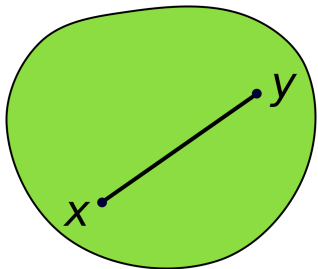


# Optimization in Machine Learning

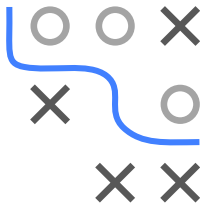
## Mathematical Concepts

### Convexity



#### Learning goals

- Convex sets
- Convex functions



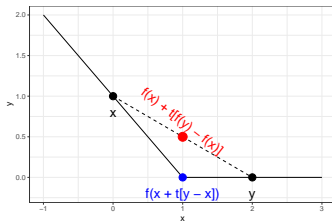
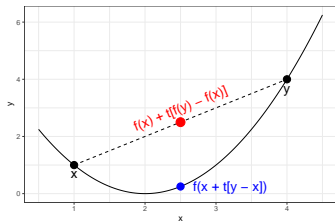


# CONVEX FUNCTIONS

Let  $f : \mathcal{S} \rightarrow \mathbb{R}$ ,  $\mathcal{S}$  convex.  $f$  is **convex** if for all  $\mathbf{x}, \mathbf{y} \in \mathcal{S}$  and all  $t \in [0, 1]$

$$f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) \leq f(\mathbf{x}) + t(f(\mathbf{y}) - f(\mathbf{x})).$$

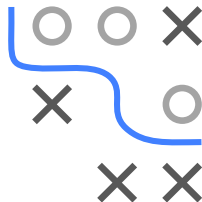
Intuitively: Connecting line lies above function.



**Left:** Strictly convex function. **Right:** Convex, but not strictly.

**Strictly convex** if “ $<$ ” instead of “ $\leq$ ”. **Concave** (strictly) if the inequality holds with “ $\geq$ ” (“ $>$ ”), respectively.

**Note:**  $f$  (strictly) concave  $\Leftrightarrow -f$  (strictly) convex.



# EXAMPLES

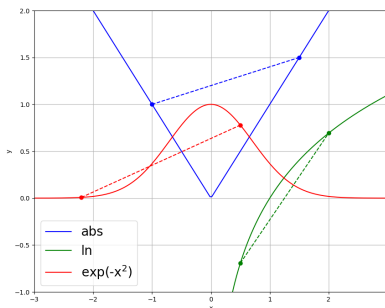
**Convex function:**  $f(x) = |x|$

**Proof:**

$$\begin{aligned}f(x + t(y - x)) &= |x + t(y - x)| = |(1 - t)x + t \cdot y| \\&\leq |(1 - t)x| + |t \cdot y| = (1 - t)|x| + t|y| \\&= |x| + t \cdot (|y| - |x|) = f(x) + t \cdot (f(y) - f(x))\end{aligned}$$

**Concave function:**  $f(x) = \log(x)$

**Neither nor:**  $f(x) = \exp(-x^2)$  (but log-concave)

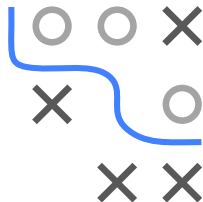


# OPERATIONS PRESERVING CONVEXITY

- **Nonnegatively weighted summation:** Weights  $w_1, \dots, w_n \geq 0$ , convex functions  $f_1, \dots, f_n$ :  $w_1 f_1 + \dots + w_n f_n$  also convex  
In particular: Sum of convex functions also convex
- **Composition:**  $g$  convex,  $f$  linear:  $h = g \circ f$  also convex  
**Proof:**

$$\begin{aligned}h(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) &= g(f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))) \\&= g(f(\mathbf{x}) + t(f(\mathbf{y}) - f(\mathbf{x}))) \\&\leq g(f(\mathbf{x})) + t(g(f(\mathbf{y})) - g(f(\mathbf{x}))) \\&= h(\mathbf{x}) + t(h(\mathbf{y}) - h(\mathbf{x}))\end{aligned}$$

- **Elementwise maximization:**  $f_1, \dots, f_n$  convex functions:  
 $g(\mathbf{x}) = \max \{f_1(\mathbf{x}), \dots, f_n(\mathbf{x})\}$  also convex



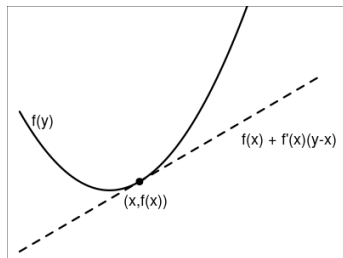
# FIRST ORDER CONDITION

Prove convexity via **gradient**:  
Let  $f$  be differentiable.

$f$  (strictly) convex



$$f(\mathbf{y}) \stackrel{(>)}{\geq} f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x}) \text{ for all } \mathbf{x}, \mathbf{y} \in \mathcal{S} \text{ (s.t. } \mathbf{x} \neq \mathbf{y})$$



## SECOND ORDER CONDITION

Matrix  $A$  is **positive (semi)definite** (p.(s.)d.) if  $\mathbf{v}^T A \mathbf{v} \stackrel{(\geq)}{>} 0$  for all  $\mathbf{v} \neq 0$ .

**Notation:**  $A \stackrel{(\succ)}{\succ} 0$  for  $A$  p.(s.)d. and  $B \stackrel{(\succ)}{\succ} A$  if  $B - A \stackrel{(\succ)}{\succ} 0$

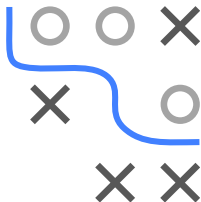
Prove convexity via **Hessian**:

Let  $f \in \mathcal{C}^2$  and  $H(\mathbf{x})$  be its Hessian.

$$f \text{ (strictly) convex} \iff H(\mathbf{x}) \stackrel{(\succ)}{\succ} 0 \text{ for all } \mathbf{x} \in \mathcal{S}$$

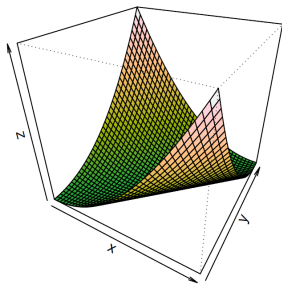
**Alternatively:** Since  $H(\mathbf{x})$  symmetric for  $f \in \mathcal{C}^2$ :

$$H(\mathbf{x}) \stackrel{(\succ)}{\succ} 0 \iff \text{all eigenvalues of } H(\mathbf{x}) \geq 0$$



## SECOND ORDER CONDITION / 2

**Example:**  $f(\mathbf{x}) = x_1^2 + x_2^2 - 2x_1x_2$ ,  $\nabla f(\mathbf{x}) = \begin{pmatrix} 2x_1 - 2x_2 \\ 2x_2 - 2x_1 \end{pmatrix}$ ,  $H(\mathbf{x}) = \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix}$ .



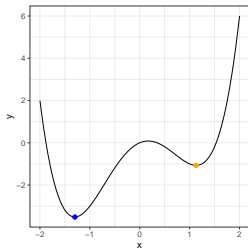
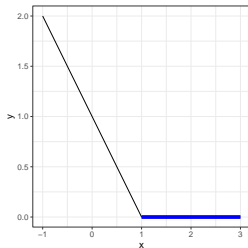
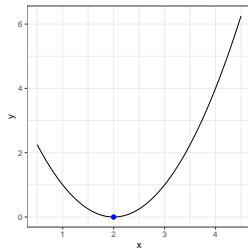
$f$  is convex since  $H(\mathbf{x})$  is p.s.d. for all  $\mathbf{x} \in \mathcal{S}$ :

$$\begin{aligned} \mathbf{v}^T \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix} \mathbf{v} &= \mathbf{v}^T \begin{pmatrix} 2v_1 - 2v_2 \\ -2v_1 + 2v_2 \end{pmatrix} = 2v_1^2 - 2v_1v_2 - 2v_1v_2 + 2v_2^2 \\ &= 2v_1^2 - 4v_1v_2 + 2v_2^2 = 2(v_1 - v_2)^2 \geq 0. \end{aligned}$$



# CONVEX FUNCTIONS IN OPTIMIZATION

- For a convex function, every local optimum is also a global one  
⇒ No need for involved global optimizers, local ones are enough
- A strictly convex function has at most one optimal point
- Example for strictly convex function without optimum:  $\exp$  on  $\mathbb{R}$



**Left:** Strictly convex; exactly one local minimum, which is also global. **Middle:** Convex, but not strictly; all local optima are also global ones but not unique. **Right:** Not convex.

# CONVEX FUNCTIONS IN OPTIMIZATION / 2

“... in fact, the great watershed in optimization isn’t between linearity and nonlinearity, but convexity and nonconvexity.”

SIAM REVIEW  
Vol. 35, No. 2, pp. 183–238, June 1993

## LAGRANGE MULTIPLIERS AND OPTIMALITY\*

**Abstract.** Lagrange multipliers used to be viewed as auxiliary variables introduced in a problem of constrained minimization in order to write first-order optimality conditions formally as a system of equations. Modern applications, with their emphasis on numerical methods and more complicated side conditions than equations, have demanded deeper understanding of the concept and how it fits into a larger theoretical picture.

**Key words.** Lagrange multipliers, optimization, saddle points, dual problems, augmented Lagrangian, constraint qualifications, normal cones, subgradients, nonsmooth analysis

**AMS subject classifications.** 49K99, 58C20, 90C99, 49M29