Computational Intelligence

LMU Munich

winter term 2024/2025

Thomas Gabor
Claudia Linnhoff-Popien

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

schedule update

49	2024-12-03 Lecture #8	2024-12-05 Lecture #9
50	2024-12-10 Writing Exercise #4	2024-12-12 Writing Exercise #5
51	2024-12-17 Lecture #10	2024-12-19 Reading Exercise #3

Algorithm 5 (gradient descent). Let $\mathcal{D} = (\mathcal{X}, \mathcal{T}, \tau, e, \langle x_u \rangle_{0 \le u \le t})$ be an optimization process. Let $\mathcal{T}$ be continuous ($\mathcal{T} = \mathbb{R}$, e.g.) and let $\tau' : \mathcal{X} \to \mathcal{T}$ be the first derivative of τ. The process $\mathcal{D}$ continues via gradient descent (with update rate $\alpha \in \mathbb{R}^+$) if e is of the form

$$e(\langle x_u \rangle_{0 \le u \le t}, \tau) = x_{t+1} = x_t - \alpha \cdot \tau'(x_t).$$

The learning rate α can also be given as a function, usually $\alpha : \mathbb{N} \to \mathbb{R}$ so that $e(\langle x_u \rangle_{0 \le u \le t}, \tau) = x_{t+1} = x_t - \alpha(t) \cdot \tau'(x_t)$.
If τ is stochastic, this process is called stochastic gradient descent (SGD).

recap

Algorithm 6 (gradient descent (policy)). Let π_θ be a policy π that depends on vector of continuous parameters $\theta \in \Theta$ such that usually $\Theta = \mathbb{R}^N$ for some N. Let $\tau : \Theta \to \mathcal{T}$ be a target function on the parameters θ of a policy π_θ. Let $\mathcal{T}$ be continuous ($\mathcal{T} = \mathbb{R}$, e.g.) and let $\tau' : \Theta \to \mathcal{T}$ be the first derivative of τ, i.e., $\tau'(\theta) = \frac{\partial \tau(\theta)}{\partial \theta}$. If $\mathcal{D} = (\Theta, \mathcal{T}, \tau, e, \langle x_u \rangle_{0 \le u \le t})$ is an optimization process that continues via gradient descent, $\mathcal{D}$ is a process of policy optimization via gradient descent.

Many Variants of Gradient Descent

recap

source: https://optimization.cbe.cornell.edu/index.php?title=AdaGrad

Algorithm 1: AdaGrad general algorithm

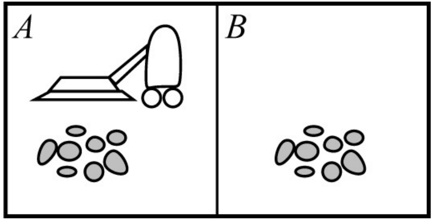
η: Stepsize ;
$f(x)$: Stochastic objective function ;
x_1: Initial parameter vector;
for $t = 1$ *to* T **do**
  Evaluate $f_t(x_t)$;
  Get and save g_t ;
  $G_t \leftarrow \sum_{\tau=1}^{t} g_\tau g_\tau^\top$;
  $x_{t+1} \leftarrow x_t - \eta G_t^{-1/2} g_t$;
end
return x_t

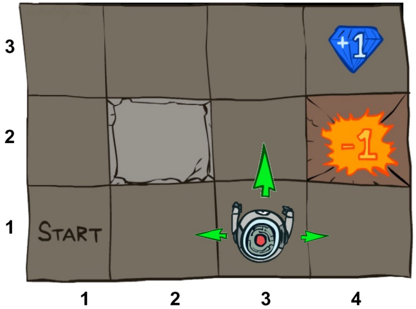
source: https://arxiv.org/pdf/1412.6980.pdf%5D

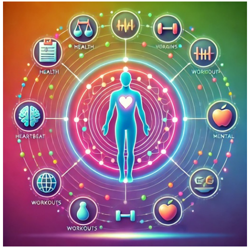
Algorithm 1: *Adam*, our proposed algorithm for stochastic optimization. See section 2 for details, and for a slightly more efficient (but less clear) order of computation. g_t^2 indicates the elementwise square $g_t \odot g_t$. Good default settings for the tested machine learning problems are $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. All operations on vectors are element-wise. With β_1^t and β_2^t we denote β_1 and β_2 to the power t.

Require: α: Stepsize
Require: $\beta_1, \beta_2 \in [0, 1)$: Exponential decay rates for the moment estimates
Require: $f(\theta)$: Stochastic objective function with parameters θ
Require: θ_0: Initial parameter vector
  $m_0 \leftarrow 0$ (Initialize 1st moment vector)
  $v_0 \leftarrow 0$ (Initialize 2nd moment vector)
  $t \leftarrow 0$ (Initialize timestep)
  while θ_t not converged **do**
    $t \leftarrow t + 1$
    $g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ (Get gradients w.r.t. stochastic objective at timestep t)
    $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (Update biased first moment estimate)
    $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (Update biased second raw moment estimate)
    $\widehat{m}_t \leftarrow m_t / (1 - \beta_1^t)$ (Compute bias-corrected first moment estimate)
    $\widehat{v}_t \leftarrow v_t / (1 - \beta_2^t)$ (Compute bias-corrected second raw moment estimate)
    $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \widehat{m}_t / (\sqrt{\widehat{v}_t} + \epsilon)$ (Update parameters)
  end while
  return θ_t (Resulting parameters)

running examples







Russel, Norvig. Artificial Intelligence – A Modern Approach. Third Edition. 2016.
https://inst.eecs.berkeley.edu/~cs188/fa22/
chatgpt.com

encoding policies...

Differentiable Programming

Jax (Python)

https://github.com/google/jax

Zygote (Julia)

```
julia> using Zygote

julia> f(x) = 5x + 3

julia> f(10), f'(10)
(53, 5.0)
```

https://github.com/FluxML/Zygote.jl

Neural Networks

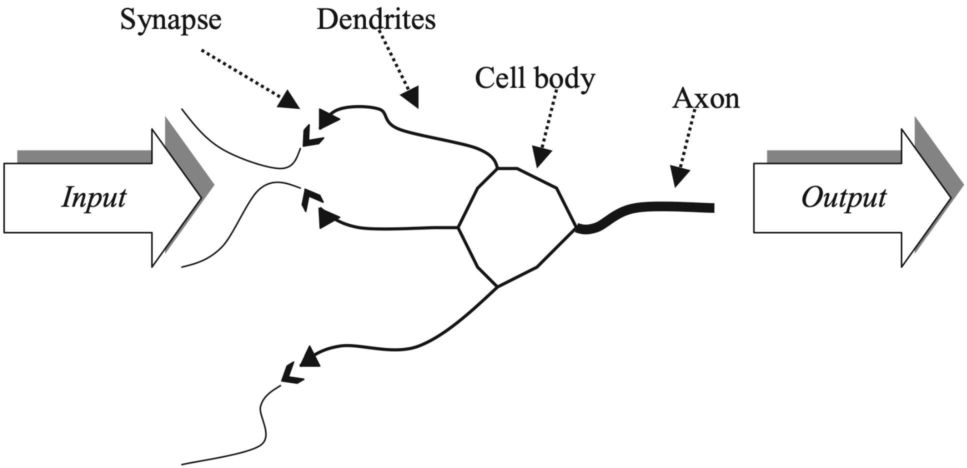
PyTorch

https://pytorch.org

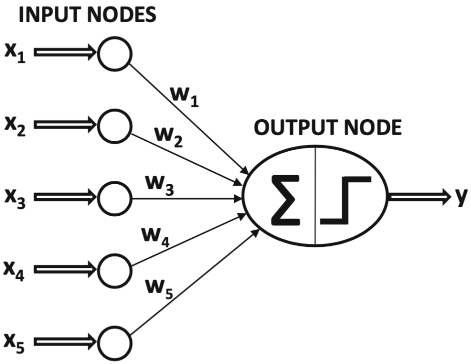
TensorFlow

https://www.tensorflow.org

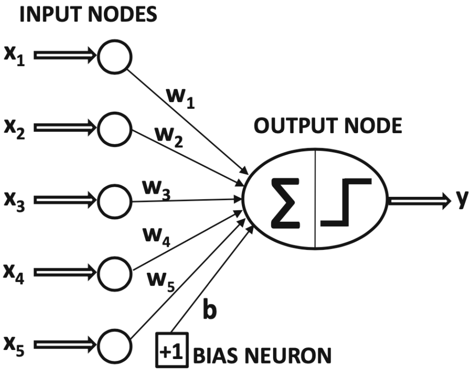
Neural Networks



source: Brabazon, O'Neill, McGarraghy. Natural Computing Algorithms. 2015.

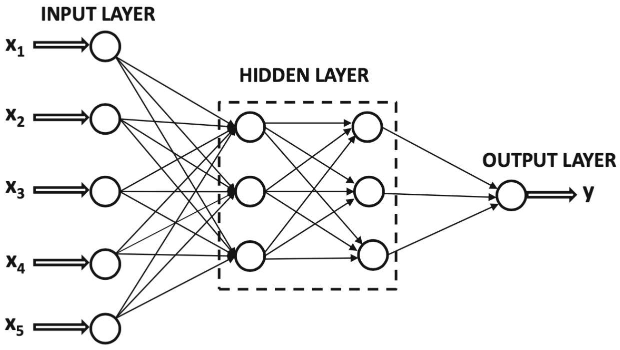


(a) Perceptron without bias

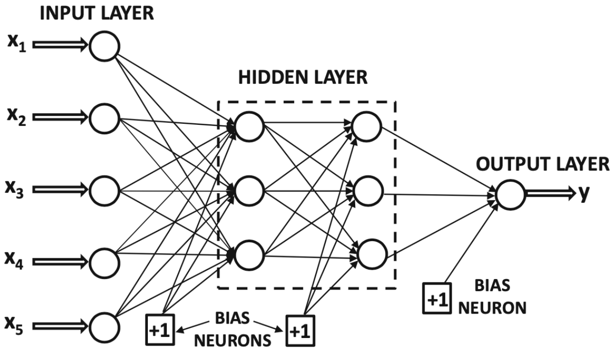


(b) Perceptron with bias

source: Aggarwal. Neural Networks and Deep Learning. 2

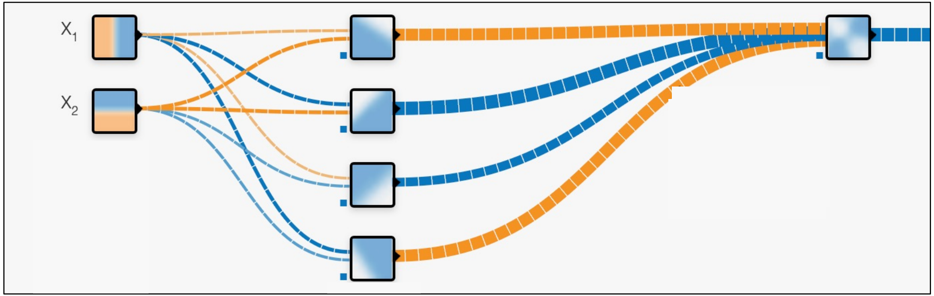


(a) No bias neurons



(b) With bias neurons

Let's try!



https://playground.tensorflow.org

Definition 8 (neural network)**.** A neural network (NN) is a function $\mathcal{N} : \mathbb{R}^p \to \mathbb{R}^q$ with p inputs and q outputs. This function is defined via a graph made up of r layers $L_1, ..., L_r$ where each layer L_l consists of $|L_l|$ cells $C_{l,1}, ..., C_{l,|L_l|}$, which make up the graph's vertices, and each cell $C_{l,c}$ of the layer L_l is connected to all cells of the previous layer, i.e., $C_{l-1,d}$ for $d = 1, ..., |L_{l-1}|$, via the graph's edges. Each edge of a cell $C_{l,c}$ is assigned an edge weight $E_{l,c,e} \in \mathbb{R}, e = 1, ..., |L_{l-1}|$. Given a fixed graph structure and activation function $f : \mathbb{R} \to \mathbb{R}$, the vector of all edge weights

$$\mathbf{w} = \langle E_{l,c,e} \rangle_{\, l=1,...,r,\;\; c=1,...,|L_l|,\;\; e=1,...,|L_{l-1}|}$$

and the vector of all cell biases

$$\mathbf{b} = \langle B_{l,c} \rangle_{\, l=1,...,r,\;\; c=1,...,|L_l|}$$

with $B_{l,c} \in \mathbb{R}$ define the network's functionality. The combined vector $\overline{\mathcal{N}} = \mathbf{w} +\!\!\!+ \mathbf{b}$ is called the network $\mathcal{N}$'s parameters.

Definition 8 (neural network)**.** A neural network (NN) is a function $\mathcal{N} : \mathbb{R}^p \to \mathbb{R}^q$ with p inputs and q outputs. This function is defined via a graph made up of r layers $L_1, ..., L_r$ where each layer L_l consists of $|L_l|$ cells $C_{l,1}, ..., C_{l,|L_l|}$, which make up the graph's vertices, and each cell $C_{l,c}$ of the layer L_l is connected to all cells of the previous layer, i.e., $C_{l-1,d}$ for $d = 1, ..., |L_{l-1}|$, via the graph's edges. Each edge of a cell $C_{l,c}$ is assigned an edge weight $E_{l,c,e} \in \mathbb{R}, e = 1, ..., |L_{l-1}|$. Given a fixed graph structure and activation function $f : \mathbb{R} \to \mathbb{R}$, the vector of all edge weights

$$\mathbf{w} = \langle E_{l,c,e} \rangle_{\, l=1,\ldots,r,\;\; c=1,\ldots,|L_l|,\;\; e=1,\ldots,|L_{l-1}|}$$

and the vector of all cell biases

$$\mathbf{b} = \langle B_{l,c} \rangle_{\, l=1,\ldots,r,\;\; c=1,\ldots,|L_l|}$$

with $B_{l,c} \in \mathbb{R}$ define the network's functionality. The combined vector $\overline{\mathcal{N}} = \mathbf{w} \mathbin{+\!\!\!+} \mathbf{b}$ is called the network $\mathcal{N}$'s parameters.
A network's output given an input $\mathbf{x} \in \mathbb{R}^p$ is given via

$$\mathbf{y} = \mathcal{N}(\mathbf{x}) = \langle O(r, c) \rangle_{c=1,\ldots,|L_r|} \in \mathbb{R}^q$$

$$\text{where } O(l, c) = \begin{cases} x_c & \text{if } l = 0, \\ f\big(B_{l,c} + \sum_{i=1}^{|L_{l-1}|} E_{l,c,i} \cdot O(l-1, i)\big) & \text{otherwise.} \end{cases}$$