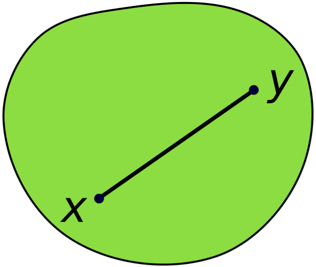
Optimization in Machine Learning

Mathematical Concepts
Convexity



Learning goals

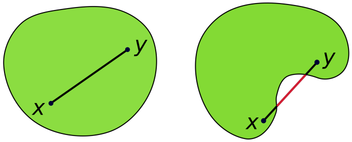
- Convex sets
- Convex functions

CONVEX SETS

A set of $\mathcal{S} \subseteq \mathbb{R}^d$ is **convex**, if for all $\mathbf{x}, \mathbf{y} \in \mathcal{S}$ and all $t \in [0, 1]$ the following holds:

$$\mathbf{x} + t(\mathbf{y} - \mathbf{x}) \in \mathcal{S}$$

Intuitively: Connecting line between any $\mathbf{x}, \boldsymbol{y} \in \mathcal{S}$ lies completely in $\mathcal{S}$.



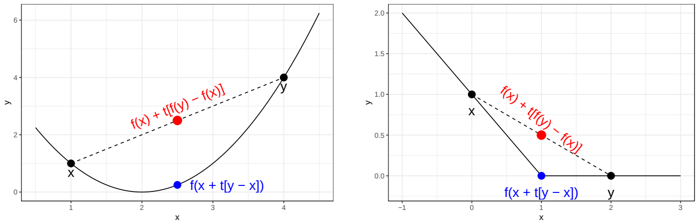
Left: convex set. **Right:** not convex. (Source: Wikipedia)

CONVEX FUNCTIONS

Let $f : \mathcal{S} \to \mathbb{R}$, $\mathcal{S}$ convex. f is **convex** if 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

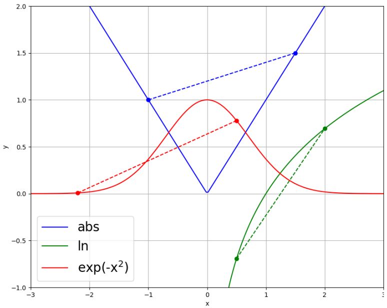
Convex function: $f(x) = |x|$

Proof:

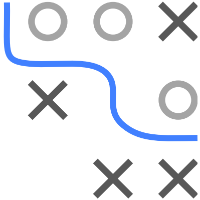
$$\begin{aligned} f\left(x + t(y-x)\right) &= |x + t(y-x)| = |(1-t)x + t\cdot y| \\ &\leq |(1-t)x| + |t\cdot y| = (1-t)|x| + t|y| \\ &= |x| + t\cdot(|y| - |x|) = f(x) + t\cdot(f(y) - f(x)) \end{aligned}$$

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

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



OPERATIONS PRESERVING CONVEXITY



- **Nonnegatively weighted summation:** Weights $w_1, \ldots, w_n \geq 0$, convex functions $f_1, \ldots, f_n$: $w_1 f_1 + \cdots + w_n f_n$ 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:** $f_1, \ldots, f_n$ convex functions:
  $g(\mathbf{x}) = \max\{f_1(\mathbf{x}), \ldots, f_n(\mathbf{x})\}$ also convex

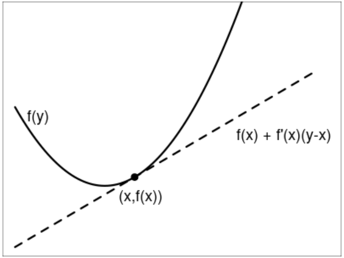
FIRST ORDER CONDITION

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

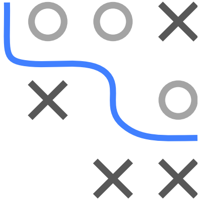
$$f \text{ (strictly) convex}$$

$$\iff$$

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



SECOND ORDER CONDITION



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

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

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

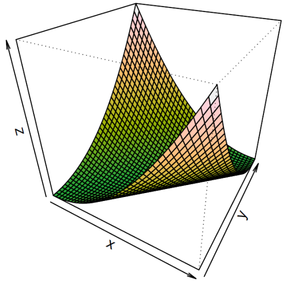
$$f \text{ (strictly) convex} \iff H(\mathbf{x}) \overset{(\succ)}{\succcurlyeq} 0 \text{ for all } \mathbf{x} \in \mathcal{S}$$

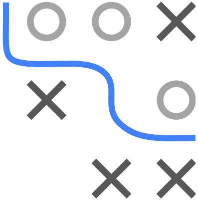
Alternatively: Since $H(\mathbf{x})$ symmetric for $f \in \mathcal{C}^2$:

$$H(\mathbf{x}) \succcurlyeq 0 \Leftrightarrow \text{all eigenvalues of } H(\mathbf{x}) \geq 0$$

SECOND ORDER CONDITION / 2

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

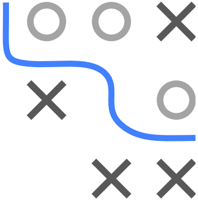




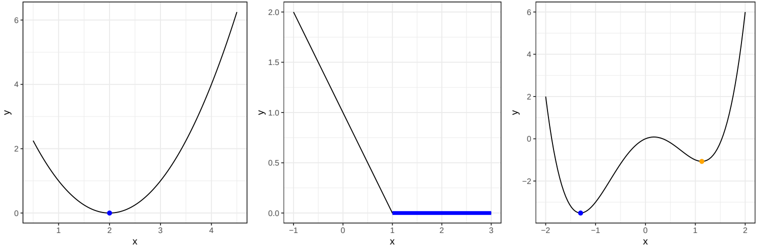
f is convex since $H(\mathbf{x})$ is p.s.d. for all $\mathbf{x} \in \mathcal{S}$:

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

CONVEX FUNCTIONS IN OPTIMIZATION



- 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
- Example for strictly convex function without optimum: $\exp$ on $\mathbb{R}$

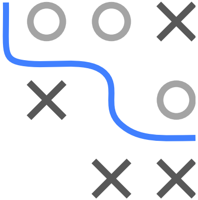


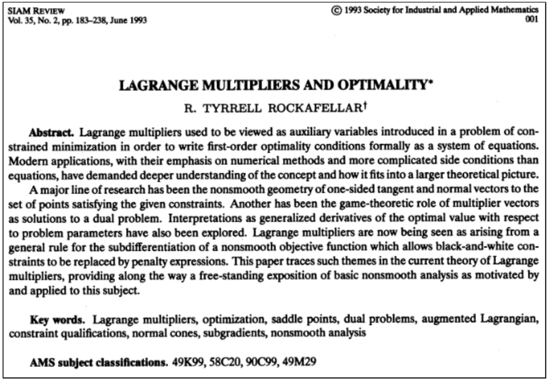
Left: Strictly convex; exactly one local minimum, which is also global. **Middle:** Convex, but not strictly; all local optima are also global ones but not unique. **Right:** Not convex.

CONVEX FUNCTIONS IN OPTIMIZATION / 2

"... in fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity."

– R. Tyrrell Rockafellar. *SIAM Review*, 1993.





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

© 1993 Society for Industrial and Applied Mathematics
001

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