Lecture 2. Convex Optimization Basics

Advanced Optimization (Fall 2023)

Peng Zhao

zhaop@lamda.nju.edu.cn

Nanjing University

(Constrained) Optimization Problem

- We adopt a ***minimization*** language

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{X} \end{aligned}$$

  - optimization variable $\mathbf{x} \in \mathbb{R}^d$
  - objective function: $f : \mathbb{R}^d \mapsto \mathbb{R}$
  - feasible domain: $\mathcal{X} \subseteq \mathbb{R}^d$

Unconstrained Optimization

- The optimization variable is feasible over the whole $\mathbb{R}^d$-space.

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \mathbb{R}^d \end{aligned}$$

- It is one of ***the most basic*** forms of mathematical optimization and serves as the foundations.

--- *"any optimization problem can be regarded as an unconstrained one"*

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{X} \end{aligned} \quad \Longrightarrow \quad \begin{aligned} \min \quad & h(\mathbf{x}) \triangleq f(\mathbf{x}) + \delta_{\mathcal{X}}(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \mathbb{R}^d \end{aligned}$$

barrier/indicator function

$$\delta_{\mathcal{X}}(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \in \mathcal{X}, \\ \infty, & \mathbf{x} \notin \mathcal{X}. \end{cases}$$

Convex Optimization

- This lecture focuses on the following simplified setting:
  - Language: ***minimization*** problem
  - Objective function: ***continuous*** and ***convex***
  - Feasible domain: a ***convex*** subset of ***Euclidean space***

- What is a convex set?
- What is a convex function?
- How to minimize?

Outline

- Convex Set and Convex Function
- Convex Optimization Problem
- Optimality Condition
- Function Properties

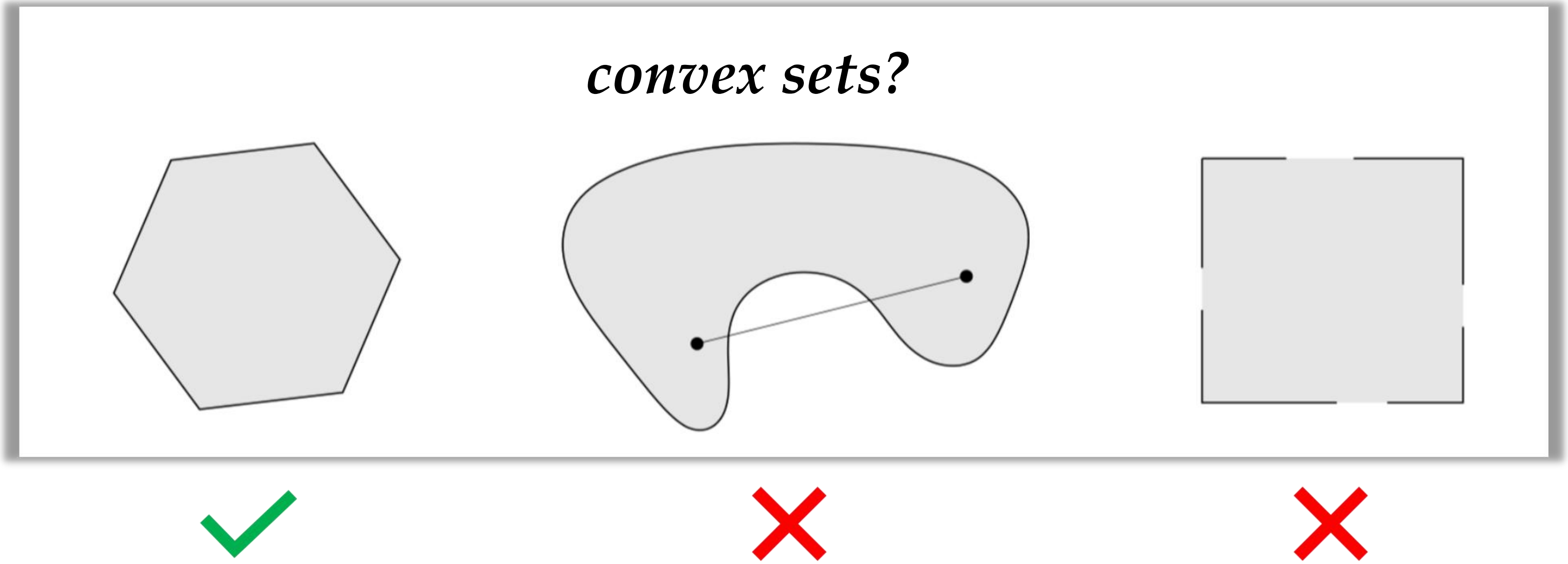
Part 1. Convex Set and Convex Function

- Definition
- Ball and Ellipsoid
- Convex Hull and Projection
- Convex/Concave Function
- Zeroth, First and Second-order Condition

Convex Set

Definition 1 (Convex Set). A set $\mathcal{X}$ is convex if for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, all the points on the line segment connecting $\mathbf{x}$ and $\mathbf{y}$ also belong to $\mathcal{X}$, i.e.,

$$\forall \alpha \in [0,1],\ \alpha\mathbf{x} + (1-\alpha)\mathbf{y} \in \mathcal{X}.$$



Examples

- A line segment is convex.
- A ray, which has the form $\{\mathbf{x}_0 + \theta\mathbf{v} \mid \theta \geq 0\}$, where $\mathbf{v} \neq \mathbf{0}$, is convex.
- Any subspace is convex.

Convex Set

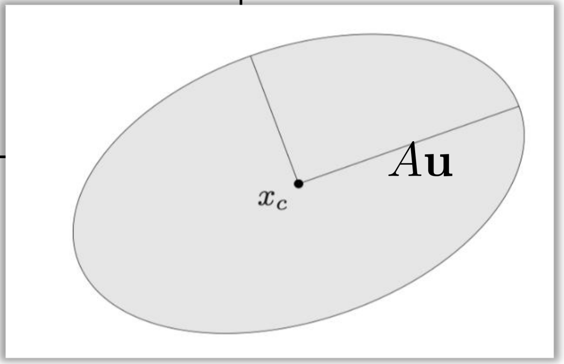
Definition 2 (Ball). A (Euclidean) ball (or just ball) in $\mathbb{R}^d$ has the form

$$\mathbb{B}(\mathbf{x}_c, r) = \{\mathbf{x}_c + r\mathbf{u} \mid \|\mathbf{u}\|_2 \le 1\}.$$

Definition 3 (Ellipsoids). A ellipsoid in $\mathbb{R}^d$ has the form

$$\mathcal{E}(\mathbf{x}_c, A) = \{\mathbf{x}_c + A\mathbf{u} \mid \|\mathbf{u}\|_2 \le 1\},$$

where A is assumed to be symmetric and positive definite.

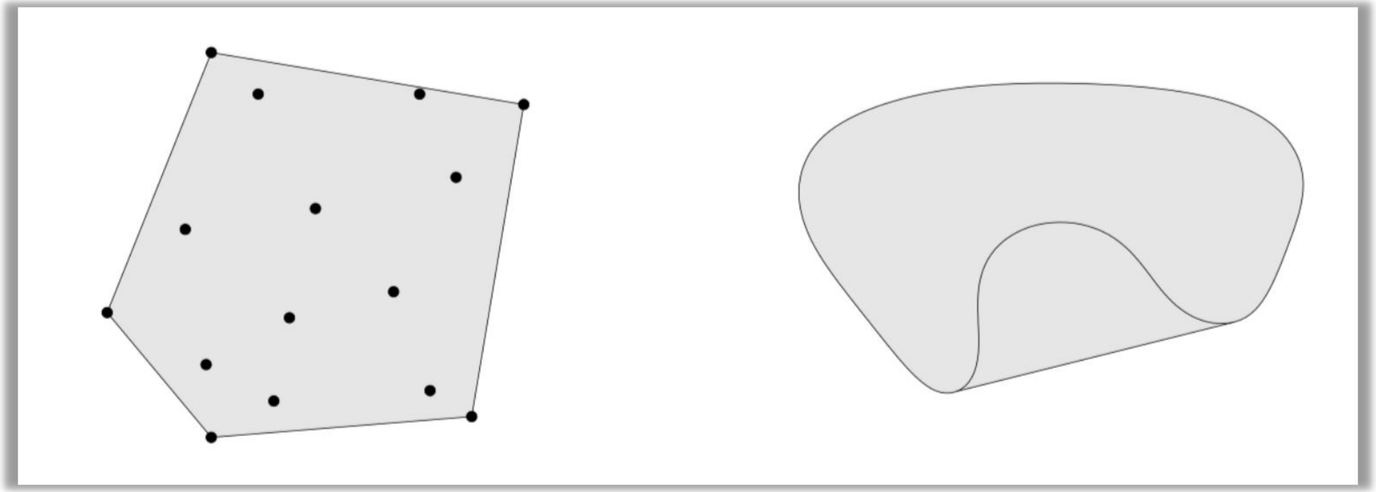


Convex Set

Definition 4 (Convex Hull). The convex hull of a set $\mathcal{X}$, denoted conv $\mathcal{X}$, is the set of all convex combinations of points in $\mathcal{X}$:

$$\operatorname{conv} \mathcal{X} = \{\theta_1 \mathbf{x}_1 + \cdots + \theta_k \mathbf{x}_k \mid \mathbf{x}_i \in \mathcal{X}, \theta_i \geq 0, i \in [k], \theta_1 + \cdots + \theta_k = 1\}.$$

Examples:



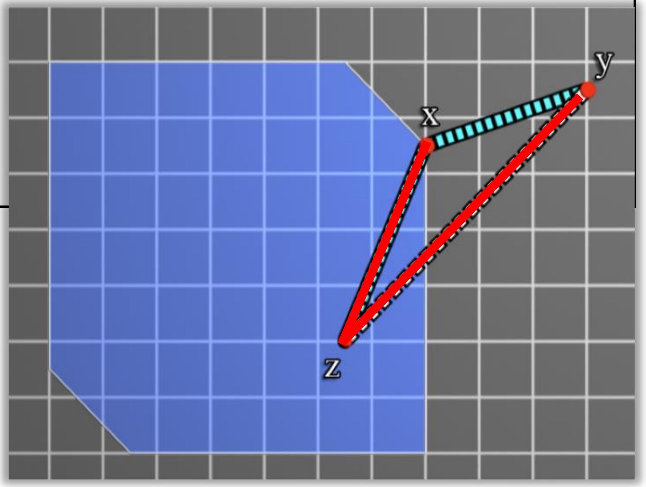
Projection onto Convex Sets

Definition 5 (Projection)**.** The projection $\mathbf{x}$ of a given point $\mathbf{y}$ onto a convex set $\mathcal{X}$ is defined as the closest point inside the convex set. Formally,

$$\mathbf{x} = \Pi_{\mathcal{X}}[\mathbf{y}] \triangleq \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - \mathbf{y}\|.$$

Theorem 1 (Pythagoras Theorem)**.** *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set, $\mathbf{y} \in \mathbb{R}^d$ and $\mathbf{x} = \Pi_{\mathcal{X}}[\mathbf{y}]$. Then for any $\mathbf{z} \in \mathcal{X}$ we have*

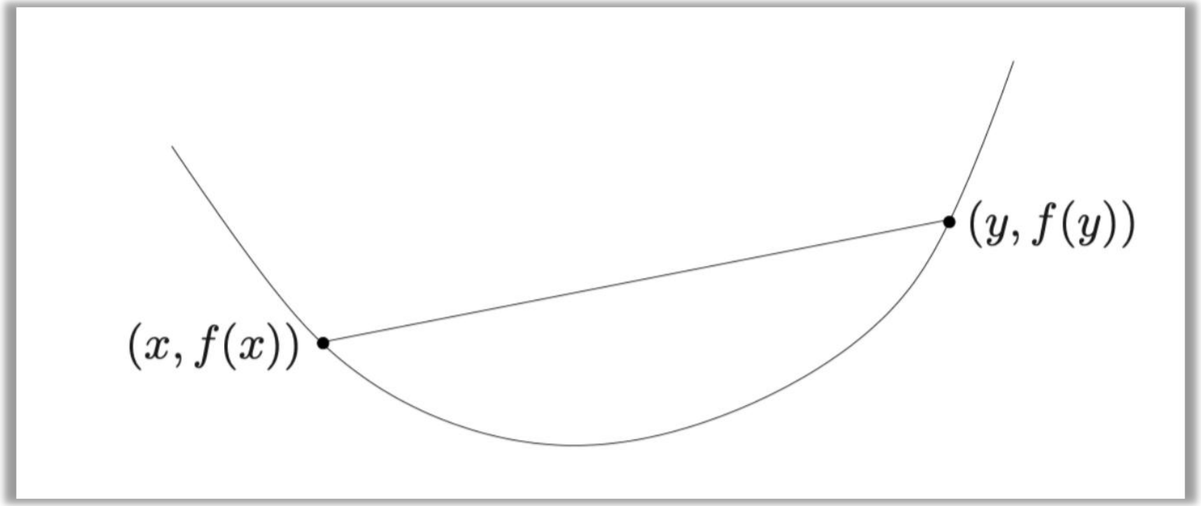
$$\|\mathbf{y} - \mathbf{z}\| \geq \|\Pi_{\mathcal{X}}[\mathbf{y}] - \mathbf{z}\|.$$



Convex Function

Definition 6 (Convex Function)**.** A function $f : \mathcal{X} \mapsto \mathbb{R}$ is convex if for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$,

$$\forall \alpha \in [0, 1], \quad f((1-\alpha)\mathbf{x} + \alpha \mathbf{y}) \leq (1-\alpha) f(\mathbf{x}) + \alpha f(\mathbf{y}).$$



a convex function

Convex/Concave Function

Definition 6 (Convex Function)**.** A function $f : \mathcal{X} \mapsto \mathbb{R}$ is ***convex*** if for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$,

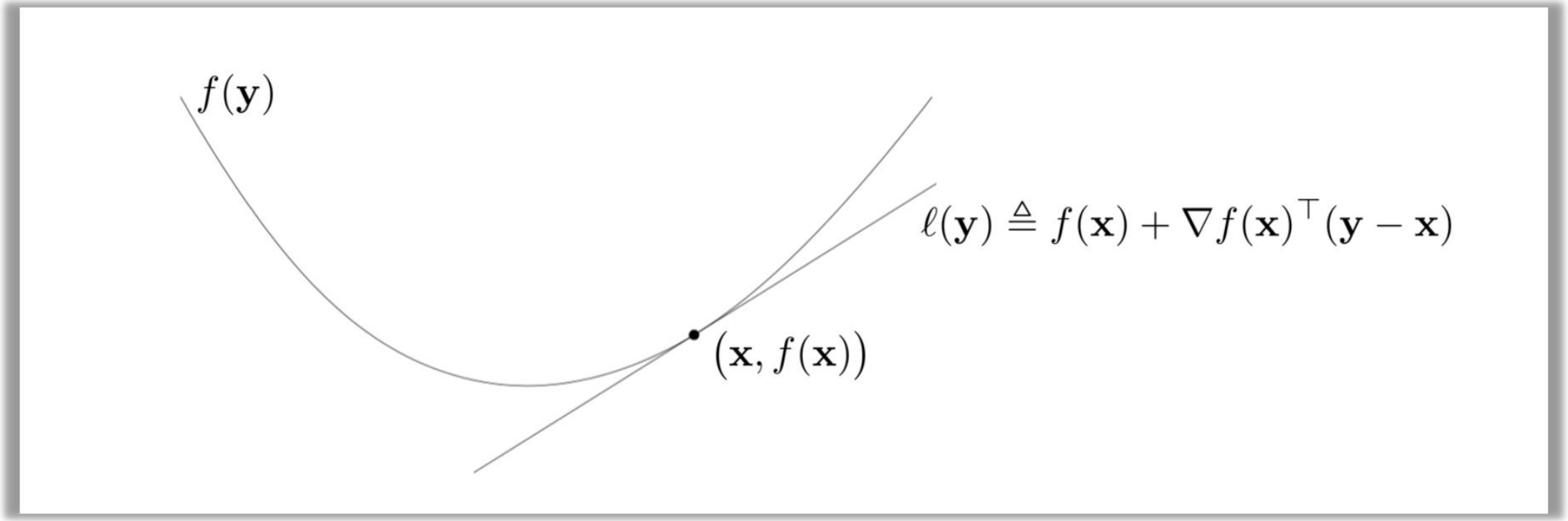
$$\forall \alpha \in [0,1], \quad f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}).$$

Definition 7 (Concave Function)**.** A function $f : \mathcal{X} \mapsto \mathbb{R}$ is ***concave*** if for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$,

$$\forall \alpha \in [0,1], \quad f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \geq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}).$$

- Both definitions have already assume a *convex* feasible domain.
- We focus on the *"convex" language*, since the negative of concave functions are convex.

Convex Function



If f is convex and differentiable, then $\underline{f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle} \leq f(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$.

the first-order Taylor approximation of f near $\mathbf{x}$

A commonly used equivalent form: $f(\mathbf{x}) - f(\mathbf{y}) \leq \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle$.

Convex Function

How to check whether a function is convex or not?

Theorem 2. *A function f is convex **if and only if** $\mathrm{dom}\, f$ **is convex** and one of the following properties hold, for all $\mathbf{x}, \mathbf{y} \in \mathrm{dom}\, f$ and $\alpha \in [0, 1]$,*

(i) Zeroth order condition: $f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \leq (1-\alpha) f(\mathbf{x}) + \alpha f(\mathbf{y})$.

(ii) First order condition: $f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq f(\mathbf{y})$.

(iii) Second order condition: $\nabla^2 f(x) \succeq 0$.

Convex Function

Examples on $\mathbb{R}$:

- Exponential: e^{ax}, where $a \in \mathbb{R}$.
- Powers: x^a, where $a \geq 1$ or $a \leq 0$.
- Powers of absolute value: $|x|^p$, where $p \geq 1$.
- Negative logarithm: $-\log x$.
- Negative entropy: $x \log x$.

Convex Function

Examples on $\mathbb{R}^d$:

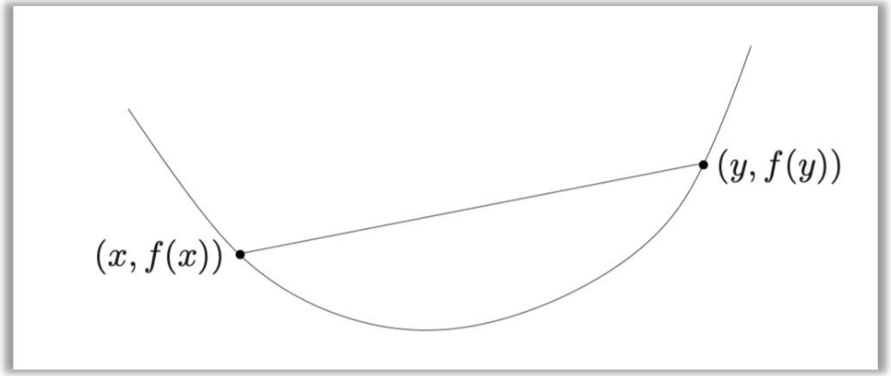
- norm: $f(\mathbf{x}) = \|\mathbf{x}\|$.
- maximum: $f(\mathbf{x}) = \max\{x_1, \dots, x_n\}$.
- Log-sum-exp: $f(\mathbf{x}) = \log\left(e^{x_1} + \cdots + e^{x_n}\right)$.

Jensen's Inequality

Theorem 3 (Jensen's Inequality). *If X is a random variable such that $X \in \operatorname{dom} f$ with probability one, and f is convex, then we have*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

Intuition:



Convexity: $f\left(\underbrace{\theta_1 \mathbf{x}_1+\cdots+\theta_k \mathbf{x}_k}_{\mathbb{E}[X]}\right) \leq \underbrace{\theta_1 f\left(\mathbf{x}_1\right)+\cdots+\theta_k f\left(\mathbf{x}_k\right)}_{\mathbb{E}[f(X)]}$

Part 2. Convex Optimization Problem

- Problem
- Subgradients
- Why Convexity?

Convex Optimization Problem

- We adopt a ***minimization*** language

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & g_i(\mathbf{x}) \le 0, \;\; i = 1, \cdots, m \\ & \mathbf{a}_i^\top \mathbf{x} = b_i, \;\; i = 1, \cdots, n \end{aligned}$$

- optimization variable $\mathbf{x} \in \mathbb{R}^d$
- ***convex*** objective function: $f : \mathbb{R}^d \mapsto \mathbb{R}$
- ***convex*** inequality constraints: $g_1, \ldots, g_m$

Convex Optimization Problem

- We adopt a ***minimization*** language

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & g_i(\mathbf{x}) \leq 0, \quad i = 1, \cdots, m \\ & \mathbf{a}_i^\top \mathbf{x} = b_i, \quad i = 1, \cdots, n \end{aligned}$$

Example 1 (SVM).

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1, \quad i = 1, \cdots, n \end{aligned}$$

Convex Optimization Problem

- We adopt a ***minimization*** language

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & g_i(\mathbf{x}) \leq 0, \quad i = 1, \cdots, m \\ & \mathbf{a}_i^\top \mathbf{x} = b_i, \quad i = 1, \cdots, n \end{aligned}$$

Example 2 (NMF decomposition).

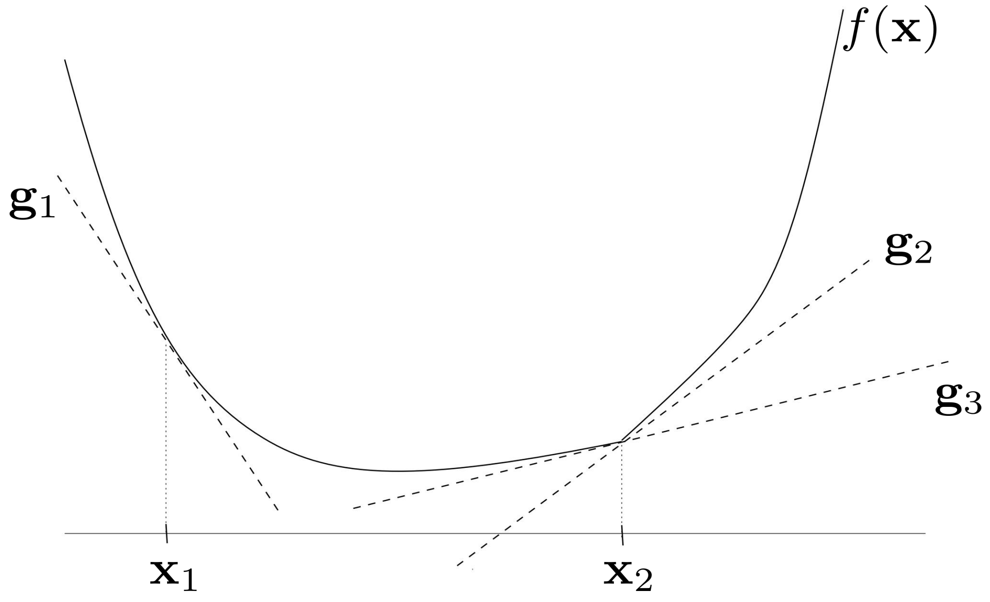
$$\begin{aligned} \min_{U,V} \quad & \left\| X - UV^\top \right\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & U_{i,j}, V_{i,j} \geq 0 \end{aligned}$$

Subgradient

Definition 8 (Subgradient)**.** Let $f : \mathcal{X} \mapsto \mathbb{R}$ be a proper function and let $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$. A vector $\mathbf{g} \in \mathbb{R}^d$ is called a *subgradient* of f at $\mathbf{x}$ if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle, \text{ for all } \mathbf{y} \in \mathbb{R}^d.$$

Intuition: subgradient $\mathbf{g} \in \partial f(\mathbf{x})$ can be any variable that makes the line $f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle$ below the curve f.



Subdifferential

Definition 8 (Subgradient). Let $f : \mathcal{X} \mapsto \mathbb{R}$ be a proper function and let $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$. A vector $\mathbf{g} \in \mathbb{R}^d$ is called a *subgradient* of f at $\mathbf{x}$ if

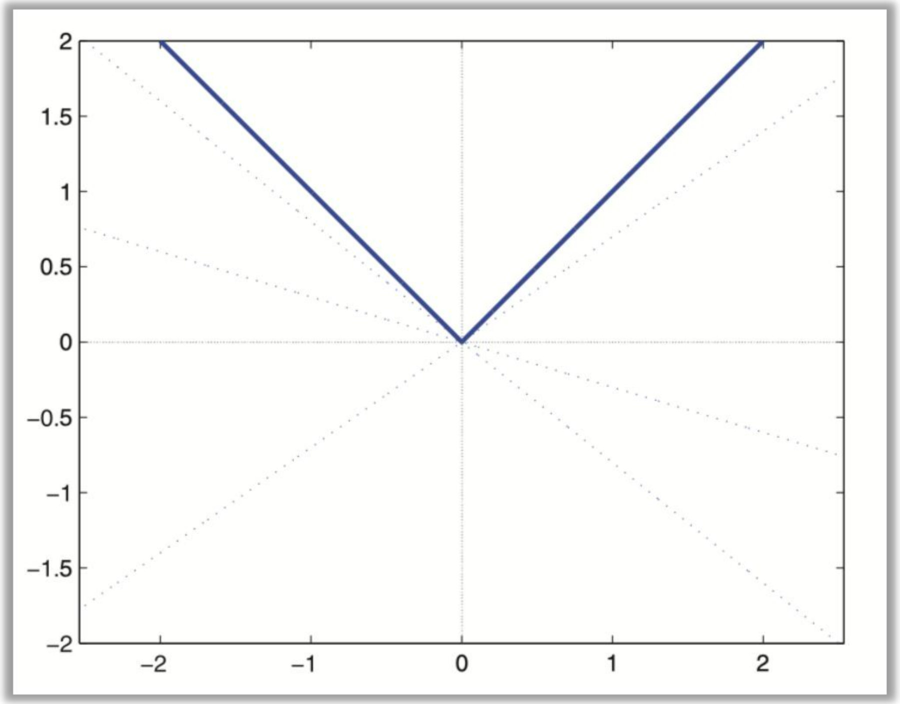
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle, \text{ for all } \mathbf{y} \in \mathbb{R}^d.$$

Definition 9 (Subdifferential). The set of all subgradients of f at $\mathbf{x}$ is called the *subdifferential* of f at $\mathbf{x}$ and is denoted by $\partial f(\mathbf{x})$,

$$\partial f(\mathbf{x}) \triangleq \{\mathbf{g} \in \mathbb{R}^d \mid f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle, \text{ for all } \mathbf{y} \in \mathbb{R}^d\}.$$

Subgradient and Subdifferential

Example 3. The subdifferential of $f(\mathbf{x}) = \|\mathbf{x}\|$ at $\mathbf{x} = \mathbf{0}$ is the dual norm unit ball, i.e., $\partial f(\mathbf{0}) = \{\mathbf{g} \mid \|\mathbf{g}\|_* \leq 1\}$.



an illustration for 1-dim case

$$f(x) = |x|$$

Subgradient and Subdifferential

Example 3. The subdifferential of $f(\mathbf{x}) = \|\mathbf{x}\|$ at $\mathbf{x} = \mathbf{0}$ is the dual norm unit ball, i.e., $\partial f(\mathbf{0}) = \{\mathbf{g} \mid \|\mathbf{g}\|_* \le 1\}$.

Proof:

By definition, it suffices to prove that $\mathbf{g} \in \partial f(\mathbf{0})$ if and only if

$$\|\mathbf{y}\| \ge \langle \mathbf{g}, \mathbf{y} \rangle \text{ holds for all } \mathbf{y} \in \mathbb{R}^d.$$

① if $\|\mathbf{g}\|_* \le 1$, then by the Cauchy-Schwarz inequality,

$$\langle \mathbf{g}, \mathbf{y} \rangle \le \|\mathbf{y}\| \, \|\mathbf{g}\|_* \le \|\mathbf{y}\|.$$

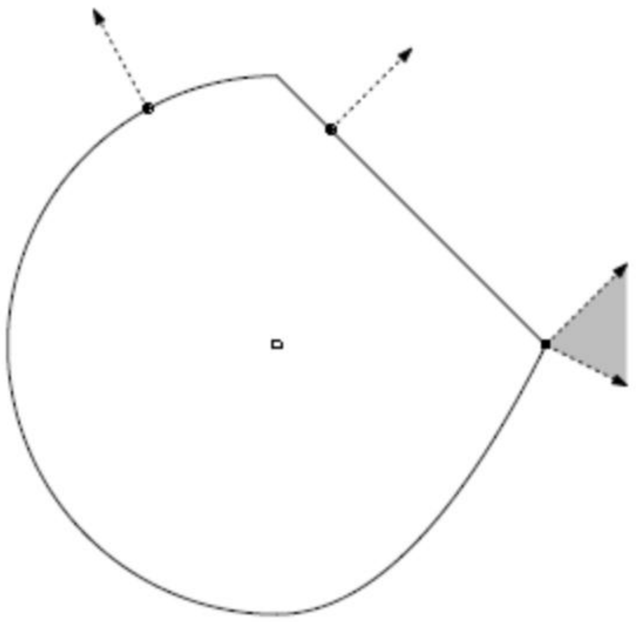
② if $\|\mathbf{y}\| \ge \langle \mathbf{g}, \mathbf{y} \rangle$ is true, then by the definition of dual norm,

$$\|\mathbf{g}\|_* \triangleq \sup\{\langle \mathbf{g}, \mathbf{y} \rangle \mid \|\mathbf{y}\| \le 1\} \le \sup\{\|\mathbf{y}\| \mid \|\mathbf{y}\| \le 1\} \le 1. \qquad \square$$

Subgradient and Subdifferential

Example 4. For indicator function $f(\mathbf{x}) = \delta_{\mathcal{X}}(\mathbf{x})$, its subdifferential at any point $\mathbf{x} \in \mathcal{X}$ is $N_{\mathcal{X}}(\mathbf{x}) = \partial f(\mathbf{x}) = \{\mathbf{g} \mid \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \leq 0, \forall \mathbf{y} \in \mathcal{X}\}$.

called normal cone



Subgradient

- Relationship between ***Lipschitzness*** and ***bounded subgradient***

Theorem 1. *Let $f : \mathcal{X} \to \mathbb{R}$ be a convex function. Consider the following two claims:*

(i) Lipschitzness: $|f(\mathbf{x}) - f(\mathbf{y})| \leq G\|\mathbf{x} - \mathbf{y}\|$ for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.

(ii) Bounded subgradient: $\|\mathbf{g}\|_* \leq G$ for any $\mathbf{g} \in \partial f(\mathbf{x}), \mathbf{x} \in \mathcal{X}$.

Then

(a) (ii) $\Rightarrow$ (i).

(b) if $\mathcal{X}$ is open, then (i) $\Leftrightarrow$ (ii).

Existence of Subgradient

- ***Existence of subgradients*** implies ***convexity***.

> **Theorem 5.** *Let $f : \mathcal{X} \mapsto \mathbb{R}$ be a proper function and assume $\mathcal{X}$ is convex. If **for any** $\mathbf{x} \in \mathcal{X}$, its subgradients exist, then f is convex.*

- A ***sufficient condition*** for deciding a convex function.
- The reverse direction is ***not*** always correct (example on the next page).

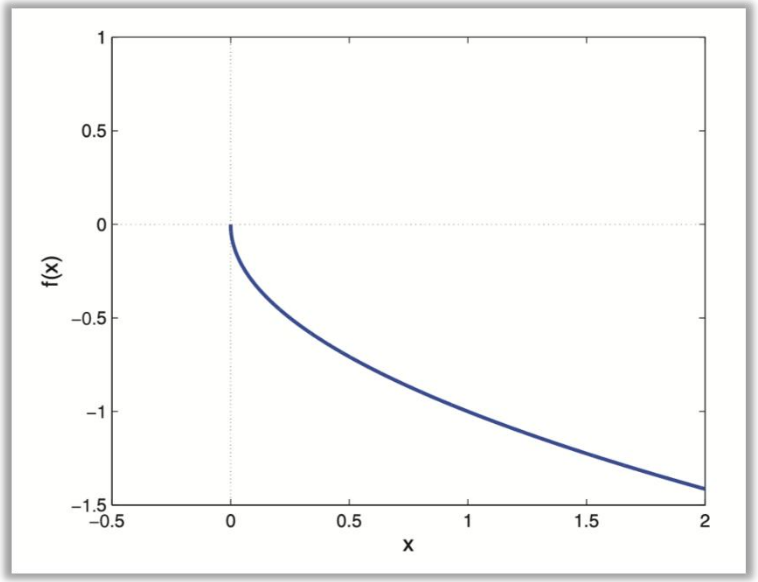
Existence of Subgradient

- Convexity ***doesn't*** always imply existence of subgradients.

Example 5. Consider function $f : \mathbb{R} \to (-\infty, \infty]$ defined by

$$f(x) = \begin{cases} -\sqrt{x}, & x \geq 0 \\ \infty, & \text{else} \end{cases},$$

it is convex but does not have a subgradient at $x = 0$.



Existence of Subgradient

- Nevertheless, if we only care about the ***interior*** of feasible domain, convexity ***does*** imply existent subgradients.

Theorem 6. *Let* $f : \mathcal{X} \mapsto \mathbb{R}$ *be a convex function and assume the feasible domain* $\mathcal{X}$ *is convex. Consider any interior point* $\mathbf{x} \in \text{int}(\mathcal{X})$*. Then* $\partial f(\mathbf{x})$ *is nonempty.*

How to Compute Subgradient

- General principle: unfortunately, hard to give :(
- Ad-hoc calculations: see earlier examples.
- **Good news**: easy for ***convex and differential*** functions.

Theorem 7. *Let $f : \mathcal{X} \mapsto \mathbb{R}$ be a proper and convex function and assume $\mathcal{X}$ is convex.*

1. *If f is differentiable at $\mathbf{x}$, then $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.*
2. *Conversely, if f has a unique subgradient, then it is differentiable at $\mathbf{x}$ and $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$.*

How to Compute Subgradient

Example 6. The subdifferential of ℓ_2-norm $f(\mathbf{x}) = \|\mathbf{x}\|_2$ is

$$\partial f(\mathbf{x}) = \begin{cases} \left\{ \frac{\mathbf{x}}{\|\mathbf{x}\|_2} \right\}, & \mathbf{x} \neq \mathbf{0} \ \text{(gradient of norm)} \\ \{\mathbf{g} \mid \|\mathbf{g}\|_2 \leq 1\}, & \mathbf{x} = \mathbf{0} \ \text{(discussed earlier)} \end{cases}$$

Why Convexity?

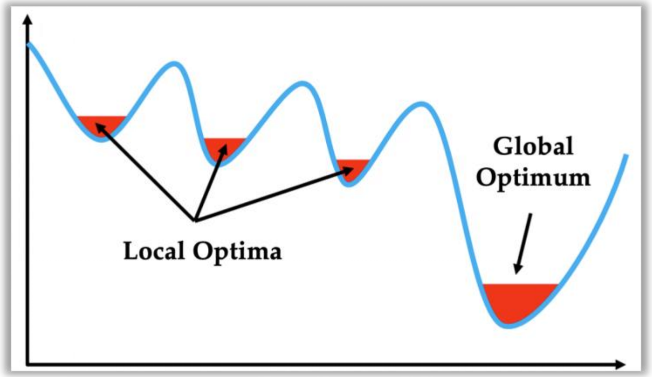
- **Local to Global Phenomenon**

  For convex (and differentiable) functions, ***gradient is highly informative***.

$$\nabla f(\mathbf{x}) \in \partial f(\mathbf{x})$$

  - **Local**: the gradient $\nabla f(\mathbf{x})$ contains a priori only ***local*** information about the function f around $\mathbf{x}$;
  - **Global**: the subdifferential $\partial f(\mathbf{x})$ gives a global information in the form of a linear lower bound on the ***entire*** function.

Why Convexity?



- **Local to Global Phenomenon**

For convex (unconstrained) optimization, ***local minima are global minima***.

> **Theorem 8.** *Let f be convex. If $\mathbf{x}$ is a local minimum of f then $\mathbf{x}$ is a global minimum of f.*

A simple proof:

Assume that $\mathbf{x}$ is local minimum of f. Then for γ small enough, for any $\mathbf{y}$,

(local minima)

$$\underline{f(\mathbf{x}) \le f((1-\gamma)\mathbf{x} + \gamma\mathbf{y})} \le (1-\gamma)f(\mathbf{x}) + \gamma f(\mathbf{y}),$$

which implies $f(\mathbf{x}) \le f(\mathbf{y})$ and thus $\mathbf{x}$ is a global minimum of f.

Part 3. Optimality Condition

- Fermat's Optimality Condition
- First-order Optimality Condition
- Some Corollaries

Fermat's Optimality Condition

- ***Unconstrained*** case

Theorem 9 (Fermat's Optimality Condition). *Let $f : \mathbb{R}^d \to (-\infty, \infty]$ be a proper convex function. Then*

$$\mathbf{x}^\star \in \operatorname{argmin}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathbb{R}^d\}$$

if and only if $\mathbf{0} \in \partial f(\mathbf{x}^\star)$.

A simple proof:

Combining $f(\mathbf{x}) \geq f(\mathbf{x}^\star)$ and $f(\mathbf{x}) \geq f(\mathbf{x}^\star) + \langle \mathbf{g}, \mathbf{x} - \mathbf{x}^\star \rangle, \mathbf{g} \in \partial f(\mathbf{x}^\star)$ finishes the proof.

Example

Example 7 (Median). Suppose that we are given n different and ordered numbers $a_1 < a_2 < \cdots < a_n$. Denote $A = \{a_1, a_2, \ldots, a_n\} \subseteq \mathbb{R}$. The median of A is a number satisfying

$$\text{median}(A) = \begin{cases} a_{\frac{n+1}{2}}, & n \text{ odd} \\ \left[a_{\frac{n}{2}}, a_{\frac{n}{2}+1}\right], & n \text{ even} \end{cases}.$$

Solving the optimization problem:

From an optimization perspective, solving medians equals to solving the following optimization problem.

$$\text{median}(A) = \arg\min_x \left\{ f(x) \triangleq \sum_{i=1}^{n} |x - a_i| \right\}$$

Example

- ***Proof of median***

From an optimization perspective, solving medians equals to solving the following optimization problem.

$$\text{median}(A) = \arg\min_{x} \left\{ f(x) \triangleq \sum_{i=1}^{n} |x - a_i| \right\}$$

Denote $f_i(x) = |x - a_i|$, then it hold that $f(x) = f_1(x) + f_2(x) + \cdots + f_n(x)$ and

$$\partial f_i(x) = \begin{cases} 1, & x > a_i \\ -1, & x < a_i \\ [-1, 1], & x = a_i \end{cases}$$

Example

- ***Proof of median***

Denote $f_i(x) = |x - a_i|$, then it hold that $f(x) = f_1(x) + f_2(x) + \cdots + f_n(x)$ and

$$\partial f_i(x) = \begin{cases} 1, & x > a_i \\ -1, & x < a_i \\ [-1, 1], & x = a_i \end{cases}$$

$$\begin{aligned} \partial f(x) &= \partial f_1(x) + \partial f_2(x) + \cdots + \partial f_n(x) \\ &= \begin{cases} \#\{i : a_i < x\} - \#\{i : a_i > x\}, & x \notin A, \\ \#\{i : a_i < x\} - \#\{i : a_i > x\} + [-1, 1], & x \in A. \end{cases} \end{aligned}$$

Example

- ***Proof of median***

$$\partial f(x) = \partial f_1(x) + \partial f_2(x) + \cdots + \partial f_n(x)$$
$$= \begin{cases} \#\{i : a_i < x\} - \#\{i : a_i > x\}, & x \notin A, \\ \#\{i : a_i < x\} - \#\{i : a_i > x\} + [-1, 1], & x \in A. \end{cases}$$

$$\partial f(x) = \begin{cases} i - (n - i) = 2i - n, & x \in (a_i, a_{i+1}) \\ (i - 1) - (n - i) + [-1, 1] = 2i - 1 - n + [-1, 1], & x = a_i \\ -n, & x < a_1 \\ n, & x > a_n \end{cases}$$

Example

- ***Proof of median***

$$\partial f(x) = \begin{cases} i-(n-i) = 2i-n, & x \in (a_i, a_{i+1}) \\ (i-1)-(n-i)+[-1,1] = 2i-1-n+[-1,1], & x = a_i \\ -n, & x < a_1 \\ n, & x > a_n \end{cases}$$

① Suppose $x = a_i$. Then,

$$0 \in \partial f(x) = 2i-1-n+[-1,1] \Leftrightarrow |2i-1-n| \leq 1 \Leftrightarrow \frac{n}{2} \leq i \leq \frac{n}{2}+1 \Leftrightarrow x = \left[a_{\frac{n}{2}}, a_{\frac{n}{2}+1}\right]$$

② Suppose $x \in (a_i, a_{i+1})$. Then, $0 \in \partial f(x) = 2i-n \Leftrightarrow i = \frac{n}{2} \Leftrightarrow x \in \left(a_{\frac{n}{2}}, a_{\frac{n}{2}+1}\right)$

Combining the two cases finishes the proof (by further checking n is odd or even). $\square$

First-order Optimality Condition

- ***Constrained*** Case

> **Theorem 10** (First-order Optimality Condition). *Let f be convex and $\mathcal{X}$ a closed convex set on which f is differentiable. Then $\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$ if and only if there exists $\mathbf{g} \in \partial f(\mathbf{x}^\star)$ such that*
>
> $$\langle \mathbf{g}, \mathbf{x} - \mathbf{x}^\star \rangle \geq 0, \forall \mathbf{x} \in \mathcal{X}.$$

A simple proof: derived from the ***Fermat's optimality condition***.

⇨ deploying the Fermat's optimility condition on the unconstrained "surrogate" objective

$$h(\mathbf{x}) \triangleq f(\mathbf{x}) + \delta_{\mathcal{X}}(\mathbf{x})$$

First-order Optimality Condition

- ***Constrained*** Case

Theorem 10 (First-order Optimality Condition). *Let f be convex and $\mathcal{X}$ a closed convex set on which f is differentiable. Then $\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$ if and only if there exists $\mathbf{g} \in \partial f(\mathbf{x}^\star)$ such that*

$$\langle \mathbf{g}, \mathbf{x} - \mathbf{x}^\star \rangle \geq 0, \forall \mathbf{x} \in \mathcal{X}.$$

Example 4. For indicator function $f(\mathbf{x}) = \delta_{\mathcal{X}}(\mathbf{x})$, its subdifferential at any point $\mathbf{x} \in \mathcal{X}$ is $N_{\mathcal{X}}(\mathbf{x}) = \partial f(\mathbf{x}) = \{\mathbf{g} \mid \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \leq 0, \forall \mathbf{y} \in \mathcal{X}\}$.

$$\Rightarrow \quad \partial h(\mathbf{x}) = \partial f(\mathbf{x}) + N_{\mathcal{X}}(\mathbf{x})$$

Set Addition: elementwise sum

First-order Optimality Condition

- ***Constrained*** Case

> **Theorem 10** (First-order Optimality Condition). *Let f be convex and $\mathcal{X}$ a closed convex set on which f is differentiable. Then $\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$ if and only if there exists $\mathbf{g} \in \partial f(\mathbf{x}^\star)$ such that*
>
> $$\langle \mathbf{g}, \mathbf{x} - \mathbf{x}^\star \rangle \geq 0, \forall \mathbf{x} \in \mathcal{X}.$$

Fermat's optimality condition says that $\mathbf{x}^\star$ is optimal if and only if $\mathbf{0} \in \partial f(\mathbf{x}^\star)$.

$$\mathbf{0} \in \partial h(\mathbf{x}^\star) = \partial f(\mathbf{x}^\star) + N_{\mathcal{X}}(\mathbf{x}^\star)$$

$$\Rightarrow \quad -\partial f(\mathbf{x}^\star) \cap N_{\mathcal{X}}(\mathbf{x}^\star) \neq \emptyset$$

$$\Rightarrow \quad \exists \mathbf{g} \in -\partial f(\mathbf{x}^\star) \quad \text{s.t. } \langle \mathbf{g}, \mathbf{x} - \mathbf{x}^\star \rangle \leq 0, \forall \mathbf{x} \in \mathcal{X} \qquad \square$$

Karush–Kuhn–Tucker (KKT) Conditions

Theorem 11. *Consider the minimization problem*

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & g_i(\mathbf{x}) \le 0, \;\; i \in [m], \end{aligned} \tag{1}$$

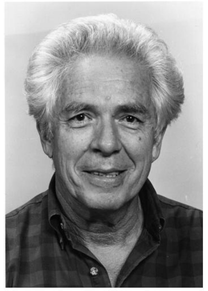
where $f, g_1, g_2, \ldots, g_m$ *are real-valued convex functions.*

1. *Let* $\mathbf{x}^\star$ *be an optimal solution of (1), and assume that Slater's condition is satisfied. Then there exist* $\lambda_1, \ldots, \lambda_m \ge 0$ *for which*

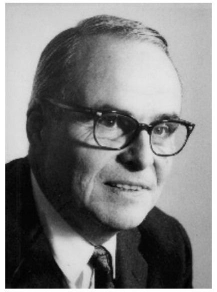
$$\mathbf{0} \in \partial f(\mathbf{x}^\star) + \sum_{i=1}^{m} \lambda_i \partial g_i(\mathbf{x}^\star) \tag{2}$$

$$\lambda_i g_i(\mathbf{x}^\star) = 0, \quad i \in [m]. \tag{3}$$

2. *If* $\mathbf{x}^\star$ *satisfies conditions (2) and (3) for some* $\lambda_1, \lambda_2, \ldots, \lambda_m \ge 0$*, then it is an optimal solution of problem (1).*



Harold Kuhn
1925-2014



Albert Tucker
1905-1995

Published conditions in 1951.

William Karush
1917-1997

Developed (necessary) conditions in 1939 in his (unpublished) MS thesis.

Part 4. Function Properties

- Smoothness
- Strong Convexity

Lipschitz Continuity

Definition 1 (Continuity). A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is continuous at $\mathbf{x} \in \operatorname{dom} f$ if for all $\epsilon > 0$ there exists a $\delta > 0$ with $\mathbf{y} \in \operatorname{dom} f$, such that

$$\|\mathbf{y} - \mathbf{x}\|_2 \leq \delta \Rightarrow \|f(\mathbf{y}) - f(\mathbf{x})\|_2 \leq \epsilon.$$

Definition 2 (Lipschitz Continuity). A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is G-Lipschitz-continuous if for all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$,

$$\|f(\mathbf{x}) - f(\mathbf{y})\| \leq G \|\mathbf{x} - \mathbf{y}\|.$$

Lipschitzness and Subgradient

- Relationship between ***Lipschitzness*** and ***bounded subgradient***

Theorem 1. *Let* $f : \mathcal{X} \to \mathbb{R}$ *be a convex function. Consider the following two claims:*

(i) Lipschitzness: $|f(\mathbf{x}) - f(\mathbf{y})| \leq G\|\mathbf{x} - \mathbf{y}\|$ *for any* $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.

(ii) Bounded subgradient: $\|\mathbf{g}\|_* \leq G$ *for any* $\mathbf{g} \in \partial f(\mathbf{x}), \mathbf{x} \in \mathcal{X}$.

Then

(a) (ii) $\Rightarrow$ *(i).*

(b) if $\mathcal{X}$ *is open, then (i)* $\Leftrightarrow$ *(ii).*

Smoothness

Definition 3 (Smoothness). A function f is L-smooth with respect to the $\|\cdot\|$ norm if, for any $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L\|\mathbf{x} - \mathbf{y}\|.$$

Smoothness is also called ***gradient Lipschitz*** in many literature.

Smoothness is defined over the primal-dual norms, which become ℓ_2-norm when specialized to Euclidean space (and then, $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\,\|\mathbf{x} - \mathbf{y}\|_2$).

The class of L-smooth functions over domain $\mathcal{X}$ is denoted by $C_L^{1,1}(\mathcal{X})$.

Smoothness

Definition 4. Let $\mathcal{X} \subseteq \mathbb{R}^d$. We denote by $C_L^{a,b}(\mathcal{X})$ the class of functions with the following properties:

(i) any $f \in C_L^{a,b}(\mathcal{X})$ is a times continuously differentiable on $\mathcal{X}$.

(ii) f's b-th derivative is Lipschitz continuous on $\mathcal{X}$ with constant L:

$$\left\|\nabla^b f(\mathbf{x}) - \nabla^b f(\mathbf{y})\right\|_* \leq L\|\mathbf{x} - \mathbf{y}\|, \ \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}.$$

Ref: Lectures on Convex Optimization, Yurii Nesterov. Page 23-24.

- Lipschitz continuous functions belong to $C_L^{0,0}(\mathcal{X})$.

- Smoothness is in fact the ***Lipshitzness of gradients***.

Smoothness

Example 1. Linear function $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + c$ is 0-smooth.

Example 2. Quadratic function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} + \mathbf{w}^\top \mathbf{x} + c$ is $\|A\|_{\mathrm{op},p}$-smooth w.r.t. $\|\cdot\|_p$ norm.

Proof. The proof is direct by the definition of smoothness and the operator norm:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_p = \|A\mathbf{x} - A\mathbf{y}\|_p \leq \|A\|_{\mathrm{op},p}\|\mathbf{x} - \mathbf{y}\|_p.$$

Definition 6 (Matrix Operator Norm). The operator norm (or called induced norm) of a matrix $A \in \mathbb{R}^{m\times n}$ is defined by

$$\|A\|_{\mathrm{op},p} \triangleq \max\left\{\frac{\|A\mathbf{x}\|_p}{\|\mathbf{x}\|_p} \;\middle|\; \mathbf{x} \in \mathbb{R}^d, \mathbf{x} \neq \mathbf{0}\right\}.$$

Smoothness

Example 3. Log-sum-exp function $f(\mathbf{x}) = \log\left(e^{x_1} + e^{x_2} + \cdots + e^{x_n}\right)$ is 1-smooth w.r.t. ℓ_2-norm and ℓ_∞-norm.

Example 4. Function $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_p^2$ is $(p-1)$-smooth w.r.t. ℓ_p-norm.

Example 5. Function $f(\mathbf{x}) = \sqrt{1 + \|\mathbf{x}\|_2^2}$ is 1-smooth w.r.t. ℓ_2-norm.

Example 6. Function $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \Pi_{\mathcal{X}}[\mathbf{x}]\|^2$ is 1-smooth w.r.t. ℓ_2-norm, where $\Pi_{\mathcal{X}}[\mathbf{x}]$ denotes the Euclidean projection of $\mathbf{x}$ onto a *convex* domain $\mathcal{X}$.

Smoothness

Example 5. Function $f(\mathbf{x}) = \sqrt{1 + \|\mathbf{x}\|_2^2}$ is 1-smooth w.r.t. ℓ_2-norm.

Proof:

$$\nabla f(\mathbf{x}) = \frac{\mathbf{x}}{\sqrt{\|\mathbf{x}\|_2^2 + 1}}$$

$$\Longrightarrow \nabla^2 f(\mathbf{x}) = \frac{1}{\sqrt{\|\mathbf{x}\|_2^2 + 1}} \left(I - \frac{\mathbf{x}\mathbf{x}^\top}{\|\mathbf{x}\|_2^2 + 1} \right) \preceq \frac{1}{\sqrt{\|\mathbf{x}\|_2^2 + 1}} I \preceq I \qquad \square$$

Example 6. Function $f(\mathbf{x}) = \frac{1}{2} \|\mathbf{x} - \Pi_{\mathcal{X}}[\mathbf{x}]\|^2$ is 1-smooth w.r.t. ℓ_2-norm, where $\Pi_{\mathcal{X}}[\mathbf{x}]$ denotes the Euclidean projection of $\mathbf{x}$ onto a *convex* domain $\mathcal{X}$.

Smoothness

The next lemma is an ***equivalent*** condition of smoothness.

Lemma 1 (Descent Lemma). *Let f be an L-smooth function over a given convex set $\mathcal{X}$. Then for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

Proof: $f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle \mathrm{d}t$ (calculus)

$$\Longrightarrow f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle = \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \mathrm{d}t$$

$$\text{(Cauchy-Schwarz)} \leq \int_0^1 \|\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})\| \, \|\mathbf{y} - \mathbf{x}\| \, \mathrm{d}t$$

$$\text{(smoothness)} \leq L\|\mathbf{y} - \mathbf{x}\|^2 \int_0^1 t \mathrm{d}t \leq \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2 \qquad \square$$

Smoothness

Theorem 2 (***First-order*** **Characterizations of L-smoothness).** *Let $f : \mathcal{X} \to \mathbb{R}$ be a convex function, differentiable over $\mathcal{X}$. Then the following claims are equivalent:*

(i) f *is L-smooth.*

(ii) $f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2$ *for all* $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.

(iii) $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_*^2$ *for all* $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.

(iv) $\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_*^2$ *for all* $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.

(v) $f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \geq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}) - \frac{L}{2}\lambda(1 - \lambda)\|\mathbf{x} - \mathbf{y}\|^2$ *for any* $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ *and* $\lambda \in [0, 1]$.

Smoothness

Theorem 3 (*Second-order* Characterization of L-smoothness)**.** *Let f be a twice continuously differentiable function over $\mathbb{R}^d$. Then for a given $L \geq 0$, L-smoothness w.r.t. the ℓ_p-norm ($p \in [1, \infty]$) is equivalent to*

$$\left\|\nabla^2 f(\mathbf{x})\right\|_{op,p} \leq L,$$

for any $\mathbf{x} \in \mathbb{R}^d$.

Strong Convexity

Definition 5 (Strong Convexity). A function f is σ-strongly convex if, for any $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$ and $\lambda \in [0, 1]$,

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) - \frac{\sigma}{2}\lambda(1-\lambda)\|\mathbf{x} - \mathbf{y}\|^2.$$

- Clearly, for generally convex functions, $\sigma = 0$.

Examples:

- $f(\mathbf{x}) = \|\mathbf{x}\|_p^2$ is 2-strongly-convex with respect to norm $\|\cdot\|_p$.
- Negative entropy $f(\mathbf{x}) = \sum_{i=1}^d x_i \ln x_i$ over probability distribution (i.e., $x_i \in [0, 1]$ and $\sum_{i=1}^d x_i = 1$) is 1-strongly-convex.

Strong Convexity

Theorem 3 (*First-order* Characterizations of Strong Convexity). *Let f be a proper closed and convex function. Then for a given $\sigma > 0$, the followings equal:*

(i) f *is* σ*-strongly convex.*

(ii) *For any* $\mathbf{x} \in \mathrm{dom}(\partial f), \mathbf{y} \in \mathrm{dom}(f)$ *and* $\mathbf{g} \in \partial f(\mathbf{x})$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle + \frac{\sigma}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

commonly used

(iii) *For any* $\mathbf{x}, \mathbf{y} \in \mathrm{dom}(\partial f)$, *and* $\mathbf{g}_\mathbf{x} \in \partial f(\mathbf{x}), \mathbf{g}_\mathbf{y} \in \partial f(\mathbf{y})$,

$$\langle \mathbf{g}_\mathbf{x} - \mathbf{g}_\mathbf{y}, \mathbf{x} - \mathbf{y} \rangle \geq \sigma\|\mathbf{x} - \mathbf{y}\|^2.$$

(iv) *Function* $f(\cdot) - \frac{\sigma}{2}\|\cdot\|^2$ *is convex.*

Strong Convexity

Proof: (i)→(ii)

$$f(\lambda \mathbf{y} + (1-\lambda)\mathbf{x}) \leq \lambda f(\mathbf{y}) + (1-\lambda) f(\mathbf{x}) - \frac{\sigma}{2}\lambda(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2$$

$$\Rightarrow \frac{f(\mathbf{x}+\lambda(\mathbf{y}-\mathbf{x})) - f(\mathbf{x})}{\lambda} \leq f(\mathbf{y}) - f(\mathbf{x}) - \frac{\sigma}{2}(1-\lambda)\|\mathbf{x}-\mathbf{y}\|^2 \quad \text{(rearrange)}$$

$$\Rightarrow f'(\mathbf{x};\mathbf{y}-\mathbf{x}) \triangleq \lim_{\lambda \to 1} \frac{f(\mathbf{x}+\lambda(\mathbf{y}-\mathbf{x})) - f(\mathbf{x})}{\lambda} \leq f(\mathbf{y}) - f(\mathbf{x}) - \frac{\sigma}{2}\|\mathbf{x}-\mathbf{y}\|^2$$

$f'(\mathbf{x};\mathbf{y}-\mathbf{x})$: the ***directional derivative*** of f at point $\mathbf{x}$ along direction $\mathbf{y}-\mathbf{x}$

$$\forall \mathbf{g} \in \partial f(\mathbf{x}), \quad \langle \mathbf{g}, \mathbf{y}-\mathbf{x}\rangle \leq f'(\mathbf{x};\mathbf{y}-\mathbf{x})$$

Plugging $\mathbf{g} = \nabla f(\mathbf{x})$ finishes the proof. □

Strong Convexity

> **Theorem 4.** *Let $\mathcal{X}$ be a Euclidean space. Then f is σ-strongly convex if and only if the function $f(\cdot) - \frac{\sigma}{2}\|\cdot\|^2$ is convex.*

f is "as least as convex" as a quadratic function.

Example 8. $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} + \mathbf{w}^\top\mathbf{x} + c$ is σ-strongly convex w.r.t. the ℓ_2-norm if and only if $A \succeq \sigma I$.

Proof: f is σ-strongly convex if and only if $h(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top(A - \sigma I)\mathbf{x} + \mathbf{w}^\top\mathbf{x} + c$ is convex

$$\Longrightarrow \nabla^2 h(\mathbf{x}) = A - \sigma I \succeq 0 \qquad \square$$

Strong Convexity

Theorem 5 (*Second-order* Characterization of Strong Convexity). *Let $\mathcal{X}$ be a Euclidean space. Then f is σ-strongly convex if and only if for any $\mathbf{x}, \mathbf{w} \in \mathcal{X}$,*

$$\underline{\mathbf{w}^\top \nabla^2 f(\mathbf{x}) \mathbf{w}} \geq \sigma \|\mathbf{w}\|^2 .$$

a more familiar form: $\|\mathbf{w}\|^2_{\nabla^2 f(\mathbf{x})}$

Furthermore, when using ℓ_2-norm, it is equivalent to $\nabla^2 f(\mathbf{x}) \succeq \sigma I$.

- Negative entropy $f(\mathbf{x}) = \sum_{i=1}^d x_i \ln x_i$ over probability distribution (i.e., $x_i \in [0,1]$ and $\sum_{i=1}^d x_i = 1$) is 1-strongly-convex.

Strong Convexity

Theorem 6. *Let f be a proper closed and σ-strongly convex function. Then*

- f *has a unique minimizer, denoted by* $\mathbf{x}^\star$.
- $f(\mathbf{x}) - f(\mathbf{x}^\star) \geq \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}^\star\|^2$ *for all* $\mathbf{x} \in \mathrm{dom}(f)$.

Strongly Convex and Smooth

If function f is both σ-strongly convex and L-smooth w.r.t. ℓ_2-norm, then

- $\sigma I \preccurlyeq \nabla^2 f(\mathbf{x}) \preccurlyeq LI$
- f is ***γ-well-conditioned*** where $\gamma \triangleq \sigma/L \leq 1$ is called the condition number.

Relationship

Theorem 7 (Conjugate Correspondence). *Consider the conjugate function:*

$$f^*(\mathbf{y}) \triangleq \max_{\mathbf{x} \in \mathcal{X}} \left\{ \langle \mathbf{y}, \mathbf{x} \rangle - f(\mathbf{x}) \right\}.$$

(a) If the function f is convex and $\frac{1}{\sigma}$-smooth w.r.t. the norm $\|\cdot\|$, then its conjugate f^* is σ-strongly convex w.r.t. the dual norm $\|\cdot\|_*$.

(b) If f is proper closed σ-strongly convex w.r.t. the norm $\|\cdot\|$, then f^* is $\frac{1}{\sigma}$-smooth w.r.t. the dual norm $\|\cdot\|_*$.

Summary

- **CONVEXITY**
  - **CONVEX SET**
    - Definition
    - Ball and Ellipsoid
    - Convex Hull
    - Projection
  - **CONVEX FUNCTION**
    - Definition
    - Concave Function
    - Zeroth, First and Second-order Condition
- **OPTIMALITY CONDITION**
  - Fermat's Optimality Condition
  - First-order Optimality Condition
- **CONVEX OPTIMIZATION PROBLEM**
  - Convex Optimization
  - Subgradients
  - Existence of Subgradients
  - How to Compute Subgradients
  - Why Convexity?
- **FUNCTION PROPERTIES**
  - Smoothness
  - Strong Convexity

Q & A

Thanks!