



# Lecture 2. Convex Problems

Advanced Optimization (Fall 2026)

**Peng Zhao**

[zhaop@lamda.nju.edu.cn](mailto:zhaop@lamda.nju.edu.cn)

Nanjing University

# Outline

- Performance Measure
- Subgradient
- Optimality Condition
- Function Properties
- Gradient Descent Template

# Convex Optimization Problem

- We adopt a minimization language

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

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

# Part 1. Performance Measure

- Iterated Optimization

$$\mathbf{x}_1 \rightarrow \mathbf{x}_2 \rightarrow \cdots \mathbf{x}_t \rightarrow \mathbf{x}_{t+1} \rightarrow \cdots \mathbf{x}_T$$

To output a sequence  $\{\bar{\mathbf{x}}_t\}_{t=1}^T$  such that  $\bar{\mathbf{x}}_t$  **approximates**  $\mathbf{x}^*$  when  $t$  goes larger.

where  $\{\bar{\mathbf{x}}_t\}_{t=1}^T$  can be *statistics* of the original sequence  $\{\mathbf{x}_t\}_{t=1}^T$ ,

and  $\mathbf{x}^* \arg \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$  denotes the optimal solution.

- Measure

- Function-value level:  $f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) \leq \varepsilon(T)$

- Optimizer-value level:  $\|\bar{\mathbf{x}}_T - \mathbf{x}^*\| \leq \varepsilon(T)$

and  $\varepsilon(T)$  is the *approximation error* and is a function of iterations  $T$ .

# Performance Measure

- In general, there are two performance measures (essentially same).

**Convergence:**  $f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) \leq \varepsilon(T)$ ,

- **Qualitatively:**  $\varepsilon(T) \rightarrow 0$  when  $T \rightarrow \infty$
- **Quantitatively:**  $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right) / \mathcal{O}\left(\frac{1}{T}\right) / \mathcal{O}\left(\frac{1}{T^2}\right) / \mathcal{O}\left(\frac{1}{e^T}\right) / \dots$

**Complexity:**

- **Definition:** number of iterations required to achieve  $f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) \leq \varepsilon$ .
- **Quantitatively:**  $\mathcal{O}\left(\frac{1}{\varepsilon^2}\right) / \mathcal{O}\left(\frac{1}{\varepsilon}\right) / \mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}}\right) / \mathcal{O}\left(\log \frac{1}{\varepsilon}\right) / \dots$

corresponds to  $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right) / \mathcal{O}\left(\frac{1}{T}\right) / \mathcal{O}\left(\frac{1}{T^2}\right) / \mathcal{O}\left(\frac{1}{e^T}\right) / \dots$

# More Concretely

## Iterated Complexity *vs* Computational Complexity

- Computational complexity requires further consideration of the per-round runtime (i.e., the number of elementary operations that the algorithm needs to do); like FLOPS.
- Examples
  - first-order method ( $\nabla f(\mathbf{x}_t)$ ) *vs* second-order method ( $\nabla^2 f(\mathbf{x}_t)$ );
  - operations: Euclidean ( $\Pi_{\mathcal{X}}[\mathbf{y}]$ ) *vs* Mahalanobis ( $\Pi_{\mathcal{X}}^M[\mathbf{y}]$ ) projection

# Key Factors

- Consider the function-value level:

$$f(\bar{\mathbf{x}}_T) - f(\mathbf{x}^*) \leq \varepsilon(T)$$

where  $\{\bar{\mathbf{x}}_t\}_{t=1}^T$  can be *statistics* of the original sequence  $\{\mathbf{x}_t\}_{t=1}^T$ ,

- There are key quantities to consider
  - **update:**  $\mathbf{x}_t$  to  $\mathbf{x}_{t+1}$   $\rightarrow$  (sub)gradient
  - **comparator:**  $\mathbf{x}^*$   $\rightarrow$  optimality condition
  - **function:**  $f$   $\rightarrow$  function property

# Part 2. Subgradient

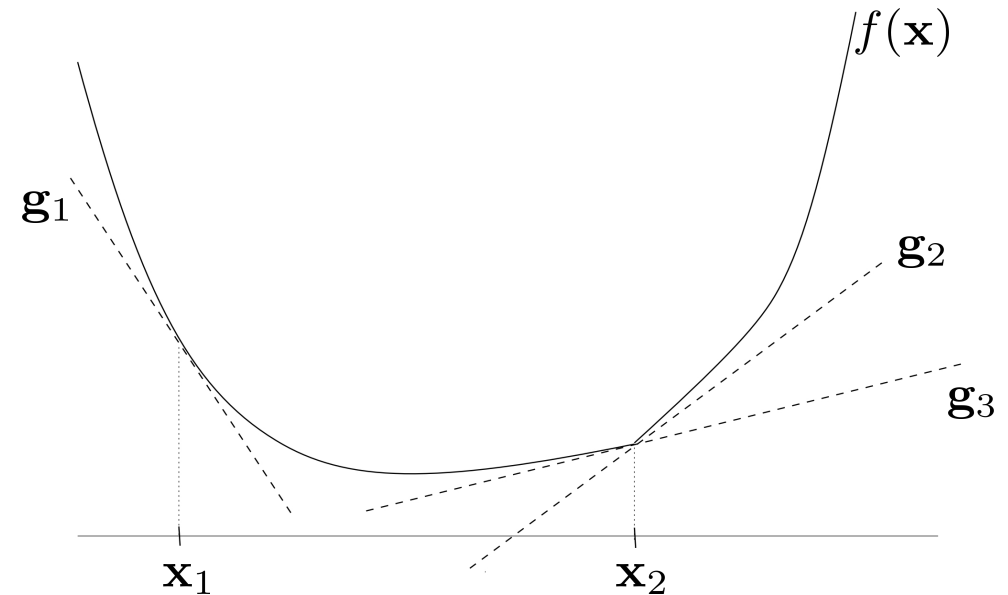
- Subgradient
- Subdifferential
- Existence and Calculation

# Subgradient

**Definition 1** (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 1** (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

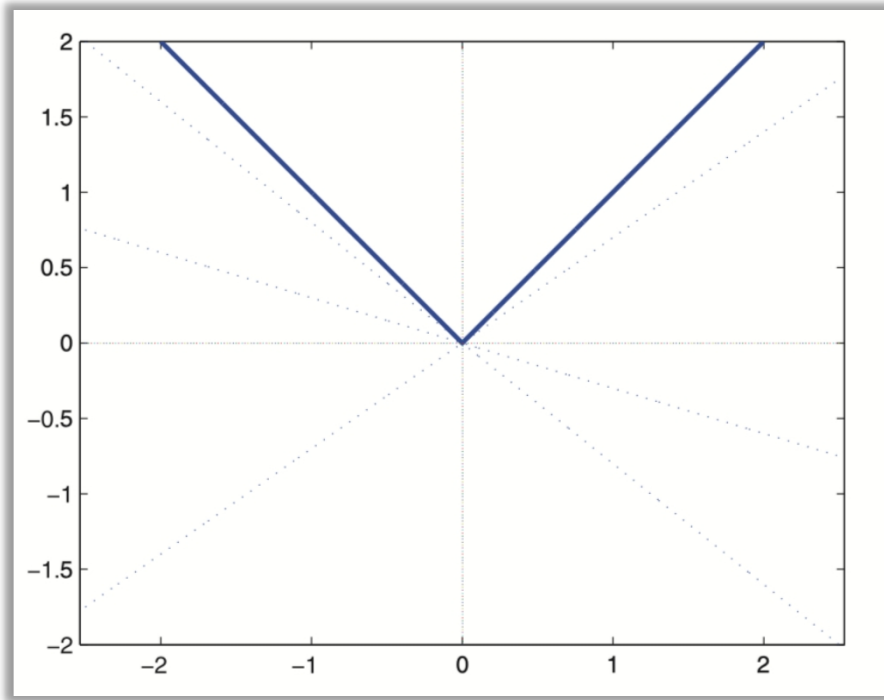
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle, \text{ for all } \mathbf{y} \in \mathbb{R}^d.$$

**Definition 2** (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 1.** 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 1.** The subdifferential of  $f(\mathbf{x}) = \|\mathbf{x}\|$  at  $\mathbf{x} = \mathbf{0}$  is the dual norm unit ball, i.e.,  $\partial f(\mathbf{0}) = \{\mathbf{g} \mid \|\mathbf{g}\|_* \leq 1\}$ .

*Proof:*

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

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

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

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

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

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

□

# Subgradient and Subdifferential

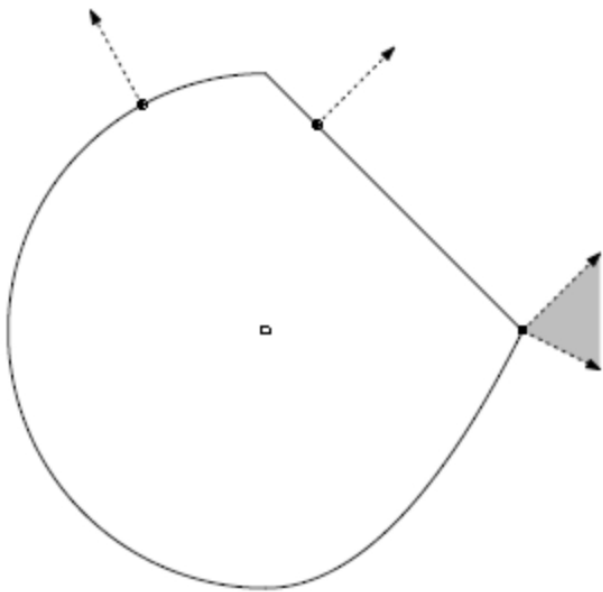
**Example 2.** For indicator function  $f(\mathbf{x}) = \delta_{\mathcal{X}}(\mathbf{x})$ , its subdifferential at any point  $\mathbf{x} \in \mathcal{X}$  is  $N_{\mathcal{X}}(\mathbf{x}) = \partial f(\mathbf{x}) = \{\mathbf{g} \mid \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \leq 0, \forall \mathbf{y} \in \mathcal{X}\}$ . *called normal cone*

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

By definition,  $\mathbf{g} \in \partial \delta_{\mathcal{X}}(\mathbf{x})$  if and only if  $\delta_{\mathcal{X}}(\mathbf{y}) \geq \delta_{\mathcal{X}}(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle$  for any  $\mathbf{y}$ .

If  $\mathbf{x} \in \mathcal{X}$ , then  $\delta_{\mathcal{X}}(\mathbf{x}) = 0$ .

- For  $\mathbf{y} \notin \mathcal{X}$  the inequality is trivial ( $+\infty \geq \dots$ ).
- For  $\mathbf{y} \in \mathcal{X}$  it becomes  $0 \geq \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle$ , i.e.  $\langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \leq 0$  for all  $\mathbf{y} \in \mathcal{X}$ . That's exactly the normal cone's definition.  $\square$



# Existence of Subgradient

- *Existence of subgradients* implies *convexity*.

**Theorem 1.** 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).

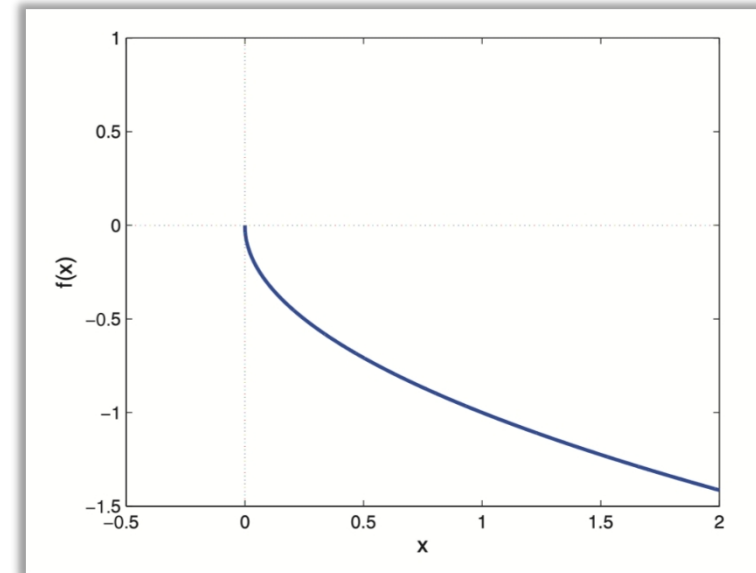
# Existence of Subgradient

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

**Example 3.** Consider function  $f : \mathbb{R} \rightarrow (-\infty, \infty]$  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

- Convexity *doesn't* always imply existence of subgradients.
- Nevertheless, if we only care about the *interior* of feasible domain, convexity does imply the *existence* of subgradients.

**Theorem 2.** *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 3.** *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 4.** The subdifferential of  $\ell_2$ -norm  $f(\mathbf{x}) = \|\mathbf{x}\|_2$  is

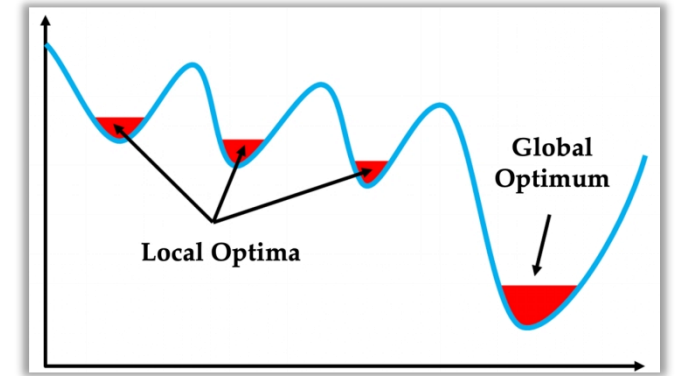
$$\partial f(\mathbf{x}) = \begin{cases} \left\{ \frac{\mathbf{x}}{\|\mathbf{x}\|_2} \right\}, & \mathbf{x} \neq \mathbf{0} \text{ (gradient of norm)} \\ \{\mathbf{g} \mid \|\mathbf{g}\|_2 \leq 1\}, & \mathbf{x} = \mathbf{0} \text{ (discussed earlier)} \end{cases}$$

*Proof can be found in Example 3.34 of Amir Beck's book.*

# Why Convexity?

- **Local to Global Phenomenon**

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



**Theorem 4.** *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  $\gamma$  small enough, for any  $\mathbf{y}$ ,

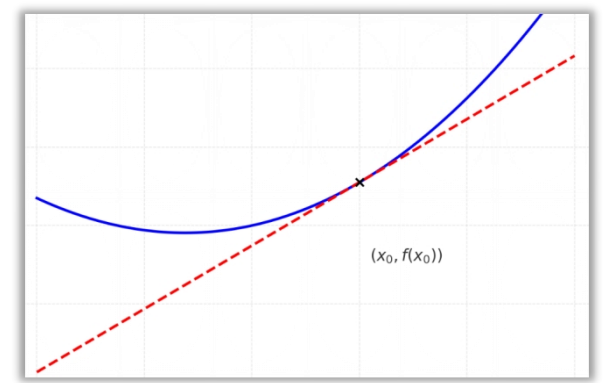
$$\underbrace{f(\mathbf{x})}_{\text{(local minima)}} \leq f((1 - \gamma)\mathbf{x} + \gamma\mathbf{y}) \leq (1 - \gamma)f(\mathbf{x}) + \gamma f(\mathbf{y}),$$

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

# Why Convexity?

- **Local to Global Phenomenon - II**

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



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

- **Local:** the gradient  $\nabla f(\mathbf{x})$  is computed using only infinitesimal/**local** information of  $f(\cdot)$  at  $\mathbf{x}$ ;
- **Global:** the subdifferential  $\partial f(\mathbf{x})$  gives **global** information in the form of a linear lower bound on the entire function.

# Part 3. Optimality Condition

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

# Why Optimality Condition?

- Given a point, can you verify if it is optimal?
- ➔ Optimality condition; we start from the simplest *unconstrained* case.

**Theorem 5** (Fermat's Optimality Condition). *Let  $f : \mathbb{R}^d \rightarrow (-\infty, \infty]$  be a proper convex function. Then*

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

*if and only if*

$$\mathbf{0} \in \partial f(\mathbf{x}^*).$$

**Proof:** Simply combining  $f(\mathbf{x}) \geq f(\mathbf{x}^*)$

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \langle \mathbf{g}, \mathbf{x} - \mathbf{x}^* \rangle, \mathbf{g} \in \partial f(\mathbf{x}^*) \quad \square$$

# First-order Optimality Condition

- *Constrained* Case

**Theorem 6** (First-order Optimality Condition). *Let  $f$  be convex and  $\mathcal{X}$  a closed convex set on which  $f$  is differentiable. Then  $\mathbf{x}^* \in \arg \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$  if and only if there exists  $\mathbf{g} \in \partial f(\mathbf{x}^*)$  such that*

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

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

⇒ deploying the Fermat's optimality condition on the unconstrained “surrogate” objective

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

*Proof can be found in Theorem 3.67 of Amir Beck's book.*

# First-order Optimality Condition

- *Constrained* Case

**Theorem 6** (First-order Optimality Condition). *Let  $f$  be convex and  $\mathcal{X}$  a closed convex set on which  $f$  is differentiable. Then  $\mathbf{x}^* \in \arg \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$  if and only if there exists  $\mathbf{g} \in \partial f(\mathbf{x}^*)$  such that*

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

**Example 2.** For indicator function  $f(\mathbf{x}) = \delta_{\mathcal{X}}(\mathbf{x})$ , its subdifferential at any point  $\mathbf{x} \in \mathcal{X}$  is  $N_{\mathcal{X}}(\mathbf{x}) = \partial f(\mathbf{x}) = \{\mathbf{g} \mid \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle \leq 0, \forall \mathbf{y} \in \mathcal{X}\}$ .

$$\Rightarrow \partial h(\mathbf{x}) = \partial f(\mathbf{x}) + N_{\mathcal{X}}(\mathbf{x}) \quad \text{Set Addition: elementwise sum}$$

# First-order Optimality Condition

- *Constrained* Case

**Theorem 6** (First-order Optimality Condition). *Let  $f$  be convex and  $\mathcal{X}$  a closed convex set on which  $f$  is differentiable. Then  $\mathbf{x}^* \in \arg \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$  if and only if there exists  $\mathbf{g} \in \partial f(\mathbf{x}^*)$  such that*

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

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

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

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

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

□

# Karush–Kuhn–Tucker (KKT) Conditions

**Theorem 7.** Consider the minimization problem

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

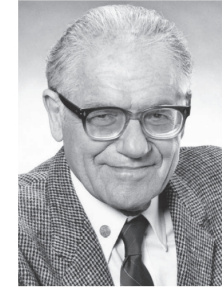
where  $f, g_1, g_2, \dots, g_m$  are real-valued convex functions.

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

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

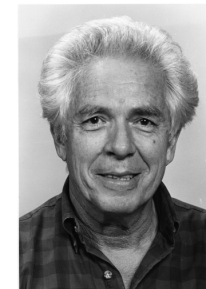
$$\lambda_i g_i(\mathbf{x}^*) = 0, \quad i \in [m]. \tag{3}$$

2. **Sufficient condition:** If a point  $\mathbf{x}^*$  satisfies conditions (2) and (3) for some  $\lambda_1, \lambda_2, \dots, \lambda_m \geq 0$ , then it is an optimal solution of problem (1).

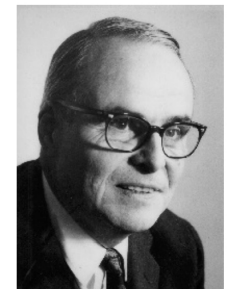


**William Karush**  
1917-1997

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



**Harold Kuhn**  
1925-2014



**Albert Tucker**  
1905-1995

*Published conditions in 1951.*

# Proof Sketch

(1) Inequalities  $\mathcal{X} = \{\mathbf{x} \mid g_i(\mathbf{x}) \leq 0\}$ : At  $\mathbf{x}$ , let  $\mathcal{I}(\mathbf{x}) = \{i \mid g_i(\mathbf{x}) = 0\}$  be the active set. Then

$$N_{\mathcal{X}}(\mathbf{x}) = \left\{ \sum_{i \in \mathcal{I}(\mathbf{x})} \lambda_i \nabla g_i(\mathbf{x}) \mid \lambda_i \geq 0 \right\}.$$

(Inactive constraints don't contribute)

(2) Equalities  $\mathcal{X} = \{\mathbf{x} \mid h_j(\mathbf{x}) = 0\}$  (affine manifold):

$$N_{\mathcal{X}}(\mathbf{x}) = \left\{ \sum_j \mu_j \nabla h_j(\mathbf{x}) \mid \mu_j \in \mathbb{R} \right\}.$$

(3) Mixed case  $g_i(\mathbf{x}) \leq 0, h_j(\mathbf{x}) = 0$ : combining the two gives

$$N_{\mathcal{X}}(\mathbf{x}) = \left\{ \sum_{i \in \mathcal{I}(\mathbf{x})} \lambda_i \nabla g_i(\mathbf{x}) + \sum_j \mu_j \nabla h_j(\mathbf{x}) \mid \lambda_i \geq 0 \right\}.$$

# Proof Sketch

$$N_{\mathcal{X}}(\mathbf{x}) = \left\{ \sum_{i \in \mathcal{I}(\mathbf{x})} \lambda_i \nabla g_i(\mathbf{x}) + \sum_j \mu_j \nabla h_j(\mathbf{x}) \mid \lambda_i \geq 0 \right\}.$$

Plugging  $N_{\mathcal{X}}(\mathbf{x})$  into the condition

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

yields the KKT stationarity together with primal feasibility, dual feasibility (  $\lambda_i \geq 0$  )  
and complementary slackness (  $\lambda_i g_i(\mathbf{x}) = 0$  ).

□

# Understanding KKT Conditions

- On the one hand, KKT conditions depict properties of the optimization solution (consider the dual form and interpretation in SVM).

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

$$\mathbf{0} \in \partial f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i \partial g_i(\mathbf{x}^*)$$
$$\lambda_i g_i(\mathbf{x}^*) = 0, \quad i \in [m].$$

- On the other hand, many optimization methods can be thought of as iterative approximations to solve the KKT conditions.

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

# Part 4. Function Properties

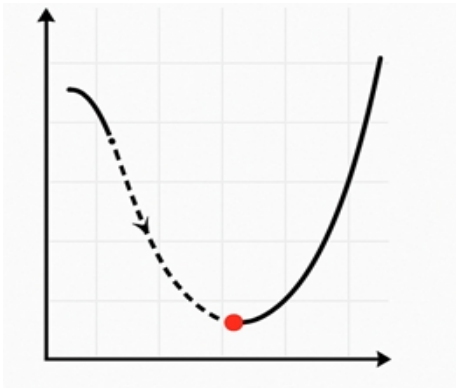
- Smoothness
- Strong Convexity

# Function

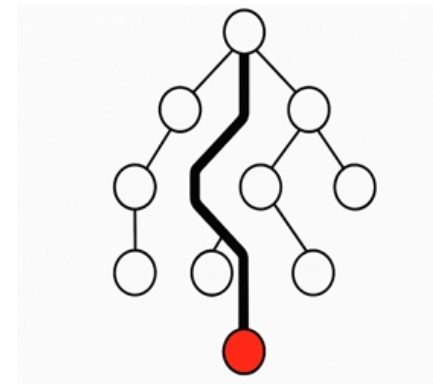
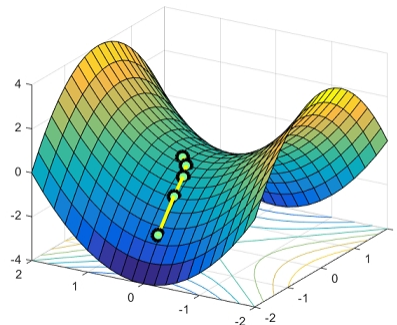
- Function mapping  $f : \text{dom } f \subseteq \mathcal{X} \subseteq \mathbb{R}^n \rightarrow \mathcal{Y} \subseteq \mathbb{R}^m$

**Definition 3** (Continuous Function). A function  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is continuous at  $\mathbf{x} \in \text{dom } f$  if for all  $\epsilon > 0$  there exists a  $\delta > 0$  with  $\mathbf{y} \in \text{dom } f$ , such that

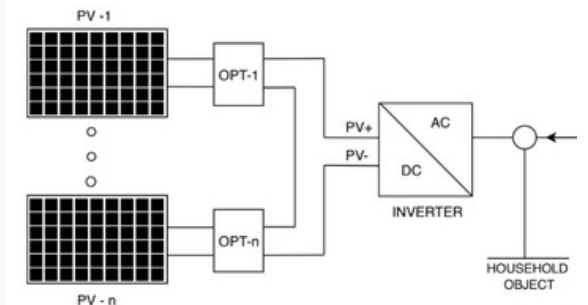
$$\|\mathbf{y} - \mathbf{x}\|_2 \leq \delta \Rightarrow \|f(\mathbf{y}) - f(\mathbf{x})\|_2 \leq \epsilon.$$



Continuous Optimization



Discrete Optimization



# Lipschitz Continuity

**Definition 3** (Continuity). A function  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is continuous at  $\mathbf{x} \in \text{dom } f$  if for all  $\epsilon > 0$  there exists a  $\delta > 0$  with  $\mathbf{y} \in \text{dom } f$ , such that

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

**Definition 4** (Lipschitz Continuity). A function  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is  $G$ -Lipschitz-continuous if for all  $\mathbf{x}, \mathbf{y} \in \text{dom } f$ ,

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

# Lipschitzness and Subgradient

- Relationship between *Lipschitzness* and *bounded subgradient*

**Theorem 8.** Let  $f : \mathcal{X} \rightarrow \mathbb{R}$  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).

# Optimization Complexity: Easy vs Hard

A gentle start: Quadratic Functions:  $f(x) = ax^2 + bx + c$

- Very easy to optimize, in fact, we have a close-form solution.
- Any benign property (compared to general convex problem)?

# Smoothness

**Definition 5** (Smoothness). A function  $f$  is  $L$ -smooth with respect to the  $\|\cdot\|$  norm if, for any  $\mathbf{x}, \mathbf{y} \in \text{dom } f$ ,

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

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

Smoothness is defined over the primal-dual norms, which become  $\ell_2$ -norm when specialized to Euclidean space (and then,  $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L \|\mathbf{x} - \mathbf{y}\|_2$ ).

# 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:**  $f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle dt$  (calculus)

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

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

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

□

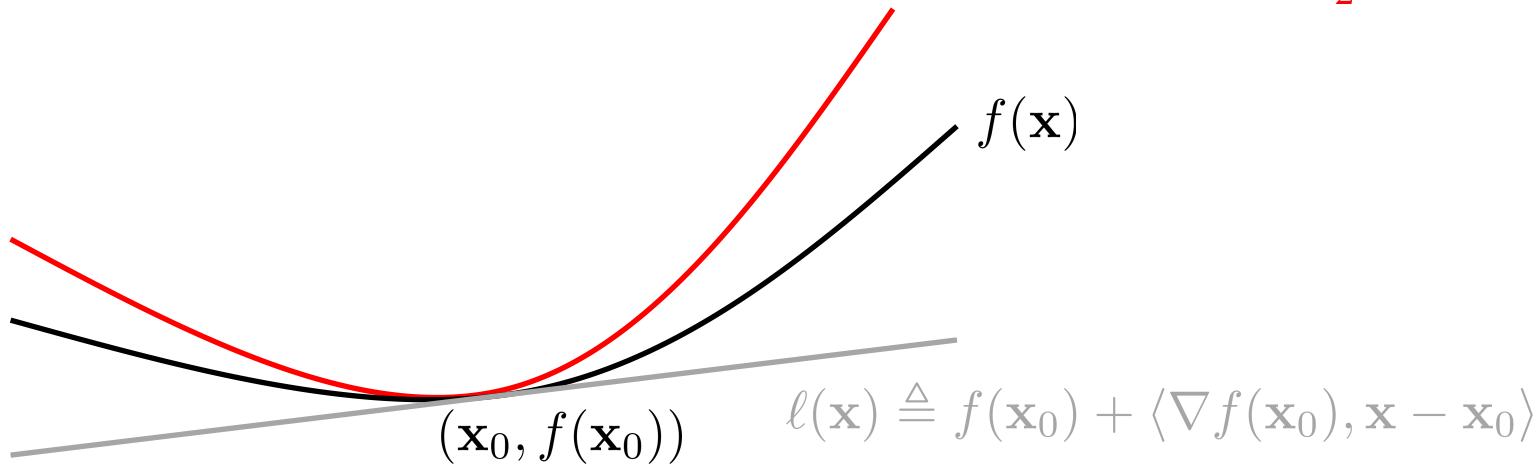
# Smoothness

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  $\mathbf{x}, \mathbf{y} \in \mathcal{X}$*

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

$$h(\mathbf{x}) \triangleq f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{x}_0\|_2^2$$



In Optimization theory:  
smooth vs nonsmooth  
“平滑” vs “非平滑”

# 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  $\|A\|_{\text{op},p}$ -smooth w.r.t.  $\|\cdot\|_p$  norm.

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

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

**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

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

# Smoothness

**Example 3.** Log-sum-exp function  $f(\mathbf{x}) = \log(e^{x_1} + e^{x_2} + \dots + e^{x_n})$  is 1-smooth w.r.t.  $\ell_2$ -norm and  $\ell_\infty$ -norm.

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

**Example 5.** Function  $f(\mathbf{x}) = \sqrt{1 + \|\mathbf{x}\|_2^2}$  is 1-smooth w.r.t.  $\ell_2$ -norm.

**Example 6.** Function  $f(\mathbf{x}) = \frac{1}{2} \|\mathbf{x} - \Pi_{\mathcal{X}}[\mathbf{x}]\|^2$  is 1-smooth w.r.t.  $\ell_2$ -norm, where  $\Pi_{\mathcal{X}}[\mathbf{x}]$  denotes the Euclidean projection of  $\mathbf{x}$  onto a *convex* domain  $\mathcal{X}$ .

# Smoothness

**Theorem 9** (*First-order* Characterizations of  $L$ -smoothness). Let  $f : \mathcal{X} \rightarrow \mathbb{R}$  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  $\lambda \in [0, 1]$ . (essentially zero-th order characterization)

*Proofs can be found below Theorem 5.8 of Amir Beck's book.*

# Smoothness

**Theorem 10** (*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  $\ell_p$ -norm ( $p \in [1, \infty]$ ) is equivalent to*

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

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

**Example 5.** Function  $f(\mathbf{x}) = \sqrt{1 + \|\mathbf{x}\|_2^2}$  is 1-smooth w.r.t.  $\ell_2$ -norm.

*Proof:*

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

□

# Strong Convexity

**Definition 7 (Strong Convexity).** A function  $f$  is  $\sigma$ -strongly convex with respect to norm  $\|\cdot\|$  if, for any  $\mathbf{x}, \mathbf{y} \in \text{dom } f$  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:*

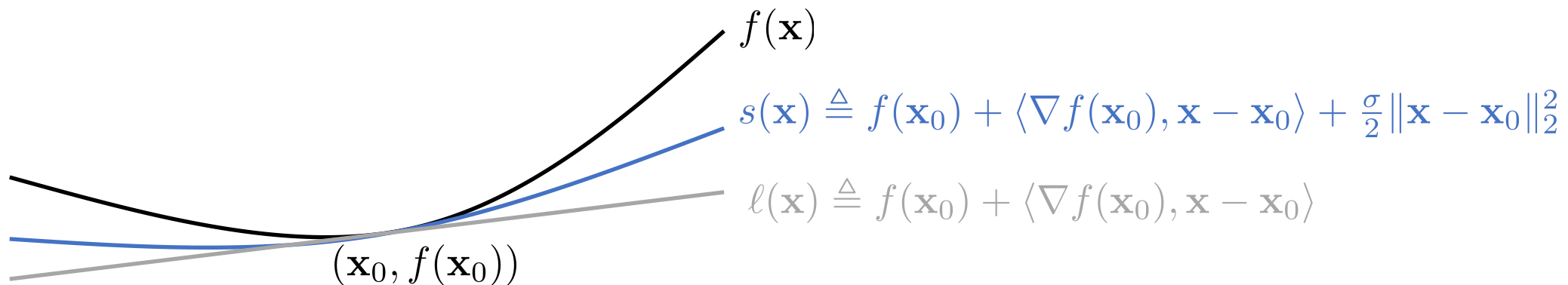
- $f(\mathbf{x}) = \|\mathbf{x}\|_p^2$  is 2-strongly-convex with respect to norm  $\|\cdot\|_p$ .
- Negative entropy  $f(\mathbf{x}) = \sum_{i=1}^d x_i \ln x_i$  over probability distribution (i.e.,  $x_i \in [0, 1]$  and  $\sum_{i=1}^d x_i = 1$ ) is 1-strongly-convex with respect to norm  $\|\cdot\|_1$ .

# Strong Convexity

The most commonly used property for strongly convex functions.

**Theorem 11.** *Let  $f$  be a proper closed and  $\sigma$ -strongly convex function. Then for any  $\mathbf{x} \in \text{dom}(\partial f)$ ,  $\mathbf{y} \in \text{dom}(f)$  and  $\mathbf{g} \in \partial f(\mathbf{x})$ ,*

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



# Strong Convexity

**Theorem 11** (*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  $\sigma$ -strongly convex.

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

$$\begin{aligned} f(\lambda \mathbf{y} + (1 - \lambda)\mathbf{x}) &\leq \lambda f(\mathbf{y}) + (1 - \lambda)f(\mathbf{x}) - \frac{\sigma}{2}\lambda(1 - \lambda)\|\mathbf{x} - \mathbf{y}\|^2 \\ \Rightarrow \frac{f(\mathbf{x} + \lambda(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{\lambda} &\leq f(\mathbf{y}) - f(\mathbf{x}) - \frac{\sigma}{2}(1 - \lambda)\|\mathbf{x} - \mathbf{y}\|^2 \quad (\text{rearrange}) \\ \Rightarrow f'(\mathbf{x}; \mathbf{y} - \mathbf{x}) &\triangleq \lim_{\lambda \rightarrow 0} \frac{f(\mathbf{x} + \lambda(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{\lambda} \leq f(\mathbf{y}) - f(\mathbf{x}) - \frac{\sigma}{2}\|\mathbf{x} - \mathbf{y}\|^2 \end{aligned}$$

$f'(\mathbf{x}; \mathbf{y} - \mathbf{x})$ : 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 12.** Let  $\mathcal{X}$  be a Euclidean space. Then  $f$  is  $\sigma$ -strongly convex with respect to norm  $\|\cdot\|$  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  $\sigma$ -strongly convex w.r.t. the  $\ell_2$ -norm if and only if  $A \succeq \sigma I$ .

**Proof:**  $f$  is  $\sigma$ -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

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

□

# Strong Convexity

**Theorem 14.** *Let  $f$  be a proper closed and  $\sigma$ -strongly convex function. Then*

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

Be careful about the usage of  $\mathbf{x}^* = \arg \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$  (you need to ensure the uniqueness).

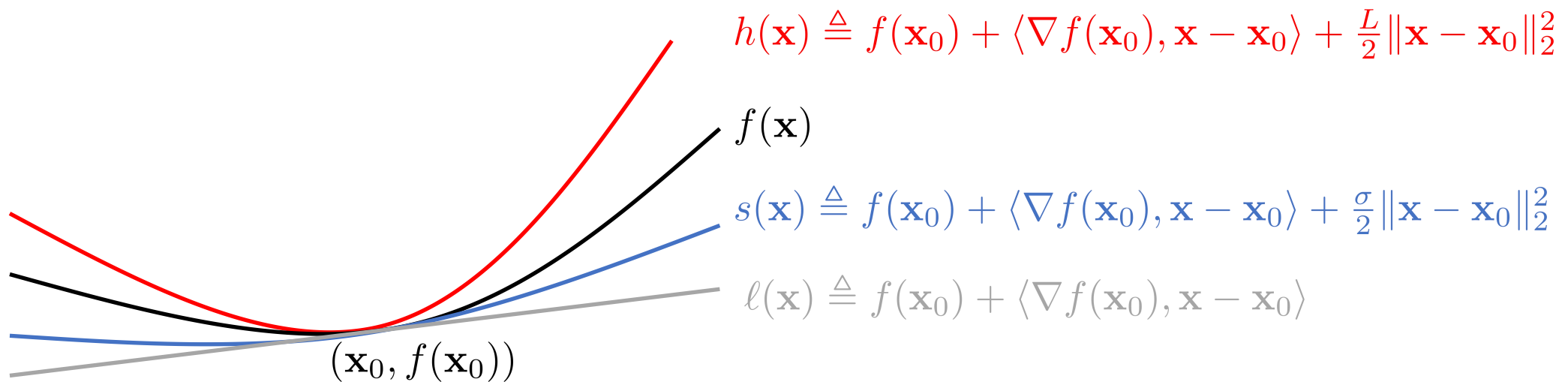
The *function-value convergence* is more essential for strongly convex optimization.

# Strongly Convex and Smooth

If function  $f$  is both  $\sigma$ -strongly convex and  $L$ -smooth w.r.t.  $\ell_2$ -norm, then

- $\sigma I \preceq \nabla^2 f(\mathbf{x}) \preceq LI$
- $f$  is  $\gamma$ -well-conditioned with  $\gamma = \kappa^{-1}$ , where  $\kappa \triangleq L/\sigma \geq 1$  is called the *condition number*.

The smaller the condition number  $\kappa$  is, the easier the function is.



# Relationship

**Theorem 15** (Conjugate Correspondence). *Consider the conjugate function:*

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

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

*Reference:* Kakade et al., [On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization](#). 2009.

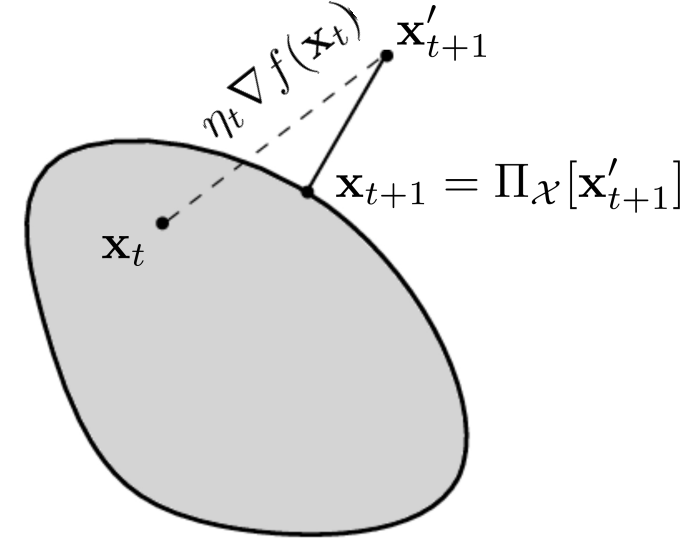
# Part 5. Gradient Descent

- Gradient Descent
- Surrogate Optimization

# Gradient Descent

- GD Template:

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} [\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)]$$



- $\mathbf{x}_1$  can be an arbitrary point inside the domain.
- $\eta_t > 0$  is the potentially time-varying *step size* (or called *learning rate*).
- Projection  $\Pi_{\mathcal{X}}[\mathbf{y}] = \arg \min_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x} - \mathbf{y}\|$  ensures the feasibility.

# Why Gradient Descent?

- For simplicity, we consider the *unconstrained* setting.

- **A General Idea:** *Surrogate Optimization*

We aim to find a sequence of *local upper bounds*  $U_1, \dots, U_T$ , where the surrogate function  $U_t : \mathbb{R}^d \mapsto \mathbb{R}$  may depend on  $\mathbf{x}_t$  such that

- (i)  $f(\mathbf{x}_t) = U_t(\mathbf{x}_t)$ ;
- (ii)  $f(\mathbf{x}) \leq U_t(\mathbf{x})$  holds for all  $\mathbf{x} \in \mathbb{R}^d$ ;
- (iii)  $U_t(\mathbf{x})$  should be simple enough to minimize.

$\Rightarrow$  Then, our proposed algorithm would be  $\mathbf{x}_{t+1} = \arg \min_{\mathbf{x}} U_t(\mathbf{x})$

# Why Gradient Descent?

- Following the *surrogate optimization* principle, let's invent GD for convex and *smooth* functions.

**Proposition 1.** *Suppose that  $f$  is convex and differentiable. Moreover, suppose that  $f$  is  $L$ -smooth with respect to  $\ell_2$ -norm. Define the surrogate  $U_t : \mathbb{R}^d \mapsto \mathbb{R}$  as*

$$U_t(\mathbf{x}) \triangleq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{x}_t\|_2^2.$$

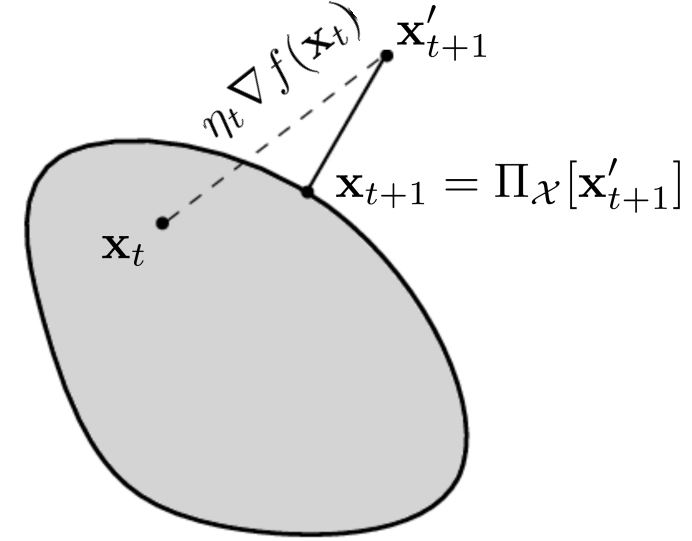
*Then, we have*

- (i)  $f(\mathbf{x}_t) = U_t(\mathbf{x}_t)$ ;
- (ii)  $f(\mathbf{x}) \leq U_t(\mathbf{x})$  holds for all  $\mathbf{x} \in \mathbb{R}^d$ ;
- (iii)  $\mathbf{x}_{t+1} = \arg \min_{\mathbf{x}} U_t(\mathbf{x})$  is equivalent to  $\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{L} \nabla f(\mathbf{x}_t)$ .

# Gradient Descent

- GD Template:

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} [\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)]$$



- $\mathbf{x}_1$  can be an arbitrary point inside the domain.
- $\eta_t > 0$  is the potentially time-varying *step size* (or called *learning rate*).
- Projection  $\Pi_{\mathcal{X}}[\mathbf{y}] = \arg \min_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x} - \mathbf{y}\|$  ensures the feasibility.

⇒ It only forms a general template, and there are still much complexity yet to specify.

# GD Template

GD template:  $\mathbf{x}_{t+1} = \Pi_{\mathcal{X}} [\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)]$

- step size  $\eta_t$
- output sequence  $\bar{\mathbf{x}}_t$  based on  $\{\mathbf{x}_s\}_{s=1}^t$
- stochastic case: gradient estimates
- GD is only a template, and many variants exist
- ...

*Key requirements:* provable guarantees, particularly with finite-sample (non-asymptotic) rates

# Summary

## PERFORMANCE MEASURE

- Complexity Measure
- Convergence Rate
- Iteration Complexity
- Computational Complexity

## SUBGRADIENT

- Subgradient
- Subdifferential

## OPTIMALITY CONDITION

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

## FUNCTION PROPERTIES

- Continuous Functions
- Smoothness/Strong Convexity
- Condition Number
- Conjugate Correspondence

## GRADIENT DESCENT TEMPLATE

- Gradient Descent
- Surrogate Optimization

Q & A

*Thanks!*