Deep Learning 1

Foma Shipilov @fomazu foma@shipilov.ru

ОКР(0.3 × ЭКЗ + $\underbrace{0{,}3\text{ БДЗ} + 0{,}25\text{ МДЗ} + 0{,}15\text{ ПР}}_{\text{НАКОП} \geq 7{,}5}$)

- Backpropagation
- Fully-connected networks
- ? Classification

$$a(x) = w_1 x_1 + \ldots + w_d x_d = f(x, w) \qquad x \in \mathbb{R}^d$$

$w \in W$ weights

$\uparrow$ input features

$\{(x_i, y_i)\}_{i=1}^{\ell}$ training set

$L(\hat{y}, y)$ — loss function

$$Q(w) = \frac{1}{\ell} \sum_{i=1}^{\ell} L\big(f(x_i, w), y_i\big) \to \min_w$$

Backprop. alternatives

Historic:
- Evolutionary algorithm 1950s
- Reservoir computing (2000) → ~~~ → lin regr →

Modern:
- Direct feedback alignment (2016)

$$y = f(g(x)) \qquad \frac{dy}{dx}(x) = \left.\frac{df}{dg}\right|_{g(x)} \cdot \left.\frac{dg}{dx}\right|_{x} \leftarrow \text{fixed random}$$

- Forward-forward (2022)

Backpropagation

1. $y = f(g(x))$ $\qquad \frac{dy}{dx} = \frac{df}{dg} \cdot \frac{dg}{dx}$

2. $y = f(g_1(x), g_2(x), \dots, g_n(x))$

$$\frac{dy}{dx} = \sum_{i=1}^{n} D_i f \frac{dg_i}{dx}$$

$$D_i f = \frac{df}{dz_i} \qquad f(z_1 \dots z_n)$$

symbolic $\qquad x^2 \to \underbrace{2}\underbrace{x} \quad \forall x$

numeric $\qquad x^2 \to \{x := 3\} \to 6$

Computational Graph

$$f = x_1^2(x_2x_3 - 2x_1^2) = x_1^2x_2x_3 - 2x_1^4$$

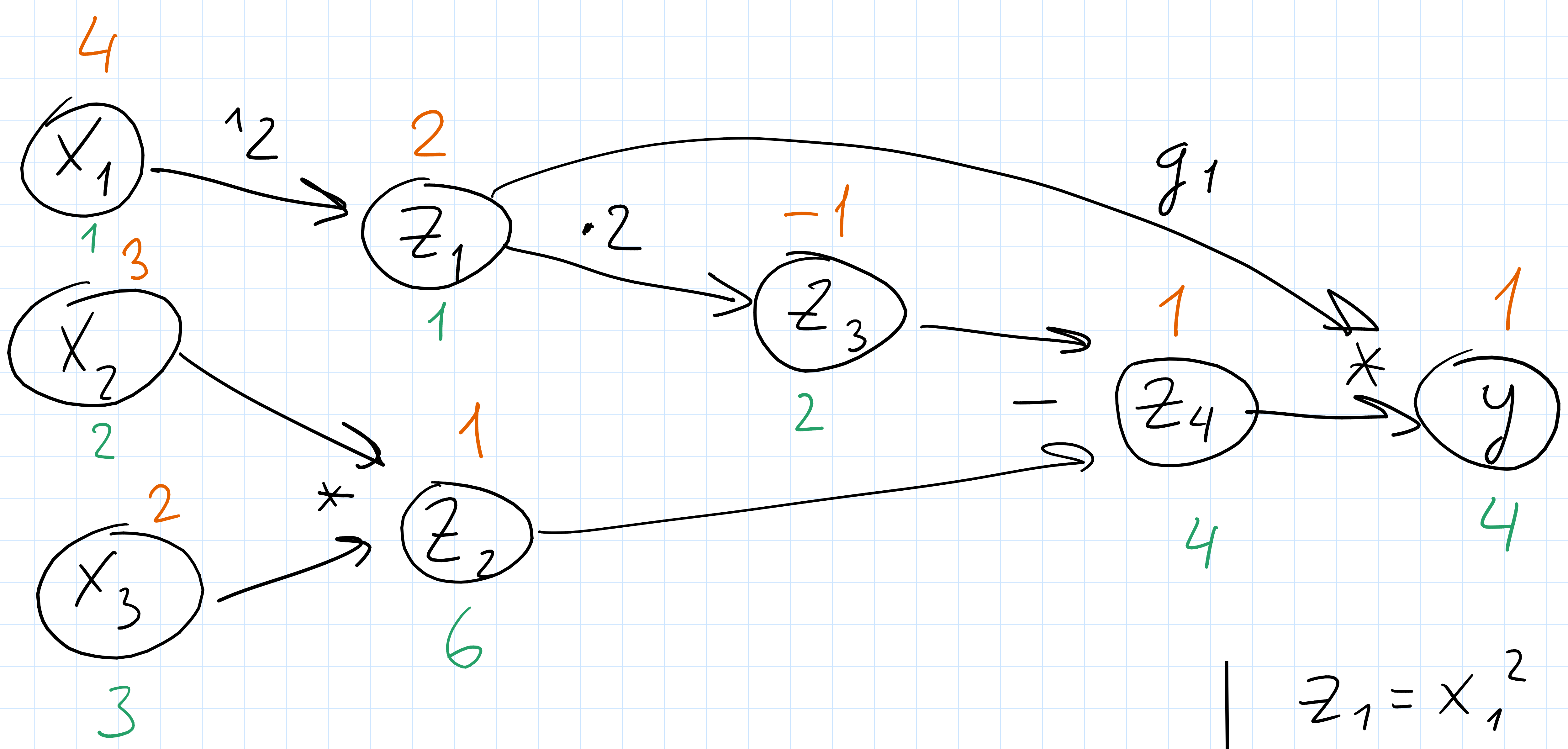
$$\frac{df}{dx_1} = 2x_1x_2x_3 - 8x_1^3 = 4$$

$$\frac{df}{dx_2} = x_1^2x_3 = 3$$

$$\frac{df}{dx_3} = x_1^2x_2 = 2$$

$$x = (1, 2, 3)$$

$$y = x_1^2(x_2x_3 - 2x_1^2)$$



$$z_1 = x_1^2$$

$$z_2 = x_2 x_3$$

$$z_3 = 2 z_1$$

$$z_4 = z_2 - z_3$$

$$y = z_1 z_4$$

$$\frac{dy}{dy} = 1$$

!

$$\frac{dy}{dz_4} = \frac{dy}{dy} \cdot \frac{dy}{dz_4} = 1 \cdot z_1 = 1$$

$$\frac{dy}{dz_3} = \frac{dy}{dz_4} \cdot \frac{dz_4}{dz_3} = 1 \cdot (-1) = -1$$

$$\frac{dy}{dz_4} \cdot \frac{dz_4}{dz_3}$$

$$\frac{dy}{dz_1} = z_4 \cdot 1 + \frac{dy}{dz_3} \cdot \frac{dz_3}{dz_1} = 4 + (-1) \cdot 2 = 2$$

$$y = f(g_1(z_1), g_2(z_1)) = g_1 g_2 = z_1 z_4$$

$$g_1(z_1) = z_1 \qquad g_2(z_1) = z_4(z_1)$$

$$\frac{dg_1}{dz_1} = 1 \qquad \frac{dg_2}{dz_1} = \frac{dz_4}{dz_1} = \frac{dz_4}{dz_3} \cdot \frac{dz_3}{dz_1}$$

$$\frac{dy}{dz_2} = \frac{dy}{dz_4} \cdot \frac{dz_4}{dz_2} = 1 \cdot 1 = 1$$

$$\frac{dy}{dx_1} = \frac{dy}{dz_1} \cdot \frac{dz_1}{dx_1} = 2 \cdot 2x_1 = 4$$

$$\frac{dy}{dx_2} = \frac{dy}{dz_2} \cdot \frac{dz_2}{dx_2} = 1 \cdot x_3 = 3$$

$$\frac{dy}{dx_3} = \frac{dy}{dz_2} \cdot \frac{dz_2}{dx_3} = 1 \cdot x_2 = 2$$

$x \in \mathbb{R}^d$ $\quad a(x) = \langle w, x \rangle = w^T x \quad w \in \mathbb{R}^d$

$w \in \mathbb{R}^d$

$z_1 = W_1 x + b_1 \qquad W_1 \in \mathbb{R}^{d_1 \times d} \qquad b_1 \in \mathbb{R}^{d_1}$

$z_2 = W_2 z_1 + b_2 \qquad W_2 \in \mathbb{R}^{d_2 \times d_1} \qquad b_2 \in \mathbb{R}^{d_2}$

...

$z_n = W_n z_{n-1} + b_n \qquad W_n \in \mathbb{R}^{d_n \times d_{n-1}} \qquad b_n \in \mathbb{R}^{d_n}$

$$z_n = \underbrace{W_n \cdot \ldots \cdot W_2 W_1}_{W \in \mathbb{R}^{d_n \times d}} x + b$$

$\sigma(\cdot) : \mathbb{R} \to \mathbb{R}$ activation function (non-linear)

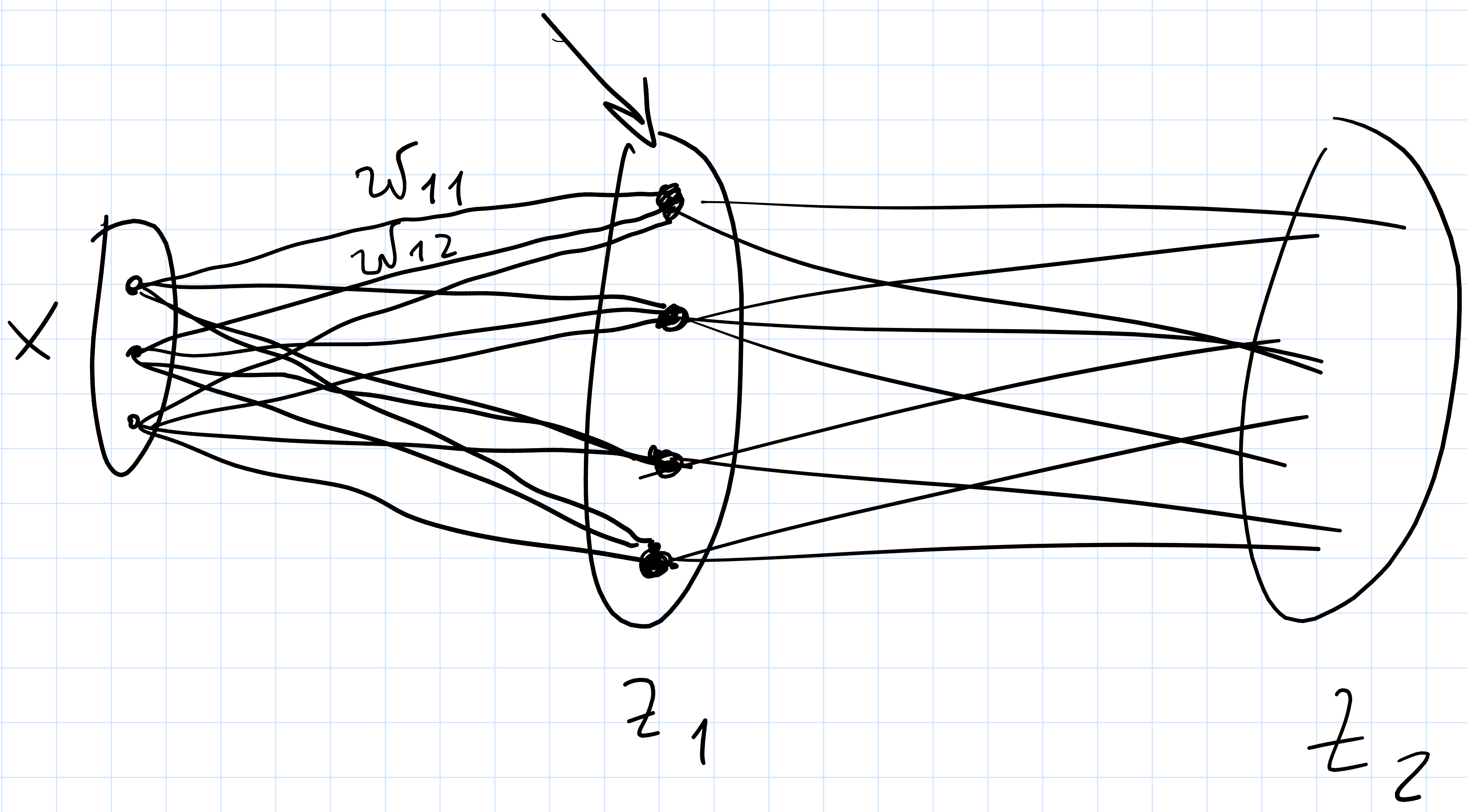
$z_i = \sigma(W_i z_{i-1} + b_i)$

dense / linear / fully-connected layer

z_i - latent representation

$$z_i = \left(z_i^{(1)}, \dots, z_i^{(d_i)}\right)$$

neuron

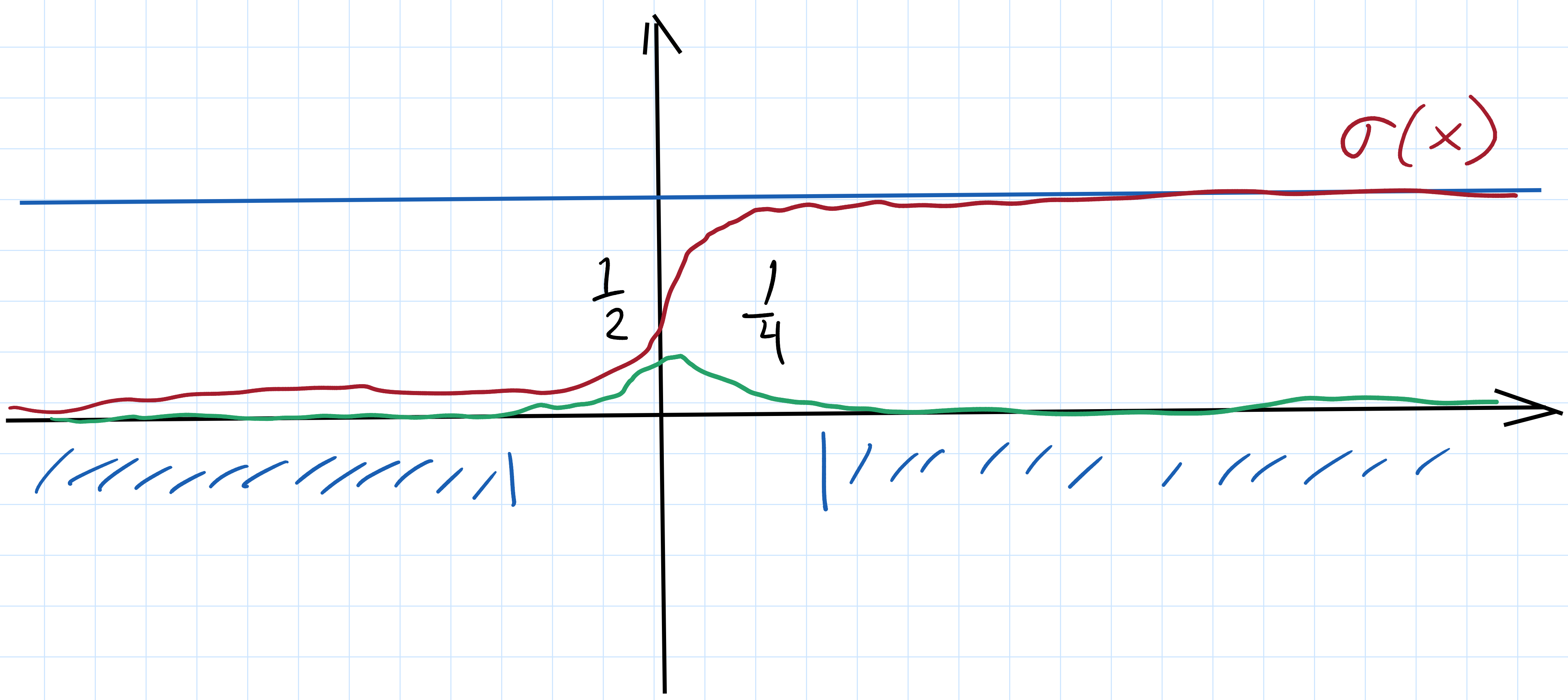


MLP Multi Layer Perceptron

FCN Fully Connected Network

1. Sigmoid $\qquad \sigma(x) = \dfrac{1}{1+e^{-x}} \qquad \sigma'(x) = \sigma(1-\sigma)$

$$\frac{e^{-x}}{(1+e^{-x})^2}$$

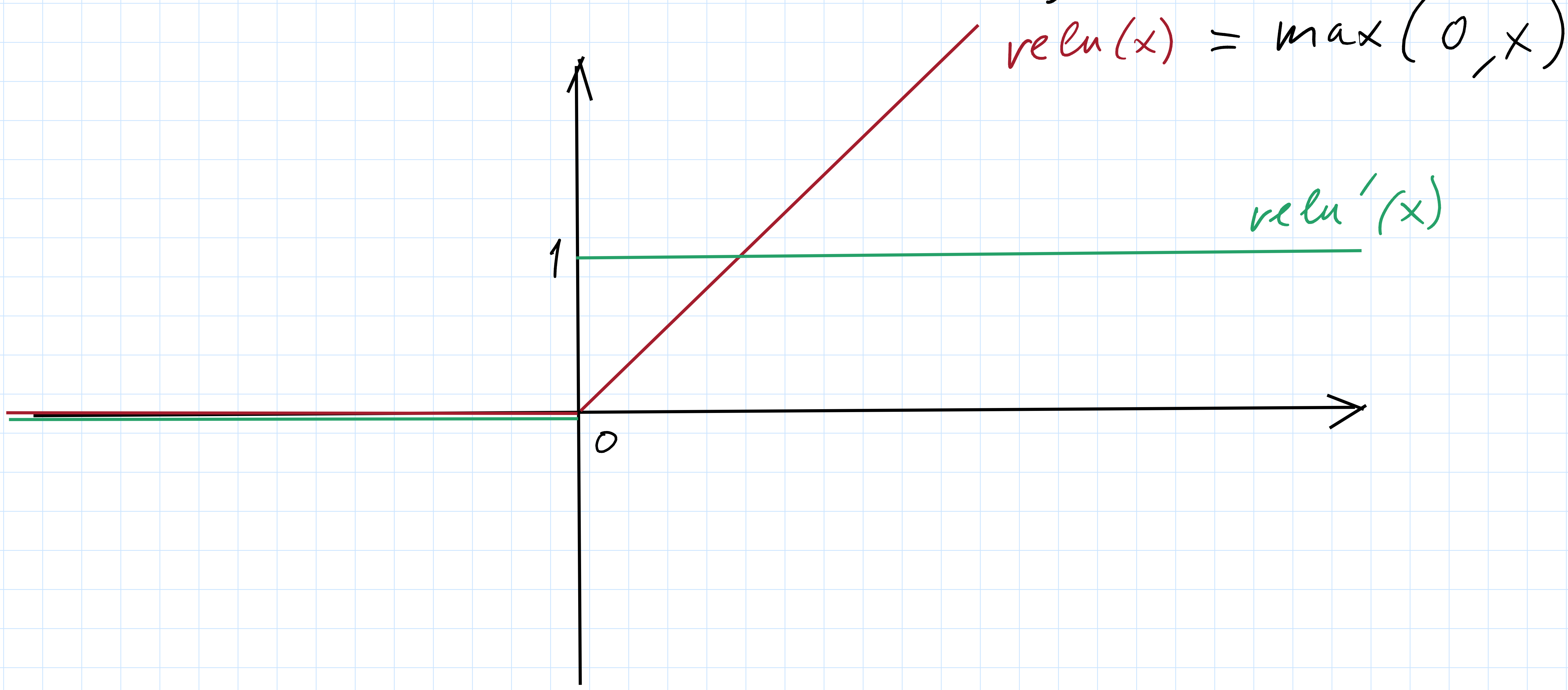


$$u_i = W_i z_i + b_i \qquad \frac{dl}{dz_i} = \frac{dl}{d_{i+1}} \cdot \underbrace{\sigma'(u_i)}_{\approx 0} \cdot \frac{du}{dz_i}$$

$$z_{i+1} = \sigma(u_i)$$

vanishing gradients

2. ReLU (rectified linear unit)



3. Leaky ReLU (α) = max(αx, x)