Deep Learning - Theory and Practice

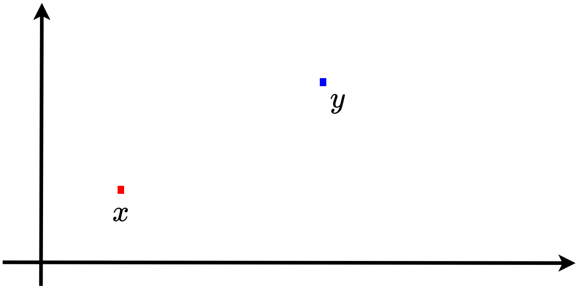
IE 643

Lectures 11, 12 - Part 2

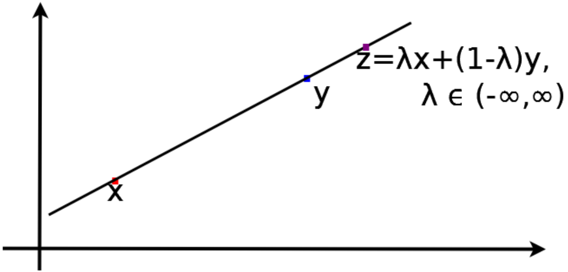
September 9 & 13, 2022

Convex Sets

Points in a 2D space

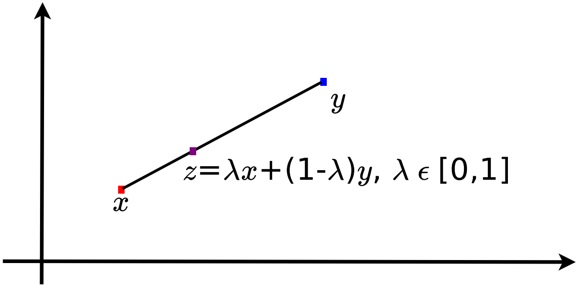


Affine combination of two points



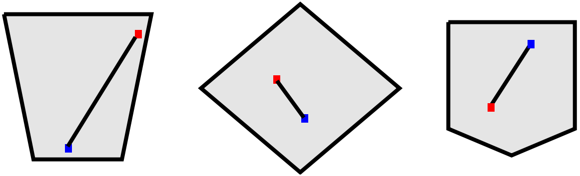
- z is an arbitrary point on the line passing through x and y.

Convex combination of two points



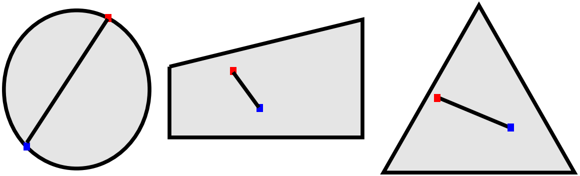
- z is an arbitrary point on the line segment connecting x and y.

Convex Sets



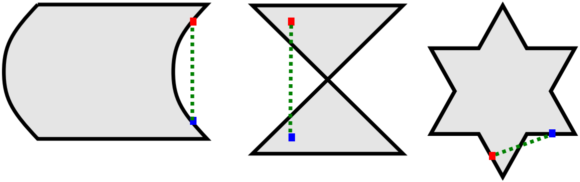
- A set $\mathcal{C}$ is convex if $\lambda x + (1-\lambda)y \in \mathcal{C}$, $\forall x, y \in \mathcal{C}$, $\forall \lambda \in [0,1]$.
- The line segment connecting x and y in $\mathcal{C}$ lies entirely within $\mathcal{C}$.

Convex Sets



- A set $\mathcal{C}$ is convex if $\lambda x + (1 - \lambda) y \in \mathcal{C}$, $\forall x, y \in \mathcal{C}$, $\lambda \in [0, 1]$.
- The line segment connecting x and y in $\mathcal{C}$ lies entirely within $\mathcal{C}$.

Non-convex Sets



- A set $\mathcal{C}$ is not convex if there exist two points x and y in $\mathcal{C}$ such that $\lambda x + (1-\lambda)y \notin \mathcal{C}$, for some $\lambda \in [0,1]$.
- The line segment connecting x and y in $\mathcal{C}$ does not entirely lie within $\mathcal{C}$.

Convex Sets and Convex Combination

Going beyond two points

- Let x, y, z be points in a set $\mathcal{C}$.
- How to extend the definition of convex combination to these three points?

Convex Sets and Convex Combination

Going beyond two points

- Let x, y, z be points in a set $\mathcal{C}$.
- How to extend the definition of convex combination to these three points? (Homework!)

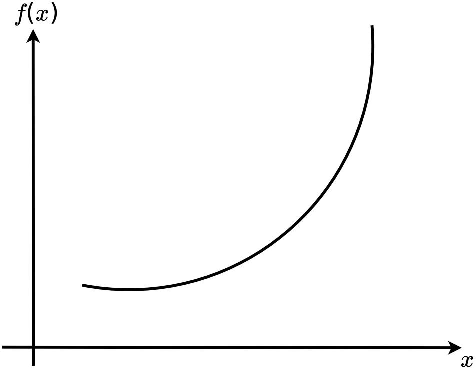
Convex Sets and Convex Combination

Going beyond two points

- More generally, let $x^1, x^2, \ldots, x^m$ be m points in a set $\mathcal{C}$.
- How to extend the definition of convex combination to these m points? (Homework!)

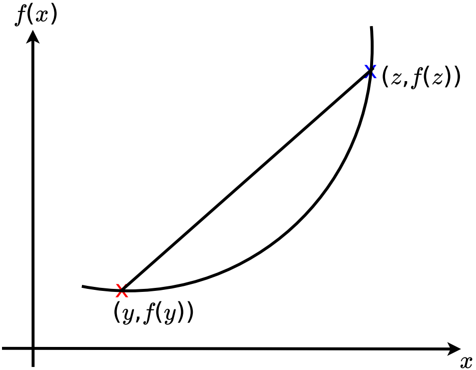
Convex Functions

Convex Function - Definition



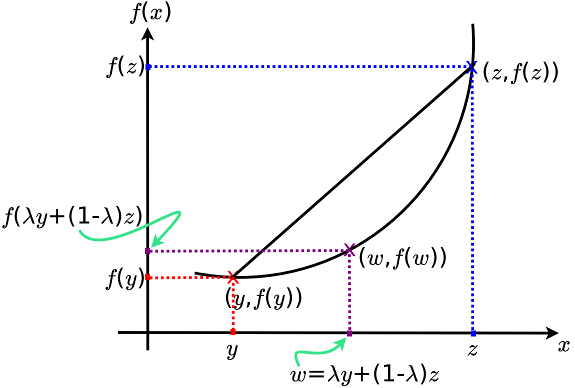
- A function $f : \mathcal{C} \to \mathbb{R}$, defined over a convex set $\mathcal{C} \subseteq \mathbb{R}$ is called convex if $f(\lambda y + (1-\lambda)z) \leq \lambda f(y) + (1-\lambda)f(z)$, $\forall y, z \in \mathcal{C}$, $\forall \lambda \in [0,1]$.

Convex Function - Definition



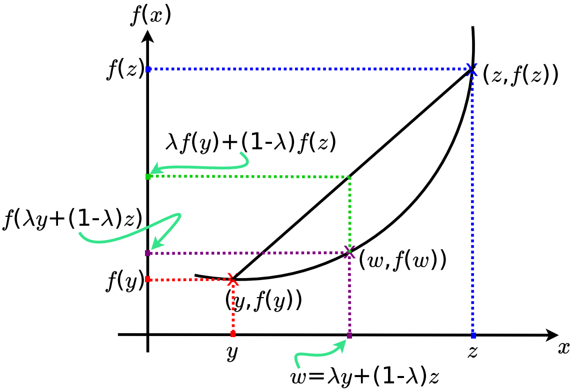
- A function $f : \mathcal{C} \to \mathbb{R}$, defined over a convex set $\mathcal{C} \subseteq \mathbb{R}$ is called convex if $f(\lambda y + (1-\lambda)z) \leq \lambda f(y) + (1-\lambda) f(z)$, $\forall y, z \in \mathcal{C}$, $\forall \lambda \in [0, 1]$.
- Chord over-estimates the graph of function.

Convex Function - Definition



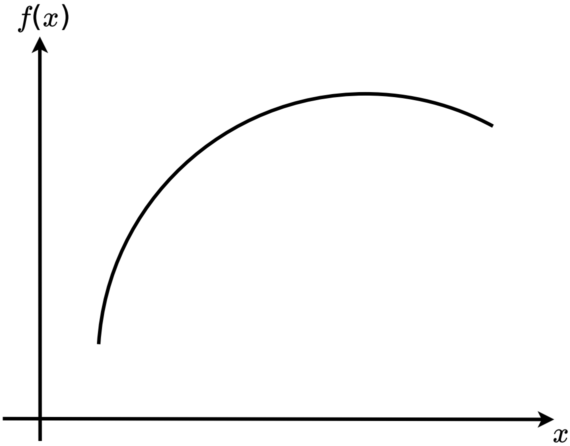
- A function $f : \mathcal{C} \to \mathbb{R}$, defined over a convex set $\mathcal{C} \subseteq \mathbb{R}$ is called convex if $f(\lambda y + (1-\lambda)z) \leq \lambda f(y) + (1-\lambda)f(z)$, $\forall y, z \in \mathcal{C}$, $\forall \lambda \in [0, 1]$.
- Chord over-estimates the graph of function.

Convex Function - Definition



- A function $f : \mathcal{C} \to \mathbb{R}$, defined over a convex set $\mathcal{C} \subseteq \mathbb{R}$ is called convex if $f(\lambda y + (1-\lambda)z) \leq \lambda f(y) + (1-\lambda)f(z)$, $\forall y, z \in \mathcal{C}$, $\forall \lambda \in [0, 1]$.
- Chord over-estimates the graph of function.

Concave Function



- Concave function is a close relative of convex function.
- A function $f : \mathcal{C} \to \mathbb{R}$, defined over a convex set $\mathcal{C} \subseteq \mathbb{R}$ is called concave if -f is convex over $\mathcal{C}$.

Note: Concave functions are also defined over convex sets.

Convex Function - Characterization

Extending convex function definition to more than two points.

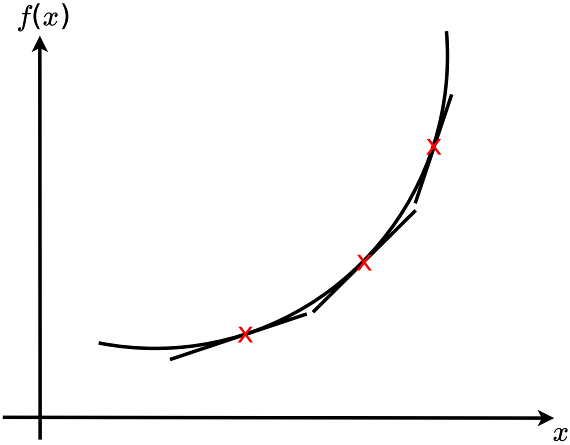
How to extend the definition of convex functions to a set of points $\{x^1, x^2, \ldots, x^m\} \subset \mathcal{C}$?

Convex Function - Characterization

Extending convex function definition to more than two points.

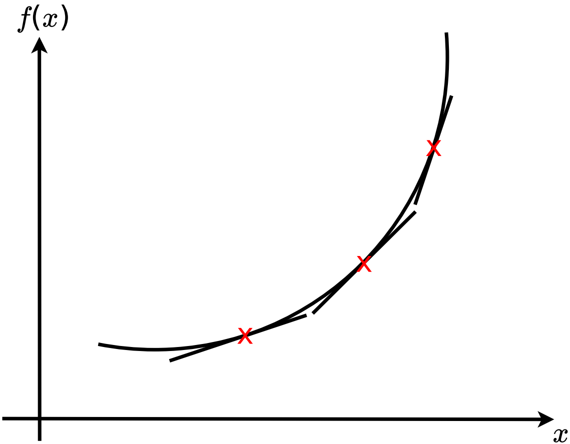
How to extend the definition of convex functions to a set of points $\{x^1, x^2, \dots, x^m\} \subset \mathcal{C}$? (Homework !)

Convex Function - First Order Characterization



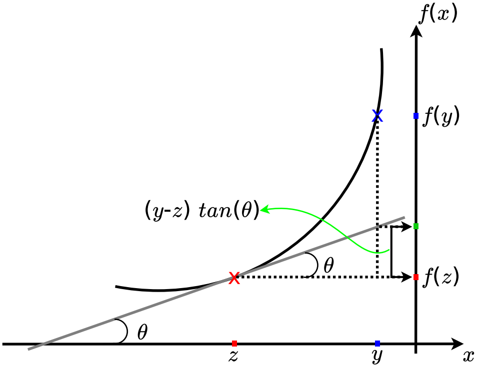
- For differentiable convex functions, the tangent lines under-estimate the graph of function.
- Recall: Tangent at a point is a first-order approximation of a function.

Convex Function - First Order Characterization



- Recall: First order approximation of a function f at y in the vicinity of point z:
  - $f(y) \approx f(z) + (y - z)f'(z)$.

Convex Function - First Order Characterization



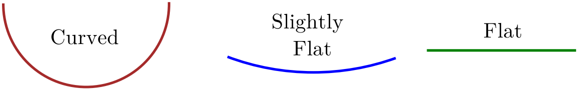
- Let $\mathcal{C} \subseteq \mathbb{R}$ be an open interval. Let $f : \mathcal{C} \to \mathbb{R}$ be a continuously differentiable function. Then f is convex if and only if

$$f(y) \geq f(z) + (y - z)f'(z), \ \forall y, z \in \mathcal{C}.$$

- $f'(z)$ is the derivative of f at z.

Note: $\mathcal{C} \subseteq \mathbb{R}$ is assumed to be an open interval.

Convex Function - Second Order Characterization



- Let $\mathcal{C} \subseteq \mathbb{R}$ be an open interval. Let $f : \mathcal{C} \to \mathbb{R}$ be a twice continuously differentiable function. Then f is convex if and only if $f''(x) \geq 0, \forall x \in \mathcal{C}$.
- $f''(x)$ is the double derivative of f at x.
- $f''(x) \geq 0$ indicates non-negative curvature.

Convex Function - Interesting Properties

Convex functions enjoy several interesting properties.

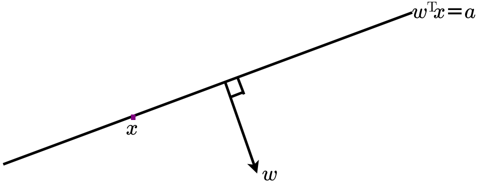
- If f_i are convex for $i = 1, \ldots, m$, their weighted sum $\sum_{i=1}^{m} \theta_i f_i$ is convex, when $\theta_i \geq 0$, $\forall i = 1, \ldots m$.
- If f_i are convex for $i = 1, \ldots, m$, $\max_{i=1,\ldots,m} f_i$ is convex.
- If f is convex then $g(x) = f(ax + b), a, b \in \mathbb{R}$ is also convex. (**Affine invariance**)

Moving Towards Higher Dimensions...

Convex Sets In High Dimensions

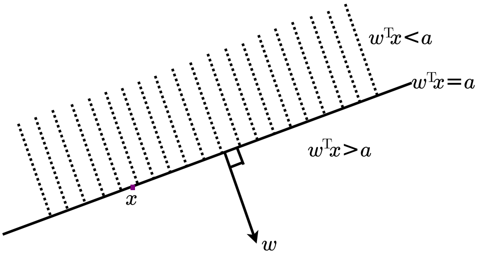
- Extreme cases: the empty set $\emptyset$ and the full space $\mathbb{R}^d$.

Convex Sets In High Dimensions



- Hyperplane: $\{x \in \mathbb{R}^d : w^\top x = a\}$, for some $\mathbf{0} \neq w \in \mathbb{R}^d$ and $a \in \mathbb{R}$.

Convex Sets In High Dimensions



- Closed Halfspace:
  - $\{x \in \mathbb{R}^d : w^\top x \geq a\}$, for some $\mathbf{0} \neq w \in \mathbb{R}^d$ and $a \in \mathbb{R}$.
  - $\{x \in \mathbb{R}^d : w^\top x \leq a\}$, for some $\mathbf{0} \neq w \in \mathbb{R}^d$ and $a \in \mathbb{R}$.
- Open Halfspace:
  - $\{x \in \mathbb{R}^d : w^\top x > a\}$, for some $\mathbf{0} \neq w \in \mathbb{R}^d$ and $a \in \mathbb{R}$.
  - $\{x \in \mathbb{R}^d : w^\top x < a\}$, for some $\mathbf{0} \neq w \in \mathbb{R}^d$ and $a \in \mathbb{R}$.

High Dimensional Representation - Notations

- Gradient of a function $f : \mathbb{R}^d \to \mathbb{R}$ at a point x

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \vdots \\ \vdots \\ \frac{\partial f(x)}{\partial x_d} \end{pmatrix}$$

High Dimensional Representation - Notations

- Hessian Matrix of a function $f : \mathbb{R}^d \to \mathbb{R}$ at a point x

$$H = \nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_d} \\ \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_d \partial x_1} & \frac{\partial^2 f(x)}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_d \partial x_d} \end{pmatrix}$$

- Note the size of H: $d \times d$, we will denote this as $H \in \mathbb{R}^{d \times d}$.
- Note also that H is symmetric. (why?)

Transpose Of A Matrix

Matrix A of size $d \times d$

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1d} \\ \\ a_{21} & a_{22} & \dots & a_{2d} \\ \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots \\ a_{d1} & a_{d2} & \dots & a_{dd} \end{pmatrix}$$

Transpose of A (of same size)

$$A^\top = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{d1} \\ \\ a_{12} & a_{22} & \dots & a_{d2} \\ \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots \\ a_{1d} & a_{2d} & \dots & a_{dd} \end{pmatrix}$$

Note: Rows of matrix A are columns of matrix $A^\top$.

Symmetric Matrix

- A matrix A is symmetric if $A = A^\top$.

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1d} \\ \\ a_{12} & a_{22} & \dots & a_{2d} \\ \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots \\ a_{1d} & a_{2d} & \dots & a_{dd} \end{pmatrix} = A^\top$$

Symmetric Positive (Semi-)definite Matrix

- A symmetric matrix $A \in \mathbb{R}^{d \times d}$ is called positive semi-definite (denoted by $A \succeq 0$) if

$$x^\top Ax \geq 0, \ \forall x \ \in \mathbb{R}^d.$$

Caution: This definition is non-intuitive.

- A symmetric matrix $A \in \mathbb{R}^{d \times d}$ is called positive definite (denoted by $A \succ 0$) if

$$x^\top Ax > 0, \ \forall x \ \in \mathbb{R}^d \text{ such that } x \neq 0.$$

Symmetric Positive (Semi-)definite Matrix

Computation-friendly definitions

- $A \succeq 0 \iff$ all eigen values of A are non-negative.
- $A \succ 0 \iff$ all eigen values of A are positive.
- **Recall**:
  - A (non-zero) vector x is called an eigen vector of matrix $A \in \mathbb{R}^{d \times d}$ with a corresponding eigen value β, if $Ax = \beta x$.
  - An eigen value of a symmetric matrix is always real. (Why?)

Convex Function - Characterization In 1D

Let $\mathcal{C} \subseteq \mathbb{R}$ be an open convex set and let $f : \mathcal{C} \to \mathbb{R}$ be a convex function.
Recall:

- Zero-th Order Characterization

$$f(\lambda y + (1-\lambda)z) \leq \lambda f(y) + (1-\lambda)f(z), \ \forall y, z \in \mathcal{C}, \ \forall \lambda \in [0,1].$$

- First Order Characterization

f continuously differentiable, $f(y) \geq f(z) + f'(z)(y-z), \forall y, z \in \mathcal{C}$.

- Second Order Characterization

f twice continuously differentiable and $f''(x) \geq 0, \ \forall x \in \mathcal{C}$.

Convex Function - Characterization In High Dimensions

Let $\mathcal{C} \subseteq \mathbb{R}^d$ be an open convex set and let $f : \mathcal{C} \to \mathbb{R}$ be a convex function.

- Zero-th Order Characterization

$$f(\lambda y + (1-\lambda)z) \leq \lambda f(y) + (1-\lambda)f(z), \ \forall y, z \in \mathcal{C}, \ \forall \lambda \in [0,1].$$

- First Order Characterization

f continuously differentiable, $f(y) \geq f(z) + \nabla f(z)^\top (y-z), \forall y, z \in \mathcal{C}.$

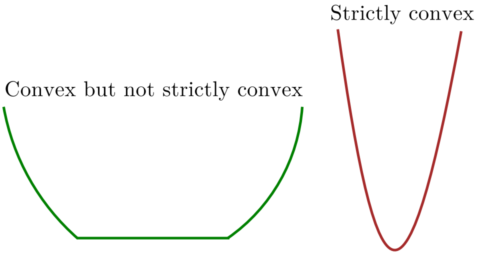
- Second Order Characterization

$$f \text{ twice continuously differentiable and } \nabla^2 f \succeq 0.$$

Other Flavors Of Convex Function

- Strictly convex function
- Strongly convex function

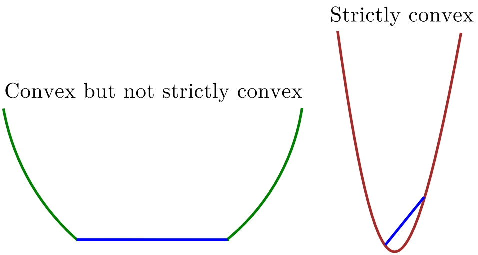
Strictly Convex Function



A function $f : \mathcal{C} \to \mathbb{R}$ defined over a convex set $\mathcal{C}$ is called strictly convex if

$$f(\lambda y + (1-\lambda)z) < \lambda f(y) + (1-\lambda) f(z), \ \forall y, z \in \mathcal{C} \text{ s.t. } x \neq y, \ \forall \lambda \in (0,1).$$

Strictly Convex Function

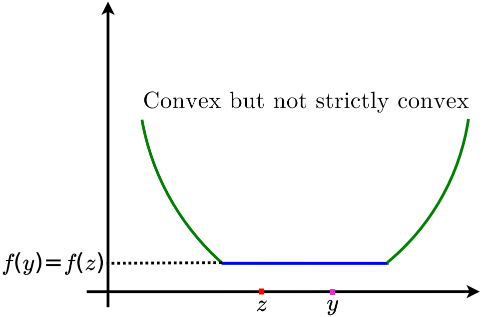


A function $f : \mathcal{C} \rightarrow \mathbb{R}$ defined over a convex set $\mathcal{C}$ is called strictly convex if

$$f(\lambda y + (1-\lambda)z) < \lambda f(y) + (1-\lambda) f(z), \ \forall y, z \in \mathcal{C} \text{ s.t. } x \neq y, \ \forall \lambda \in (0,1).$$

- Graph of the function should be **strictly below** the chord !

Strictly Convex Function - First Order Characterization



- Let $\mathcal{C} \subseteq \mathbb{R}$ be an open convex set. Let $f : \mathcal{C} \to \mathbb{R}$ be a continuously differentiable function. Then f is strictly convex if and only if

$$f(y) > f(z) + \nabla f(z)^\top (y - z), \forall y, z, \in \mathcal{C}, \;\; y \neq z.$$

Strictly Convex Function - Second Order Characterization

- Let $\mathcal{C} \subseteq \mathbb{R}$ be an open convex set. Let $f : \mathcal{C} \to \mathbb{R}$ be a twice continuously differentiable function. f is strictly convex if

$$\nabla^2 f \succ 0.$$

- Important note: This positive definiteness condition is sufficient but not necessary.
  - *e.g.* $f(x) = x^4$, $x \in \mathbb{R}$ is strictly convex but $f''(x) = 0$ at $x = 0$.

Strictly Convex Function

Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a convex set and let $f : \mathcal{C} \to \mathbb{R}$ be a strictly convex function.

- Zero-th Order Characterization

$$f(\lambda y + (1-\lambda)z) < \lambda f(y) + (1-\lambda)f(z), \;\; \forall y, z \in \mathcal{C}, \;\; y \neq z \; \forall \lambda \in (0,1).$$

- First Order Characterization

  f continuously differentiable in $int(\mathcal{C})$ and

$$\forall z \in int(\mathcal{C}), \;\; f(y) > f(z) + \nabla f(z)^\top (y - z), \forall y \neq z \in \mathcal{C}.$$

Strictly Convex Function

Let $\mathcal{C} \subseteq \mathbb{R}^d$ be an open convex set and let $f : \mathcal{C} \to \mathbb{R}$.

- Second Order Characterization **(sufficient condition)**

$$f \text{ twice continuously differentiable and } \nabla^2 f \succ 0, \implies$$

f is strictly convex.

Convex Optimization Problem

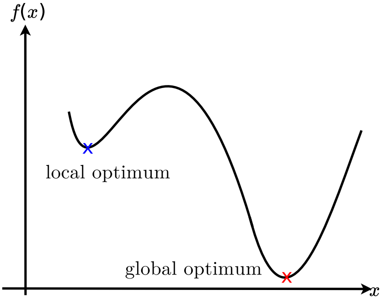
General Optimization Problem

$$\min_{x \in \mathcal{C}} f(x)$$

- f is called objective function and $\mathcal{C}$ is called feasible set.
- Let $f^* = \min_{x \in \mathcal{C}} f(x)$ denote the **optimal objective function value**.
- **Optimal Solution Set** $X^* = \{x \in \mathcal{C} : f(x) = f^*\}$.
- Let us denote by x^* an optimal solution in X^*.

General Optimization Problem

$$\min_{x \in \mathcal{C}} f(x) \tag{OP}$$



General Optimization Problem

$$\min_{x \in \mathcal{C}} f(x) \tag{OP}$$

Local Optimal Solution

A solution z to (OP) is called local optimal solution if $f(z) \leq f(\hat{z})$, $\forall \hat{z} \in \mathcal{N}_\epsilon(z)$ for some $\epsilon > 0$.

Recall: $\mathcal{N}_\epsilon(z)$ denotes the ϵ-neighborhood of z with respect to a suitable distance metric.

Global Optimal Solution

A solution z to (OP) is called global optimal solution if $f(z) \leq f(\hat{z})$, $\forall \hat{z} \in \mathcal{C}$.

Local vs. Global Optimal Solutions

First-order necessary condition for optimality

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function. If a point $x^* \in \mathbb{R}^d$ is a local optimal solution of $\min_{x \in \mathbb{R}^d} f(x)$ then $\nabla f(x^*) = \mathbf{0}$.

General Optimization Problem

$$\min_{x \in \mathcal{C}} f(x)$$

- $\mathcal{C} \subseteq \mathbb{R}^d$ is a convex set
- $f : \mathcal{C} \to \mathbb{R}$ is a convex function

Convex Optimization Regime!

What Is So Special About Convex Optimization?

- Appealing geometry in small dimensions
- Nice properties from an optimization perspective
  - Every local optimal solution (if it exists) is a global optimal solution.

Local vs. Global Optimal Solutions

Proposition

Consider the convex optimization problem $\min_{x \in \mathbb{R}^d} f(x)$ and let $X^* \neq \phi$ (recall: $X^* = \{x \in \mathbb{R}^d : f(x) \leq f(z) \ \forall z \in \mathbb{R}^d\}$ denotes the set of optimal solutions of the optimization problem).
Then every local optimal solution of the problem $\min_{x \in \mathbb{R}^d} f(x)$ is a global optimal solution.

Local vs. Global Optimal Solutions

First-order necessary and sufficient condition for optimality

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable convex function. A point $x^* \in \mathbb{R}^d$ is an optimal solution of $\min_{x \in \mathbb{R}^d} f(x)$ **if and only if** $\nabla f(x^*) = \mathbf{0}$.

Local vs. Global Optimal Solutions

Proposition

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable convex function. A point $x^* \in \mathbb{R}^d$ is an optimal solution of $\min_{x \in \mathbb{R}^d} f(x)$ **if and only if** $\nabla f(x^*) = \mathbf{0}$.

Note: The zero gradient condition is necessary and sufficient for optimality!

Optimal solutions for Strictly Convex Functions

Uniqueness of solution

Consider the convex optimization problem $\min_{x \in \mathbb{R}^d} f(x)$ where f is strictly convex. If the set $X^* = \text{argmin}_{x \in \mathbb{R}^d} f(x)$ is non-empty, then X^* contains **exactly one** element.