Preliminaries for Optimization Algorithm Design and Analysis

Hailiang ZHAO @ ZJU-CS

http://hliangzhao.me

October 31, 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

Outline

Cauchy–Schwartz Inequality

Proposition A.1 (Cauchy-Schwartz Inequality)

For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, *we have* $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \leq \|\boldsymbol{x}\| \|\boldsymbol{y}\|$.

Lemma A.1

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

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \frac{1}{2}\left(\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - \|\boldsymbol{x} - \boldsymbol{y}\|^2\right) \tag{1}$$

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \frac{1}{2}\left(\|\boldsymbol{x} + \boldsymbol{y}\|^2 - \|\boldsymbol{x}\|^2 - \|\boldsymbol{y}\|^2\right) \tag{2}$$

$$\begin{aligned} \langle \boldsymbol{x} - \boldsymbol{z}, \boldsymbol{y} - \boldsymbol{w} \rangle &= \frac{1}{2}\left(\|\boldsymbol{x} - \boldsymbol{w}\|^2 + \|\boldsymbol{z} - \boldsymbol{y}\|^2\right) \\ &\quad - \frac{1}{2}\left(\|\boldsymbol{z} - \boldsymbol{w}\|^2 + \|\boldsymbol{x} - \boldsymbol{y}\|^2\right). \end{aligned} \tag{3}$$

Singular Value Decomposition (SVD)

Definition A.1 (Singular Value Decomposition, SVD)

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

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{4}$$

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* $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, ..., \sigma_r)$ *with* $\sigma_1 \geq ... \geq \sigma_r > 0$.

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 σ_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

$$\mathbf{L}ij = \begin{cases} |\mathcal{V}_i| & \text{if } i = j, \\ -1 & \text{if } i \neq j \text{ and } (i, j) \in \mathcal{E}, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

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. $rank(\mathbf{L}) = n - c$*, 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 $\boldsymbol{\xi}$, we have

$$\mathbb{E}\left[\|\boldsymbol{\xi} - \mathbb{E}[\boldsymbol{\xi}]\|^2\right] \leq \mathbb{E}\left[\|\boldsymbol{\xi}^2\|\right]. \tag{6}$$

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

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

$$f(\mathbb{E}[\boldsymbol{\xi}]) \leq \mathbb{E}[f(\boldsymbol{\xi})]. \tag{7}$$

Outline

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* $\boldsymbol{x}, \boldsymbol{y} \in C$ *and* $\alpha \in [0, 1]$ *we have* $\alpha \boldsymbol{x} + (1 - \alpha)\boldsymbol{y} \in C$*.*

Definition B.2 (Convex Function)

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

$$f\big(\alpha \boldsymbol{x} + (1 - \alpha)\boldsymbol{y}\big) \leq \alpha f(\boldsymbol{x}) + (1 - \alpha) f(\boldsymbol{y}). \tag{8}$$

Definitions Evolved in Convex Analysis

Definition B.3 (Concave Function)

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

Definition B.4 (Strictly Convex Function)

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

$$f\big(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}\big) < \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}). \tag{9}$$

Definition B.5 (Strongly Convex Function)

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

$$f\Big(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}\Big) \leq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}) \tag{10}$$

$$- \frac{\mu\alpha(1-\alpha)}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2. \tag{11}$$

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

Jensen's Inequality

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

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

$$f\left(\sum_{i=1}^{m} \alpha_i \boldsymbol{x}_i\right) \leq \sum_{i=1}^{m} \alpha_i f(\boldsymbol{x}_i). \tag{12}$$

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

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{y} - \boldsymbol{x}\|, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in C. \tag{13}$$

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

Properties of L-smooth Functions

Proposition B.2

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

$$\left|f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle\right| \leq \frac{L}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in C. \tag{14}$$

If f is both L-smooth and convex, then

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{1}{2L}\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|^2. \tag{15}$$

Subgradients

Definition B.8 (Subgradient of a Convex Function)

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

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \mathbf{g}, \boldsymbol{y} - \boldsymbol{x} \rangle, \forall \boldsymbol{y} \in C. \tag{16}$$

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

Proposition B.3

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

Inequalities with Functions' Smoothness

Proposition B.4

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

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \mathbf{g}, \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\mu}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2, \quad \forall \mathbf{g} \in \partial f(\boldsymbol{x}). \tag{17}$$

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

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \geq \frac{\mu}{2} \|\boldsymbol{x} - \boldsymbol{x}^*\|^2. \tag{18}$$

Inequalities with Functions' Smoothness

Proposition B.4 (Cont'd)

On the other hand, if f is differentiable and μ-strongly convex, we have

$$f(\boldsymbol{x}^*) \geq f(\boldsymbol{x}) - \frac{1}{2\mu}\|\nabla f(\boldsymbol{x})\|^2. \tag{19}$$

We can further have

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq \mu \|\boldsymbol{x} - \boldsymbol{y}\|^2. \tag{20}$$

In particular,

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \geq \mu \|\boldsymbol{x} - \boldsymbol{y}\|. \tag{21}$$

Other Definitions used in Convex Analysis

Definition B.9 (Epigraph)

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

$$epi\, f = \left\{ (\boldsymbol{x}, t) \mid \boldsymbol{x} \in C, t \geq f(\boldsymbol{x}) \right\}. \tag{22}$$

Definition B.10 (Closed Function)

If $epi\, f$ 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 $f : C \to 2^{\mathbb{R}^n}$ (also denoted as $f : C \rightrightarrows \mathbb{R}^n$ for brevity) is called a monotone operator if

$$\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{u} - \boldsymbol{v} \rangle \geq 0, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in C \text{ and } \boldsymbol{u} \in f(\boldsymbol{x}), \boldsymbol{v} \in f(\boldsymbol{y}). \tag{23}$$

In particular, if f is single-valued and

$$\langle \boldsymbol{x} - \boldsymbol{y}, f(\boldsymbol{x}) - f(\boldsymbol{y}) \rangle \geq 0, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in C, \tag{24}$$

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

$$Graph(\mathcal{T}) = \{(\boldsymbol{x}, \boldsymbol{u}) \mid \boldsymbol{x} \in C, \boldsymbol{u} \in \mathcal{T}(\boldsymbol{x})\}. \tag{25}$$

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

Proposition B.5

If $\mathcal{T}$ is a maximal monotone operator, the 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 $f : C \to \mathbb{R}^n$ is convex, then $\partial f(\boldsymbol{x})$ is a monotone operator. If f is further μ-strongly convex, then

$$\langle \boldsymbol{x}_1 - \boldsymbol{x}_2, \mathbf{g}_1 - \mathbf{g}_2 \rangle \geq \mu \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2 \tag{26}$$

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

Bregman Distance

Definition B.13 (Bregman Distance)

Given a differentiable convex function ϕ, the associated Bregman distance is defined as

$$D_\phi(\boldsymbol{y}, \boldsymbol{x}) = \phi(\boldsymbol{y}) - \phi(\boldsymbol{x}) - \langle \nabla \phi(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle. \tag{27}$$

If ϕ is convex but not differentiable, then the associated Bregman Distance is defined as

$$D_\phi^{\boldsymbol{v}}(\boldsymbol{y}, \boldsymbol{x}) = \phi(y) - \phi(\boldsymbol{x}) - \langle \boldsymbol{v}, \boldsymbol{y} - \boldsymbol{x} \rangle, \tag{28}$$

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

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

$$D_\phi(\boldsymbol{y}, \boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2. \tag{29}$$

Bregman Distance

Lemma B.1

The Bregman distance D_ϕ has the following properties:

1. *When ϕ is μ-strongly convex, we have*

$$D_\phi(\boldsymbol{y}, \boldsymbol{x}) \geq \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2. \tag{30}$$

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

$$\begin{aligned}\langle \nabla\phi(\boldsymbol{u}) - \nabla\phi(\boldsymbol{v}), \boldsymbol{w} - \boldsymbol{u}\rangle &= D_\phi(\boldsymbol{w}, \boldsymbol{v}) \\ &\quad - \Big(D_\phi(\boldsymbol{w}, \boldsymbol{u}) + D_\phi(\boldsymbol{u}, \boldsymbol{v})\Big).\end{aligned} \tag{31}$$

Conjugate Function

Definition B.14 (Conjugate Function)

Given $f : C \to \mathbb{R}$, its conjugate function is defined as

$$f^*(\boldsymbol{u}) = \sup_{\boldsymbol{z} \in C} \big(\langle \boldsymbol{z}, \boldsymbol{u} \rangle - f(\boldsymbol{z}) \big). \tag{32}$$

The domain of f^* is

$$dom\, f^* = \{\boldsymbol{u} \mid f^*(\boldsymbol{u}) < +\infty\}. \tag{33}$$

Properties of Conjugate Function

Proposition B.7 (Properties of Conjugate Function)

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

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

Proposition B.8 (Fenchel-Young Inequality)

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

$$f(\boldsymbol{x}) + f^*(\boldsymbol{y}) \geq \langle \boldsymbol{x}, \boldsymbol{y} \rangle. \tag{34}$$

Lagrangian Function

Definition B.15 (Lagrangian Function)

Given a constrained problem:

$$\begin{aligned} \min_{\boldsymbol{x}\in\mathbb{R}^n} \quad & f(\boldsymbol{x}) \\ s.t. \quad & \mathbf{A}\boldsymbol{x} = \boldsymbol{b}, \\ & \mathbf{g}(\boldsymbol{x}) \leq \mathbf{0}, \end{aligned} \tag{35}$$

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

$$L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) = f(\boldsymbol{x}) + \langle \boldsymbol{u}, \mathbf{A}\boldsymbol{x} - \boldsymbol{b}\rangle + \langle \boldsymbol{v}, \mathbf{g}(\boldsymbol{x})\rangle, \tag{36}$$

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

Lagrange Dual Function

Definition B.16 (Lagrange Dual Function)

Given a constrained problem (35)*, the Lagrange dual function is* $d(\boldsymbol{u}, \boldsymbol{v}) = \min_{\boldsymbol{x} \in C} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})$*, where* C *is the intersection of the domains of* f *and* g*. The domain of the dual function is*

$$\mathcal{D} = \left\{ (\boldsymbol{u}, \boldsymbol{v}) \mid \boldsymbol{v} \geq 0, d(\boldsymbol{u}, \boldsymbol{v}) > -\infty \right\}. \tag{37}$$

Definition B.17 (Dual Problem)

Given a constrained problem (35)*, the dual problem is*

$$\max_{\boldsymbol{u}, \boldsymbol{v}} d(\boldsymbol{u}, \boldsymbol{v}), \quad s.t. \quad (\boldsymbol{u}, \boldsymbol{v}) \in \mathcal{D}. \tag{38}$$

Correspondingly, (35) *is called the primal problem.*

Slater's Condition

Definition B.18 (Slater's Condition)

For convex primal problem (35)*, if there exists an* $\boldsymbol{x}_0$ *such that*

$$\mathbf{A}\boldsymbol{x}_0 = \boldsymbol{b}, \tag{39}$$

$$g_i(\boldsymbol{x}_0) \leq 0, \forall i \in \mathcal{I}_1, \tag{40}$$

$$g_j(\boldsymbol{x}_0) < 0, \forall j \in \mathcal{I}_2, \tag{41}$$

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. *$d(\boldsymbol{u}, \boldsymbol{v})$ is always a concave function, even if the primal problem* (35) *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 $\boldsymbol{x}(\boldsymbol{u}, \boldsymbol{v}) \in \operatorname{argmin}_{\boldsymbol{x} \in C} L(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})$, then*

$$\big(\mathbf{A}\boldsymbol{x}(\boldsymbol{u}, \boldsymbol{v}) - \boldsymbol{b}, \mathbf{g}(\boldsymbol{x}(\boldsymbol{u}, \boldsymbol{v}))\big) \in \partial d(\boldsymbol{u}, \boldsymbol{v}). \tag{42}$$

Properties of Dual Problem

Proof Sketch of Proposition B.9.2

We consider a problem with inequality constraints:

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

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 $\boldsymbol{x}$ has no solution, then we can say v is a lower bound of f:

$$\begin{cases} f(\boldsymbol{x}) < v \\ g_i(\boldsymbol{x}) \leq 0, i \in [m] \end{cases} \tag{43}$$

Properties of Dual Problem

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

If (43) has a solution, then, for any $\boldsymbol{\lambda} \geq \mathbf{0}$, the following equation of $\boldsymbol{x}$

$$f(\boldsymbol{x}) + \sum_{i \in [m]} \lambda_i g_i(\boldsymbol{x}) < v \tag{44}$$

has a solution. According to the equivalence of contrapositives, we have: For any $\boldsymbol{\lambda} \geq \mathbf{0}$, if (44) has no solution, then (43) has no solution. On the other hand, (44) has no solution for any given $\boldsymbol{\lambda} \geq 0$ *iff* the following inequality holds for any given $\boldsymbol{\lambda} \geq 0$:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) + \sum_{i \in [m]} \lambda_i g_i(\boldsymbol{x}) \geq v. \tag{45}$$

Properties of Dual Problem

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

Combing the above results, we have: If (45) 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.

$$v^* = \max_{\boldsymbol{\lambda} \geq 0} \underbrace{\left(\min_{\boldsymbol{x}} \left[\overbrace{f(x) + \sum_{i \in [m]} \lambda_i g_i(\boldsymbol{x})}^{L(\boldsymbol{x},\boldsymbol{\lambda})} \right] \right)}_{d(\boldsymbol{\lambda}) := \min_{\boldsymbol{x}} L(\boldsymbol{x},\boldsymbol{\lambda})}. \tag{46}$$

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

$$\min_{\boldsymbol{x}^*} f(\boldsymbol{x}^*) \geq \max_{\boldsymbol{\lambda}^*} d(\boldsymbol{\lambda}^*). \tag{47}$$

KKT Point and KKT Condition

Definition B.19 (KKT Point and KKT Condition)

$(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v})$ *is called a Karush-Kuhn-Tucker (KKT) point of problem* (35) *if*

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

The above conditions are called the KKT condition of problem (35). They are the optimality condition of problem (35) when problem (35) is convex and satisfies the Slater's condition.

KKT Point and KKT Condition

Proposition B.10

When $f(\boldsymbol{x})$ *and* $g_i(\boldsymbol{x}), i \in [p]$ *in problem* (35) *are all convex, then*

1. *every KKT point is a saddle point of the Lagrangian function, and*
2. $(\boldsymbol{x}^*, \boldsymbol{u}^*, \boldsymbol{v}^*)$ *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 $\mathcal{S}$ of $\mathbb{R}^n$ 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 conv$(\mathcal{X})$, is the set of all convex combinations of points in $\mathcal{X}$:

$$conv(\mathcal{X}) = \left\{ \sum_{i=1}^{k} \alpha_i \boldsymbol{x}_i \mid \boldsymbol{x}_i \in \mathcal{X}, \alpha_i \geq 0, i \in [k], \sum_{i=1}^{k} \alpha_i = 1 \right\}. \tag{48}$$

Danskin's Theorem

Theorem B.1 (Danskin's Theorem)

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

$$\mathcal{Z}(\boldsymbol{x}) = \left\{ \bar{\boldsymbol{z}} \mid \phi(\boldsymbol{x}, \bar{\boldsymbol{z}}) = \max_{\boldsymbol{z} \in \mathcal{Z}} \phi(\boldsymbol{x}, \boldsymbol{z}) \right\}. \tag{49}$$

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

$$\partial f(\boldsymbol{x}) = conv\left\{ \nabla_x \phi(\boldsymbol{x}, \boldsymbol{z}) \mid \boldsymbol{z} \in \mathcal{Z}(\boldsymbol{x}) \right\}, \forall \boldsymbol{x} \in \mathbb{R}^n. \tag{50}$$

Saddle Point

Definition B.22 (Saddle Point)

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

$$f(\boldsymbol{x}^*, \boldsymbol{\lambda}) \leq f(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) \leq f(\boldsymbol{x}, \boldsymbol{\lambda}^*), \forall \boldsymbol{x} \in C, \boldsymbol{\lambda} \in D. \tag{51}$$

Hoffman's Bound

Lemma B.2 (Hoffman's Bound)

Consider the non-empty polyhedron

$$\mathcal{X} = \{\boldsymbol{x} \mid \mathbf{A}\boldsymbol{x} = \boldsymbol{a}, \mathbf{B}\boldsymbol{x} \leq \boldsymbol{b}\}. \tag{52}$$

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

$$dist(\boldsymbol{x}, \mathcal{X})^2 \leq \theta^2 \left(\|\mathbf{A}\boldsymbol{x} - \boldsymbol{a}\|^2 + \|[\mathbf{B}\boldsymbol{x} - \boldsymbol{b}]_+\|^2 \right)^2, \tag{53}$$

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

Outline

Several Functions

Definition C.1 (Proper Function)

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

Definition C.2 (Lower Semicontinuous Function)

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

$$\liminf_{\boldsymbol{x} \to \boldsymbol{x}_0} g(\boldsymbol{x}) \geq g(\boldsymbol{x}_0). \tag{54}$$

Definition C.3 (Coercive Function)

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

Subdifferential

Definition C.4 (Subdifferential)

Let f be a proper and lower semicontinuous function.

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

$$\liminf_{\boldsymbol{y} \neq \boldsymbol{x}, \boldsymbol{y} \to \boldsymbol{x}} \frac{f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \boldsymbol{u}, \boldsymbol{y} - \boldsymbol{x} \rangle}{\|\boldsymbol{y} - \boldsymbol{x}\|} \geq 0. \tag{55}$$

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

$$\begin{aligned} \partial f(\boldsymbol{x}) = \Big\{ \boldsymbol{u} \in \mathbb{R}^n \,|\, \exists \boldsymbol{x}_k \to \boldsymbol{x}, f(\boldsymbol{x}_k) \to f(\boldsymbol{x}), \\ \boldsymbol{u}_k \in \hat{\partial} f(\boldsymbol{x}_k) \to \boldsymbol{u}, k \to \infty \Big\}. \end{aligned} \tag{56}$$

Critical Point and Properties of Subdifferential

Definition C.5 (Critical Point)

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

Lemma C.1

Some properties of subdifferential:

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

$$\partial(f + g)(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) + \partial g(\boldsymbol{x}). \tag{57}$$

Outline

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.