Calculus
Gradient Descent

Arno Kimeswenger

Intitute of Mathematics and Statistics
University of Applied Sciences Wiener Neustadt

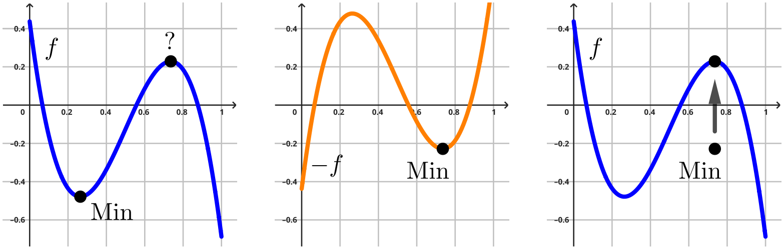
April 28, 2026

Gradient Descent

- **Gradient Descent:**
  Numerical method to find minimum of a function $f(x, y)$ (or more general $f(\mathbf{x})$ with $\mathbf{x} = (x_1, \ldots, x_n)$)

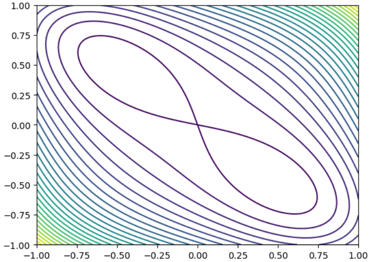
Gradient Descent

- Gradient Descent:
  Numerical method to find minimum of a function $f(x, y)$ (or more general $f(\mathbf{x})$ with $\mathbf{x} = (x_1, \ldots, x_n)$)
- **If we have a method to find the minimum**
  $\Rightarrow$ Method for maximum: Use $-f$ and compute minimum to get maximum of f



Gradient Descent

- **Example.** Find Minimum of $f(x,y) = x^4 + y^4 + \frac{1}{2}\cdot\left(x^2+y^2\right) + 2xy$
  - Gradient: $\nabla f = \begin{pmatrix} 4x^3+x+2y \\ 4y^3+y+2x \end{pmatrix}$
  - Always plot if possible
  - Method 1: $\nabla f = 0$
    We get a solution $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\begin{pmatrix} 1/2 \\ -1/2 \end{pmatrix}$ and $\begin{pmatrix} -1/2 \\ 1/2 \end{pmatrix}$
  - Often solving the nonlinear system of equations is too complicated $\Rightarrow$ Newton's method, but we need second derivatives $\Rightarrow$ inapplicable
  - Methode 2: Gradient Descent



Gradient Descent

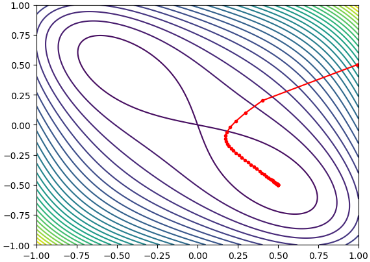
- Gradient Descent
  - ∇f points in direction of steepest ascent
  - $\Rightarrow -\nabla f$ points in direction of steepest descent
  - Idea: Go in direction of steepest descent

    old approximation $\mathrm{v}_k = \begin{pmatrix} x_k \\ y_k \end{pmatrix}$

    new approximation $\mathrm{v}_{k+1} = \begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix}$

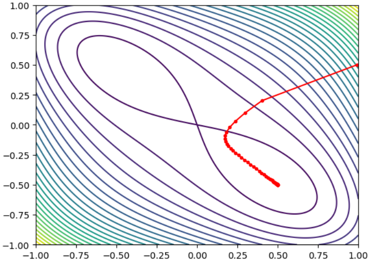
    step size η_k

$$\mathrm{v}_{k+1} = \mathrm{v}_k - \eta_k \cdot \nabla f(\mathrm{v}_k)$$



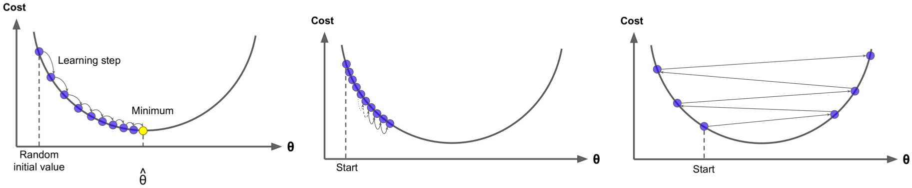
Gradient Descent

- How to define the step size?
  - In our example we used $\eta_k = 0.1$, constant
  - It would be possible to use optimal step size $\Rightarrow$ 1d problem $\Rightarrow$ to complicated
  - Sometimes use $\eta_k = \frac{\eta_0}{1+k}$ or $\eta_k = \eta_0 \cdot 0.5^k$
  - Armijo-Condition:
    In step k of gradient descent:
    Choose start value for η, e.g. $\eta = 1$,
    $\sigma \in (0,1)$ and $\varrho \in (0,1)$
    while $f(x_{old} - \eta \cdot \nabla f(x_{old})) > f(x_{old}) - \sigma \cdot \eta \cdot |\nabla f(x_{old})|^2$
    $\eta = \varrho \cdot \eta$
    end
    Choose $\eta_k = \eta$



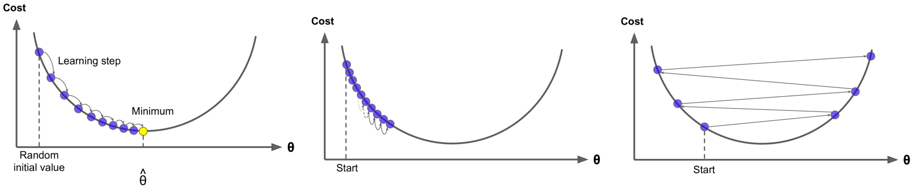
Gradient Descent

- **Step size is critical, from left to right: okay, too small, too large**

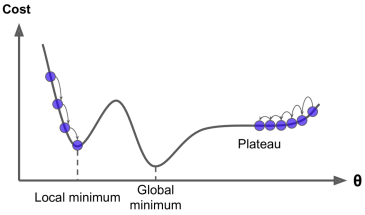


Gradient Descent

- Step size is critical, from left to right: okay, too small, too large

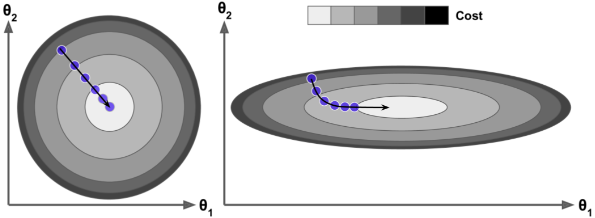


- **Gradient Descent uses local information only**
  $\Rightarrow$ It is possible that solution converges to local minimum or saddle point



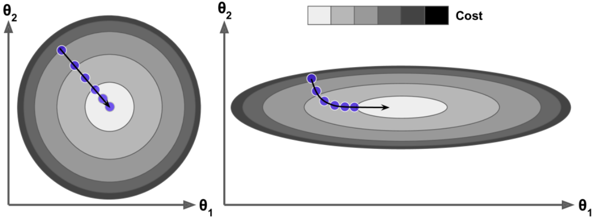
Gradient Descent

- **Gradient Descent is influenced by scaling**



Gradient Descent

- Gradient Descent is influenced by scaling



- **Gradient Descent is basis for optimization of parameters of neural networks (backtracking algorithm)!**
- **Often modified variants of Gradient Descent are used**