

# Preliminaries for Optimization Algorithm Design and Analysis

**Hailiang ZHAO @ ZJU-CS**

*<http://hliangzhao.me>*

October 17, 2022

*The contents in this slide are used very frequently, and they should be kept firmly in mind. I will update the slide aperiodically, if necessary.*

# Outline

## A Algebra and Probability

- Cauchy–Schwartz Inequality

- Singular Value Decomposition

- Laplacian Matrix

- Inequalities on Expectation

## B Convex Analysis

- Convex Set and Convex Functions

- Smooth and Lipschitz Continuous Functions

- Monotone Operator and Monotone Function

- Lagrangian Function, Dual Problem, and KKT Conditions

## C Non-Convex Analysis

- Lower Semicontinuous Function

- Subdifferential

## References

# Outline

## A Algebra and Probability

Cauchy–Schwartz Inequality

Singular Value Decomposition

Laplacian Matrix

Inequalities on Expectation

## B Convex Analysis

Convex Set and Convex Functions

Smooth and Lipschitz Continuous Functions

Monotone Operator and Monotone Function

Lagrangian Function, Dual Problem, and KKT Conditions

## C Non-Convex Analysis

Lower Semicontinuous Function

Subdifferential

## References

# Cauchy–Schwartz Inequality

## Proposition A.1 (Cauchy–Schwartz Inequality)

For any  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ , we have

$$\langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\| \|\mathbf{y}\|. \quad (1)$$

## Lemma A.1

For any  $\mathbf{x}, \mathbf{y}, \mathbf{z}$  and  $\mathbf{w} \in \mathbb{R}^n$ , we have the three identities:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{2} (\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \|\mathbf{x} - \mathbf{y}\|^2) \quad (2)$$

$$\langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{2} (\|\mathbf{x} + \mathbf{y}\|^2 - \|\mathbf{x}\|^2 - \|\mathbf{y}\|^2) \quad (3)$$

$$\begin{aligned} \langle \mathbf{x} - \mathbf{z}, \mathbf{y} - \mathbf{w} \rangle &= \frac{1}{2} (\|\mathbf{x} - \mathbf{w}\|^2 + \|\mathbf{z} - \mathbf{y}\|^2) \\ &\quad - \frac{1}{2} (\|\mathbf{z} - \mathbf{w}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2). \end{aligned} \quad (4)$$

# Singular Value Decomposition (SVD)

## Definition A.1 (Singular Value Decomposition, SVD)

*Suppose that  $\mathbf{A} \in \mathbb{R}^{m \times n}$  with  $\text{rank}(\mathbf{A}) = r$ . Then  $\mathbf{A}$  can be factorized as*

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{5}$$

*where  $\mathbf{U} \in \mathbb{R}^{m \times r}$  satisfies  $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ ,  $\mathbf{V} \in \mathbb{R}^{n \times r}$  satisfies  $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ , and  $\mathbf{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_r)$  with  $\sigma_1 \geq \dots \geq \sigma_r > 0$ .*

The above factorization is called the economical singular value decomposition (SVD) of  $\mathbf{A}$ . The columns of  $\mathbf{U}$  are called left singular vectors of  $\mathbf{A}$ , the columns of  $\mathbf{V}$  are right singular vectors, and the numbers  $\sigma_i$  are the singular values.

# Laplacian Matrix

## Definition A.2 (Laplacian Matrix of a Graph)

*Denote a graph as  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , where  $\mathcal{V}$  and  $\mathcal{E}$  are the node and the edge sets, respectively.  $e_{ij} = (i, j) \in \mathcal{E}$  indicates that nodes  $i$  and  $j$  are connected. Define  $\mathcal{V}_i = \{j \in \mathcal{V} \mid (i, j) \in \mathcal{E}\}$  to be the neighborhood of node  $i$ , i.e., the index set of the nodes that are connected to node  $i$ . The Laplacian matrix  $\mathbf{L}$  of the graph is defined as*

$$\mathbf{L}_{ij} = \begin{cases} |\mathcal{V}_i| & \text{if } i = j, \\ -1 & \text{if } i \neq j \text{ and } (i, j) \in \mathcal{E}, \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

# Laplacian Matrix

## Proposition A.2 (Properties of Laplacian Matrix)

*A Laplacian matrix  $\mathbf{L}$  of a graph with  $n$  nodes has the following properties:*

1.  $\mathbf{L} \succeq \mathbf{0}$ ;
2.  $\text{rank}(\mathbf{L}) = n - c$ , where  $c$  is the number of connected components in the graph, and the eigenvector associated to 0 is  $\mathbf{1}_n$ .

# Expectation

## Proposition A.3

*Given random vector  $\xi$ , we have*

$$\mathbb{E}[\xi - \mathbb{E}[\xi]]^2 \leq \mathbb{E}[\xi]^2. \quad (7)$$

## Proposition A.4 (Jensen's Inequality: Continuous Case)

*if  $f : C \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$  is convex and  $\xi$  is a random vector over  $C$ , then*

$$f(\mathbb{E}[\xi]) \leq \mathbb{E}[f(\xi)]. \quad (8)$$



# Outline

## A Algebra and Probability

Cauchy–Schwartz Inequality

Singular Value Decomposition

Laplacian Matrix

Inequalities on Expectation

## B Convex Analysis

Convex Set and Convex Functions

Smooth and Lipschitz Continuous Functions

Monotone Operator and Monotone Function

Lagrangian Function, Dual Problem, and KKT Conditions

## C Non-Convex Analysis

Lower Semicontinuous Function

Subdifferential

## References

# Definitions Evolved in Convex Analysis

In the following, we only consider convex analysis on  $n$  dimensional Euclidean spaces.

## Definition B.1 (Convex Set)

*A set  $C \subseteq \mathbb{R}^n$  is called convex if for all  $\mathbf{x}, \mathbf{y} \in C$  and  $\alpha \in [0, 1]$  we have  $\alpha\mathbf{x} + (1 - \alpha)\mathbf{y} \in C$ .*

## Definition B.2 (Convex Function)

*A function  $f : C \rightarrow \mathbb{R}$  is called convex if  $C$  is a convex set and for all  $\mathbf{x}, \mathbf{y} \in C$  and  $\alpha \in [0, 1]$  we have*

$$f(\alpha\mathbf{x} + (1 - \alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1 - \alpha)f(\mathbf{y}). \quad (9)$$

# Definitions Evolved in Convex Analysis

## Definition B.3 (Concave Function)

*A function  $f : C \rightarrow \mathbb{R}$  is called concave if  $-f$  is convex.*

## Definition B.4 (Strictly Convex Function)

*A function  $f : C \rightarrow \mathbb{R}$  is called strictly convex if  $C$  is a convex set and for all  $\mathbf{x} \neq \mathbf{y}$  and  $\alpha \in (0, 1)$  we have*

$$f(\alpha \mathbf{x} + (1 - \alpha) \mathbf{y}) < \alpha f(\mathbf{x}) + (1 - \alpha) f(\mathbf{y}). \quad (10)$$

# Definitions Evolved in Convex Analysis

## Definition B.5 (Strongly Convex Function)

*A function  $f : C \rightarrow \mathbb{R}$  is called strongly convex if  $C$  is a convex set and there exists a constant  $\mu > 0$  such that for all  $\mathbf{x}, \mathbf{y} \in C$  and  $\alpha \in [0, 1]$  we have*

$$f(\alpha \mathbf{x} + (1 - \alpha) \mathbf{y}) \leq \alpha f(\mathbf{x}) + (1 - \alpha) f(\mathbf{y}) \quad (11)$$

$$- \frac{\mu \alpha (1 - \alpha)}{2} \|\mathbf{y} - \mathbf{x}\|^2. \quad (12)$$

$\mu$  is called the strongly convexity modules of  $f$ . We call  $f$  a  $\mu$ -strongly convex function. If a function is not strongly convex, we call it a generally convex function.

# Jensen's Inequality

## Proposition B.1 (Jensen's Inequality: Discrete Case)

If  $f : C \rightarrow \mathbb{R}$  is convex,  $\mathbf{x}_i \in C$ ,  $\alpha_i \geq 0$ ,  $i \in [m]$ , and  $\sum_{i=1}^m \alpha_i = 1$ , then

$$f\left(\sum_{i=1}^m \alpha_i \mathbf{x}_i\right) \leq \sum_{i=1}^m \alpha_i f(\mathbf{x}_i). \quad (13)$$

# Smooth and Lipschitz Continuous Functions

## Definition B.6 (Smooth Function)

*A function is (informally) called smooth if it is continuously differentiable.*

## Definition B.7 (Function with Lipschitz Continuous Gradients)

*A differentiable function  $f : C \rightarrow \mathbb{R}$  is called to have Lipschitz continuous gradients if there exists  $L > 0$  such that*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{y} - \mathbf{x}\|, \quad \forall \mathbf{x}, \mathbf{y} \in C. \quad (14)$$

*We call  $f$  is an  $L$ -smooth function.*

# Properties of $L$ -smooth Functions

## Proposition B.2

*If  $f : C \rightarrow \mathbb{R}$  is  $L$ -smooth, then*

$$|f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| \leq \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in C. \quad (15)$$

*If  $f$  is both  $L$ -smooth and convex, then*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2L} \|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|^2. \quad (16)$$

# Subgradients

## Definition B.8 (Subgradient of a Convex Function)

*A vector  $\mathbf{g}$  is called a subgradient of a convex function  $f : C \rightarrow \mathbb{R}$  at  $\mathbf{x} \in C$  if*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle, \forall \mathbf{y} \in C. \quad (17)$$

*The set of subgradients at  $\mathbf{x}$  is denoted as  $\partial f(\mathbf{x})$ .*

## Proposition B.3

*For convex function  $f : C \rightarrow \mathbb{R}$ , its subgradient exists at every interior point of  $C$ . It is differentiable at  $\mathbf{x}$  iff  $\partial f(\mathbf{x})$  is a singleton.*



# Inequalities with Functions' Smoothness

## Proposition B.4

*If  $f : C \rightarrow \mathbb{R}$  is  $\mu$ -strongly convex, then*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|^2, \quad \forall \mathbf{g} \in \partial f(\mathbf{x}). \quad (18)$$

*In particular, if  $f$  is  $\mu$ -strongly convex and  $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x} \in C} f(\mathbf{x})$ , then*

$$f(\mathbf{x}) - f(\mathbf{x}^*) \geq \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}^*\|^2. \quad (19)$$

# Inequalities with Functions' Smoothness

## Proposition B.4 (Cont'd)

*On the other hand, if  $f$  is differentiable and  $\mu$ -strongly convex, we have*

$$f(\mathbf{x}^*) \geq f(\mathbf{x}) - \frac{1}{2\mu} \|\nabla f(\mathbf{x})\|^2. \quad (20)$$

*We can further have*

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \mu \|\mathbf{x} - \mathbf{y}\|^2. \quad (21)$$

*In particular,*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \geq \mu \|\mathbf{x} - \mathbf{y}\|. \quad (22)$$

## Other Definitions used in Convex Analysis

### Definition B.9 (Epigraph)

*The epigraph of  $f : C \rightarrow \mathbb{R}$  is defined as*

$$\text{epi } f = \{(\mathbf{x}, t) \mid \mathbf{x} \in C, t \geq f(\mathbf{x})\}. \quad (23)$$

### Definition B.10 (Closed Function)

*If  $\text{epi } f$  is a closed set, then  $f$  is called a closed function.*

## Other Definitions used in Convex Analysis

### Definition B.11 (Monotone Operator and Monotone Function)

A set-valued mapping  $f : C \rightarrow 2^{\mathbb{R}^n}$  (also denoted as  $f : C \rightrightarrows \mathbb{R}^n$  for brevity) is called a monotone operator if

$$\langle \mathbf{x} - \mathbf{y}, \mathbf{u} - \mathbf{v} \rangle \geq 0, \quad \forall \mathbf{x}, \mathbf{y} \in C \text{ and } \mathbf{u} \in f(\mathbf{x}), \mathbf{v} \in f(\mathbf{y}). \quad (24)$$

In particular, if  $f$  is single-valued and

$$\langle \mathbf{x} - \mathbf{y}, f(\mathbf{x}) - f(\mathbf{y}) \rangle \geq 0, \quad \forall \mathbf{x}, \mathbf{y} \in C, \quad (25)$$

then it is called a monotone function.

## Other Definitions used in Convex Analysis

### Definition B.12 (Maximal Monotone Operator)

*Define the graph of an operator  $\mathcal{T}$  as*

$$\text{Graph}(\mathcal{T}) = \{(\mathbf{x}, \mathbf{u}) \mid \mathbf{x} \in C, \mathbf{u} \in \mathcal{T}(\mathbf{x})\}. \quad (26)$$

*For a monotone operator  $\mathcal{T}$ , if it has the property: For any monotone operator  $\mathcal{T}'$ ,  $\text{Graph}(\mathcal{T}) \subseteq \text{Graph}(\mathcal{T}')$  implies  $\mathcal{T} = \mathcal{T}'$ , then it is called a maximal monotone operator.*

### Proposition B.5

*If  $\mathcal{T}$  is a maximal monotone operator, then its resolvent  $(\mathcal{I} + \mathcal{T})^{-1}$  is single-valued. Note that  $\mathcal{I}$  is the identity operator.*

# Monotonicity of Subgradient

## Proposition B.6 (Monotonicity of Subgradient)

*If  $f : C \rightarrow \mathbb{R}^n$  is convex, then  $\partial f(\mathbf{x})$  is a monotone operator. If  $f$  is further  $\mu$ -strongly convex, then*

$$\langle \mathbf{x}_1 - \mathbf{x}_2, \mathbf{g}_1 - \mathbf{g}_2 \rangle \geq \mu \|\mathbf{x}_1 - \mathbf{x}_2\|^2 \quad (27)$$

*holds for any  $\mathbf{x}_1, \mathbf{x}_2 \in C$  and  $\mathbf{g}_1 \in \partial f(\mathbf{x}_1), \mathbf{g}_2 \in \partial f(\mathbf{x}_2)$ . If  $f$  is closed and convex, then  $\partial f(\mathbf{x})$  is a maximal monotone operator.*

# Bregman Distance

## Definition B.13 (Bregman Distance)

*Given a differentiable convex function  $\phi$ , the associated Bregman distance is defined as*

$$D_{\phi}(\mathbf{y}, \mathbf{x}) = \phi(\mathbf{y}) - \phi(\mathbf{x}) - \langle \nabla \phi(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle. \quad (28)$$

*If  $\phi$  is convex but not differentiable, then the associated Bregman Distance is defined as*

$$D_{\phi}^{\mathbf{v}}(\mathbf{y}, \mathbf{x}) = \phi(\mathbf{y}) - \phi(\mathbf{x}) - \langle \mathbf{v}, \mathbf{y} - \mathbf{x} \rangle, \quad (29)$$

*where  $\mathbf{v}$  is a particular subgradient in  $\partial\phi(\mathbf{x})$ .*

The squared Euclidean distance is obtained when  $\phi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$ , in which case

$$D_{\phi}(\mathbf{y}, \mathbf{x}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2. \quad (30)$$

# Bregman Distance

## Lemma B.1

*The Bregman distance  $D_\phi$  has the following properties:*

1. *When  $\phi$  is  $\mu$ -strongly convex, we have*

$$D_\phi(\mathbf{y}, \mathbf{x}) \geq \frac{\mu}{2} \|\mathbf{y} - \mathbf{x}\|^2. \quad (31)$$

2. *For any  $\mathbf{u}$ ,  $\mathbf{v}$ , and  $\mathbf{w}$ , we have*

$$\begin{aligned} \langle \nabla \phi(\mathbf{u}) - \nabla \phi(\mathbf{v}), \mathbf{w} - \mathbf{u} \rangle &= D_\phi(\mathbf{w}, \mathbf{v}) \\ &\quad - \left( D_\phi(\mathbf{w}, \mathbf{u}) + D_\phi(\mathbf{u}, \mathbf{v}) \right). \end{aligned} \quad (32)$$



# Conjugate Function

## Definition B.14 (Conjugate Function)

*Given  $f : C \rightarrow \mathbb{R}^n$ , its conjugate function is defined as*

$$f^*(\mathbf{u}) = \sup_{\mathbf{z} \in C} (\langle \mathbf{z}, \mathbf{u} \rangle - f(\mathbf{z})). \quad (33)$$

*The domain of  $f^*$  is*

$$\text{dom } f^* = \{\mathbf{u} \mid f^*(\mathbf{u}) < +\infty\}. \quad (34)$$

# Properties of Conjugate Function

## Proposition B.7 (Properties of Conjugate Function)

*Given  $f : C \rightarrow \mathbb{R}^n$ , its conjugate function  $f^*$  has the following properties:*

1.  $f^*$  is always a convex function.
2.  $f^{**}(\mathbf{x}) \leq f(\mathbf{x}), \forall \mathbf{x} \in C$ .
3. If  $f$  is closed and convex, then  $f^{**}(\mathbf{x}) = f(\mathbf{x}), \forall \mathbf{x} \in C$ .
4. If  $f$  is  $L$ -smooth, then  $f^*$  is  $L^{-1}$ -strongly convex on  $\text{dom } f^*$ .  
Conversely, if  $f$  is  $\mu$ -strongly convex, then  $f^*$  is  $\mu^{-1}$ -smooth on  $\text{dom } f^*$ .
5. If  $f$  is closed and convex, then  $\mathbf{y} \in \partial f(\mathbf{x})$  iff  $\mathbf{x} \in \partial f^*(\mathbf{y})$ .

## Proposition B.8 (Fenchel-Young Inequality)

*Let  $f^*$  be the conjugate function of  $f$ , then*

$$f(\mathbf{x}) + f^*(\mathbf{y}) \geq \langle \mathbf{x}, \mathbf{y} \rangle. \quad (35)$$

# Lagrangian Function

## Definition B.15 (Lagrangian Function)

*Given a constrained problem:*

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^n} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{Ax} = \mathbf{b}, \\ & \mathbf{g}(\mathbf{x}) \leq \mathbf{0}, \end{aligned} \tag{36}$$

*where  $\mathbf{A} \in \mathbb{R}^{m \times n}$  and  $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), \dots, g_p(\mathbf{x})]^T$ , the Lagrangian function is*

$$L(\mathbf{x}, \mathbf{u}, \mathbf{v}) = f(\mathbf{x}) + \langle \mathbf{u}, \mathbf{Ax} - \mathbf{b} \rangle + \langle \mathbf{v}, \mathbf{g}(\mathbf{x}) \rangle, \tag{37}$$

*where  $\mathbf{v} \geq 0$ .*

# Lagrange Dual Function

## Definition B.16 (Lagrange Dual Function)

*Given a constrained problem (36), the Lagrange dual function is  $d(\mathbf{u}, \mathbf{v}) = \min_{\mathbf{x} \in C} L(\mathbf{x}, \mathbf{u}, \mathbf{v})$ , where  $C$  is the intersection of the domains of  $f$  and  $g$ . The domain of the dual function is*

$$\mathcal{D} = \{(\mathbf{u}, \mathbf{v}) \mid \mathbf{v} \geq 0, d(\mathbf{u}, \mathbf{v}) > -\infty\}. \quad (38)$$

## Definition B.17 (Dual Problem)

*Given a constrained problem (36), the dual problem is*

$$\max_{\mathbf{u}, \mathbf{v}} d(\mathbf{u}, \mathbf{v}), \quad \text{s.t.} \quad (\mathbf{u}, \mathbf{v}) \in \mathcal{D}. \quad (39)$$

*Correspondingly, (36) is called the primal problem.*

# Slater's Condition

## Definition B.18 (Slater's Condition)

*For convex primal problem (36), if there exists an  $\mathbf{x}_0$  such that*

$$\mathbf{A}\mathbf{x}_0 = \mathbf{b}, \tag{40}$$

$$g_i(\mathbf{x}_0) \leq 0, \forall i \in \mathcal{I}_1, \tag{41}$$

$$g_j(\mathbf{x}_0) < 0, \forall j \in \mathcal{I}_2, \tag{42}$$

*where  $\mathcal{I}_1$  and  $\mathcal{I}_2$  are the sets of indices of linear and nonlinear inequality constraints, respectively, then the Slater's condition holds.*

# Properties of Dual Problem

## Proposition B.9 (Properties of Dual Problem)

1.  $d(\mathbf{u}, \mathbf{v})$  is always a concave function, even if the primal problem (36) is not convex.
2. The primal and the dual optimal values,  $f^*$  and  $d^*$ , always satisfy the weak duality:  $f^* \geq d^*$ .
3. When the Slater's condition holds, the strong duality holds:  $f^* = d^*$ .
4. Let  $\mathbf{x}(\mathbf{u}, \mathbf{v}) \in \operatorname{argmin}_{\mathbf{x} \in C} L(\mathbf{x}, \mathbf{u}, \mathbf{v})$ , then

$$(\mathbf{Ax}(\mathbf{u}, \mathbf{v}) - \mathbf{b}, \mathbf{g}(\mathbf{x}(\mathbf{u}, \mathbf{v}))) \in \partial d(\mathbf{u}, \mathbf{v}). \quad (43)$$

# Properties of Dual Problem

## Proof Sketch of Proposition B.9.2

We consider a problem with inequality constraints:

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

Our target is to find the optimal (maximal) lower bound of  $f$ . Firstly, for any  $v \in \mathbb{R}$ , how to make it be a lower bound of  $f$ ? Actually, if the following equation system on  $\mathbf{x}$  has no solution, then we can say  $v$  is a lower bound of  $f$ :

$$\begin{cases} f(\mathbf{x}) < v \\ g_i(\mathbf{x}) \leq 0, i \in [m] \end{cases} \quad (44)$$

# Properties of Dual Problem

## Proof Sketch of Proposition B.9.2 (Cont'd)

If (44) has a solution, then, for any  $\lambda \geq \mathbf{0}$ , the following equation of  $\mathbf{x}$

$$f(\mathbf{x}) + \sum_{i \in [m]} \lambda_i g_i(\mathbf{x}) < v \quad (45)$$

has a solution. According to the equivalence of contrapositives, we have: For any  $\lambda \geq \mathbf{0}$ , if (45) has no solution, then (44) has no solution. On the other hand, (45) has no solution for any given  $\lambda \geq \mathbf{0}$  *iff* the following inequality holds for any given  $\lambda \geq \mathbf{0}$ :

$$\min_{\mathbf{x}} f(\mathbf{x}) + \sum_{i \in [m]} \lambda_i g_i(\mathbf{x}) \geq v. \quad (46)$$



# Properties of Dual Problem

## Proof Sketch of Proposition B.9.2 (Cont'd)

Combining the above results, we have: If (46) holds for any given  $\boldsymbol{\lambda} \geq \mathbf{0}$ , then  $v$  is a lower bound of  $f$ . Note that we want to find the maximal lower bound of  $f$ , i.e.

$$v^* = \max_{\boldsymbol{\lambda} \geq \mathbf{0}} \left( \underbrace{\min_{\mathbf{x}} \left[ \overbrace{f(\mathbf{x}) + \sum_{i \in [m]} \lambda_i g_i(\mathbf{x})}^{L(\mathbf{x}, \boldsymbol{\lambda})} \right]}_{d(\boldsymbol{\lambda}) := \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda})} \right). \quad (47)$$

As a infimum of  $f$ , we have  $v^* = \min_{\mathbf{x}^*} f(\mathbf{x}^*)$ . Therefore, we have:

$$\min_{\mathbf{x}^*} f(\mathbf{x}^*) \geq \max_{\boldsymbol{\lambda}^*} d(\boldsymbol{\lambda}^*). \quad (48)$$

# KKT Point and KKT Condition

## Definition B.19 (KKT Point and KKT Condition)

$(\mathbf{x}, \mathbf{u}, \mathbf{v})$  is called a *Karush-Kuhn-Tucker (KKT) point* of problem (36) if

1. *Stationary*:  $\mathbf{0} \in \partial f(\mathbf{x}) + \mathbf{A}^T \mathbf{u} + \sum_{i=1}^p v_i \partial g_i(\mathbf{x})$ .
2. *Primal feasibility*:  $\mathbf{Ax} = \mathbf{b}, g_i(\mathbf{x}) \leq 0, \forall i \in [p]$ .
3. *Complementary slackness*:  $v_i g_i(\mathbf{x}) = 0, \forall i \in [p]$ .
4. *Dual feasibility*:  $v_i \geq 0, \forall i \in [p]$ .

The above conditions are called the KKT condition of problem (36). They are the optimality condition of problem (36) when problem (36) is convex and satisfies the Slater's condition.

# KKT Point and KKT Condition

## Proposition B.10

*When  $f(\mathbf{x})$  and  $g_i(\mathbf{x})$ ,  $i \in [p]$  in problem (36) are all convex, then*

- 1. every KKT point is a saddle point of the Lagrangian function, and*
- 2.  $(\mathbf{x}^*, \mathbf{u}^*, \mathbf{v}^*)$  is a pair of the primal and the dual solutions with zero dual gap iff it satisfies the KKT condition.*

# Compact Set and Convex Hull

## Definition B.20 (Compact Set)

*A subset  $S$  of  $\mathcal{R}^n$  is called compact if it is both bounded and closed.*

## Definition B.21 (Convex Hull)

*The convex hull of a set  $\mathcal{X}$ , denoted as  $\text{conv}(\mathcal{X})$ , is the set of all convex combinations of points in  $\mathcal{X}$ :*

$$\text{conv}(\mathcal{X}) = \left\{ \sum_{i=1}^k \alpha_i \mathbf{x}_i \mid \mathbf{x}_i \in \mathcal{X}, \alpha_i \geq 0, i \in [k], \sum_{i=1}^k \alpha_i = 1 \right\}. \quad (49)$$

# Danskin's Theorem

## Theorem B.1 (Danskin's Theorem)

Let  $\mathcal{Z}$  be a compact subset of  $\mathbb{R}^m$ , and let  $\phi : \mathbb{R}^n \times \mathcal{Z} \rightarrow \mathbb{R}$  be continuous and such that  $\phi(\cdot, \mathbf{z}) : \mathbb{R}^n \rightarrow \mathbb{R}$  is convex for each  $\mathbf{z} \in \mathcal{Z}$ . Define  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  by  $f(\mathbf{x}) = \max_{\mathbf{z} \in \mathcal{Z}} \phi(\mathbf{x}, \mathbf{z})$  and

$$\mathcal{Z}(\mathbf{x}) = \left\{ \bar{\mathbf{z}} \mid \phi(\mathbf{x}, \bar{\mathbf{z}}) = \max_{\mathbf{z} \in \mathcal{Z}} \phi(\mathbf{x}, \mathbf{z}) \right\}. \quad (50)$$

If  $\phi(\cdot, \mathbf{z})$  is differentiable for all  $\mathbf{z} \in \mathcal{Z}$  and  $\nabla_{\mathbf{x}} \phi(\mathbf{x}, \cdot)$  is continuous on  $\mathcal{Z}$  for each  $\mathbf{x}$ , then

$$\partial f(\mathbf{x}) = \text{conv} \left\{ \nabla_{\mathbf{x}} \phi(\mathbf{x}, \mathbf{z}) \mid \mathbf{z} \in \mathcal{Z}(\mathbf{x}) \right\}, \forall \mathbf{x} \in \mathbb{R}^n. \quad (51)$$

# Saddle Point

## Definition B.22 (Saddle Point)

$(\mathbf{x}^*, \boldsymbol{\lambda}^*)$  is called a saddle point of function  $f(\mathbf{x}, \boldsymbol{\lambda}) : C \times D \rightarrow \mathbb{R}$  if it satisfies the following inequalities:

$$f(\mathbf{x}^*, \boldsymbol{\lambda}) \leq f(\mathbf{x}^*, \boldsymbol{\lambda}^*) \leq f(\mathbf{x}, \boldsymbol{\lambda}^*), \forall \mathbf{x} \in C, \boldsymbol{\lambda} \in D. \quad (52)$$

# Hoffman's Bound

## Lemma B.2 (Hoffman's Bound)

*Consider the non-empty polyhedron*

$$\mathcal{X} = \{\mathbf{x} \mid \mathbf{Ax} = \mathbf{a}, \mathbf{Bx} \leq \mathbf{b}\}. \quad (53)$$

*Then there exists a constant  $\theta$ , depending only on  $[\mathbf{A}^T, \mathbf{B}^T]^T$ , such that for any  $\mathbf{x}$  we have*

$$\text{dist}(\mathbf{x}, \mathcal{X})^2 \leq \theta^2 \left( \|\mathbf{Ax} - \mathbf{a}\|^2 + \|[\mathbf{Bx} - \mathbf{b}]_+\|^2 \right)^2, \quad (54)$$

*where  $[\cdot]_+$  means the projection to the non-negative orthant, i.e.,  $[\cdot]_+ = \max\{\cdot, \mathbf{0}\}$ .*

# Outline

## A Algebra and Probability

Cauchy–Schwartz Inequality

Singular Value Decomposition

Laplacian Matrix

Inequalities on Expectation

## B Convex Analysis

Convex Set and Convex Functions

Smooth and Lipschitz Continuous Functions

Monotone Operator and Monotone Function

Lagrangian Function, Dual Problem, and KKT Conditions

## C Non-Convex Analysis

Lower Semicontinuous Function

Subdifferential

## References



# Several Functions

## Definition C.1 (Proper Function)

*A function  $g : \mathbb{R}^n \rightarrow (-\infty, +\infty]$  is said to be proper if  $\text{dom } g \neq \emptyset$ , where  $\text{dom } g = \{\mathbf{x} \in \mathbb{R}^n \mid g(\mathbf{x}) < +\infty\}$ .*

## Definition C.2 (Lower Semicontinuous Function)

*A function  $g : \mathbb{R}^n \rightarrow (-\infty, +\infty]$  is said to be lower semicontinuous at point  $\mathbf{x}_0$  if*

$$\liminf_{\mathbf{x} \rightarrow \mathbf{x}_0} g(\mathbf{x}) \geq g(\mathbf{x}_0). \quad (55)$$

## Definition C.3 (Coercive Function)

*$f$  is called coercive if  $\lim_{\|\mathbf{x}\| \rightarrow \infty} f(\mathbf{x}) \rightarrow \infty$ .*

# Subdifferential

## Definition C.4 (Subdifferential)

Let  $f$  be a proper and lower semicontinuous function.

1. For a given  $\mathbf{x} \in \text{dom } f$ , the Fréchet subdifferential of  $f$  at  $\mathbf{x}$ , written as  $\hat{\partial}f(\mathbf{x})$ , is the set of all vectors  $\mathbf{u} \in \mathbb{R}^n$ , which satisfies

$$\liminf_{\mathbf{y} \neq \mathbf{x}, \mathbf{y} \rightarrow \mathbf{x}} \frac{f(\mathbf{y}) - f(\mathbf{x}) - \langle \mathbf{u}, \mathbf{y} - \mathbf{x} \rangle}{\|\mathbf{y} - \mathbf{x}\|} \geq 0. \quad (56)$$

2. The limiting subdifferential, or simply the subdifferential, of  $f$  at  $\mathbf{x} \in \mathbb{R}^n$ , written as  $\partial f(\mathbf{x})$ , is defined through the following closure process:

$$\partial f(\mathbf{x}) = \left\{ \mathbf{u} \in \mathbb{R}^n \mid \exists \mathbf{x}_k \rightarrow \mathbf{x}, f(\mathbf{x}_k) \rightarrow f(\mathbf{x}), \right. \\ \left. \mathbf{u}_k \in \hat{\partial}f(\mathbf{x}_k) \rightarrow \mathbf{u}, k \rightarrow \infty \right\}. \quad (57)$$

# Critical Point and Properties of Subdifferential

## Definition C.5 (Critical Point)

A point  $\mathbf{x}$  is called a critical point of function  $f$  if  $\mathbf{0} \in \partial f(\mathbf{x})$ .

## Lemma C.1

Some properties of subdifferential:

1. *In the nonconvex context, Fermat's rule remains unchanged: If  $\mathbf{x} \in \mathbb{R}^n$  is a local minimizer of  $g$ , then  $\mathbf{0} \in \partial g(\mathbf{x})$ .*
2. *Let  $(\mathbf{x}_k, \mathbf{u}_k)$  be a sequence such that  $\mathbf{x}_k \rightarrow \mathbf{x}$ ,  $\mathbf{u}_k \rightarrow \mathbf{u}$ ,  $g(\mathbf{x}_k) \rightarrow g(\mathbf{x})$ , and  $\mathbf{u}_k \in \partial g(\mathbf{x}_k)$ , then  $\mathbf{u} \in \partial g(\mathbf{x})$ .*
3. *If  $f$  is a continuously differentiable function, then*

$$\partial(f + g)(\mathbf{x}) = \nabla f(\mathbf{x}) + \partial g(\mathbf{x}). \quad (58)$$

# Outline

## A Algebra and Probability

Cauchy–Schwartz Inequality

Singular Value Decomposition

Laplacian Matrix

Inequalities on Expectation

## B Convex Analysis

Convex Set and Convex Functions

Smooth and Lipschitz Continuous Functions

Monotone Operator and Monotone Function

Lagrangian Function, Dual Problem, and KKT Conditions

## C Non-Convex Analysis

Lower Semicontinuous Function

Subdifferential

## References

# References

1. Lin, Zhouchen, Huan Li, and Cong Fang. *Alternating Direction Method of Multipliers for Machine Learning*. Springer Nature, 2022.
2. Boyd, Stephen, Stephen P. Boyd, and Lieven Vandenberghe. *Convex Optimization*. Cambridge university press, 2004.