Office Hours: Linear Classifiers Part Deux

By Abraham Kang

Before starting your work, please review [eCornell's policy regarding plagiarism](https://s3.amazonaws.com/ecornell/global/eCornellPlagiarismPolicy.pdf) (the presentation of someone else's work as your own without source credit - https://s3.amazonaws.com/ecornell/global/eCornellPlagiarismPolicy.pdf).

Sigmoid Function

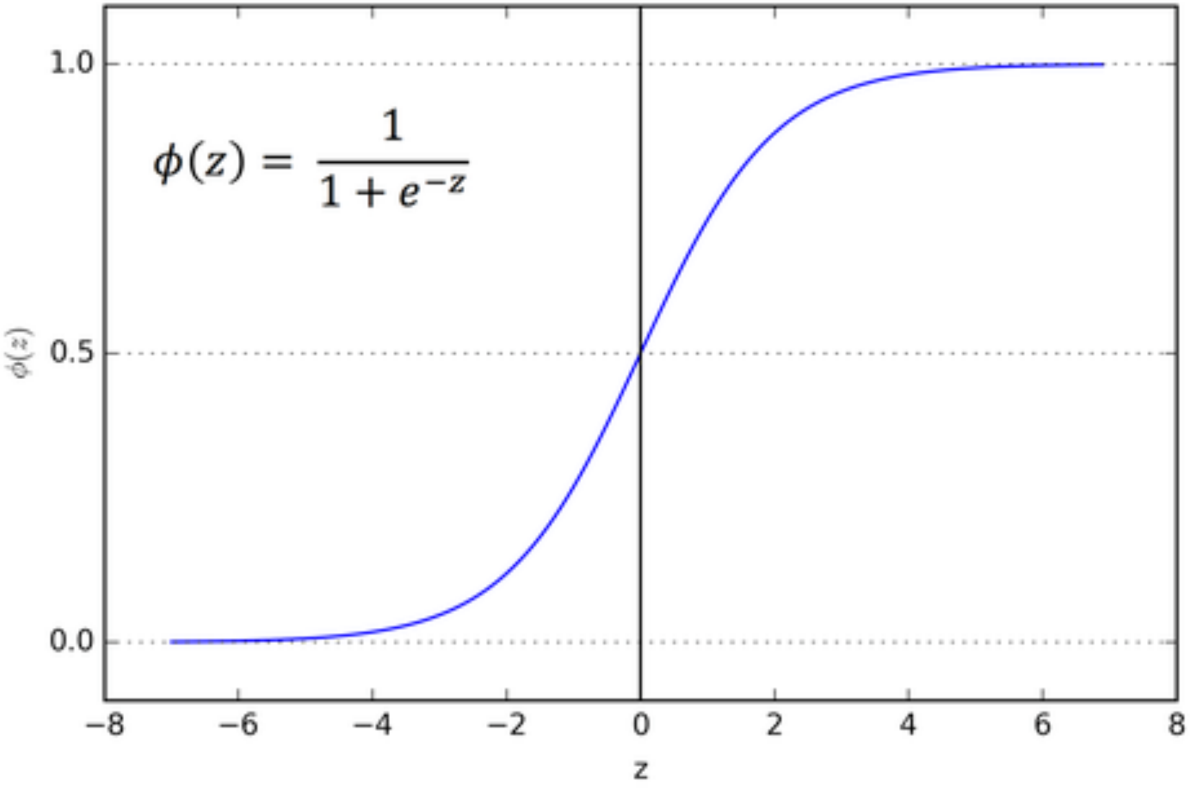


Fig: Sigmoid Function

Finding the Optimal Weight Values

- Loss Functions
- Partial Derivatives
- Stochastic Gradient Decent

Loss Functions

MAE = Mean Absolute Error $$MAE = \frac{\sum_{i=1}^{n} \mid y_i - \hat{y}_i \mid}{n}$$

MSE = Mean Squared Error $$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}$$

Cross Entropy Loss = $$-(y_i log(\hat{y}_i) + (1 - y_i) log(1 - \hat{y}_i))$$

Carlini and Wagner = $$\text{minimize} \;\; \|x' - x\|_2^2 + c \cdot \ell(x') \text{ where the loss function } \ell \text{ is defined as}$$
$$\ell(x') = \max(\max\{Z(x')_i : i \neq t\} - Z(x')_t, -\kappa)$$

Fast Gradient Sign Method (FGSM) = $$x' = x - \epsilon \cdot \mathbf{sign}(\nabla \mathrm{loss}_{F,t}(x))$$

Loss Functions

$$\mathbf{w}^\top \mathbf{x}_i + b$$

$$l(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \mathbf{w} - y_i)^2$$

$$\mathbf{w} = \arg\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x_i}^T \mathbf{w} - y_i)^2$$

Partial Derivatives

- What do they tell you
- Which partial derivatives are you interested in if you are trying to correct w and b?

$$\mathbf{w}^\top \mathbf{x}_i + b$$

$$l(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \mathbf{w} - y_i)^2$$

SDG

$$l(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \mathbf{w} - y_i)^2$$

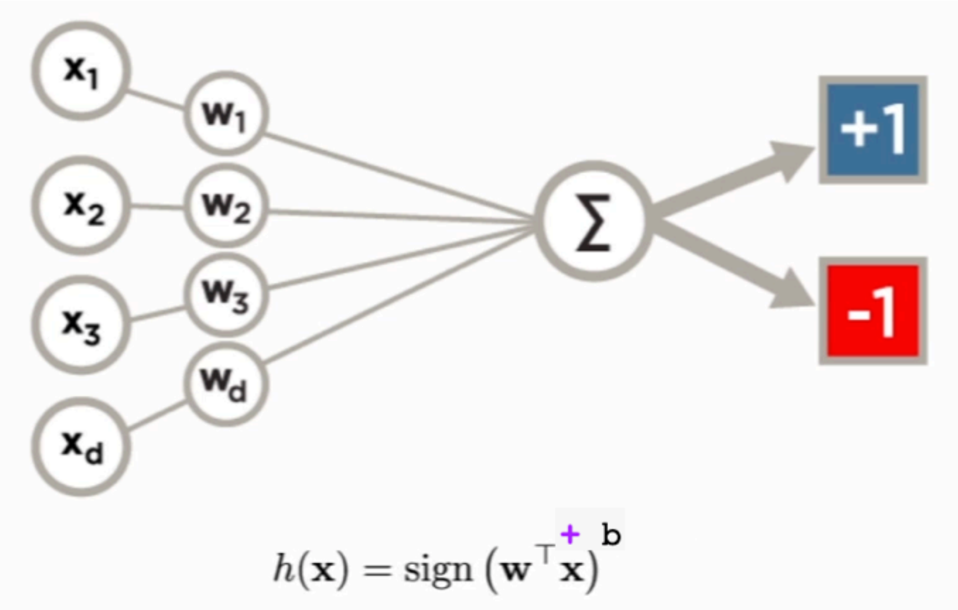
$$^{*}W_x = W_x - \alpha\left(\frac{\partial Error}{\partial W_x}\right)$$

Build-a-Spam-Email-Classifier Part I

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

Build-a-Spam-Email-Classifier Part II

Before

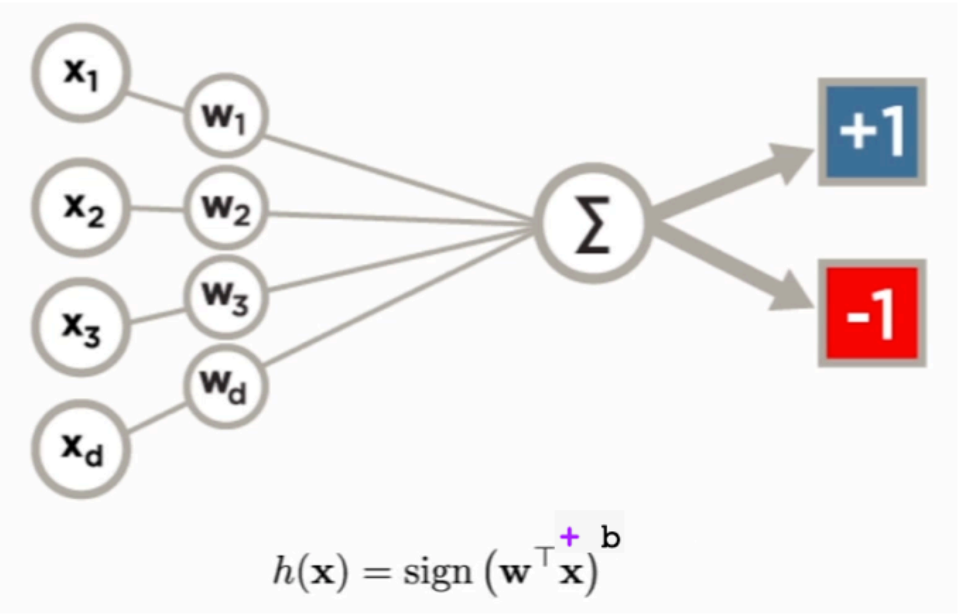


Now…So what is the difference?

$$P(y \mid \mathbf{x}; \mathbf{w}) = \sigma(y(\mathbf{w}^\top \mathbf{x} + b))$$

Build-a-Spam-Email-Classifier Part II

Before



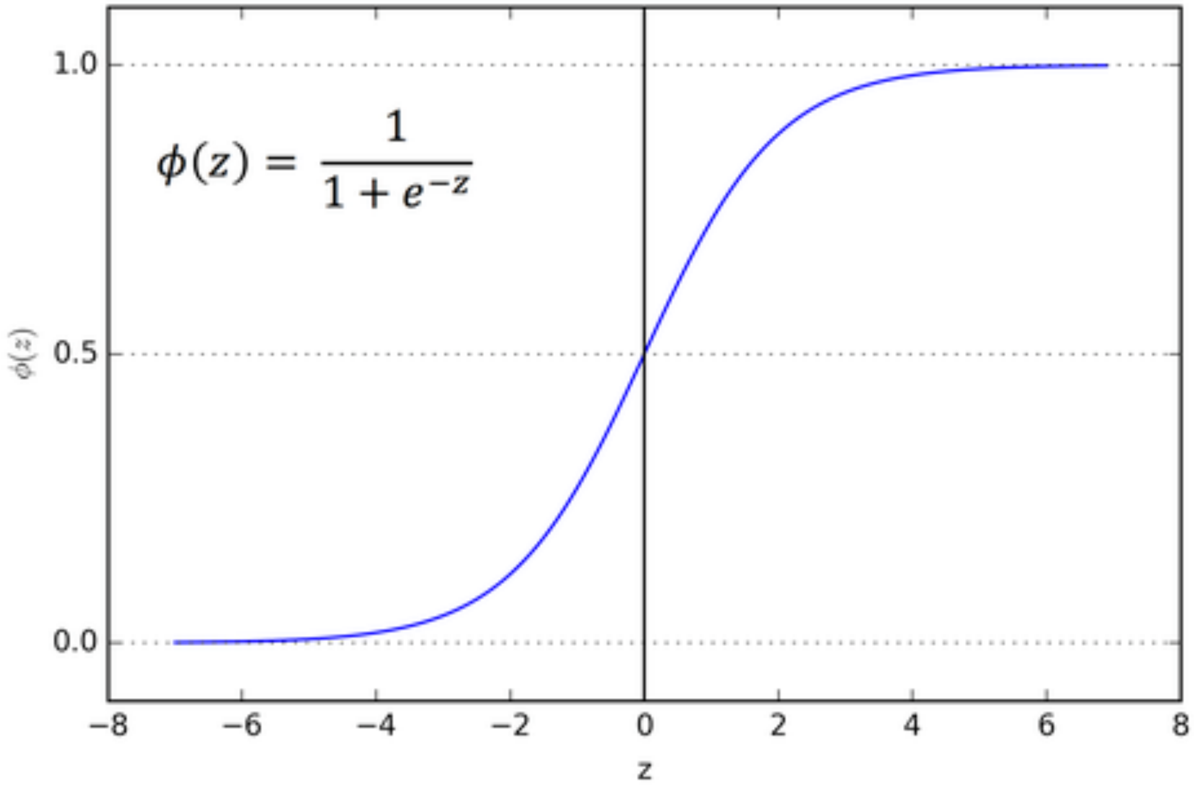


Fig: Sigmoid Function

Now…So what is the difference?

$$P(y \mid \mathbf{x}; \mathbf{w}) = \sigma(y(\mathbf{w}^\top\mathbf{x} + b))$$

Build-a-Spam-Email-Classifier Part III

$$NLL = -\log P(\mathbf{y} \mid \mathbf{X}; \mathbf{w}, b) = -\sum_{i=1}^{n} \log((P(y_i \mid \mathbf{x}_i; \mathbf{w}, b))) = -\sum_{i=1}^{n} \log\left(\sigma(y_i(\mathbf{w}^\top \mathbf{x}_i + b))\right)$$

Build-a-Spam-Email-Classifier Part IV

$$\frac{\partial NLL(\mathbf{X}, \mathbf{y}, \mathbf{w}, b)}{\partial \mathbf{w}} = \sum_{i=1}^{n} -y_i \sigma(-y_i(\mathbf{w}^\top \mathbf{x}_i + b))\mathbf{x}_i.$$

$$\frac{\partial NLL(\mathbf{X}, \mathbf{y}, \mathbf{w}, b)}{\partial b} = \sum_{i=1}^{n} -y_i \sigma(-y_i(\mathbf{w}^\top \mathbf{x}_i + b)).$$

What are these derivatives tied to?

Build-a-Spam-Email-Classifier Part IV

$$\frac{\partial NLL(\mathbf{X}, \mathbf{y}, \mathbf{w}, b)}{\partial \mathbf{w}} = \sum_{i=1}^{n} -y_i \sigma(-y_i(\mathbf{w}^\top \mathbf{x}_i + b))\mathbf{x}_i.$$

$$\frac{\partial NLL(\mathbf{X}, \mathbf{y}, \mathbf{w}, b)}{\partial b} = \sum_{i=1}^{n} -y_i \sigma(-y_i(\mathbf{w}^\top \mathbf{x}_i + b)).$$

What are these derivatives tied to?

$$^{*}W_x = W_x - \alpha \left(\frac{\partial Error}{\partial W_x}\right)$$

Build-a-Spam-Email-Classifier Part V

$$*W_x = W_x - \alpha \left(\frac{\partial Error}{\partial W_x}\right)$$

$$b' = b - \eta \frac{\partial L}{\partial b}$$

1. Calc gradients using previous gradient(…)
2. Update weights and biases
3. Save loss for that step in losses using the log_loss(…)

Math Review – Chain Rule - Computational Procedure

- If you are given the following functions that are chained together):

y = 2x then z = 3y then a=4z then

output = 7a and finally Error=(actual - output)^2

- Because the inputs are chained together, we can mathematically derive the effect each input has on the overall Error by finding the derivative of the Error with respect to that particular input. $\frac{\partial Error}{\partial output}$ $\frac{\partial Error}{\partial a}$ $\frac{\partial Error}{\partial z}$ $\frac{\partial Error}{\partial y}$ $\frac{\partial Error}{\partial x}$

- We can start from the back of the chain and move toward the beginning to figure each of these out

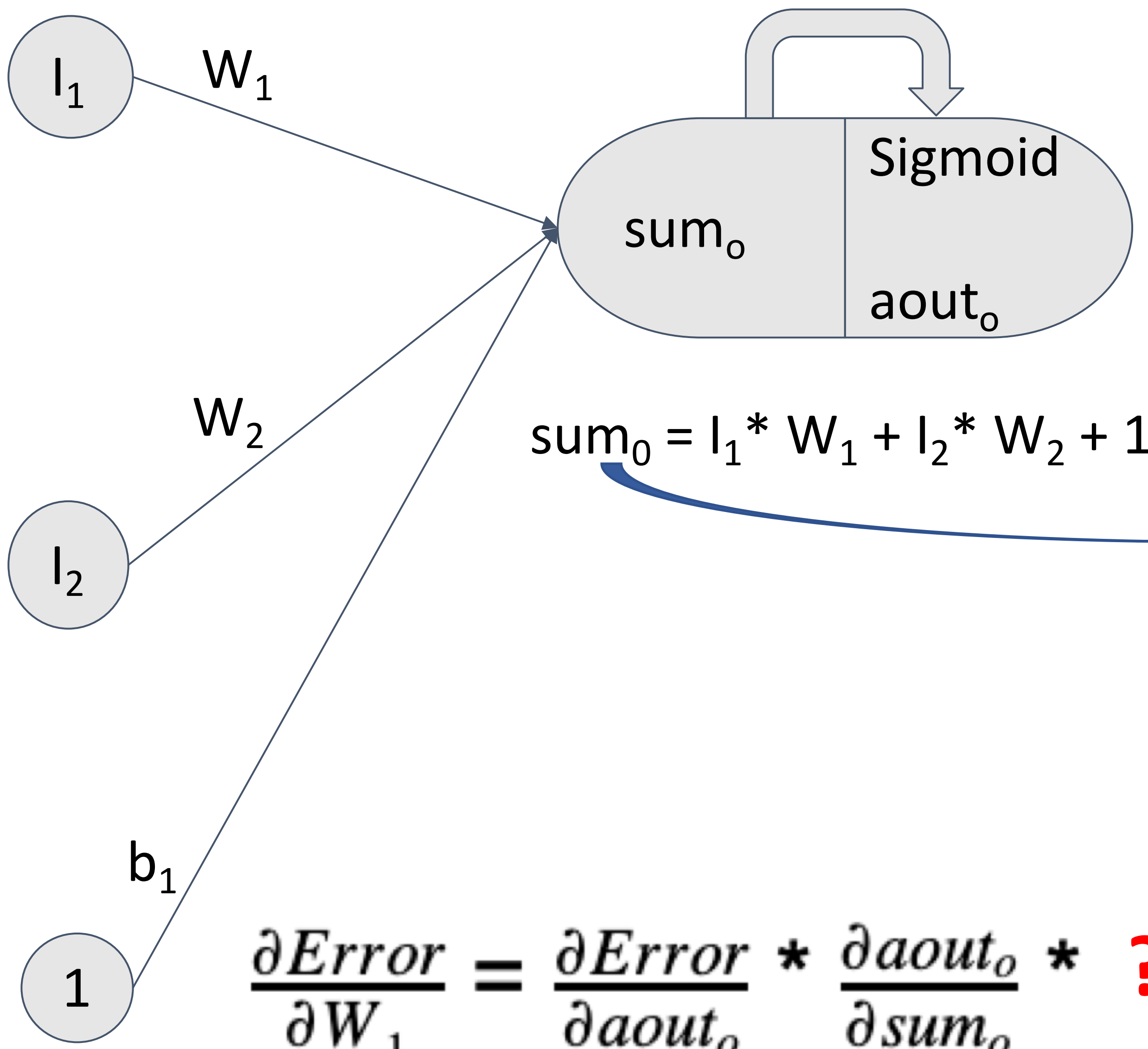
$$\frac{\partial Error}{\partial a} = \frac{\partial Error}{\partial output} * \frac{\partial output}{\partial a}$$

$$\frac{\partial Error}{\partial z} = \frac{\partial Error}{\partial output} * \frac{\partial output}{\partial a} * \frac{\partial a}{\partial z}$$

$$\frac{\partial Error}{\partial y} = \frac{\partial Error}{\partial output} * \frac{\partial output}{\partial a} * \frac{\partial a}{\partial z} * \frac{\partial z}{\partial y}$$

$$\frac{\partial Error}{\partial x} =$$

Forward Propagation



$sum_0 = I_1 * W_1 + I_2 * W_2 + 1 * b_1$

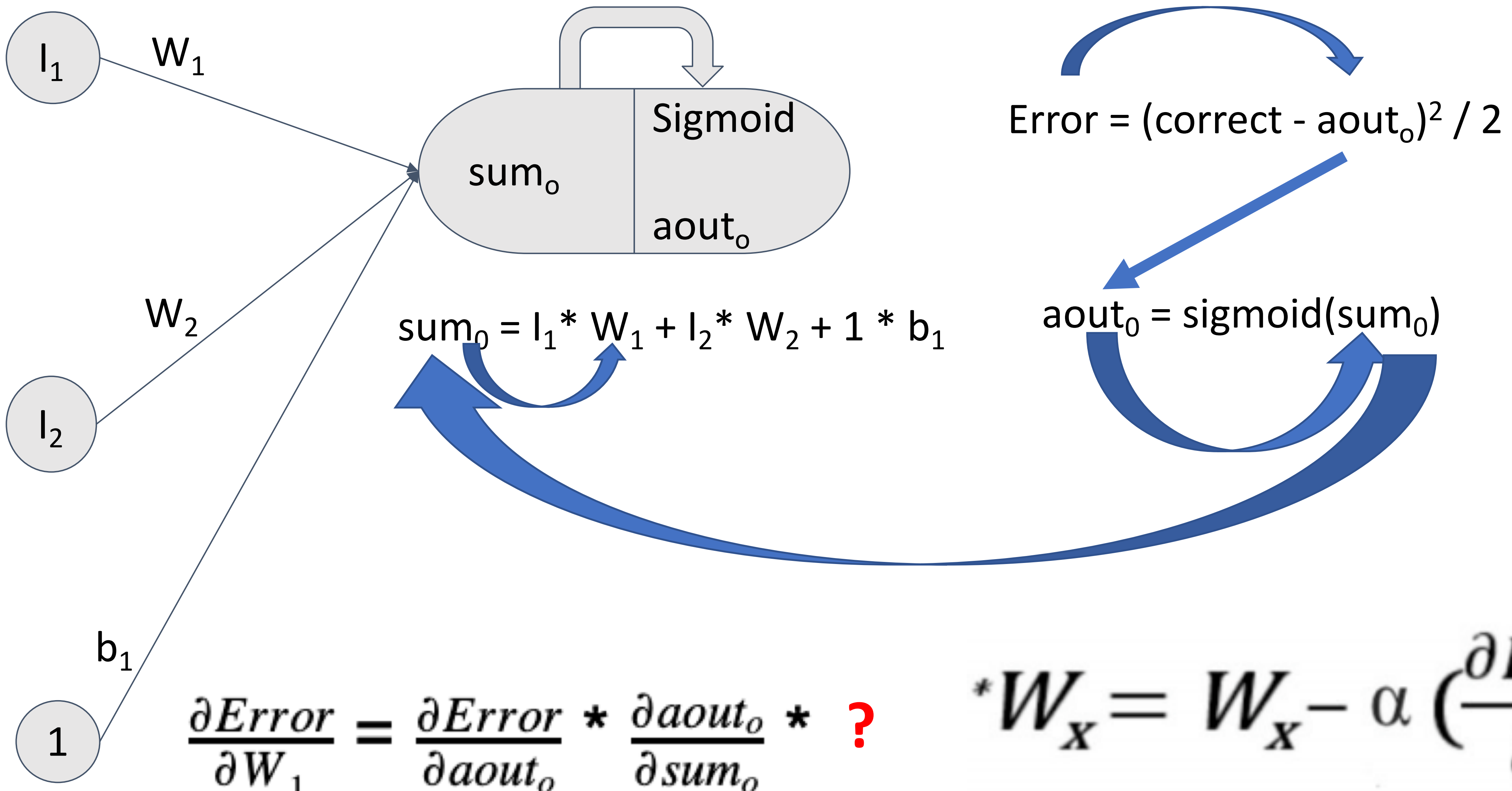
$aout_0 = \text{sigmoid}(sum_0)$

Error = (correct - $aout_o)^2$ / 2

$$\frac{\partial Error}{\partial W_1} = \frac{\partial Error}{\partial aout_o} * \frac{\partial aout_o}{\partial sum_o} * \ ?$$

$$^{*}W_x = W_x - \alpha \left(\frac{\partial Error}{\partial W_x}\right)$$

Back Propagation

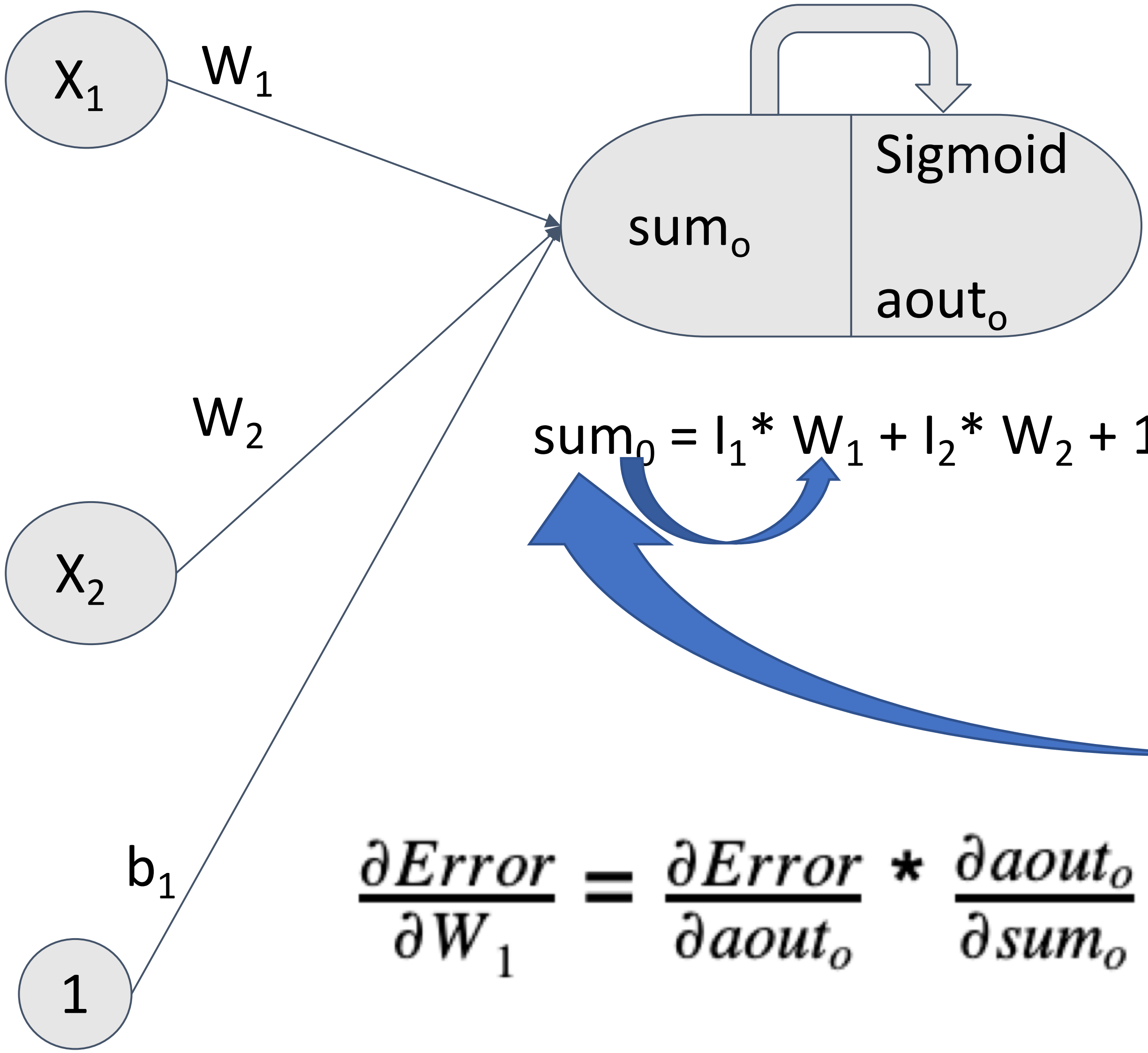


$$\frac{\partial Error}{\partial W_1} = \frac{\partial Error}{\partial aout_o} * \frac{\partial aout_o}{\partial sum_o} * \textbf{?}$$

$${}^{*}W_x = W_x - \alpha \left(\frac{\partial Error}{\partial W_x}\right)$$

Back Propagation

$$ {}^{*}W_x = W_x - \alpha\left(\frac{\partial Error}{\partial W_x}\right) $$



Error = $(y - aout_o)^2 / 2$

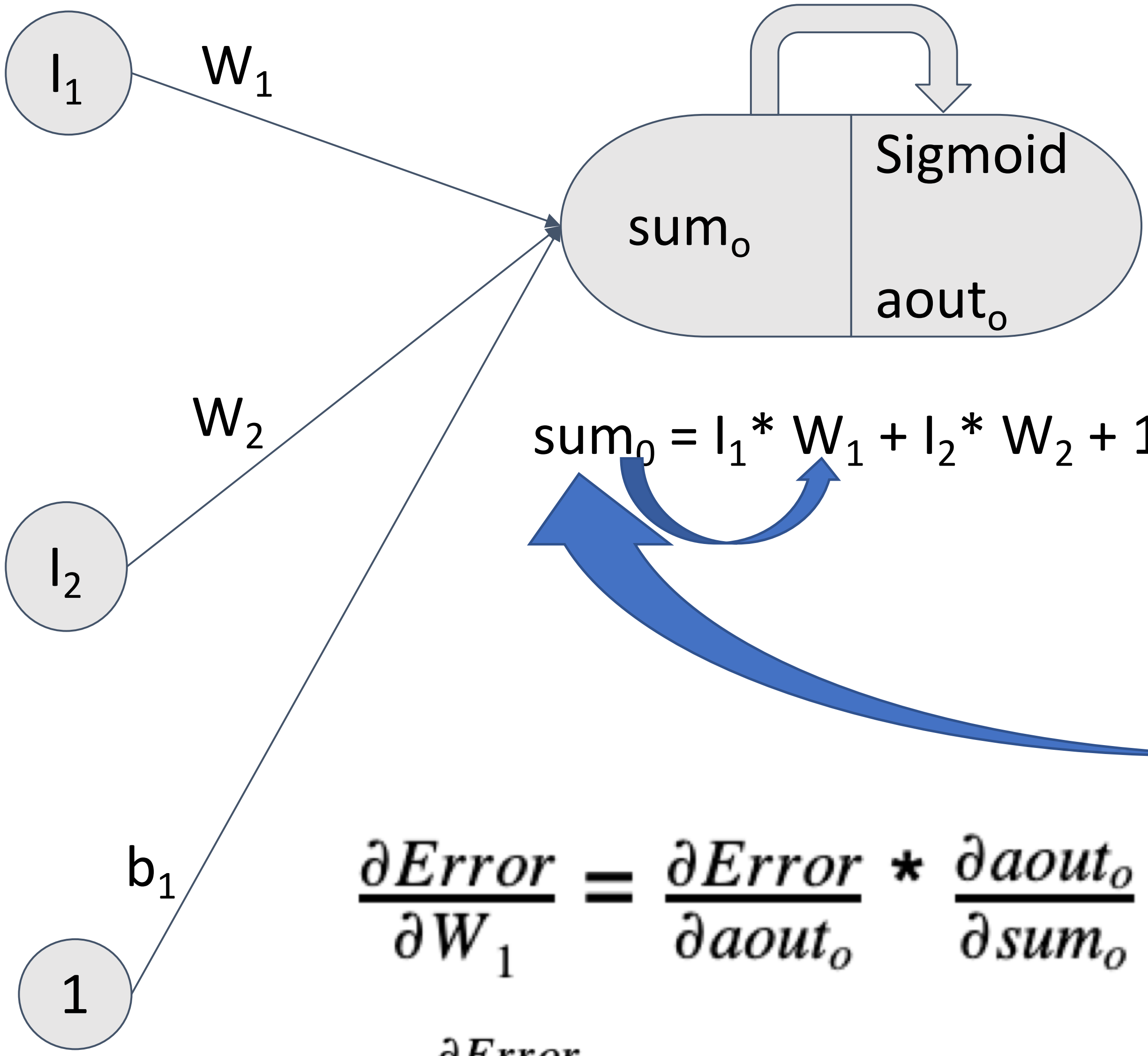
$aout_0$ = sigmoid(sum_0)

$sum_0 = I_1 * W_1 + I_2 * W_2 + 1 * b_1$

$$ \frac{\partial Error}{\partial W_1} = \frac{\partial Error}{\partial aout_o} * \frac{\partial aout_o}{\partial sum_o} * \frac{\partial sum_o}{\partial W_1} $$

Back Propagation

$$ {}^{*}W_x = W_x - \alpha\left(\frac{\partial Error}{\partial W_x}\right) $$



$$ Error = (correct - aout_o)^2 / 2 $$

$$ aout_0 = sigmoid(sum_0) $$

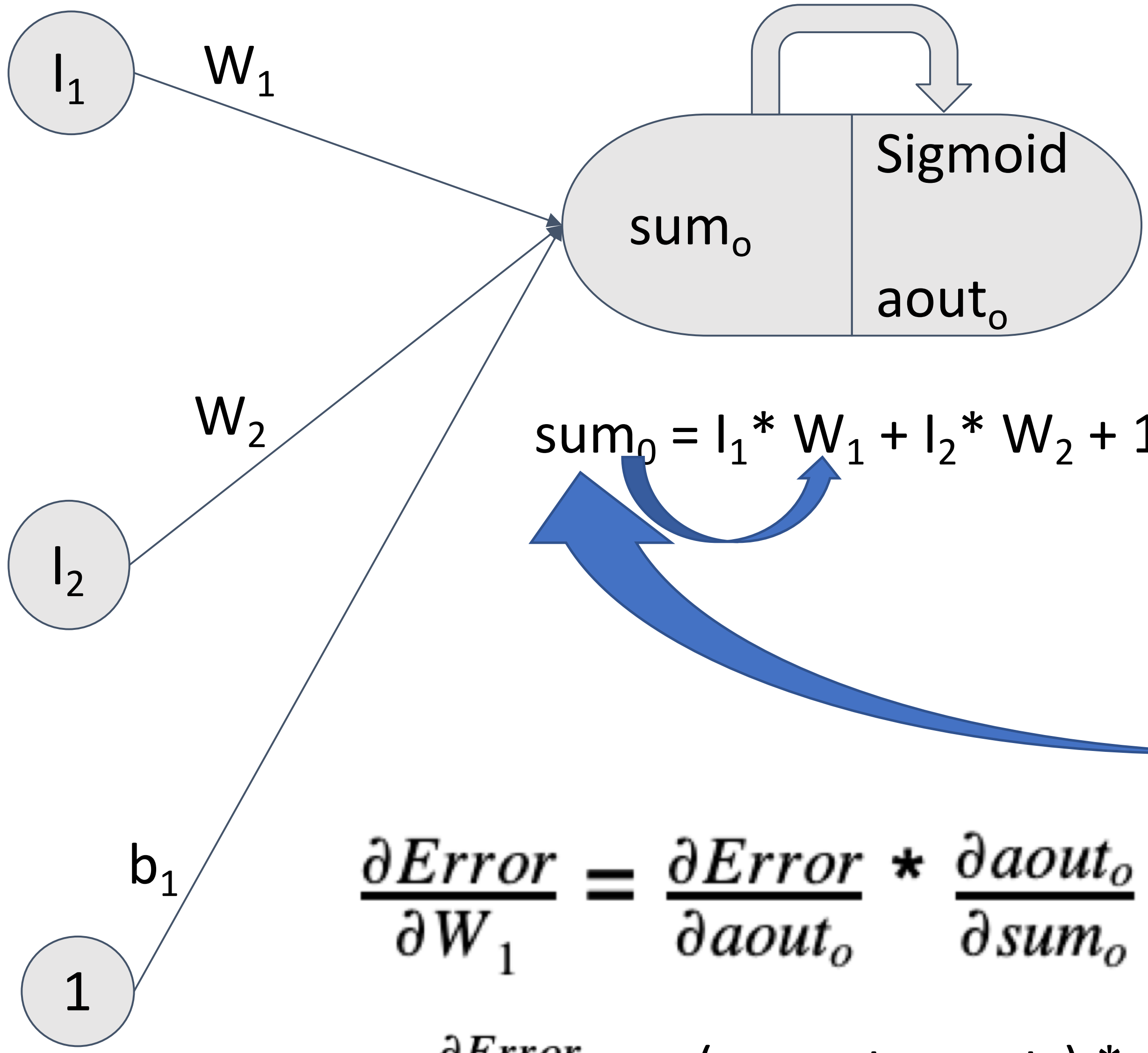
$$ sum_0 = I_1 * W_1 + I_2 * W_2 + 1 * b_1 $$

$$ \frac{\partial Error}{\partial W_1} = \frac{\partial Error}{\partial aout_o} * \frac{\partial aout_o}{\partial sum_o} * \frac{\partial sum_o}{\partial W_1} $$

$$ \frac{\partial Error}{\partial W_1} $$

Back Propagation

$$*W_x = W_x - \alpha \left(\frac{\partial Error}{\partial W_x}\right)$$



$$Error = (correct - aout_o)^2 / 2$$

$$aout_0 = sigmoid(sum_0)$$

$$sum_0 = I_1 * W_1 + I_2 * W_2 + 1 * b_1$$

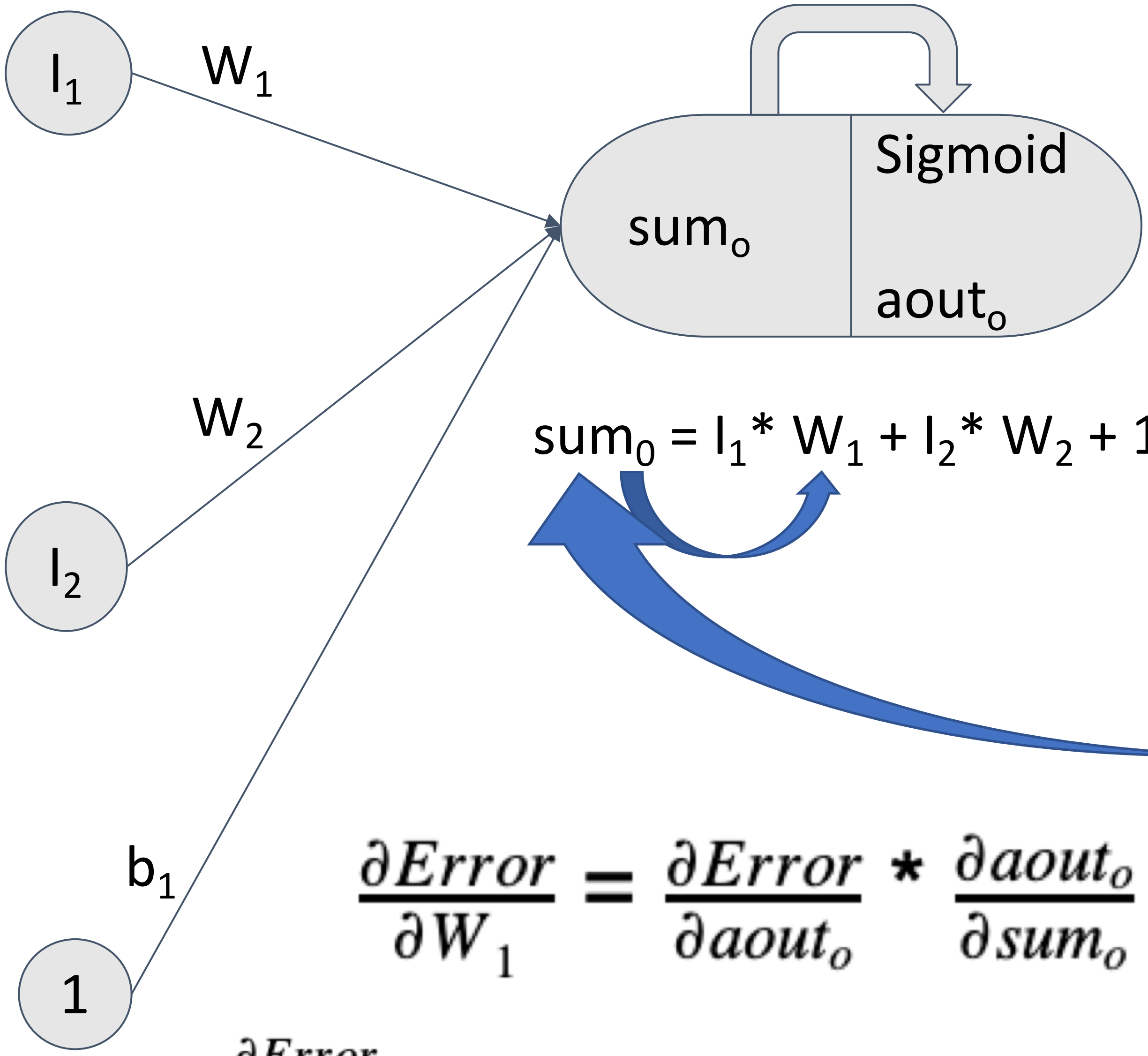
$$\frac{\partial Error}{\partial W_1} = \frac{\partial Error}{\partial aout_o} * \frac{\partial aout_o}{\partial sum_o} * \frac{\partial sum_o}{\partial W_1}$$

$Sigmoid'(sum_0) =$

$$\frac{\partial Error}{\partial W_1} = -(correct - aout_o) *$$

Back Propagation

$$^{*}W_x = W_x - \alpha \left(\frac{\partial Error}{\partial W_x}\right)$$



Error = (correct - $aout_o$)^2 / 2

$aout_0$ = sigmoid(sum_0)

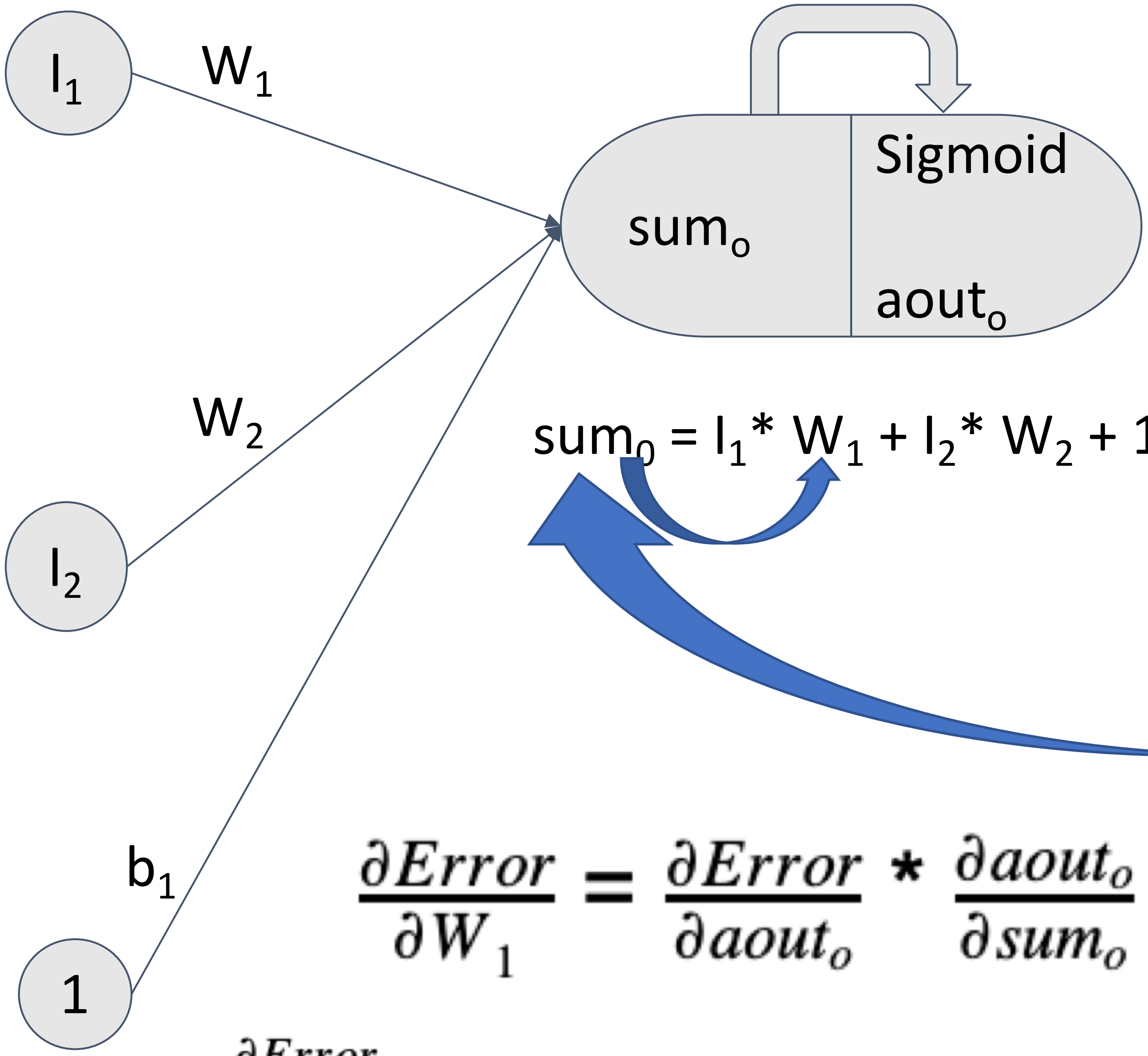
$sum_0 = I_1 * W_1 + I_2 * W_2 + 1 * b_1$

$$\frac{\partial Error}{\partial W_1} = \frac{\partial Error}{\partial aout_o} * \frac{\partial aout_o}{\partial sum_o} * \frac{\partial sum_o}{\partial W_1}$$

$\frac{\partial Error}{\partial W_1}$ = -(correct - $aout_o$) * (sigmoid(sum_0)*(1-sigmoid(sum_0))) *

Back Propagation

$$^{*}W_x = W_x - \alpha\left(\frac{\partial Error}{\partial W_x}\right)$$



Error = (correct - $aout_o)^2$ / 2

$aout_0$ = sigmoid(sum_0)

$sum_0 = I_1 * W_1 + I_2 * W_2 + 1 * b_1$

$$\frac{\partial Error}{\partial W_1} = \frac{\partial Error}{\partial aout_o} * \frac{\partial aout_o}{\partial sum_o} * \frac{\partial sum_o}{\partial W_1}$$

$$\frac{\partial Error}{\partial W_1} = -(correct - aout_o) * (sigmoid(sum_0)*(1-sigmoid(sum_0))) * I_1$$