Backpropagation

CS1090B Data Science II – Spring 2025

Pavlos Protopapas, Natesh Pillai, and Chris Gumb

Sanil Edwin

Gradient Descent Considerations

- We still need to calculate the derivatives.
- We need to set the learning rate.
- Local vs global minima.
- The full likelihood function includes summing up all individual '*errors*'. Sometimes this includes hundreds of thousands of examples.

From Part A

Can we do it? Can we calculate the derivative of any loss function?

Wolfram Alpha can do it for us!

However, we need a formalism to deal with these derivatives.

And we set up a nice formalism using the chain rule!
But now let us talk about details on how to implement this.

From Part A | Chain Rule

Chain rule for computing gradients:

$$y = g(x) \qquad z = f(y) = f\left(g(x)\right)$$

$$\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}$$

$$\boldsymbol{y} = g(\boldsymbol{x}) \qquad z = f(\boldsymbol{y}) = f\left(g(\boldsymbol{x})\right)$$

$$\frac{\partial z}{\partial x_i} = \sum_j \frac{\partial z}{\partial y_j}\frac{\partial y_j}{\partial x_i}$$

For longer chains:

$$\frac{\partial z}{\partial x_i} = \sum_{j_1} \dots \sum_{j_m} \frac{\partial z}{\partial y_{j_1}} \dots \frac{\partial y_{j_m}}{\partial x_i}$$

Logistic Regression Revisited

X → **Affine** → $h = \beta_0 + \beta_1 X$ → **Activation** → $p = \dfrac{1}{1 + e^{-h}}$ → **Loss Fun** → $\mathcal{L}(\beta) = \sum_{i}^{n} \mathcal{L}_i(\beta)$

Logistic Regression Revisited

$$\mathcal{L}_i = -y \log p - (1-y) \log (1-p)$$

X → **Affine** → $h = \beta_0 + \beta_1 X$ → **Activation** → $p = \dfrac{1}{1+e^{-h}}$ → **Loss Fun** → $\mathcal{L}(\beta) = \sum_i^n \mathcal{L}_i(\beta)$

Logistic Regression Revisited

$$\mathcal{L}_i = -y \log p - (1-y) \log (1-p)$$

X → Affine → $h = \beta_0 + \beta_1 X$ → Activation → $p = \dfrac{1}{1+e^{-h}}$ → Loss Fun → $\mathcal{L}(\beta) = \sum_i^n \mathcal{L}_i(\beta)$

$$\frac{\partial \mathcal{L}_i}{\partial p}$$

$$\frac{\partial \mathcal{L}_i}{\partial p} = -y\frac{1}{p} + (1-y)\frac{1}{1-p}$$

Logistic Regression Revisited

$$\mathcal{L}_i = -y \log p - (1-y)\log(1-p)$$

$X \rightarrow$ **Affine** $\rightarrow h = \beta_0 + \beta_1 X \rightarrow$ **Activation** $\rightarrow p = \dfrac{1}{1+e^{-h}} \rightarrow$ **Loss Fun** $\rightarrow \mathcal{L}(\beta) = \sum_i^n \mathcal{L}_i(\beta)$

$$\frac{\partial \mathcal{L}_i}{\partial p}$$

$$\frac{\partial \mathcal{L}_i}{\partial p} = -y\frac{1}{p} + (1-y)\frac{1}{1-p}$$

$\leftarrow$

$$\frac{\partial \mathcal{L}_i}{\partial p}\frac{\partial p}{\partial h}$$

$$\frac{\partial p}{\partial h} = \sigma(h)(1-\sigma(h))$$

Logistic Regression Revisited

$$\mathcal{L}_i = -y \log p - (1-y) \log (1-p)$$

X → Affine → $h = \beta_0 + \beta_1 X$ → Activation → $p = \dfrac{1}{1+e^{-h}}$ → Loss Fun → $\mathcal{L}(\beta) = \sum_i^n \mathcal{L}_i(\beta)$

$$\frac{\partial \mathcal{L}_i}{\partial p}\frac{\partial p}{\partial h}\frac{\partial h}{\partial \beta}$$

$$\frac{\partial h}{\partial \beta_1} = X, \frac{dh}{d\beta_0} = 1$$

←

$$\frac{\partial \mathcal{L}_i}{\partial p}\frac{\partial p}{\partial h}$$

$$\frac{\partial p}{\partial h} = \sigma(h)(1-\sigma(h))$$

←

$$\frac{\partial \mathcal{L}_i}{\partial p}$$

$$\frac{\partial \mathcal{L}_i}{\partial p} = -y\frac{1}{p} + (1-y)\frac{1}{1-p}$$

Logistic Regression Revisited

$$\mathcal{L}_i = -y \log p - (1-y)\log(1-p)$$

$X \rightarrow$ Affine $\rightarrow h = \beta_0 + \beta_1 X \rightarrow$ Activation $\rightarrow p = \frac{1}{1+e^{-h}} \rightarrow$ Loss Fun $\rightarrow \mathcal{L}(\beta) = \sum_i^n \mathcal{L}_i(\beta)$

$$\frac{\partial \mathcal{L}_i}{\partial p}\frac{\partial p}{\partial h}\frac{\partial h}{\partial \beta}$$

$$\frac{\partial h}{\partial \beta_1} = X, \frac{dh}{d\beta_0} = 1$$

$\leftarrow$

$$\frac{\partial \mathcal{L}_i}{\partial p}\frac{\partial p}{\partial h}$$

$$\frac{\partial p}{\partial h} = \sigma(h)(1-\sigma(h))$$

$\leftarrow$

$$\frac{\partial \mathcal{L}_i}{\partial p}$$

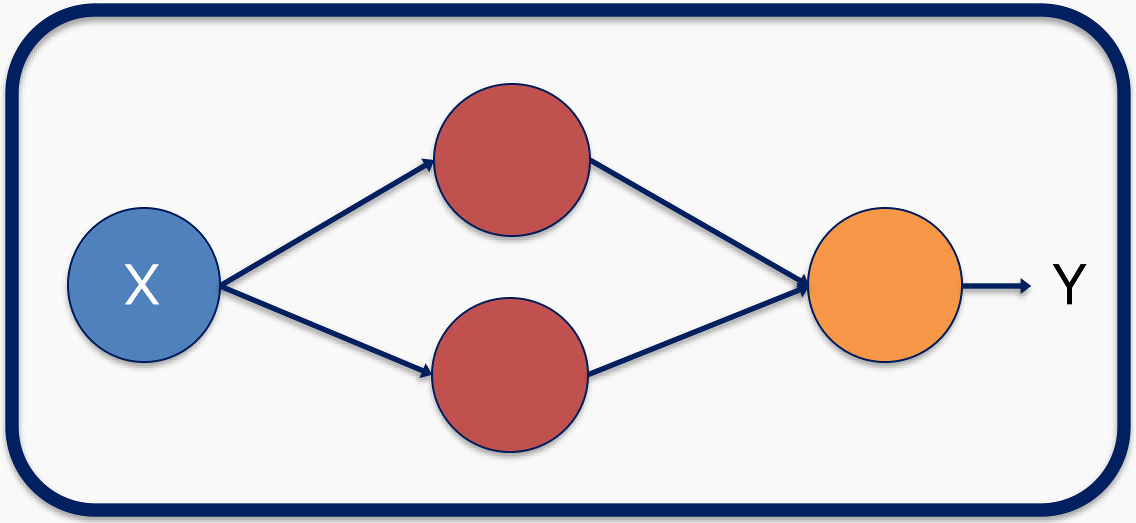
$$\frac{\partial \mathcal{L}_i}{\partial p} = -y\frac{1}{p} + (1-y)\frac{1}{1-p}$$

$$\boxed{\frac{\partial \mathcal{L}_i}{\partial \beta_1} = \frac{\partial \mathcal{L}_i}{\partial p}\frac{\partial p}{\partial h}\frac{\partial h}{\partial \beta_1} = -X\sigma(h)\big(1-\sigma(h)\big)\left[y\frac{1}{p} - (1-y)\frac{1}{1-p}\right]}$$

$$\boxed{\frac{\partial \mathcal{L}_i}{\partial \beta_0} = \frac{\partial \mathcal{L}_i}{\partial p}\frac{\partial p}{\partial h}\frac{\partial h}{\partial \beta_0} = -\sigma(h)\big(1-\sigma(h)\big)\left[y\frac{1}{p} - (1-y)\frac{1}{1-p}\right]}$$

Motivation | Backpropagation

Forward pass
(Making predictions)



Loss Function

Calculate **loss**

We want to minimize this loss. How do we do that?

All the weights are adjusted such that the loss is minimized. The error is propagated in the **backwards** direction from the rear end to the layers at the front.

Hence the name, "**Back**propagation" !

Motivation | Backpropagation

- We now know that the derivatives need to be evaluated at specific values of X, y and W.

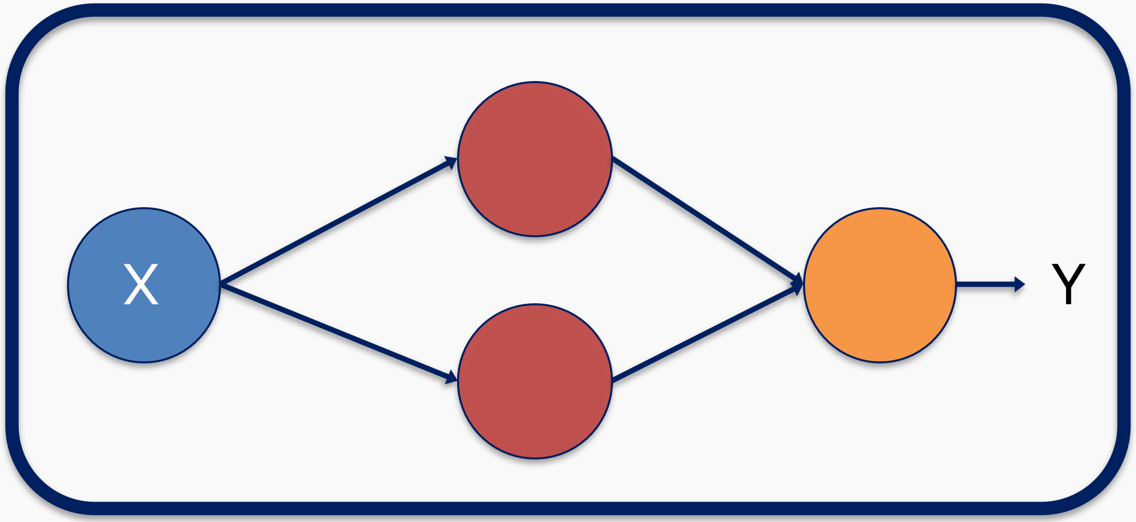
- Since we have an expression for the derivative, we can build a function that takes as input X, y and W and returns the derivatives. We can then use gradient descent to update the weights.

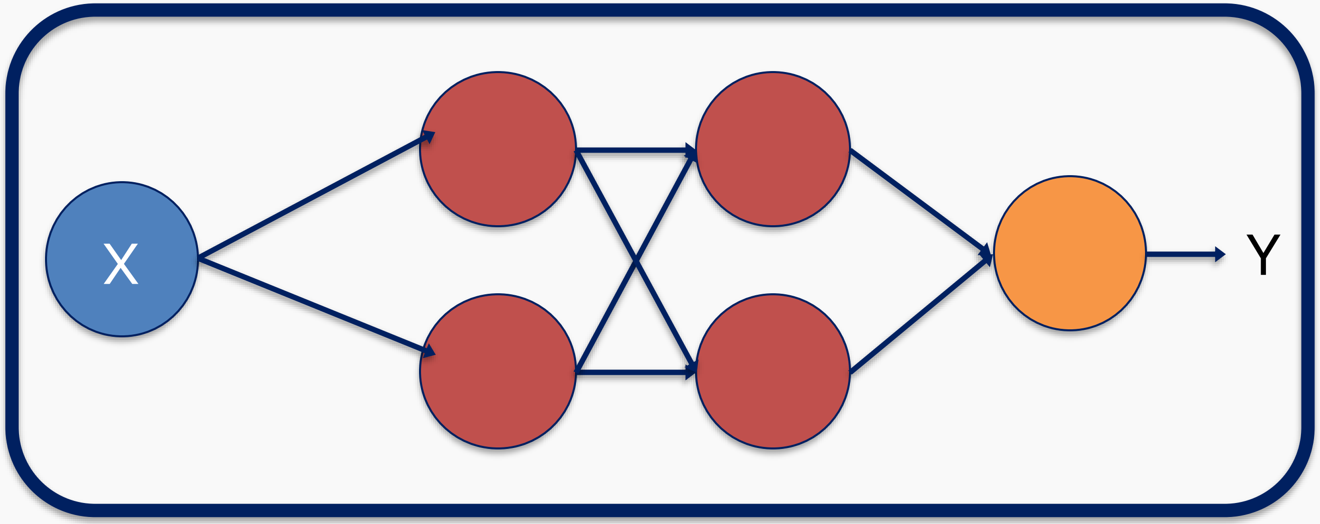
But what is wrong with this approach?

Motivation | Backpropagation

This approach works well, but it does not generalize.

For example, if the network is changed, we need to write a new function to evaluate the derivatives.

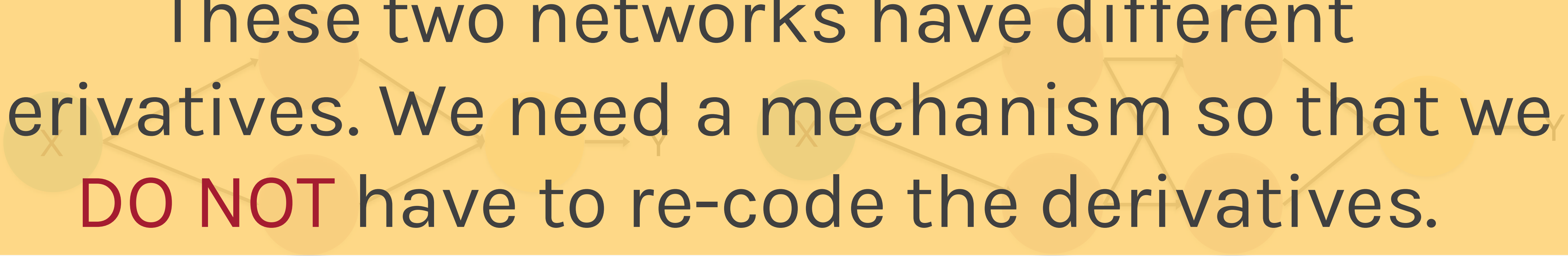




Motivation | Backpropagation

This approach works well, but it does not generalize.

For example, if the network is changed, we need to write a new function to evaluate the derivatives.



These two networks have different derivatives. We need a mechanism so that we DO NOT have to re-code the derivatives.

Motivation | Backpropagation

We need to find a formalism to calculate the derivatives of the loss w.r.t weights that is:

1. Flexible enough that adding a node or a layer or changing something in the network will not require re-deriving the functional form from scratch.
2. It is exact.
3. It is computationally efficient.

Auto-differentiation to the rescue!

For example, for input *X={3}, y=1* and weight *W=3,* we evaluate the values of the variables, partial derivatives and the chain up to this point as shown below.

Variables	Derivatives	Value of the variable	Value of the derivative	$\frac{d\xi_i}{dW}$
$\xi_1 = -W^T X$	$\frac{\partial \xi_1}{\partial W} = -X$	-9	-3	-3
$\xi_2 = e^{\xi_1} = e^{-W^T X}$	$\frac{\partial \xi_2}{\partial \xi_1} = e^{\xi_1}$	e^{-9}	e^{-9}	-3e^{-9}
$\xi_3 = 1 + \xi_2 = 1 + e^{-W^T X}$	$\frac{\partial \xi_3}{\partial \xi_2} = 1$	1+e^{-9}	1	-3e^{-9}
$\xi_4 = \frac{1}{\xi_3} = \frac{1}{1 + e^{-W^T X}} = p$	$\frac{\partial \xi_4}{\partial \xi_3} = -\frac{1}{\xi_3^2}$	$\frac{1}{1 + e^{-9}}$	$-\left(\frac{1}{1 + e^{-9}}\right)^2$	$3e^{-9}\left(\frac{1}{1+e^{-9}}\right)^2$
$\xi_5 = \log \xi_4 = \log p \ = \log \frac{1}{1 + e^{-W^T X}}$	$\frac{\partial \xi_5}{\partial \xi_4} = \frac{1}{\xi_4}$	$\log \frac{1}{1 + e^{-9}}$	$1 + e^{-9}$	$3e^{-9}\left(\frac{1}{1+e^{-9}}\right)$
$\mathcal{L}_i^A = -y\xi_5$	$\frac{\partial \mathcal{L}}{\partial \xi_5} = -y$	$-\log \frac{1}{1 + e^{-9}}$	-1	-3$e^{-9}\left(\frac{1}{1+e^{-9}}\right)$
$\frac{\partial \mathcal{L}_i^A}{\partial W} = \frac{\partial \mathcal{L}_i}{\partial \xi_5}\frac{\partial \xi_5}{\partial \xi_4}\frac{\partial \xi_4}{\partial \xi_3}\frac{\partial \xi_3}{\partial \xi_2}\frac{\partial \xi_2}{\partial \xi_1}\frac{\partial \xi_1}{\partial W}$				-0.00037018372

BUT we still need to know the derivatives 😱

Variables	Derivatives	Value of the variable	Value of the derivative	$\frac{d\xi_i}{dW}$
$\xi_1 = -W^T X$	$\frac{\partial \xi_1}{\partial W} = -X$	-9	-3	-3
$\xi_2 = e^{\xi_1} = e^{-W^T X}$	$\frac{\partial \xi_2}{\partial \xi_1} = e^{\xi_1}$	e^{-9}	e^{-9}	-3e^{-9}
$\xi_3 = 1 + \xi_2 = 1 + e^{-W^T X}$	$\frac{\partial \xi_3}{\partial \xi_2} = 1$	1+e^{-9}	1	-3e^{-9}
$\xi_4 = \frac{1}{\xi_3} = \frac{1}{1 + e^{-W^T X}} = p$	$\frac{\partial \xi_4}{\partial \xi_3} = -\frac{1}{\xi_3^2}$	$\frac{1}{1 + e^{-9}}$	$-\left(\frac{1}{1 + e^{-9}}\right)^2$	3$e^{-9}\left(\frac{1}{1+e^{-9}}\right)^2$
$\xi_5 = \log \xi_4 = \log p \ = \log \frac{1}{1 + e^{-W^T X}}$	$\frac{\partial \xi_5}{\partial \xi_4} = \frac{1}{\xi_4}$	$\log \frac{1}{1 + e^{-9}}$	$1 + e^{-9}$	3$e^{-9}\left(\frac{1}{1+e^{-9}}\right)$
$\mathcal{L}_i^A = -y\xi_5$	$\frac{\partial \mathcal{L}}{\partial \xi_5} = -y$	$-\log \frac{1}{1 + e^{-9}}$	-1	-3$e^{-9}\left(\frac{1}{1+e^{-9}}\right)$
$\frac{\partial \mathcal{L}_i^A}{\partial W} = \frac{\partial \mathcal{L}_i}{\partial \xi_5} \frac{\partial \xi_5}{\partial \xi_4} \frac{\partial \xi_4}{\partial \xi_3} \frac{\partial \xi_3}{\partial \xi_2} \frac{\partial \xi_2}{\partial \xi_1} \frac{\partial \xi_1}{\partial W}$				-0.00037018372

BUT we still need to know the derivatives 😱

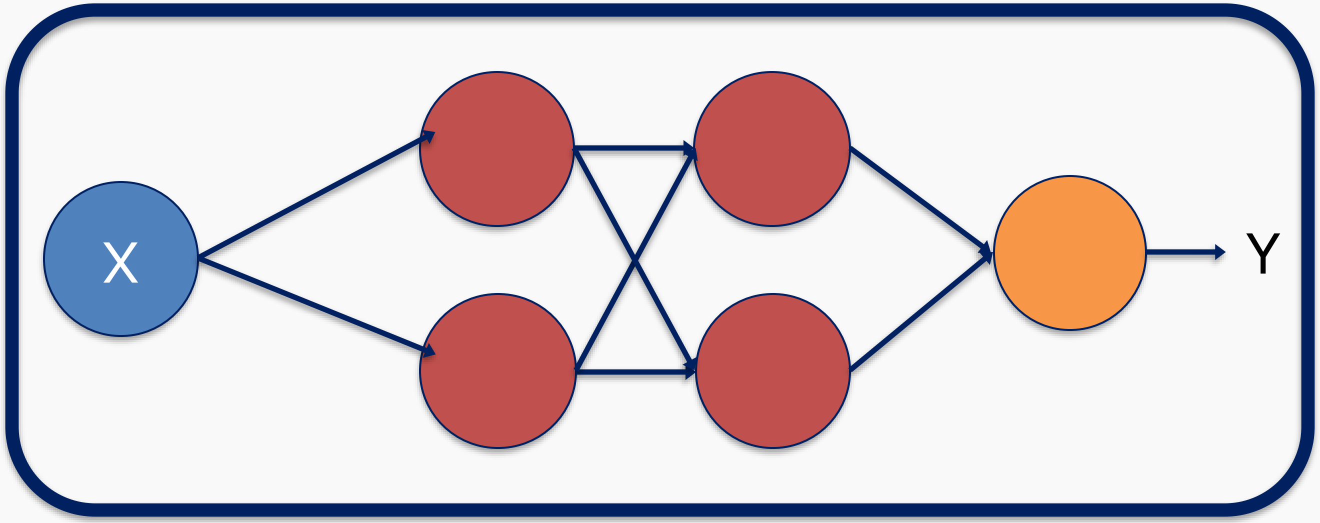
Variables	Derivatives	Value of the variable	Value of the derivative	$\frac{d\xi_i}{dW}$
$\xi_1 = -W^T X$	$\frac{\partial \xi_1}{\partial W} = -X$	-9	-3	-3
$\xi_2 = e^{\xi_1} = e^{-W^T X}$	$\frac{\partial \xi_2}{\partial \xi_1} = e^{\xi_1}$	e^{-9}	e^{-9}	-3e^{-9}
$\xi_3 = 1 + \xi_2 = 1 + e^{-W^T X}$	$\frac{\partial \xi_3}{\partial \xi_2} = 1$	1+e^{-9}	1	-3e^{-9}
$\xi_4 = \frac{1}{\xi_3} = \frac{1}{1 + e^{-W^T X}} = p$	$\frac{\partial \xi_4}{\partial \xi_3} = -\frac{1}{\xi_3^2}$	$\frac{1}{1+e^{-9}}$	$-\left(\frac{1}{1+e^{-9}}\right)^2$	$3e^{-9}\left(\frac{1}{1+e^{-9}}\right)^2$
$\xi_5 = \log \xi_4 = \log p = \log \frac{1}{1 + e^{-W^T X}}$	$\frac{\partial \xi_5}{\partial \xi_4} = \frac{1}{\xi_4}$	$\log \frac{1}{1+e^{-9}}$	$1 + e^{-9}$	$3e^{-9}\left(\frac{1}{1+e^{-9}}\right)$
$\mathcal{L}_i^A = -y\xi_5$	$\frac{\partial \mathcal{L}}{\partial \xi_5} = -y$	$-\log \frac{1}{1+e^{-9}}$	-1	-3$e^{-9}\left(\frac{1}{1+e^{-9}}\right)$
$\frac{\partial \mathcal{L}_i^A}{\partial W} = \frac{\partial \mathcal{L}_i}{\partial \xi_5}\frac{\partial \xi_5}{\partial \xi_4}\frac{\partial \xi_4}{\partial \xi_3}\frac{\partial \xi_3}{\partial \xi_2}\frac{\partial \xi_2}{\partial \xi_1}\frac{\partial \xi_1}{\partial W}$				-0.00037018372

Notice though those are basic functions (simpleton functions) which are easy to code.

$\xi_0 = X$	$\frac{\partial \xi_0}{\partial X} = 1$	`def x0(x):` `    return x`	`def derx0():` `    return 1`
$\xi_1 = -W^T \xi_0$	$\frac{\partial \xi_1}{\partial W} = -X$	`def x1(a,x):` `    return -a*x`	`def derx1(a,x):` `    return -a`
$\xi_2 = \mathrm{e}^{\xi_1}$	$\frac{\partial \xi_2}{\partial \xi_1} = e^{\xi_1}$	`def x2(x):` `    return np.exp(x)`	`def derx2(x):` `    return np.exp(x)`
$\xi_3 = 1 + \xi_2$	$\frac{\partial \xi_3}{\partial \xi_2} = 1$	`def x3(x):` `    return 1+x`	`def derx3(x):` `    return 1`
$\xi_4 = \frac{1}{\xi_3}$	$\frac{\partial \xi_4}{\partial \xi_3} = -\frac{1}{\xi_3^2}$	`def x4(x):` `    return 1/(x)`	`def derx4(x):` `    return  -(1/x)**(2)`
$\xi_5 = \log \xi_4$	$\frac{\partial \xi_5}{\partial \xi_4} = \frac{1}{\xi_4}$	`def x5(x):` `    return np.log(x)`	`def derx5(x)` `    return 1/x`
$\mathcal{L}_i^A = -y\xi_5$	$\frac{\partial \mathcal{L}}{\partial \xi_5} = -y$	`def L(y,x):` `    return -y*x`	`def derL(y):` `    return -y`

Putting it altogether

1. We specify the network structure.



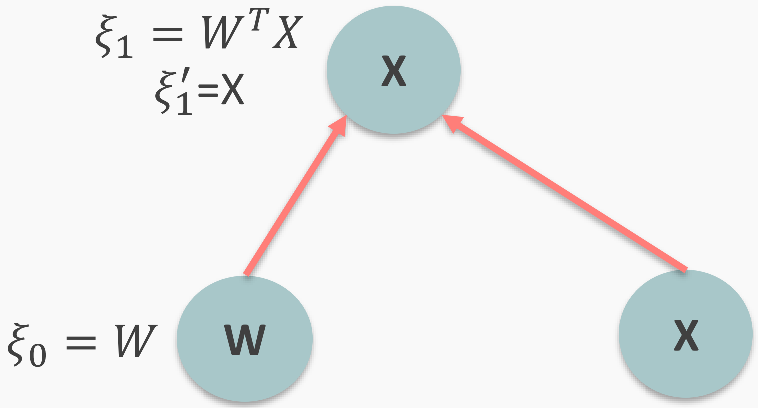
This magic happens when we do **model.compile()**

2. Build the computational graph.

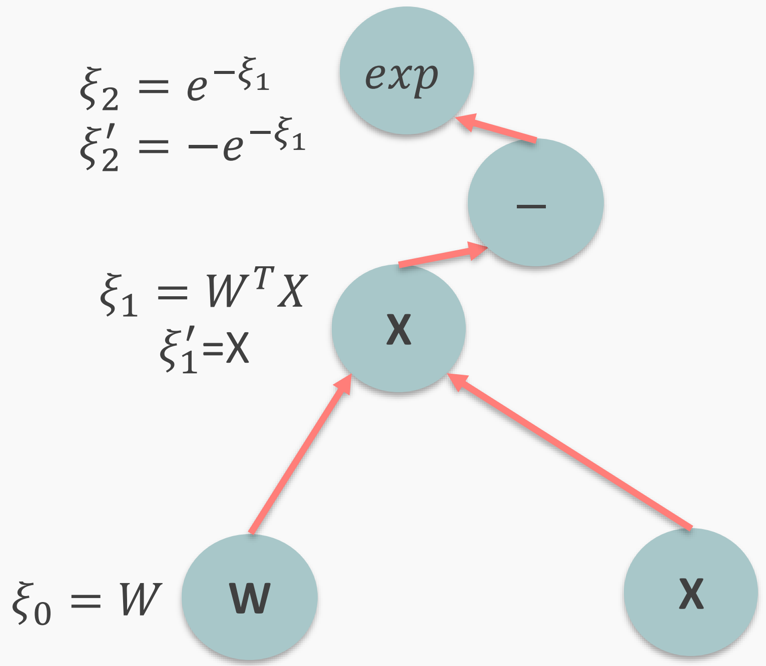
At each node of the graph, we build two functions: the evaluation of the variable and its partial derivative for the previous variables.

Computational Graph

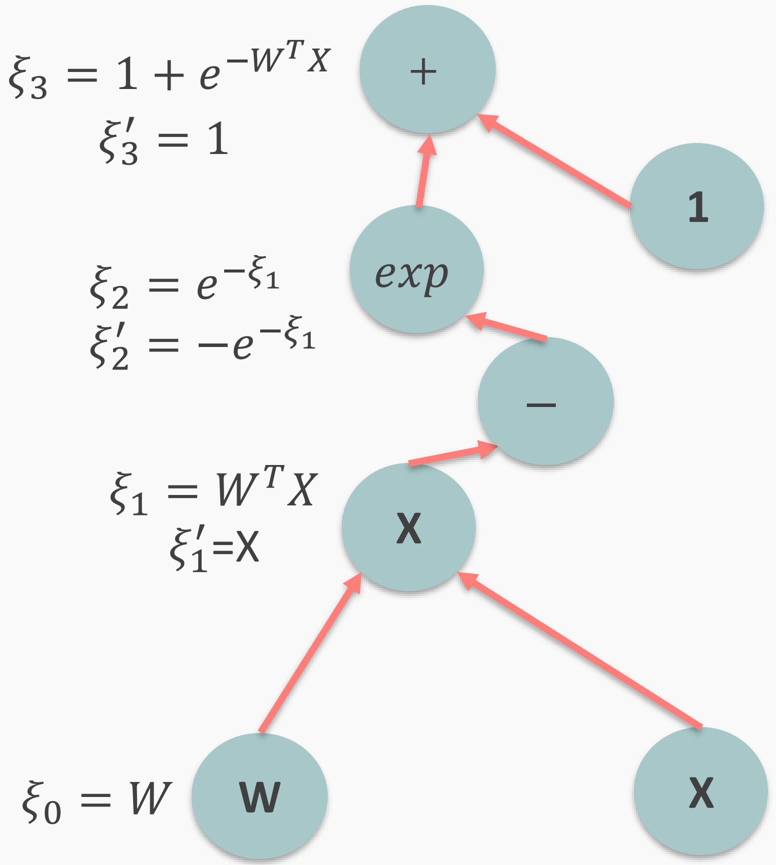
Computational Graph



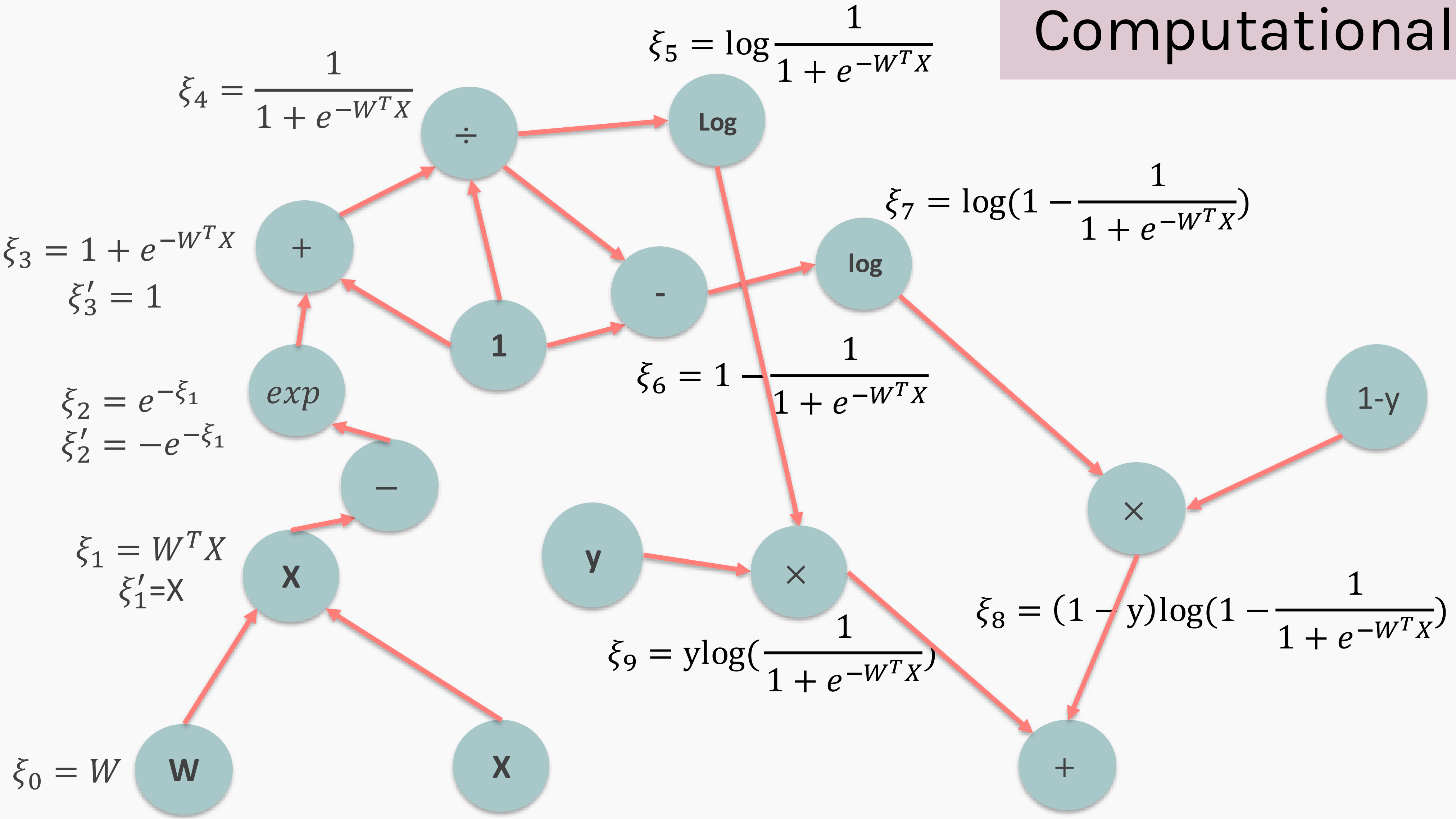
Computational Graph



Computational Graph

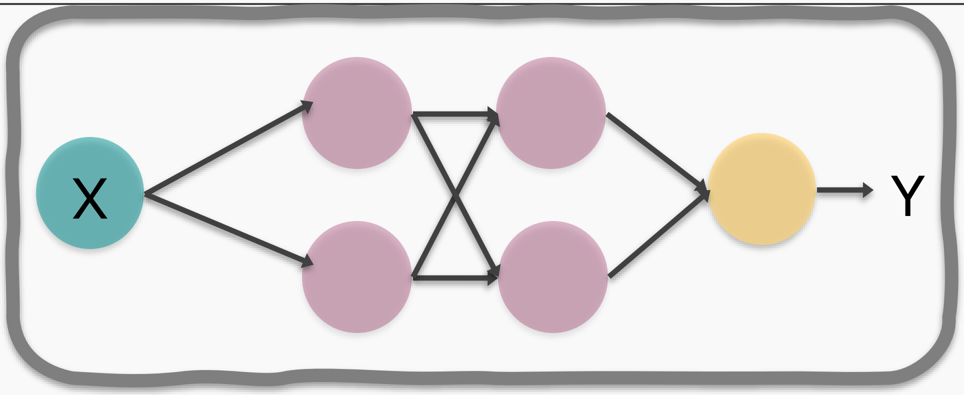


Computational Graph



$$-\mathcal{L} = \xi_9 = \mathrm{y}\log(\frac{1}{1+e^{-W^T X}}) + (1-\mathrm{y})\log(1 - \frac{1}{1+e^{-W^T X}})$$

Putting it altogether

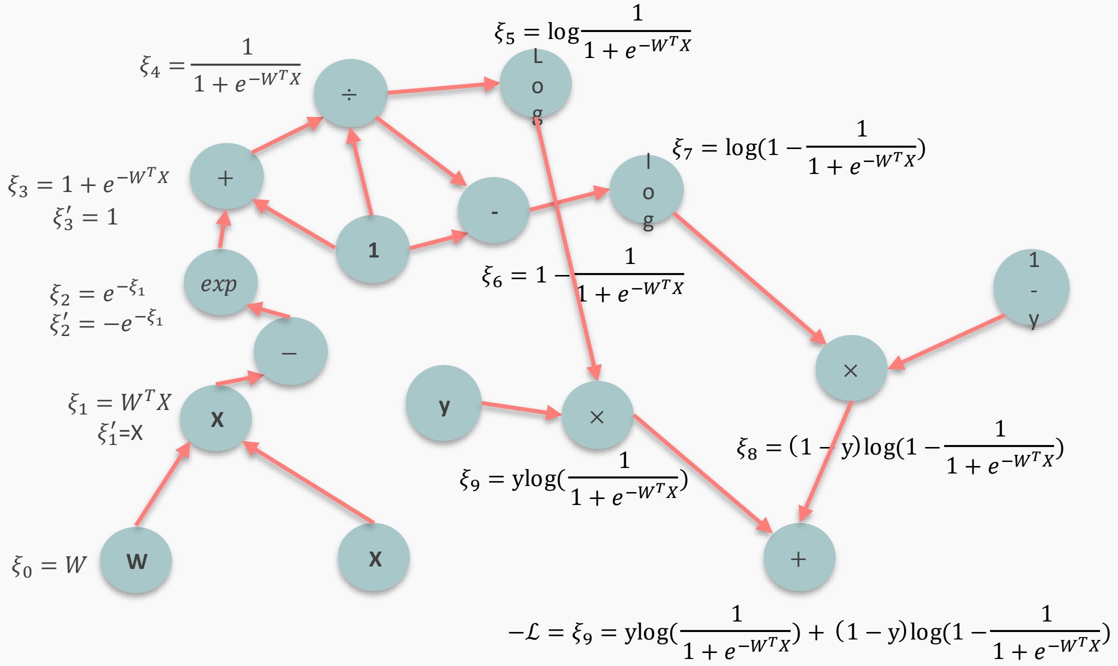


1. We specify the network structure.

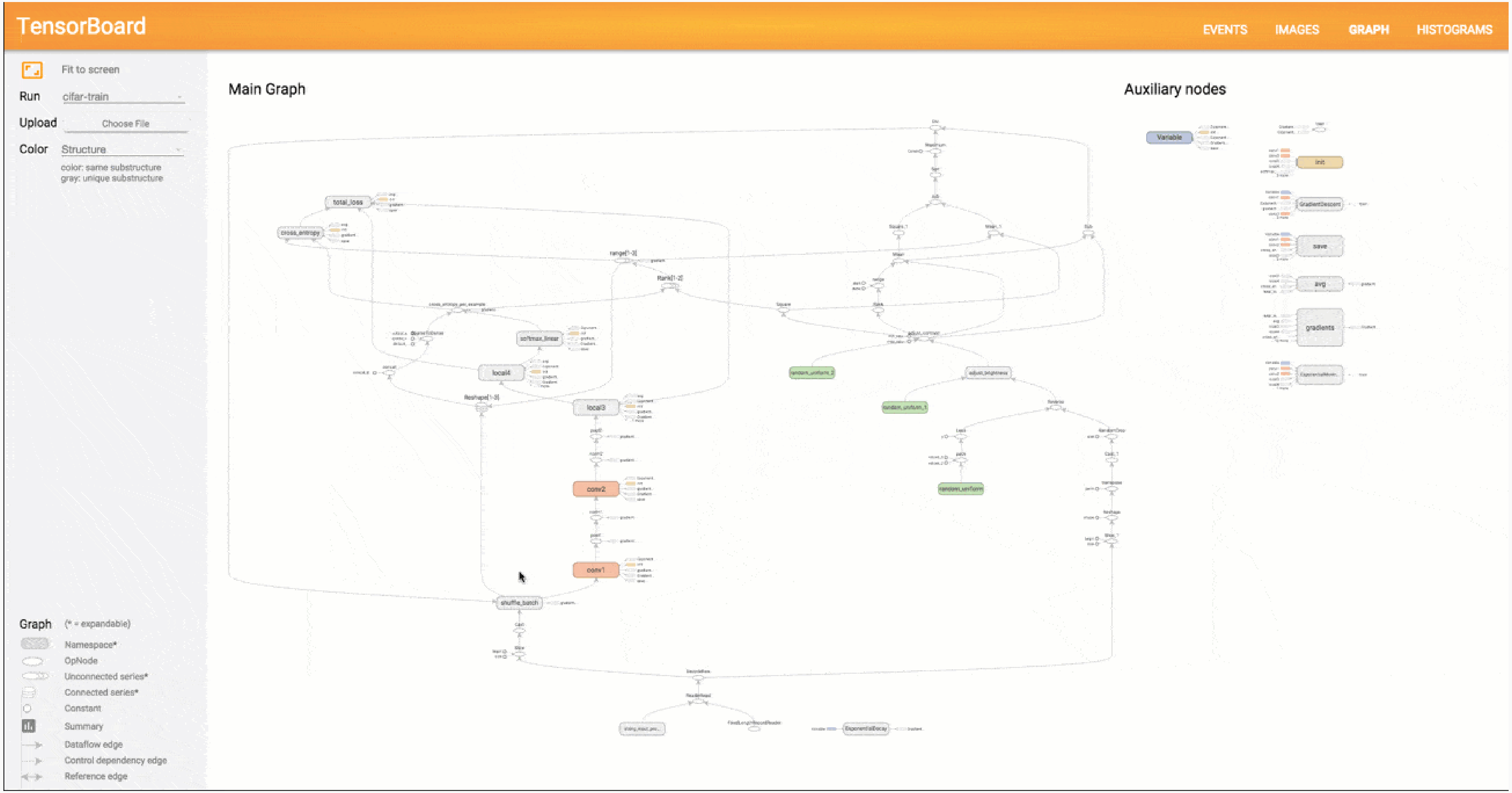
This magic happens when we do model.compile()

2. Build the computational graph.

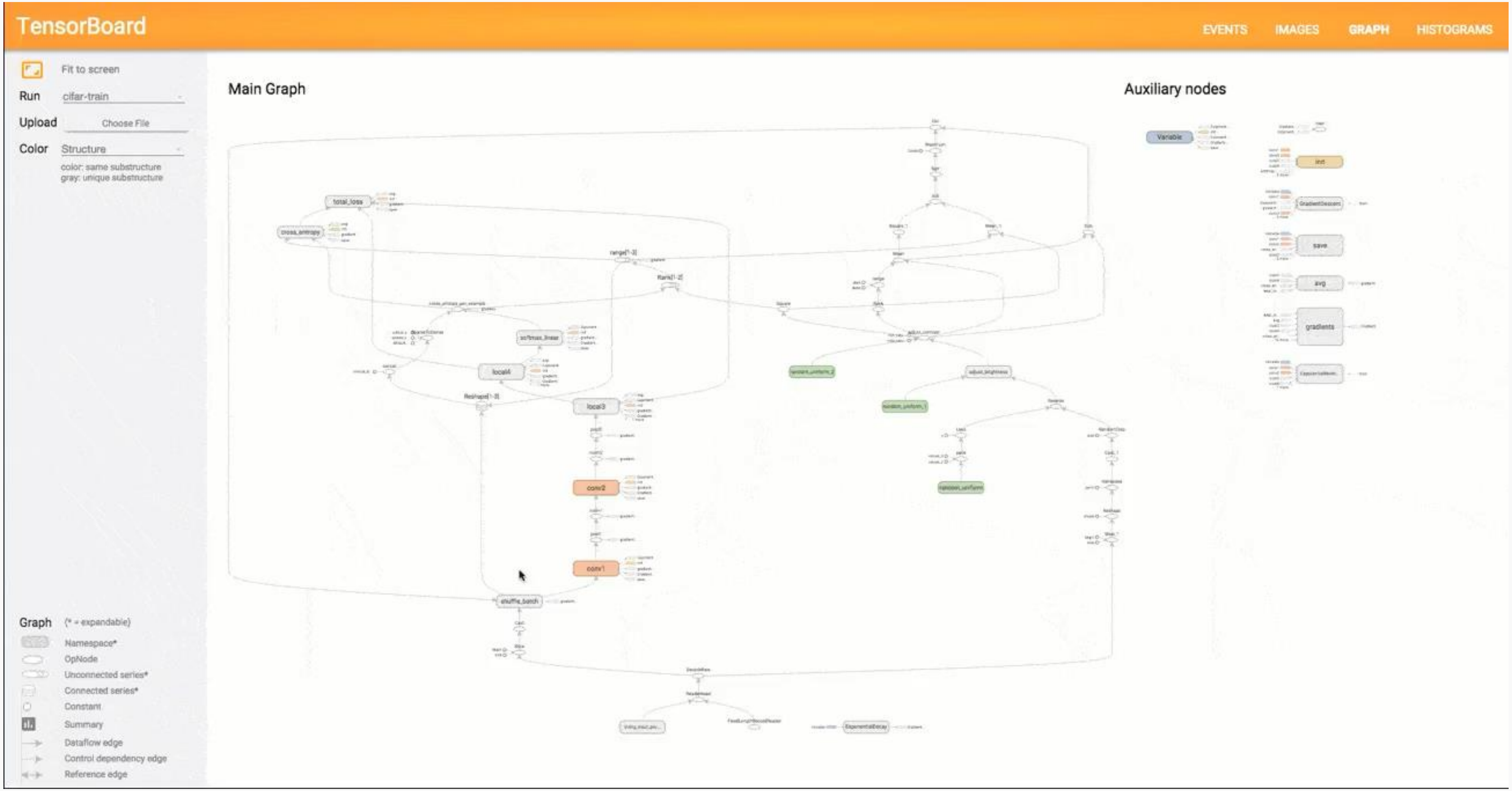
At each node of the graph, we build two functions: the evaluation of the variable and its partial derivative with respect to the previous variables (as shown in the table a few slides back)



Computational Graph - Tensorboard



Computational Graph - Tensorboard



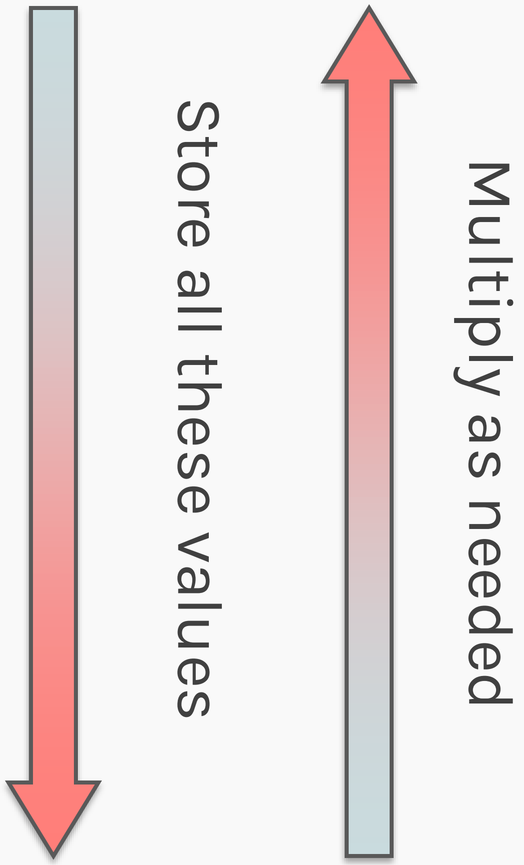
source

Forward mode: Evaluate the derivative at: $X=\{3\}, y=1, W=3$

Variables	Derivative	Value of the variable	Value of the derivative	$\frac{d\xi_i}{dW}$
$\xi_1 = -W^T X$	$\frac{\partial \xi_1}{\partial W} = -X$	-9	-3	-3
$\xi_2 = e^{\xi_1} = e^{-W^T X}$	$\frac{\partial \xi_2}{\partial \xi_1} = e^{\xi_1}$	e^{-9}	e^{-9}	-3e^{-9}
$\xi_3 = 1 + \xi_2 = 1 + e^{-W^T X}$	$\frac{\partial \xi_3}{\partial \xi_2} = 1$	1+e^{-9}	1	-3e^{-9}
$\xi_4 = \frac{1}{\xi_3} = \frac{1}{1 + e^{-W^T X}} = p$	$\frac{\partial \xi_4}{\partial \xi_3} = -\frac{1}{\xi_3^2}$	$\frac{1}{1 + e^{-9}}$	$-\left(\frac{1}{1 + e^{-9}}\right)^2$	$3e^{-9}\left(\frac{1}{1+e^{-9}}\right)^2$
$\xi_5 = \log \xi_4 = \log p = \log \frac{1}{1 + e^{-W^T}}$	$\frac{\partial \xi_5}{\partial \xi_4} = \frac{1}{\xi_4}$	$\log \frac{1}{1 + e^{-9}}$	$1 + e^{-9}$	$3e^{-9}\left(\frac{1}{1+e^{-9}}\right)$
$\mathcal{L}_i^A = -y\xi_5$	$\frac{\partial \mathcal{L}}{\partial \xi_5} = -y$	$-\log \frac{1}{1 + e^{-9}}$	-1	-3$e^{-9}\left(\frac{1}{1+e^{-9}}\right)$
$\frac{\partial \mathcal{L}_i^A}{\partial W} = \frac{\partial \mathcal{L}_i}{\partial \xi_5}\frac{\partial \xi_5}{\partial \xi_4}\frac{\partial \xi_4}{\partial \xi_3}\frac{\partial \xi_3}{\partial \xi_2}\frac{\partial \xi_2}{\partial \xi_1}\frac{\partial \xi_1}{\partial W}$				-0.00037018372

Reverse mode: Evaluate the derivative at: $X=\{3\}, y=1, W=3$

Variables	Derivatives	Value of the variable	Value of the derivative
$\xi_1 = -W^T X$	$\frac{\partial \xi_1}{\partial W} = -X$	-9	-3
$\xi_2 = e^{\xi_1} = e^{-W^T X}$	$\frac{\partial \xi_2}{\partial \xi_1} = e^{\xi_1}$	e^{-9}	e^{-9}
$\xi_3 = 1 + \xi_2 = 1 + e^{-W^T X}$	$\frac{\partial \xi_3}{\partial \xi_2} = 1$	1+e^{-9}	1
$\xi_4 = \frac{1}{\xi_3} = \frac{1}{1+e^{-W^T X}} = p$	$\frac{\partial \xi_4}{\partial \xi_3} = -\frac{1}{\xi_3^2}$	$\frac{1}{1+e^{-9}}$	$-\left(\frac{1}{1+e^{-9}}\right)^2$
$\xi_5 = \log \xi_4 = \log p = \log \frac{1}{1+e^{-W^T X}}$	$\frac{\partial \xi_5}{\partial \xi_4} = \frac{1}{\xi_4}$	$\log \frac{1}{1+e^{-9}}$	$1 + e^{-9}$
$\mathcal{L}_i^A = -y\xi_5$	$\frac{\partial \mathcal{L}}{\partial \xi_5} = -y$	$-\log \frac{1}{1+e^{-9}}$	-1
$\frac{\partial \mathcal{L}_i^A}{\partial W} = \frac{\partial \mathcal{L}_i}{\partial \xi_5}\frac{\partial \xi_5}{\partial \xi_4}\frac{\partial \xi_4}{\partial \xi_3}\frac{\partial \xi_3}{\partial \xi_2}\frac{\partial \xi_2}{\partial \xi_1}\frac{\partial \xi_1}{\partial W}$			



Forward v/s Reverse mode

m is the number of outputs
n is the number of inputs

1. When doing automatic differentiation, using the forward mode will be helpful when $m > n$,

2. But usually, we have more input features than outputs, i.e. $m < n$. So, we use the reverse mode as it reduces the computational complexity.

3. Backprop originally meant just using the chain rule, but usually it refers to reverse mode automatic differentiation.

Thank you