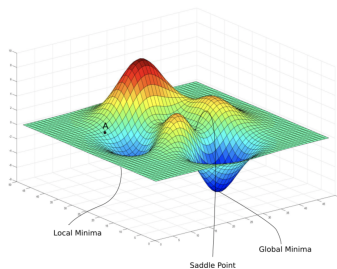


# Optimization in Machine Learning

## Mathematical Concepts: Conditions for optimality



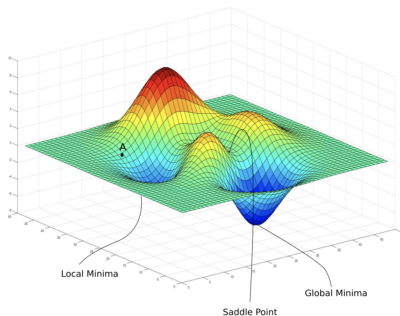
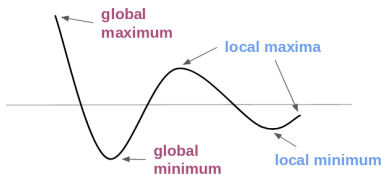
### Learning goals

- Local and global optima
- First & second order conditions

# DEFINITION LOCAL AND GLOBAL MINIMUM

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

- $f$  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  $\mathbf{x}^* \in \mathcal{S}$ , if  $\epsilon > 0$  exists s.t.  $f(\mathbf{x}^*) \leq f(\mathbf{x})$  for all  $\mathbf{x} \in B_\epsilon(\mathbf{x}^*)$  (“ $\epsilon$ ”-ball around  $\mathbf{x}^*$ ).



Source (left): [https://en.wikipedia.org/wiki/Maxima\\_and\\_minima](https://en.wikipedia.org/wiki/Maxima_and_minima).

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

# EXISTENCE OF OPTIMA

We regard the two main cases of  $f : \mathcal{S} \rightarrow \mathbb{R}$ :

- **$f$  continuous:** If  $\mathcal{S}$  is **compact**,  $f$  attains a minimum and a maximum (extreme value theorem).
- **$f$  discontinuous:** **No general** statement possible about existence of optima.

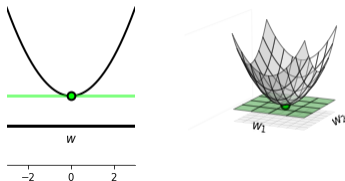
**Example:**  $\mathcal{S} = [0, 1]$  compact,  $f$  discontinuous with

$$f(x) = \begin{cases} 1/x & \text{if } x > 0, \\ 0 & \text{if } x = 0. \end{cases}$$

# FIRST ORDER CONDITION FOR OPTIMALITY

**Observation:** At an interior local optimum of  $f \in \mathcal{C}^1$ , first order Taylor approximation is flat, i.e., first order derivatives are zero.

This condition is therefore **necessary** and called **first order**.



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, *Machine Learning Refined*, 2020)

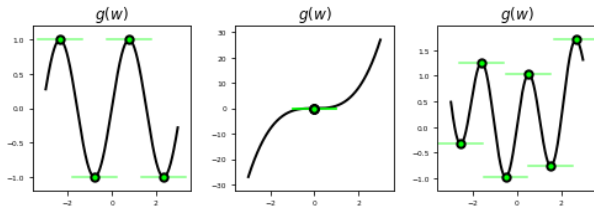
# FIRST ORDER CONDITION FOR OPTIMALITY

**First order condition:** Gradient of  $f$  at local optimum  $\mathbf{x}^* \in \mathcal{S}$  is zero:

$$\nabla f(\mathbf{x}^*) = (0, \dots, 0)^T$$

Points with zero first order derivative are called **stationary**.

Condition is **not sufficient**: Not all stationary points are local optima.



**Left:** Four points fulfill the necessary condition and are indeed optima.

**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

**Second order condition:** Hessian of  $f \in \mathcal{C}^2$  at stationary point  $\mathbf{x}^* \in \mathcal{S}$  is positive or negative definite:

$$H(\mathbf{x}^*) \succ 0 \text{ or } H(\mathbf{x}^*) \prec 0$$

**Interpretation:** Curvature of  $f$  at local optimum is either positive in all directions or negative in all directions.

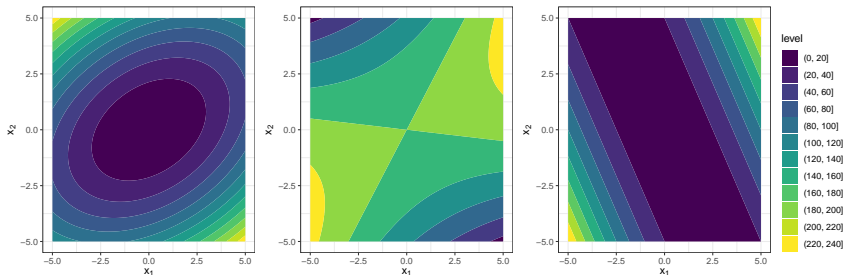
The second order condition is **sufficient** for a stationary point.

**Proof:** Later.

# CONDITIONS FOR OPTIMALITY AND CONVEXITY

Let  $f : \mathcal{S} \rightarrow \mathbb{R}$  be **convex**. Then:

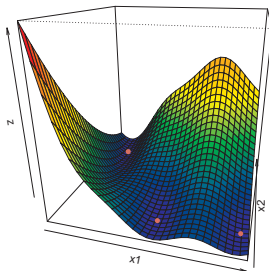
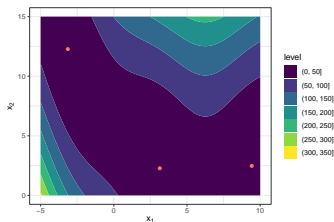
- Any local minimum is **also global** minimum
- If  $f$  **strictly convex**,  $f$  has **exactly one** local minimum which is also unique global minimum on  $\mathcal{S}$



Three quadratic forms. **Left:**  $H(\mathbf{x}^*)$  has two positive eigenvalues. **Middle:**  $H(\mathbf{x}^*)$  has positive and negative eigenvalue. **Right:**  $H(\mathbf{x}^*)$  has positive and a zero eigenvalue.

# CONDITIONS FOR OPTIMALITY AND CONVEXITY

## Example: Branin function



Spectra of Hessians (numerically computed):

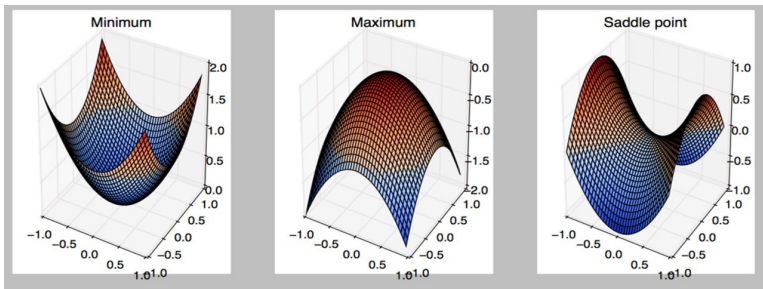
	$\lambda_1$	$\lambda_2$
Left	22.29	0.96
Middle	11.07	1.73
Right	11.33	1.69



# CONDITIONS FOR OPTIMALITY AND CONVEXITY

Definition: **Saddle point** at  $\mathbf{x}$

- $\mathbf{x}$  stationary (necessary)
- $H(\mathbf{x})$  indefinite, i.e., positive and negative eigenvalues (sufficient)



# CONDITIONS FOR OPTIMALITY AND CONVEXITY

## Examples:

- $f(x, y) = x^2 - y^2, \nabla f(x, y) = (2x, -2y)^T,$

$$H_f(x, y) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$$

$\implies$  Saddle point at  $(0, 0)$  (sufficient condition met)

- $g(x, y) = x^4 - y^4, \nabla g(x, y) = (4x^3, -4y^3)^T,$

$$H_g(x, y) = \begin{pmatrix} 12x^2 & 0 \\ 0 & -12y^2 \end{pmatrix}$$

$\implies$  Saddle point at  $(0, 0)$  (sufficient condition **not** met)