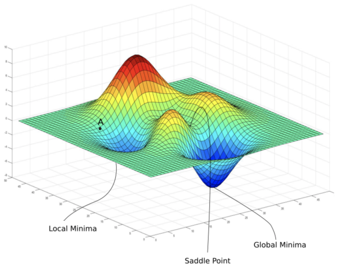
Optimization in Machine Learning

Mathematical Concepts
Conditions for optimality

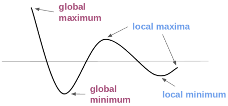


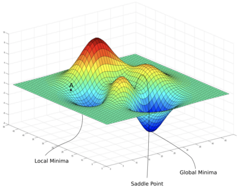
Learning goals

- Local and global optima
- First & second order conditions

EXTREMA AND SADDLE POINTS

- Given $\mathcal{S} \subseteq \mathbb{R}^d$, $f : \mathcal{S} \to \mathbb{R}$
- Global minimum at $\boldsymbol{x}^*$: $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{S}$
- Local minimum at $\boldsymbol{x}^*$: $\exists\, \epsilon > 0$ s.t. $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in S \cap B_\epsilon(\boldsymbol{x}^*)$ (ϵ-ball)
- Analogously for global and local max
- We call $\boldsymbol{x}^*$ saddle point if in feasible portion of every eps-ball $S \cap B_\epsilon(\boldsymbol{x}^*)$, is at least a strictly better and a strictly worse point





Source (left): `https://en.wikipedia.org/wiki/Maxima_and_minima`

Source (right): `https://wngaw.github.io/linear-regression/`

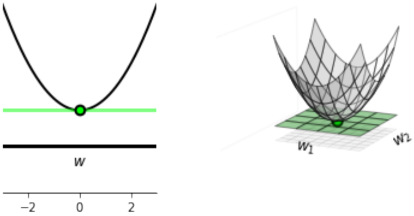
EXISTENCE OF OPTIMA

- $f : \mathcal{S} \to \mathbb{R}$
- If f continuous and $\mathcal{S}$ compact: minimum and maximum exist (extreme value theorem)
- If f discontinuous: no general existence statement
- Negative example, with $\mathcal{S} = [0, 1]$:

$$f(x) = \begin{cases} 1/x & x > 0 \\ 0 & x = 0 \\ -1/x & x < 0 \end{cases}$$

FIRST-ORDER CONDITION

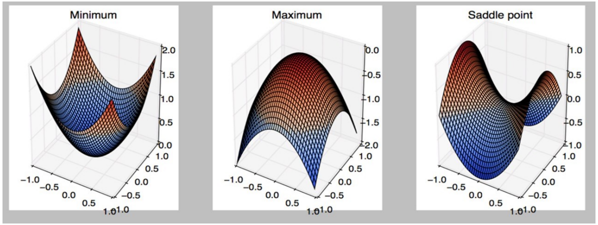
- Let $f : \mathcal{S} \to \mathbb{R}$, f differentiable, $\boldsymbol{x}^*$ interior point of $\mathcal{S}$
- Necessary condition:
  If $\boldsymbol{x}^*$ is a local extremum, then $\nabla f(\boldsymbol{x}^*) = 0$
- Such points are called ‘stationary’
- Intuition: at a local extremum, the function must be flat, otherwise we can find a direction to move to a better value
- Not sufficient, e.g. saddle points are possible



Source: Watt (2020)

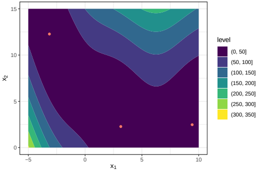
SECOND-ORDER CONDITION

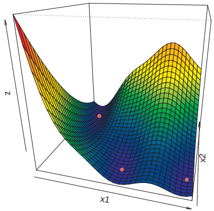
- Let $f : \mathcal{S} \to \mathbb{R}$, $f \in \mathcal{C}^2$, $\boldsymbol{x}^*$ interior point of $\mathcal{S}$
- If $H(\boldsymbol{x}^*)$ is definite, then $\boldsymbol{x}^*$ is a strict local extremum
- If $H(\boldsymbol{x}^*)$ is semi-definite, then $\boldsymbol{x}^*$ is a local extremum
- If $H(\boldsymbol{x}^*)$ is indefinite, then $\boldsymbol{x}^*$ is a saddle point
- If $H(\boldsymbol{x}^*)$ is p(s)d, then $\boldsymbol{x}^*$ is a (strict) local min
  this implies f is locally (strictly) convex
- If $H(\boldsymbol{x}^*)$ is n(s)d, then $\boldsymbol{x}^*$ is a (strict) local max
  this implies f is locally (strictly) concave
- Interpretation: curvature pos (or neg) in all directions



EXAMPLE: BRANIN FUNCTION

- Branin function with 3 local minima

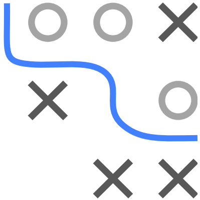




- EVs of Hessian at local minima:

	λ_1	λ_2
Left	22.29	0.96
Middle	11.07	1.73
Right	11.33	1.69

CONVEXITY AND OPTIMA



- $f : \mathcal{S} \to \mathbb{R}$ convex on convex set $\mathcal{S}$
- Any local minimum is global
- The set of minima is convex
- If f strictly convex: at most one local minimum (unique global on $\mathcal{S}$, if it exists)
- Analogously for concave functions

EXAMPLE

- $f(x, y) = x^4 + y^4 - x^2 - y^2$
- $\nabla f(x, y) = (4x^3 - 2x, 4y^3 - 2y)$
- $H_{(}x, y) = \begin{pmatrix} 12x^2 - 2 & 0 \\ 0 & 12y^2 - 2 \end{pmatrix}$
- At $(0, 0)^T$ we have strict local max
- At $(\pm\frac{1}{\sqrt{2}}, \pm\frac{1}{\sqrt{2}})^T$ we have 4 strict local min
- At $(0, \pm\frac{1}{\sqrt{2}})^T$, $(\pm\frac{1}{\sqrt{2}}, 0)^T$ we have 4 saddle points