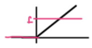
3- neural networks

biological neuron = all spikes (brief impulses) have the same magnitude and duration
↳ information is coded in the rate of the spikes

synapses = inputs (x) ⟶ dendrites = weighs (w) ⟶ soma = transfer func (Σ) ⟶ activation treshold

history of artificial neuron
↳ linear threshold logic unit = $x+y+2 \to$ and , $x+y+1 \to$ or , $-x \to$ not (no solution for xor)
↳ perceptron = $w^{next} = w^{curr} + \eta (y_i - \bar{y}_i) x_i$ $\quad \bar{y}_i \begin{cases} 1 & \text{if } wx_i > 0 \\ 0 & \text{otherwise} \end{cases}$
↳ adaline = $w^{next} = w^{curr} + \eta (y_i - w^{curr} \cdot x_i) x_i$ $\quad y_i \in \{-1, 1\}$ ⟶ improved learning rule
↳ back propagation = application of the chain rule in calculus

activation functions
↳ sigmoid = $\sigma(x) = \frac{1}{1+e^{-x}}$ $\quad \frac{d\sigma(x)}{dx} = \sigma(x) \cdot (1-\sigma(x))$  derivatives $\quad R \to [0,1]$
  ↳ since it is always positive, it introduces a bias for the next layer, which is not good
↳ hyperbolic tangent (tanh) = $\frac{d \tanh(x)}{dx} = 1 - \tanh^2(x)$  $R \to [-1, 1]$
↳ rectified linear units (relu) = $\varphi(x) = \max(0, x)$ $\quad \frac{d\varphi(x)}{dx} = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{else} \end{cases}$ 
  ↳ converges 6x faster than sigmoid / tanh
↳ leaky relu (ReLU) = $f(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha x & \text{otherwise} \end{cases}$ $\quad \alpha \to$ learned during training 
  ↳ parametric relu
↳ maxout = $\max(w_1^T x + b_1 ,\ w_2^T x + b_2 \ldots)$ generalization of ReLU and LeakyReLu

stochastic gradient descent

★ in stochastic gradient descent it is necessary to decrease the learning rate over time
↳ because noise (the random sampling of m training examples) may not vanish even the minimum is reached

momentum = helps accelate gradients vectors in the right direction
↳ without = $w^{(t+1)} = w^{(t)} - \eta g^{(t)}$
↳ with momentum = $w^{(t+1)} = w^{(t)} + v^{(t)}$ $\quad v^{(t)} = \alpha v^{(t-1)} - \eta g^{(t)}$ (exponential decay)
↳ size of the step depends on how large and aligned the subgradients are