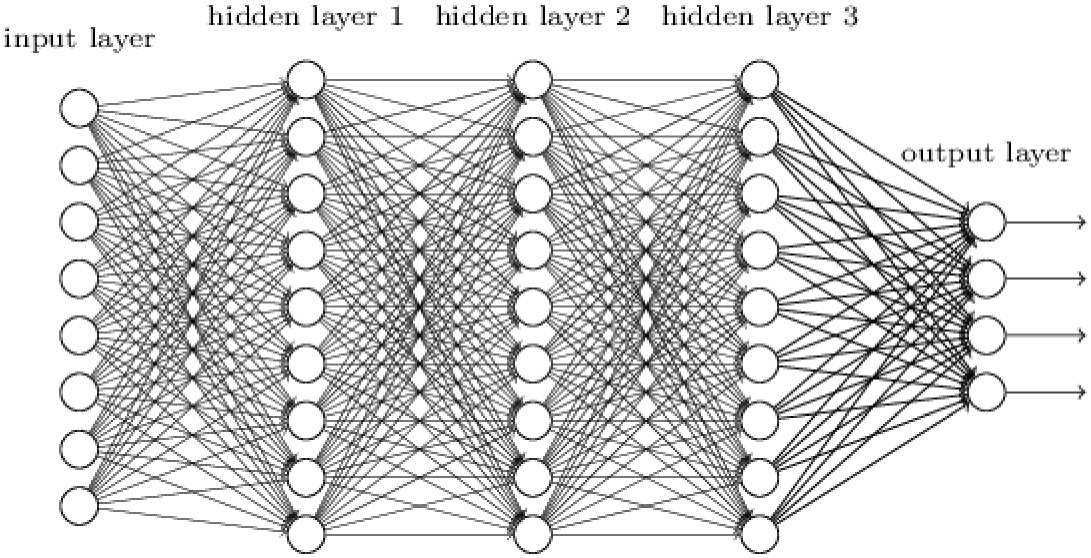
DEEP LEARNING INTRODUCTION



Reda Igbaria & Itay Evron

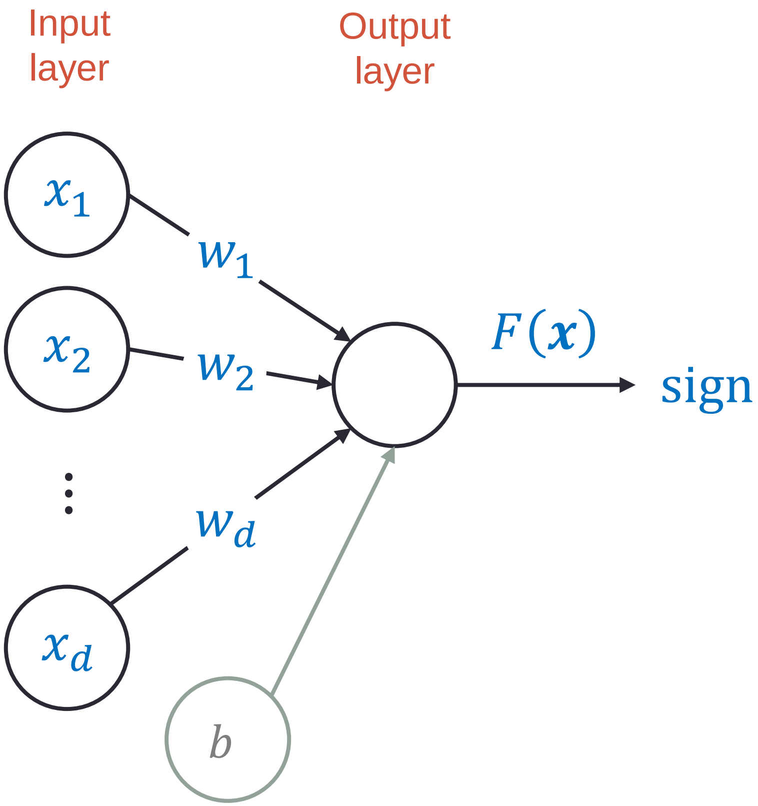
Outline

- From perceptrons to neural networks
- Training neural networks
- Backpropagation
- Playground

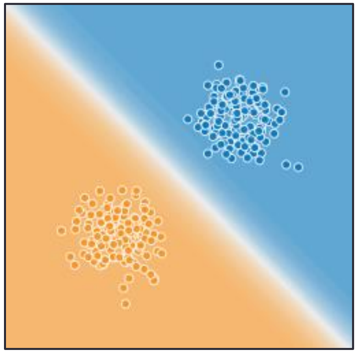
Perceptron

- Recall the perceptron linear model

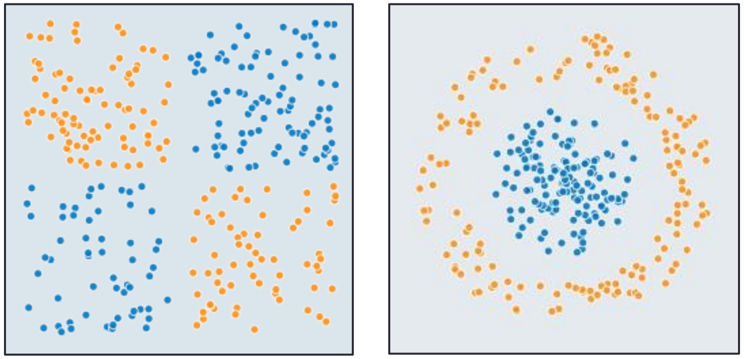
$$F(\boldsymbol{x}) = \boldsymbol{w}^{\top}\boldsymbol{x} + b = \sum_{i=1}^{d} x_i w_i + b$$



- Creates a linear decision boundary.

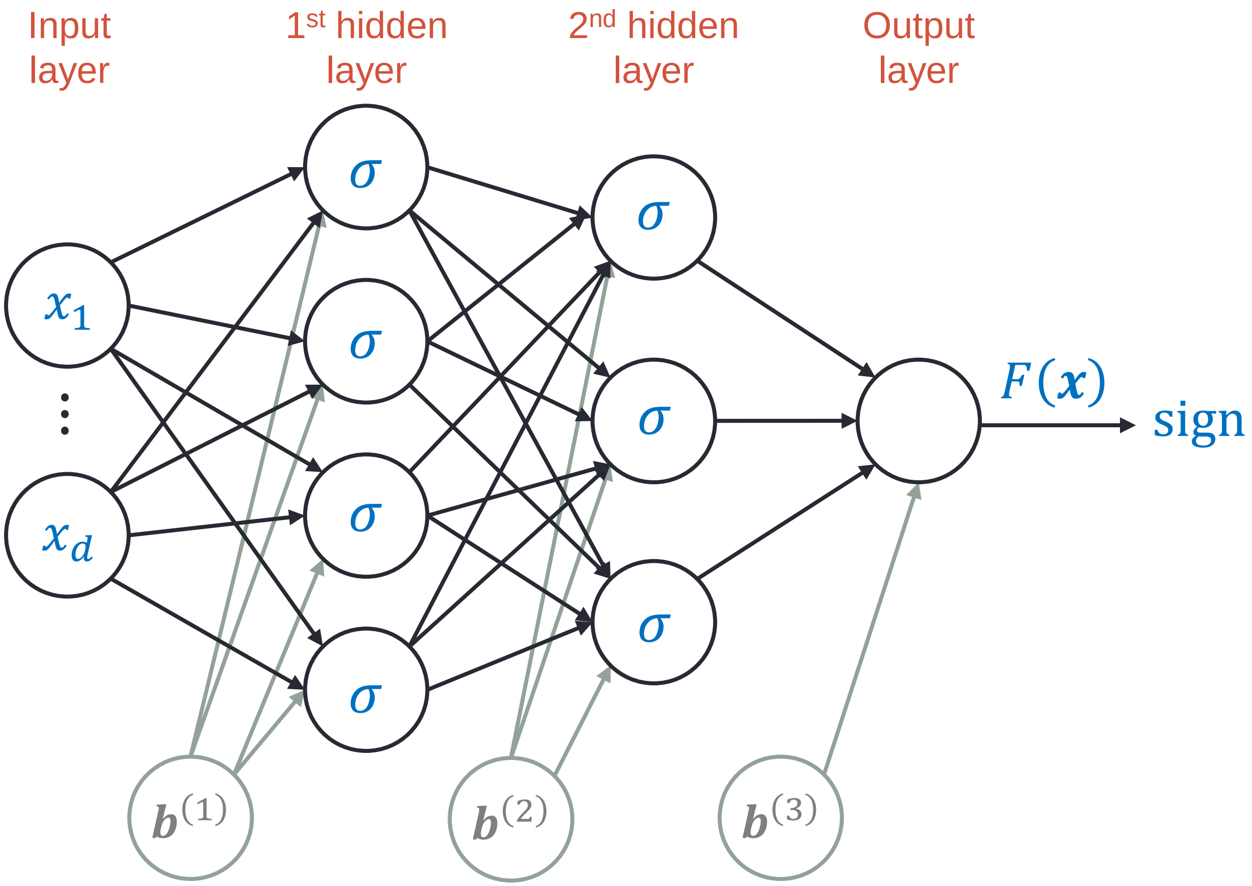


- How can we create more complicated decision boundaries?

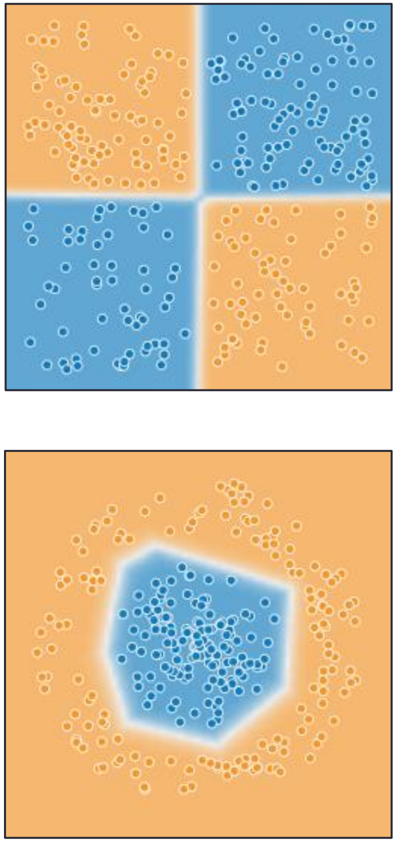


Neural networks (Multilayer Perceptron)

- We can add more layers!
- For instance, $F(\boldsymbol{x}) = \mathbf{w}^{(3)^{\top}}\sigma\left(\mathbf{W}^{(2)^{\top}}\sigma\left(\mathbf{W}^{(1)^{\top}}\boldsymbol{x}+\boldsymbol{b}^{(1)}\right)+\boldsymbol{b}^{(2)}\right)+\boldsymbol{b}^{(3)}$



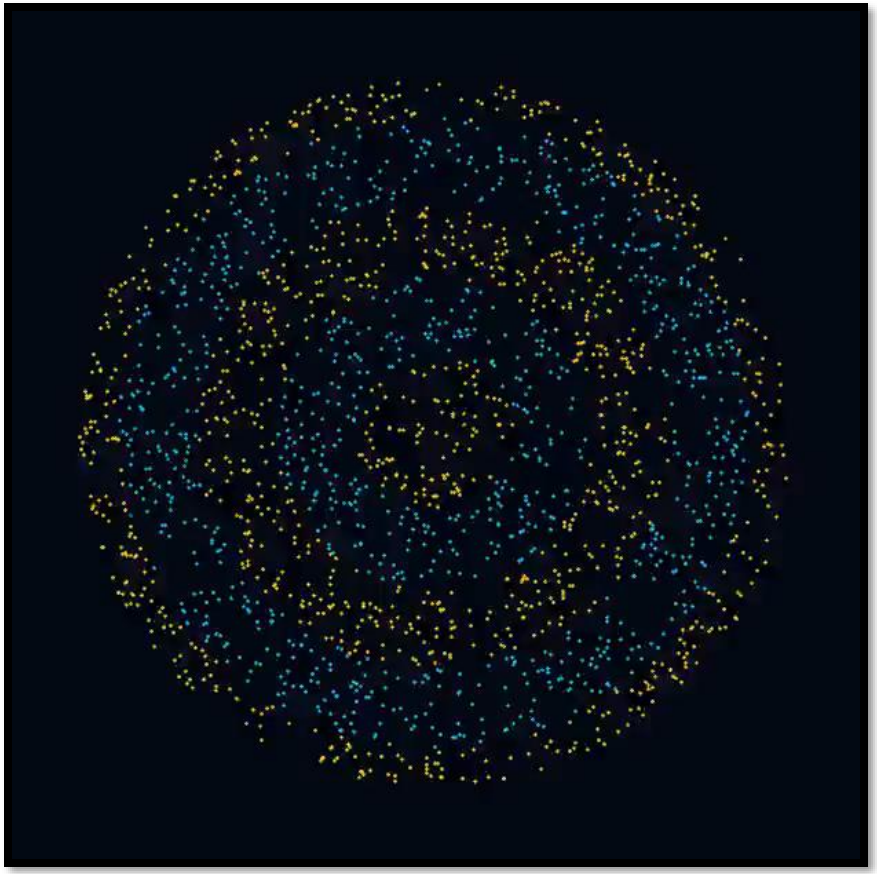
- Can now fit richer functions



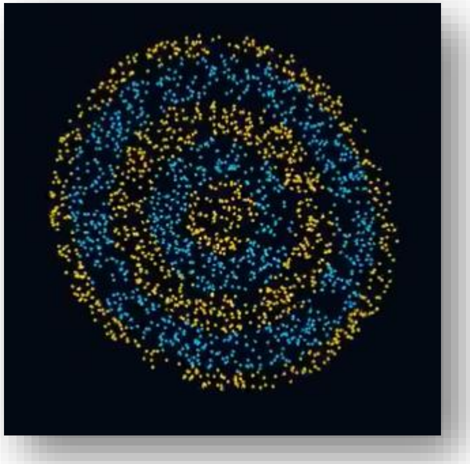
Created on TF playground

Demo: “Feature” learning

- Two-layer neural network tries to separate “circles” by learning a single new feature.



Source: Matt Henderson on Twitter



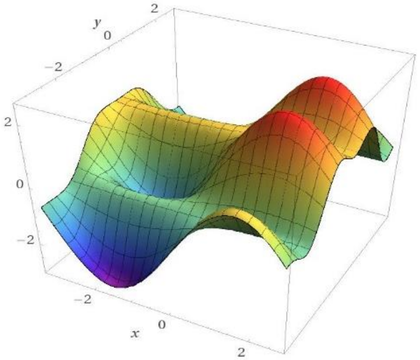
Training neural networks

- Define a network F parameterized by $\Theta = \left(\mathbf{W}^{(1)}, \boldsymbol{b}^{(1)}, \mathbf{W}^{(2)}, \boldsymbol{b}^{(2)}, \ldots\right)$. Which?
- Define a loss over network outputs, e.g., MSE: $\mathcal{L} = \frac{1}{m}\sum_{i=1}^{m}(F(\boldsymbol{x}_i) - y_i)^2$ Which?
- Gradient descent general scheme:
  1. Initialize parameters randomly How?
  2. While model has not converged: When?
     1. Compute gradient $\frac{\partial}{\partial\Theta}\mathcal{L}(\Theta)$ How?
     2. Update weights $\Theta \leftarrow \Theta - \eta\frac{\partial}{\partial\Theta}\mathcal{L}(\Theta)$ How?

Today's focus

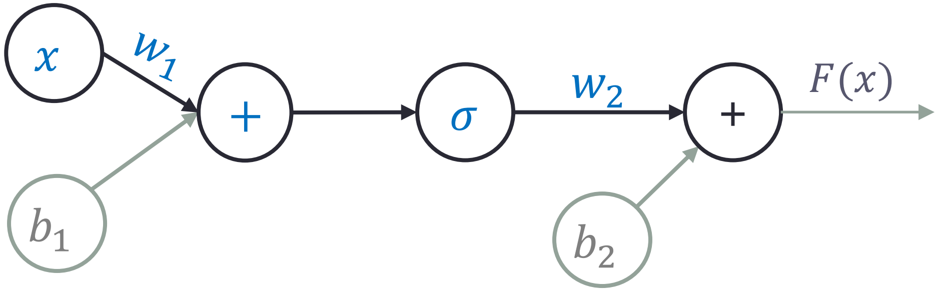
- The loss is usually not convex w.r.t Θ,

  but practically we often converge to a good minimum!



Simple regression example

- Consider a simple network $F(x) = w_2 \cdot \sigma(w_1 x + b_1) + b_2$.



- The architecture is defined only after choosing the activation function.

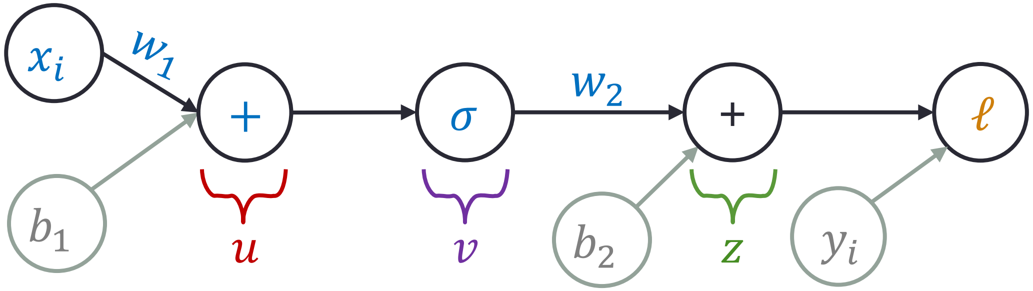
  We will use the sigmoid function $\sigma(a) = \frac{1}{1+e^{-a}}$.

- Trying to solve a regression problem, we minimize the squared loss

$$\mathcal{L} = \sum_{i=1}^{m} (F(x_i) - y_i)^2 = \sum_{i=1}^{m} \ell(x_i, y_i)$$

Simple regression example

- Consider a simple network $F(x) = w_2 \cdot \sigma(w_1 x + b_1) + b_2$.

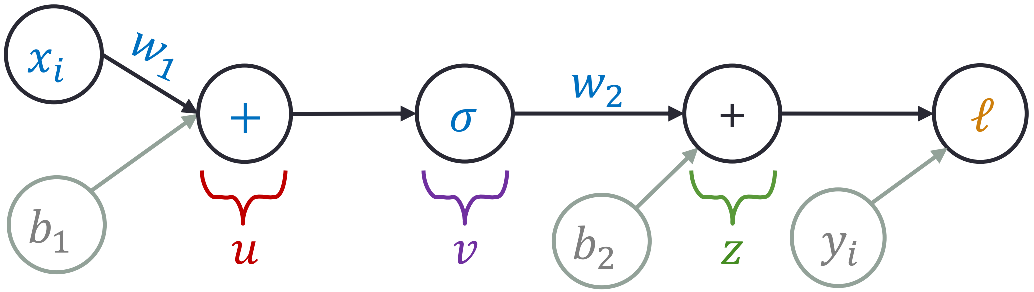


where $\sigma(a) = \frac{1}{1+e^{-a}}$ and the loss is $\mathcal{L} = \sum_{i=1}^{m} (F(x_i) - y_i)^2 = \sum_{i=1}^{m} \ell(x_i, y_i)$.

- Denote: $u = w_1 x + b_1$, $v = \sigma(u)$, $z = w_2 v + b_2 \big(= F(x)\big)$, $\ell = (z - y_i)^2$
- Compute all partial derivatives using the chain rule:
  - $\frac{\partial}{\partial w_2} \ell = \frac{\partial \ell}{\partial z} \frac{\partial z}{\partial w_2} =$ ______________________
  - $\frac{\partial}{\partial b_2} \ell = \frac{\partial \ell}{\partial z} \frac{\partial z}{\partial b_2} =$ ______________________

Simple regression example

- Consider a simple network $F(x) = w_2 \cdot \sigma(w_1 x + b_1) + b_2$.

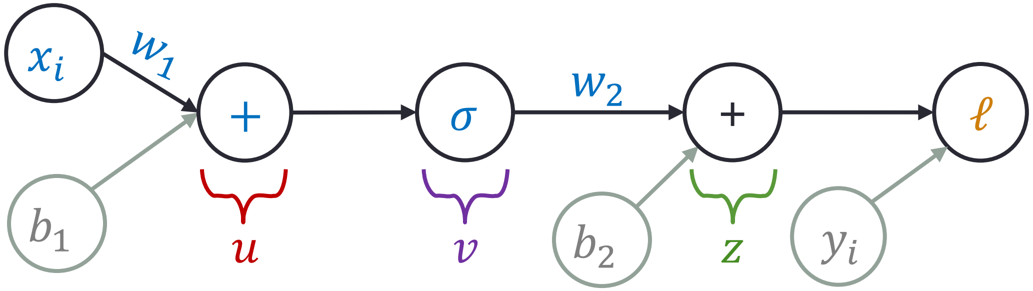


where $\sigma(a) = \frac{1}{1+e^{-a}}$ and the loss is $\mathcal{L} = \sum_{i=1}^{m}(F(x_i) - y_i)^2 = \sum_{i=1}^{m} \ell(x_i, y_i)$.

- Denote: $u = w_1 x + b_1$, $v = \sigma(u)$, $z = w_2 v + b_2 \big(= F(x)\big)$, $\ell = (z - y_i)^2$
- Compute all partial derivatives using the chain rule:
  - $\frac{\partial}{\partial w_2}\ell = \frac{\partial \ell}{\partial z}\frac{\partial z}{\partial w_2} = 2(z - y_i) \cdot v$
  - $\frac{\partial}{\partial b_2}\ell = \frac{\partial \ell}{\partial z}\frac{\partial z}{\partial b_2} = 2(z - y_i) \cdot 1$

Simple regression example

- Consider a simple network $F(x) = w_2 \cdot \sigma(w_1 x + b_1) + b_2$.

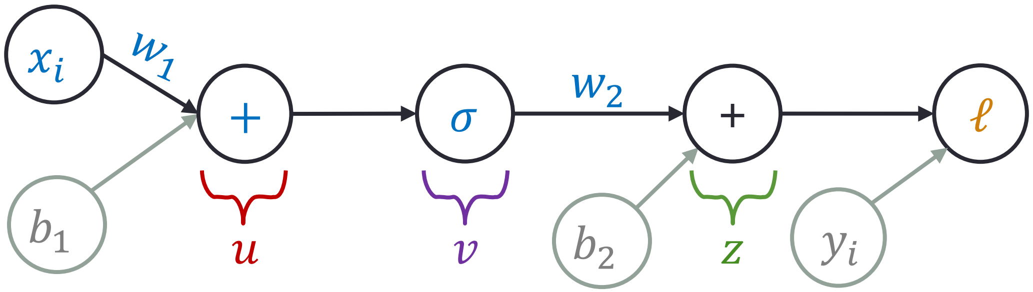


where $\sigma(a) = \frac{1}{1+e^{-a}}$ and the loss is $\mathcal{L} = \sum_{i=1}^{m}(F(x_i) - y_i)^2 = \sum_{i=1}^{m} \ell(x_i, y_i)$.

- Denote: $u = w_1 x + b_1$, $v = \sigma(u)$, $z = w_2 v + b_2 \big(= F(x)\big)$, $\ell = (z - y_i)^2$
- Compute all partial derivatives using the chain rule:
  - $\frac{\partial}{\partial w_1}\ell = \frac{\partial \ell}{\partial z}\frac{\partial z}{\partial v}\frac{\partial v}{\partial u}\frac{\partial u}{\partial w_1} =$ ____________________
  - $\frac{\partial}{\partial b_1}\ell = \frac{\partial \ell}{\partial z}\frac{\partial z}{\partial v}\frac{\partial v}{\partial u}\frac{\partial u}{\partial b_1} =$ ____________________

Simple regression example

- Consider a simple network $F(x) = w_2 \cdot \sigma(w_1 x + b_1) + b_2$.



where $\sigma(a) = \frac{1}{1+e^{-a}}$ and the loss is $\mathcal{L} = \sum_{i=1}^{m}(F(x_i) - y_i)^2 = \sum_{i=1}^{m} \ell(x_i, y_i)$.

- Denote: $u = w_1 x + b_1$, $v = \sigma(u)$, $z = w_2 v + b_2 (= F(x))$, $\ell = (z - y_i)^2$
- Compute all partial derivatives using the chain rule:

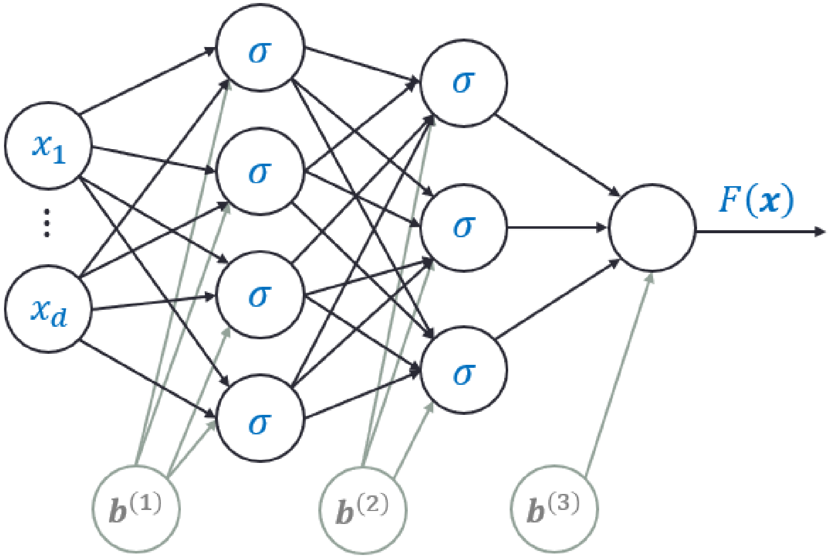
$$\frac{f(x)}{g(x)} = \frac{f'(x) \cdot g(x) - g'(x) \cdot f(x)}{g^2(x)}$$

$$\left(\frac{1}{1+e^{-x}}\right)' = \frac{0 + e^{-x}}{(1+e^{-x})^2} = \frac{1}{(1+e^{-x})} \cdot \frac{e^{x}}{1+e^{-x}}$$

  - $\frac{\partial}{\partial w_1}\ell = \frac{\partial \ell}{\partial z}\frac{\partial z}{\partial v}\frac{\partial v}{\partial u}\frac{\partial u}{\partial w_1} = 2(z - y_i) \cdot w_2 \cdot \sigma(u)(1 - \sigma(u)) \cdot x$
  - $\frac{\partial}{\partial b_1}\ell = \frac{\partial \ell}{\partial z}\frac{\partial z}{\partial v}\frac{\partial v}{\partial u}\frac{\partial u}{\partial b_1} = 2(z - y_i) \cdot w_2 \cdot \sigma(u)(1 - \sigma(u)) \cdot 1$

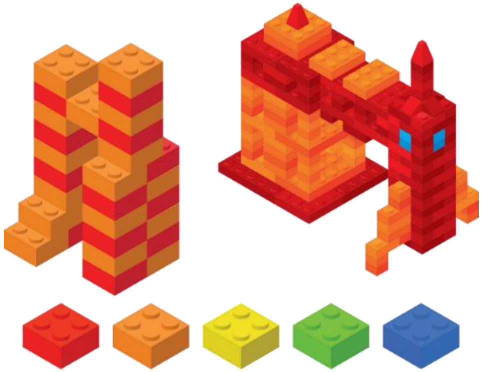
Larger models

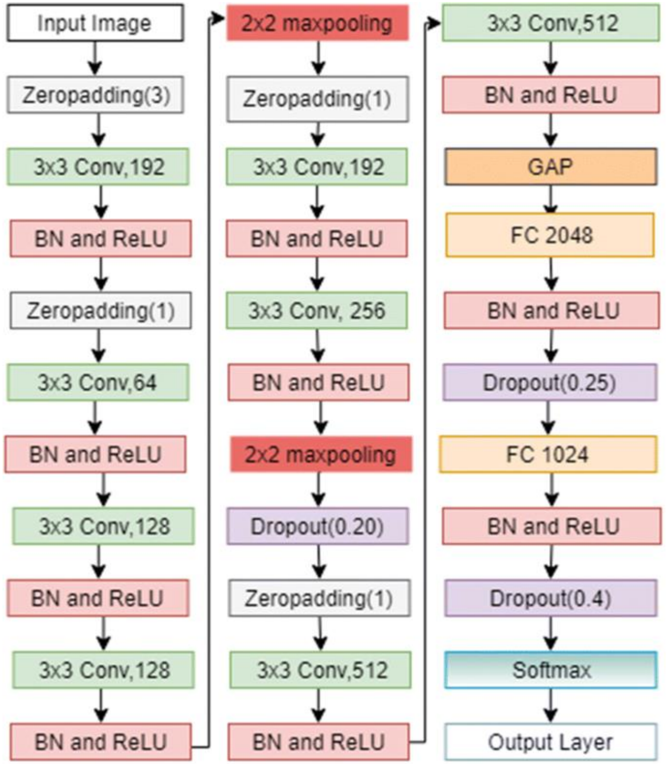
- Same idea!
- Compute all partial derivatives of the loss w.r.t the parameters.
  - For a large network, this is time consuming.
  - Backpropagation does this efficiently!
  - We will present the modular approach.



Modular Approach: Code layers, not networks!

- Layer specification – each layer should provide three functions:
  1. The layer output given its input (Forward)
  2. Loss gradient w.r.t. the input (Backward)
  3. Loss gradient w.r.t. parameters





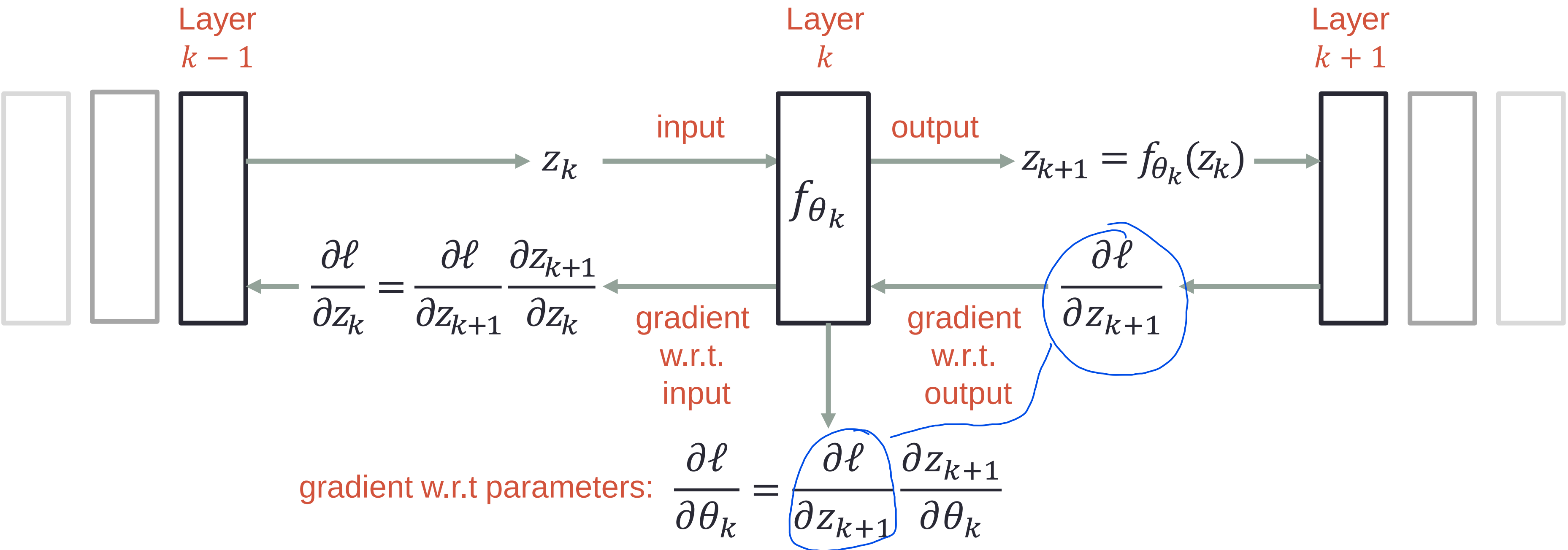
Input Image
Zeropadding(3)
3x3 Conv,192
BN and ReLU
Zeropadding(1)
3x3 Conv,64
BN and ReLU
3x3 Conv,128
BN and ReLU
3x3 Conv,128
BN and ReLU

2x2 maxpooling
Zeropadding(1)
3x3 Conv,192
BN and ReLU
3x3 Conv, 256
BN and ReLU
2x2 maxpooling
Dropout(0.20)
Zeropadding(1)
3x3 Conv,512
BN and ReLU

3x3 Conv,512
BN and ReLU
GAP
FC 2048
BN and ReLU
Dropout(0.25)
FC 1024
BN and ReLU
Dropout(0.4)
Softmax
Output Layer

Modular Approach: Code layers not networks!

- Layer specification – each layer should provide three functions:
  1. The layer output given its input (Forward)
  2. Loss gradient w.r.t. the input (Backward)
  3. Loss gradient w.r.t. parameters
- For instance, layer k implements a function f_{θ_k} with its parameters θ_k.



Layer $k-1$

Layer k

Layer $k+1$

input

output

z_k

$z_{k+1} = f_{\theta_k}(z_k)$

f_{θ_k}

$$\frac{\partial \ell}{\partial z_k} = \frac{\partial \ell}{\partial z_{k+1}} \frac{\partial z_{k+1}}{\partial z_k}$$

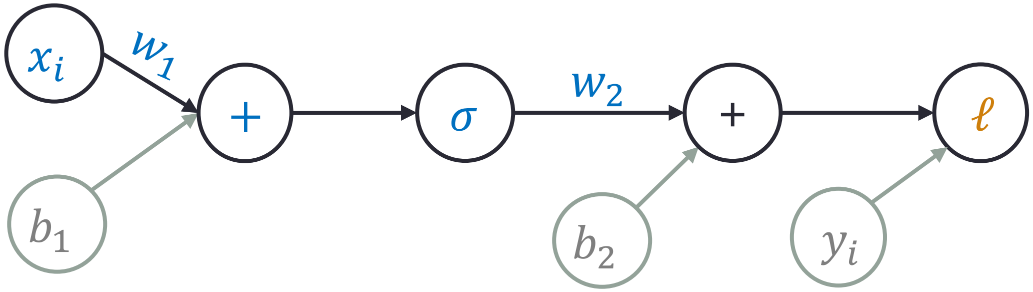
gradient w.r.t. input

$$\frac{\partial \ell}{\partial z_{k+1}}$$

gradient w.r.t. output

gradient w.r.t parameters: $\frac{\partial \ell}{\partial \theta_k} = \frac{\partial \ell}{\partial z_{k+1}} \frac{\partial z_{k+1}}{\partial \theta_k}$

- Let’s build the regression network we saw earlier with the modular approach



Linear

$$f(z) = w_1 z + b_1$$

$$\frac{\partial f}{\partial z} = w_1$$

$$\frac{\partial f}{\partial w_1} = z$$

$$\frac{\partial f}{\partial b_1} = 1$$

ReLU

$$f(z) = \max\{0, z\}$$

$$\frac{\partial f}{\partial z} = \mathbb{I}_{z>0}$$

Linear

$$f(z) = w_2 z + b_2$$

$$\frac{\partial f}{\partial z} = w_2$$

$$\frac{\partial f}{\partial w_2} = z$$

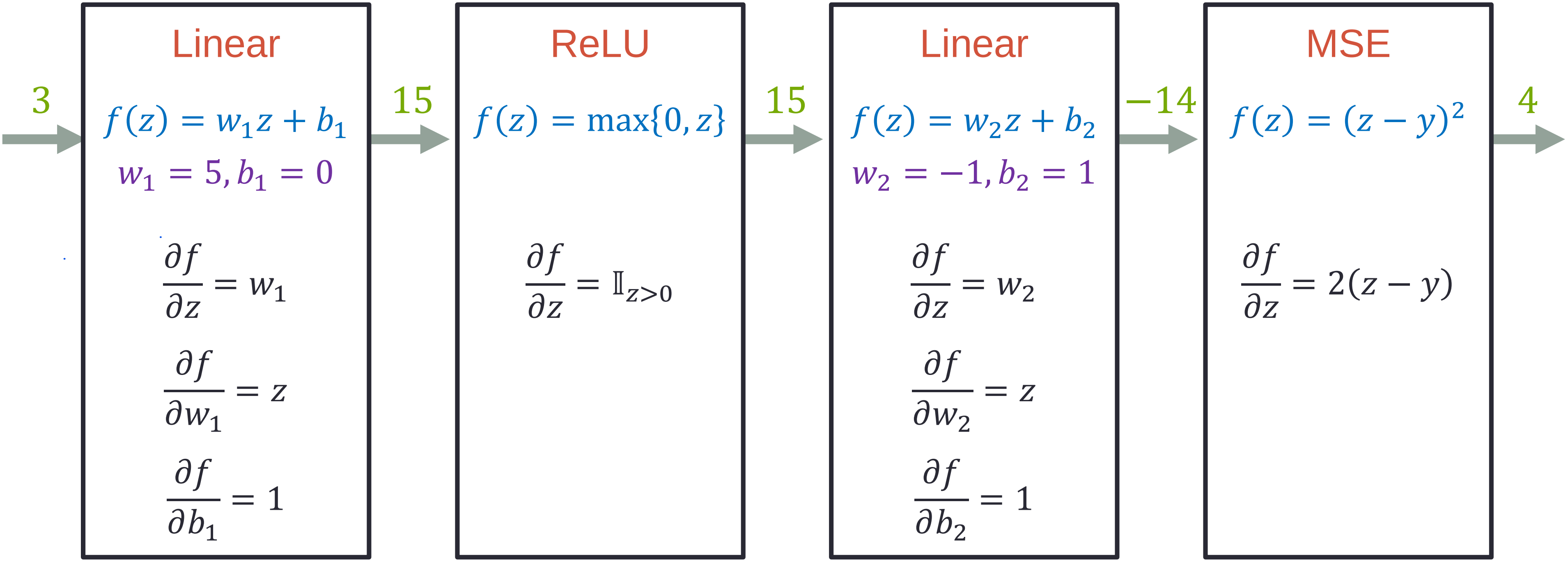
$$\frac{\partial f}{\partial b_2} = 1$$

MSE

$$f(z) = (z - y)^2$$

$$\frac{\partial f}{\partial z} = 2(z - y)$$

- Initialization: $w_1 = 5, w_2 = -1, b_1 = 0, b_2 = 1$
- Input: $x = 3,\ y = -12$
- First, we compute the forward pass.



3

Linear

$f(z) = w_1 z + b_1$

$w_1 = 5, b_1 = 0$

$$\frac{\partial f}{\partial z} = w_1$$

$$\frac{\partial f}{\partial w_1} = z$$

$$\frac{\partial f}{\partial b_1} = 1$$

15

ReLU

$f(z) = \max\{0, z\}$

$$\frac{\partial f}{\partial z} = \mathbb{I}_{z>0}$$

15

Linear

$f(z) = w_2 z + b_2$

$w_2 = -1, b_2 = 1$

$$\frac{\partial f}{\partial z} = w_2$$

$$\frac{\partial f}{\partial w_2} = z$$

$$\frac{\partial f}{\partial b_2} = 1$$

-14

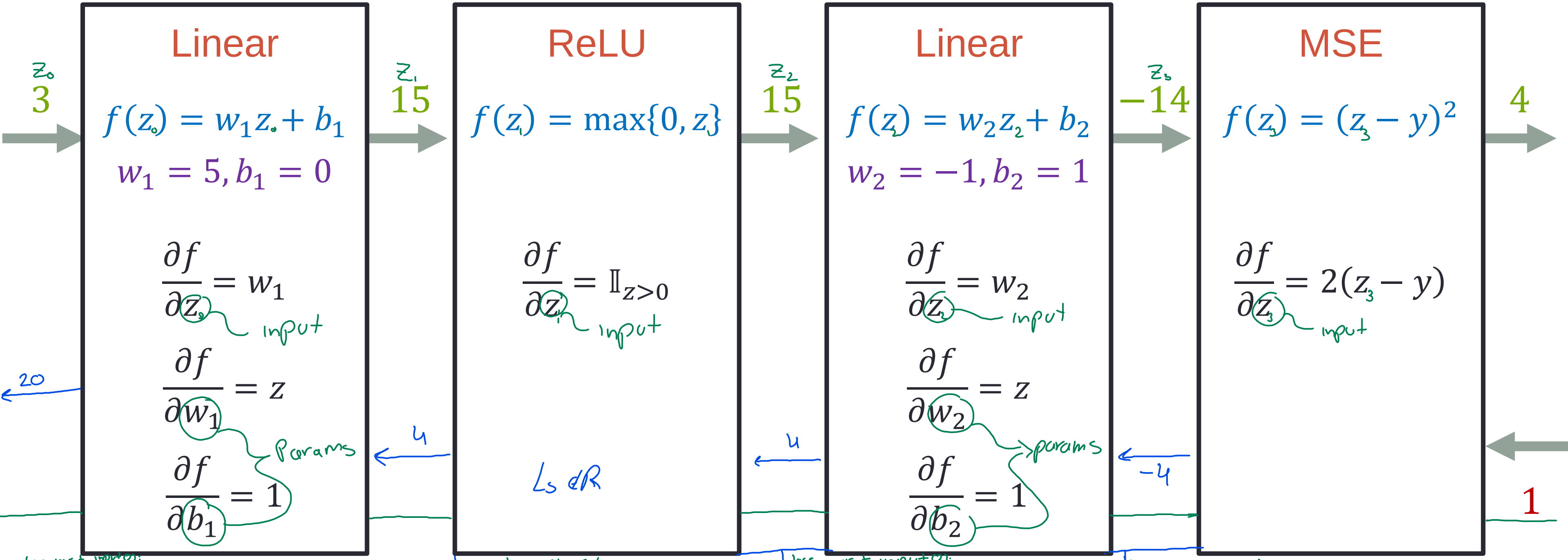
MSE

$f(z) = (z - y)^2$

$$\frac{\partial f}{\partial z} = 2(z - y)$$

4

- Initialization: $w_1 = 5, w_2 = -1, b_1 = 0, b_2 = 1$
- Input: $x = 3,\ y = -12$
- First, we compute the forward pass.
- Then, to compute gradients w.r.t θ, we perform a backward pass.



loss w.r.t input(2):

$$\frac{\partial \ell}{\partial z_0} = \frac{\partial \ell}{\partial z_1} \cdot \frac{\partial z_1}{\partial z_0} = 4 \cdot 5 = 20$$

loss w.r.t params(3):

$$\frac{\partial \ell}{\partial w_1} = \frac{\partial \ell}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_1} = 4 \cdot 3 = 12 \qquad \frac{\partial \ell}{\partial b_1} = \frac{\partial \ell}{\partial z_1} \cdot \frac{\partial z_1}{\partial b_1} = 4 \cdot 1 = 4$$

loss w.r.t input: (2):

$$\frac{\partial \ell}{\partial z_1} = \frac{\partial \ell}{\partial z_2} \cdot \frac{\partial z_2}{\partial z_1} = 4 \cdot 1$$

loss w.r.t input(2):

$$\frac{\partial \ell}{\partial z_2} = \frac{\partial \ell}{\partial z_3} \cdot \boxed{\frac{\partial z_3}{\partial z_2}} = (-4) \cdot (-1) = 4$$

w_2

loss w.r.t params(3):

$$\frac{\partial \ell}{\partial w_2} = \frac{\partial \ell}{\partial z_3} \cdot \frac{\partial z_3}{\partial w_2} = (-4) \cdot 15 = -60$$

loss w.r.t input(2):

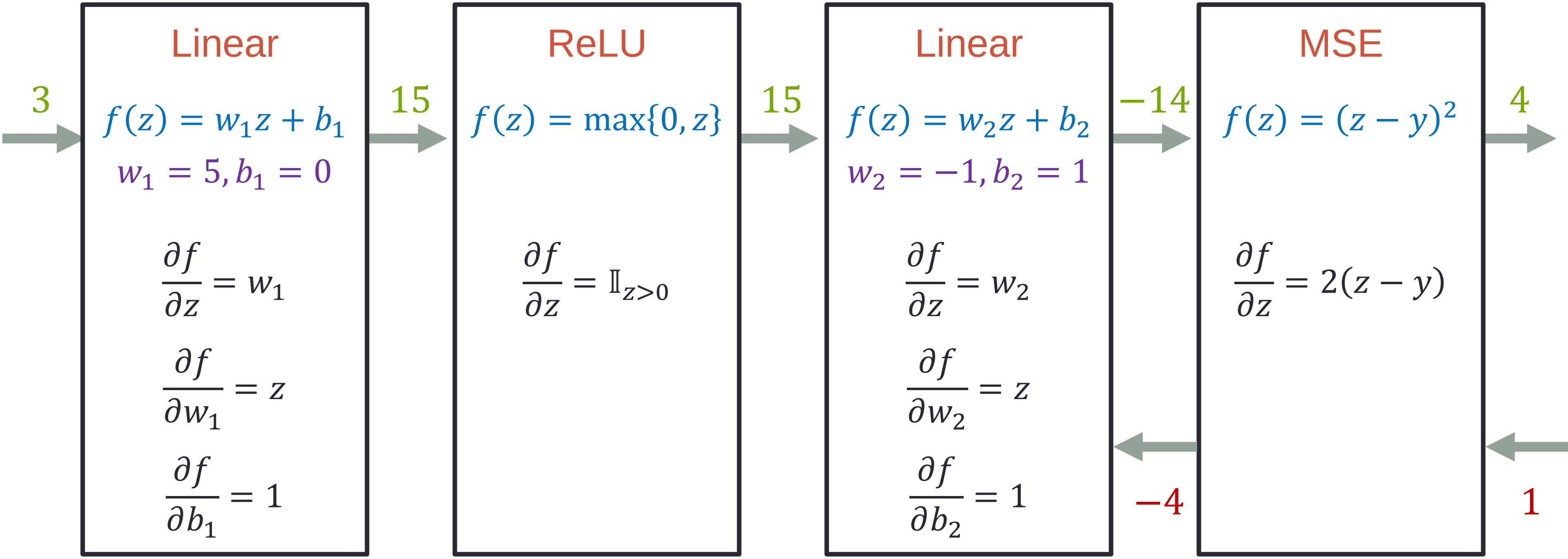
$$\frac{\partial \ell}{\partial z_3} = 2(-14 - (-12)) = -4$$

(3):

$$\frac{\partial \ell}{\partial b_2} = \frac{\partial \ell}{\partial z_1} \cdot \frac{\partial z_1}{\partial b_2} = (-4) \cdot 1 = -4$$

$$\frac{\partial \ell}{\partial \ell}$$

- Initialization: $w_1 = 5, w_2 = -1, b_1 = 0, b_2 = 1$
- Input: $x = 3,\ y = -12$
- First, we compute the forward pass.
- Then, to compute gradients w.r.t θ, we perform a backward pass.



Linear

$f(z) = w_1 z + b_1$

$w_1 = 5, b_1 = 0$

$$\frac{\partial f}{\partial z} = w_1$$

$$\frac{\partial f}{\partial w_1} = z$$

$$\frac{\partial f}{\partial b_1} = 1$$

ReLU

$f(z) = \max\{0, z\}$

$$\frac{\partial f}{\partial z} = \mathbb{I}_{z>0}$$

Linear

$f(z) = w_2 z + b_2$

$w_2 = -1, b_2 = 1$

$$\frac{\partial f}{\partial z} = w_2$$

$$\frac{\partial f}{\partial w_2} = z$$

$$\frac{\partial f}{\partial b_2} = 1$$

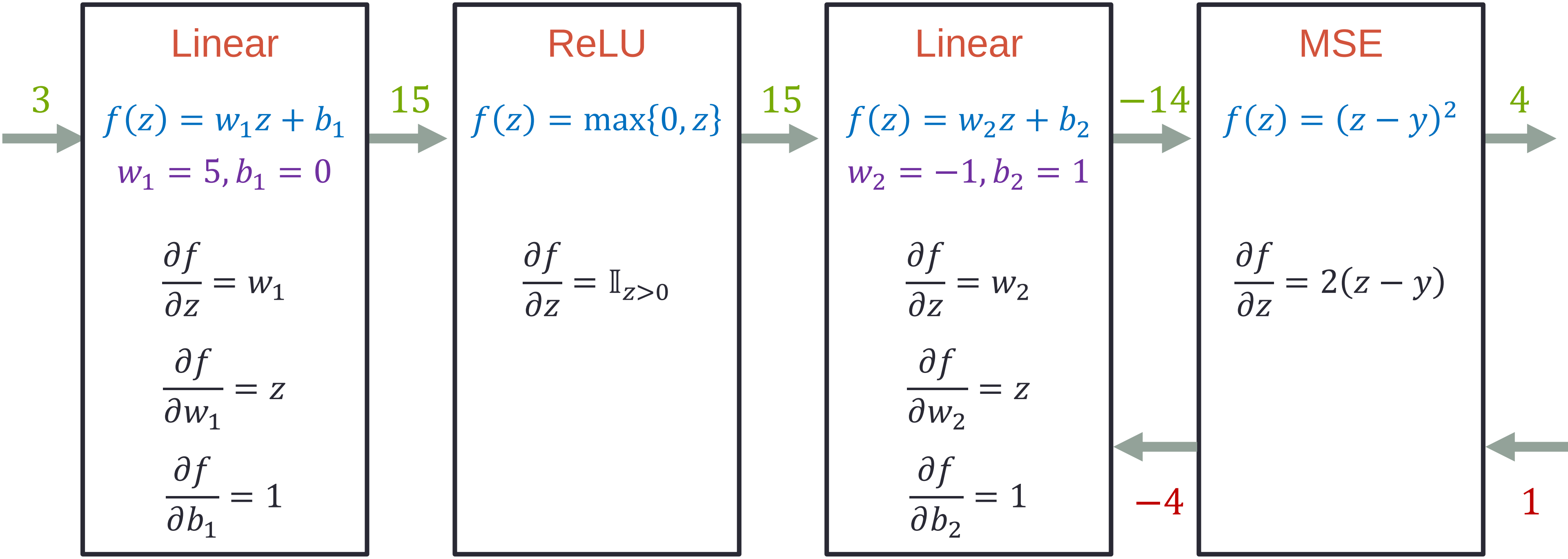
MSE

$f(z) = (z - y)^2$

$$\frac{\partial f}{\partial z} = 2(z - y)$$

$$\frac{\partial \ell}{\partial z} = \frac{\partial \ell}{\partial f}\frac{\partial f}{\partial z} = 1 \cdot -4$$

- Initialization: $w_1 = 5, w_2 = -1, b_1 = 0, b_2 = 1$
- Input: $x = 3,\ y = -12$
- First, we compute the forward pass.
- Then, to compute gradients w.r.t θ, we perform a backward pass.



3

Linear

$f(z) = w_1 z + b_1$

$w_1 = 5, b_1 = 0$

$$\frac{\partial f}{\partial z} = w_1$$

$$\frac{\partial f}{\partial w_1} = z$$

$$\frac{\partial f}{\partial b_1} = 1$$

15

ReLU

$f(z) = \max\{0, z\}$

$$\frac{\partial f}{\partial z} = \mathbb{I}_{z>0}$$

15

Linear

$f(z) = w_2 z + b_2$

$w_2 = -1, b_2 = 1$

$$\frac{\partial f}{\partial z} = w_2$$

$$\frac{\partial f}{\partial w_2} = z$$

$$\frac{\partial f}{\partial b_2} = 1$$

-14

-4

MSE

$f(z) = (z - y)^2$

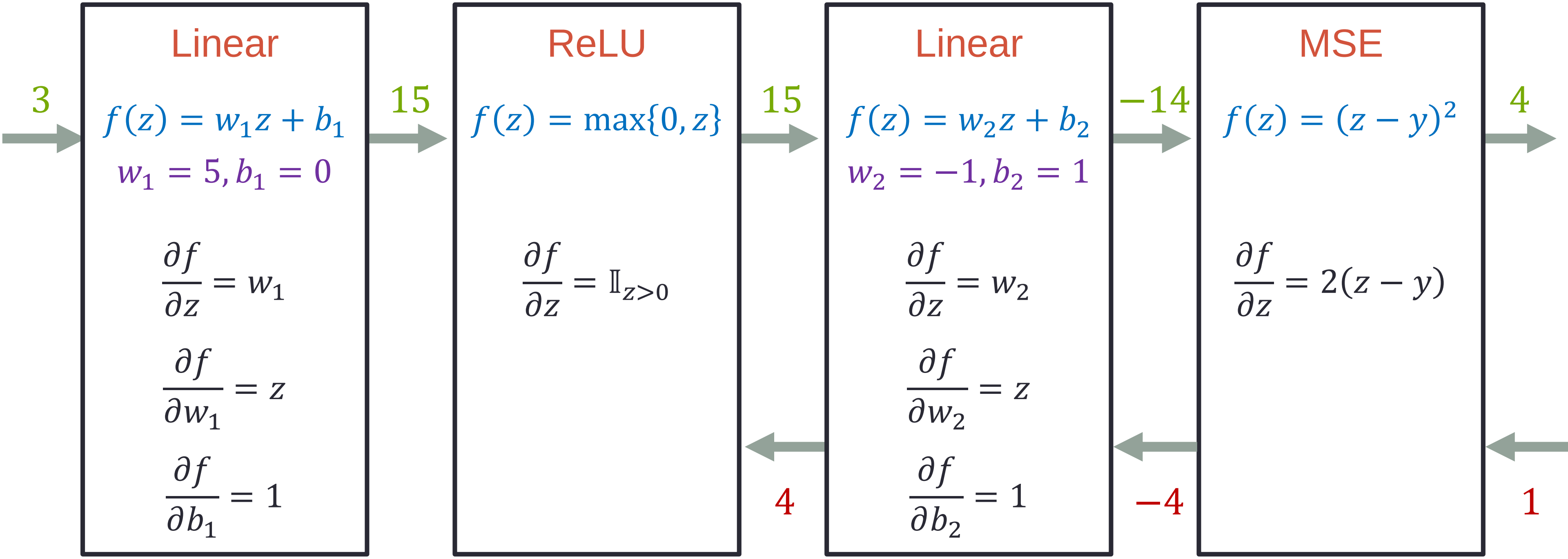
$$\frac{\partial f}{\partial z} = 2(z - y)$$

4

1

(no parameters for this layer)

- Initialization: $w_1 = 5, w_2 = -1, b_1 = 0, b_2 = 1$
- Input: $x = 3,\ y = -12$
- First, we compute the forward pass.
- Then, to compute gradients w.r.t θ, we perform a backward pass.



Linear

$f(z) = w_1 z + b_1$

$w_1 = 5, b_1 = 0$

$$\frac{\partial f}{\partial z} = w_1$$

$$\frac{\partial f}{\partial w_1} = z$$

$$\frac{\partial f}{\partial b_1} = 1$$

ReLU

$f(z) = \max\{0, z\}$

$$\frac{\partial f}{\partial z} = \mathbb{I}_{z>0}$$

Linear

$f(z) = w_2 z + b_2$

$w_2 = -1, b_2 = 1$

$$\frac{\partial f}{\partial z} = w_2$$

$$\frac{\partial f}{\partial w_2} = z$$

$$\frac{\partial f}{\partial b_2} = 1$$

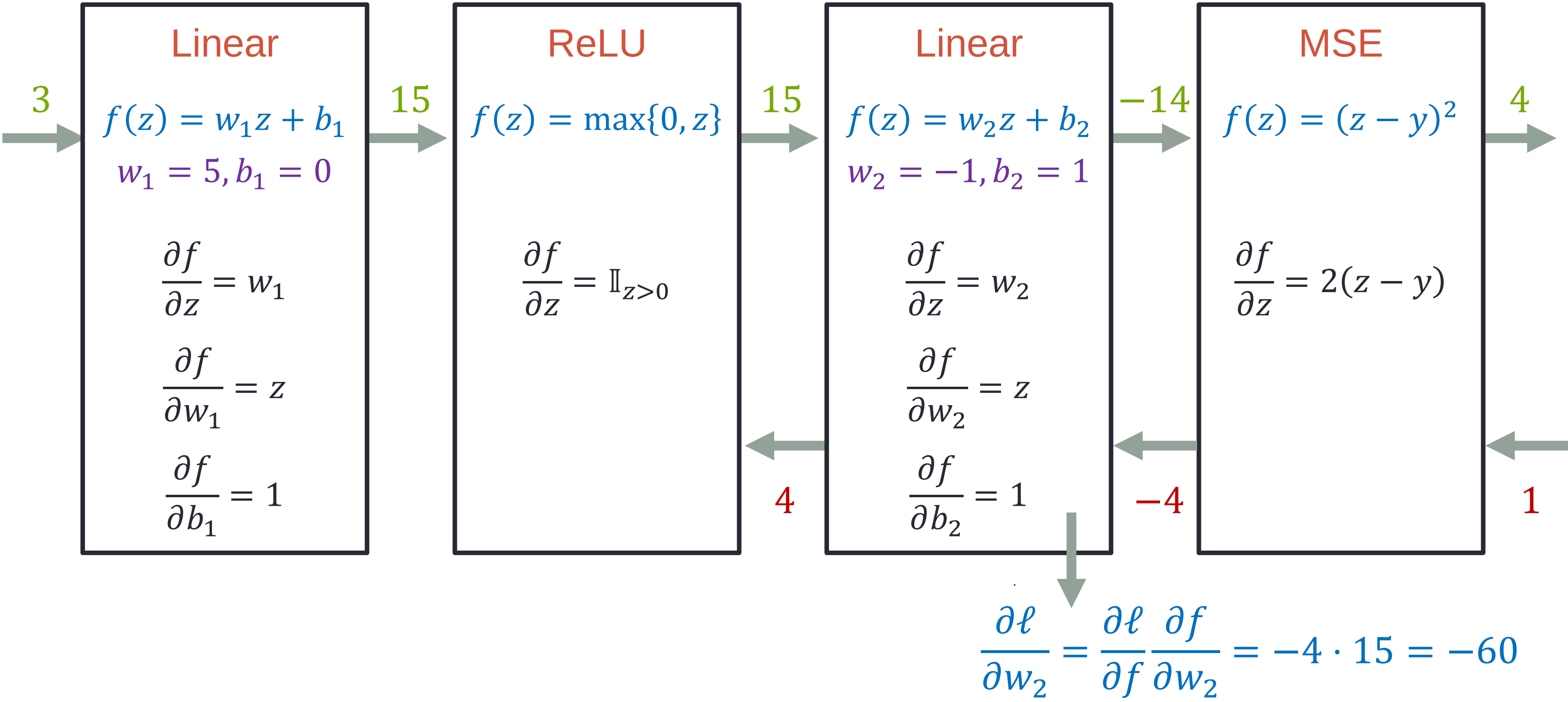
MSE

$f(z) = (z - y)^2$

$$\frac{\partial f}{\partial z} = 2(z - y)$$

$$\frac{\partial \ell}{\partial z} = \frac{\partial \ell}{\partial f}\frac{\partial f}{\partial z} = -4 \cdot (-1)$$

- Initialization: $w_1 = 5, w_2 = -1, b_1 = 0, b_2 = 1$
- Input: $x = 3,\ y = -12$
- First, we compute the forward pass.
- Then, to compute gradients w.r.t θ, we perform a backward pass.



Linear

$f(z) = w_1 z + b_1$

$w_1 = 5, b_1 = 0$

$$\frac{\partial f}{\partial z} = w_1$$

$$\frac{\partial f}{\partial w_1} = z$$

$$\frac{\partial f}{\partial b_1} = 1$$

ReLU

$f(z) = \max\{0, z\}$

$$\frac{\partial f}{\partial z} = \mathbb{I}_{z>0}$$

Linear

$f(z) = w_2 z + b_2$

$w_2 = -1, b_2 = 1$

$$\frac{\partial f}{\partial z} = w_2$$

$$\frac{\partial f}{\partial w_2} = z$$

$$\frac{\partial f}{\partial b_2} = 1$$

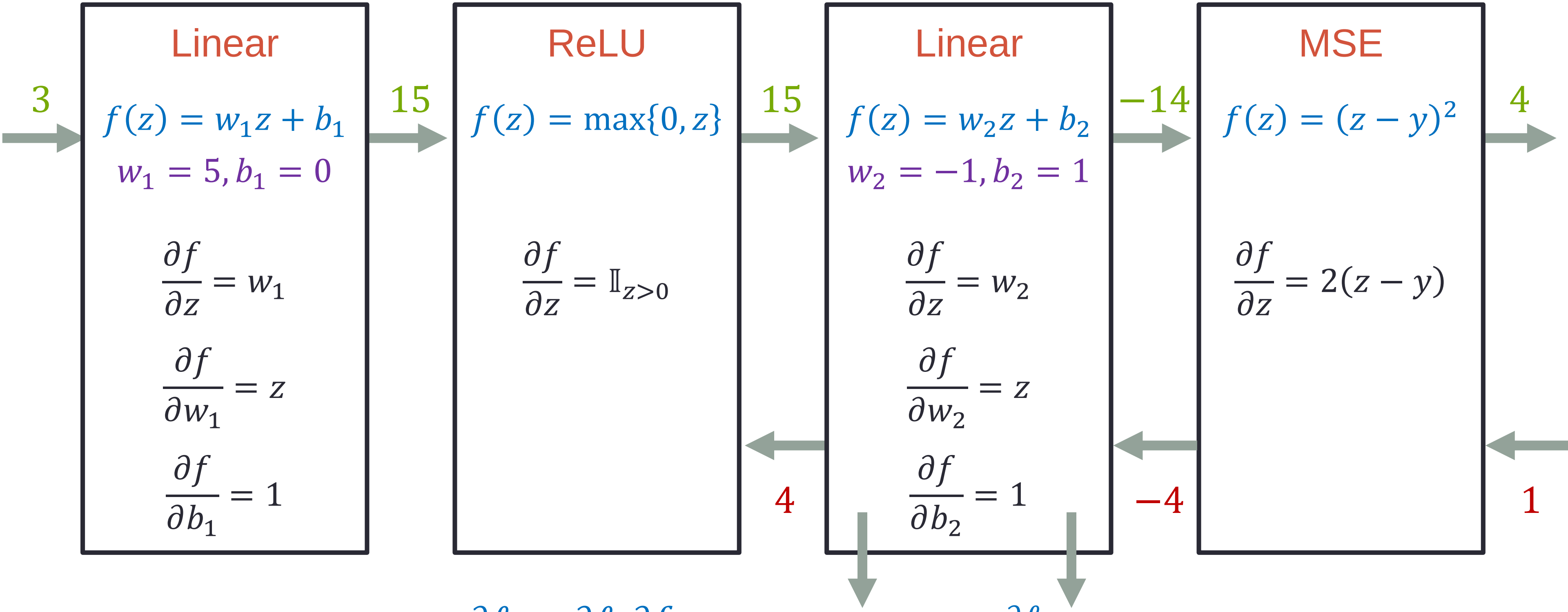
MSE

$f(z) = (z - y)^2$

$$\frac{\partial f}{\partial z} = 2(z - y)$$

$$\frac{\partial \ell}{\partial w_2} = \frac{\partial \ell}{\partial f}\frac{\partial f}{\partial w_2} = -4 \cdot 15 = -60$$

- Initialization: $w_1 = 5, w_2 = -1, b_1 = 0, b_2 = 1$
- Input: $x = 3,\ y = -12$
- First, we compute the forward pass.
- Then, to compute gradients w.r.t θ, we perform a backward pass.



Linear

$f(z) = w_1 z + b_1$

$w_1 = 5, b_1 = 0$

$$\frac{\partial f}{\partial z} = w_1$$

$$\frac{\partial f}{\partial w_1} = z$$

$$\frac{\partial f}{\partial b_1} = 1$$

ReLU

$f(z) = \max\{0, z\}$

$$\frac{\partial f}{\partial z} = \mathbb{I}_{z>0}$$

Linear

$f(z) = w_2 z + b_2$

$w_2 = -1, b_2 = 1$

$$\frac{\partial f}{\partial z} = w_2$$

$$\frac{\partial f}{\partial w_2} = z$$

$$\frac{\partial f}{\partial b_2} = 1$$

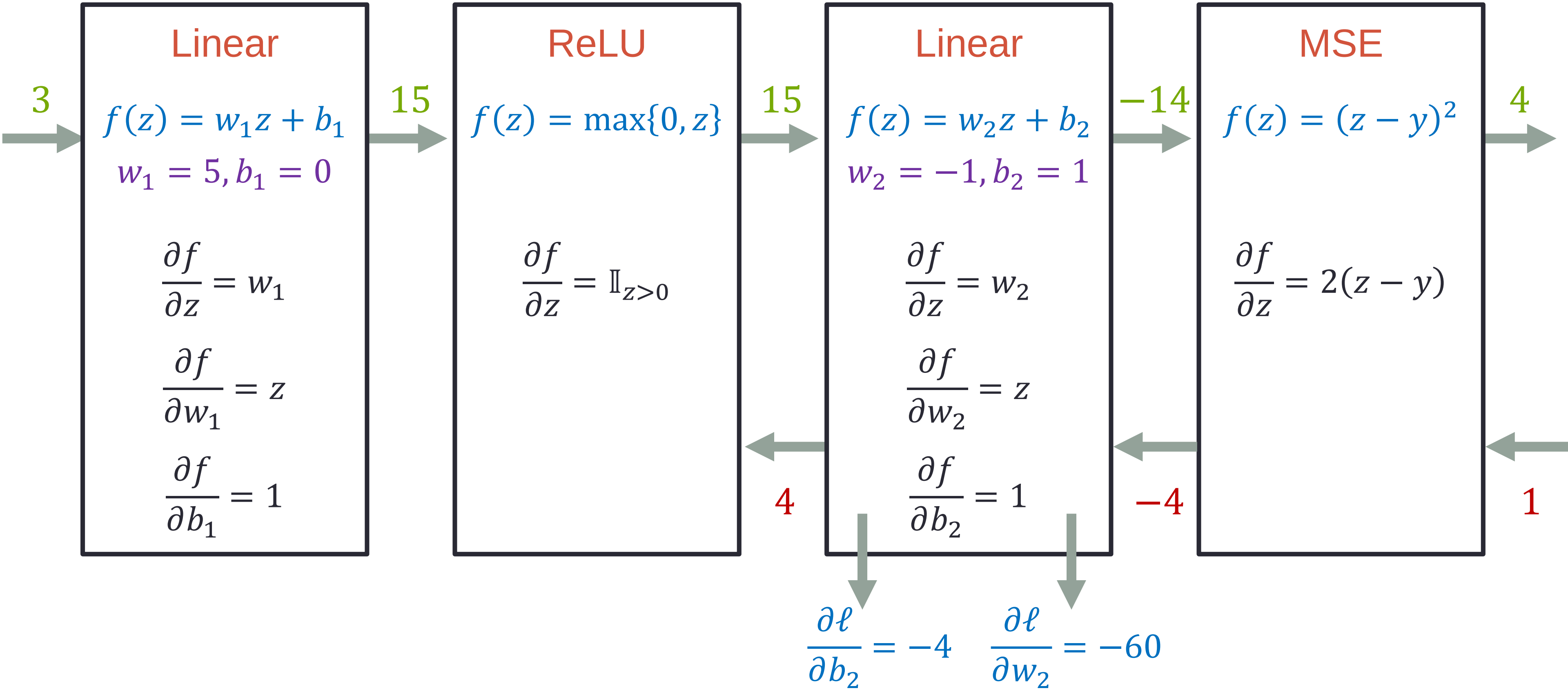
MSE

$f(z) = (z - y)^2$

$$\frac{\partial f}{\partial z} = 2(z - y)$$

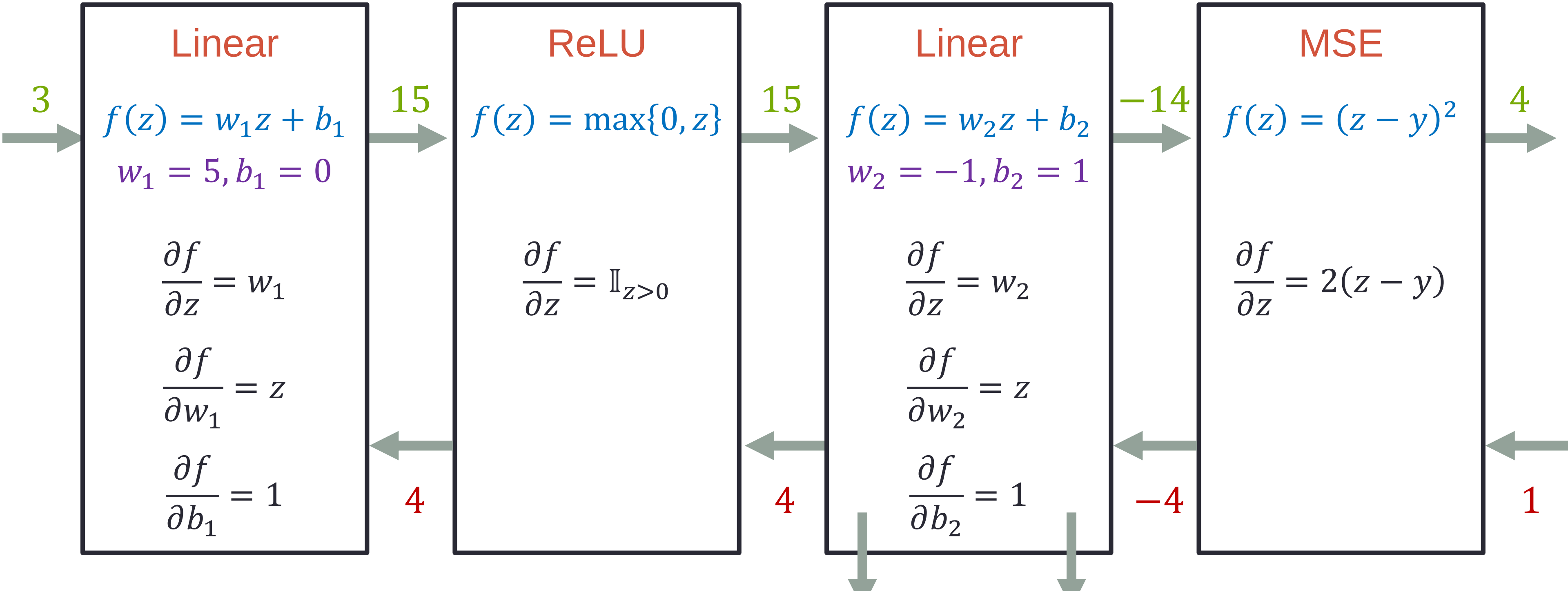
$$\frac{\partial \ell}{\partial b_2} = \frac{\partial \ell}{\partial f}\frac{\partial f}{\partial b_2} = -4 \cdot 1 = -4 \qquad \frac{\partial \ell}{\partial w_2} = -60$$

- Initialization: $w_1 = 5, w_2 = -1, b_1 = 0, b_2 = 1$
- Input: $x = 3,\ y = -12$
- First, we compute the forward pass.
- Then, to compute gradients w.r.t θ, we perform a backward pass.



$$\frac{\partial \ell}{\partial b_2} = -4 \qquad \frac{\partial \ell}{\partial w_2} = -60$$

- Initialization: $w_1 = 5, w_2 = -1, b_1 = 0, b_2 = 1$
- Input: $x = 3,\ y = -12$
- First, we compute the forward pass.
- Then, to compute gradients w.r.t θ, we perform a backward pass.



3

Linear

$f(z) = w_1 z + b_1$

$w_1 = 5, b_1 = 0$

$\frac{\partial f}{\partial z} = w_1$

$\frac{\partial f}{\partial w_1} = z$

$\frac{\partial f}{\partial b_1} = 1$

15

4

ReLU

$f(z) = \max\{0, z\}$

$\frac{\partial f}{\partial z} = \mathbb{I}_{z>0}$

15

4

Linear

$f(z) = w_2 z + b_2$

$w_2 = -1, b_2 = 1$

$\frac{\partial f}{\partial z} = w_2$

$\frac{\partial f}{\partial w_2} = z$

$\frac{\partial f}{\partial b_2} = 1$

-14

-4

MSE

$f(z) = (z - y)^2$

$\frac{\partial f}{\partial z} = 2(z - y)$

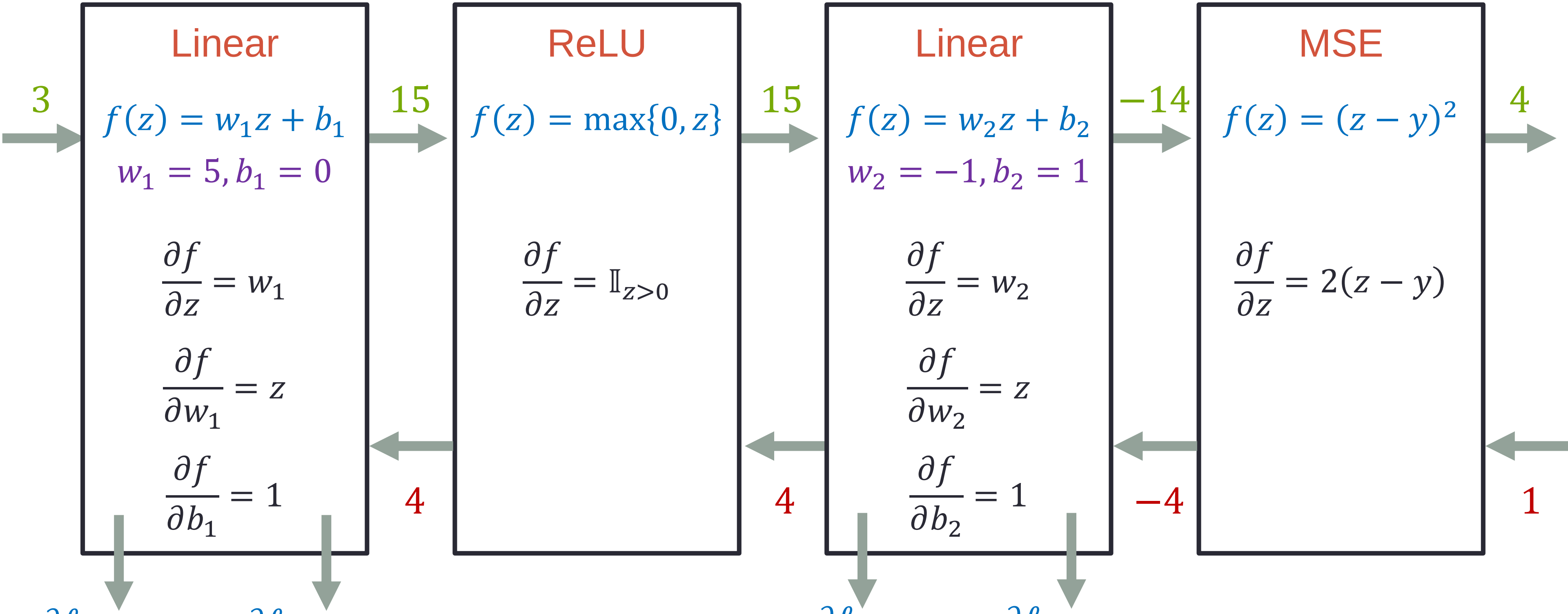
4

1

$$\frac{\partial \ell}{\partial z} = \frac{\partial \ell}{\partial f}\frac{\partial f}{\partial z} = 4 \cdot 1$$

$$\frac{\partial \ell}{\partial b_2} = -4 \qquad \frac{\partial \ell}{\partial w_2} = -60$$

- Initialization: $w_1 = 5, w_2 = -1, b_1 = 0, b_2 = 1$
- Input: $x = 3,\ y = -12$
- First, we compute the forward pass.
- Then, to compute gradients w.r.t θ, we perform a backward pass.



3

Linear

$f(z) = w_1 z + b_1$

$w_1 = 5, b_1 = 0$

$$\frac{\partial f}{\partial z} = w_1$$

$$\frac{\partial f}{\partial w_1} = z$$

$$\frac{\partial f}{\partial b_1} = 1$$

15

4

ReLU

$f(z) = \max\{0, z\}$

$$\frac{\partial f}{\partial z} = \mathbb{I}_{z>0}$$

15

4

Linear

$f(z) = w_2 z + b_2$

$w_2 = -1, b_2 = 1$

$$\frac{\partial f}{\partial z} = w_2$$

$$\frac{\partial f}{\partial w_2} = z$$

$$\frac{\partial f}{\partial b_2} = 1$$

-14

-4

MSE

$f(z) = (z - y)^2$

$$\frac{\partial f}{\partial z} = 2(z - y)$$

4

1

$$\frac{\partial \ell}{\partial b_1} = 4 \cdot 1 \qquad \frac{\partial \ell}{\partial w_1} = 4 \cdot 3 = 12$$

$$\frac{\partial \ell}{\partial b_2} = -4 \qquad \frac{\partial \ell}{\partial w_2} = -60$$

Let’s play!

- Simple regression task
- Tensorflow playground (extra)

Summary

- Neural networks can represent complicated functions
- Backpropagation:
  - Efficiently computes the gradients of the loss w.r.t the network parameters.
  - Widely used for gradient optimization methods

- Recommended read/watch: 3Blue1Brown – Neural networks