,SSR}{d\,Predicted}} \times \frac{d\,Predicted}{d\,w_3} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i)$$

$$\frac{d\,SSR}{d\,w_4} = \boxed{\frac{d\,SSR}{d\,Predicted}} \times \frac{d\,Predicted}{d\,w_4} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i)$$

So we just plug it in.

$$\frac{d\,SSR}{d\,Predicted} = \frac{d}{d\,Predicted}\sum_{i=1}^{n=3}(Observed_i - Predicted_i)^2 = \boxed{\sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i)}$$

$SSR = \sum_{i=1}^{n=3}(Observed_i - Predicted_i)^2$ $\quad$ $Predicted_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$



$$\frac{d\,SSR}{d\,w_3} = \frac{d\,SSR}{d\,Predicted} \times \boxed{\frac{d\,Predicted}{d\,w_3}} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i)$$

$$\frac{d\,SSR}{d\,w_4} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_4} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i)$$

Now, to solve for the derivative of the **Predicted** values with respect to ***w₃***…

$$\frac{d\,Predicted}{d\,w_3}$$

$SSR = \sum_{i=1}^{n=3}(Observed_i - Predicted_i)^2$ $\quad$ $Predicted_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$



$$\frac{d\,SSR}{d\,w_3} = \frac{d\,SSR}{d\,Predicted} \times \boxed{\frac{d\,Predicted}{d\,w_3}} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i)$$

$$\frac{d\,SSR}{d\,w_4} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_4} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i)$$

…we plug in the equation for the **Predicted** values.

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

$SSR = \sum_{i=1}^{n=3}(Observed_i - Predicted_i)^2$ $\quad$ $\boxed{Predicted_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3}$

$$\frac{d\ SSR}{d\ w_3} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ w_3} = \sum_{i=1}^{n=3} -2 \times (\text{Observed}_i - \text{Predicted}_i)$$

$$\frac{d\ SSR}{d\ w_4} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ w_4} = \sum_{i=1}^{n=3} -2 \times (\text{Obse}$$

And the derivative of the first term with respect to $w_3$ is $y_{1,i}$…

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

$$\text{SSR} = \sum_{i=1}^{n=3}(\text{Observed}_i - \text{Predicted}_i)^2 \quad \text{Predicted}_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$$



$$\frac{d\ SSR}{d\ w_3} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ w_3} = \sum_{i=1}^{n=3} -2 \times (\text{Observed}_i - \text{Predicted}_i)$$

$$\frac{d\ SSR}{d\ w_4} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ w_4} = \sum_{i=1}^{n=3} -2 \times (\text{Ob}$$

…and the derivatives of the other terms are both **0** since they do not contain $w_3$.

$$\frac{d\ Predicted}{d\ w_3} = \frac{d}{d\ w_3}\ \text{green squiggle} = \frac{d}{d\ w_3}(y_{1,i}w_3 + y_{2,i}w_4 + b_3) = y_{1,i} + 0$$

$$\text{SSR} = \sum_{i=1}^{n=3}(\text{Observed}_i - \text{Predicted}_i)^2 \quad \text{Predicted}_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$$



$$\frac{d\ SSR}{d\ w_3} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ w_3} = \sum_{i=1}^{n=3} -2 \times (\text{Observed}_i - \text{Predicted}_i)$$

$$\frac{d\ SSR}{d\ w_4} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ w_4} = \sum_{i=1}^{n=3} -2 \times (\text{Ob}$$

…and the derivatives of the other terms are both **0** since they do not contain $w_3$.

$$\frac{d\ Predicted}{d\ w_3} = \frac{d}{d\ w_3}\ \text{green squiggle} = \frac{d}{d\ w_3}(y_{1,i}w_3 + y_{2,i}w_4 + b_3) = y_{1,i} + 0 + 0$$

$$\text{SSR} = \sum_{i=1}^{n=3}(\text{Observed}_i - \text{Predicted}_i)^2 \quad \text{Predicted}_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$$



$$\frac{d\ SSR}{d\ w_3} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ w_3} = \sum_{i=1}^{n=3} -2 \times (\text{Observed}_i - \text{Predicted}_i)$$

$$\frac{d\ SSR}{d\ w_4} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ w_4} = \sum_{i=1}^{n=3} -2 \times (\text{Observed}_i - \text{Predicted}_i)$$

And we end up with just $y_{1,i}$.

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

$$\text{SSR} = \sum_{i=1}^{n=3}(\text{Observed}_i - \text{Predicted}_i)^2 \quad \text{Predicted}_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$$

$$\frac{d\,SSR}{d\,w_3} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_3} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{1,i}$$

$$\frac{d\,SSR}{d\,w_4} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_4} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i)$$

So we multiply the derivative of the **SSR** with respect to the **Predicted** values by $y_{1,i}$.

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

$$SSR = \sum_{i=1}^{n=3}(Observed_i - Predicted_i)^2 \qquad Predicted_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$$

$$\frac{d\,SSR}{d\,w_3} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_3} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{1,i}$$

$$\frac{d\,SSR}{d\,w_4} = \frac{d\,SSR}{d\,Predicted} \times \boxed{\frac{d\,Predicted}{d\,w_4}} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i)$$

Likewise, the derivative of the **Predicted** values with respect to $w_4$ is…

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

$$SSR = \sum_{i=1}^{n=3}(Observed_i - Predicted_i)^2 \qquad Predicted_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$$

$$\frac{d\,SSR}{d\,w_3} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_3} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{1,i}$$

$$\frac{d\,SSR}{d\,w_4} = \frac{d\,SSR}{d\,Predicted} \times \boxed{\frac{d\,Predicted}{d\,w_4}} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i)$$

…**0** for the first term…

$$\frac{d\,Predicted}{d\,w_4} = \frac{d}{d\,w_4}\text{ green squiggle } = \frac{d}{d\,w_4}(y_{1,i}w_3 + y_{2,i}w_4 + b_3) = 0$$

$$SSR = \sum_{i=1}^{n=3}(Observed_i - Predicted_i)^2 \qquad Predicted_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$$

$$\frac{d\,SSR}{d\,w_3} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_3} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{1,i}$$

$$\frac{d\,SSR}{d\,w_4} = \frac{d\,SSR}{d\,Predicted} \times \boxed{\frac{d\,Predicted}{d\,w_4}} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i)$$

…plus $y_{2,i}$ for the second term…

$$\frac{d\,Predicted}{d\,w_4} = \frac{d}{d\,w_4}\text{ green squiggle } = \frac{d}{d\,w_4}(y_{1,i}w_3 + y_{2,i}w_4 + b_3) = 0 + y_{2,i}$$

$$SSR = \sum_{i=1}^{n=3}(Observed_i - Predicted_i)^2 \qquad Predicted_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$$

$$\frac{d\,SSR}{d\,w_3} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_3} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{1,i}$$

$$\frac{d\,SSR}{d\,w_4} = \frac{d\,SSR}{d\,Predicted} \times \boxed{\frac{d\,Predicted}{d\,w_4}} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i)$$

…plus **0** for the third term…

$$\frac{d\,Predicted}{d\,w_4} = \frac{d}{d\,w_4}\,\text{green squiggle} = \frac{d}{d\,w_4}(y_{1,i}w_3 + y_{2,i}w_4 + b_3) = 0 + y_{2,i} + 0$$

$$SSR = \sum_{i=1}^{n=3}(Observed_i - Predicted_i)^2 \quad Predicted_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$$



$$\frac{d\,SSR}{d\,w_3} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_3} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{1,i}$$

$$\frac{d\,SSR}{d\,w_4} = \frac{d\,SSR}{d\,Predicted} \times \boxed{\frac{d\,Predicted}{d\,w_4}} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i)$$

…which is just $y_{2,i}$.

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

$$SSR = \sum_{i=1}^{n=3}(Observed_i - Predicted_i)^2 \quad Predicted_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$$

Draft Sharing is Strictly Prohibited!

$$\frac{d\,SSR}{d\,w_3} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_3} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{1,i}$$

$$\frac{d\,SSR}{d\,w_4} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_4} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{2,i}$$

So we multiply the derivative of the **SSR** with respect to the **Predicted** values by $y_{2,i}$.

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

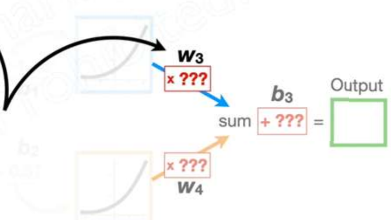
$$SSR = \sum_{i=1}^{n=3}(Observed_i - Predicted_i)^2 \quad Predicted_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$$

Draft Sharing is Strictly Prohibited!

$$\frac{d\,SSR}{d\,w_3} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_3} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{1,i}$$

$$\frac{d\,SSR}{d\,w_4} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_4} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{2,i}$$

Now that we have the derivatives of the **SSR** with respect to $w_3$…

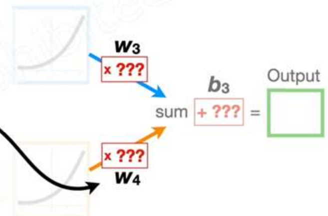



Draft Sharing is Strictly Prohibited!

$$\frac{d\,SSR}{d\,w_3} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_3} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{1,i}$$

$$\frac{d\,SSR}{d\,w_4} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_4} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{2,i}$$

…$w_4$…

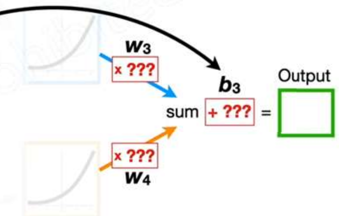



$$\frac{d\,SSR}{d\,w_3} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_3} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{1,i}$$

$$\frac{d\,SSR}{d\,w_4} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_4} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{2,i}$$

$$\frac{d\,SSR}{d\,b_3} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,b_3} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times \mathbf{1}$$

…and $b_3$…

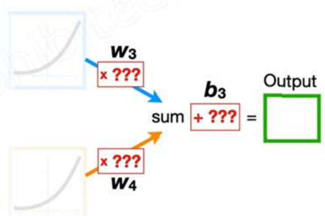



$$\frac{d\,SSR}{d\,w_3} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_3} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{1,i}$$

$$\frac{d\,SSR}{d\,w_4} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_4} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{2,i}$$

$$\frac{d\,SSR}{d\,b_3} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,b_3} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times \mathbf{1}$$

…we can plug them into **Gradient Descent** to optimize $w_3$, $w_4$ and $b_3$.

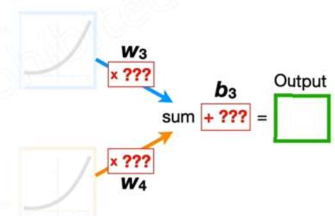


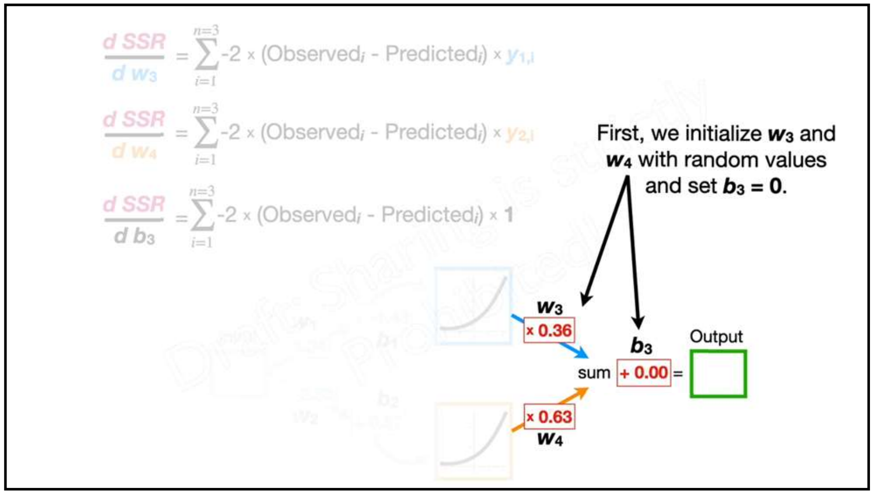

$$\frac{d\,SSR}{d\,w_3} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{1,i}$$

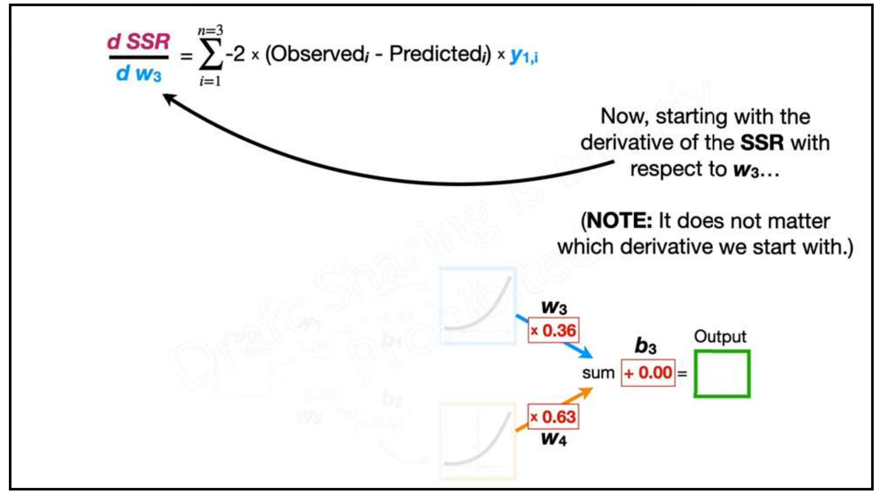
$$\frac{d\,SSR}{d\,w_4} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{2,i}$$

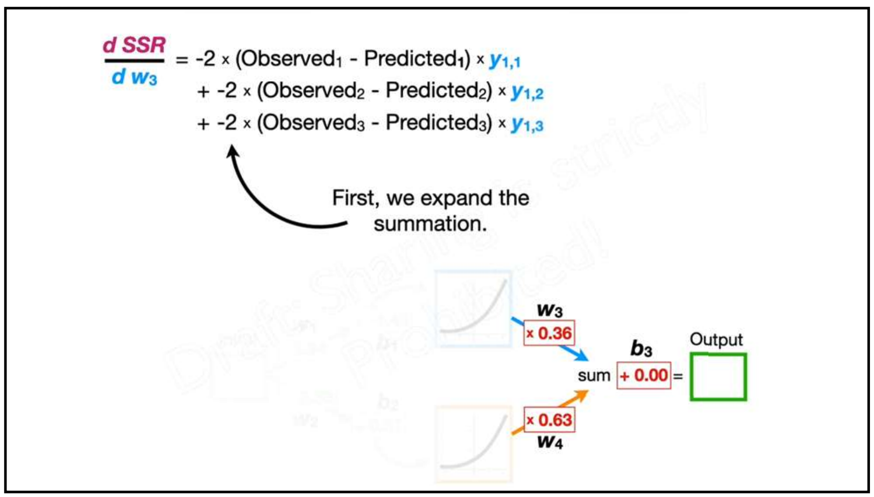
$$\frac{d\,SSR}{d\,b_3} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times \mathbf{1}$$

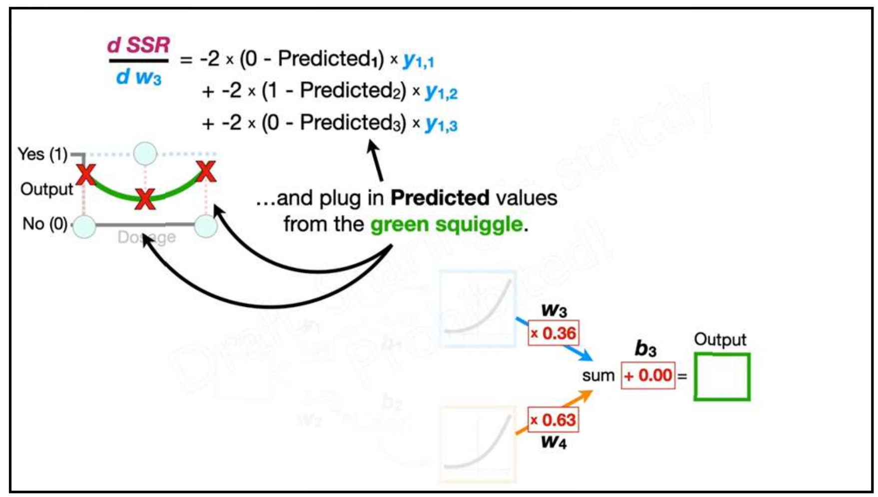
…we can plug them into **Gradient Descent** to optimize $w_3$, $w_4$ and $b_3$.

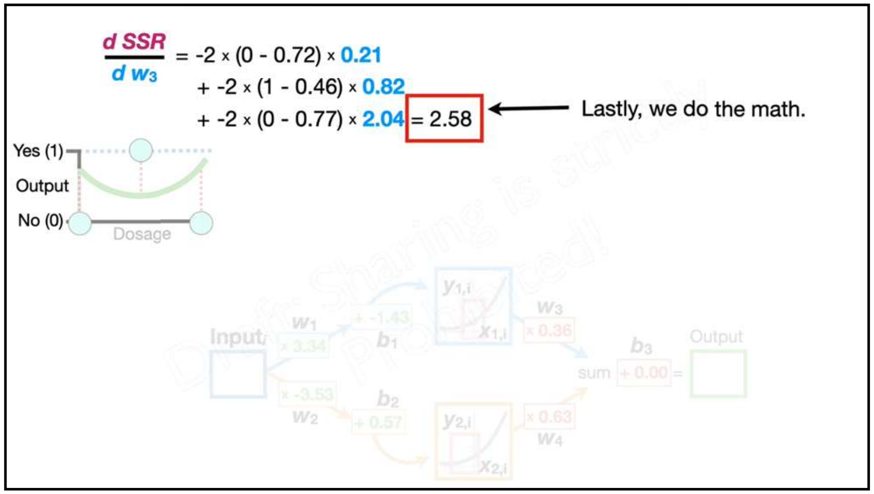


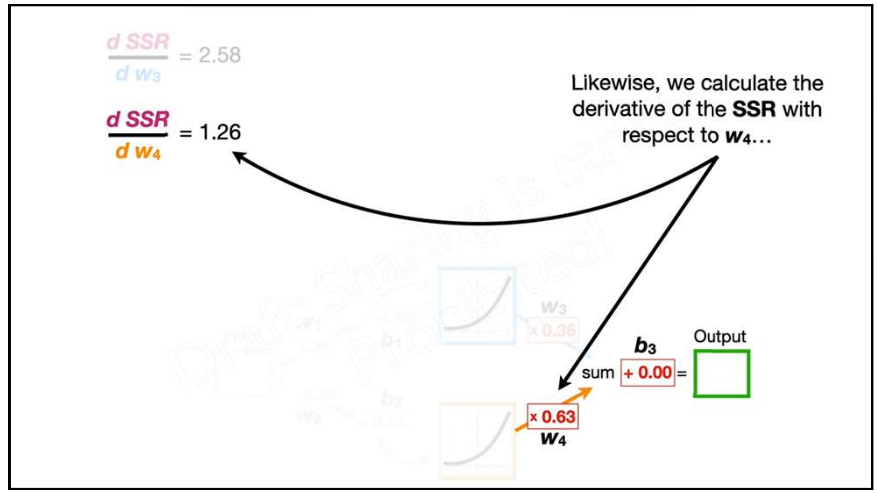

d SSR / d w3 = Σ(i=1 to n=3) -2 x (Observed_i - Predicted_i) x y1,i
d SSR / d w4 = Σ(i=1 to n=3) -2 x (Observed_i - Predicted_i) x y2,i
d SSR / d b3 = Σ(i=1 to n=3) -2 x (Observed_i - Predicted_i) x 1
First, we initialize w3 and w4 with random values and set b3 = 0.
w3
x 0.36
b3
sum + 0.00 =
Output
x 0.63
w4

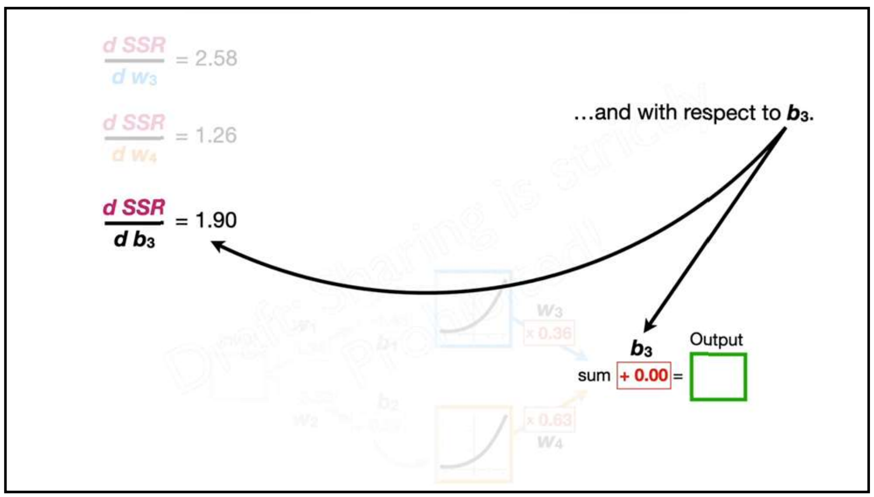

d SSR / d w3 = Σ(i=1 to n=3) -2 x (Observed_i - Predicted_i) x y1,i
Now, starting with the derivative of the SSR with respect to w3…
(NOTE: It does not matter which derivative we start with.)
w3
x 0.36
b3
sum + 0.00 =
Output
x 0.63
w4

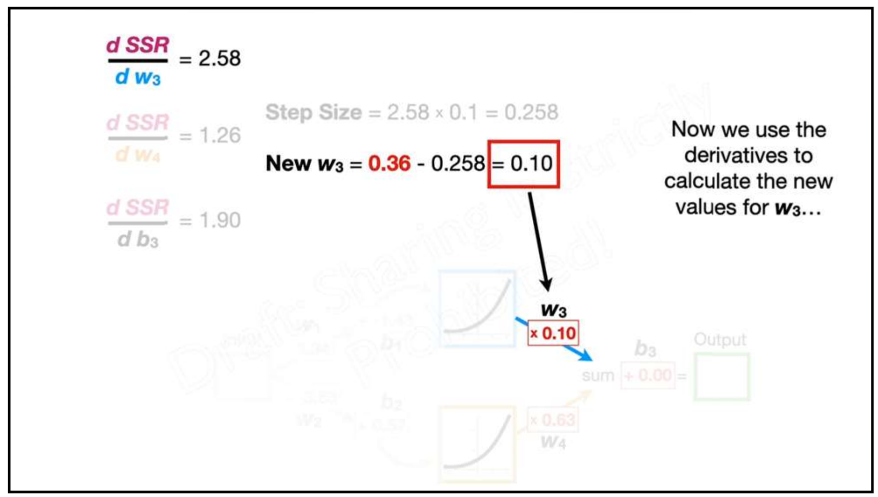

d SSR / d w3 = -2 x (Observed1 - Predicted1) x y1,1
+ -2 x (Observed2 - Predicted2) x y1,2
+ -2 x (Observed3 - Predicted3) x y1,3
First, we expand the summation.
w3
x 0.36
b3
sum + 0.00 =
Output
x 0.63
w4


d SSR / d w3 = -2 x (0 - Predicted1) x y1,1
+ -2 x (1 - Predicted2) x y1,2
+ -2 x (0 - Predicted3) x y1,3
Yes (1)
Output
No (0)
Dosage
…and plug in Predicted values from the green squiggle.
w3
x 0.36
b3
sum + 0.00 =
Output
x 0.63
w4

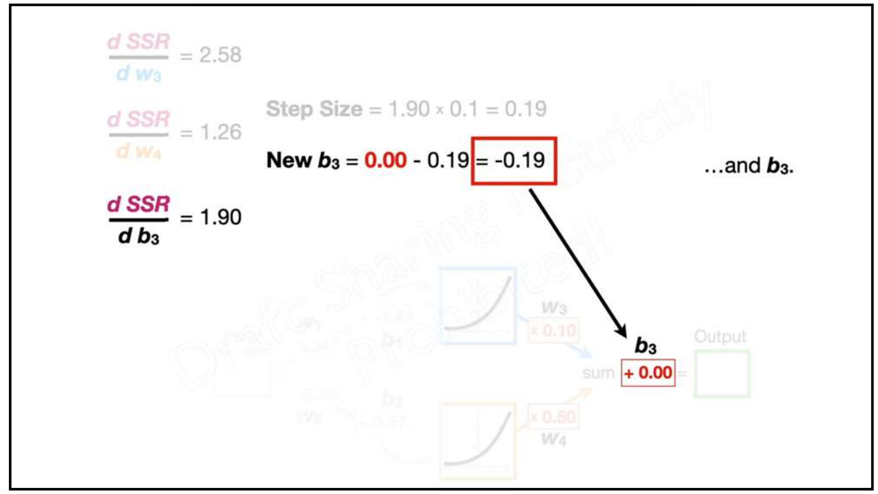

d SSR / d w3 = -2 × (0 - 0.72) × 0.21
+ -2 × (1 - 0.46) × 0.82
+ -2 × (0 - 0.77) × 2.04 = 2.58
Lastly, we do the math.
Yes (1)
Output
No (0)
Dosage

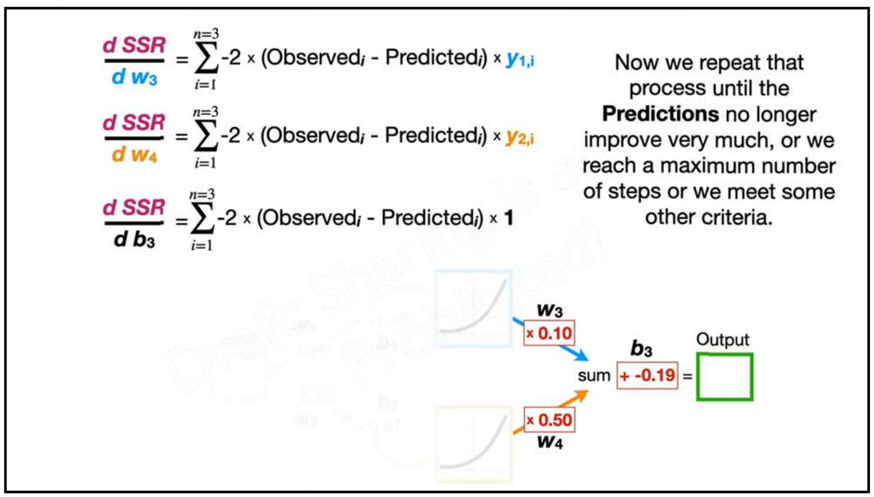

d SSR / d w3 = 2.58
d SSR / d w4 = 1.26
Likewise, we calculate the derivative of the SSR with respect to w4…
b3
Output
sum + 0.00 =
x 0.63
w4

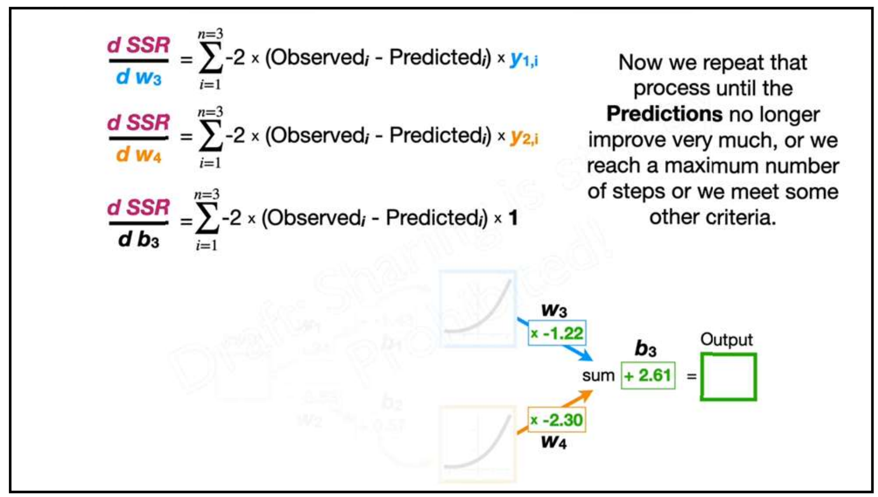

d SSR / d w3 = 2.58
d SSR / d w4 = 1.26
d SSR / d b3 = 1.90
…and with respect to b3.
b3
Output
sum + 0.00 =


d SSR / d w3 = 2.58
Step Size = 2.58 × 0.1 = 0.258
d SSR / d w4 = 1.26
New w3 = 0.36 - 0.258 = 0.10
d SSR / d b3 = 1.90
Now we use the derivatives to calculate the new values for w3…
w3
x 0.10


d SSR / d w3 = 2.58
d SSR / d w4 = 1.26
d SSR / d b3 = 1.90
Step Size = 1.26 × 0.1 = 0.126
New w4 = 0.63 - 0.126 = 0.50
…w4…
w3
b3
Output
sum
× 0.63
w4


d SSR / d w3 = 2.58
d SSR / d w4 = 1.26
d SSR / d b3 = 1.90
Step Size = 1.90 × 0.1 = 0.19
New b3 = 0.00 - 0.19 = -0.19
…and b3.
w3
× 0.10
b3
+ 0.00
Output
sum
× 0.50
w4

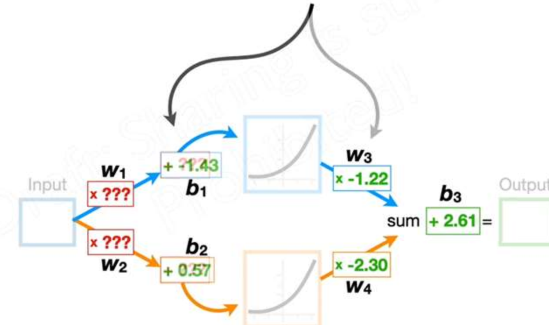

d SSR / d w3 = Σ(i=1 to n=3) -2 × (Observed_i - Predicted_i) × y1,i
d SSR / d w4 = Σ(i=1 to n=3) -2 × (Observed_i - Predicted_i) × y2,i
d SSR / d b3 = Σ(i=1 to n=3) -2 × (Observed_i - Predicted_i) × 1
Now we repeat that process until the Predictions no longer improve very much, or we reach a maximum number of steps or we meet some other criteria.
w3
× 0.10
b3
Output
sum
+ -0.19
=
× 0.50
w4


d SSR / d w3 = Σ(i=1 to n=3) -2 × (Observed_i - Predicted_i) × y1,i
d SSR / d w4 = Σ(i=1 to n=3) -2 × (Observed_i - Predicted_i) × y2,i
d SSR / d b3 = Σ(i=1 to n=3) -2 × (Observed_i - Predicted_i) × 1
Now we repeat that process until the Predictions no longer improve very much, or we reach a maximum number of steps or we meet some other criteria.
w3
× -1.22
b3
Output
sum
+ 2.61
=
× -2.30
w4

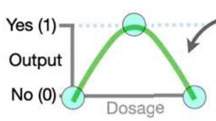
# Neural Networks...

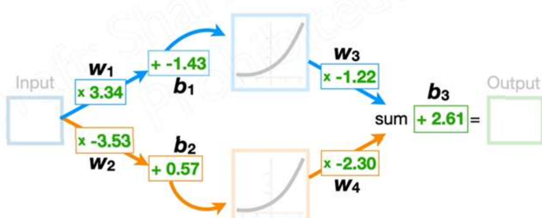
## UPDATING ALL PARAMETERS

...in order to optimize all of the **Weights** and **Biases** in this **Neural Network**...

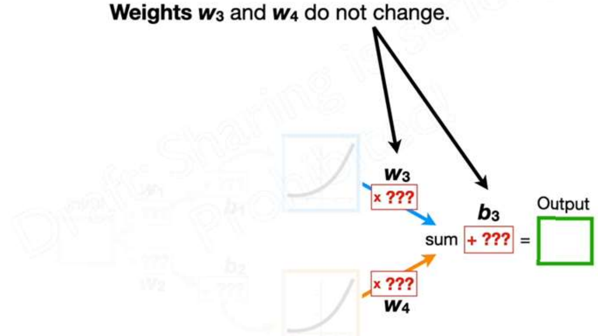


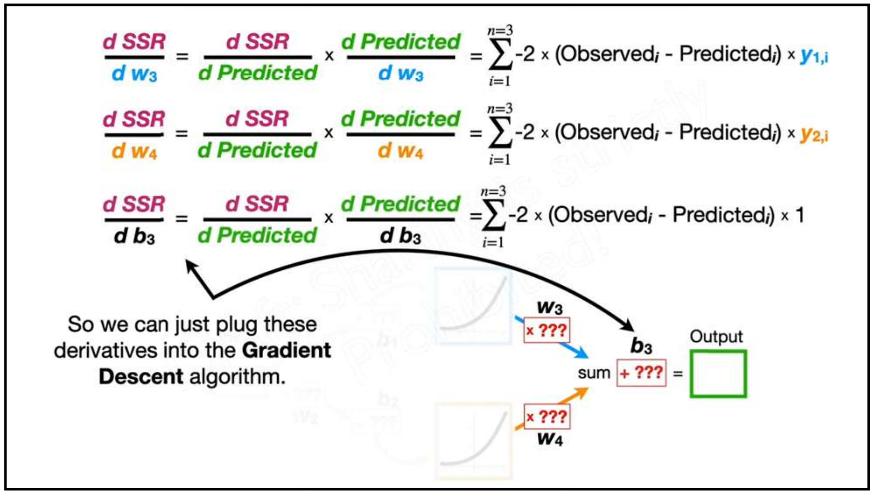

...so that it can fit a **green squiggle** to this data.



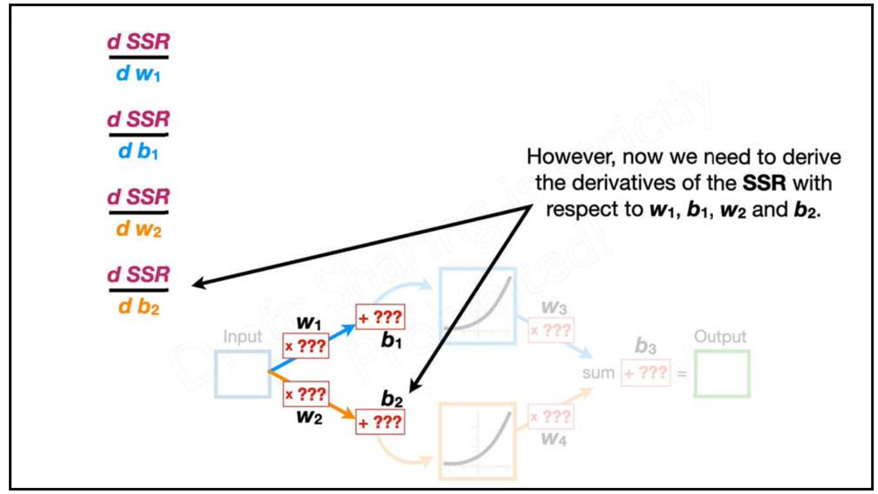




**NOTE:** The derivatives that we derived in **Part 1** for **Bias** $b_3$ and **Weights** $w_3$ and $w_4$ do not change.

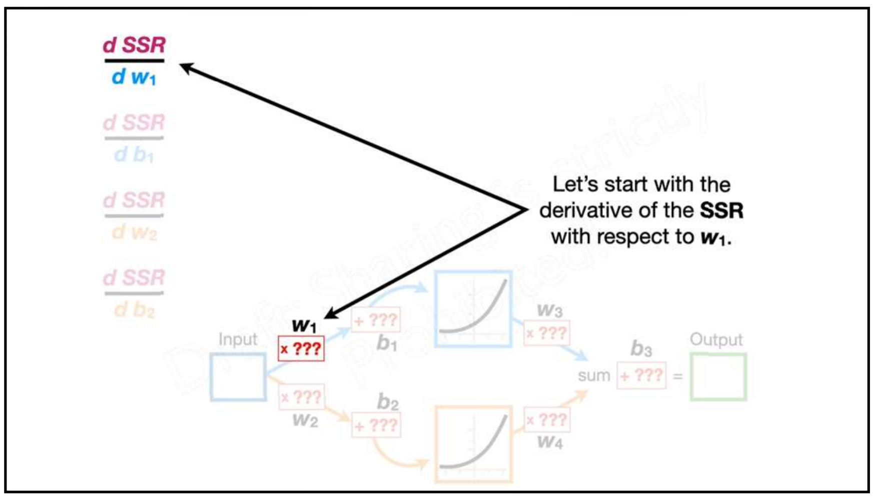




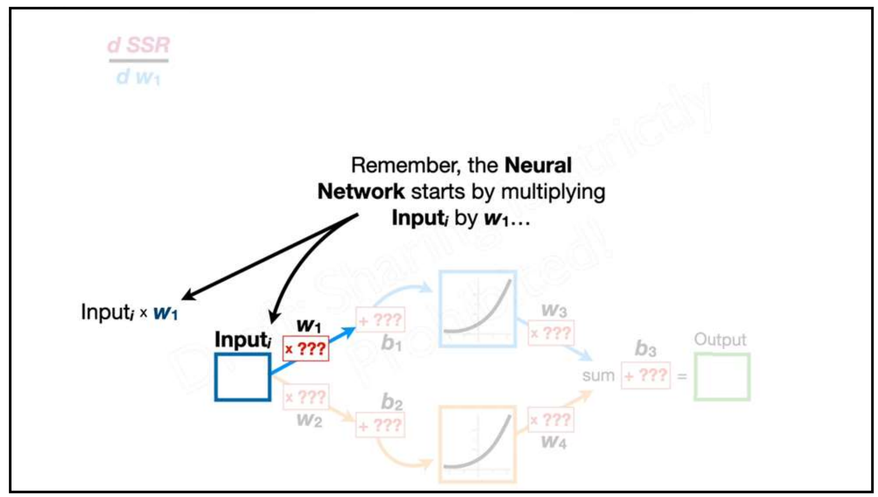
$\frac{d\,SSR}{d\,w_3} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_3} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{1,i}$
$\frac{d\,SSR}{d\,w_4} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,w_4} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{2,i}$
$\frac{d\,SSR}{d\,b_3} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,b_3} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times 1$
So we can just plug these derivatives into the **Gradient Descent** algorithm.
$w_3$
x ???
$b_3$
sum + ??? =
Output
x ???
$w_4$

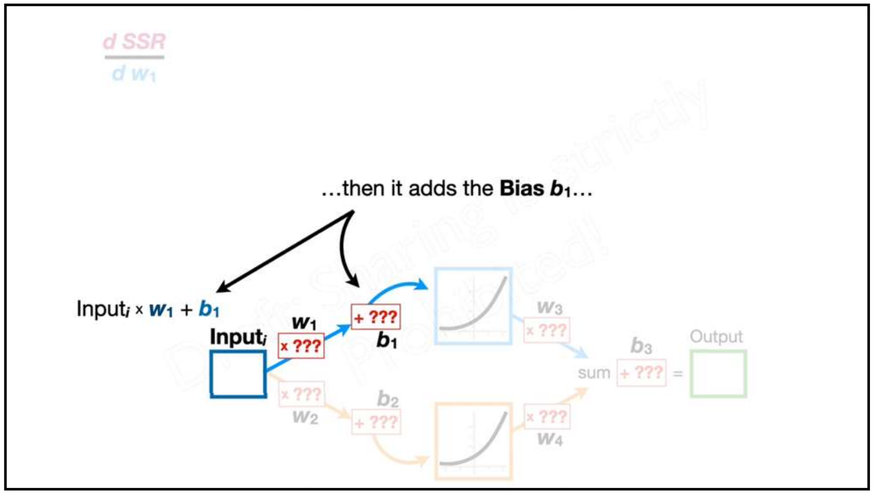
$\frac{d\,SSR}{d\,w_1}$
$\frac{d\,SSR}{d\,b_1}$
$\frac{d\,SSR}{d\,w_2}$
$\frac{d\,SSR}{d\,b_2}$
However, now we need to derive the derivatives of the **SSR** with respect to $w_1$, $b_1$, $w_2$ and $b_2$.
Input
$w_1$
x ???
+ ???
$b_1$
x ???
$w_2$
$b_2$
+ ???

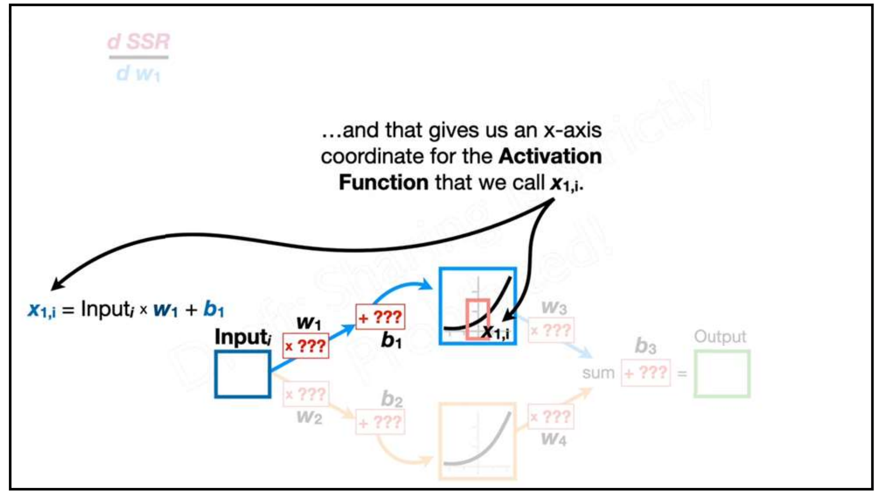



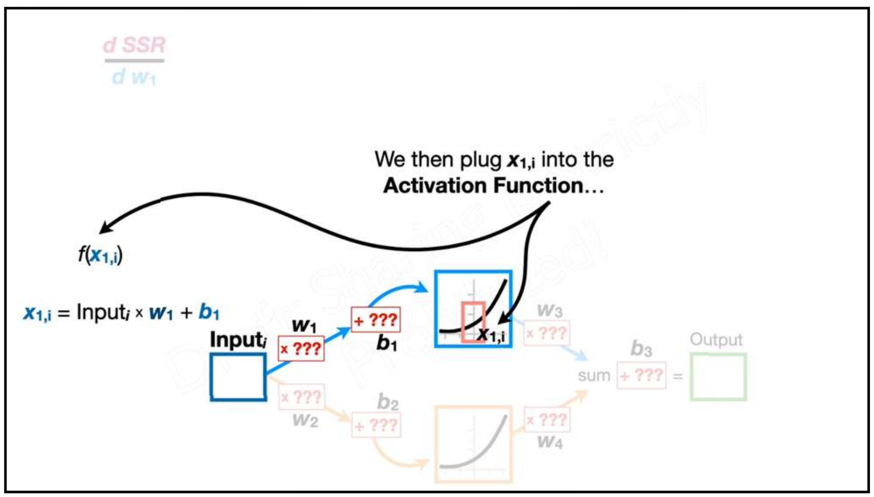
$\frac{d\,SSR}{d\,w_1}$
$\frac{d\,SSR}{d\,b_1}$
$\frac{d\,SSR}{d\,w_2}$
$\frac{d\,SSR}{d\,b_2}$
Let's start with the derivative of the **SSR** with respect to $w_1$.
Input
$w_1$
x ???

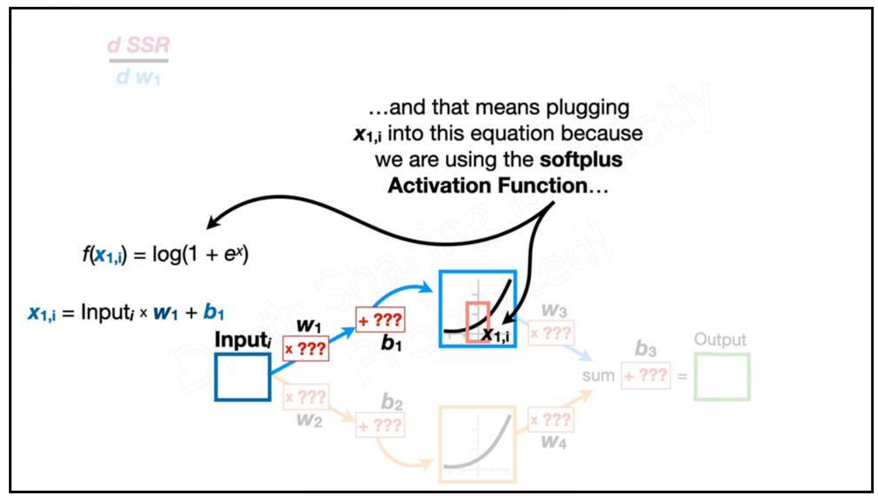



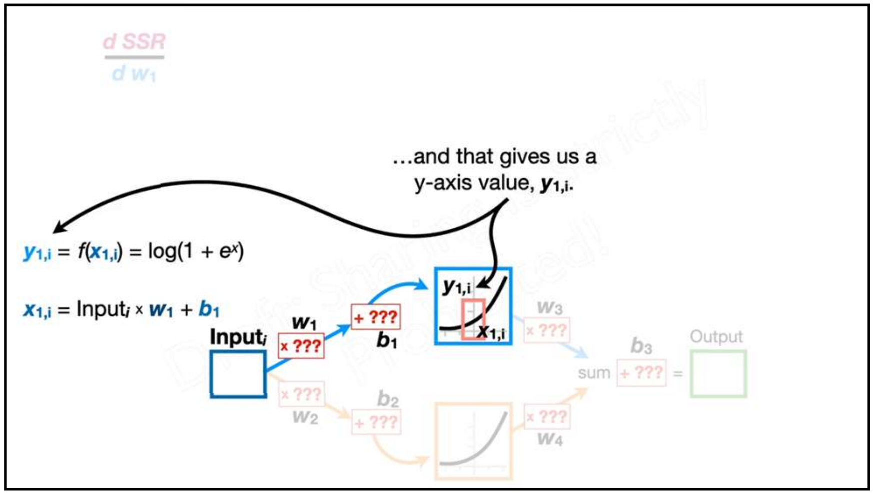
Remember, the **Neural Network** starts by multiplying $Input_i$ by $w_1$…
$Input_i \times w_1$
$Input_i$
$w_1$
x ???


d SSR / d w1
…then it adds the Bias b1…
Inputi × w1 + b1
Inputi
w1
× ???
+ ???
b1
w3
× ???
b3
sum + ??? =
Output
× ???
w2
b2
+ ???
× ???
w4

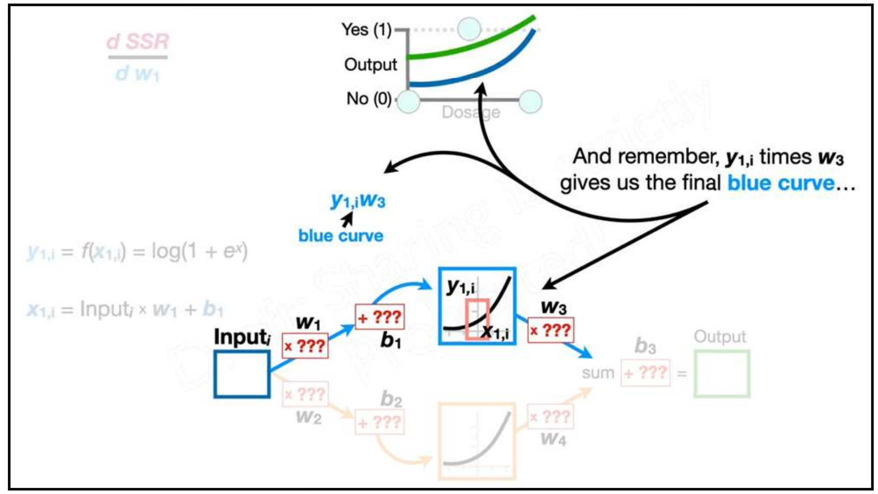

d SSR / d w1
…and that gives us an x-axis coordinate for the Activation Function that we call x1,i.
x1,i = Inputi × w1 + b1
Inputi
w1
× ???
+ ???
b1
x1,i
w3
× ???
b3
sum + ??? =
Output
× ???
w2
b2
+ ???
× ???
w4


d SSR / d w1
We then plug x1,i into the Activation Function…
f(x1,i)
x1,i = Inputi × w1 + b1
Inputi
w1
× ???
+ ???
b1
x1,i
w3
× ???
b3
sum + ??? =
Output
× ???
w2
b2
+ ???
× ???
w4

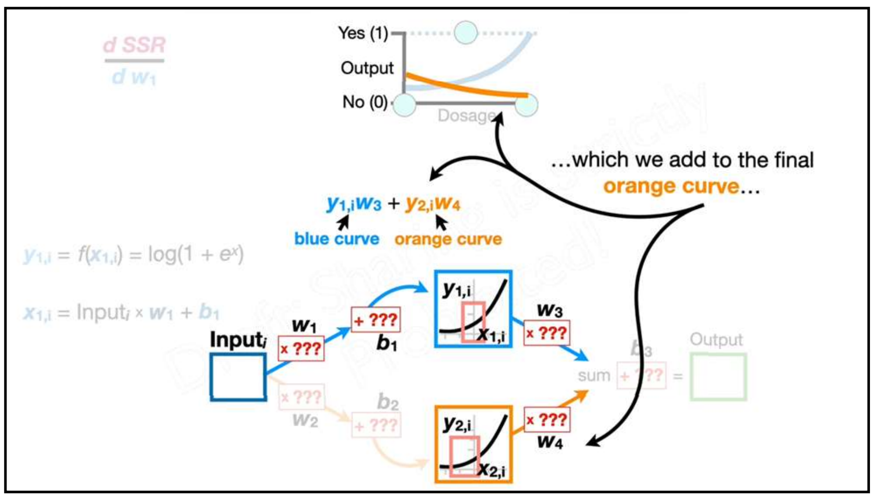

d SSR / d w1
…and that means plugging x1,i into this equation because we are using the softplus Activation Function…
f(x1,i) = log(1 + e^x)
x1,i = Inputi × w1 + b1
Inputi
w1
× ???
+ ???
b1
x1,i
w3
× ???
b3
sum + ??? =
Output
× ???
w2
b2
+ ???
× ???
w4



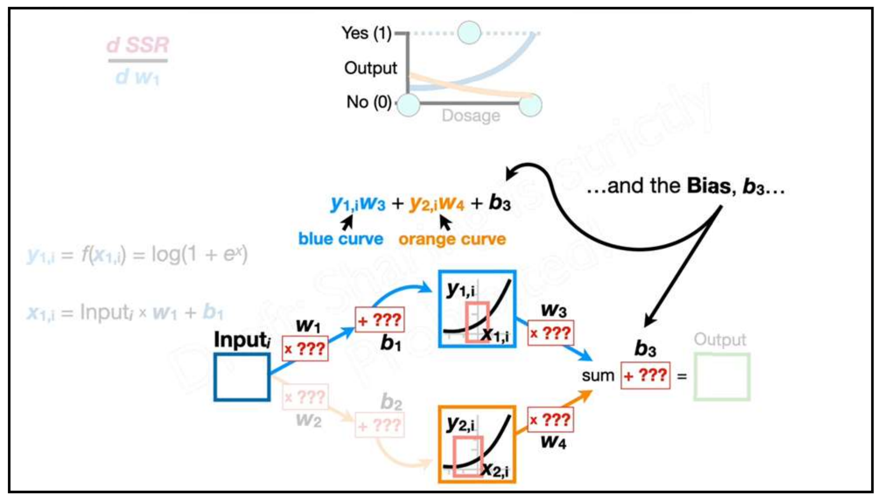
d SSR
d w1
…and that gives us a y-axis value, y1,i.
y1,i = f(x1,i) = log(1 + e^x)
x1,i = Inputi × w1 + b1
Inputi
w1
x ???
+ ???
b1
y1,i
x1,i
w3
x ???
b3
sum
+ ???
=
Output
x ???
w2
b2
+ ???
x ???
w4

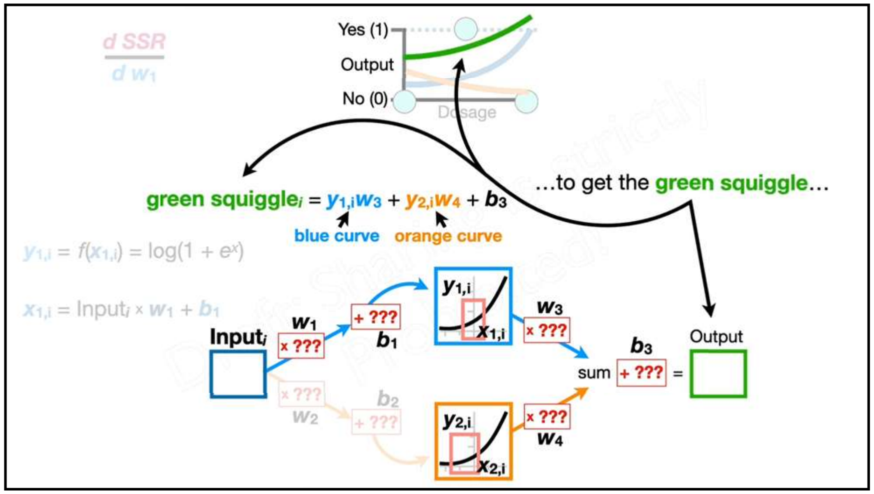
d SSR
d w1
Yes (1)
Output
No (0)
Dosage
And remember, y1,i times w3 gives us the final blue curve…
y1,iw3
blue curve
y1,i = f(x1,i) = log(1 + e^x)
x1,i = Inputi × w1 + b1
Inputi
w1
x ???
+ ???
b1
y1,i
x1,i
w3
x ???
b3
sum
+ ???
=
Output
x ???
w2
b2
+ ???
x ???
w4

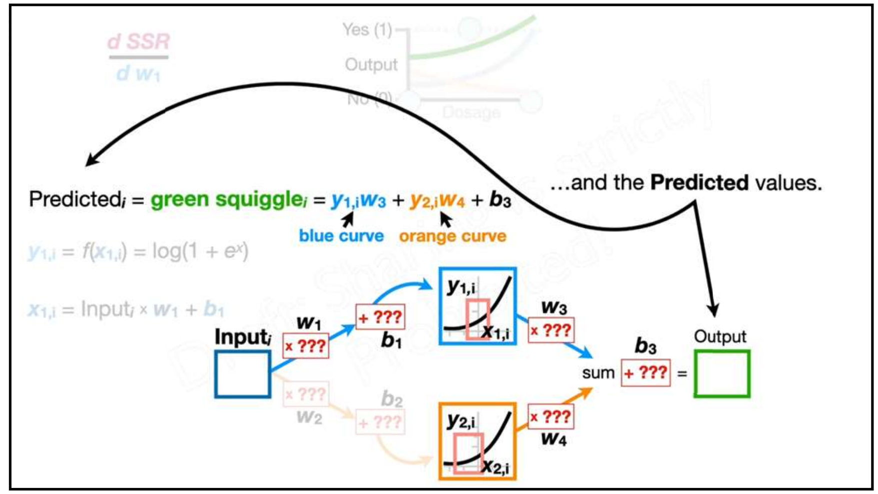



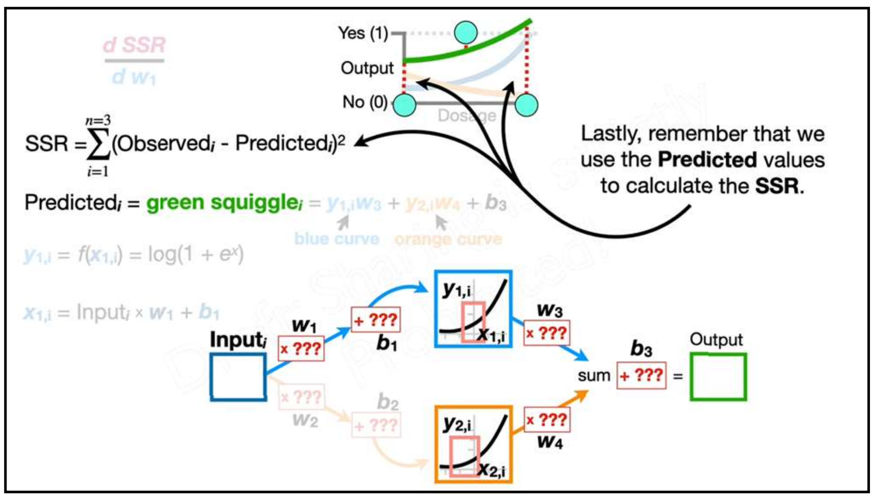
d SSR
d w1
Yes (1)
Output
No (0)
Dosage
…which we add to the final orange curve…
y1,iw3 + y2,iw4
blue curve
orange curve
y1,i = f(x1,i) = log(1 + e^x)
x1,i = Inputi × w1 + b1
Inputi
w1
x ???
+ ???
b1
y1,i
x1,i
w3
x ???
b3
sum
Output
x ???
w2
b2
+ ???
y2,i
x2,i
x ???
w4

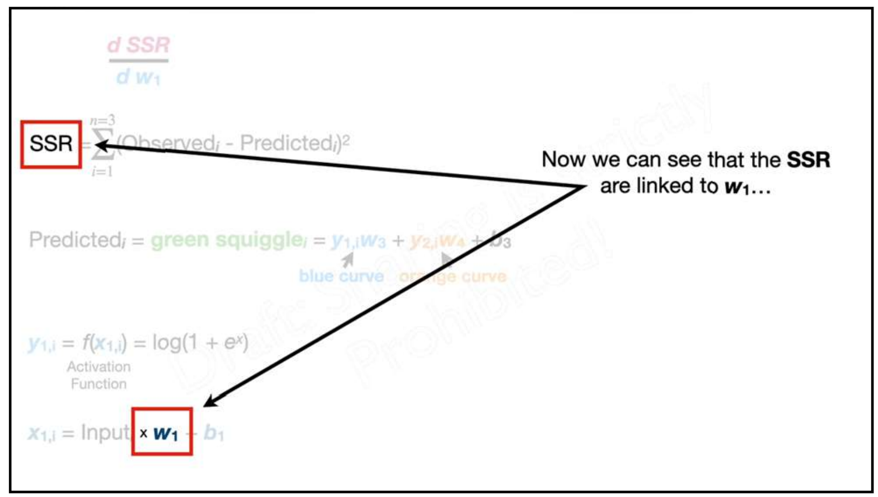



d SSR
d w1
Yes (1)
Output
No (0)
Dosage
…and the Bias, b3…
y1,iw3 + y2,iw4 + b3
blue curve
orange curve
y1,i = f(x1,i) = log(1 + e^x)
x1,i = Inputi × w1 + b1
Inputi
w1
x ???
+ ???
b1
y1,i
x1,i
w3
x ???
b3
sum
+ ???
=
Output
x ???
w2
b2
+ ???
y2,i
x2,i
x ???
w4


Yes (1)
Output
No (0)
Dosage
green squiggle$_i$ = $y_{1,i}w_3$ + $y_{2,i}w_4$ + $b_3$
blue curve
orange curve
…to get the green squiggle…
Input$_i$
$w_1$
× ???
+ ???
$b_1$
$y_{1,i}$
$x_{1,i}$
$w_3$
× ???
$b_3$
sum + ??? =
Output
$y_{2,i}$
$x_{2,i}$
× ???
$w_4$


Predicted$_i$ = green squiggle$_i$ = $y_{1,i}w_3$ + $y_{2,i}w_4$ + $b_3$
blue curve
orange curve
…and the **Predicted** values.
Input$_i$
$w_1$
× ???
+ ???
$b_1$
$y_{1,i}$
$x_{1,i}$
$w_3$
× ???
$b_3$
sum + ??? =
Output
$y_{2,i}$
$x_{2,i}$
× ???
$w_4$


Yes (1)
Output
No (0)
Dosage
SSR = $\sum_{i=1}^{n=3}$(Observed$_i$ - Predicted$_i$)$^2$
Predicted$_i$ = green squiggle$_i$
Lastly, remember that we use the **Predicted** values to calculate the **SSR**.
Input$_i$
$w_1$
× ???
+ ???
$b_1$
$y_{1,i}$
$x_{1,i}$
$w_3$
× ???
$b_3$
sum + ??? =
Output
$y_{2,i}$
$x_{2,i}$
× ???
$w_4$


SSR
Now we can see that the **SSR** are linked to $w_1$…
× $w_1$

$\frac{d\ SSR}{d\ w_1}$

$\text{SSR} = \sum_{i=1}^{n=3}(\text{Observed}_i - \boxed{\text{Predicted}_i})^2$

$\boxed{\text{Predicted}_i} = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$

blue curve orange curve

…first, by the **Predicted** values…

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

Activation Function

$x_{1,i} = \text{Input}_i \times w_1 + b_1$

$\frac{d\ SSR}{d\ w_1}$

$\text{SSR} = \sum_{i=1}^{n=3}(\text{Observed}_i - \boxed{\text{Predicted}_i})^2$

$\boxed{\text{Predicted}_i} = \text{green squiggle}_i = \boxed{y_{1,i}}w_3 + y_{2,i}w_4 + b_3$

blue curve orange curve

…then $y_{1,i}$, the y-axis values that come from the **Activation Function…**

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

Activation Function

$x_{1,i} = \text{Input}_i \times w_1 + b_1$

$\frac{d\ SSR}{d\ w_1}$

$\text{SSR} = \sum_{i=1}^{n=3}(\text{Observed}_i - \boxed{\text{Predicted}_i})^2$

$\boxed{\text{Predicted}_i} = \text{green squiggle}_i = \boxed{y_{1,i}}w_3 + y_{2,i}w_4 + b_3$

blue curve orange curve

…and lastly, $x_{1,i}$, the x-axis values that are input for the **Activation Function.**

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

Activation Function

$\boxed{x_{1,i}} = \text{Input}_i \times w_1 + b_1$

$\frac{d\ SSR}{d\ w_1}$

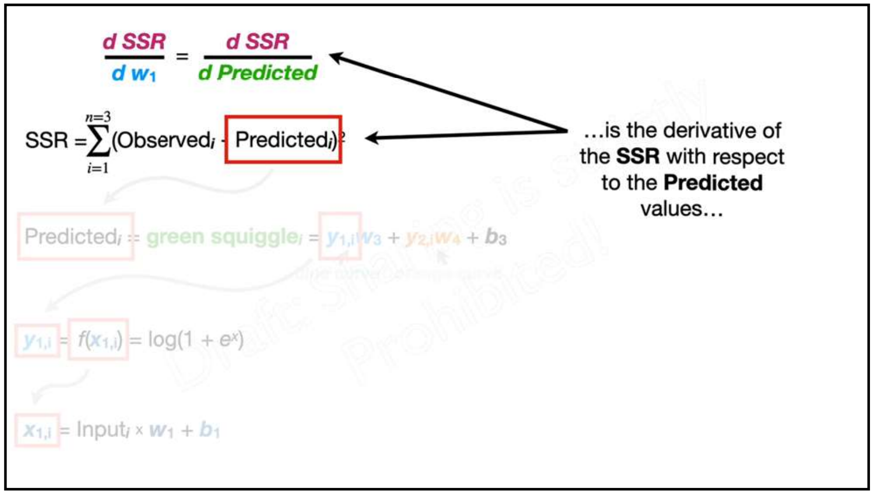
Thus, **The Chain Rule** tells us that the derivative of the **SSR** with respect to $w_1$…

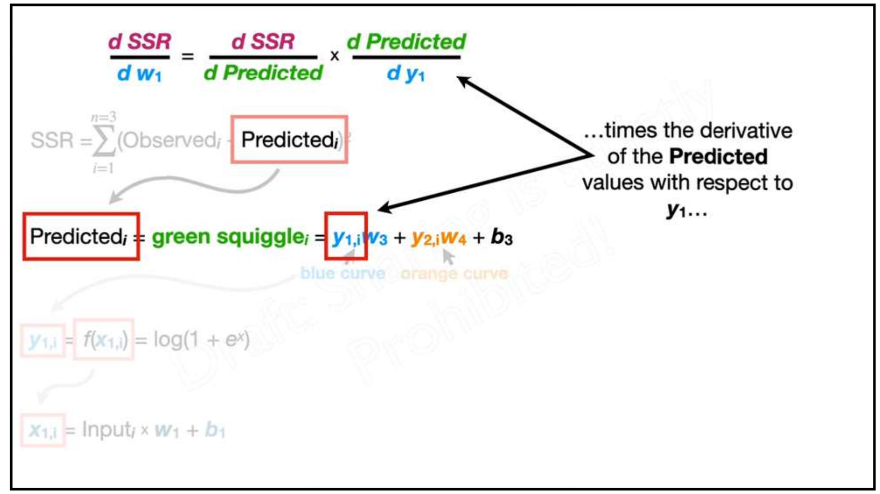
$\text{SSR} = \sum_{i=1}^{n=3}(\text{Observed}_i - \boxed{\text{Predicted}_i})^2$

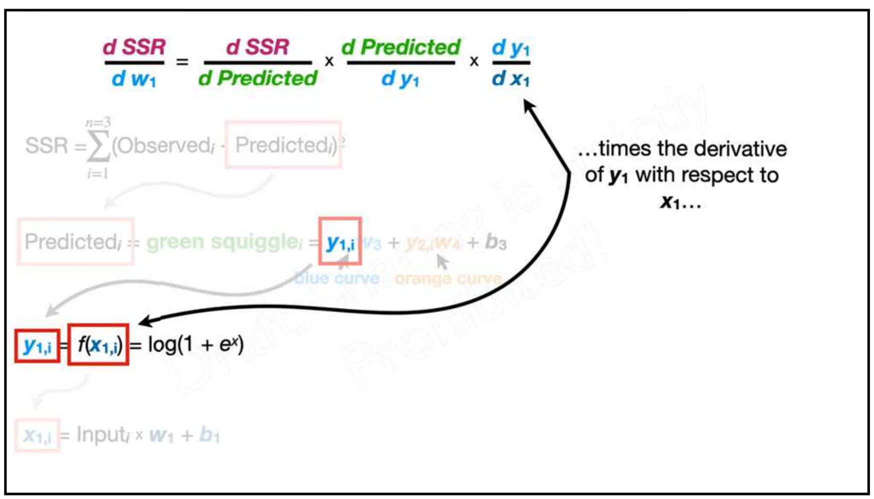
$\boxed{\text{Predicted}_i} = \text{green squiggle}_i = \boxed{y_{1,i}}w_3 + y_{2,i}w_4 + b_3$

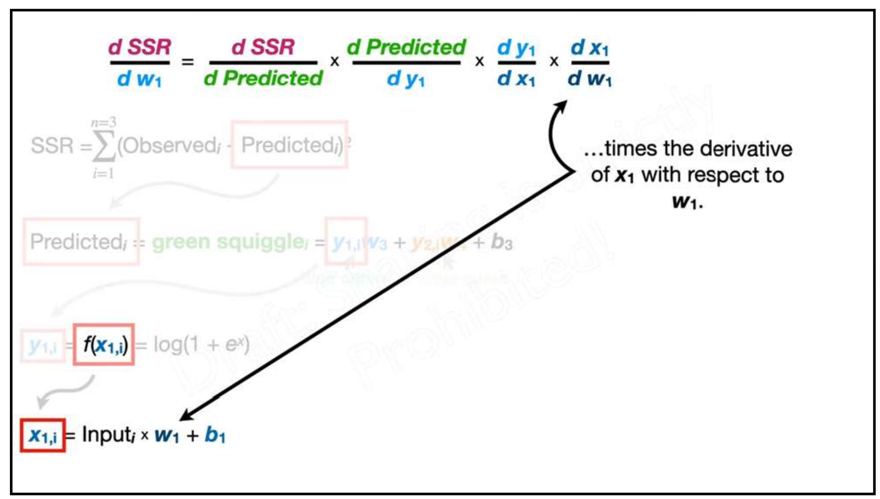
$\boxed{y_{1,i}} = \boxed{f(x_{1,i})} = \log(1 + e^x)$

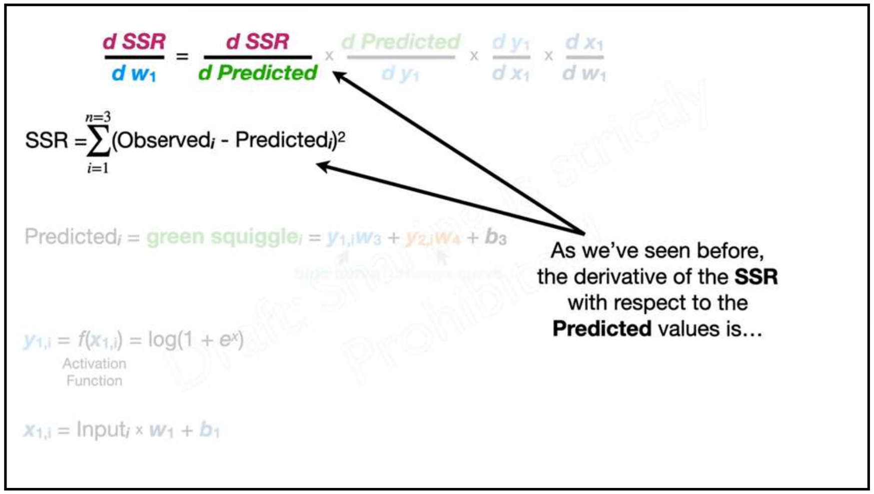
$\boxed{x_{1,i}} = \text{Input}_i \times w_1 + b_1$

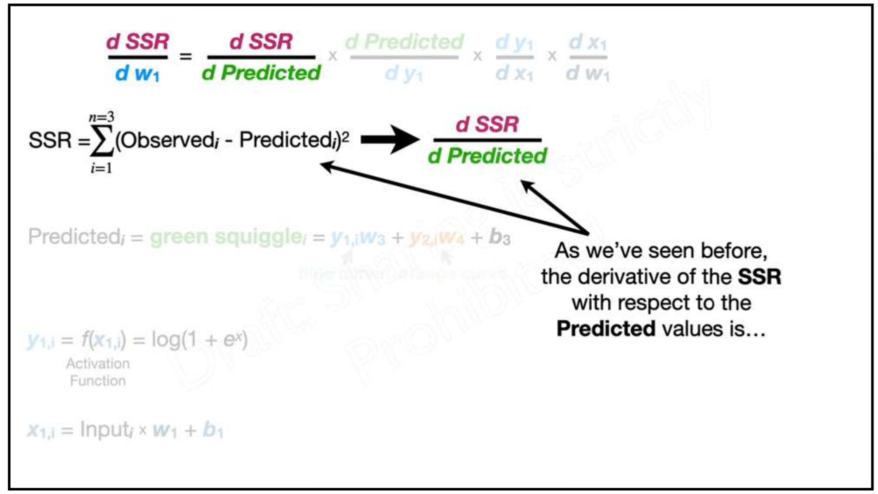

d SSR / d w1 = d SSR / d Predicted
SSR = ∑ i=1 n=3 (Observedi - Predictedi)²
…is the derivative of the SSR with respect to the Predicted values…
Predictedi = green squigglei = y1,iw3 + y2,iw4 + b3
y1,i = f(x1,i) = log(1 + e^x)
x1,i = Inputi × w1 + b1

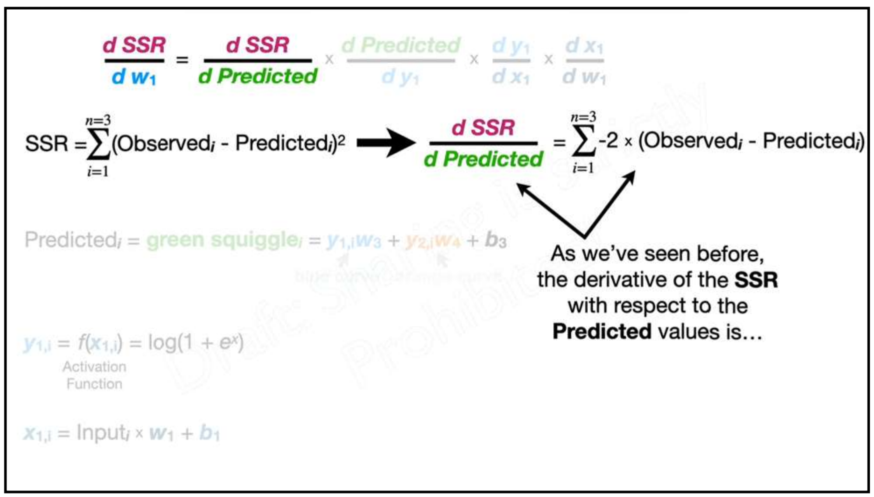

d SSR / d w1 = d SSR / d Predicted x d Predicted / d y1
SSR = ∑ i=1 n=3 (Observedi - Predictedi)²
…times the derivative of the Predicted values with respect to y1…
Predictedi = green squigglei = y1,iw3 + y2,iw4 + b3
blue curve
orange curve
y1,i = f(x1,i) = log(1 + e^x)
x1,i = Inputi × w1 + b1

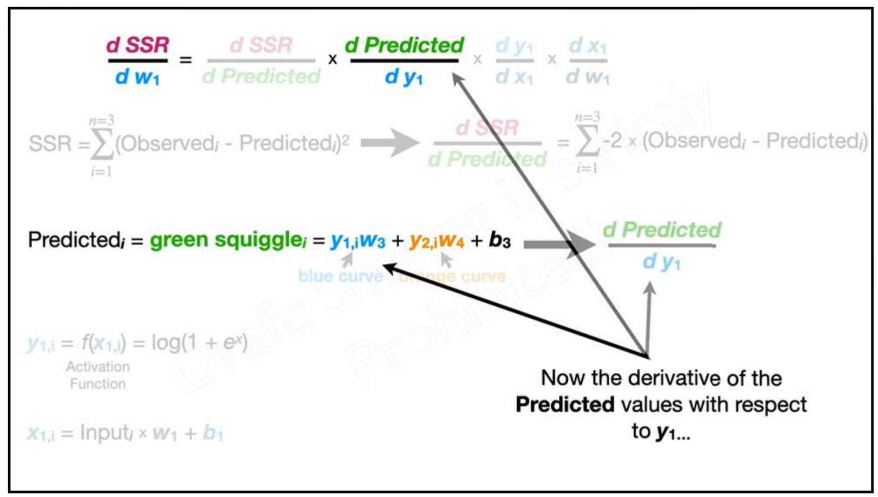

d SSR / d w1 = d SSR / d Predicted x d Predicted / d y1 x d y1 / d x1
SSR = ∑ i=1 n=3 (Observedi - Predictedi)²
…times the derivative of y1 with respect to x1…
Predictedi = green squigglei = y1,iw3 + y2,iw4 + b3
blue curve
orange curve
y1,i = f(x1,i) = log(1 + e^x)
x1,i = Inputi × w1 + b1


d SSR / d w1 = d SSR / d Predicted x d Predicted / d y1 x d y1 / d x1 x d x1 / d w1
SSR = ∑ i=1 n=3 (Observedi - Predictedi)²
…times the derivative of x1 with respect to w1.
Predictedi = green squigglei = y1,iw3 + y2,iw4 + b3
y1,i = f(x1,i) = log(1 + e^x)
x1,i = Inputi × w1 + b1


$\frac{d\,SSR}{d\,w_1} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,y_1} \times \frac{d\,y_1}{d\,x_1} \times \frac{d\,x_1}{d\,w_1}$
$SSR = \sum_{i=1}^{n=3}(Observed_i - Predicted_i)^2$
$Predicted_i$ = green $squiggle_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$
blue curve orange curve
$y_{1,i} = f(x_{1,i}) = \log(1 + e^x)$
Activation Function
$x_{1,i} = Input_i \times w_1 + b_1$
As we've seen before, the derivative of the **SSR** with respect to the **Predicted** values is…


$\frac{d\,SSR}{d\,w_1} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,y_1} \times \frac{d\,y_1}{d\,x_1} \times \frac{d\,x_1}{d\,w_1}$
$SSR = \sum_{i=1}^{n=3}(Observed_i - Predicted_i)^2 \rightarrow \frac{d\,SSR}{d\,Predicted}$
$Predicted_i$ = green $squiggle_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$
blue curve orange curve
$y_{1,i} = f(x_{1,i}) = \log(1 + e^x)$
Activation Function
$x_{1,i} = Input_i \times w_1 + b_1$
As we've seen before, the derivative of the **SSR** with respect to the **Predicted** values is…


$\frac{d\,SSR}{d\,w_1} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,y_1} \times \frac{d\,y_1}{d\,x_1} \times \frac{d\,x_1}{d\,w_1}$
$SSR = \sum_{i=1}^{n=3}(Observed_i - Predicted_i)^2 \rightarrow \frac{d\,SSR}{d\,Predicted} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i)$
$Predicted_i$ = green $squiggle_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$
blue curve orange curve
$y_{1,i} = f(x_{1,i}) = \log(1 + e^x)$
Activation Function
$x_{1,i} = Input_i \times w_1 + b_1$
As we've seen before, the derivative of the **SSR** with respect to the **Predicted** values is…


$\frac{d\,SSR}{d\,w_1} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,y_1} \times \frac{d\,y_1}{d\,x_1} \times \frac{d\,x_1}{d\,w_1}$
$SSR = \sum_{i=1}^{n=3}(Observed_i - Predicted_i)^2 \rightarrow \frac{d\,SSR}{d\,Predicted} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i)$
$Predicted_i$ = green $squiggle_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3 \rightarrow \frac{d\,Predicted}{d\,y_1}$
blue curve orange curve
$y_{1,i} = f(x_{1,i}) = \log(1 + e^x)$
Activation Function
$x_{1,i} = Input_i \times w_1 + b_1$
Now the derivative of the **Predicted** values with respect to $y_1$…

$$\frac{d\,SSR}{d\,w_1} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,y_1} \times \frac{d\,y_1}{d\,x_1} \times \frac{d\,x_1}{d\,w_1}$$

$$SSR = \sum_{i=1}^{n=3}(Observed_i - Predicted_i)^2 \rightarrow \frac{d\,SSR}{d\,Predicted} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i)$$

$$Predicted_i = \text{green squiggle}_i = \underbrace{y_{1,i}}_{\text{blue curve}}w_3 + \underbrace{y_{2,i}}_{\text{orange curve}}w_4 + b_3 \rightarrow \frac{d\,Predicted}{d\,y_1} = w_3$$

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

Activation Function

$$x_{1,i} = Input_i \times w_1 + b_1$$

So the derivative of the **Predicted** values with respect to $y_1$ is $w_3$.



$$\frac{d\,SSR}{d\,w_1} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,y_1} \times \frac{d\,y_1}{d\,x_1} \times \frac{d\,x_1}{d\,w_1}$$

$$SSR = \sum_{i=1}^{n=3}(Observed_i - Predicted_i)^2 \rightarrow \frac{d\,SSR}{d\,Predicted} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i)$$

$$Predicted_i = \text{green squiggle}_i = \underbrace{y_{1,i}}_{\text{blue curve}}w_3 + \underbrace{y_{2,i}}_{\text{orange curve}}w_4 + b_3 \rightarrow \frac{d\,Predicted}{d\,y_1} = w_3$$

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

Activation Function

$$x_{1,i} = Input_i \times w_1 + b_1$$

Now we need to solve for the derivative of $y_1$ with respect to $x_1$.



$$\frac{d\,SSR}{d\,w_1} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,y_1} \times \frac{d\,y_1}{d\,x_1} \times \frac{d\,x_1}{d\,w_1}$$

$$SSR = \sum_{i=1}^{n=3}(Observed_i - Predicted_i)^2 \rightarrow \frac{d\,SSR}{d\,Predicted} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i)$$

$$Predicted_i = \text{green squiggle}_i = \underbrace{y_{1,i}}_{\text{blue curve}}w_3 + \underbrace{y_{2,i}}_{\text{orange curve}}w_4 + b_3 \rightarrow \frac{d\,Predicted}{d\,y_1} = w_3$$

$$y_{1,i} = f(x_{1,i}) = \log(1 + e^x) \rightarrow \frac{d\,y_1}{d\,x_1}$$

Activation Function

...and the derivative of $e^x$ is $e^x$.

$$x_{1,i} = Input_i \times w_1 + b_1$$

**NOTE:** $\frac{d}{d\,z}\log(z) = \frac{1}{z}$ $\quad \frac{d}{d\,x}e^x = e^x$



$$\frac{d\,SSR}{d\,w_1} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,y_1} \times \frac{d\,y_1}{d\,x_1} \times \frac{d\,x_1}{d\,w_1}$$

$$SSR = \sum_{i=1}^{n=3}(Observed_i - Predicted_i)^2 \rightarrow \frac{d\,SSR}{d\,Predicted} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i)$$

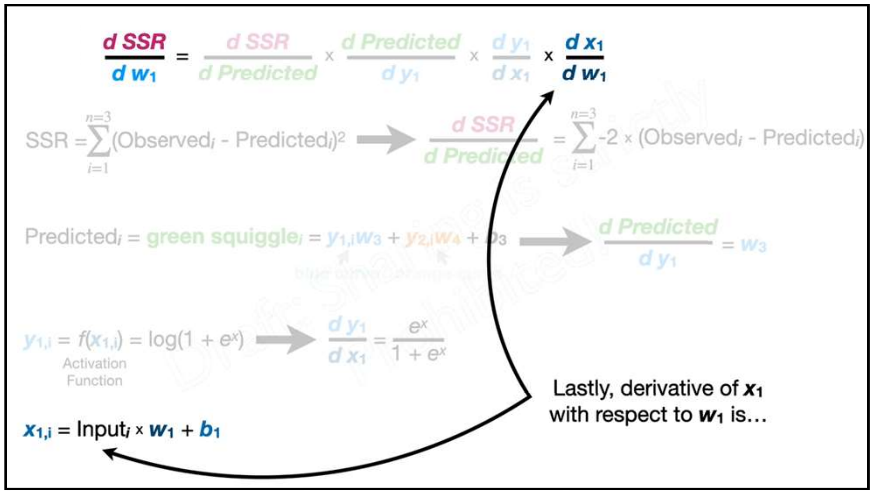
$$Predicted_i = \text{green squiggle}_i = \underbrace{y_{1,i}}_{\text{blue curve}}w_3 + \underbrace{y_{2,i}}_{\text{orange curve}}w_4 + b_3 \rightarrow \frac{d\,Predicted}{d\,y_1} = w_3$$

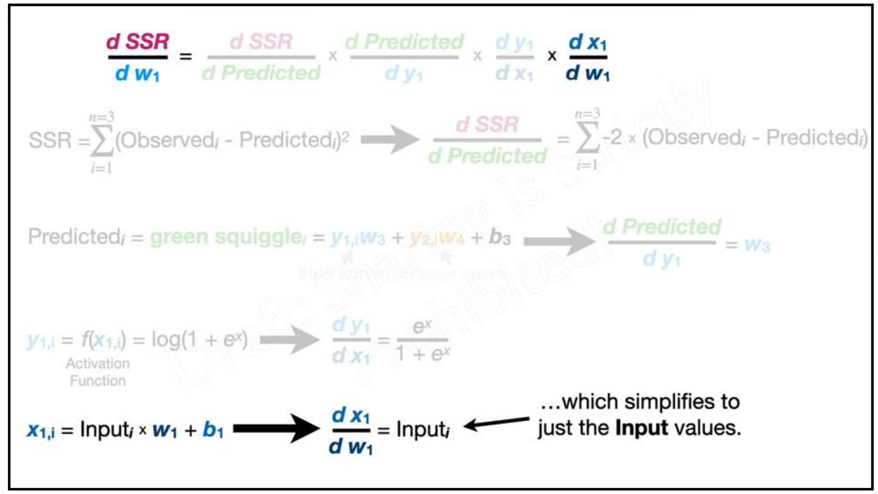
$$y_{1,i} = f(x_{1,i}) = \log(1 + e^x) \rightarrow \boxed{\frac{d\,y_1}{d\,x_1} = \frac{e^x}{1 + e^x}}$$

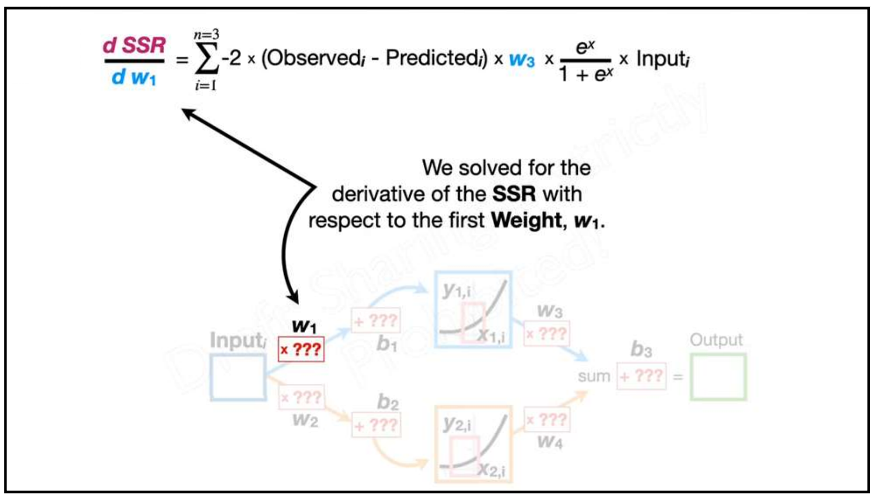
Activation Function

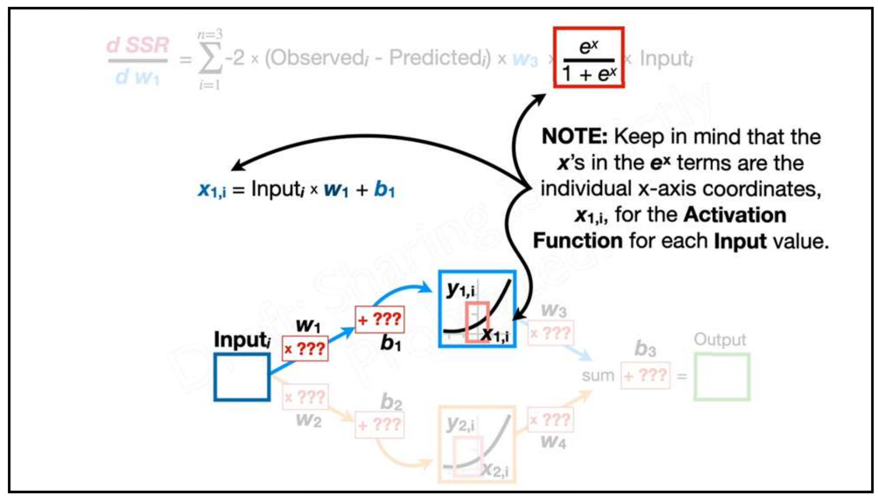
And this whole thing simplifies to…

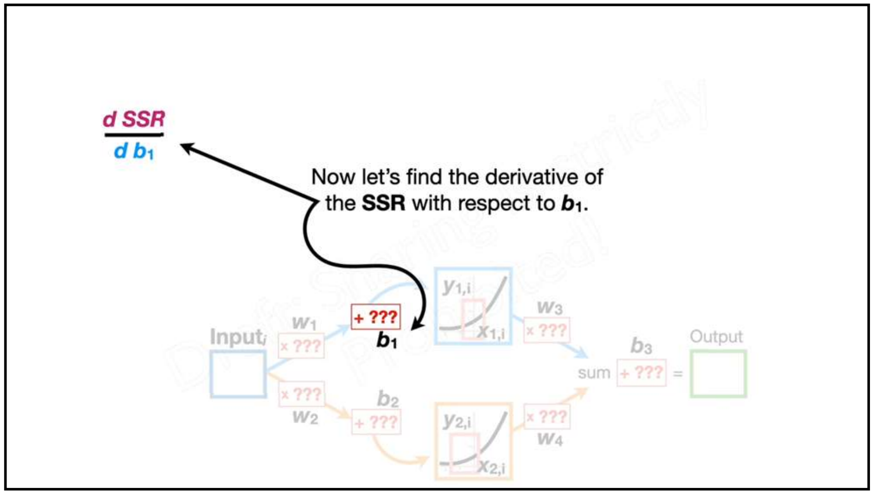
$$x_{1,i} = Input_i \times w_1 + b_1$$

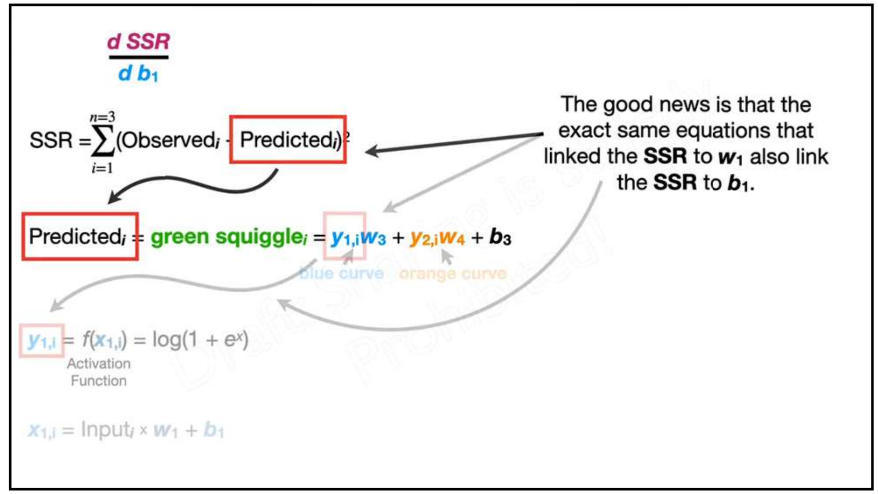

d SSR / d w1 = d SSR / d Predicted x d Predicted / d y1 x d y1 / d x1 x d x1 / d w1
SSR = Σ(Observedi - Predictedi)2 → d SSR / d Predicted = Σ -2 x (Observedi - Predictedi)
Predictedi = green squigglei = y1,iw3 + y2,iw4 + b3 → d Predicted / d y1 = w3
y1,i = f(x1,i) = log(1 + ex) → d y1 / d x1 = ex / (1 + ex)
Activation Function
x1,i = Inputi x w1 + b1
Lastly, derivative of x1 with respect to w1 is…

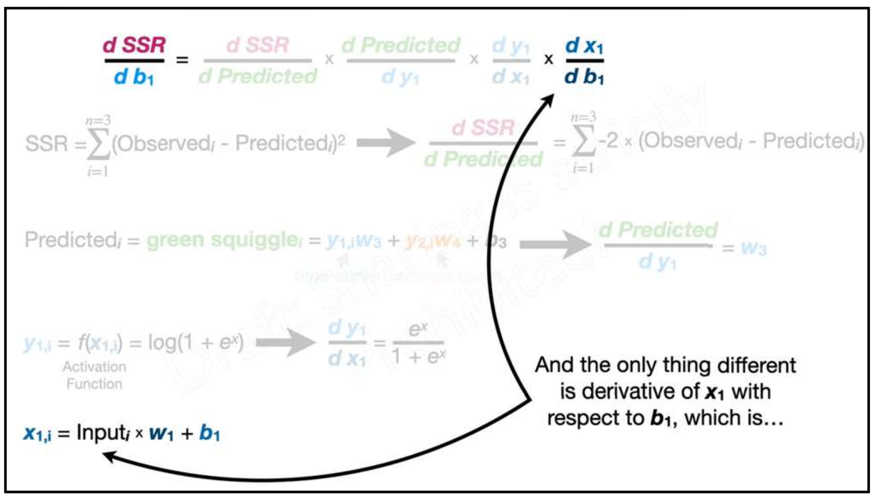

d SSR / d w1 = d SSR / d Predicted x d Predicted / d y1 x d y1 / d x1 x d x1 / d w1
SSR = Σ(Observedi - Predictedi)2 → d SSR / d Predicted = Σ -2 x (Observedi - Predictedi)
Predictedi = green squigglei = y1,iw3 + y2,iw4 + b3 → d Predicted / d y1 = w3
y1,i = f(x1,i) = log(1 + ex) → d y1 / d x1 = ex / (1 + ex)
Activation Function
x1,i = Inputi x w1 + b1 → d x1 / d w1 = Inputi
…which simplifies to just the Input values.

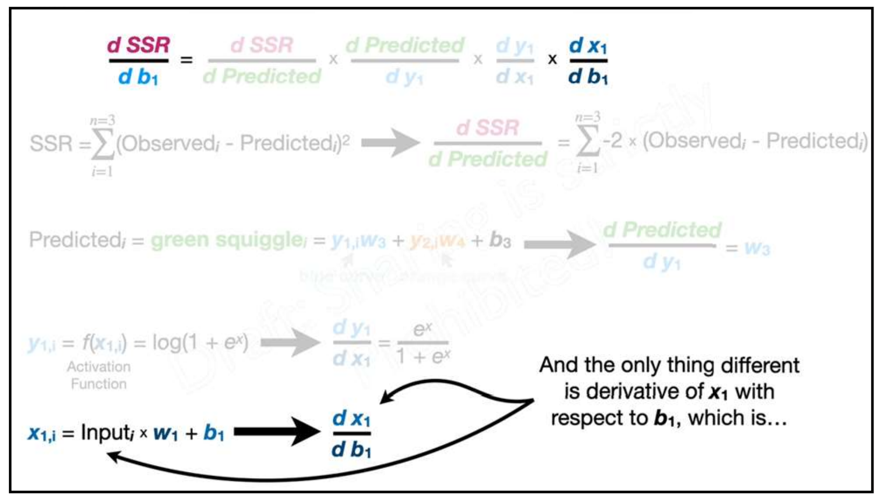

d SSR / d w1 = Σ -2 x (Observedi - Predictedi) x w3 x ex / (1 + ex) x Inputi
We solved for the derivative of the SSR with respect to the first Weight, w1.
Inputi
w1
x ???
+ ???
b1
y1,i
x1,i
w3
x ???
w2
b2
y2,i
x2,i
w4
b3
sum
Output

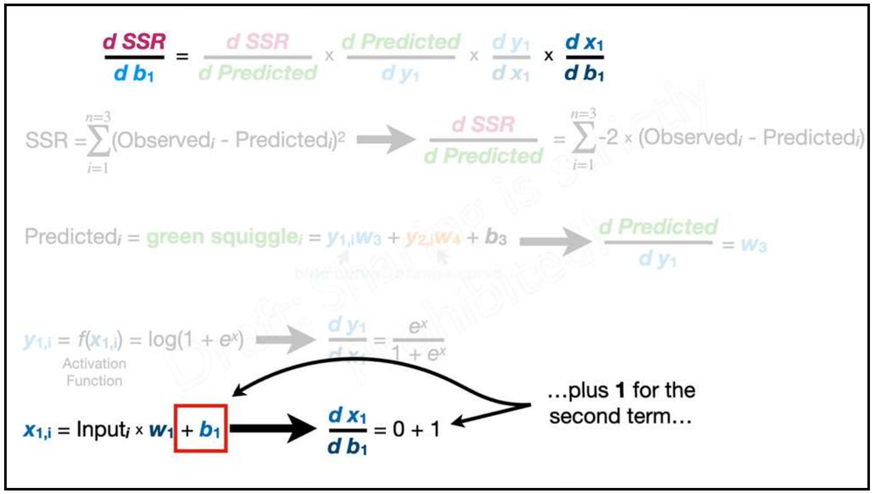

d SSR / d w1 = Σ -2 x (Observedi - Predictedi) x w3 x ex / (1 + ex) x Inputi
x1,i = Inputi x w1 + b1
NOTE: Keep in mind that the x's in the ex terms are the individual x-axis coordinates, x1,i, for the Activation Function for each Input value.
Inputi
w1
x ???
+ ???
b1
y1,i
x1,i
w3
w2
b2
y2,i
x2,i
w4
b3
sum
Output

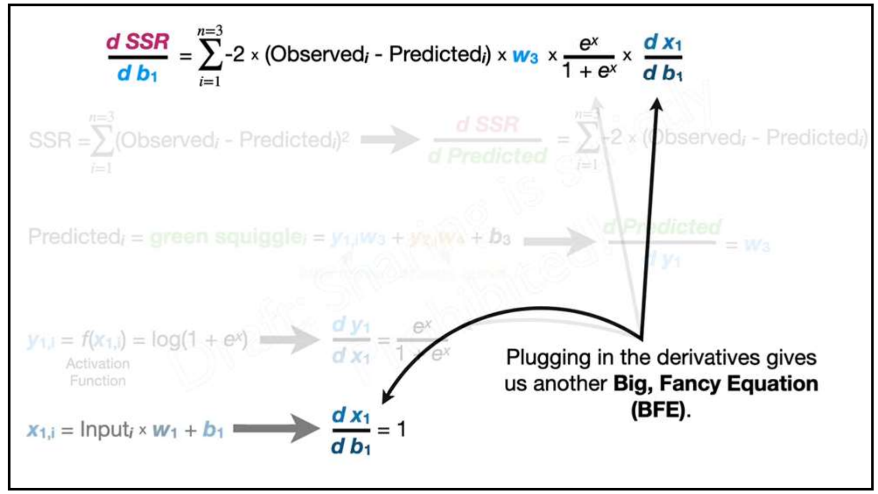

d SSR / d b₁
Now let's find the derivative of the SSR with respect to b₁.
Inputᵢ
w₁
x ???
+ ???
b₁
y₁,ᵢ
x₁,ᵢ
w₃
x ???
b₃
sum + ??? =
Output
x ???
w₂
b₂
+ ???
y₂,ᵢ
x₂,ᵢ
x ???
w₄

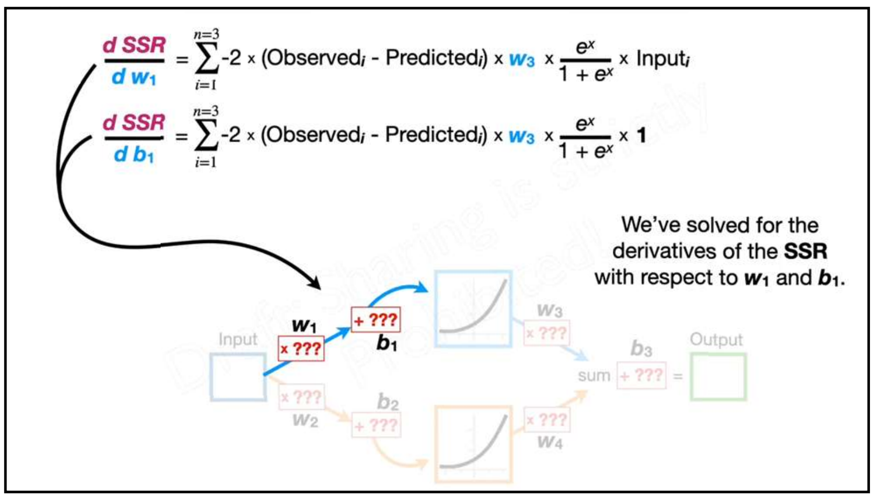

d SSR / d b₁
SSR = Σ_{i=1}^{n=3} (Observedᵢ - Predictedᵢ)²
The good news is that the exact same equations that linked the SSR to w₁ also link the SSR to b₁.
Predictedᵢ = green squiggleᵢ = y₁,ᵢw₃ + y₂,ᵢw₄ + b₃
blue curve
orange curve
y₁,ᵢ = f(x₁,ᵢ) = log(1 + eˣ)
Activation Function
x₁,ᵢ = Inputᵢ × w₁ + b₁

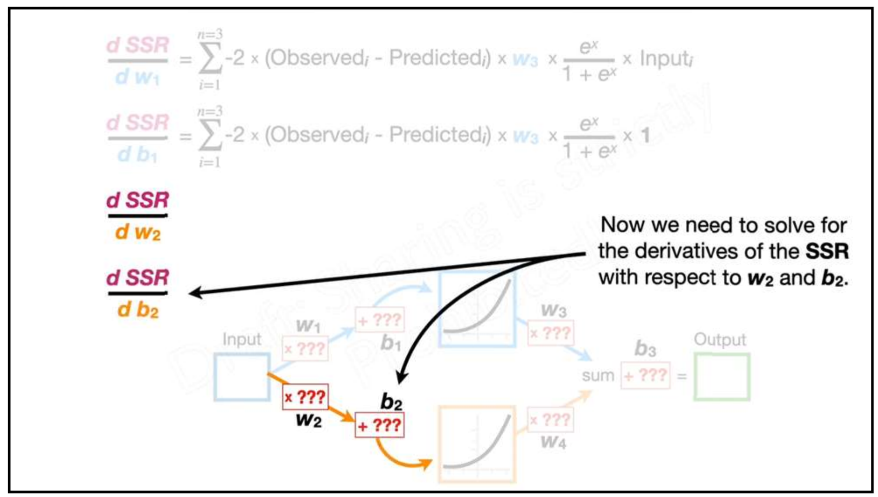

d SSR / d b₁ = d SSR / d Predicted × d Predicted / d y₁ × d y₁ / d x₁ × d x₁ / d b₁
SSR = Σ_{i=1}^{n=3} (Observedᵢ - Predictedᵢ)² → d SSR / d Predicted = Σ_{i=1}^{n=3} -2 × (Observedᵢ - Predictedᵢ)
Predictedᵢ = green squiggleᵢ = y₁,ᵢw₃ + y₂,ᵢw₄ + b₃ → d Predicted / d y₁ = w₃
y₁,ᵢ = f(x₁,ᵢ) = log(1 + eˣ) → d y₁ / d x₁ = eˣ / (1 + eˣ)
Activation Function
And the only thing different is derivative of x₁ with respect to b₁, which is…
x₁,ᵢ = Inputᵢ × w₁ + b₁

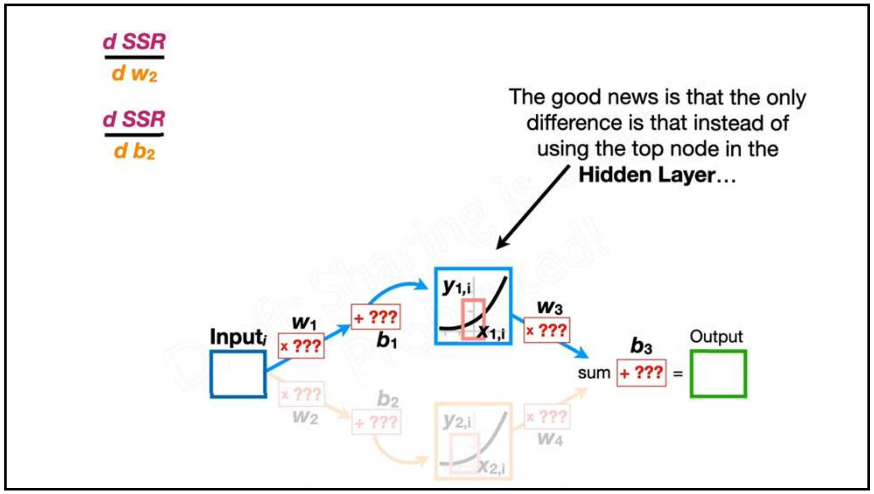

d SSR / d b₁ = d SSR / d Predicted × d Predicted / d y₁ × d y₁ / d x₁ × d x₁ / d b₁
SSR = Σ_{i=1}^{n=3} (Observedᵢ - Predictedᵢ)² → d SSR / d Predicted = Σ_{i=1}^{n=3} -2 × (Observedᵢ - Predictedᵢ)
Predictedᵢ = green squiggleᵢ = y₁,ᵢw₃ + y₂,ᵢw₄ + b₃ → d Predicted / d y₁ = w₃
y₁,ᵢ = f(x₁,ᵢ) = log(1 + eˣ) → d y₁ / d x₁ = eˣ / (1 + eˣ)
Activation Function
And the only thing different is derivative of x₁ with respect to b₁, which is…
x₁,ᵢ = Inputᵢ × w₁ + b₁ → d x₁ / d b₁

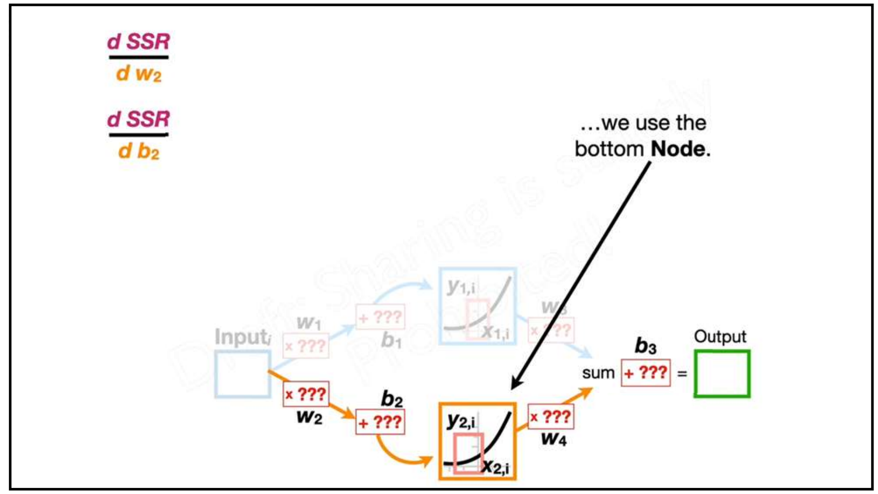

$\frac{d\ SSR}{d\ b_1} = \frac{d\ SSR}{d\ Predicted} \times \frac{d\ Predicted}{d\ y_1} \times \frac{d\ y_1}{d\ x_1} \times \frac{d\ x_1}{d\ b_1}$
$SSR = \sum_{i=1}^{n=3}(\text{Observed}_i - \text{Predicted}_i)^2 \rightarrow \frac{d\ SSR}{d\ Predicted} = \sum_{i=1}^{n=3} -2 \times (\text{Observed}_i - \text{Predicted}_i)$
$\text{Predicted}_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3 \rightarrow \frac{d\ Predicted}{d\ y_1} = w_3$
$y_{1,i} = f(x_{1,i}) = \log(1 + e^x) \rightarrow \frac{d\ y_1}{d\ x_1} = \frac{e^x}{1 + e^x}$
Activation Function
$x_{1,i} = \text{Input}_i \times w_1 + b_1 \rightarrow \frac{d\ x_1}{d\ b_1} = 0 + 1$
…plus 1 for the second term…

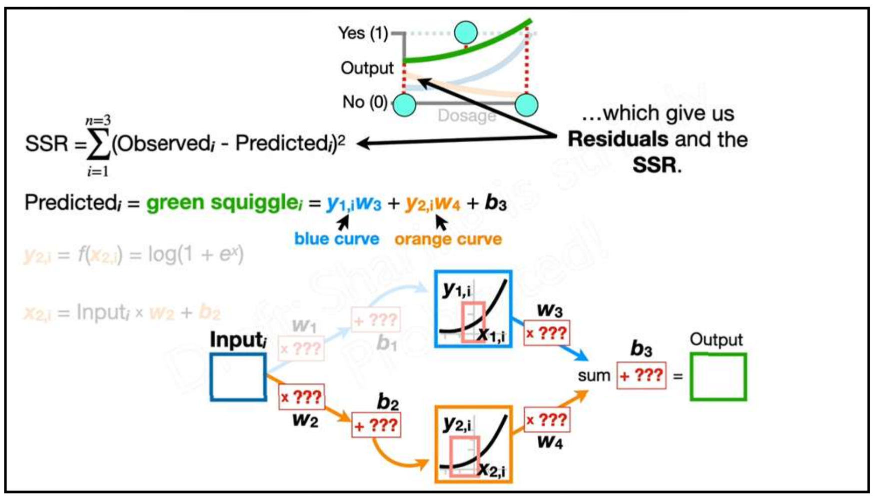

$\frac{d\ SSR}{d\ b_1} = \sum_{i=1}^{n=3} -2 \times (\text{Observed}_i - \text{Predicted}_i) \times w_3 \times \frac{e^x}{1 + e^x} \times \frac{d\ x_1}{d\ b_1}$
$SSR = \sum_{i=1}^{n=3}(\text{Observed}_i - \text{Predicted}_i)^2 \rightarrow \frac{d\ SSR}{d\ Predicted} = \sum_{i=1}^{n=3} -2 \times (\text{Observed}_i - \text{Predicted}_i)$
$\text{Predicted}_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3 \rightarrow \frac{d\ Predicted}{d\ y_1} = w_3$
$y_{1,i} = f(x_{1,i}) = \log(1 + e^x) \rightarrow \frac{d\ y_1}{d\ x_1} = \frac{e^x}{1 + e^x}$
Activation Function
$x_{1,i} = \text{Input}_i \times w_1 + b_1 \rightarrow \frac{d\ x_1}{d\ b_1} = 1$
Plugging in the derivatives gives us another Big, Fancy Equation (BFE).

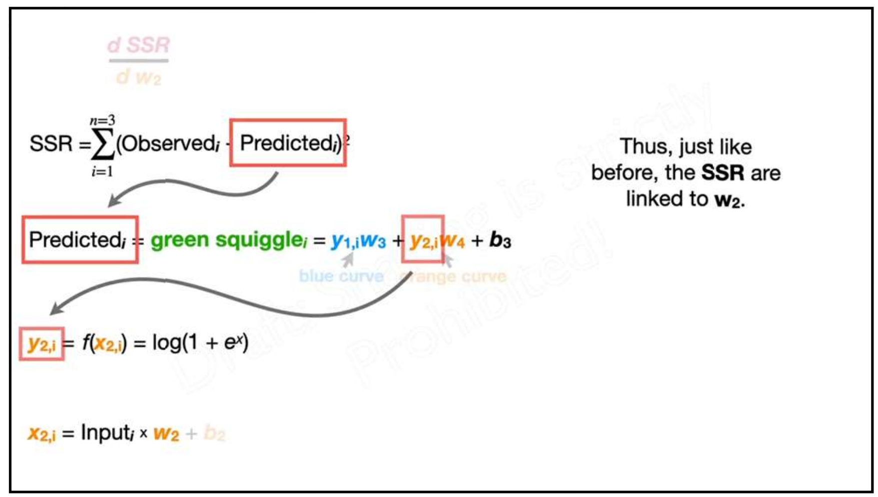

$\frac{d\ SSR}{d\ w_1} = \sum_{i=1}^{n=3} -2 \times (\text{Observed}_i - \text{Predicted}_i) \times w_3 \times \frac{e^x}{1 + e^x} \times \text{Input}_i$
$\frac{d\ SSR}{d\ b_1} = \sum_{i=1}^{n=3} -2 \times (\text{Observed}_i - \text{Predicted}_i) \times w_3 \times \frac{e^x}{1 + e^x} \times 1$
We've solved for the derivatives of the SSR with respect to $w_1$ and $b_1$.
Input
$w_1$
x ???
+ ???
$b_1$
$w_3$
x ???
$b_3$
Output
sum + ??? =
x ???
$w_2$
$b_2$
+ ???
x ???
$w_4$


$\frac{d\ SSR}{d\ w_1} = \sum_{i=1}^{n=3} -2 \times (\text{Observed}_i - \text{Predicted}_i) \times w_3 \times \frac{e^x}{1 + e^x} \times \text{Input}_i$
$\frac{d\ SSR}{d\ b_1} = \sum_{i=1}^{n=3} -2 \times (\text{Observed}_i - \text{Predicted}_i) \times w_3 \times \frac{e^x}{1 + e^x} \times 1$
$\frac{d\ SSR}{d\ w_2}$
$\frac{d\ SSR}{d\ b_2}$
Now we need to solve for the derivatives of the SSR with respect to $w_2$ and $b_2$.
Input
$w_1$
x ???
+ ???
$b_1$
$w_3$
x ???
$b_3$
Output
sum + ??? =
x ???
$w_2$
$b_2$
+ ???
x ???
$w_4$


d SSR / d w2
d SSR / d b2
The good news is that the only difference is that instead of using the top node in the Hidden Layer…
Inputi
w1
× ???
+ ???
b1
y1,i
x1,i
w3
× ???
b3
sum
+ ???
=
Output
w2
b2
y2,i
x2,i
w4


d SSR / d w2
d SSR / d b2
…we use the bottom Node.
Inputi
w1
b1
y1,i
x1,i
× ???
w2
b2
+ ???
y2,i
x2,i
w4
b3
sum
+ ???
=
Output


Yes (1)
Output
No (0)
Dosage
…which give us Residuals and the SSR.
SSR = Σ_{i=1}^{n=3} (Observedi - Predictedi)²
Predictedi = green squigglei = y1,iw3 + y2,iw4 + b3
blue curve
orange curve
y2,i = f(x2,i) = log(1 + e^x)
x2,i = Inputi × w2 + b2
Inputi
y1,i
x1,i
w3
× ???
b3
sum
+ ???
=
Output
w2
b2
y2,i
x2,i
w4


d SSR / d w2
SSR = Σ_{i=1}^{n=3} (Observedi - Predictedi)²
Thus, just like before, the SSR are linked to w2.
Predictedi = green squigglei = y1,iw3 + y2,iw4 + b3
blue curve
orange curve
y2,i = f(x2,i) = log(1 + e^x)
x2,i = Inputi × w2 + b2

$$\frac{d\,SSR}{d\,w_2} = \frac{d\,SSR}{d\,Predicted} \times \frac{d\,Predicted}{d\,y_2} \times \frac{d\,y_2}{d\,x_2} \times \frac{d\,x_2}{d\,w_2}$$

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

$$Predicted_i = \text{green squiggle}_i = y_{1,i}w_3 + y_{2,i}w_4 + b_3$$

$$y_{2,i} = f(x_{2,i}) = \log(1 + e^x)$$

$$x_{2,i} = Input_i \times w_2 + b_2$$

And we use **The Chain Rule** to derive the derivative of the **SSR** with respect to $\mathbf{w_2}$.



$$\frac{d\,SSR}{d\,w_2} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times w_4 \times \frac{e^x}{1 + e^x} \times Input_i$$

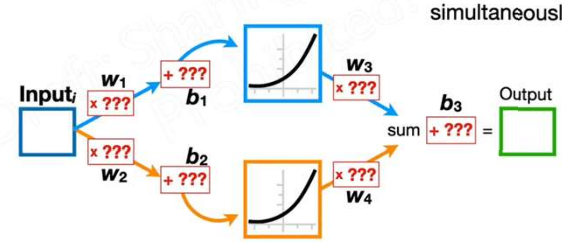
$$\frac{d\,SSR}{d\,b_2} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times w_4 \times \frac{e^x}{1 + e^x} \times \mathbf{1}$$

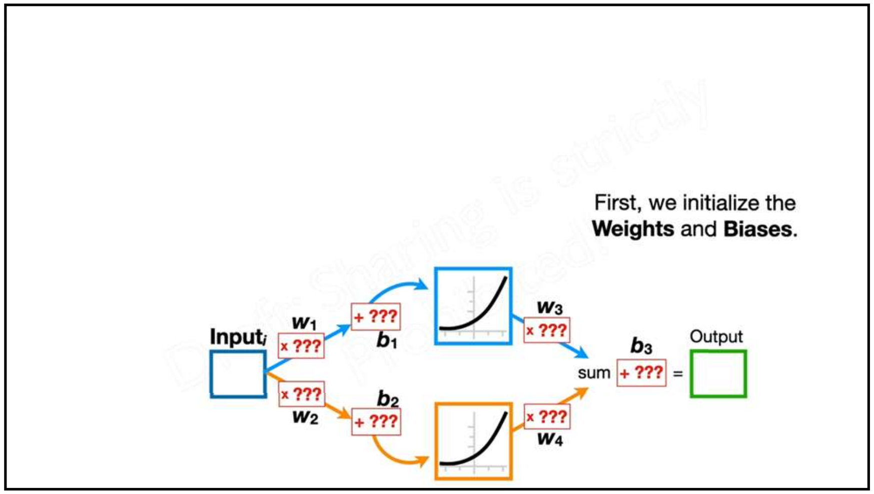
$$\frac{d\,SSR}{d\,w_3} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{1,i}$$

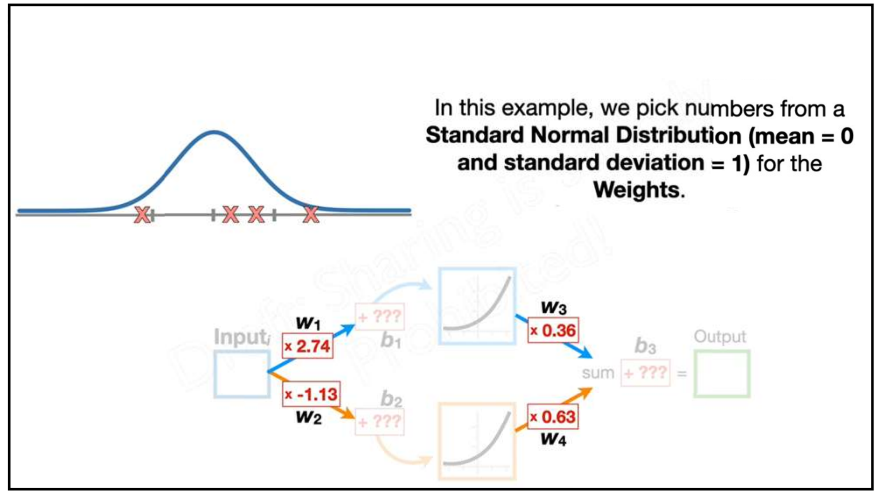
$$\frac{d\,SSR}{d\,w_4} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times y_{2,i}$$

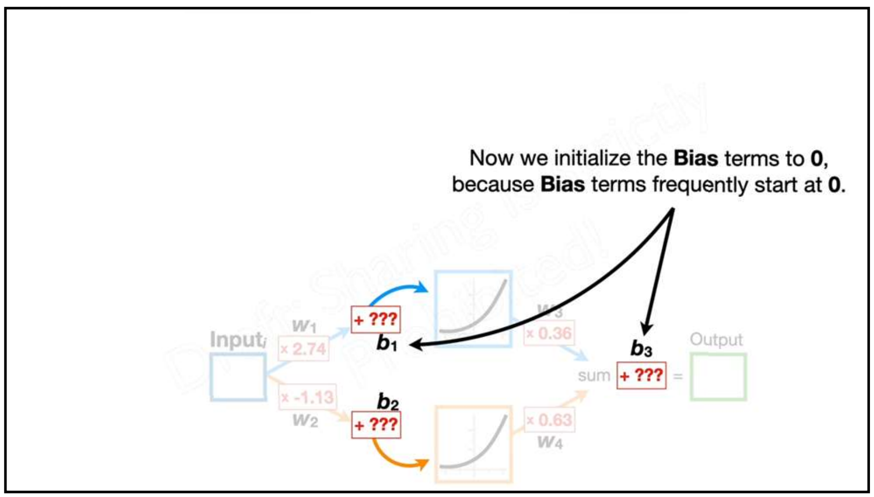
$$\frac{d\,SSR}{d\,b_3} = \sum_{i=1}^{n=3} -2 \times (Observed_i - Predicted_i) \times \mathbf{1}$$

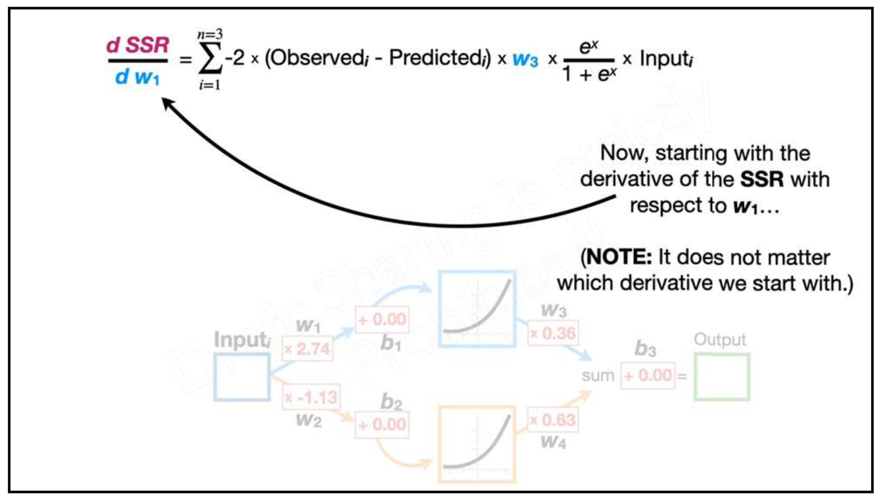
use **Gradient Descent** to optimize all of the parameters simultaneously.

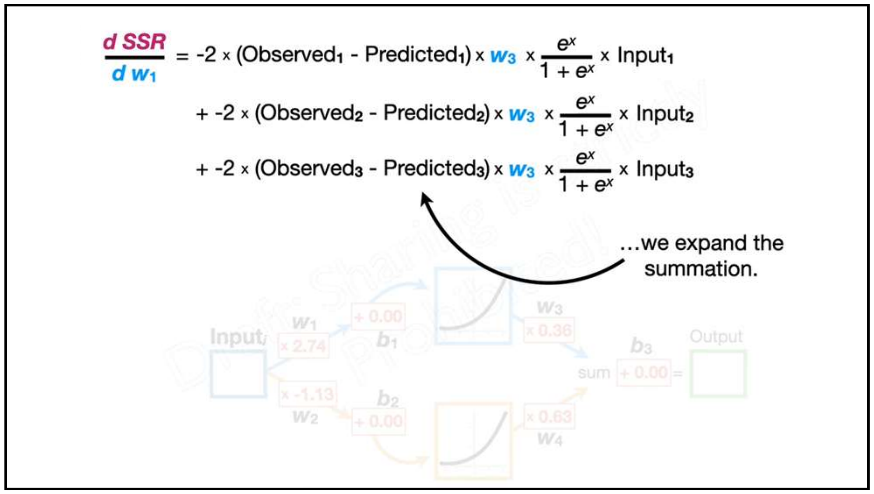


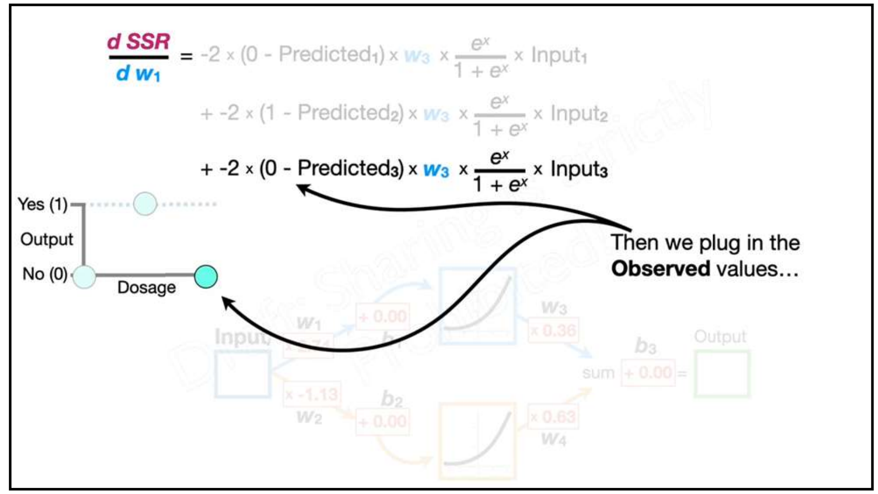

First, we initialize the **Weights** and **Biases**.
Input$_i$
$w_1$
x ???
$w_2$
x ???
+ ???
$b_1$
$b_2$
+ ???
$w_3$
x ???
$w_4$
x ???
$b_3$
sum + ??? =
Output

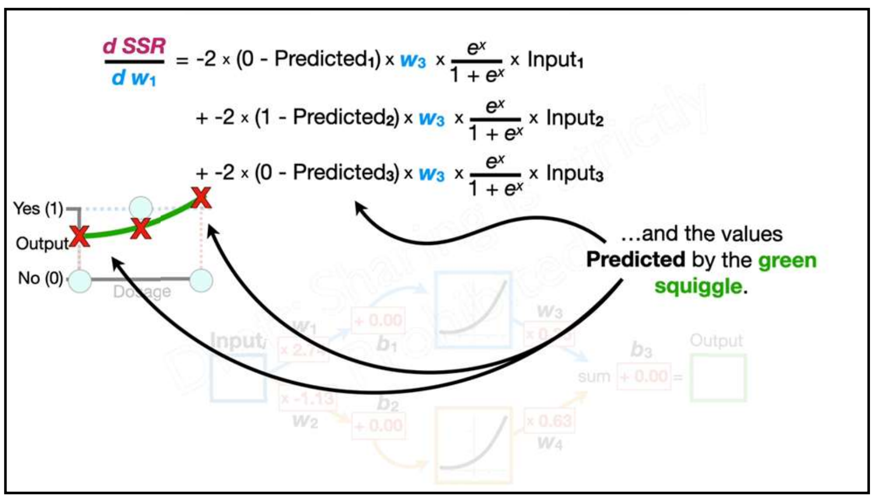

In this example, we pick numbers from a **Standard Normal Distribution (mean = 0 and standard deviation = 1)** for the **Weights**.
Input$_i$
$w_1$
x 2.74
$w_2$
x -1.13
+ ???
$b_1$
$b_2$
+ ???
$w_3$
x 0.36
$w_4$
x 0.63
$b_3$
sum + ??? =
Output

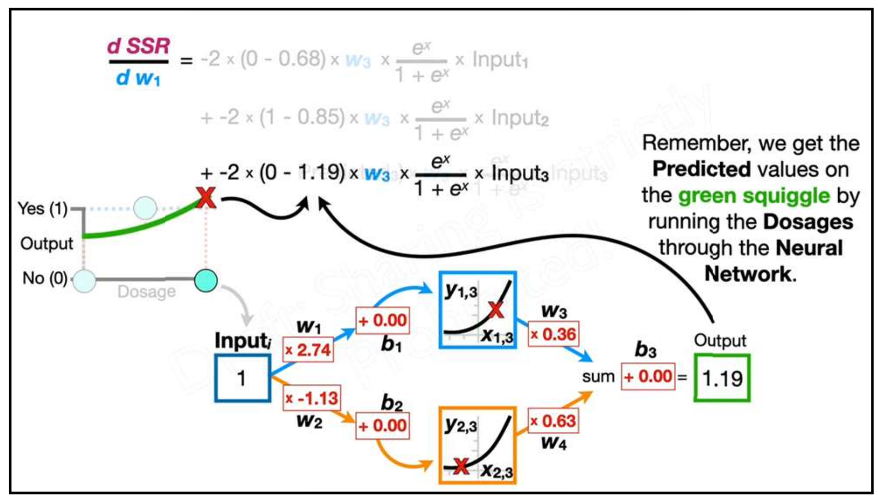

Now we initialize the **Bias** terms to **0**, because **Bias** terms frequently start at **0**.
Input$_i$
$w_1$
x 2.74
$w_2$
x -1.13
+ ???
$b_1$
$b_2$
+ ???
$w_3$
x 0.36
$w_4$
x 0.63
$b_3$
sum + ??? =
Output

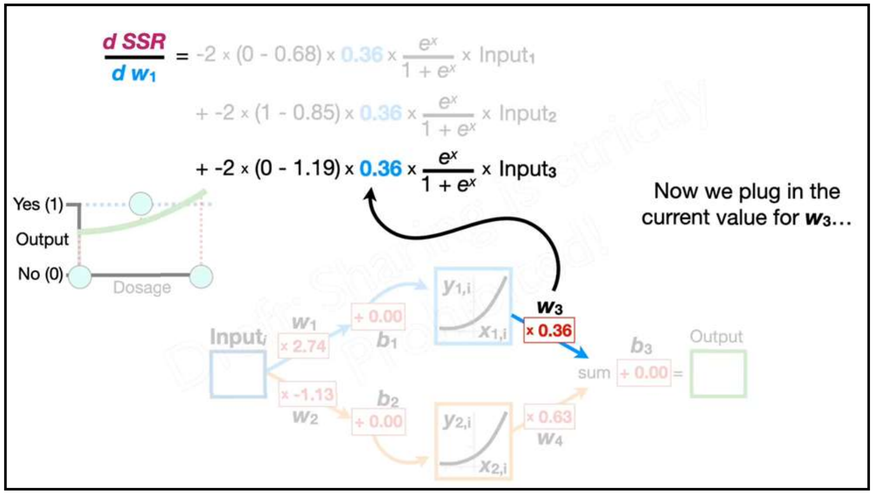

$\frac{d\ SSR}{d\ w_1} = \sum_{i=1}^{n=3} -2 \times (\text{Observed}_i - \text{Predicted}_i) \times w_3 \times \frac{e^x}{1+e^x} \times \text{Input}_i$
Now, starting with the derivative of the **SSR** with respect to $w_1$…
(**NOTE:** It does not matter which derivative we start with.)
Input$_i$
$w_1$
x 2.74
$w_2$
x -1.13
+ 0.00
$b_1$
$b_2$
+ 0.00
$w_3$
x 0.36
$w_4$
x 0.63
$b_3$
sum + 0.00 =
Output

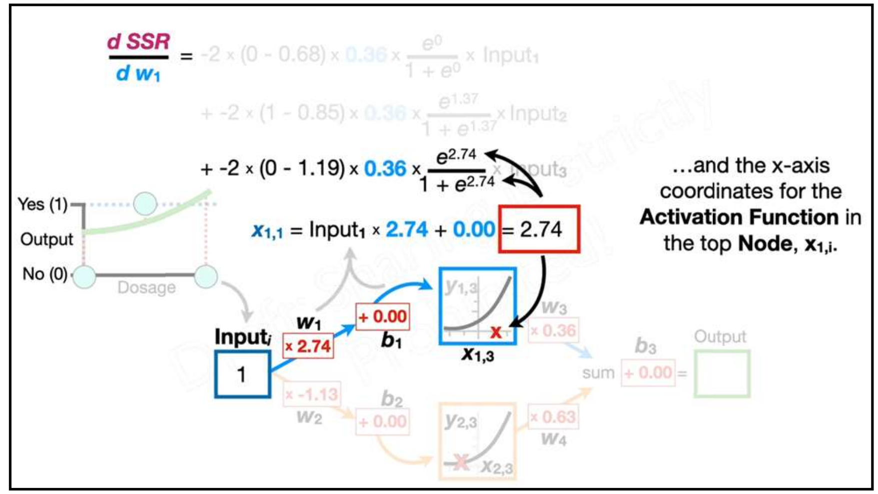

$\frac{d\,SSR}{d\,w_1} = -2 \times (\text{Observed}_1 - \text{Predicted}_1) \times w_3 \times \frac{e^x}{1+e^x} \times \text{Input}_1$
$+ -2 \times (\text{Observed}_2 - \text{Predicted}_2) \times w_3 \times \frac{e^x}{1+e^x} \times \text{Input}_2$
$+ -2 \times (\text{Observed}_3 - \text{Predicted}_3) \times w_3 \times \frac{e^x}{1+e^x} \times \text{Input}_3$
…we expand the summation.

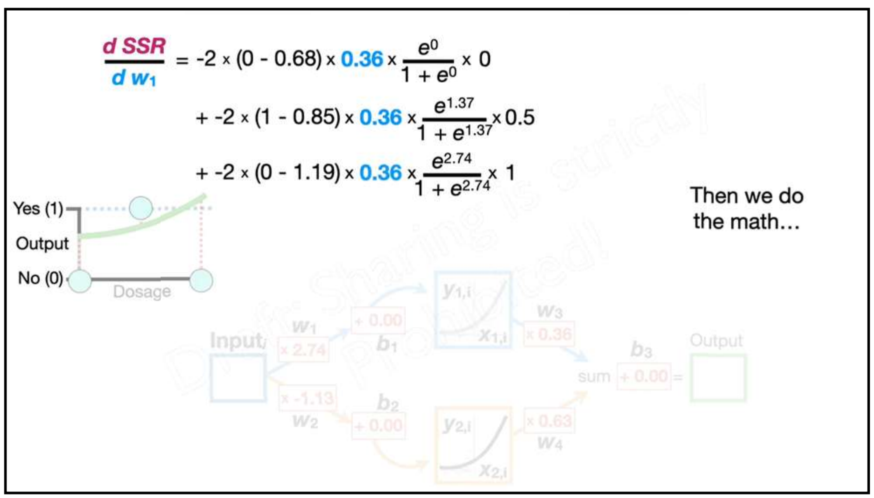

$\frac{d\,SSR}{d\,w_1} = -2 \times (0 - \text{Predicted}_1) \times w_3 \times \frac{e^x}{1+e^x} \times \text{Input}_1$
$+ -2 \times (1 - \text{Predicted}_2) \times w_3 \times \frac{e^x}{1+e^x} \times \text{Input}_2$
$+ -2 \times (0 - \text{Predicted}_3) \times w_3 \times \frac{e^x}{1+e^x} \times \text{Input}_3$
Yes (1)
Output
No (0)
Dosage
Then we plug in the **Observed** values…

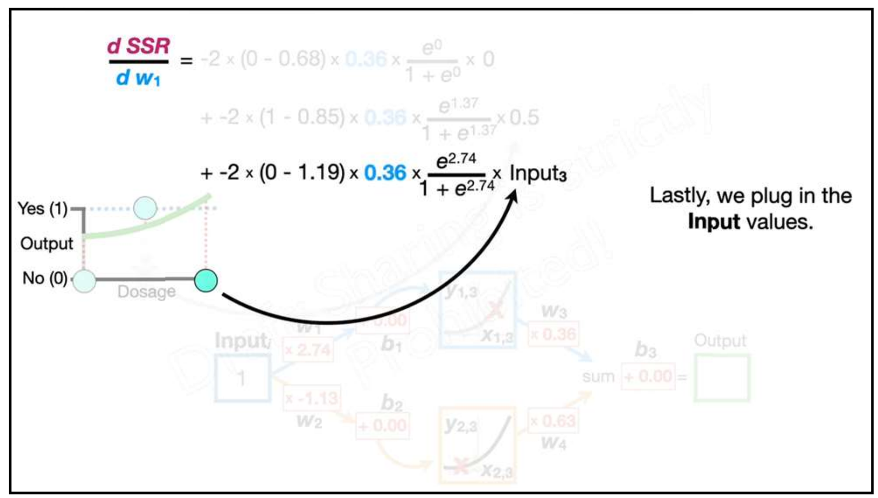

$\frac{d\,SSR}{d\,w_1} = -2 \times (0 - \text{Predicted}_1) \times w_3 \times \frac{e^x}{1+e^x} \times \text{Input}_1$
$+ -2 \times (1 - \text{Predicted}_2) \times w_3 \times \frac{e^x}{1+e^x} \times \text{Input}_2$
$+ -2 \times (0 - \text{Predicted}_3) \times w_3 \times \frac{e^x}{1+e^x} \times \text{Input}_3$
Yes (1)
Output
No (0)
Dosage
…and the values **Predicted** by the **green squiggle**.

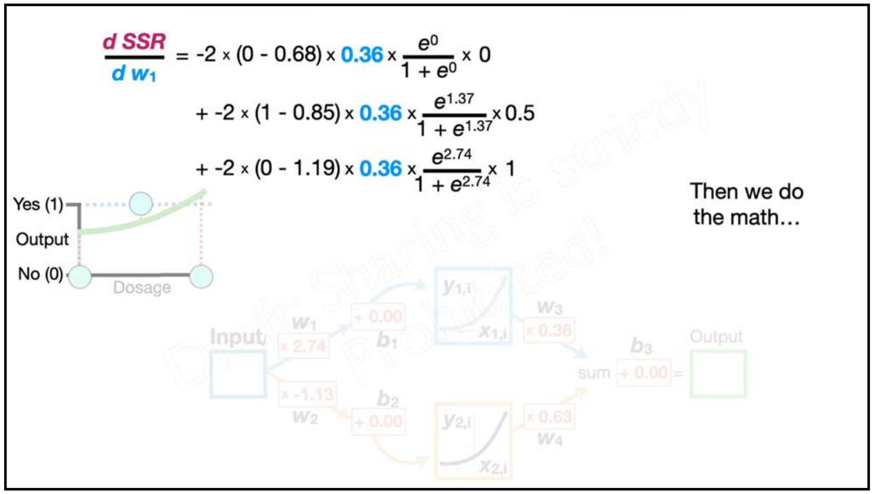

$\frac{d\,SSR}{d\,w_1} = -2 \times (0 - 0.68) \times w_3 \times \frac{e^x}{1+e^x} \times \text{Input}_1$
$+ -2 \times (1 - 0.85) \times w_3 \times \frac{e^x}{1+e^x} \times \text{Input}_2$
$+ -2 \times (0 - 1.19) \times w_3 \times \frac{e^x}{1+e^x} \times \text{Input}_3$
Remember, we get the **Predicted** values on the **green squiggle** by running the **Dosages** through the **Neural Network**.
Yes (1)
Output
No (0)
Dosage
Input$_i$
1
$w_1$
x 2.74
+ 0.00
$b_1$
$y_{1,3}$
$x_{1,3}$
$w_3$
x 0.36
x -1.13
$w_2$
$b_2$
+ 0.00
$y_{2,3}$
$x_{2,3}$
x 0.63
$w_4$
$b_3$
sum + 0.00 =
Output
1.19

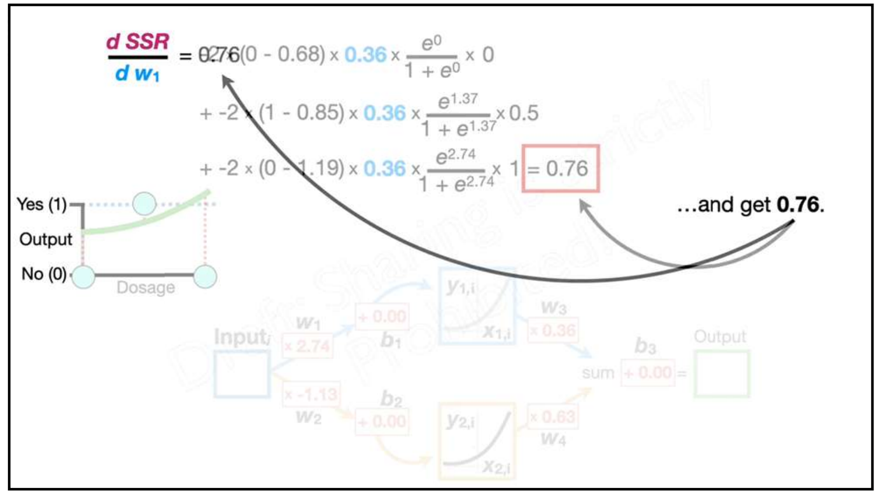

d SSR / d w1 = -2 x (0 - 0.68) x 0.36 x e^x / (1 + e^x) x Input1
+ -2 x (1 - 0.85) x 0.36 x e^x / (1 + e^x) x Input2
+ -2 x (0 - 1.19) x 0.36 x e^x / (1 + e^x) x Input3
Now we plug in the current value for w3…
Yes (1)
Output
No (0)
Dosage
w3
x 0.36

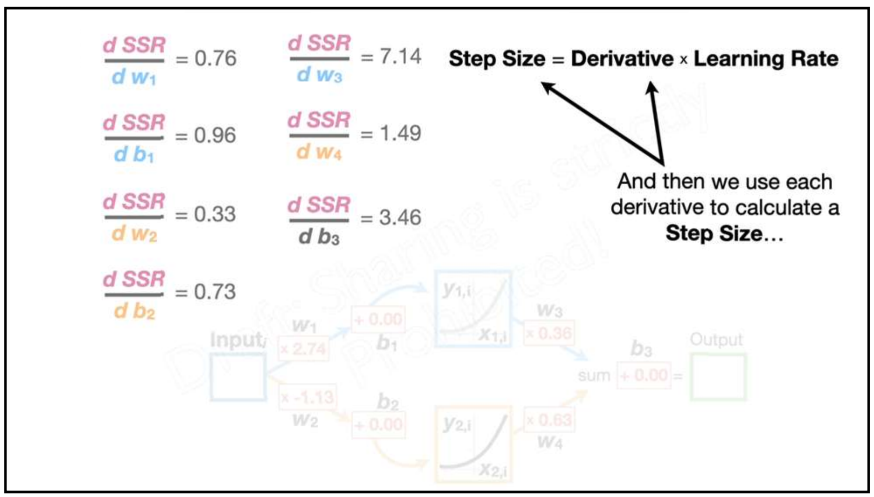

d SSR / d w1 = -2 x (0 - 0.68) x 0.36 x e^0 / (1 + e^0) x Input1
+ -2 x (1 - 0.85) x 0.36 x e^1.37 / (1 + e^1.37) x Input2
+ -2 x (0 - 1.19) x 0.36 x e^2.74 / (1 + e^2.74) x Input3
x1,1 = Input1 x 2.74 + 0.00 = 2.74
…and the x-axis coordinates for the Activation Function in the top Node, x1,i.
Yes (1)
Output
No (0)
Dosage
Inputi
1
w1
x 2.74
+ 0.00
b1
y1,3
x1,3

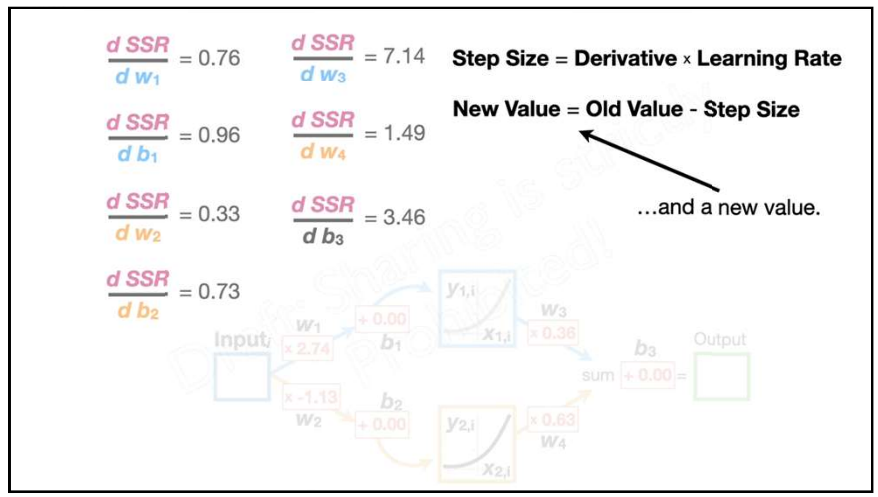

d SSR / d w1 = -2 x (0 - 0.68) x 0.36 x e^0 / (1 + e^0) x 0
+ -2 x (1 - 0.85) x 0.36 x e^1.37 / (1 + e^1.37) x 0.5
+ -2 x (0 - 1.19) x 0.36 x e^2.74 / (1 + e^2.74) x 1
Then we do the math…
Yes (1)
Output
No (0)
Dosage

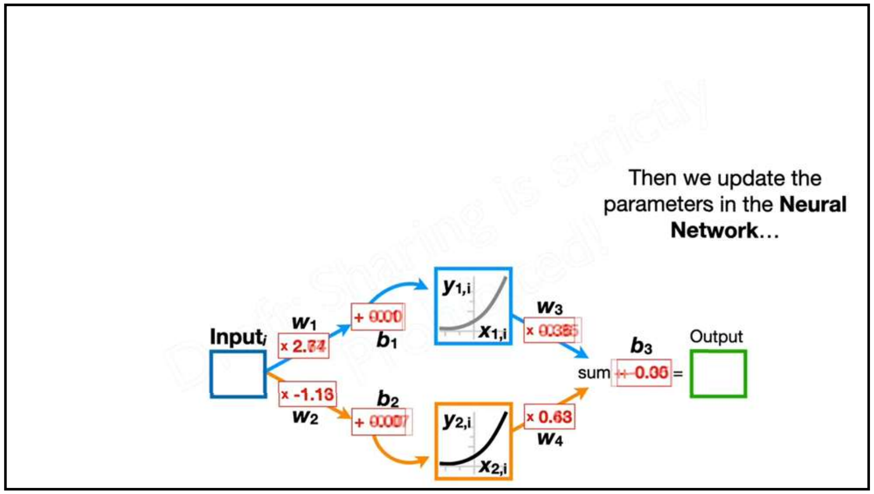

d SSR / d w1 = -2 x (0 - 0.68) x 0.36 x e^0 / (1 + e^0) x 0
+ -2 x (1 - 0.85) x 0.36 x e^1.37 / (1 + e^1.37) x 0.5
+ -2 x (0 - 1.19) x 0.36 x e^2.74 / (1 + e^2.74) x Input3
Lastly, we plug in the Input values.
Yes (1)
Output
No (0)
Dosage

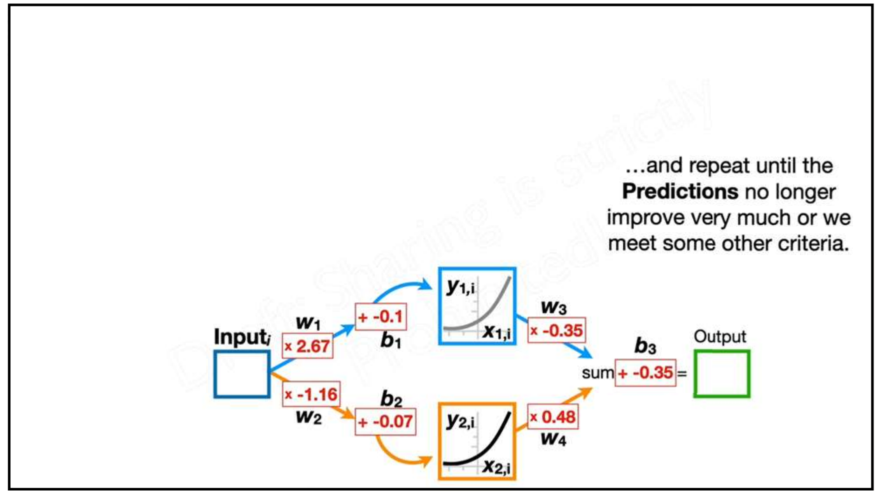

$\frac{d\ SSR}{d\ w_1} = -2 \times (0 - 0.68) \times 0.36 \times \frac{e^0}{1 + e^0} \times 0$
$+ -2 \times (1 - 0.85) \times 0.36 \times \frac{e^{1.37}}{1 + e^{1.37}} \times 0.5$
$+ -2 \times (0 - 1.19) \times 0.36 \times \frac{e^{2.74}}{1 + e^{2.74}} \times 1$
Then we do the math…
Yes (1)
Output
No (0)
Dosage

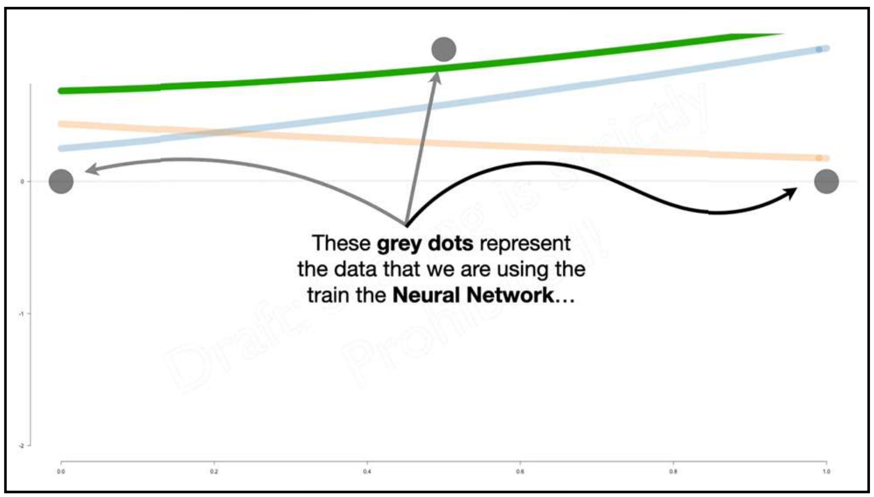

$\frac{d\ SSR}{d\ w_1} = 0.76$
$-2 \times (0 - 0.68) \times 0.36 \times \frac{e^0}{1 + e^0} \times 0$
$+ -2 \times (1 - 0.85) \times 0.36 \times \frac{e^{1.37}}{1 + e^{1.37}} \times 0.5$
$+ -2 \times (0 - 1.19) \times 0.36 \times \frac{e^{2.74}}{1 + e^{2.74}} \times 1 = 0.76$
…and get **0.76**.
Yes (1)
Output
No (0)
Dosage

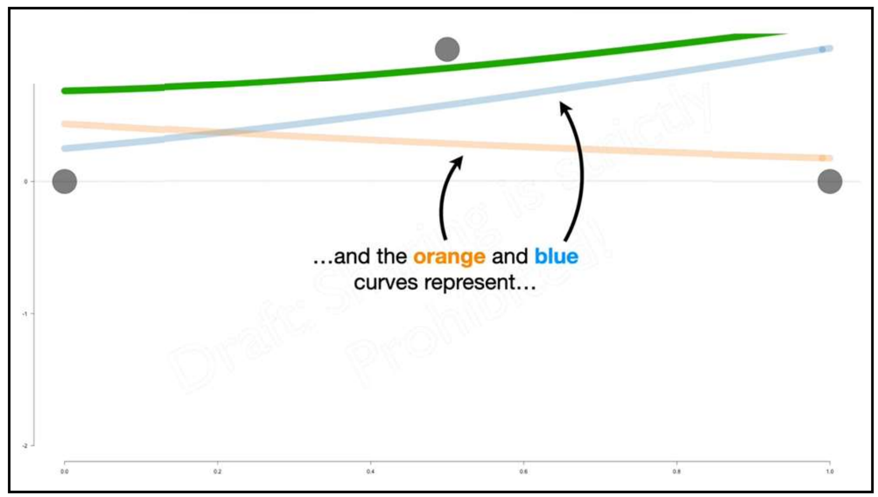

$\frac{d\ SSR}{d\ w_1} = 0.76$
$\frac{d\ SSR}{d\ b_1} = 0.96$
$\frac{d\ SSR}{d\ w_2} = 0.33$
$\frac{d\ SSR}{d\ b_2} = 0.73$
$\frac{d\ SSR}{d\ w_3} = 7.14$
$\frac{d\ SSR}{d\ w_4} = 1.49$
$\frac{d\ SSR}{d\ b_3} = 3.46$
**Step Size = Derivative × Learning Rate**
And then we use each derivative to calculate a **Step Size**…

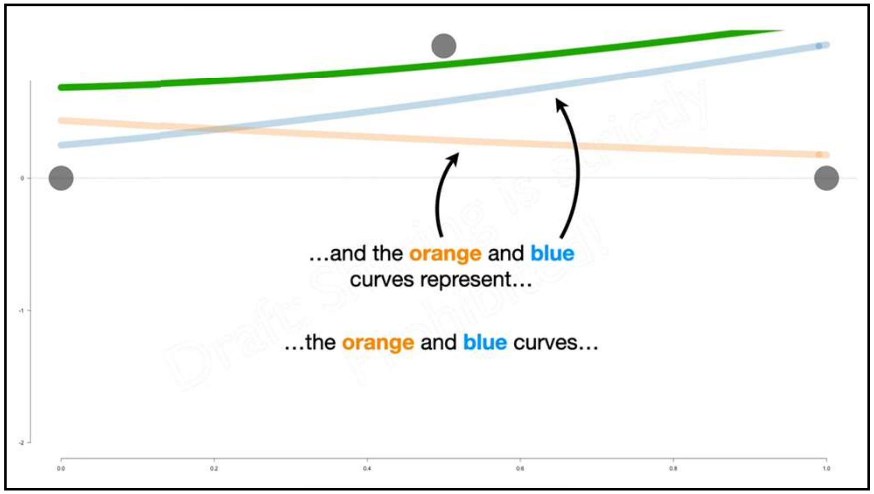

$\frac{d\ SSR}{d\ w_1} = 0.76$
$\frac{d\ SSR}{d\ b_1} = 0.96$
$\frac{d\ SSR}{d\ w_2} = 0.33$
$\frac{d\ SSR}{d\ b_2} = 0.73$
$\frac{d\ SSR}{d\ w_3} = 7.14$
$\frac{d\ SSR}{d\ w_4} = 1.49$
$\frac{d\ SSR}{d\ b_3} = 3.46$
**Step Size = Derivative × Learning Rate**
**New Value = Old Value - Step Size**
…and a new value.

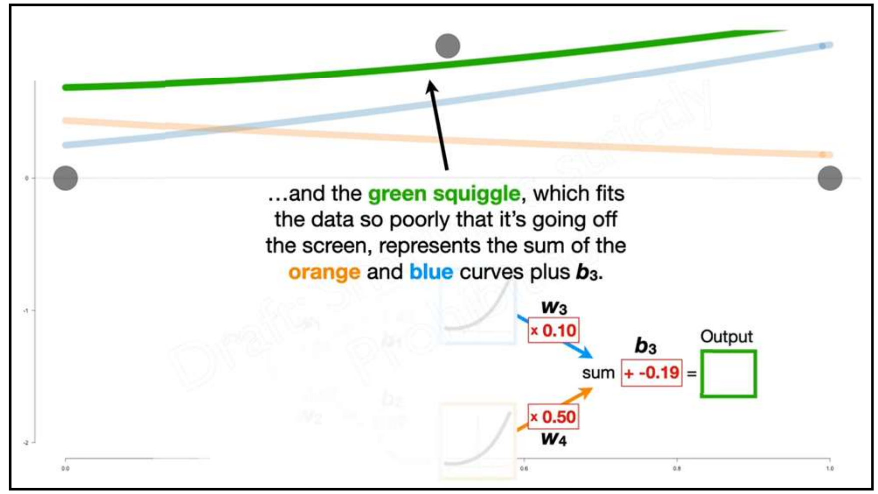

Then we update the parameters in the **Neural Network**…
Input$_i$
$w_1$
$b_1$
$y_{1,i}$
$x_{1,i}$
$w_3$
$b_3$
Output
sum
=
x -1.16
$w_2$
$b_2$
$y_{2,i}$
$x_{2,i}$
x 0.63
$w_4$

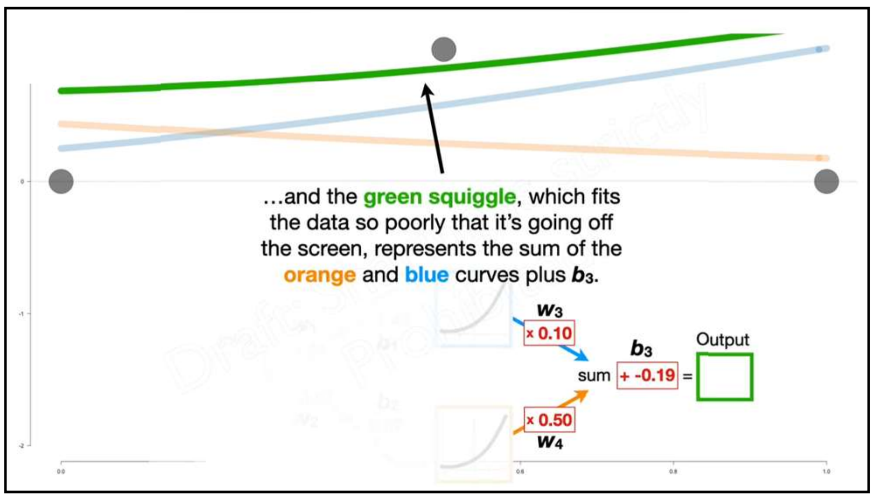

…and repeat until the **Predictions** no longer improve very much or we meet some other criteria.
Input$_i$
$w_1$
x 2.67
+ -0.1
$b_1$
$y_{1,i}$
$x_{1,i}$
$w_3$
x -0.35
$b_3$
Output
sum
+ -0.35
=
x -1.16
$w_2$
$b_2$
+ -0.07
$y_{2,i}$
$x_{2,i}$
x 0.48
$w_4$

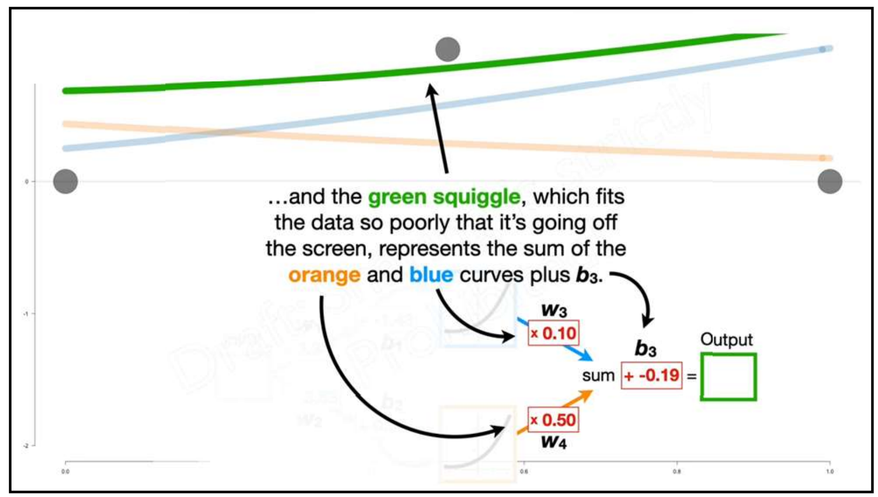

These **grey dots** represent the data that we are using the train the **Neural Network**…

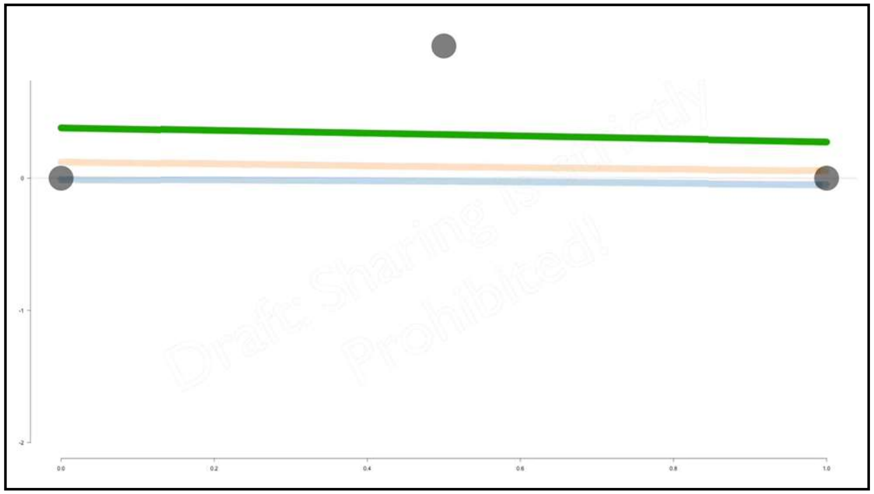

…and the **orange** and **blue** curves represent…

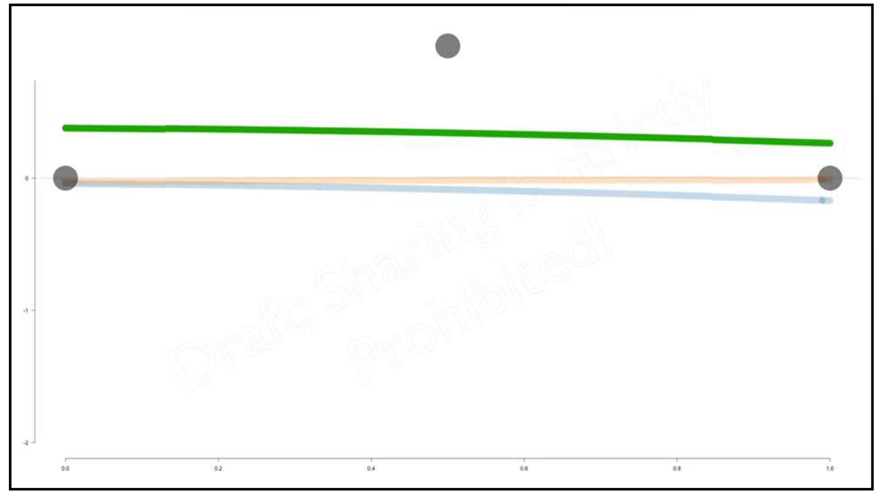

...and the orange and blue curves represent...
...the orange and blue curves...

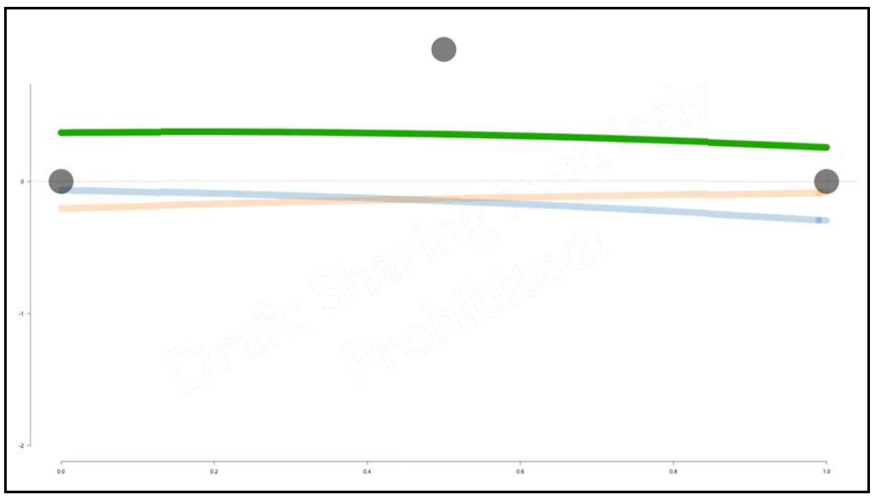

...and the green squiggle, which fits the data so poorly that it's going off the screen, represents the sum of the orange and blue curves plus $b_3$.
$w_3$
x 0.10
$b_3$
Output
sum
+ -0.19
=
x 0.50
$w_4$

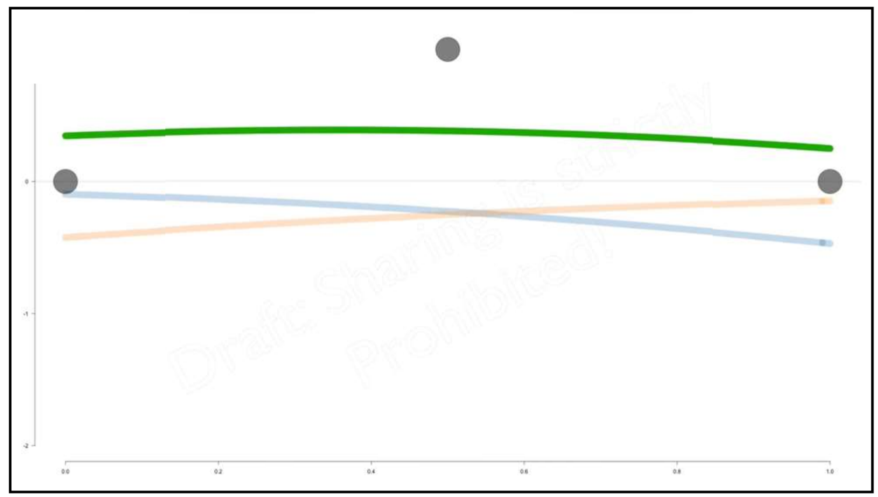

...and the green squiggle, which fits the data so poorly that it's going off the screen, represents the sum of the orange and blue curves plus $b_3$.
$w_3$
x 0.10
$b_3$
Output
sum
+ -0.19
=
x 0.50
$w_4$


...and the green squiggle, which fits the data so poorly that it's going off the screen, represents the sum of the orange and blue curves plus $b_3$.
$w_3$
x 0.10
$b_3$
Output
sum
+ -0.19
=
x 0.50
$w_4$

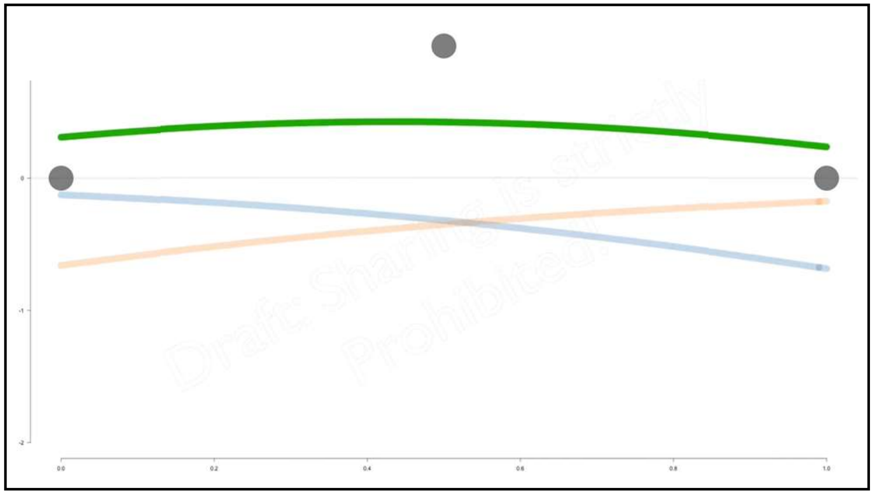


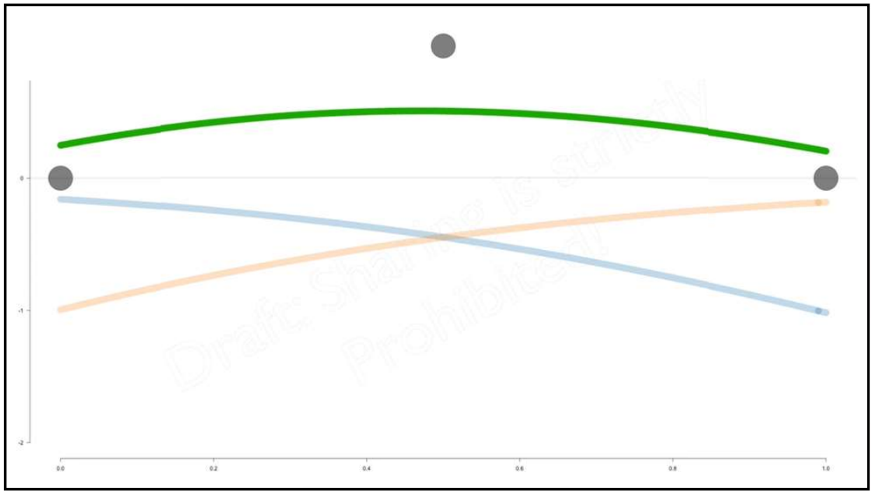


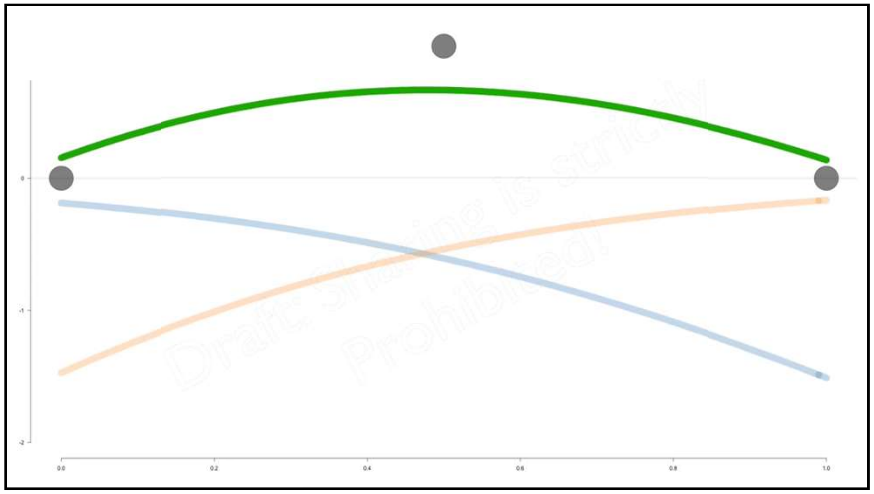


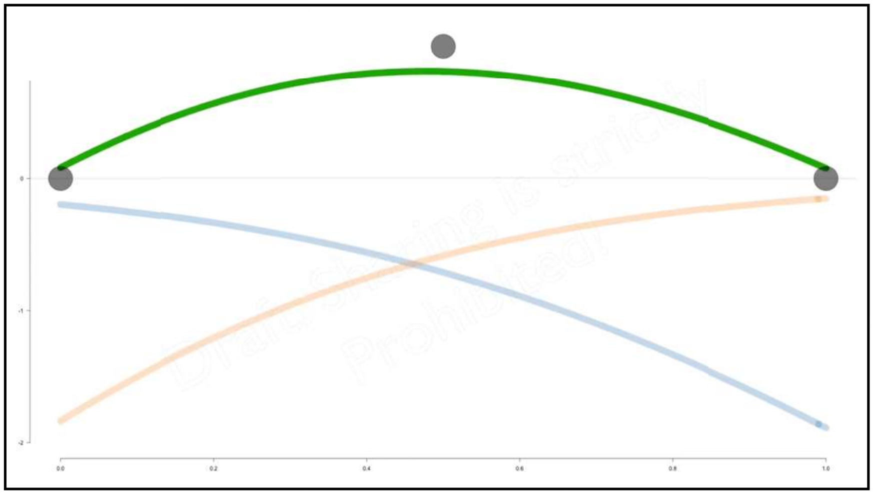


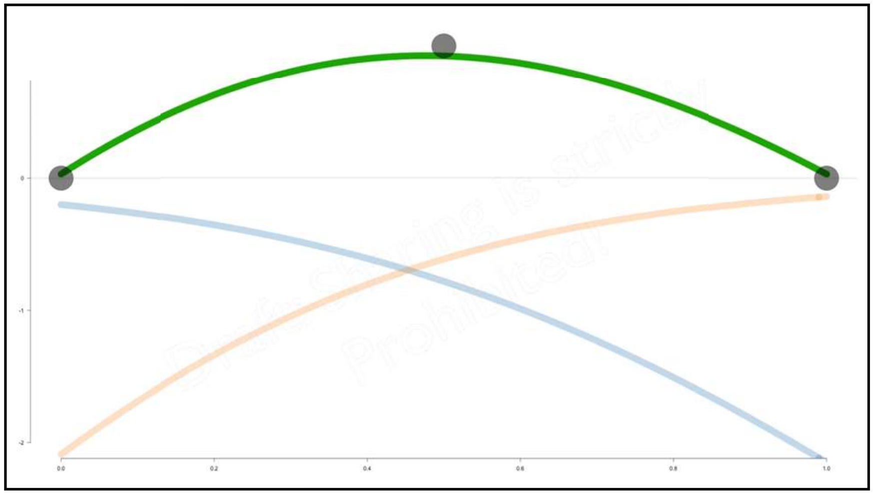
3/8/2021

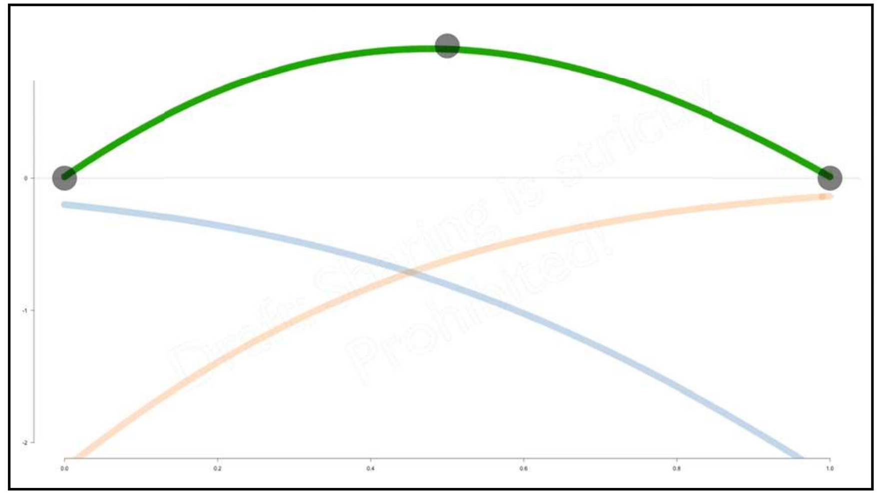


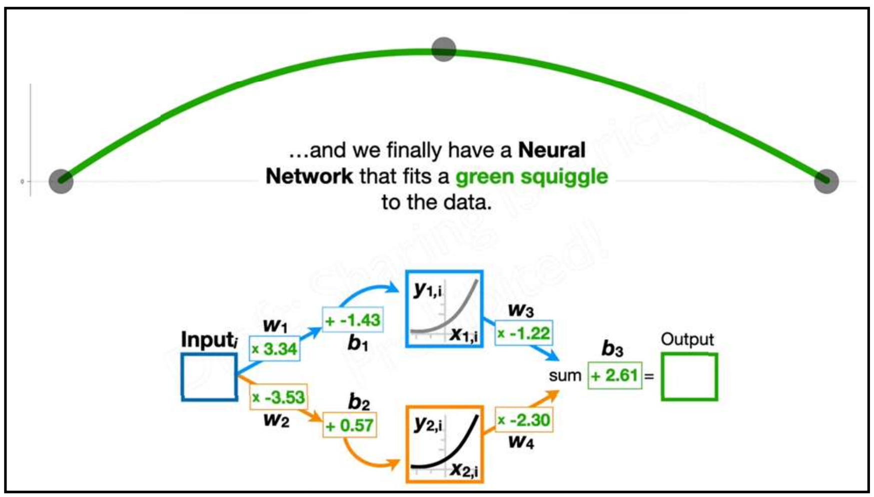











...and we finally have a **Neural Network** that fits a **green squiggle** to the data.
Input$_i$
$w_1$
x 3.34
+ -1.43
$b_1$
$y_{1,i}$
$x_{1,i}$
$w_3$
x -1.22
x -3.53
$w_2$
$b_2$
+ 0.57
$y_{2,i}$
$x_{2,i}$
x -2.30
$w_4$
$b_3$
sum
+ 2.61
=
Output

THANK YOU!