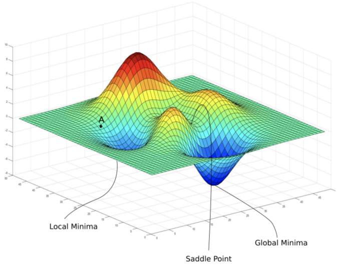
Optimization in Machine Learning

Mathematical Concepts:
Conditions for optimality



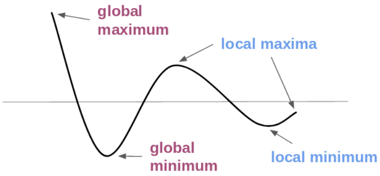
Learning goals

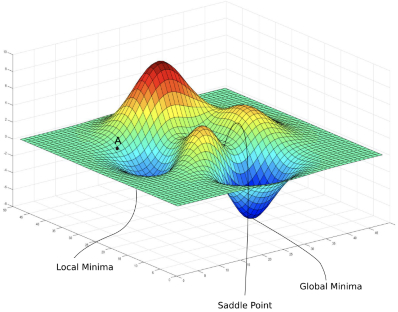
- Local and global
- First & second order conditions

DEFINITION LOCAL AND GLOBAL MINIMUM

Given $\mathcal{S} \subseteq \mathbb{R}^d$, $f : \mathcal{S} \to \mathbb{R}$:

- f has **global minimum** in $\mathbf{x}^* \in \mathcal{S}$, if $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{S}$
- f has a **local minimum** in $\mathbf{x}^*$, if $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\mathbf{x} \in B_\epsilon(\mathbf{x}^*)$, with $B_\epsilon(\mathbf{x}^*) := \{\mathbf{x} \in \mathcal{S} \mid \|\mathbf{x} - \mathbf{x}^*\| < \epsilon\}$ ("ϵ"-ball round $\mathbf{x}^*$).





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

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

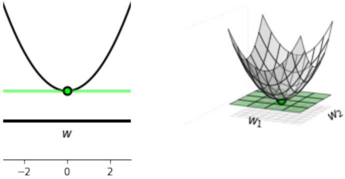
EXISTENCE OF OPTIMA

$$f : \mathcal{S} \to \mathbb{R}$$

- f continous:
  - A real-valued function f defined on a **compact set** must attain a minimum and a maximum (extreme value theorem).
- f not continuous:
  - In general no statement possible about existence of maximum/minimum.

FIRST ORDER CONDITION FOR OPTIMALITY

Let $f \in \mathcal{C}^1$. **Observation:** At a local minimum (for an interior point) 1st order Taylor series approx is perfectly flat; 1st order derivs are 0.



(Strictly) convex functions (left: univariate; right: multivariate) with unique local minimum, which is the global one. Tangent (hyperplane) is perfectly flat at the optimum.

Source: Watt, 2020, Machine Learning Refined.

FIRST ORDER CONDITION FOR OPTIMALITY

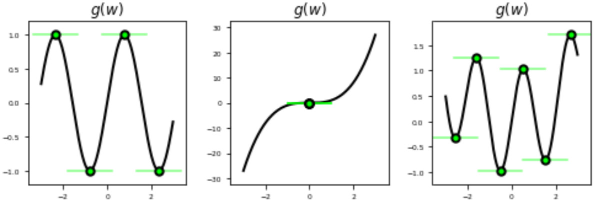
At every (interior) local minimum $\mathbf{x}^*$ the first derivative is necessarily always zero; it is therefore called **first-order** or **necessary** condition.

- **First-order condition (univariate):** Let $\mathbf{x}^* \in \mathbb{R}$ be a local minimum of f. Then:
$$f'(\mathbf{x}^*) = 0$$
- **First-order condition (multivariate):** Let $\mathbf{x}^* \in \mathbb{R}^d$ be a local minimum of f. Then:
$$\nabla f(\mathbf{x}^*) = (0, 0, ..., 0)^\top$$

The points at which the first order derivative is zero are called **stationary points**.

FIRST ORDER CONDITION FOR OPTIMALITY

The condition is **not sufficient**: Not every stationary point ($\nabla f(\mathbf{x}) = 0$) is a local minimum.



Left: Four points fulfill the necessary conditions; but two of the points are local maxima (not minima). Middle: One point fulfills the necessary condition, but is not a local optimum. Right: Multiple local minima and maxima.

Source: Watt, 2020, Machine Learning Refined.

SECOND ORDER CONDITION FOR OPTIMALITY

Let $f \in \mathcal{C}^2$. If the function is locally convex, so:

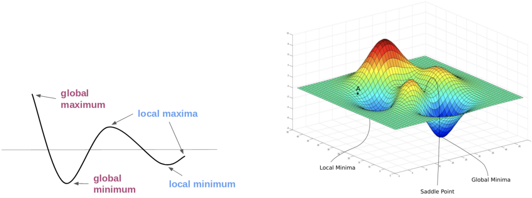
- **Second-order condition**: A **stationary** point $\mathbf{x}^* \in \mathcal{S} \subseteq \mathbb{R}$ fulfills

$$f''(x^*) > 0 \quad (d = 1)$$

$$\nabla^2 f(\mathbf{x}^*) \text{ is positive definite} \quad (d > 1)$$

(all EVs positive), hence curvature is positive in all directions.

Then the second-order condition is **sufficient** to prove a local minimum.

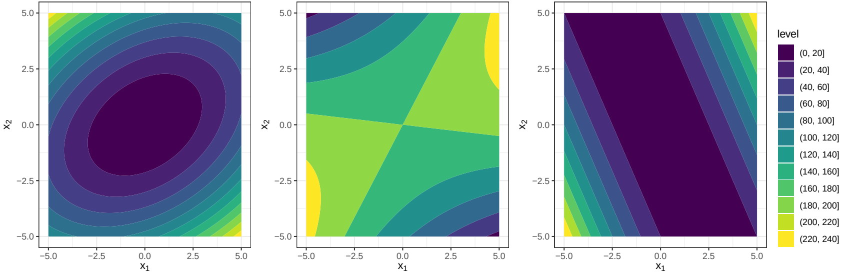


Two functions that are locally convex, but not globally convex.

CONDITIONS FOR OPTIMALITY AND CONVEXITY

Let $f : \mathcal{S} \to \mathbb{R}$ be convex on convex set $\mathcal{S}$. Then the following holds:

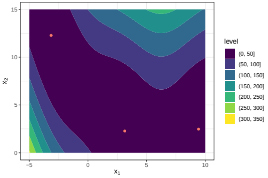
- Any local minimum is also global minimum
- If f strictly convex, f has exactly one local minimum which is also unique global minimum on $\mathcal{S}$

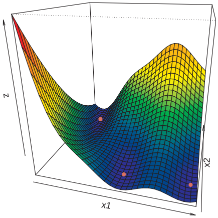


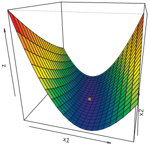
Three different quadratic forms. Left: Hessian has a two positive Eigenvalues ($\lambda_1 = 2$, $\lambda_2 = 5$). Middle: Hessian has positive and negative Eigenvalue ($\lambda_1 = 2$, $\lambda_2 = -5$). Right: Hessian has positive and a zero Eigenvalue ($\lambda_1 = 2$, $\lambda_2 = 0$).

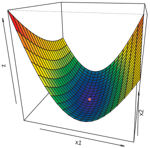
CONDITIONS FOR OPTIMALITY AND CONVEXITY

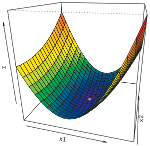
Example: Branin function











Gradient and Hessian have been computed numerically at the minima / red points (R package numDeriv). Gradients are 0, function is locally convex. Eigenspectra:

$\lambda_1 = 22.29, \lambda_2 = 0.96$ (Opt. 1)

$\lambda_1 = 11.07, \lambda_2 = 1.73$ (Opt. 2)

$\lambda_1 = 11.33, \lambda_2 = 1.69$ (Opt. 3)

CONDITIONS FOR OPTIMALITY AND CONVEXITY

Def.: **Saddle point**

- Gradient of 0
- If H is indefinite at stationary point, so pos and neg Eigenvalues occur in H, we have a saddle point
- The latter is only a sufficient condition, but not a necessary one