# DD2418 Language Engineering
# 8a: Neural networks

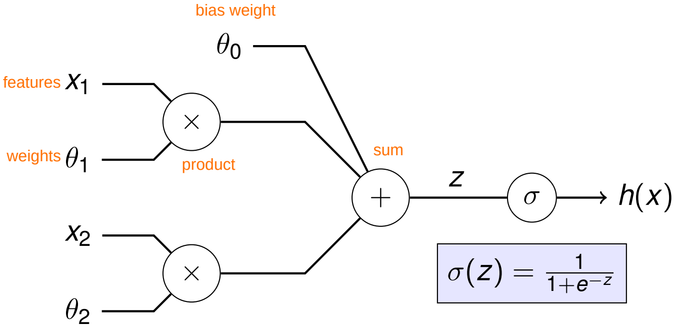
Johan Boye, KTH

April 30, 2020

# Binary logistic regression

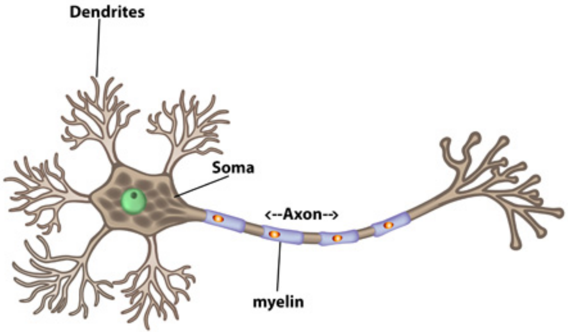
- Represent data as $n$-ary vectors of features $x = (x_1, \ldots, x_n)$.
- The model consists of weights $\theta_0, \theta_1, \ldots, \theta_n$.
- The result $h(x)$ is interpreted as the probability that $x$ belongs to the positive class.

0<= h(x) <=1



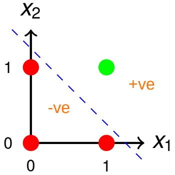

$$\boxed{\sigma(z) = \frac{1}{1+e^{-z}}}$$

# Biological inspiration: The neuron



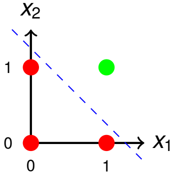
# An artificial neuron cannot compute XOR

AND is linearly separable:

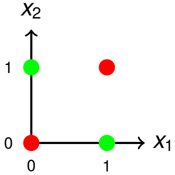


# An artificial neuron cannot compute XOR

AND is linearly separable:

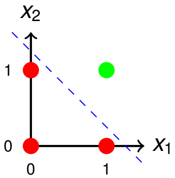



XOR is *not* linearly separable:

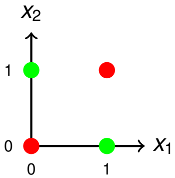


# An artificial neuron cannot compute XOR

AND is linearly separable:

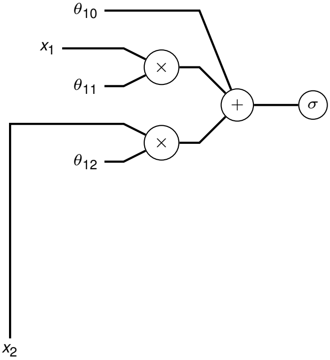



XOR is *not* linearly separable:

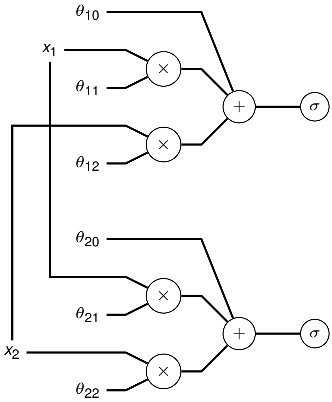



Solution: Use several connected artificial neurons.

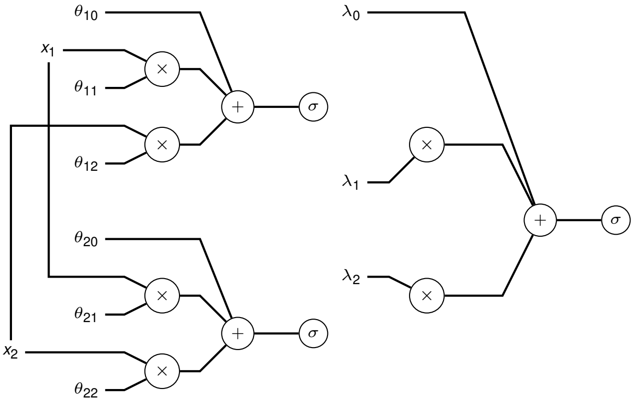
# Using three neurons



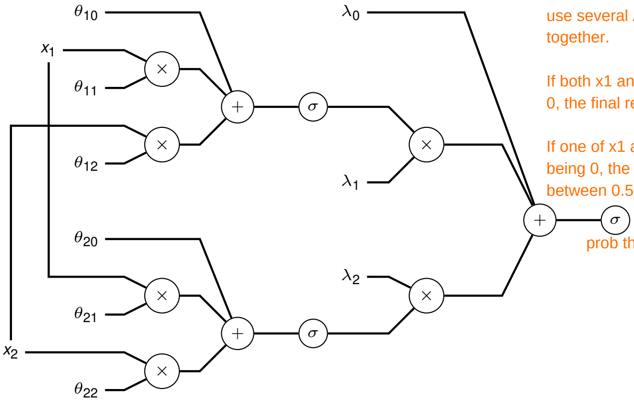
# Using three neurons



# Using three neurons



# Using three neurons




One solution to compute XOR is to use several ANN connected together.

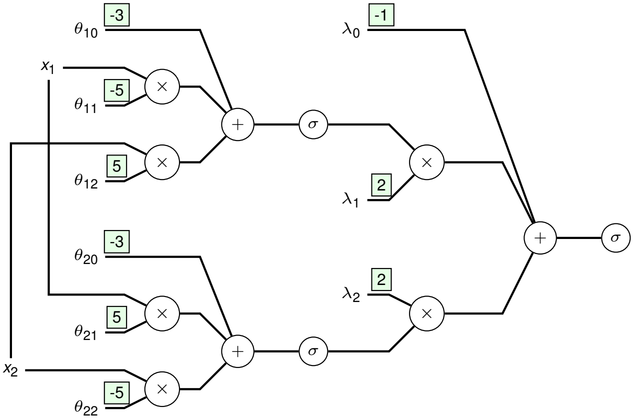
If both x1 and x2 are 1, or both are 0, the final result will be 0 and 0.5

If one of x1 and x2 is 1, the other being 0, the final result will be between 0.5 and 1

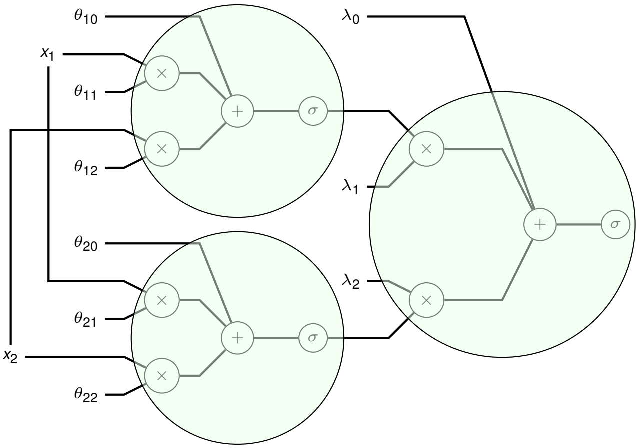
prob that this is a +ve class

# Solving XOR using three neurons

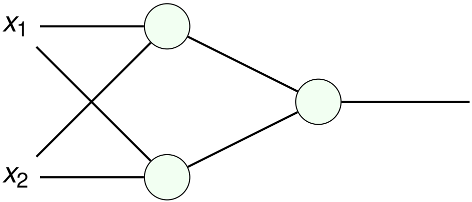
Assume x1 and x2 are both equal to 1 or 0 --> result = 0.288 which is <0.5
Assume x1 =1 and x2 = 0 --> result = 0.682 which is > 0.5
Assume x1 = 0 and x2= 1 --> result = 0. 682 which is > 0.5



# Simplifying the picture

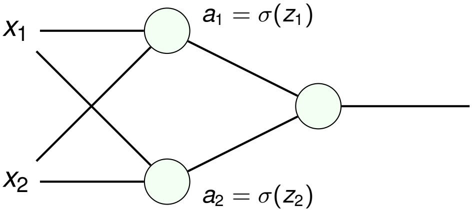


# Simplifying the picture



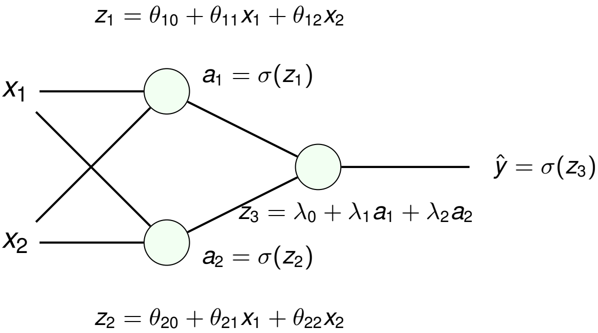
# Simplifying the picture

$$z_1 = \theta_{10} + \theta_{11} x_1 + \theta_{12} x_2$$

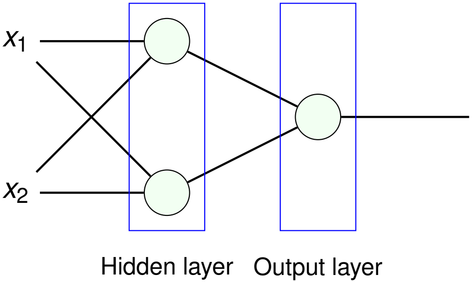



$$z_2 = \theta_{20} + \theta_{21} x_1 + \theta_{22} x_2$$

# Simplifying the picture

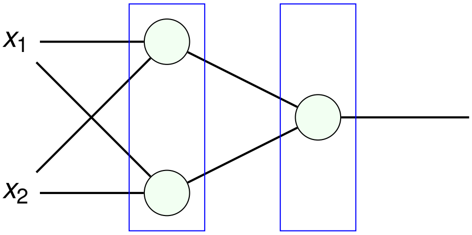


# Layers



# Vector notation

$$a = \sigma(\theta x) \qquad \hat{y} = \sigma(\lambda a)$$

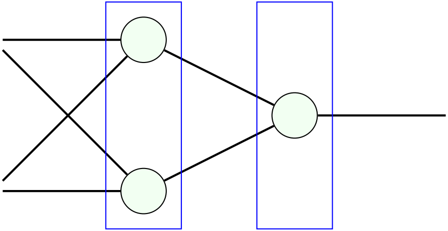



Hidden layer    Output layer

# Example revisited

$$\theta = \begin{array}{c c} \begin{array}{ccc} \text{bias w} & \text{w assoc with x1} & \text{w assoc with x2} \end{array} & \\ \begin{pmatrix} -3 & -5 & 5 \\ -3 & 5 & -5 \end{pmatrix} & \begin{array}{l} \text{w assoc with theta 1} \\ \text{w assoc with theta 2} \end{array} \end{array} \qquad \lambda = \begin{pmatrix} -1 & 2 & 2 \end{pmatrix}$$

$$x = \begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix}$$



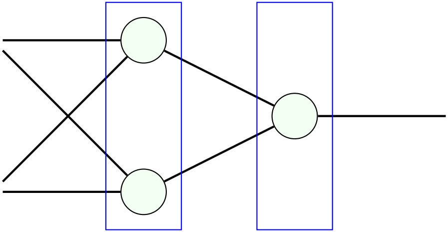
# Alternative notation

The one I learnt in CZ4042

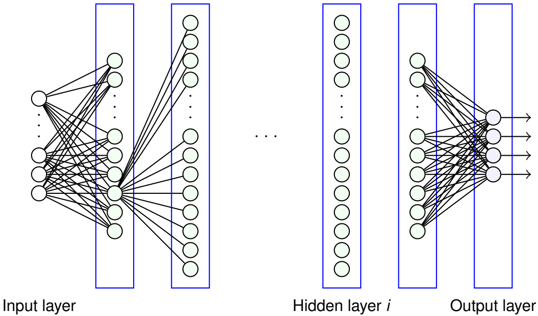
$$\theta = \begin{pmatrix} -5 & 5 \\ 5 & -5 \end{pmatrix} \quad b_\theta = \begin{pmatrix} -3 \\ -3 \end{pmatrix} \quad \lambda = \begin{pmatrix} 2 & 2 \end{pmatrix} \quad b_\lambda = -1$$

$$a = \sigma(\theta x + b_\theta) \qquad \hat{y} = \sigma(\lambda a + b_\lambda)$$

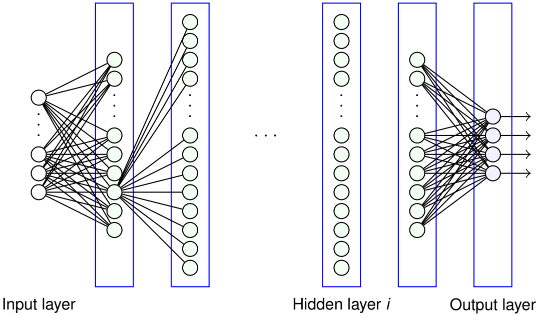
$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$



# Feed-forward networks



# Feed-forward networks




For hidden layer 1:
$z_1 = \theta_1 x$
$a_1 = g_1(z_1)$

For hidden layer $i$:
$z_i = \theta_i a_{i-1}$
$a_i = g_i(z_i)$

For the output layer:
$z_n = \theta_n a_{n-1}$
$\hat{y} = g_n(z_n)$

where each $g_i$ is some non-linear function.

# Feed-forward networks

- The network computes a non-linear function of the input: $\hat{y} = f(x)$
- Each layer computes a linear and a non-linear transformation of the input
- The network thus computes a composition of functions

$$f = f_1 \circ f_2 \circ \ldots \circ f_n$$

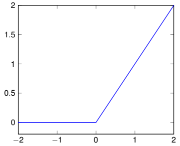
where each function $f_i$ is parametrized by $\theta_i$.

# Activation functions

The activation function is a non-linear function.
Most commonly used are:

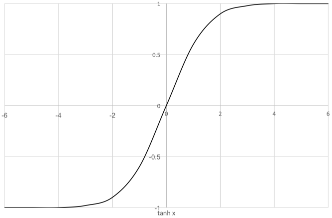
[Rectified Linear Unit (RELU)

$z = \max(0, x)$

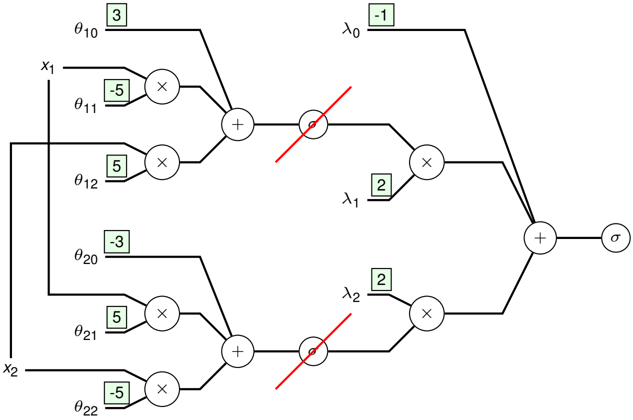



Hyperbolic tangent (tanh)

$z = \frac{e^x - e^{-x}}{e^x + e^{-x}}$



# Are non-linearities essential?



# Are non-linearities essential?

Yes.

Otherwise, we would have in each layer

$$a_i = \theta_i a_{i-1}$$

and thus

$$\hat{y} = \theta_n(\theta_{n-1}(\ldots \theta_1 x \ldots))$$

But then we could simply multiply the matrices:

$$\theta = \theta_n \theta_{n-1} \ldots \theta_1$$

and let $\hat{y} = \theta x$.

That is, a multi-layer network without non-linear transformations is equivalent to a single neuron!

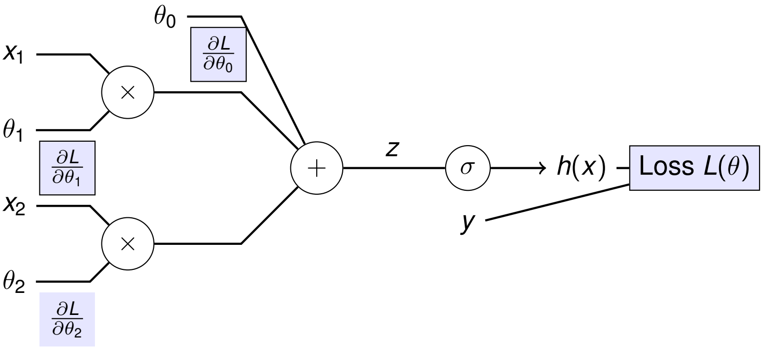
# DD2418 Language Engineering
# 8b: Training neural networks

Johan Boye, KTH

# Learning in logistic regression

To do gradient descent, we need to...

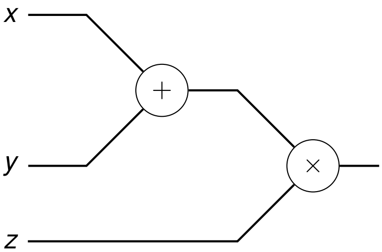
- ... do a *forward pass* to compute the predicted value,
- ... followed by a *backward pass* where we compute the gradient of the loss function



# Backward differentiation (backpropagation)

Consider a simpler example (borrowed from A. Karpathy):

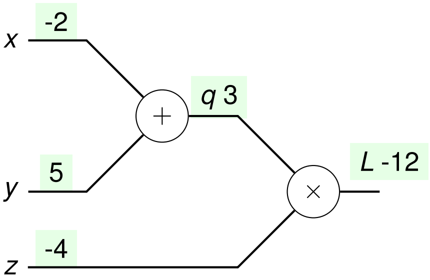
$$L(x, y, z) = (x + y)z$$



# Backward differentiation (backpropagation)

$L(x, y, z) = (x + y)z$

$x = -2, y = 5, z = -4$

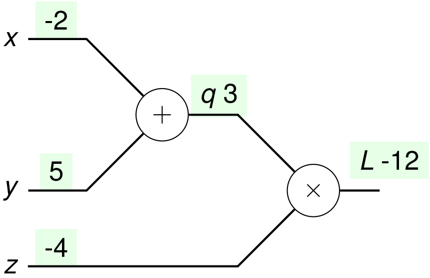


# Backward differentiation (backpropagation)

$L(x, y, z) = (x + y)z$

$x = -2, y = 5, z = -4$

$$\boxed{q = x + y,\ \frac{\partial q}{\partial x} = 1,\ \frac{\partial q}{\partial y} = 1}$$



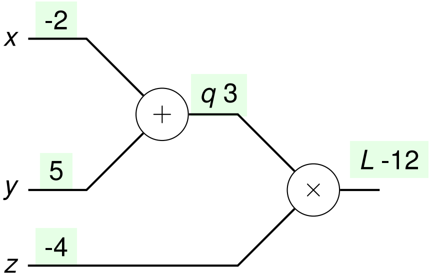
# Backward differentiation (backpropagation)

$$L(x, y, z) = (x + y)z$$

$$x = -2, y = 5, z = -4$$

$$\boxed{q = x + y,\ \frac{\partial q}{\partial x} = 1,\ \frac{\partial q}{\partial y} = 1}$$

$$\boxed{L = qz,\ \frac{\partial L}{\partial q} = z,\ \frac{\partial L}{\partial z} = q}$$



# Backward differentiation (backpropagation)

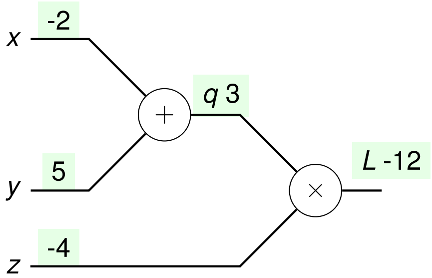
$$L(x, y, z) = (x + y)z$$

$$x = -2, y = 5, z = -4$$

$$\boxed{q = x + y,\ \frac{\partial q}{\partial x} = 1,\ \frac{\partial q}{\partial y} = 1}$$

$$\boxed{L = qz,\ \frac{\partial L}{\partial q} = z,\ \frac{\partial L}{\partial z} = q}$$

We seek $\frac{\partial L}{\partial x}, \frac{\partial L}{\partial y}, \frac{\partial L}{\partial z}$



# Backward differentiation (backpropagation)

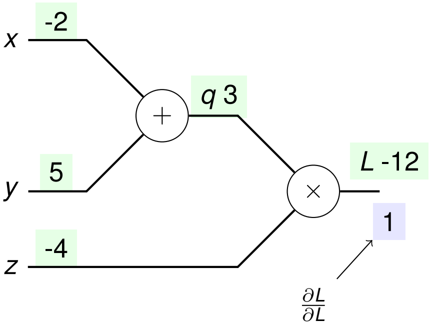
$$L(x, y, z) = (x + y)z$$

$$x = -2, y = 5, z = -4$$

$$\boxed{q = x + y,\ \frac{\partial q}{\partial x} = 1,\ \frac{\partial q}{\partial y} = 1}$$

$$\boxed{L = qz,\ \frac{\partial L}{\partial q} = z,\ \frac{\partial L}{\partial z} = q}$$

We seek $\frac{\partial L}{\partial x}, \frac{\partial L}{\partial y}, \frac{\partial L}{\partial z}$



# Backward differentiation (backpropagation)

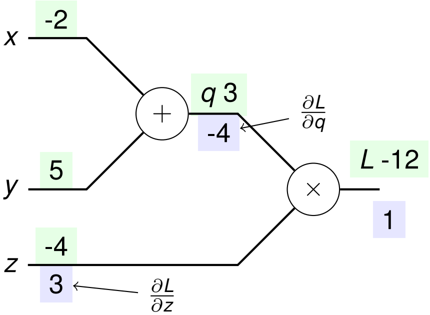
$$L(x, y, z) = (x + y)z$$

$$x = -2, y = 5, z = -4$$

$$\boxed{q = x + y,\ \frac{\partial q}{\partial x} = 1,\ \frac{\partial q}{\partial y} = 1}$$

$$\boxed{L = qz,\ \frac{\partial L}{\partial q} = z,\ \frac{\partial L}{\partial z} = q}$$

We seek $\frac{\partial L}{\partial x}, \frac{\partial L}{\partial y}, \frac{\partial L}{\partial z}$



# Backward differentiation (backpropagation)

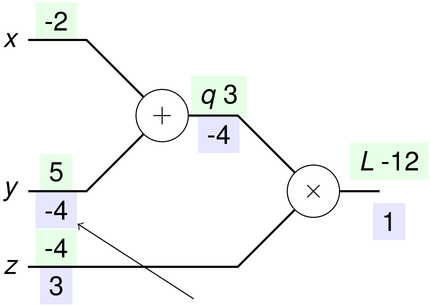
$L(x,y,z) = (x+y)z$

$\boxed{q = x + y,\ \frac{\partial q}{\partial x} = 1,\ \frac{\partial q}{\partial y} = 1}$

$x = -2, y = 5, z = -4$

$\boxed{L = qz,\ \frac{\partial L}{\partial q} = z,\ \frac{\partial L}{\partial z} = q}$

We seek $\frac{\partial L}{\partial x}, \frac{\partial L}{\partial y}, \frac{\partial L}{\partial z}$




$$\frac{\partial L}{\partial y} = \frac{\partial L}{\partial q}\frac{\partial q}{\partial y}$$

# Backward differentiation (backpropagation)

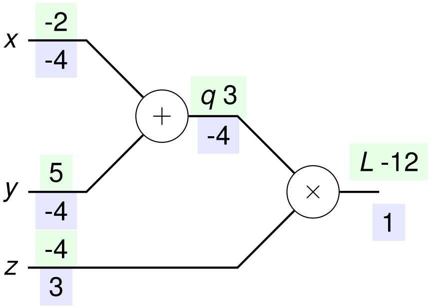
$L(x, y, z) = (x + y)z$

$x = -2, y = 5, z = -4$

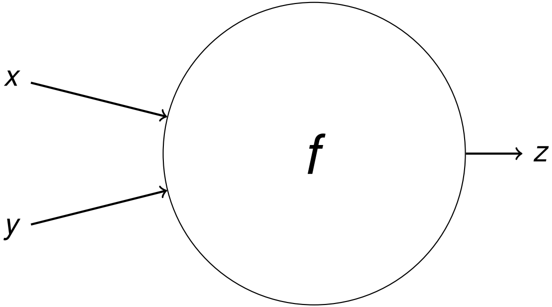
$\boxed{q = x + y,\ \frac{\partial q}{\partial x} = 1,\ \frac{\partial q}{\partial y} = 1}$

$\boxed{L = qz,\ \frac{\partial L}{\partial q} = z,\ \frac{\partial L}{\partial z} = q}$

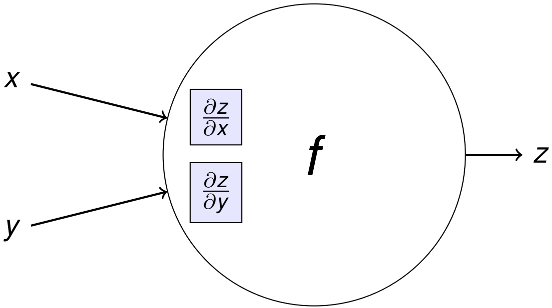
We seek $\frac{\partial L}{\partial x}, \frac{\partial L}{\partial y}, \frac{\partial L}{\partial z}$



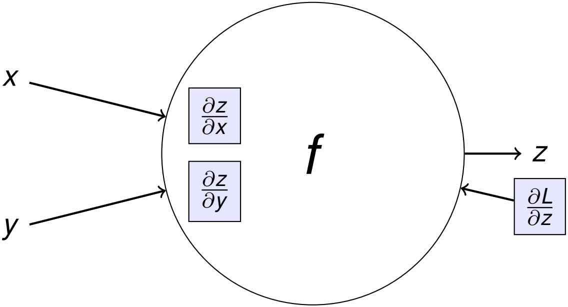
# Backward differentiation (backpropagation)



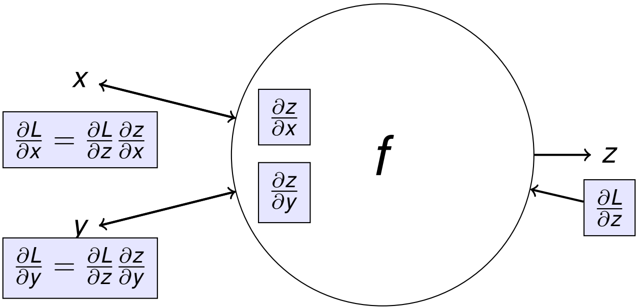
# Backward differentiation (backpropagation)



# Backward differentiation (backpropagation)

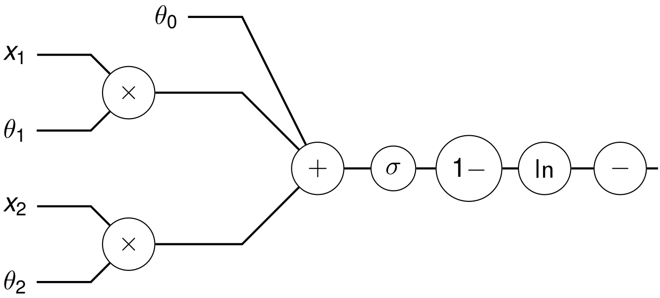


# Backward differentiation (backpropagation)



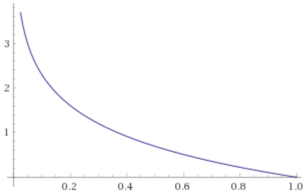
# Backpropagation again

Suppose $x = (1, 1)$ and $y = 0$.
Then the loss is $-\ln(1 - \sigma(\theta^T x))$.

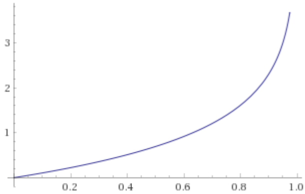


# Cross-entropy loss function

$$\ell(x^{(i)}, y^{(i)}) = \begin{cases} -\log(\sigma(\theta^T x^{(i)})) & \text{if } y^{(i)} = 1 \\ -\log(1 - \sigma(\theta^T x^{(i)})) & \text{if } y^{(i)} = 0 \end{cases}$$




$-\log(\sigma(\theta^T x^{(i)}))$



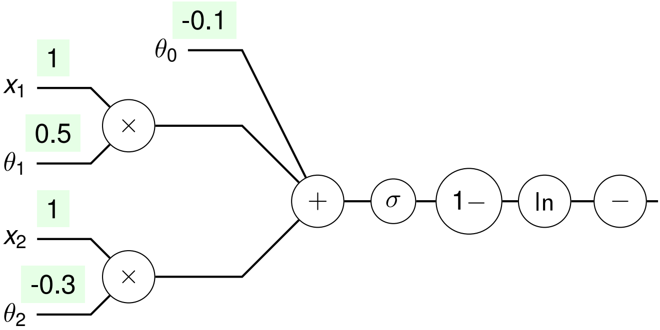

$-\log(1 - \sigma(\theta^T x^{(i)}))$

Since either $y^{(i)} = 1$ or $y^{(i)} = 0$:

$$\ell(\theta) = \frac{1}{m} \sum_{i=0}^{m} [-y^{(i)} \log(\sigma(\theta^T x^{(i)})) - (1 - y^{(i)}) \log(1 - (\sigma(\theta^T x^{(i)})))]$$

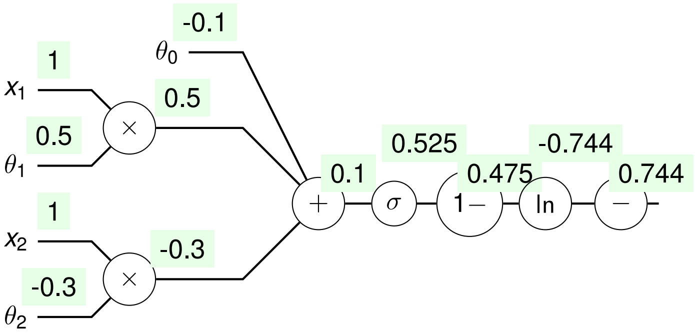
# Backpropagation again

Suppose $\theta = (-0.1, 0.5, -0.3)$. First we do the forward pass.



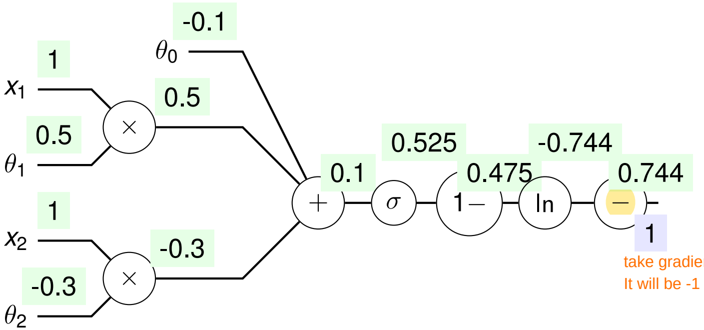
# Backpropagation again

Suppose $\theta = (-0.1, 0.5, -0.3)$. First we do the forward pass.



# Backpropagation again

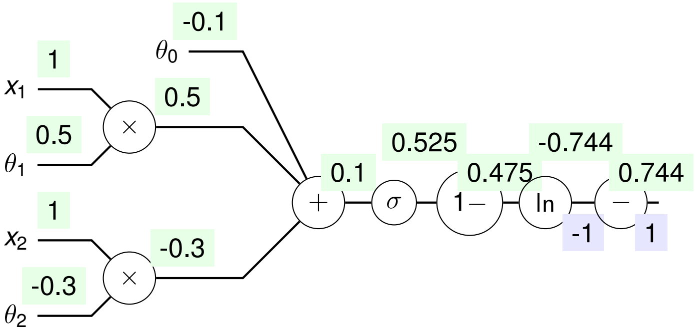
Now do the backward pass.




take gradient 0f -x here.
It will be -1 * 1 = -1

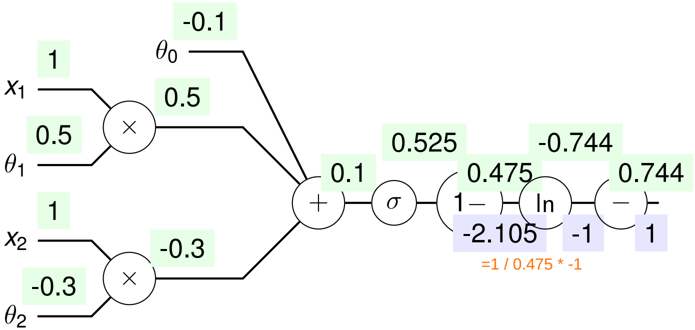
# Backpropagation again

Now do the backward pass.



# Backpropagation again

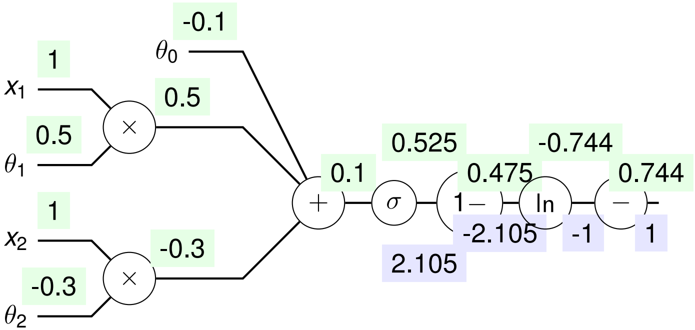
Now do the backward pass.




(

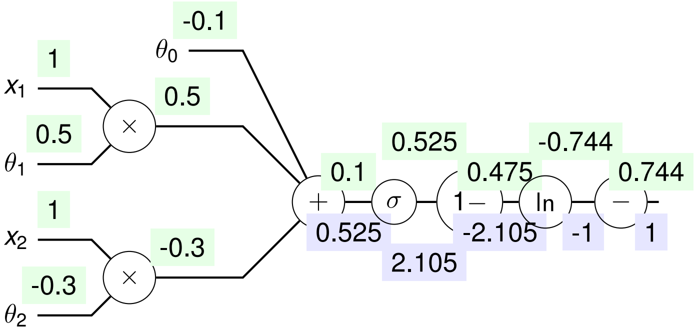
# Backpropagation again

Now do the backward pass.



# Backpropagation again

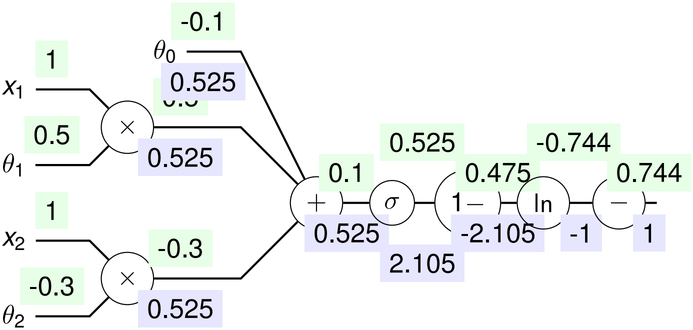
Now do the backward pass.




$$0.525 = \sigma(0.1)(1 - \sigma(0.1))(2.105)$$

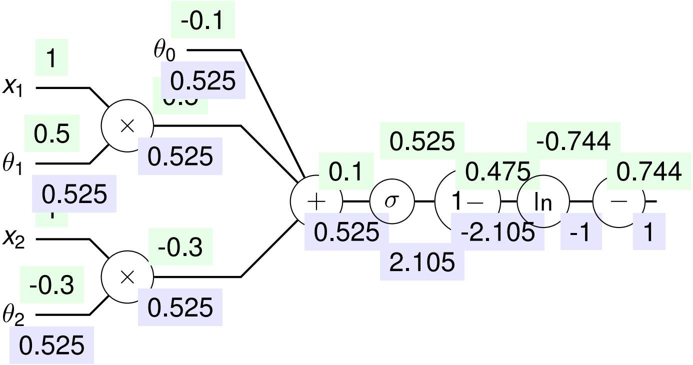
# Backpropagation again

Now do the backward pass.



# Backpropagation again

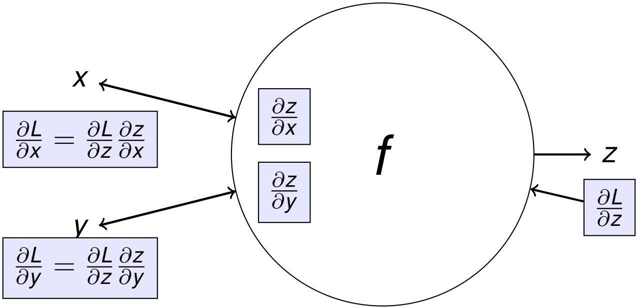
Now do the backward pass.



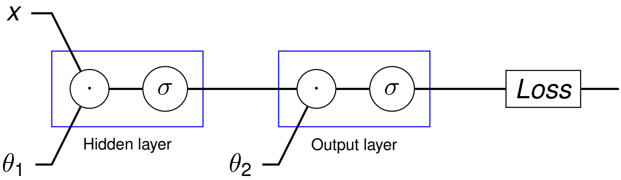
# Backward differentiation (backpropagation)

if

$x, y, z$ are vectors. $\frac{\partial z}{\partial x}$ is now a (Jacobian) matrix: the derivative of every element of $z$ w.r.t. every element of $x$.

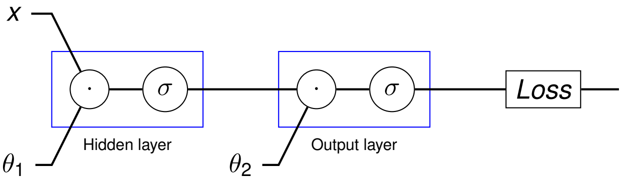


# Backpropagation with a hidden layer



# Backpropagation with a hidden layer

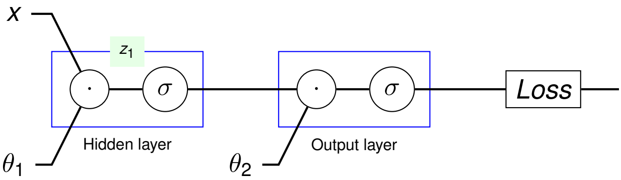
in this example, assume no bias weight




$$x = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad y = 1$$

$$\theta_1 = \begin{pmatrix} 0.45 & 0.05 \\ -0.38 & 0.74 \end{pmatrix} \quad \theta_2 = \begin{pmatrix} -0.13 & 0.95 \end{pmatrix}$$

# Backpropagation with a hidden layer

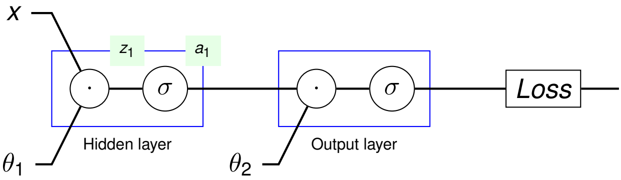



$$x = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad y = 1$$

$$\theta_1 = \begin{pmatrix} 0.45 & 0.05 \\ -0.38 & 0.74 \end{pmatrix} \qquad \theta_2 = \begin{pmatrix} -0.13 & 0.95 \end{pmatrix}$$

$$z_1 = \begin{pmatrix} 0.5 \\ 0.36 \end{pmatrix}$$

# Backpropagation with a hidden layer



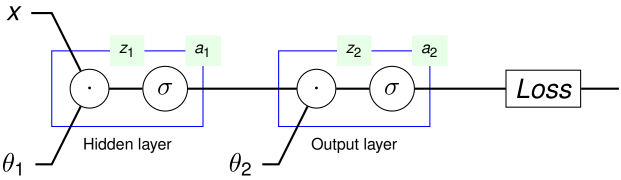

$$x = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad y = 1$$

$$\theta_1 = \begin{pmatrix} 0.45 & 0.05 \\ -0.38 & 0.74 \end{pmatrix} \qquad \theta_2 = \begin{pmatrix} -0.13 & 0.95 \end{pmatrix}$$

$$z_1 = \begin{pmatrix} 0.5 \\ 0.36 \end{pmatrix}$$

$$a_1 = \begin{pmatrix} 0.62 \\ 0.59 \end{pmatrix}$$

# Backpropagation with a hidden layer



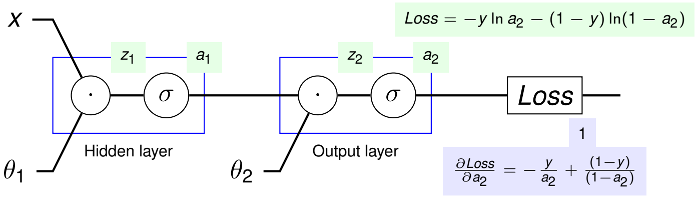

$$x = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad y = 1$$

$$\theta_1 = \begin{pmatrix} 0.45 & 0.05 \\ -0.38 & 0.74 \end{pmatrix} \qquad \theta_2 = \begin{pmatrix} -0.13 & 0.95 \end{pmatrix}$$

$$z_1 = \begin{pmatrix} 0.5 \\ 0.36 \end{pmatrix} \qquad z2 = 0.48$$

$$a_1 = \begin{pmatrix} 0.62 \\ 0.59 \end{pmatrix} \qquad a2 = 0.62$$

# Backpropagation with a hidden layer



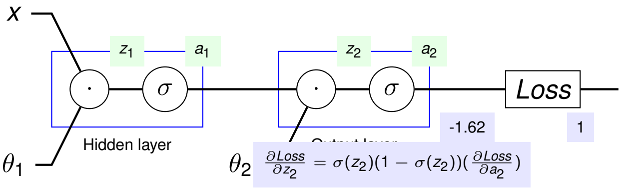

$$x = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad y = 1$$

$$\theta_1 = \begin{pmatrix} 0.45 & 0.05 \\ -0.38 & 0.74 \end{pmatrix} \qquad \theta_2 = \begin{pmatrix} -0.13 & 0.95 \end{pmatrix} \qquad \frac{\partial Loss}{\partial a_2} = -1.62$$

$$z_1 = \begin{pmatrix} 0.5 \\ 0.36 \end{pmatrix} \qquad z2 = 0.48$$

$$a_1 = \begin{pmatrix} 0.62 \\ 0.59 \end{pmatrix} \qquad a2 = 0.62$$

# Backpropagation with a hidden layer




$$x = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad y = 1$$

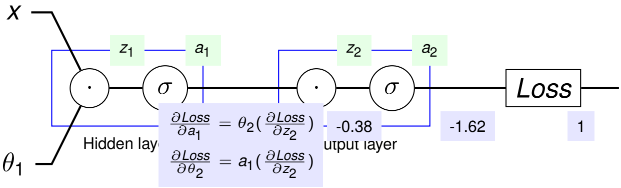
$$\theta_1 = \begin{pmatrix} 0.45 & 0.05 \\ -0.38 & 0.74 \end{pmatrix} \qquad \theta_2 = \begin{pmatrix} -0.13 & 0.95 \end{pmatrix} \qquad \frac{\partial Loss}{\partial a_2} = -1.62$$

$$z_1 = \begin{pmatrix} 0.5 \\ 0.36 \end{pmatrix} \qquad z2 = 0.48$$

$$a_1 = \begin{pmatrix} 0.62 \\ 0.59 \end{pmatrix} \qquad a2 = 0.62$$

$$\frac{\partial Loss}{\partial z_2} = -0.38$$

# Backpropagation with a hidden layer




$$x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$y = 1$$

$$\theta_1 = \begin{pmatrix} 0.45 & 0.05 \\ -0.38 & 0.74 \end{pmatrix}$$

$$z_1 = \begin{pmatrix} 0.5 \\ 0.36 \end{pmatrix}$$

$$a_1 = \begin{pmatrix} 0.62 \\ 0.59 \end{pmatrix}$$

$$\theta_2 = \begin{pmatrix} -0.13 & 0.95 \end{pmatrix}$$

$$z2 = 0.48$$

$$a2 = 0.62$$

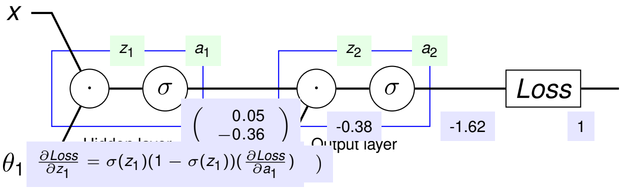
$$\frac{\partial Loss}{\partial z_2} = -0.38$$

$$\frac{\partial Loss}{\partial \theta_2} = \begin{pmatrix} -0.24 & -0.22 \end{pmatrix}$$

$$\frac{\partial Loss}{\partial a_1} = \begin{pmatrix} 0.05 \\ -0.36 \end{pmatrix}$$

$$\frac{\partial Loss}{\partial a_2} = -1.62$$

# Backpropagation with a hidden layer




$$x = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad y = 1$$

$$\theta_1 = \begin{pmatrix} 0.45 & 0.05 \\ -0.38 & 0.74 \end{pmatrix} \qquad \theta_2 = \begin{pmatrix} -0.13 & 0.95 \end{pmatrix} \qquad \frac{\partial Loss}{\partial a_2} = -1.62$$

$$z_1 = \begin{pmatrix} 0.5 \\ 0.36 \end{pmatrix} \qquad z2 = 0.48$$

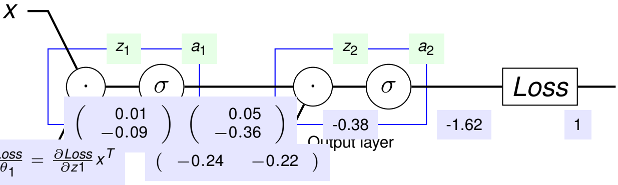
$$a_1 = \begin{pmatrix} 0.62 \\ 0.59 \end{pmatrix} \qquad a2 = 0.62$$

$$\frac{\partial Loss}{\partial z_1} = \begin{pmatrix} 0.01 \\ -0.09 \end{pmatrix} \qquad \frac{\partial Loss}{\partial z_2} = -0.38$$

$$\frac{\partial Loss}{\partial \theta_2} = \begin{pmatrix} -0.24 & -0.22 \end{pmatrix}$$

$$\frac{\partial Loss}{\partial a_1} = \begin{pmatrix} 0.05 \\ -0.36 \end{pmatrix}$$

# Backpropagation with a hidden layer




$$x = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad y = 1$$

$$\theta_1 = \begin{pmatrix} 0.45 & 0.05 \\ -0.38 & 0.74 \end{pmatrix} \qquad \theta_2 = \begin{pmatrix} -0.13 & 0.95 \end{pmatrix} \qquad \frac{\partial Loss}{\partial a_2} = -1.62$$

$$z_1 = \begin{pmatrix} 0.5 \\ 0.36 \end{pmatrix} \qquad z2 = 0.48$$

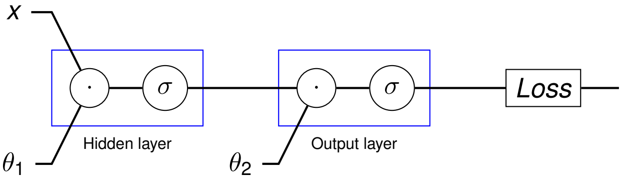
$$a_1 = \begin{pmatrix} 0.62 \\ 0.59 \end{pmatrix} \qquad a2 = 0.62$$

$$\frac{\partial Loss}{\partial z_1} = \begin{pmatrix} 0.01 \\ -0.09 \end{pmatrix} \qquad \frac{\partial Loss}{\partial z_2} = -0.38$$

$$\frac{\partial Loss}{\partial \theta_1} = \begin{pmatrix} 0.01 & 0.01 \\ -0.09 & -0.09 \end{pmatrix} \qquad \frac{\partial Loss}{\partial \theta_2} = \begin{pmatrix} -0.24 & -0.22 \end{pmatrix}$$

$$\frac{\partial Loss}{\partial a_1} = \begin{pmatrix} 0.05 \\ -0.36 \end{pmatrix}$$

# Backpropagation with a hidden layer




$$x = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad y = 1$$

$$\theta_1 = \begin{pmatrix} 0.45 & 0.05 \\ -0.38 & 0.74 \end{pmatrix} \qquad \theta_2 = \begin{pmatrix} -0.13 & 0.95 \end{pmatrix} \qquad \frac{\partial Loss}{\partial a_2} = -1.62$$

$$z_1 = \begin{pmatrix} 0.5 \\ 0.36 \end{pmatrix} \qquad z2 = 0.48$$

$$a_1 = \begin{pmatrix} 0.62 \\ 0.59 \end{pmatrix} \qquad a2 = 0.62$$

$$\frac{\partial Loss}{\partial z_1} = \begin{pmatrix} 0.01 \\ -0.09 \end{pmatrix} \qquad \frac{\partial Loss}{\partial z_2} = -0.38$$

$$\frac{\partial Loss}{\partial \theta_1} = \begin{pmatrix} 0.01 & 0.01 \\ -0.09 & -0.09 \end{pmatrix} \qquad \frac{\partial Loss}{\partial \theta_2} = \begin{pmatrix} -0.24 & -0.22 \end{pmatrix}$$

$$\frac{\partial Loss}{\partial a_1} = \begin{pmatrix} 0.05 \\ -0.36 \end{pmatrix}$$