Lecture 2. Convex Optimization Basics

Advanced Optimization (Fall 2022)

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
- Convex Function
- Convex Optimization Problem
- Optimality Condition

Part 1. Convex Set

- Definition

- Ball and Ellipsoid

- Convex Hull

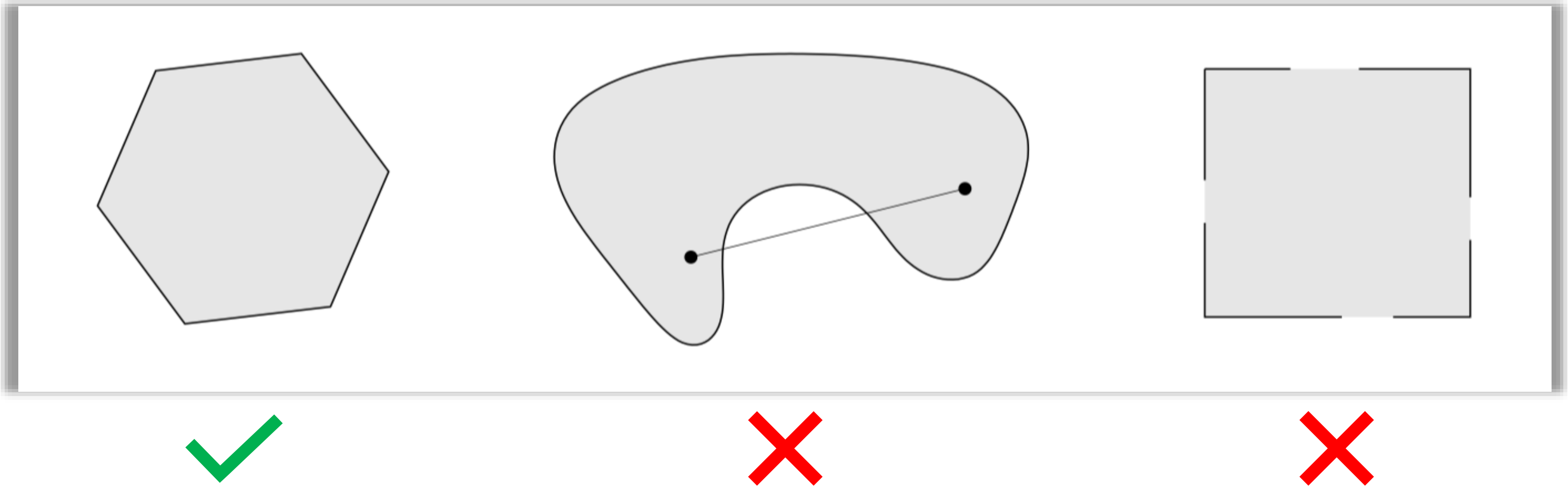
- Projection

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.,

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

Convex sets?



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

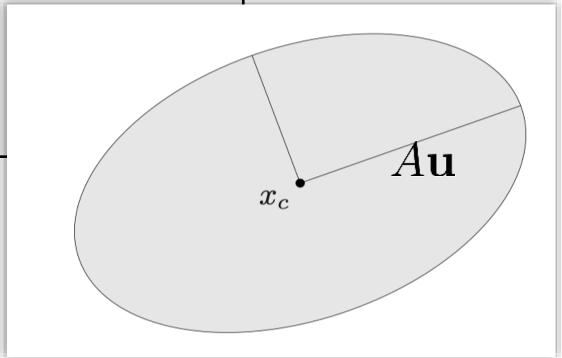
Definition 2 (Ball). A (Euclidean) ball (or just ball) in $\mathbb{R}^d$ has the form

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

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

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

where A is assumed to be symmetric and positive definite.

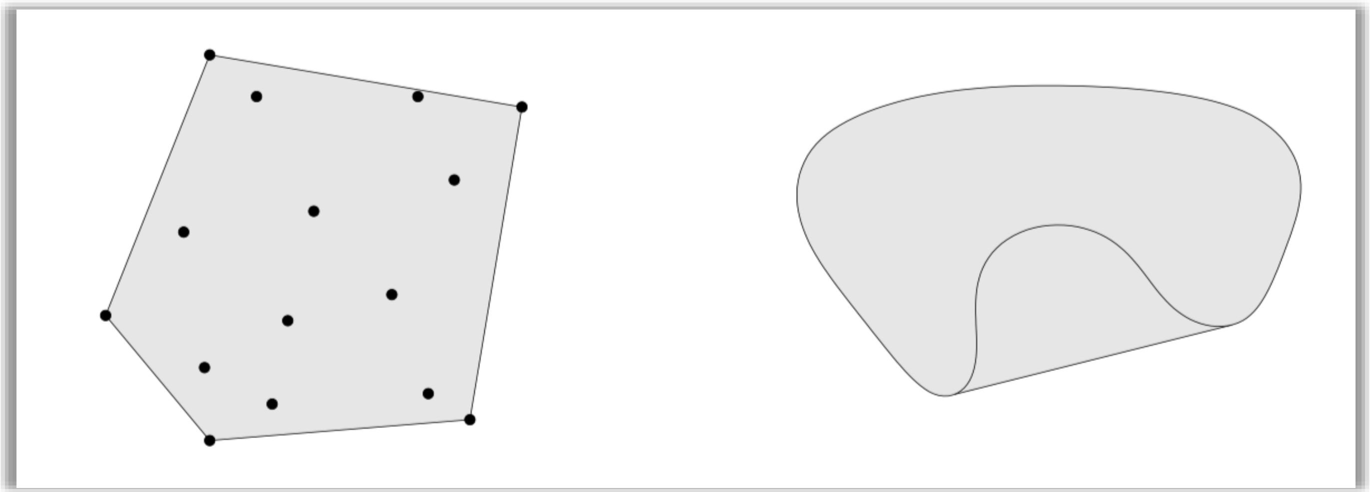


Convex Set

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:



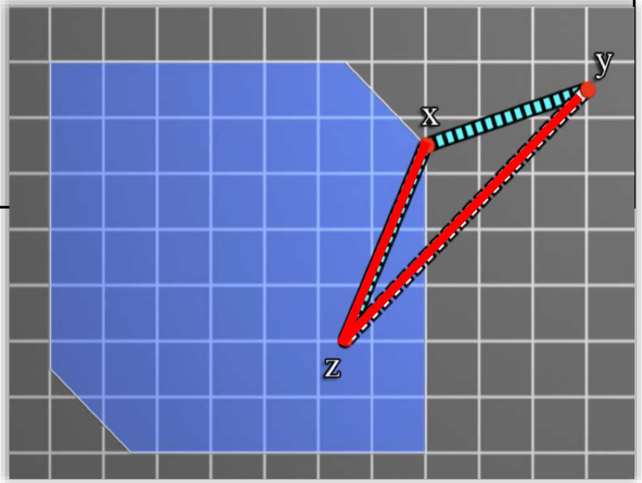
Projection onto Convex Sets

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*

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



Part 2. Convex Function

- Definition

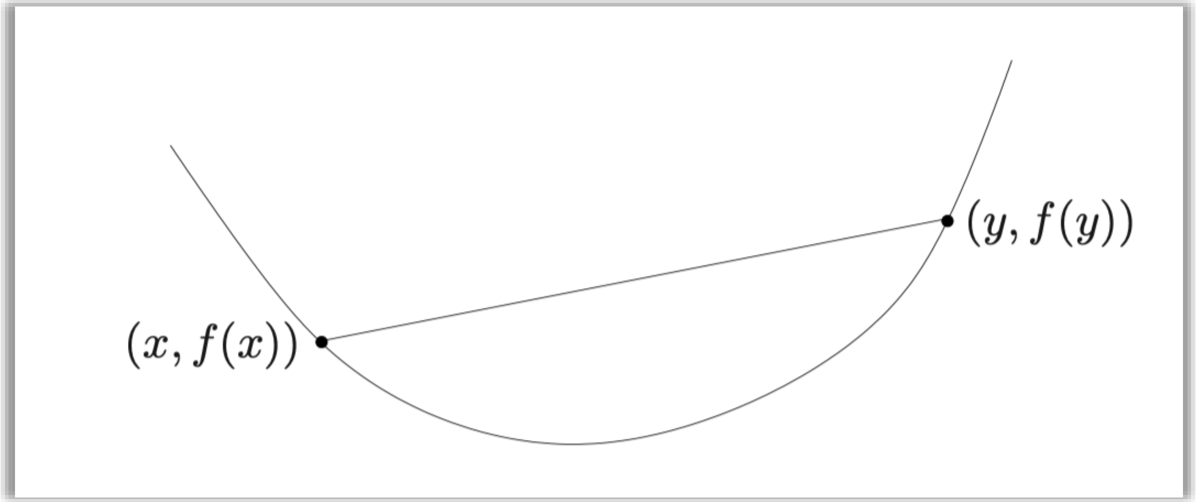
- Concave Function

- Zero-th, First and Second-order Condition

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

Concave Function

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

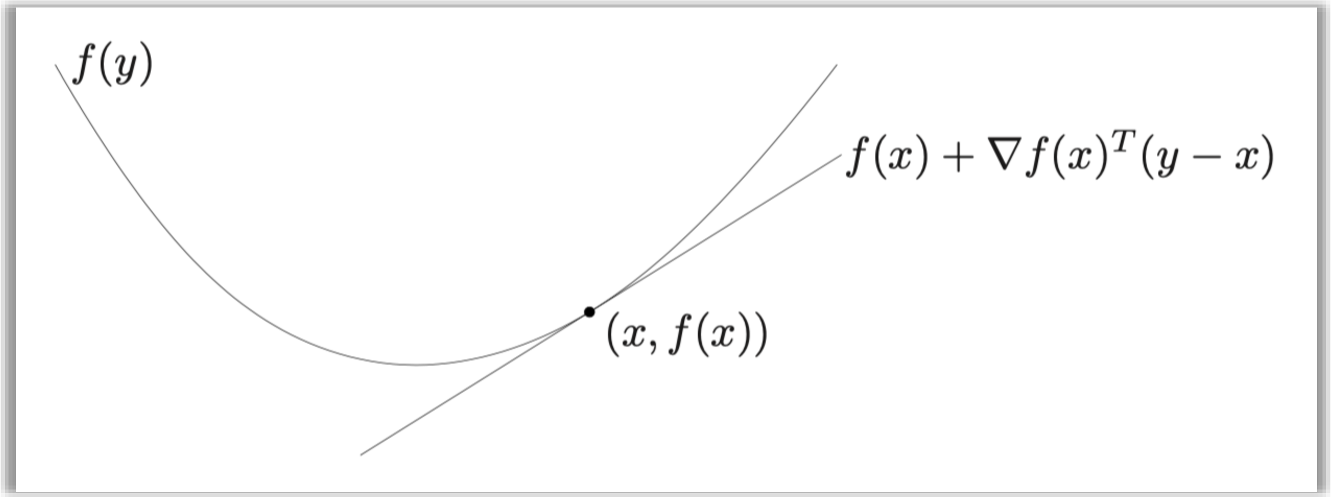
$$\forall \alpha \in [0, 1], \quad f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y}).$$

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 assume ***convex sets***.
- We focus on the ***"convex" language***, since the negative of concave functions are convex.

Convex Function



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

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

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 $\alpha \in [0, 1]$,

- Zero-th order condition: $f((1-\alpha)\mathbf{x} + \alpha\mathbf{y}) \leq (1-\alpha)f(\mathbf{x}) + \alpha f(\mathbf{y})$.
- First order condition: $f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq f(\mathbf{y})$.
- 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: $f(\mathbf{x}) = \max\{x_1, \ldots, x_n\}$.
- Log-sum-exp: $f(\mathbf{x}) = \log\left(e^{x_1} + \cdots + e^{x_n}\right)$.

Jensen's Inequality

Theorem 2 (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:

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

Part 3. Convex Optimization Problem

- Convex Optimization 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

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

Example 1 (SVM).

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

Convex Optimization Problem

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

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

Example 2 (NMF decomposition).

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

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

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

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 $\partial f(\mathbf{x})$,

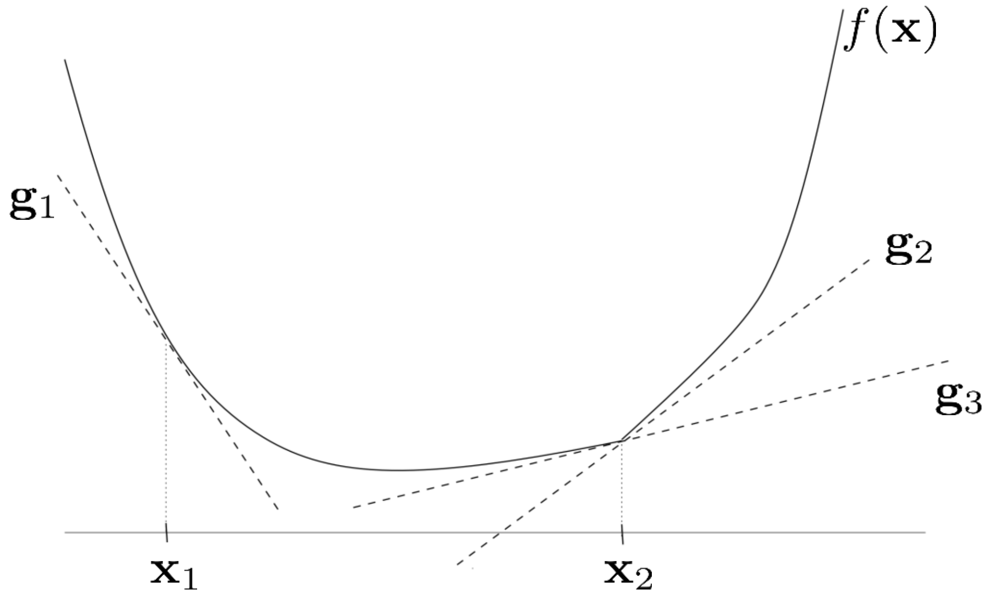
$$\partial f(\mathbf{x}) \triangleq \{\mathbf{g} \in \mathbb{R}^d \mid f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle, \text{ for all } \mathbf{y} \in \mathbb{R}^d\}.$$

Subgradient

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

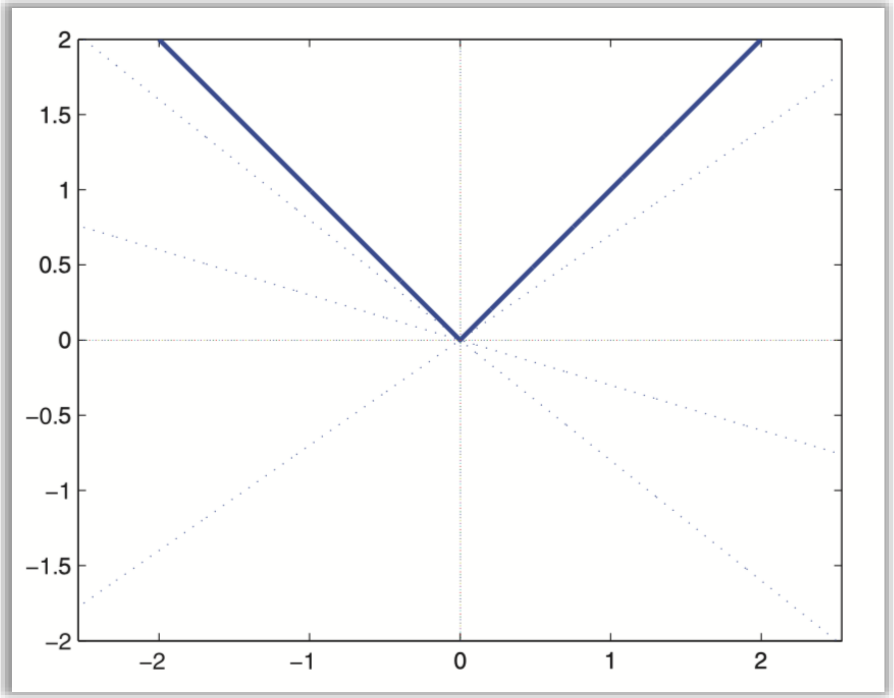
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle, \text{ for all } \mathbf{y} \in \mathbb{R}^d.$$

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.



Subgradient

Example 3. The subdifferential of ℓ_2-norm $f(\mathbf{x}) = \|\mathbf{x}\|_2$ at $\mathbf{x} = \mathbf{0}$ is the norm unit ball, i.e., $\partial f(\mathbf{0}) = \{\mathbf{g} \mid \|\mathbf{g}\|_2 \leq 1\}$.



an illustration for 1-dim case

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

Subgradient

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

called normal cone

Subgradient

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

> **Theorem 3.** *Let* $f : \mathcal{X} \to \mathbb{R}$ *be a convex function. Suppose that* $\mathcal{X} \subseteq \operatorname{int}(\operatorname{dom} f)$. *Consider the following two claims:*
>
> *(i) Lipschitzness:* $|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|$ *for any* $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.
>
> *(ii) Bounded subgradient:* $\|\mathbf{g}\| \leq L$ *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 4.** *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).

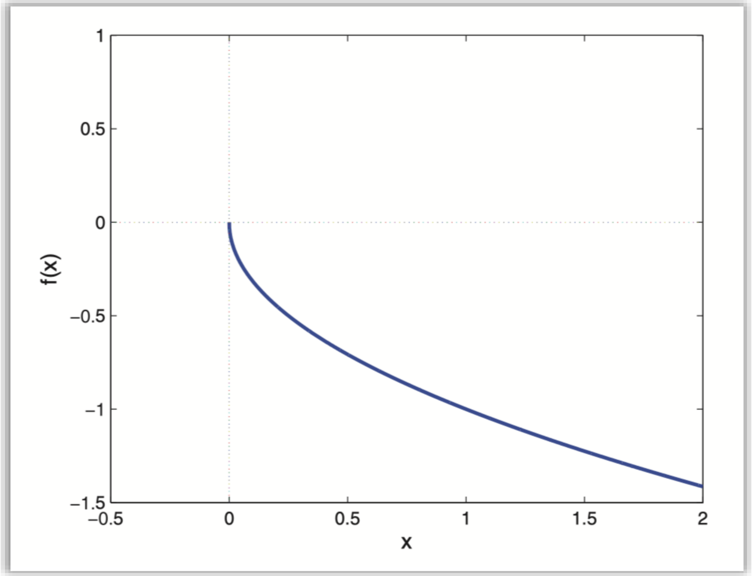
Existence of Subgradient

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

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

Why Convexity?

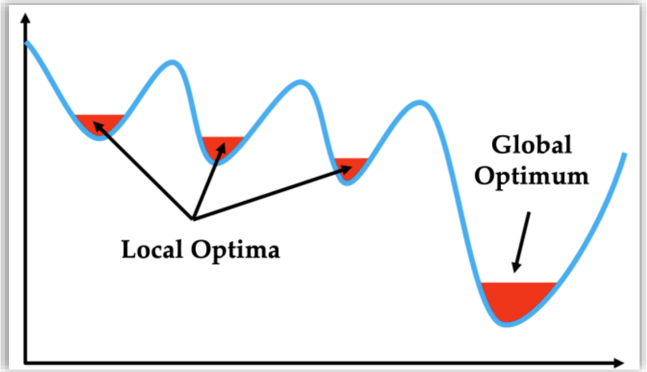
- **Local to Global Phenomenon**

  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 7.** *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)

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

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

Part 4. Optimality Condition

- Fermat's Optimality Condition
- First-order Optimality Condition
- Fritz-John Optimality Condition
- KKT Condition

Fermat's Optimality Condition

- ***Unconstrained*** case

Theorem 8 (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
$$f(\mathbf{x}) \geq f(\mathbf{x}^\star)$$
$$f(\mathbf{x}) \geq f(\mathbf{x}^\star) + \langle \mathbf{g}, \mathbf{x} - \mathbf{x}^\star \rangle, \mathbf{g} \in \partial f(\mathbf{x}^\star)$$
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 β that satisfies

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

Solving the optimization problem:

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

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

Example

- ***Proof of median***

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

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

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***

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

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

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

Combining the two cases finishes the proof. $\square$

First-order Optimality Condition

- ***Constrained*** Case

> **Theorem 9** (First-order Optimality Condition). *Let f be convex and $\mathcal{X}$ a closed convex set on which f is differentiable. Then $\mathbf{x}^\star \in \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$ 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 9 (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 $N_{\mathcal{X}}(\mathbf{x}) = \partial f(\mathbf{x}) = \{\mathbf{g} \mid \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \leq 0, \forall \mathbf{y} \in \mathcal{X}\}$.

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

Set Addition: elementwise sum

First-order Optimality Condition

- ***Constrained*** Case

> **Theorem 9** (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 10. *Consider the minimization problem*

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

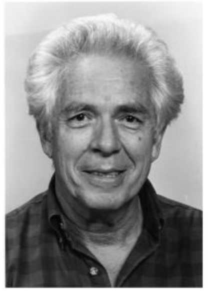
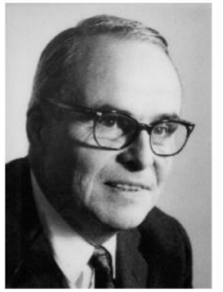
where $f, g_1, g_2, \ldots, g_m$ are real-valued convex functions.

1. *Let $\mathbf{x}^\star$ be an optimal solution of* (1)*, and assume that Slater's condition is satisfied. Then there exist $\lambda_1, \ldots, \lambda_m \geq 0$ 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 $\lambda_1, \lambda_2, \ldots, \lambda_m \geq 0$, 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.

Proof (sketch) of KKT Conditions

- We start by the ***necessity*** of KKT conditions, i.e., *suppose a point is optimal, what kind of conditions it should satisfy*.

Lemma 1. *Let* $f, g_1, g_2, \ldots, g_m : \mathcal{X} \to \mathbb{R}$ *be real-valued functions. Consider the problem*

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

Assume that the minimum value of problem (1) is finite and equal to $f^\star$ *and define*

$$F(\mathbf{x}) \triangleq \max\left\{f(\mathbf{x}) - f^\star, g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_m(\mathbf{x})\right\}.$$

Then the optimal set of problem (1) is the same as the set of minimizers of F.

another reduction from constrained opt. to unconstrained one

Proof (sketch) of KKT Conditions

Lemma 1. *Let $f, g_1, g_2, \ldots, g_m : \mathcal{X} \to \mathbb{R}$ be real-valued functions. Consider the problem*

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

Assume that the minimum value of problem (1) is finite and equal to $f^\star$ and define

$$F(\mathbf{x}) \triangleq \max\left\{f(\mathbf{x}) - f^\star, g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_m(\mathbf{x})\right\}.$$

Then the optimal set of problem (1) is the same as the set of minimizers of F.

Intuition: the optimizer $\mathbf{x}^\star$ of F will make each function inside F as small as possible, i.e., $f(\mathbf{x}^\star) \le f^\star$ and $g_i(\mathbf{x}^\star) \le 0$ for $i \in [m]$.

Proof: Denote by $\mathcal{S}^\star$ the set of optimizers of Problem (1)

Case 1: $\mathbf{x} \notin \mathcal{S}^\star$. One of the two cases must exist, which both lead to $F(\mathbf{x}) > 0$:

(1.1) $\mathbf{x}$ is not in the feasible domain, i.e., $\exists i \in [m],\ g_i(\mathbf{x}) > 0 \Rightarrow F(\mathbf{x}) > 0$.

(1.2) $\mathbf{x}$ is in the feasible domain but suboptimal, i.e., $f(\mathbf{x}) > f^\star \Rightarrow F(\mathbf{x}) > 0$.

Proof (sketch) of KKT Conditions

Lemma 1. *Let $f, g_1, g_2, \ldots, g_m : \mathcal{X} \to \mathbb{R}$ be real-valued functions. Consider the problem*

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

Assume that the minimum value of problem (1) is finite and equal to $f^\star$ and define

$$F(\mathbf{x}) \triangleq \max \{f(\mathbf{x}) - f^\star, g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_m(\mathbf{x})\}.$$

Then the optimal set of problem (1) is the same as the set of minimizers of F.

Intuition: the optimizer $\mathbf{x}^\star$ of F will make each function inside F as small as possible, i.e., $f(\mathbf{x}^\star) \leq f^\star$ and $g_i(\mathbf{x}^\star) \leq 0$ for $i \in [m]$.

Proof: Denote by $\mathcal{S}^\star$ the set of optimizers of Problem (1)

Case 2: $\mathbf{x} \in \mathcal{S}^\star$, which leads to $F(\mathbf{x}) = 0$ obviously.

$$\Longrightarrow \quad \mathcal{S}^\star = \operatorname*{arg\,min}_{\mathbf{x} \in \mathcal{X}} F(\mathbf{x}) = \{\mathbf{x} \mid F(\mathbf{x}) = 0\}. \qquad \square$$

Fritz John Optimality Conditions

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & g_i(\mathbf{x}) \le 0, \;\; i \in [m] \end{aligned} \quad \Longleftrightarrow \quad \begin{aligned} \min \quad & F(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \mathbb{R}^d \end{aligned}$$

Fritz John (1910-1994)

Theorem 5 (Fritz John Necessary Optimality Conditions). *Consider the minimization problem* $\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x})$. *Let* $\mathbf{x}^\star$ *be an optimal solution. Then there exist* $\lambda_0, \lambda_1, \ldots, \lambda_m \ge 0$, *not all zeros, such that*

$$\mathbf{0} \in \lambda_0 \partial f(\mathbf{x}^\star) + \sum_{i=1}^{m} \lambda_i \partial g_i(\mathbf{x}^\star),$$

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

Fritz John Optimality Conditions

Theorem 5 (Fritz John Necessary Optimality Conditions). *Consider the minimization problem* $\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x})$. *Let* $\mathbf{x}^\star$ *be an optimal solution. Then there exist* $\lambda_0, \lambda_1, \ldots, \lambda_m \geq 0$, *not all zeros, such that*

$$\mathbf{0} \in \lambda_0 \partial f(\mathbf{x}^\star) + \sum_{i=1}^{m} \lambda_i \partial g_i(\mathbf{x}^\star),$$

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

Proof: Using Fermat's optimality condition, the optimizer $\mathbf{x}^\star$ satisfies $\mathbf{0} \in \partial F(\mathbf{x}^\star)$.

$$F(\mathbf{x}) \triangleq \max\left\{g_0(\mathbf{x}), g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_m(\mathbf{x})\right\}, g_0(\mathbf{x}) = f(\mathbf{x}) - f^\star.$$

⇨ ***Remaining question***: computing the subgradient of a maximum of functions

Details can be found in Amir Beck's book (Chapter 3, Theorem 3.50)

Subdifferential of a Maximum of Functions

Lemma 2. *Let* $f_1, f_2, \ldots, f_m$ *be proper convex functions, and define* $f(\mathbf{x}) = \max\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})\}$. *Let* $\mathbf{x} \in \bigcap_{i=1}^{m} \operatorname{int}(\operatorname{dom}(f_i))$. *Then*

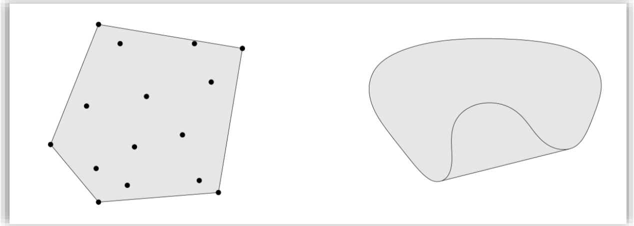
$$\partial f(\mathbf{x}) = \operatorname{conv}\left(\cup_{i \in I(\mathbf{x})} \partial f_i(\mathbf{x})\right),$$

where $I(\mathbf{x}) = \{i \in [m] \mid f_i(\mathbf{x}) = f(\mathbf{x})\}$.

conv denotes the ***convex hull***:

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

$I(\mathbf{x})$ denotes the subset of $\{f_1, \ldots, f_m\}$ that are max at $\mathbf{x}$.

Fritz John Optimality Conditions

Theorem 5 (Fritz John Necessary Optimality Conditions)**.** *Consider the minimization problem* $\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x})$*. Let* $\mathbf{x}^\star$ *be an optimal solution. Then there exist* $\lambda_0, \lambda_1, \ldots, \lambda_m \geq 0$*, not all zeros, such that*

$$\mathbf{0} \in \lambda_0 \partial f(\mathbf{x}^\star) + \sum_{i=1}^{m} \lambda_i \partial g_i(\mathbf{x}^\star),$$

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

Proof: $F(\mathbf{x}) \triangleq \max\left\{g_0(\mathbf{x}), g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_m(\mathbf{x})\right\}, g_0(\mathbf{x}) = f(\mathbf{x}) - f^\star.$

$\Rightarrow$ $\partial F(\mathbf{x}^\star) = \operatorname{conv}((\cup_{i\in I(\mathbf{x}^\star)} \partial g_i(\mathbf{x}^\star))$, where $I(\mathbf{x}^\star) = \{i \in [m] \mid g_i(\mathbf{x}^\star) = F(\mathbf{x}^\star) = 0\}$.

$\Rightarrow$ there exists $\lambda_i \geq 0$ for $i \in I(\mathbf{x}^\star)$ such that $\sum_{i\in I(\mathbf{x}^\star)} \lambda_i = 1$ and

$$\mathbf{0} \in \sum_{i\in I(\mathbf{x}^\star)} \lambda_i \partial g_i(\mathbf{x}^\star).$$

Fritz John Optimality Conditions

Theorem 5 (Fritz John Necessary Optimality Conditions)**.** *Consider the minimization problem* $\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x})$. *Let* $\mathbf{x}^\star$ *be an optimal solution. Then there exist* $\lambda_0, \lambda_1, \ldots, \lambda_m \geq 0$, *not all zeros, such that*

$$\mathbf{0} \in \lambda_0 \partial f(\mathbf{x}^\star) + \sum_{i=1}^{m} \lambda_i \partial g_i(\mathbf{x}^\star),$$

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

Proof: $\mathbf{0} \in \sum_{i\in I(\mathbf{x}^\star)} \lambda_i \partial g_i(\mathbf{x}^\star) = \lambda_0 \partial f(\mathbf{x}^\star) + \sum_{i\in I(\mathbf{x}^\star)\setminus\{0\}} \lambda_i \partial g_i(\mathbf{x}^\star)$ (plug $g_0(\mathbf{x}) = f(\mathbf{x}) - f^\star$ back)

⇨ Define $\lambda_i = 0$ for $i \notin I(\mathbf{x}^\star)$, $\lambda_i g_i(\mathbf{x}^\star) = 0$ holds in two cases:

- Case 1: $i \in I(\mathbf{x}^\star)$. $I(\mathbf{x}^\star) = \{i \in [m] \mid g_i(\mathbf{x}^\star) = F(\mathbf{x}^\star) = 0\} \Rightarrow \lambda_i g_i(\mathbf{x}^\star) = 0$.
- Case 2: $i \notin I(\mathbf{x}^\star)$. $\lambda_i = 0 \Rightarrow \lambda_i g_i(\mathbf{x}^\star) = 0$. □

Proof (sketch) of KKT Conditions

- To prove the ***necessity*** direction of KKT conditions, besides ***Fritz John conditions***, we need the ***Slater's condition***:

> There exists $\bar{\mathbf{x}} \in \mathbb{R}^d$ for which $g_i(\bar{\mathbf{x}}) < 0, \quad i \in [m]$.

Necessity:

$\mathbf{x}^\star$ is a optimizer $\Rightarrow$ ***Fritz John conditions*** + ***Slater's condition*** = ***KKT conditions***

using Slater's condition to show that $\lambda_0 \neq 0$

Sufficiency:

KKT conditions $\Rightarrow$ $\mathbf{x}^\star$ is a optimizer *(by a self-contained proof, omitted here)*

Karush–Kuhn–Tucker (KKT) Conditions

Theorem 10. *Consider the minimization problem*

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

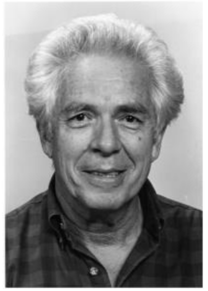
where $f, g_1, g_2, \ldots, g_m$ are real-valued convex functions.

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

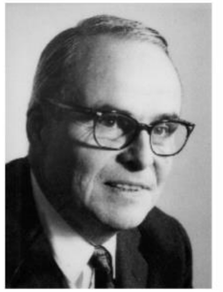
$$\mathbf{0} \in \partial f\left(\mathbf{x}^\star\right) + \sum_{i=1}^{m} \lambda_i \partial g_i\left(\mathbf{x}^\star\right) \tag{2}$$

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

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



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.

Summary

- **CONVEX SET**
  - Definition
  - Ball and Ellipsoid
  - Convex Hull
  - Projection
- **CONVEX FUNCTION**
  - Definition
  - Concave Function
  - Zero-th, First and Second-order Condition
- **CONVEX OPTIMIZATION PROBLEM**
  - Convex Optimization
  - Subgradients
  - Existence of Subgradients
  - How to Compute Subgradients
  - Why Convexity?
- **OPTIMALITY CONDITION**
  - Fermat's Optimality Condition
  - First-order Optimality Condition
  - Fritz John Optimality Conditions
  - Karush–Kuhn–Tucker (KKT) Conditions

Q & A

Thanks!