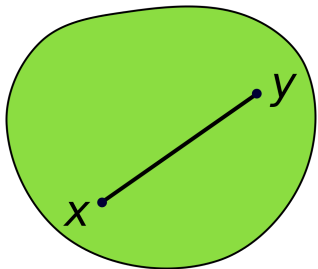# Optimization in Machine Learning
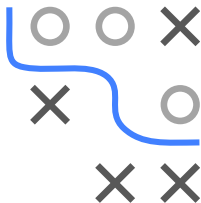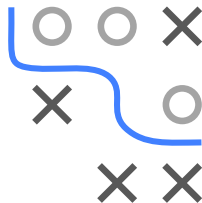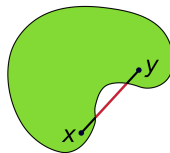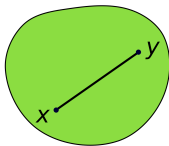
Mathematical Concepts
Convexity

Learning goals

- Convex sets
- Convex functions

CONVEX SETS

- Set $\mathcal{S} \subseteq \mathbb{R}^d$ is convex, if $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$ and $\forall t \in [0, 1]$:
  $\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}) \in \mathcal{S}$
- Intuition: Line between any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$ lies entirely in $\mathcal{S}$
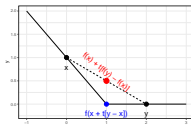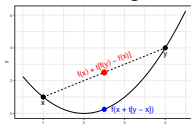- Left: convex set; Right: not convex. (Source: Wikipedia)

CONVEX FUNCTIONS

- Let $f : \mathcal{S} \to \mathbb{R}$, $\mathcal{S}$ convex
- f is convex if $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$ and $\forall t \in [0, 1]$:

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

- Intuition: Connecting line for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$ above function f

  Left: Strictly convex function; Right: Convex, but not strictly
- Strictly convex if $<$ instead of $\leq$
- Concave (strictly) if inequality holds with $\geq$ ($>$)
- NB: f (strictly) concave $\Leftrightarrow -f$ (strictly) convex
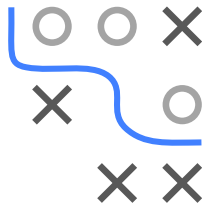
SOME EXAMPLES

- Convex: $f(x) = |x|$

$$f(x + t(y - x)) = |x + t(y - x)| = |(1 - t)x + t \cdot y|$$
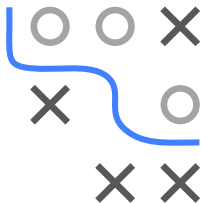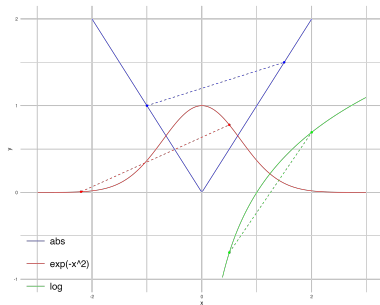$$\leq |(1 - t)x| + |t \cdot y| = (1 - t)|x| + t|y|$$
$$= |x| + t \cdot (|y| - |x|) = f(x) + t \cdot (f(y) - f(x))$$

- Concave: $f(x) = \log(x)$
- Neither: $f(x) = \exp(-x^2)$ (but log-concave)

——	abs
——	exp(-x^2)
——	log

OPERATIONS PRESERVING CONVEXITY

- Nonnegatively weighted summation:
  For $w_1, \ldots, w_n \geq 0$ and convex $f_1, \ldots, f_n$:
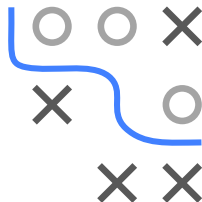 $w_1 f_1 + \cdots + w_n f_n$ is convex
  So: Sum of convex functions is also convex

- Composition: g convex, f linear: $h = g \circ f$ is also convex:

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

- Element-wise maximization: $f_1, \ldots, f_n$ convex functions:
  $g(\boldsymbol{x}) = \max \{f_1(\boldsymbol{x}), \ldots, f_n(\boldsymbol{x})\}$ is also convex

FIRST ORDER CONDITION

- For differentiable f, useful characterisation via gradient
- f convex

  $\Longleftrightarrow$

  $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x})$ for all $\boldsymbol{x} \neq \boldsymbol{y} \in \mathcal{S}$
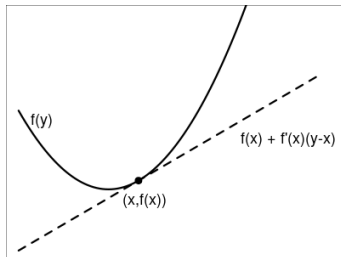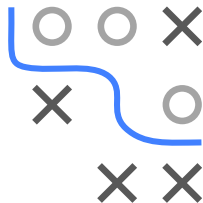

- Strictly convex if $>$ instead of $\geq$
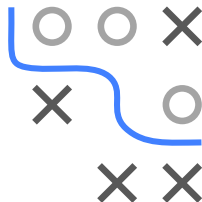
SECOND ORDER CONDITION

- For $f \in \mathcal{C}^2$: can characterize convexity via Hessian
- f convex $\iff H(\boldsymbol{x})$ psd for $\boldsymbol{x} \in \mathcal{S}$ $\forall \boldsymbol{x} \in \mathcal{S}$
- f strictly convex if $H(\boldsymbol{x})$ pd $\forall \boldsymbol{x} \in \mathcal{S}$

- To check global convexity, either verify the direct definition of psd by showing that $\boldsymbol{v}^T H(\boldsymbol{x}) \boldsymbol{v} \geq 0$ for all $\boldsymbol{v} \in \mathbb{R}^d$ and all $\boldsymbol{x}$, or, equivalently, check that all eigenvalues λ_i of all $H(\boldsymbol{x})$ satisfy $\lambda_i \geq 0$ for all $\boldsymbol{x} \in \mathcal{S}$

SECOND ORDER CONDITION

- Example:

$$f(\boldsymbol{x}) = x_1^2 + x_2^2 - 2x_1 x_2; \quad \nabla f(\boldsymbol{x}) = \begin{pmatrix} 2x_1 - 2x_2 \\ 2x_2 - 2x_1 \end{pmatrix}^T; \quad H(\boldsymbol{x}) = \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix}$$
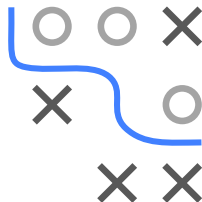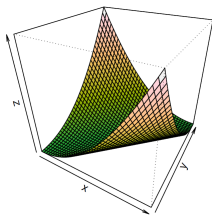
- f is convex since $H(\boldsymbol{x})$ is p.s.d. for all $\boldsymbol{x} \in \mathcal{S}$:

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

CONVEX FUNCTIONS IN OPTIMIZATION

- Will see later:
- For a convex function, every local optimum is also a global one
  $\Rightarrow$ No need for involved global optimizers, local ones are enough
- A strictly convex function has at most one optimal point
- "... in fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity."
  – R. Tyrrell Rockafellar. *SIAM Review*, 1993
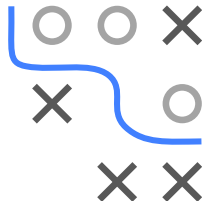
LAGRANGE MULTIPLIERS AND OPTIMALITY*

R. TYRRELL ROCKAFELLAR†

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.

A major line of research has been the nonsmooth geometry of one-sided tangent and normal vectors to the set of points satisfying the given constraints. Another has been the game-theoretic role of multiplier vectors as solutions to a dual problem. Interpretations as generalized derivatives of the optimal value with respect to problem parameters have also been explored. Lagrange multipliers are now being seen as arising from a general rule for the subdifferentiation of a nonsmooth objective function which allows black-and-white constraints to be replaced by penalty expressions. This paper traces such themes in the current theory of Lagrange multipliers, providing along the way a free-standing exposition of basic nonsmooth analysis as motivated by and applied to this subject.

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