Backpropagation

CS109B Data Science 2

Pavlos Protopapas, Mark Glickman

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.**

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

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

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

Logistic Regression Revisited

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

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

$$\boxed{\begin{gathered}\frac{\partial \mathcal{L}}{\partial p} \\ \frac{\partial \mathcal{L}}{\partial p} = -y\frac{1}{p} - (1-y)\frac{1}{1-p}\end{gathered}}$$

Logistic Regression Revisited

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

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

$$\boxed{\begin{array}{c} \dfrac{\partial \mathcal{L}}{\partial p}\dfrac{\partial p}{\partial h} \\[2ex] \dfrac{\partial p}{\partial h} = \sigma(h)(1-\sigma(h)) \end{array}} \longleftarrow \boxed{\begin{array}{c} \dfrac{\partial \mathcal{L}}{\partial p} \\[2ex] \dfrac{\partial \mathcal{L}}{\partial p} = -y\dfrac{1}{p} - (1-y)\dfrac{1}{1-p} \end{array}}$$

Logistic Regression Revisited

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

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

$$\boxed{\begin{gathered}\frac{\partial \mathcal{L}}{\partial p}\frac{\partial p}{\partial h}\frac{\partial h}{\partial \beta} \\[1em] \frac{\partial h}{\partial \beta_1} = X, \frac{d\mathcal{L}}{d\beta_0} = 1\end{gathered}} \longleftarrow \boxed{\begin{gathered}\frac{\partial \mathcal{L}}{\partial p}\frac{\partial p}{\partial h} \\[1em] \frac{\partial p}{\partial h} = \sigma(h)(1-\sigma(h))\end{gathered}} \longleftarrow \boxed{\begin{gathered}\frac{\partial \mathcal{L}}{\partial p} \\[1em] \frac{\partial \mathcal{L}}{\partial p} = -y\frac{1}{p} - (1-y)\frac{1}{1-p}\end{gathered}}$$

Logistic Regression Revisited

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

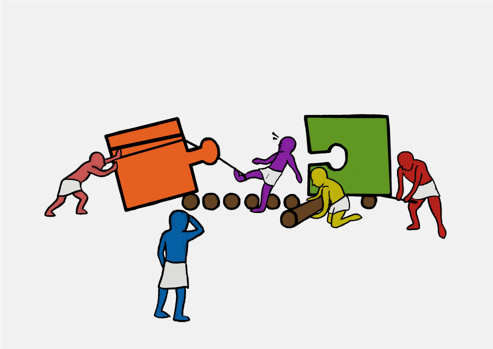
$$X \longrightarrow \boxed{\text{Affine}} \longrightarrow h = \beta_0 + \beta_1 X \longrightarrow \boxed{\text{Activation}} \longrightarrow p = \frac{1}{1+e^{-h}} \longrightarrow \boxed{\text{Loss Fun}} \longrightarrow \mathcal{L}(\beta) = \sum_i^n \mathcal{L}_i(\beta)$$

$$\boxed{\frac{\partial \mathcal{L}}{\partial p}\frac{\partial p}{\partial h}\frac{\partial h}{\partial \beta} \qquad \frac{\partial h}{\partial \beta_1} = X, \frac{d\mathcal{L}}{d\beta_0} = 1} \longleftarrow \boxed{\frac{\partial \mathcal{L}}{\partial p}\frac{\partial p}{\partial h} \qquad \frac{\partial p}{\partial h} = \sigma(h)(1-\sigma(h))} \longleftarrow \boxed{\frac{\partial \mathcal{L}}{\partial p} \qquad \frac{\partial \mathcal{L}}{\partial p} = -y\frac{1}{p} - (1-y)\frac{1}{1-p}}$$

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

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

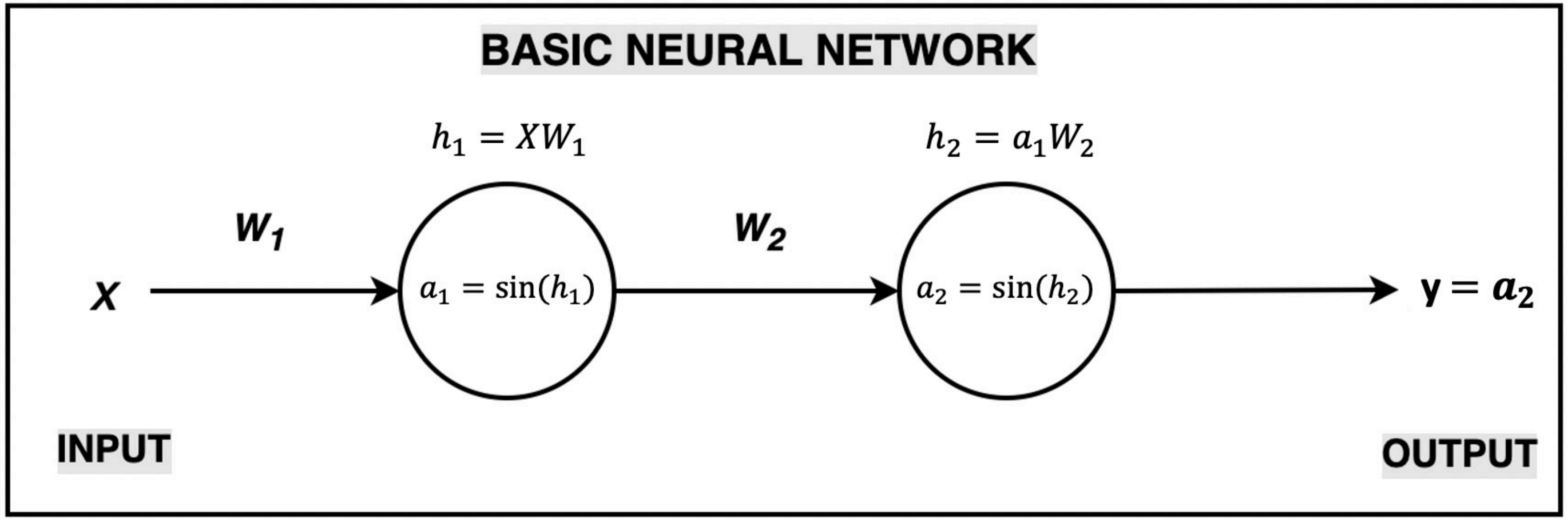
Exercise: Back-propagation by hand



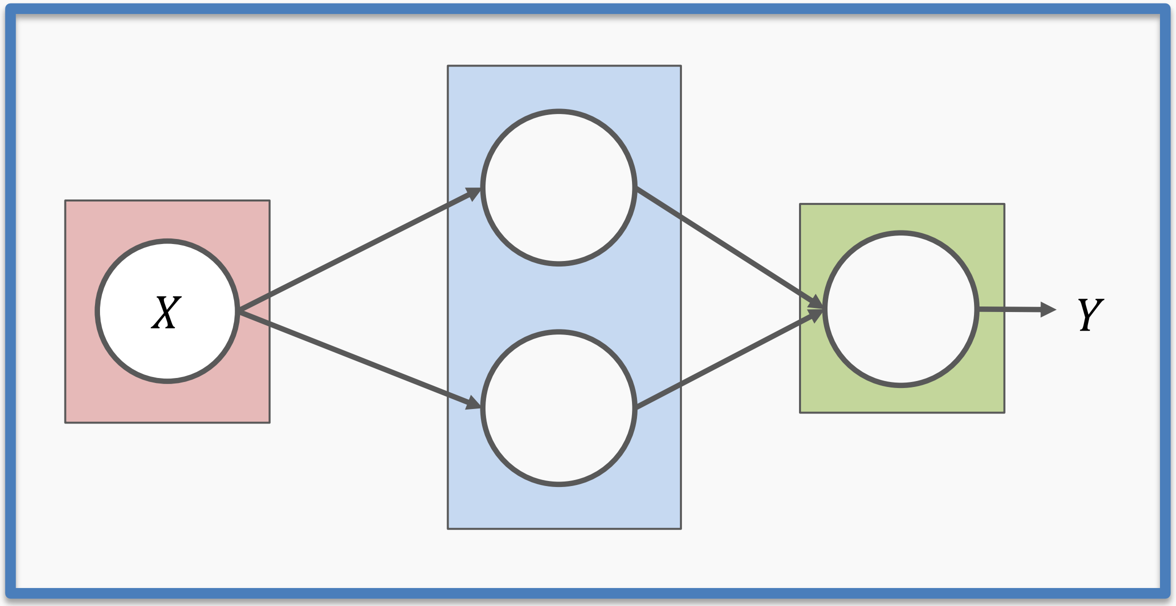
The aim of this exercise is to perform back-propagation to update the weights of a simple neural network

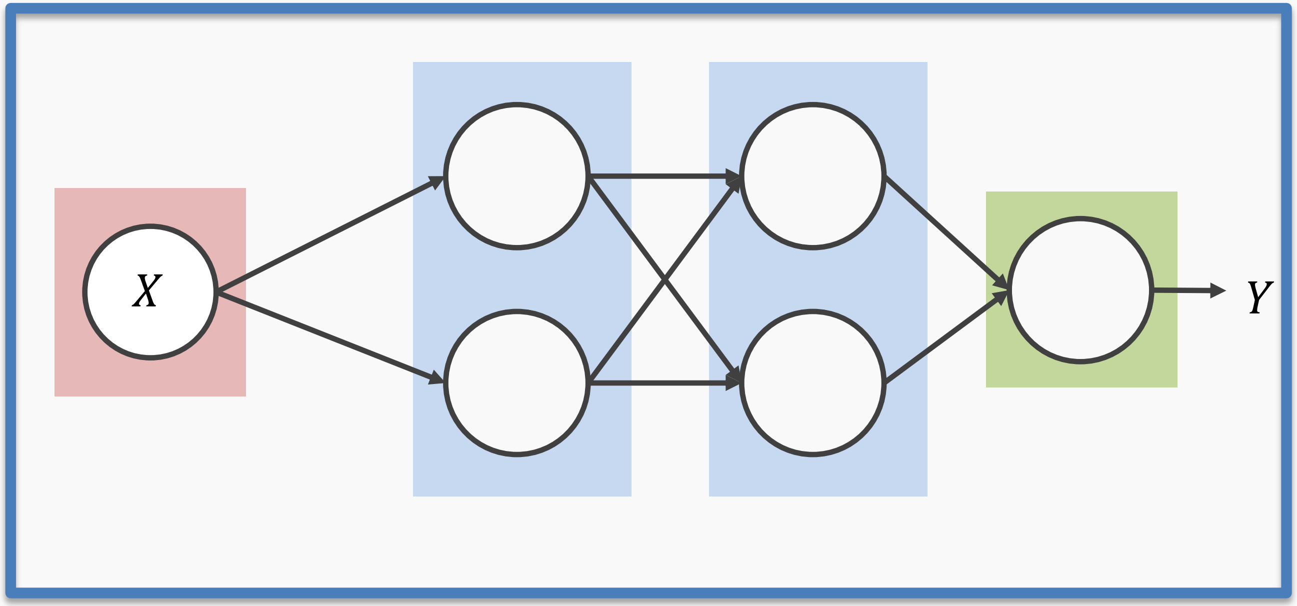
- Build a forward pass of the simple neural network with one hidden layer (see schematic below)
- Randomly initialize the weights
- Use the derivatives to update the weights
- You will need a paper and pen to derive $\frac{\partial L}{\partial W_1}$ & $\frac{\partial L}{\partial W_2}$

$$L = \frac{1}{n}\sum_{1}^{n}\left(y_{pred} - y_{true}\right)^2$$



1. Derivatives need to be evaluated at some values of *X, y,* and *Ws*.
2. But since we have an expression for the derivative, we can build a function that takes as input *X, y, W,* and returns the derivatives, and then we can use gradient descent to update.
3. 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.

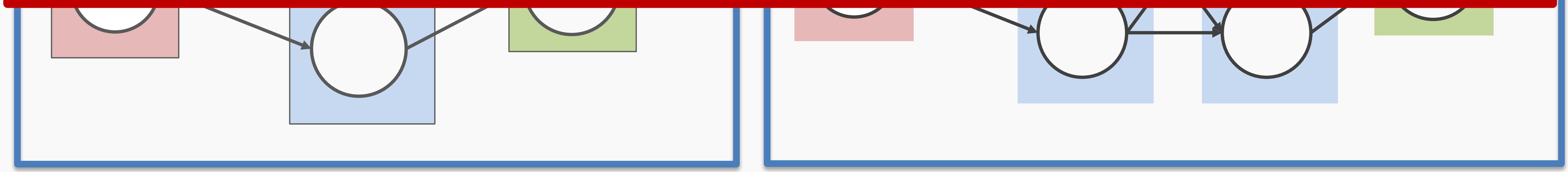




These two networks have different derivatives. We need a mechanism, so we do not need to re-code the derivatives.

1. Derivatives need to be evaluated at some values of X, y, and Ws.
2. But since we have an expression for the derivative, we can build a function that takes as input X, y, W, and returns the derivatives, and then we can use gradient descent to update.
3. This approach works well, but it does not generalize. For example, if the network is

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



Backpropagation (cont.)

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.

Hints:

1. Remember we only need to evaluate the derivatives at X_i, y_i and $W^{(k)}$.
2. We should take advantage of the chain rule we learned before.

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 partial derivative	$\frac{\partial \xi_n}{\partial W}$
$\xi_1 = -W^T X$	$\frac{\partial \xi_1}{\partial W} = -X$	-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$e^{-9}\left(\frac{1}{1+e^{-9}}\right)$
$\frac{\partial \mathcal{L}_i^A}{\partial W} = \frac{\partial \mathcal{L}_i}{\partial \xi_5}\frac{\partial \xi_5}{\partial \xi_4}\frac{\partial \xi_4}{\partial \xi_3}\frac{\partial \xi_3}{\partial \xi_2}\frac{\partial \xi_2}{\partial \xi_1}\frac{\partial \xi_1}{\partial W}$			-3	0.00037018372

😱 **BUT** we still need to specify the **derivatives** 😱

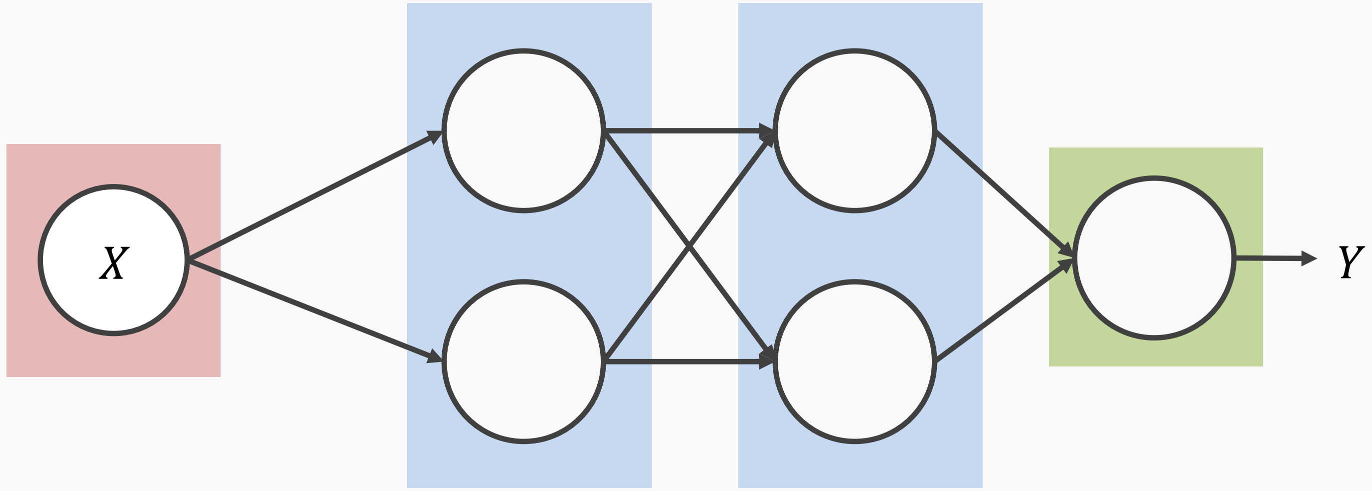
Variables	derivatives	Value of the variable	Value of the partial derivative	$\frac{\partial \xi_n}{\partial W}$
$\xi_1 = -W^T X$	$\frac{\partial \xi_1}{\partial W} = -X$	-9	-3	-3
$\xi_2 = e^{\xi_1} = e^{-W^T X}$	$\frac{\partial \xi_2}{\partial \xi_1} = e^{\xi_1}$	e^{-9}	e^{-9}	-3e^{-9}
$\xi_3 = 1 + \xi_2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

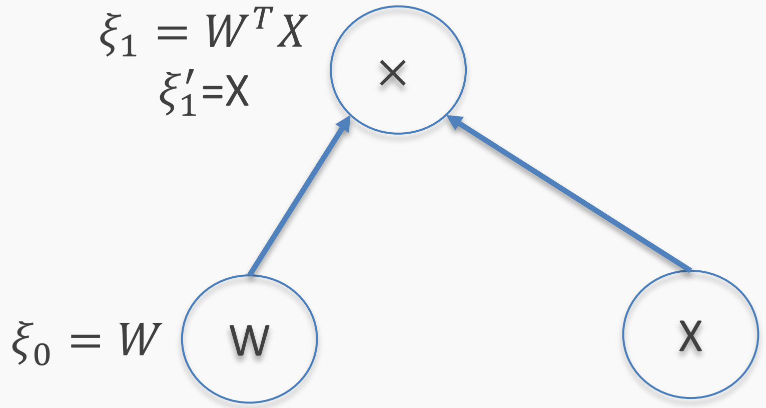
1. We specify the network structure



2. Create a computational graph …

Computational Graph

Computational Graph

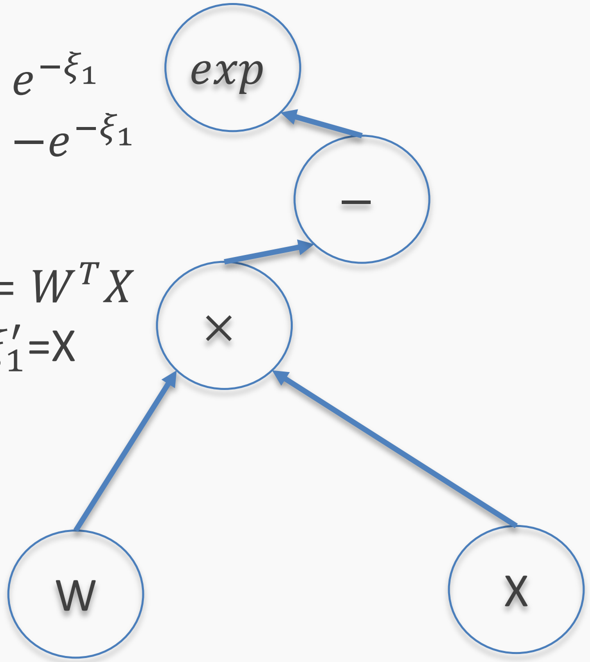


Computational Graph

$$\xi_2 = e^{-\xi_1}$$
$$\xi_2' = -e^{-\xi_1}$$

$$\xi_1 = W^T X$$
$$\xi_1' = X$$

$$\xi_0 = W$$



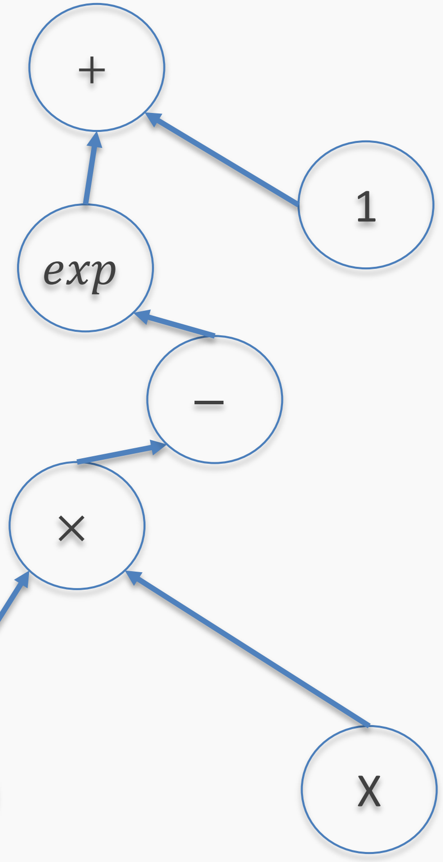
Computational Graph

$$\xi_3 = 1 + e^{-W^T X}$$
$$\xi_3' = 1$$

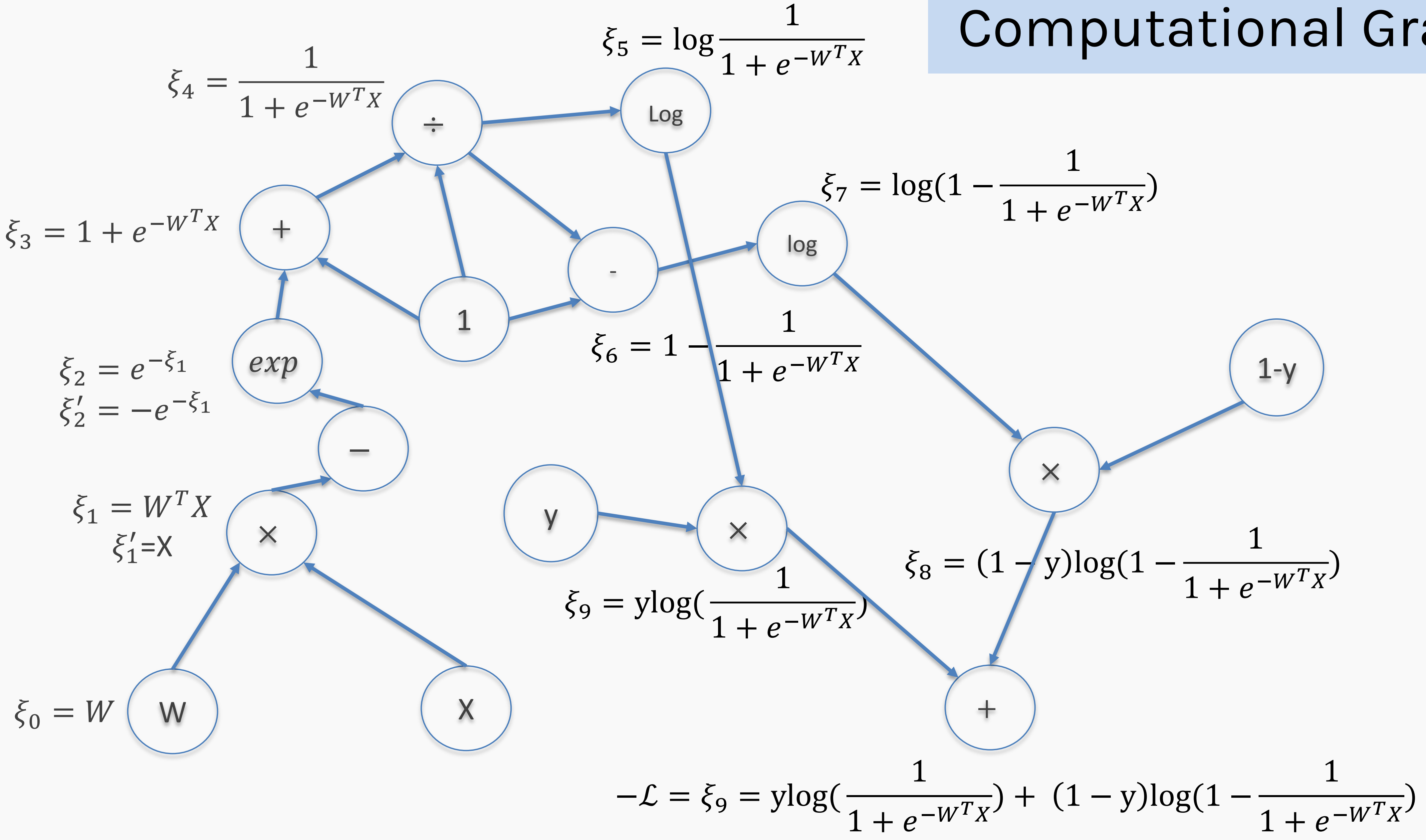
$$\xi_2 = e^{-\xi_1}$$
$$\xi_2' = -e^{-\xi_1}$$

$$\xi_1 = W^T X$$
$$\xi_1' = X$$

$$\xi_0 = W$$

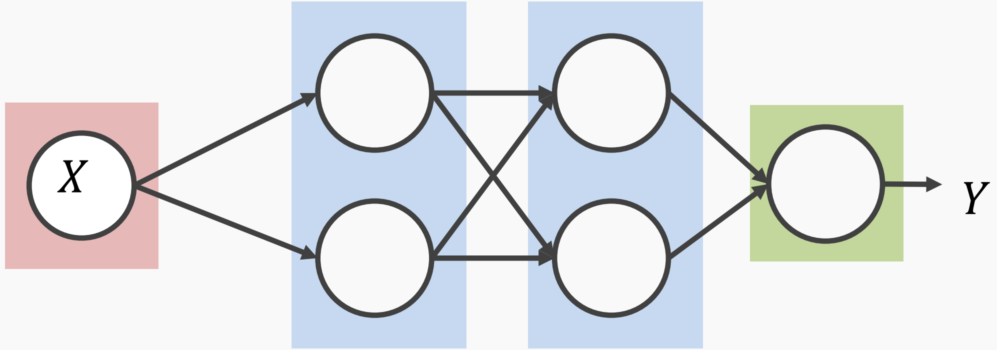


Computational Graph



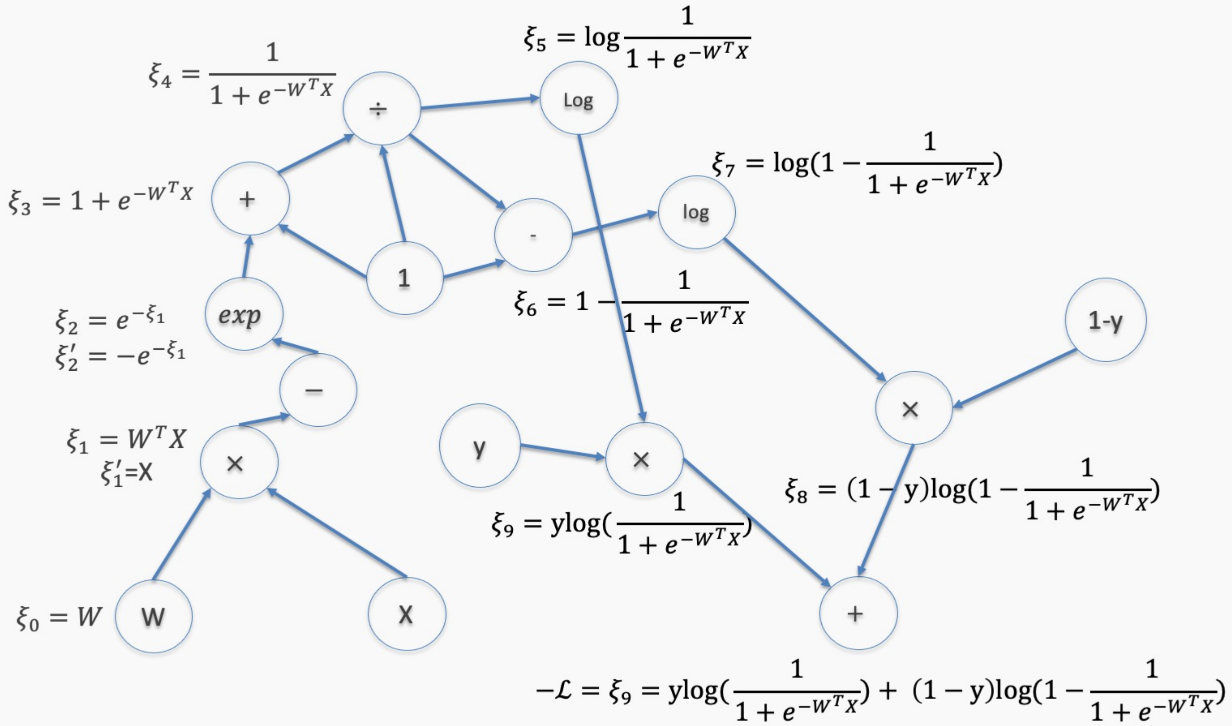
Putting it altogether

1. We specify the network structure



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 variable (as shown in the table a few slides back)



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

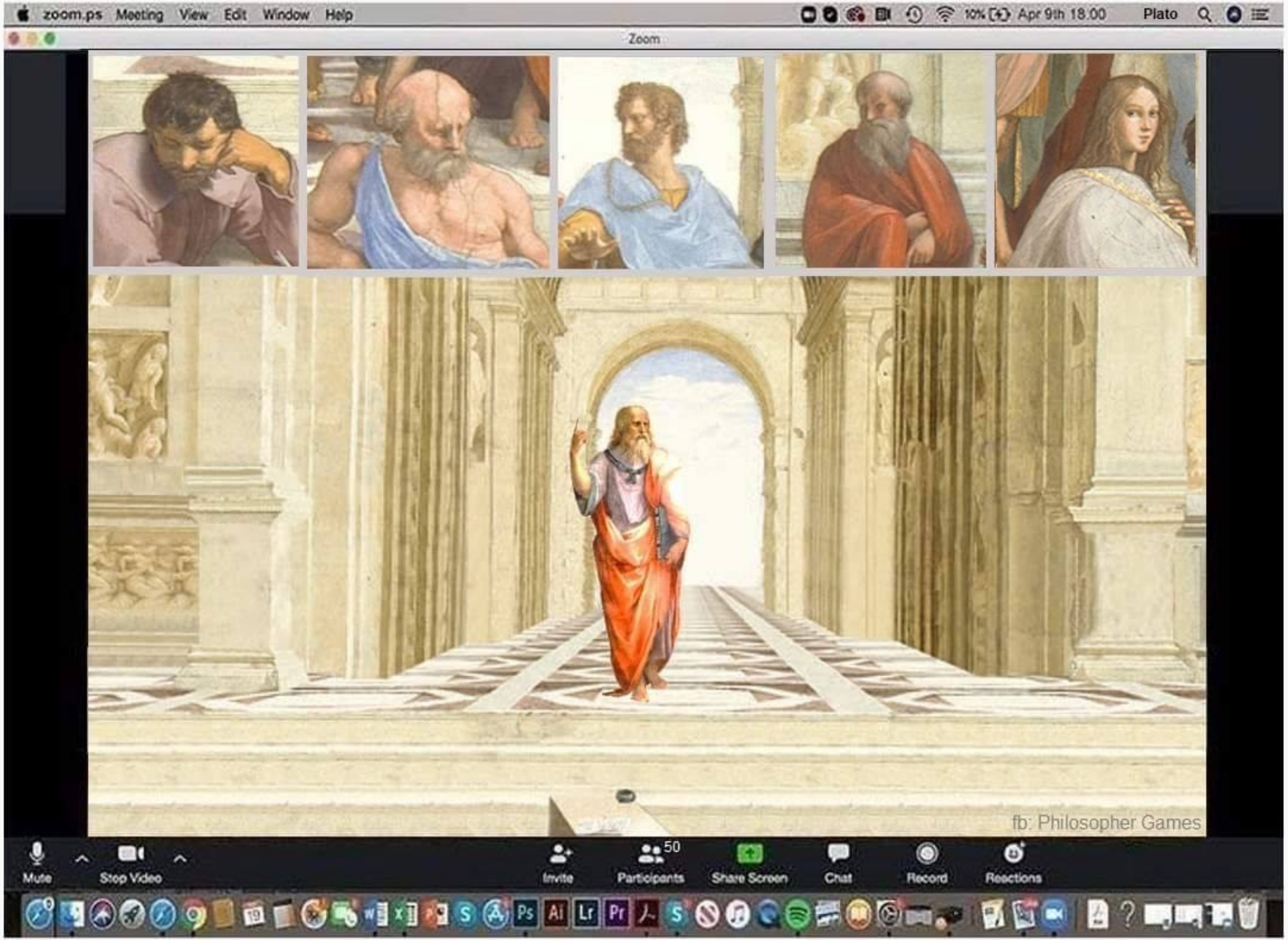
Variables	derivatives	Value of the variable	Value of the partial derivative	$\frac{d\mathcal{L}}{d\xi_n}$
$\xi_1 = -W^T X$	$\frac{\partial \xi_1}{\partial W} = -X$	-9	-3	-3
$\xi_2 = e^{\xi_1} = e^{-W^T X}$	$\frac{\partial \xi_2}{\partial \xi_1} = e^{\xi_1}$	e^{-9}	e^{-9}	-3e^{-9}
$\xi_3 = 1 + \xi_2 =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

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

Variables	derivatives	Value of the variable	Value of the partial 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
$\frac{\partial \mathcal{L}_i^A}{\partial W} = \frac{\partial \mathcal{L}_i}{\partial \xi_5} \frac{\partial \xi_5}{\partial \xi_4} \frac{\partial \xi_4}{\partial \xi_3} \frac{\partial \xi_3}{\partial \xi_2} \frac{\partial \xi_2}{\partial \xi_1} \frac{\partial \xi_1}{\partial W}$			Type equation here.

Store all these values

Multiply as needed


zoom.ps
Meeting
View
Edit
Window
Help
Apr 9th 18:00
Plato
Zoom
fb: Philosopher Games
Mute
Stop Video
Invite
50
Participants
Share Screen
Chat
Record
Reactions


HARVARD
VE RI TAS
VE RI TAS
IACS
VE RI TAS
HARVARD
Thank you