AMA4850 Final Exam Review

ZHANG, Shijun

March 24, 2026

The summary may be updated later.

The final exam will take place on **Friday, May 8**, from **12:30 PM to 2:30 PM**, in **SH2** (Main Hall of Kwong On Jubilee Sports Centre, Communal Building). Please make sure to arrive on time.

This is an open-book exam, allowing all printed materials. However, **electronic devices are not permitted**, except for a basic calculator.

Given the space constraints, the following review summary focuses only on the core topics. While exam questions will be based on these key topics, they may also include related concepts. Please ensure a comprehensive review.

All starred sections (★) are particularly important and should be reviewed carefully.

Contents

1 Key Topics Covered

1.1 ★Convexity Analysis of a Function

- Students need to determine whether a given function is **convex**.
- **Review Focus:**
  - Computing **Hessian matrices** to verify convexity.
  - Applying **matrix norm and composition rules** in convexity analysis.

1.2 ★Linear Programming (LP) Duality

- Students should be able to formulate the **dual** of a given linear program.
- **Review Focus:**
  - Understanding the process of **constructing the dual** from the primal problem.
  - Verifying **strong duality** and ensuring consistency between the primal and dual solutions.

1.3 ★Semidefinite Programming (SDP) Duality

- Students need to formulate the **dual** of a given SDP problem.
- **Review Focus:**
  - Understanding **how to derive dual constraints**.
  - Applying **primal-dual relationships** and verifying **strong duality**.

1.4 ★Reformulating a Complex Optimization Problem as an SDP

- Students should be able to rewrite inequalities in terms of **positive semidefinite matrices**.
- **Review Focus:**
  - Converting problems into **SDP form**.
  - Using **matrix inequalities** to handle nonlinear constraints.

1.5 ★Finding Stationary Points and Classifying Their Nature

- Students need to find all **stationary points** of given functions and determine their nature (local minima, maxima, or saddle points).
- **Review Focus:**
  - Computing **gradient and Hessian matrices**.
  - Using **second-order conditions** to classify stationary points.

> Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. A point $\boldsymbol{x}^*$ is a **saddle point** if:
>
> - $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ (i.e., the gradient at $\boldsymbol{x}^*$ is zero),
> - $\boldsymbol{x}^*$ is *not* a local minimum or maximum, meaning that in some directions $f(\boldsymbol{x})$ increases while in others it decreases.

1.6 ★Convergence of the Steepest Descent Method

- Students should understand the behavior of the steepest descent method when applied to smooth functions, and be able to justify convergence to stationary points under fixed step sizes.
- **Review Focus:** Arguing convergence under a **constant stepsize** without assuming exact line search.

1.7 ★Quasi-Newton Methods

- Students should be familiar with the update mechanism and iterative procedure of Quasi-Newton Methods (like BFGS) in solving unconstrained optimization problems.
- **Review Focus:** Executing multiple iterations of Quasi-Newton Methods.

1.8 ★Karush-Kuhn-Tucker (KKT) Conditions

- Students should be able to formulate and interpret the KKT conditions for constrained optimization problems.
- **Review Focus:**
  - Writing down **first-order necessary conditions** for optimality.
  - Understanding the role of **Lagrange multipliers**.
  - Knowing when KKT conditions are **necessary and/or sufficient**.
  - Recognizing constraint qualifications such as **MFCQ** and **Slater's condition**.

1.9 ★Penalty and Barrier Methods

- Students should understand how to handle constraints by reformulating them into unconstrained problems using penalty or barrier terms.
- **Review Focus:**
  - Formulating **quadratic penalty functions** and understanding their effect on solution accuracy.
  - Applying **logarithmic barrier methods** for inequality constraints.
  - Understanding the behavior as penalty/barrier parameters change.

2 General Mathematical Notations Used

2.1 Sets and Spaces

- $\mathbb{N}$: the set of natural numbers (either $\{1, 2, 3, \cdots\}$ or $\{0, 1, 2, \cdots\}$ depending on the context). In this course, we define $\mathbb{N} = \{0, 1, 2, \cdots\}$ and use $\mathbb{N}^+ = \mathbb{N}_+ = \{1, 2, \cdots\}$ to denote the set of positive natural numbers.
- $\mathbb{Z}$: Set of integers, $\{\cdots, -2, -1, 0, 1, 2, \cdots\}$
- $\mathbb{Q}$: Set of rational numbers
- $\mathbb{R}$: Set of real numbers
- $\mathbb{R}_+$: Set of nonnegative real numbers, $[0, \infty)$
- $\mathbb{R}_{++}$: Set of positive real numbers, $(0, \infty)$
- $\mathbb{R}^n$: n-dimensional real vector space
- $\mathbb{S}^n$: Set of $n \times n$ symmetric matrices
- $\mathbb{S}^n_+$: Set of $n \times n$ symmetric positive semidefinite matrices
- $A \subseteq B$: Set A is a subset of set B
- $C^k(\Omega)$: Space of functions whose (partial) derivatives up to order k exist and are continuous on Ω
- $C(\Omega) = C^0(\Omega)$: Space of continuous functions on the domain ω

2.2 Vectors and Matrices

- $\boldsymbol{x} = (x_1, \cdots, x_n) = [x_1, \cdots, x_n]^\mathsf{T} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$: A column vector
- $\boldsymbol{A} = (a_{ij})_{1 \le i \le m,\, 1 \le j \le n} = (a_{i,j})_{1 \le i \le m,\, 1 \le j \le n} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{bmatrix} \in \mathbb{R}^{m \times n}$: An $m \times n$ matrix
- $\mathbf{0}$ or $\mathbf{0}_n$, $\mathbf{1}$ or $\mathbf{1}_n$: Zero vector, all-ones vector
- $\boldsymbol{A}^\mathsf{T}$: Transpose of matrix $\boldsymbol{A}$
- $\boldsymbol{A}^{-1}$: Inverse of matrix $\boldsymbol{A}$ (if it exists)
- $\boldsymbol{I}_n$ (or simply $\boldsymbol{I}$ when the dimension is clear): the $n \times n$ identity matrix
- $\operatorname{rank}(\boldsymbol{A})$: Rank of matrix $\boldsymbol{A}$
- $\operatorname{Tr}(\boldsymbol{A})$ (or $\operatorname{tr}\boldsymbol{A}$): Trace of matrix $\boldsymbol{A}$
- $\det(\boldsymbol{A})$: Determinant of matrix $\boldsymbol{A}$
- $\lambda_{\min}(\boldsymbol{A}), \lambda_{\max}(\boldsymbol{A})$: Minimum/maximum eigenvalue of matrix $\boldsymbol{A}$
- $\boldsymbol{A} \succeq 0$: Matrix $\boldsymbol{A}$ is positive semidefinite
- $\boldsymbol{A} \succ 0$: Matrix $\boldsymbol{A}$ is positive definite
- $\boldsymbol{A} \preceq \boldsymbol{B}$ (or $\boldsymbol{B} \succeq \boldsymbol{A}$): Matrix inequality, i.e., $\boldsymbol{B} - \boldsymbol{A} \succeq 0$
- $\boldsymbol{A} \prec \boldsymbol{B}$ (or $\boldsymbol{B} \succ \boldsymbol{A}$): Matrix inequality, i.e., $\boldsymbol{B} - \boldsymbol{A} \succ 0$

2.3 Functions and Derivatives

- $f : \mathbb{R}^n \to \mathbb{R}$: A real-valued function on $\mathbb{R}^n$
- $\nabla f(\boldsymbol{x})$: Gradient of f at $\boldsymbol{x}$
- $\nabla^2 f(\boldsymbol{x})$: Hessian matrix of f at $\boldsymbol{x}$
- $\operatorname{dom}(f)$: Domain of function f

- $\text{epi}(f)$: Epigraph of function f
- $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y} = \boldsymbol{y}^\top \boldsymbol{x}$: Inner product of vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^n$

2.4 Optimization Notation

- $\min_{\boldsymbol{x}} f(\boldsymbol{x})$ (or $\text{Minimize}_{\boldsymbol{x}} f(\boldsymbol{x})$): Minimize the function $f(\boldsymbol{x})$
- $\max_{\boldsymbol{x}} f(\boldsymbol{x})$ (or $\text{Maximize}_{\boldsymbol{x}} f(\boldsymbol{x})$): Maximize the function $f(\boldsymbol{x})$
- $\arg\min_{\boldsymbol{x}} f(\boldsymbol{x})$: Value(s) of $\boldsymbol{x}$ that minimize $f(\boldsymbol{x})$
- $\arg\max_{\boldsymbol{x}} f(\boldsymbol{x})$: Value(s) of $\boldsymbol{x}$ that maximize $f(\boldsymbol{x})$

2.5 Logic and Symbols

- $\forall$: "For all"
- $\exists$: "There exists"
- $\Rightarrow$: Implies
- $\Leftrightarrow$: If and only if

3 Important Formulas and Results

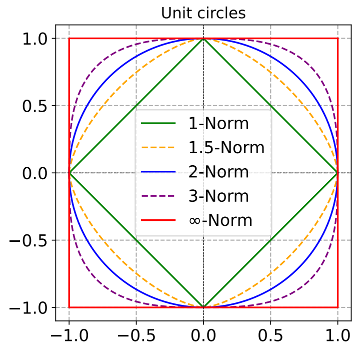
3.1 Vector norms

Definition: A function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ is called a (vector) **norm** if:

- $\|\boldsymbol{x}\| \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$.
- $\|\boldsymbol{x}\| = 0$ if and only if $\boldsymbol{x} = \boldsymbol{0}$.
- $\|\alpha \boldsymbol{x}\| = |\alpha| \cdot \|\boldsymbol{x}\|$ for any $\alpha \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^n$.
- $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.

The following are some commonly used norms:

- ⋆ ℓ^1 norm: $\|\boldsymbol{x}\|_1 := \sum_{i=1}^n |x_i|$.
- ⋆ ℓ^2 norm: $\|\boldsymbol{x}\|_2 := \sqrt{\sum_{i=1}^n |x_i|^2}$.
- ⋆ ℓ^∞ norm: $\|\boldsymbol{x}\|_\infty := \max_{1 \leq i \leq n} |x_i|$.



$$\|\boldsymbol{x}\|_\infty \leq \|\boldsymbol{x}\|_2 \leq \sqrt{n}\|\boldsymbol{x}\|_\infty, \quad \forall \boldsymbol{x} \in \mathbb{R}^n.$$
$$\|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}\|_1 \leq \sqrt{n}\|\boldsymbol{x}\|_2, \quad \forall \boldsymbol{x} \in \mathbb{R}^n.$$

$$\|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|_2 \to 0 \quad \text{as } k \to \infty \quad \Longleftrightarrow \quad \|\boldsymbol{x}^{(k)} - \boldsymbol{x}^*\|_1 \to 0 \quad \text{as } k \to \infty$$

3.2 Matrix norms

Definition: A function $\|\cdot\| : \mathbb{R}^{m \times n} \to \mathbb{R}$ is called a **matrix norm** if:

- $\|\boldsymbol{A}\| \geq 0$ for all $\boldsymbol{A} \in \mathbb{R}^{m \times n}$.
- $\|\boldsymbol{A}\| = 0$ if and only if $\boldsymbol{A} = \boldsymbol{0}$.

- $\|\alpha \boldsymbol{A}\| = |\alpha| \cdot \|\boldsymbol{A}\|$ for any $\alpha \in \mathbb{R}$ and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$.
- $\|\boldsymbol{A}+\boldsymbol{B}\| \le \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$ for any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$.

Let $\|\cdot\|$ be a vector norm. Then

$$\|\boldsymbol{A}\| := \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}\boldsymbol{x}\| = \max_{\boldsymbol{x}\neq \boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}$$

defines a matrix norm $\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}$. Such a matrix norm is said to be **induced by the vector norm** $\|\cdot\|$. Moreover, we have

$$\|\boldsymbol{A}\boldsymbol{x}\| \le \|\boldsymbol{A}\| \cdot \|\boldsymbol{x}\| \quad \text{for all } \boldsymbol{A} \in \mathbb{R}^{m\times n} \quad \text{and} \quad \boldsymbol{x} \in \mathbb{R}^n.$$

and

$$\|\boldsymbol{A}\boldsymbol{B}\| \le \|\boldsymbol{A}\| \cdot \|\boldsymbol{B}\| \quad \text{for all } \boldsymbol{A} \in \mathbb{R}^{m\times n} \quad \text{and} \quad \boldsymbol{B} \in \mathbb{R}^{n\times k}.$$

Examples of induced norms for $\boldsymbol{A} \in \mathbb{R}^{m\times n}$:

- Induced by ℓ^1 norm (maximum of the ℓ^1 norms of columns):

$$\|\boldsymbol{A}\|_1 = \max_{\|\boldsymbol{x}\|_1=1} \|\boldsymbol{A}\boldsymbol{x}\|_1 = \max_j \sum_{i=1}^m |a_{i,j}|.$$

- Induced by ℓ^2 norm (Square root of maximum eigenvalue of $\boldsymbol{A}^\top \boldsymbol{A}$):

$$\|\boldsymbol{A}\|_2 = \max_{\|x\|_2=1} \|\boldsymbol{A}\boldsymbol{x}\|_2 = \sqrt{\lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A})}.$$

Remark:

$$\|\boldsymbol{A}\|_2 = \sqrt{\lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A})} = \sqrt{\lambda_{\max}(\boldsymbol{A}\boldsymbol{A}^\top)} = \|\boldsymbol{A}^\top\|_2$$

- Induced by ℓ^∞ norm (maximum of the ℓ^1 norms of rows):

$$\|\boldsymbol{A}\|_\infty = \max_{\|\boldsymbol{x}\|_\infty=1} \|\boldsymbol{A}\boldsymbol{x}\|_\infty = \max_i \sum_{j=1}^n |a_{i,j}|.$$

3.3 Compactness and Existence of minimizers

Definition: A set $\Omega \subseteq \mathbb{R}^n$ is said to be **closed** if it contains all the limits of convergent sequences of points in Ω.

Definition: A set X is said to be **compact** if every open cover of X has a finite subcover. That is, if $\{U_\alpha\}_{\alpha\in A}$ is a collection of open sets such that: $X \subseteq \bigcup_{\alpha\in A} U_\alpha$, then there exists a finite subset $\{U_{\alpha_1}, U_{\alpha_2}, \cdots, U_{\alpha_n}\}$ such that: $X \subseteq \bigcup_{i=1}^n U_{\alpha_i}$.

$$X \in \mathbb{R}^n \text{ is closed and bounded} \qquad \Longleftrightarrow \qquad X \in \mathbb{R}^n \text{ is compact.}$$

Theorem (Existence of minimizers): Let $\Omega \subseteq \mathbb{R}^n$ be a **nonempty compact set** and let f be **continuous** on Ω. Then f achieves its infimum value over Ω, i.e., there exists $\boldsymbol{x}^* \in \Omega$ so that

$$f(\boldsymbol{x}^*) = \inf\{f(\boldsymbol{x}) : \boldsymbol{x} \in \Omega\}.$$

Theorem (Bolzano-Weierstrass): Each **infinite bounded** sequence in $\mathbb{R}^n$ has a convergent subsequence.

3.4 Positive (semi)definite matrices

Definition (Positive semidefinite matrices): Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be symmetric. We say that $\boldsymbol{A}$ is **positive semidefinite** (denoted by $\boldsymbol{A} \succeq 0$) if

$$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \geq 0 \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^n.$$

Theorem : Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be **symmetric**. The following statements are equivalent:

- All eigenvalues of $\boldsymbol{A}$ are **nonnegative**.
- There exists $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ so that $\boldsymbol{A} = \boldsymbol{M}^\top \boldsymbol{M}$.
- $\boldsymbol{A}$ is **positive semidefinite**.

Definition (Positive definite matrices): A symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is called **positive definite** (denoted by $\boldsymbol{A} \succ 0$) if

$$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} > 0 \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}.$$

$$\boldsymbol{A} \succeq 0 \quad \text{and} \quad \det(\boldsymbol{A}) \neq 0 \quad \Longrightarrow \quad \boldsymbol{A} \succ 0.$$

Note: Let $\boldsymbol{A} \succ 0$, then:

- All diagonal entries are positive.
- $\boldsymbol{A}^{-1} \succ 0$.
- $\lambda_{\min}(\boldsymbol{A}) = \inf\{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} : \|\boldsymbol{x}\|_2 = 1\}$.
- $\|\boldsymbol{A}\|_2 = \lambda_{\max}(\boldsymbol{A}) = [\lambda_{\min}(\boldsymbol{A}^{-1})]^{-1}$.

Theorem: Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be symmetric. The following statements are equivalent:

- All eigenvalues of $\boldsymbol{A}$ are **positive**.
- There exists an **invertible matrix** $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ so that $\boldsymbol{A} = \boldsymbol{M}^\top \boldsymbol{M}$.
- $\boldsymbol{A}$ is **positive definite**.

3.4.1 ★Diagonal dominance

A square matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is called **strictly row diagonally dominant** if

$$|a_{i,i}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}|,$$

i.e., the diagonal entry is larger than the (absolute) sum of all other entries in that row. If the inequality is $\geq$, then it is called **(weakly) row diagonally dominant**.

Theorem: Suppose $\boldsymbol{A} = (a_{i,j}) \in \mathbb{R}^{n\times n}$ is symmetric and strictly row diagonally dominant, and $a_{i,i} > 0$ for $i = 1, 2, \cdots, n$. Then all eigenvalues of $\boldsymbol{A}$ is positive and hence $\boldsymbol{A}$ is positive definite.

Theorem: Suppose $\boldsymbol{A} = (a_{i,j}) \in \mathbb{R}^{n\times n}$ is symmetric and (weakly) row diagonally dominant, and $a_{i,i} \geq 0$ for $i = 1, 2, \cdots, n$. Then all eigenvalues of $\boldsymbol{A}$ is non-negative and hence $\boldsymbol{A}$ is positive semidefinite.

The matrix

$$\boldsymbol{A} = \begin{bmatrix} 2+\delta & -1 & 0 \\ -1 & 2+\delta & -1 \\ 0 & -1 & 2+\delta \end{bmatrix}$$

is strictly row diagonally dominant for any $\delta > 0$ and is therefore positive definite. Matrices of this form arise often when partial differential equations.

Note that it is positive definite when $\delta = 0$, but the diagonal dominance theorem only tells us that $\lambda \geq 0$, not $\lambda > 0$ since $2 - |-1| - |-1| = 0$.

3.4.2 ★Sylvester's criterion

Theorem: Suppose $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is symmetric, then the following statements are equivalent:

(i) $\boldsymbol{A}$ is positive definite.

(ii) All the eigenvalues of $\boldsymbol{A}$ are positive.

(iii) **Sylvester's criterion:** For $1 \leq k \leq n$, $\det(\boldsymbol{A}_k) > 0$ for all the principal minor submatrix $\boldsymbol{A}_k$ (the k by k submatrix starting from the $(1,1)$ entry down to the (k,k) entry).

The matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 8 \end{bmatrix}$$

is positive definite since $1 > 0$,

$$\det \begin{bmatrix} 1 & -2 \\ -2 & 5 \end{bmatrix} = 1 > 0 \quad \text{and} \quad \det \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 8 \end{bmatrix} = 8 > 0.$$

3.5 Taylor's theorem

Theorem (Taylor's theorem in 1D with remainder tern): Suppose that f is $(n+1)$ times differentiable on an open interval containing $[a, b]$. Then

$$f(b) = f(a) + f'(a)(b-a) + \frac{f''(a)}{2!}(b-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(b-a)^n + \frac{f^{(n+1)}(\xi)}{(n+1)!}(b-a)^{n+1}$$

for some $\xi \in (a, b)$.

Theorem (Taylor's theorem in $\mathbb{R}^n$ with remainder term):

- Let $f \in C^1(\mathbb{R}^n)$, $\boldsymbol{x}$ and $\boldsymbol{y} \in \mathbb{R}^n$. Then there exists $\boldsymbol{\xi} \in \{(1-s)\boldsymbol{x} + s\boldsymbol{y} : s \in (0,1)\}$ such that

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + [\nabla f(\boldsymbol{\xi})]^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{x}).$$

- Let $f \in C^2(\mathbb{R}^n)$, $\boldsymbol{x}$ and $\boldsymbol{y} \in \mathbb{R}^n$. Then there exists $\boldsymbol{\xi} \in \{(1-s)\boldsymbol{x} + s\boldsymbol{y} : s \in (0,1)\}$ such that

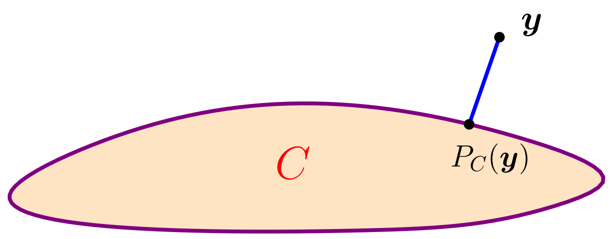
$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + [\nabla f(x)]^\mathsf{T}(\boldsymbol{y}-\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{x})^\mathsf{T}\nabla^2 f(\boldsymbol{\xi})(\boldsymbol{y}-\boldsymbol{x}).$$

3.6 Projection and Separation theorem

Theorem: Let $C \subseteq \mathbb{R}^n$ be a **nonempty closed convex** set and $\boldsymbol{y} \in \mathbb{R}^n$. Then there **exists a unique solution** to the following optimization problem:

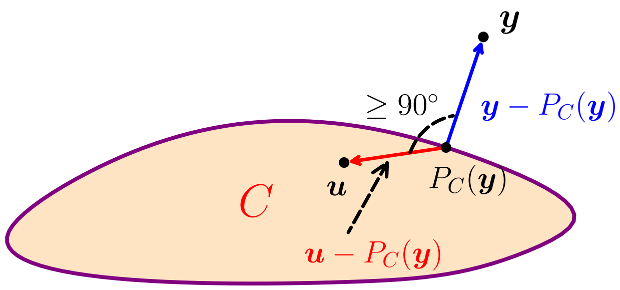
$$\text{Minimize } \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \quad \text{subject to} \quad \boldsymbol{x} \in C. \tag{1}$$

The unique solution of (1) is called the projection of $\boldsymbol{y}$ onto C, denoted by $P_C(\boldsymbol{y})$.



Theorem (Projection Inequality): Let $C \subseteq \mathbb{R}^n$ be a nonempty closed convex set, $\boldsymbol{y} \in \mathbb{R}^n$, and $\boldsymbol{u} \in C$. Then

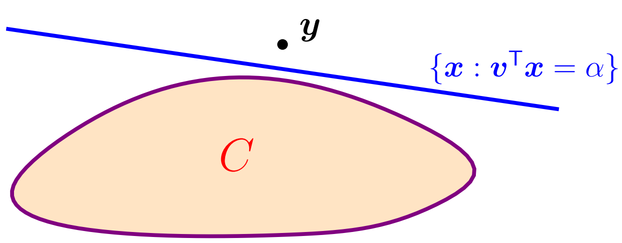
$$\left(\boldsymbol{y} - P_C(\boldsymbol{y})\right)^\mathsf{T}\left(\boldsymbol{u} - P_C(\boldsymbol{y})\right) \leq 0.$$



Theorem (Separation theorem): Let $C \subseteq \mathbb{R}^n$ be a **nonempty closed convex** set and $\boldsymbol{y} \in \mathbb{R}^n \setminus C$. Then there exists $\boldsymbol{v} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$ and $\alpha \in \mathbb{R}$ such that

$$\boldsymbol{v}^\mathsf{T}\boldsymbol{y} > \alpha > \boldsymbol{v}^\mathsf{T}\boldsymbol{u}$$

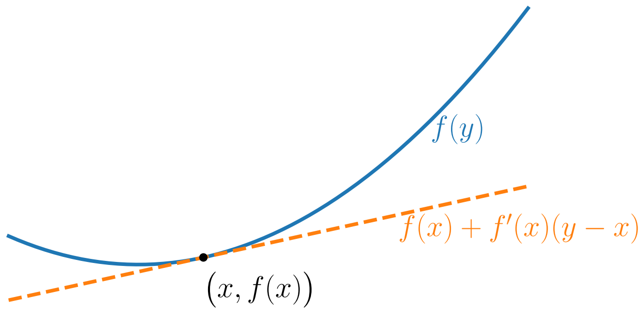
for all $\boldsymbol{u} \in C$.



3.7 ★Characterization of Convexity

Theorem (First-Order Characterization): Suppose that $f \in C^1(\Omega)$ with $\Omega \subseteq \mathbb{R}^n$ being open and convex. Then f is convex if and only if

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + [\nabla f(\boldsymbol{x})]^\mathsf{T}(\boldsymbol{y} - \boldsymbol{x}) \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \Omega.$$



Theorem (Second-Order Characterization): Suppose that $f \in C^2(\Omega)$ with $\Omega \subseteq \mathbb{R}^n$ being open and convex. Then f is convex if and only if $\nabla^2 f(\boldsymbol{x}) \succeq 0$ for all $\boldsymbol{x} \in \Omega$.

Remark: If f is convex on a convex set Ω and continuous on $\overline{\Omega}$, then f is convex on $\overline{\Omega}$. For example, $f(x) = x^s$ with $s \geq 1$ is convex on $(0, \infty)$ since $f''(x) = s(s-1)x^{s-2} \geq 0$, and it is continuous on $[0, \infty)$. Therefore, f is convex on $[0, \infty)$.

3.8 ★Calculus of Convex Functions

Proposition: Let $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$, $g : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ both be convex, $\boldsymbol{A} \in \Omega \subseteq \mathbb{R}^{n \times p}$, and $\alpha > 0$. Then the following functions are convex:

- $f + g$;
- αf;
- $f \circ \boldsymbol{A}$;
- $\max\{f, g\}$.

Proposition: Vector norms are convex.

Proposition: Suppose each component f_i of $\boldsymbol{f} = (f_1, \cdots, f_k) : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^k$ is convex, and $g : \mathbb{R}^k \to \mathbb{R}$ is convex and nondecreasing in each argument. Then $g \circ \boldsymbol{f}$ is convex.

3.9 ★Linear Programming (LP)

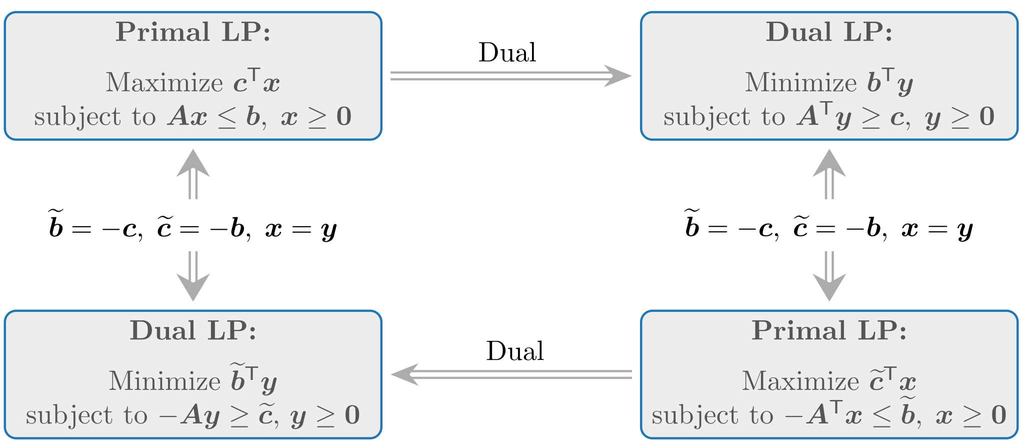
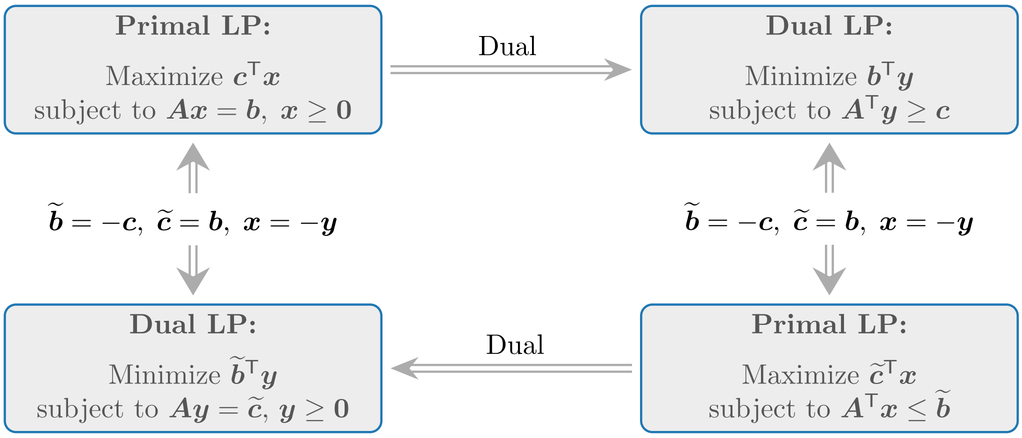


Table 1: Primal and Dual LP Problems

Primal	**Dual**
Maximize $\boldsymbol{c}^\top \boldsymbol{x}$ subject to $\boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b}, \boldsymbol{x} \ge \boldsymbol{0}$	Minimize $\boldsymbol{b}^\top \boldsymbol{y}$ subject to $\boldsymbol{A}^\top \boldsymbol{y} \ge \boldsymbol{c}, \boldsymbol{y} \ge \boldsymbol{0}$
Maximize $\boldsymbol{c}^\top \boldsymbol{x}$ subject to $\boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b}$	Minimize $\boldsymbol{b}^\top \boldsymbol{y}$ subject to $\boldsymbol{A}^\top \boldsymbol{y} = \boldsymbol{c}, \boldsymbol{y} \ge \boldsymbol{0}$
Maximize $\boldsymbol{c}^\top \boldsymbol{x}$ subject to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \ge \boldsymbol{0}$	Minimize $\boldsymbol{b}^\top \boldsymbol{y}$ subject to $\boldsymbol{A}^\top \boldsymbol{y} \ge \boldsymbol{c}$



Weak Duality: Let v_p be the optimal value of the primal LP and v_d be the optimal value of the dual LP. Then, we have

$$\begin{aligned}
v_p = \sup_{\boldsymbol{x}\ge\boldsymbol{0}}\{\boldsymbol{c}^\top\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\} &= \sup_{\boldsymbol{x}\ge\boldsymbol{0}}\inf_{\boldsymbol{y}}\{\boldsymbol{c}^\top\boldsymbol{x} - \boldsymbol{y}^\top(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\} \\
&= \sup_{\boldsymbol{x}\ge\boldsymbol{0}}\inf_{\boldsymbol{y}}\{\boldsymbol{b}^\top\boldsymbol{y} + \boldsymbol{x}^\top(\boldsymbol{c} - \boldsymbol{A}^\top\boldsymbol{y}) : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\} \\
&\le \sup_{\boldsymbol{x}\ge\boldsymbol{0}}\inf_{\boldsymbol{y}}\{\boldsymbol{b}^\top\boldsymbol{y} + \boldsymbol{x}^\top(\boldsymbol{c} - \boldsymbol{A}^\top\boldsymbol{y}) : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},\ \boldsymbol{c} \le \boldsymbol{A}^\top\boldsymbol{y}\} \\
&\le \inf_{\boldsymbol{y}}\sup_{\boldsymbol{x}\ge\boldsymbol{0}}\{\boldsymbol{b}^\top\boldsymbol{y} + \boldsymbol{x}^\top(\boldsymbol{c} - \boldsymbol{A}^\top\boldsymbol{y}) : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b},\ \boldsymbol{c} \le \boldsymbol{A}^\top\boldsymbol{y}\} \\
&\le \inf_{\boldsymbol{y}}\sup_{\boldsymbol{x}\ge\boldsymbol{0}}\{\boldsymbol{b}^\top\boldsymbol{y} + \boldsymbol{x}^\top(\boldsymbol{c} - \boldsymbol{A}^\top\boldsymbol{y}) : \boldsymbol{c} \le \boldsymbol{A}^\top\boldsymbol{y}\} \\
&= \inf_{\boldsymbol{y}}\{\boldsymbol{b}^\top\boldsymbol{y} : \boldsymbol{c} \le \boldsymbol{A}^\top\boldsymbol{y}\} =: v_d.
\end{aligned}$$

Theorem (Strong Duality): Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}^m$ and $\boldsymbol{c} \in \mathbb{R}^n$. Consider

$$v_p := \sup\{\boldsymbol{c}^\top\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \ge \boldsymbol{0}\}, \quad v_d := \inf\{\boldsymbol{b}^\top\boldsymbol{y} : \boldsymbol{c} \le \boldsymbol{A}^\top\boldsymbol{y}\}.$$

Suppose that **either**

- there exists $\widehat{\boldsymbol{x}} \ge \boldsymbol{0}$ with $\boldsymbol{A}\widehat{\boldsymbol{x}} = \boldsymbol{b}$; **or**
- there exists $\widehat{\boldsymbol{y}}$ with $\boldsymbol{c} \le \boldsymbol{A}^\top\widehat{\boldsymbol{y}}$.

Then $v_p = v_d$ and both optimal values are attained when finite.

Note: The strong duality theorem implies that if either v_p or v_d is finite, then $v_p = v_d \in \mathbb{R}$, and both optimal solutions are attained.

Remarks on Duality: Let

$$v_p := \sup\{\boldsymbol{c}^\top \boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}\}, \quad v_d := \inf\{\boldsymbol{b}^\top \boldsymbol{y} : \boldsymbol{c} \leq \boldsymbol{A}^\top \boldsymbol{y}\}.$$

Based on weak and strong duality, the values of v_p and v_d fall into one of the following four cases:

(a) $v_p = v_d \in \mathbb{R}$ (both finite and attainable).

(b) $v_p = v_d = -\infty$ (primal infeasible and dual unbounded).

(c) $v_p = v_d = \infty$ (primal unbounded and dual infeasible).

(d) $v_p = -\infty \leq v_d = \infty$ (both infeasible).

3.10 ★Semidefinite Programming

A semidefinite programming (SDP) problem has the following standard form:

$$\begin{aligned} \underset{\boldsymbol{X} \in \mathbb{S}^n}{\text{Minimize}} \quad & \operatorname{tr}(\boldsymbol{C}\boldsymbol{X}) \\ \text{subject to} \quad & \operatorname{tr}(\boldsymbol{A}_i \boldsymbol{X}) = b_i, \quad i = 1, \cdots, m, \\ & \boldsymbol{X} \succeq 0, \end{aligned}$$

Here:

- $\mathbb{S}^n = \{\boldsymbol{M} \in \mathbb{R}^{n \times n} : \boldsymbol{M}^\top = \boldsymbol{M}\}$ is the space of all real symmetric matrices.
- For any $\boldsymbol{Y} \in \mathbb{R}^{n \times n}$, the trace of $\boldsymbol{Y}$ is given by $\operatorname{tr}(\boldsymbol{Y}) := \sum_{i=1}^n y_{i,i}$.
- $\boldsymbol{C}$ and $\boldsymbol{A}_i$ are given real symmetric matrices.

Theorem (Strong Duality for SDPs): Consider the following primal-dual SDP pairs:

Primal:

$$\min_{\boldsymbol{X} \in \mathbb{S}^n} \operatorname{tr}(\boldsymbol{C}\boldsymbol{X}) \quad \text{subject to} \quad \operatorname{tr}(\boldsymbol{A}_i \boldsymbol{X}) = b_i, \ i = 1, \cdots, m, \quad \boldsymbol{X} \succeq 0.$$

Dual:

$$\max_{\boldsymbol{y} \in \mathbb{R}^m} \boldsymbol{b}^\top \boldsymbol{y} \quad \text{subject to} \quad \boldsymbol{C} - \sum_{i=1}^m y_i \boldsymbol{A}_i \succeq 0.$$

Here, $\boldsymbol{C} \in \mathbb{S}^n$ and $\boldsymbol{A}_i \in \mathbb{S}^n$ for all i. Let v_p and v_d denote their optimal values, respectively. Then the following statements hold:

(a) If there exists $\boldsymbol{X} \succ 0$ such that $\operatorname{tr}(\boldsymbol{A}_i \boldsymbol{X}) = b_i$ for all i, then $v_p = v_d$ and v_d is attained when finite.

(b) If there exists $\boldsymbol{y} \in \mathbb{R}^m$ such that $\boldsymbol{C} - \sum_{i=1}^m y_i \boldsymbol{A}_i \succ 0$, then $v_p = v_d$ and v_p is attained when finite.

In the Strong Duality Theorem, the inequality $v_p \geq v_d$ holds universally, without any assumptions. This is known as **weak duality**.

Note: In the Strong Duality Theorem for SDPs, a positive definite condition is required, which is stricter than the original positive semidefinite constraint. Thus, the Strong Duality Theorem for SDPs does not exhibit as favorable properties as its the one for LPs.

3.11 Quadratic form

Note: Let $\boldsymbol{e}_i \in \mathbb{R}^n$ be a vector whose i-th entry is 1, with all other entries being 0. Then $\boldsymbol{e}_i\boldsymbol{e}_j^\top \in \mathbb{R}^{n\times n}$ is a matrix with (i,j)-entry being 1 and all other entries being 0. Thus, a matrix $\boldsymbol{A} = (a_{i,j})$ can be expressed as

$$\boldsymbol{A} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}\boldsymbol{e}_i\boldsymbol{e}_j^\top.$$

Let $\boldsymbol{A} = (a_{i,j}) \in \mathbb{S}^n$, $\boldsymbol{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n$, Then

$$\begin{aligned}\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} &= \boldsymbol{x}^\top\left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}\boldsymbol{e}_i\boldsymbol{e}_j^\top\right)\boldsymbol{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}\boldsymbol{x}^\top\boldsymbol{e}_i\boldsymbol{e}_j^\top\boldsymbol{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{i,j}x_ix_j \\ &= \sum_{i=1}^{n}\sum_{\substack{j=1\\ i=j}}^{n} a_{i,j}x_ix_j + \sum_{i=1}^{n}\sum_{\substack{j=1\\ i<j}}^{n} a_{i,j}x_ix_j + \sum_{i=1}^{n}\sum_{\substack{j=1\\ i>j}}^{n} a_{i,j}x_ix_j = \sum_{i=1}^{n} a_{i,i}x_i^2 + \sum_{i<j} 2a_{i,j}x_ix_j.\end{aligned}$$

3.12 ★Schur Complement

The following result is crucial in reformulating problems into SDPs.

Theorem: Let $\boldsymbol{A} \in \mathbb{S}^m$, $\boldsymbol{C} \in \mathbb{S}^n$, $\boldsymbol{B} \in \mathbb{R}^{m\times n}$, and $\boldsymbol{A} \succ 0$. Then

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{C} \end{bmatrix} \succeq 0 \quad\Longleftrightarrow\quad \boldsymbol{C} - \boldsymbol{B}^\top\boldsymbol{A}^{-1}\boldsymbol{B} \succeq 0.$$

Note: We call $\boldsymbol{C} - \boldsymbol{B}^\top\boldsymbol{A}^{-1}\boldsymbol{B}$ the Schur complement of $\boldsymbol{A}$ in $\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^\top & \boldsymbol{C} \end{bmatrix}$.

Special case $(m = n = 1)$ with $a > 0$:

$$\begin{bmatrix} a & b \\ b & c \end{bmatrix} \succeq 0 \quad\Longleftrightarrow\quad ac - b^2 \geq 0 \quad\Longleftrightarrow\quad c - b^2/a \geq 0.$$

Lemma: Suppose $a, c \in \mathbb{R}$ and $\boldsymbol{b} \in \mathbb{R}^n$. Then

$$\begin{bmatrix} a\boldsymbol{I}_{n-1} & \boldsymbol{b} \\ \boldsymbol{b}^\top & c \end{bmatrix} \succeq 0 \quad\Longleftrightarrow\quad ac \geq \boldsymbol{b}^\top\boldsymbol{b} = \|\boldsymbol{b}\|_2^2, \quad a \geq 0, \quad c \geq 0.$$

Here, $\boldsymbol{I}_{n-1} \in \mathbb{R}^{(n-1)\times(n-1)}$ denotes the identity matrix.

3.13 Properties of Matrix Trace

- **Linearity:**

$$\operatorname{tr}(\boldsymbol{A} + \boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{B}), \quad \operatorname{tr}(c\boldsymbol{A}) = c\operatorname{tr}(\boldsymbol{A})$$

  for any scalar c and matrices $\boldsymbol{A}, \boldsymbol{B}$ of the same size.

- **Cyclic Property:**

$$\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{B}\boldsymbol{A})$$

  holds when $\boldsymbol{A}\boldsymbol{B}$ and $\boldsymbol{B}\boldsymbol{A}$ are defined.

For $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n\times m}$, we have

$$\operatorname{tr}(\boldsymbol{AB}) = \sum_{i=1}^{m} (\boldsymbol{AB})_{i,i} = \sum_{i=1}^{m} \underbrace{\sum_{j=1}^{n} a_{i,j} b_{j,i}}_{(\boldsymbol{AB})_{i,i}} = \sum_{j=1}^{n} \underbrace{\sum_{i=1}^{m} b_{j,i} a_{i,j}}_{(\boldsymbol{BA})_{j,j}} = \sum_{j=1}^{n} (\boldsymbol{BA})_{j,j} = \operatorname{tr}(\boldsymbol{BA}).$$

Example 1:

$$\operatorname{Diag}(\boldsymbol{c}) = \begin{bmatrix} c_1 & 0 & \cdots & 0 \\ 0 & c_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & c_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} x_1 & * & \cdots & * \\ * & x_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ * & \cdots & * & x_n \end{bmatrix}.$$

$$\boldsymbol{c}^\top \boldsymbol{x} = \sum_{i=1}^{n} c_i x_i = \operatorname{tr}[\operatorname{Diag}(\boldsymbol{c})\boldsymbol{X}].$$

$$\operatorname{Diag}(\boldsymbol{c})\boldsymbol{X} = \begin{bmatrix} c_1 & 0 & \cdots & 0 \\ 0 & c_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & c_n \end{bmatrix} \begin{bmatrix} x_1 & * & \cdots & * \\ * & x_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ * & \cdots & * & x_n \end{bmatrix} = \begin{bmatrix} c_1x_1 & * & \cdots & * \\ * & c_2x_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & * \\ * & \cdots & * & c_nx_n \end{bmatrix}.$$

Let $\boldsymbol{e}_i$ be a vector whose i-th entry is 1, with all other entries being 0.

$$\operatorname{Diag}(\boldsymbol{c}) = \sum_{i=1}^{n} c_i \boldsymbol{e}_i \boldsymbol{e}_i^\top \quad \Longrightarrow$$

$$\begin{aligned} \operatorname{tr}\big(\operatorname{Diag}(\boldsymbol{c})\boldsymbol{X}\big) = \operatorname{tr}\left(\sum_{i=1}^{n} c_i \boldsymbol{e}_i \boldsymbol{e}_i^\top \boldsymbol{X}\right) &= \sum_{i=1}^{n} c_i \cdot \operatorname{tr}\left(\boldsymbol{e}_i \boldsymbol{e}_i^\top \boldsymbol{X}\right) \\ &= \sum_{i=1}^{n} c_i \cdot \operatorname{tr}\left(\boldsymbol{e}_i^\top \boldsymbol{X} \boldsymbol{e}_i\right) = \sum_{i=1}^{n} c_i x_i = \boldsymbol{c}^\top \boldsymbol{x}. \end{aligned}$$

Example 2: Suppose $\boldsymbol{X} \in \mathbb{S}^n$ and let $\boldsymbol{E}_{i,j}$ be the symmetric matrix that is $\frac{1}{2}$ at the (i,j) and (j,i) entries, and is zero otherwise, i.e., $\boldsymbol{E}_{i,j} = \frac{1}{2}\big(\boldsymbol{e}_j \boldsymbol{e}_i^\top + \boldsymbol{e}_i \boldsymbol{e}_j^\top\big)$. Then

$$x_{i,j} = \operatorname{tr}(\boldsymbol{E}_{i,j}\boldsymbol{X}).$$

Proof. Clearly,

$$x_{i,j} = \boldsymbol{e}_i^\top \boldsymbol{X} \boldsymbol{e}_j = \operatorname{tr}(\boldsymbol{e}_i^\top \boldsymbol{X} \boldsymbol{e}_j) = \operatorname{tr}(\boldsymbol{e}_j \boldsymbol{e}_i^\top \boldsymbol{X}).$$

Similarly,

$$x_{j,i} = \operatorname{tr}(\boldsymbol{e}_i \boldsymbol{e}_j^\top \boldsymbol{X}).$$

Thus,

$$\begin{aligned} x_{i,j} = \frac{1}{2}(x_{i,j} + x_{j,i}) &= \frac{1}{2}\Big(\operatorname{tr}(\boldsymbol{e}_j \boldsymbol{e}_i^\top \boldsymbol{X}) + \operatorname{tr}(\boldsymbol{e}_i \boldsymbol{e}_j^\top \boldsymbol{X})\Big) \\ &= \operatorname{tr}\left(\frac{1}{2}(\boldsymbol{e}_j \boldsymbol{e}_i^\top + \boldsymbol{e}_i \boldsymbol{e}_j^\top)\boldsymbol{X}\right) = \operatorname{tr}(\boldsymbol{E}_{i,j}\boldsymbol{X}). \end{aligned}$$

□

3.14 Unconstrained Optimization: Optimality conditions and gradient descent

Aim: Given $f \in C^1(\mathbb{R}^n)$, solve

$$\underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{Minimize}}\, f(\boldsymbol{x})$$

3.14.1 ★Minimizer and stationary point

Minimizer and stationary point:

- We say that $\boldsymbol{x}^*$ is a global minimizer of f if

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*) \quad \forall\, \boldsymbol{x} \in \mathbb{R}^n.$$

- We say that $\boldsymbol{x}^*$ is a local minimizer of f if $\exists\ \varepsilon > 0$ so that

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*) \quad \forall\, \boldsymbol{x} \in \mathbb{R}^n \text{ with } \|\boldsymbol{x} - \boldsymbol{x}^*\|_2 < \varepsilon.$$

- We say that $\boldsymbol{x}^*$ is a stationary point of f if $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$.

Theorem (first-order Characterization): Let $f \in C^1(\mathbb{R}^n)$ and suppose that $\boldsymbol{x}^*$ is a local minimizer of f. Then $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$.

Remark: Local minimizers are stationary points for C^1 functions.

Theorem (second-order Characterization): Let $f \in C^2(\mathbb{R}^n)$.

(a) If $\boldsymbol{x}^*$ is a local minimizer of f, then $\nabla^2 f(\boldsymbol{x}^*) \succeq 0$.

(b) If $\boldsymbol{x}^*$ is a stationary point of f and $\nabla^2 f(\boldsymbol{x}^*) \succ 0$, then $\boldsymbol{x}^*$ is a local minimizer.

Remark:

- local minimizer $\implies$ $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$ and $\nabla^2 f(\boldsymbol{x}^*) \succeq 0$
- $\nabla f(\boldsymbol{x}^*) = \boldsymbol{0}$ and $\nabla^2 f(\boldsymbol{x}^*) \succ 0$ $\implies$ local minimizer

3.14.2 Steepest (gradient) descent with exact line search

Steepest descent with exact line search:
Start at $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$. For each $k = 0, 1, 2, \cdots$:

* Set $\boldsymbol{d}^{(k)} = -\nabla f(\boldsymbol{x}^{(k)})$.
* Pick α_k so that:

$$\alpha_k \in \arg\min \left\{ f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)}) : \alpha \geq 0 \right\}. \tag{2}$$

* Set $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{d}^{(k)}$.

Note:

- $\boldsymbol{d}^{(k)}$ is called search direction. In the above algorithm, $\boldsymbol{d}^{(k)} = -\nabla f(\boldsymbol{x}^{(k)})$.
- α_k is called step size. In the above algorithm, it is chosen according to the exact line search criterion (Equation (2)).
- In steepest descent with exact line search, it is implicitly assumed that a minimizer α_k exists for the exact line search subproblem (Equation (2)).

- If α_k exists and $\nabla f(\boldsymbol{x}^{(k)}) \neq 0$, then we have:

$$0 = \frac{d}{d\alpha} f(\boldsymbol{x}^{(k)} + \alpha \boldsymbol{d}^{(k)})\Big|_{\alpha=\alpha_k} = (\boldsymbol{d}^{(k)})^\mathsf{T} \nabla f(\boldsymbol{x}^{(k+1)}) = -[\nabla f(\boldsymbol{x}^{(k)})]^\mathsf{T} \nabla f(\boldsymbol{x}^{(k+1)}).$$

  New direction $\perp$ old direction: Creating zigzag path!

- Exact line search can be hard to perform.

3.14.3 Armijo Rule and Descent Direction

In contrast to exact line search, usually inexact line search strategy is performed. One commonly used rule is:

> **Armijo rule:** Let $\sigma \in (0,1)$, $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{d} \in \mathbb{R}^n$. Find $\alpha > 0$ so that
>
> $$f(\boldsymbol{x} + \alpha \boldsymbol{d}) \leq f(\boldsymbol{x}) + \alpha \sigma \nabla f(\boldsymbol{x})^\mathsf{T} \boldsymbol{d}.$$

> **Descent direction:** A vector $\boldsymbol{d}$ is a descent direction for a function $f(\boldsymbol{x})$ at $\boldsymbol{x}$ if there exists $\widetilde{\alpha} > 0$ such that
>
> $$f(\boldsymbol{x} + \alpha \boldsymbol{d}) < f(\boldsymbol{x}) \quad \text{for all } \alpha \in (0, \widetilde{\alpha}).$$

For a function $f \in C^1(\mathbb{R}^n)$, if the condition

$$\nabla f(\boldsymbol{x})^\mathsf{T} \boldsymbol{d} < 0 \quad \Longrightarrow \quad f(\boldsymbol{x} + \alpha \boldsymbol{d}) < f(\boldsymbol{x}) \quad \text{for small } \alpha > 0.$$

> **Examples:** At an $\boldsymbol{x}$ that is not stationary,
>
> - $\boldsymbol{d} = -\nabla f(\boldsymbol{x})$ is a descent direction,
>
> $$\nabla f(\boldsymbol{x})^\mathsf{T} \boldsymbol{d} = \nabla f(\boldsymbol{x})^\mathsf{T} [-\nabla f(\boldsymbol{x})] = -\|\nabla f(\boldsymbol{x})\|_2^2 < 0.$$
>
> - More generally, if $\boldsymbol{D} \succ 0$, then $\boldsymbol{d} = -\boldsymbol{D} \nabla f(\boldsymbol{x})$ is a descent direction,
>
> $$\nabla f(\boldsymbol{x})^\mathsf{T} \boldsymbol{d} = \nabla f(\boldsymbol{x})^\mathsf{T} [-\boldsymbol{D} \nabla f(\boldsymbol{x})] = -\nabla f(\boldsymbol{x})^\mathsf{T} \boldsymbol{D} \nabla f(\boldsymbol{x}) < 0.$$

> **Theorem (Armijo rule):** Let $f \in C^1(\mathbb{R}^n)$, $\boldsymbol{x} \in \mathbb{R}^n$, and $\boldsymbol{d} \in \mathbb{R}^n$ be a descent direction at $\boldsymbol{x}$. Let $\sigma \in (0,1)$. Then there exists $\alpha_1 > 0$ so that for all $\alpha \in [0, \alpha_1]$,
>
> $$f(\boldsymbol{x} + \alpha \boldsymbol{d}) \leq f(\boldsymbol{x}) + \alpha \sigma \nabla f(\boldsymbol{x})^\mathsf{T} \boldsymbol{d}.$$

3.14.4 Armijo line search by backtracking

> **Armijo line search by backtracking:** Fix $\sigma \in (0,1)$ and $\beta \in (0,1)$. Given $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{d} \in \mathbb{R}^n$ and $\bar{\alpha} > 0$.
> Find the smallest nonnegative integer $j = j_0$ so that
>
> $$f(\boldsymbol{x} + \bar{\alpha}\beta^j \boldsymbol{d}) \leq f(\boldsymbol{x}) + \bar{\alpha}\beta^j \sigma \nabla f(\boldsymbol{x})^\mathsf{T} \boldsymbol{d}. \tag{3}$$
>
> The stepsize generated is then $\bar{\alpha}\beta^{j_0}$.

- If $\boldsymbol{d}$ is a descent direction, then (3) is satisfied for all sufficiently large j.
- In practice, (3) is tested for $j = 0, 1, 2, \cdots$ successively. This is called backtracking because the stepsize $\bar{\alpha}\beta^j$ decreases with each step.

- The choice of $\bar{\alpha}$ is crucial for the efficiency of such a scheme.

Theorem (Convergence under Armijo Rule): Let $f \in C^1(\mathbb{R}^n)$ with $\inf f > -\infty$. Let $\{\bar{\alpha}_k\} \subseteq \mathbb{R}$ satisfy $0 < \inf_k \bar{\alpha}_k \le \sup_k \bar{\alpha}_k < \infty$, and fix $\sigma \in (0,1)$ and $\beta \in (0,1)$. Suppose $\{\boldsymbol{x}^{(k)}\}$ is generated as

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{d}^{(k)},$$

where

- $\boldsymbol{d}^{(k)} := -\boldsymbol{D}_k \nabla f(\boldsymbol{x}^{(k)})$; here $\{\boldsymbol{D}_k\}$ is a bounded sequence of positive definite matrices with $\boldsymbol{D}_k - \delta \boldsymbol{I} \succeq 0$ for some $\delta > 0$;
- α_k is generated via the Armijo line search by backtracking with $\boldsymbol{x} = \boldsymbol{x}^{(k)}$, $\boldsymbol{d} = \boldsymbol{d}^{(k)}$ and $\bar{\alpha} = \bar{\alpha}_k$, and σ and β defined above.

Then any accumulation point of $\{\boldsymbol{x}^{(k)}\}$ is a stationary point of f.

Remarks:

- $\boldsymbol{x}^*$ is an accumulation point of $\{\boldsymbol{x}^{(k)}\}$ if there exists a sub-sequence of $\{\boldsymbol{x}^{(k)}\}$ that converges to $\boldsymbol{x}^*$.
- If $\boldsymbol{x}^{(k)}$ is non-stationary, then $\boldsymbol{d}^{(k)}$ is a descent direction.
- The condition $\boldsymbol{D}_k - \delta \boldsymbol{I} \succeq 0$ implies that for any $\boldsymbol{y} \in \mathbb{R}^n$, we have

$$\boldsymbol{y}^\mathsf{T}(\boldsymbol{D}_k - \delta \boldsymbol{I})\boldsymbol{y} \ge 0.$$

  Hence $\boldsymbol{y}^\mathsf{T}\boldsymbol{D}_k\boldsymbol{y} \ge \delta \|\boldsymbol{y}\|^2$.
- σ is chosen to be small so that (3) may be satisfied with a small number of backtracking steps: note that each backtracking requires an evaluation of $f(\boldsymbol{x} + \alpha \boldsymbol{d})$, which adds to the main computational cost. A typical choice is $\sigma = 10^{-4}$.
- β is typically $\frac{1}{2}$.
- The choice of $\{\bar{\alpha}_k\}$ is crucial. Ideally, it should be chosen so that (3) may be satisfied with a small number of backtracking steps. Possible choices are:
  - $\bar{\alpha}_k \equiv 1$ for "Newton-like" directions.
  - $\bar{\alpha}_k = \max\{u, \min\{\ell, \alpha_{k-1}\}\}$, where u and ℓ are positive.
  - (projected, adaptive) Barzilai-Borwein stepsize.
- One can terminate when $\|\nabla f(\boldsymbol{x}^{(k)})\|_2 \le \text{tol} \cdot \max\{|f(\boldsymbol{x}^{(k)})|, 1\}$, i.e., when the gradient is small relative to the function value.

3.14.5 ★Steepest descent with constant stepsize

Theorem (Steepest descent with constant stepsize): Let $f \in C^2(\mathbb{R}^n)$ with $\inf f > -\infty$. Suppose that there exists $L > 0$ so that

$$L \ge \|\nabla^2 f(\boldsymbol{x})\|_2 \quad \text{for all } \boldsymbol{x}.$$

Fix any $\gamma \in (0,2)$ and consider the sequence generated as

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \frac{\gamma}{L} \nabla f(\boldsymbol{x}^{(k)}).$$

Then any accumulation point of $\{\boldsymbol{x}^{(k)}\}$ is a stationary point of f.

- By letting $\gamma = \alpha L$, any stepsize $\alpha \in \left(0, \frac{2}{L}\right)$ will work.
- Given L, the above algorithm can be written in one line.

- The algorithm avoids costly line search, but it may be slow as the constant stepsize can be too conservative, hindering progress.

A Chain Rule: Let $h \in C^2(\mathbb{R}^m)$ and let $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$.

Define $f(\boldsymbol{x}) := h(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})$. Then $f \in C^2(\mathbb{R}^n)$ and

$$\nabla f(\boldsymbol{x}) = \boldsymbol{A}^\top \nabla h(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}), \quad \nabla^2 f(\boldsymbol{x}) = \boldsymbol{A}^\top \nabla^2 h(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})\boldsymbol{A}.$$

In particular, if there exists L such that $L \geq \|\nabla^2 h(\boldsymbol{y})\|_2$ for all $\boldsymbol{y}$, then

$$\begin{aligned}\|\nabla^2 f(\boldsymbol{x})\|_2 &\leq \|\boldsymbol{A}^\top\|_2 \|\nabla^2 h(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})\|_2 \|\boldsymbol{A}\|_2 \\ &\leq L\|\boldsymbol{A}^\top\|_2\|\boldsymbol{A}\|_2 = L\lambda_{\max}(\boldsymbol{A}^\top\boldsymbol{A}).\end{aligned}$$

Lemma: For any $\boldsymbol{A} \in \mathbb{R}^{m \times n}$,

$$\|\boldsymbol{A}\|_2 = \sqrt{\lambda_{\max}(\boldsymbol{A}^\top\boldsymbol{A})} = \sqrt{\lambda_{\max}(\boldsymbol{A}\boldsymbol{A}^\top)} = \|\boldsymbol{A}^\top\|_2.$$

Proof sketch: The argument consists of two steps:

(a) Apply the spectral theorem to establish that

$$\|\boldsymbol{A}\|_2 = \sqrt{\lambda_{\max}(\boldsymbol{A}^\top\boldsymbol{A})}.$$

(b) Show that the nonzero eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^\top$ are identical.

3.15 ★Unconstrained Optimization: Quasi-Newton methods

Aim: Although the Newton direction $-[\nabla^2 f(\boldsymbol{x})]^{-1}\nabla f(\boldsymbol{x})$ is often a good descent direction, it is costly to compute; thus, we use a more efficient approximation.

Idea: Let $f \in C^2(\mathbb{R}^n)$. Given $\boldsymbol{x}^{(k+1)}$ and $\boldsymbol{x}^{(k)}$, we would expect

$$\nabla^2 f(\boldsymbol{x}^{(k+1)})(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) \approx \nabla f(\boldsymbol{x}^{(k+1)}) - \nabla f(\boldsymbol{x}^{(k)}).$$

Notation:

$$\boldsymbol{s}^{(k)} := \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}, \quad \boldsymbol{y}^{(k)} := \nabla f(\boldsymbol{x}^{(k+1)}) - \nabla f(\boldsymbol{x}^{(k)}).$$

This motivates us to successively construct $\boldsymbol{B}_{k+1}$ (resp., $\boldsymbol{H}_{k+1}$) to approximate $\nabla^2 f(\boldsymbol{x}^{(k+1)})$ (resp., $[\nabla^2 f(\boldsymbol{x}^{(k+1)})]^{-1}$) so that

$$\boldsymbol{B}_{k+1}\boldsymbol{s}^{(k)} = \boldsymbol{y}^{(k)} \quad (\text{resp., } \boldsymbol{H}_{k+1}\boldsymbol{y}^{(k)} = \boldsymbol{s}^{(k)}).$$

We refer to these equations as **secant equations**.

Popular Update Formulas:
Initialize $\boldsymbol{B}_0$ (or $\boldsymbol{H}_0$) at a **positive definite matrix**.

- Symmetric Rank 1 (SR1) method:

$$\boldsymbol{B}_{k+1} = \boldsymbol{B}_k + \frac{(\boldsymbol{y}^{(k)} - \boldsymbol{B}_k\boldsymbol{s}^{(k)})(\boldsymbol{y}^{(k)} - \boldsymbol{B}_k\boldsymbol{s}^{(k)})^\top}{(\boldsymbol{y}^{(k)} - \boldsymbol{B}_k\boldsymbol{s}^{(k)})^\top\boldsymbol{s}^{(k)}}$$

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k + \frac{(\boldsymbol{s}^{(k)} - \boldsymbol{H}_k \boldsymbol{y}^{(k)})(\boldsymbol{s}^{(k)} - \boldsymbol{H}_k \boldsymbol{y}^{(k)})^\top}{(\boldsymbol{s}^{(k)} - \boldsymbol{H}_k \boldsymbol{y}^{(k)})^\top \boldsymbol{y}^{(k)}}$$

- Davidon–Fletcher–Powell (DFP) formula:

$$\boldsymbol{B}_{k+1} = \left(\boldsymbol{I} - \frac{\boldsymbol{y}^{(k)} \boldsymbol{s}^{(k)\top}}{\boldsymbol{y}^{(k)\top} \boldsymbol{s}^{(k)}}\right) \boldsymbol{B}_k \left(\boldsymbol{I} - \frac{\boldsymbol{s}^{(k)} \boldsymbol{y}^{(k)\top}}{\boldsymbol{y}^{(k)\top} \boldsymbol{s}^{(k)}}\right) + \frac{\boldsymbol{y}^{(k)} \boldsymbol{y}^{(k)\top}}{\boldsymbol{y}^{(k)\top} \boldsymbol{s}^{(k)}}$$

$$\boldsymbol{H}_{k+1} = \boldsymbol{H}_k + \frac{\boldsymbol{s}^{(k)} \boldsymbol{s}^{(k)\top}}{\boldsymbol{y}^{(k)\top} \boldsymbol{s}^{(k)}} - \frac{\boldsymbol{H}_k \boldsymbol{y}^{(k)} \boldsymbol{y}^{(k)\top} \boldsymbol{H}_k}{\boldsymbol{y}^{(k)\top} \boldsymbol{H}_k \boldsymbol{y}^{(k)}}$$

- Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm:

$$\boldsymbol{B}_{k+1} = \boldsymbol{B}_k + \frac{\boldsymbol{y}^{(k)} \boldsymbol{y}^{(k)\top}}{\boldsymbol{y}^{(k)\top} \boldsymbol{s}^{(k)}} - \frac{\boldsymbol{B}_k \boldsymbol{s}^{(k)} \boldsymbol{s}^{(k)\top} \boldsymbol{B}_k}{\boldsymbol{s}^{(k)\top} \boldsymbol{B}_k \boldsymbol{s}^{(k)}}$$

$$\boldsymbol{H}_{k+1} = \left(\boldsymbol{I} - \frac{\boldsymbol{s}^{(k)} \boldsymbol{y}^{(k)\top}}{\boldsymbol{y}^{(k)\top} \boldsymbol{s}^{(k)}}\right) \boldsymbol{H}_k \left(\boldsymbol{I} - \frac{\boldsymbol{y}^{(k)} \boldsymbol{s}^{(k)\top}}{\boldsymbol{y}^{(k)\top} \boldsymbol{s}^{(k)}}\right) + \frac{\boldsymbol{s}^{(k)} \boldsymbol{s}^{(k)\top}}{\boldsymbol{y}^{(k)\top} \boldsymbol{s}^{(k)}}$$

Remark:

- DFP and BFGS are **rank-2** updates, while SR1 is a **rank-1** update.
- Since $\boldsymbol{B}_0$ and $\boldsymbol{H}_0$ were symmetric to start with, by induction, all $\boldsymbol{B}_k$ and $\boldsymbol{H}_k$ are **symmetric**.
- In practice, **BFGS** usually performs better.

Quasi-Newton Method based on $\boldsymbol{B}_k$ (or $\boldsymbol{H}_k$) for $f \in C^1(\mathbb{R}^n)$:

- Initialize at $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ and $\boldsymbol{B}_0 \succ 0$ (or $\boldsymbol{H}_0 \succ 0$).
- For $k = 0, 1, 2, \cdots$
  (a) Find $\boldsymbol{d}^{(k)}$ via $\boldsymbol{B}_k \boldsymbol{d}^{(k)} = -\nabla f(\boldsymbol{x}^{(k)})$ or $\boldsymbol{d}^{(k)} = -\boldsymbol{H}_k \nabla f(\boldsymbol{x}^{(k)})$.
  (b) Update $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \boldsymbol{d}^{(k)}$. Or, more generally, $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \alpha_k \boldsymbol{d}^{(k)}$ for some $\alpha_k > 0$.
  (c) Set $\boldsymbol{y}^{(k)} = \nabla f(\boldsymbol{x}^{(k+1)}) - \nabla f(\boldsymbol{x}^{(k)})$ and $\boldsymbol{s}^{(k)} = \boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}$. Compute $\boldsymbol{B}_{k+1}$ or $\boldsymbol{H}_{k+1}$.

3.16 Constrained Optimization: KKT conditions

Problem settings (Constrained Optimization):

$$\begin{aligned} \underset{\boldsymbol{x} \in \mathbb{R}^n}{\text{Minimize}} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & h_j(\boldsymbol{x}) = 0, \quad j \in \mathcal{J} = \{1, \cdots, p\}, \\ & g_i(\boldsymbol{x}) \le 0, \quad i \in \mathcal{I} = \{1, \cdots, m\}. \end{aligned} \tag{4}$$

Here:

- f, h_j and g_i are all C^1 functions.
- **Aim:** Find conditions to help characterize local minimizers!

Definition: We say that $\boldsymbol{x}^*$ is a local minimizer of (4) if $\boldsymbol{x}^*$ is feasible and $\exists \varepsilon > 0$ so that $f(\boldsymbol{x}) \ge f(\boldsymbol{x}^*)$ whenever $\boldsymbol{x}$ is feasible and $\|\boldsymbol{x} - \boldsymbol{x}^*\|_2 < \varepsilon$.

3.16.1 ★KKT conditions

KKT conditions: Consider (4). A point $\boldsymbol{x} \in \mathbb{R}^n$ is said to satisfy the KKT conditions if there exist $\boldsymbol{\lambda} \in \mathbb{R}^m$, $\boldsymbol{\mu} \in \mathbb{R}^p$ such that

- ⋆ (Primal feasibility) $g_i(\boldsymbol{x}) \leq 0$ for all $i \in \mathcal{I}$ and $h_j(\boldsymbol{x}) = 0$ for all $j \in \mathcal{J}$; and
- ⋆ (Stationarity) $\nabla f(\boldsymbol{x}) + \sum_{i \in \mathcal{I}} \lambda_i \nabla g_i(\boldsymbol{x}) + \sum_{j \in \mathcal{J}} \mu_j \nabla h_j(\boldsymbol{x}) = 0$; and
- ⋆ (Dual feasibility) $\lambda_i \geq 0$ for all $i \in \mathcal{I}$.
- ⋆ (Complementary slackness) $\lambda_i g_i(\boldsymbol{x}) = 0$ for all $i \in \mathcal{I}$.

The $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are called Lagrange multipliers at $\boldsymbol{x}$.

Definition (KKT point): A point $\boldsymbol{x}$ is called a KKT point (also known as a stationary point) of problem (4) if there exist multipliers $\boldsymbol{\lambda} \in \mathbb{R}^m$ and $\boldsymbol{\mu} \in \mathbb{R}^p$ such that the KKT conditions are satisfied.

3.16.2 ★MFCQ

Definition (MFCQ): Consider the feasible set of (4) and let $\boldsymbol{x}^*$ be feasible. We say that the Mangasarian-Fromovitz constraint qualification (MFCQ) holds at $\boldsymbol{x}^*$ if the following conditions hold:

$$\begin{cases} \sum\limits_{j \in \mathcal{J}} \mu_j \nabla h_j(\boldsymbol{x}^*) + \sum\limits_{i \in \mathcal{I}(\boldsymbol{x}^*)} \lambda_i \nabla g_i(\boldsymbol{x}^*) = 0 \\ \lambda_i \geq 0 \quad \forall\, i \in \mathcal{I}(\boldsymbol{x}^*) \end{cases} \implies \begin{cases} \mu_j = 0 \quad \forall\, j \in \mathcal{J}, \\ \lambda_i = 0 \quad \forall\, i \in \mathcal{I}(\boldsymbol{x}^*). \end{cases}$$

Remarks:

- The MFCQ also admits an equivalent formulation; see, for example, this link.
- If $g_i(\boldsymbol{x}^*) < 0$ for all $i \in \mathcal{I}$ so that $\mathcal{I}(\boldsymbol{x}^*) = \varnothing$, then MFCQ means $\{\nabla h_j(\boldsymbol{x}^*) : j \in \mathcal{J}\}$ is linearly independent.
- If $\mathcal{J} = \varnothing$, then MFCQ means $\{\nabla g_i(\boldsymbol{x}^*) : i \in \mathcal{I}(\boldsymbol{x}^*)\}$ is positively independent.

Theorem: Consider (4) and let $\boldsymbol{x}^*$ be a local minimizer. Suppose that MFCQ holds at $\boldsymbol{x}^*$. Then there exist $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ and $\boldsymbol{\mu}^* \in \mathbb{R}^p$ so that

- $\nabla f(\boldsymbol{x}^*) + \sum_{i \in \mathcal{I}} \lambda_i^* \nabla g_i(\boldsymbol{x}^*) + \sum_{j \in \mathcal{J}} \mu_j^* \nabla h_j(\boldsymbol{x}^*) = 0$; and
- $\lambda_i^* \geq 0$ and $\lambda_i^* g_i(\boldsymbol{x}^*) = 0$ for all $i \in \mathcal{I}$.

Remarks:

- Under MFCQ, the approximating LP is "good" around $\boldsymbol{x}^*$. We hence look at the KKT of this LP.
- If $\boldsymbol{x}^*$ is a local minimizer and if the MFCQ holds at $\boldsymbol{x}^*$, then there exist $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ and $\boldsymbol{\mu}^* \in \mathbb{R}^p$ so that $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ satisfies the KKT conditions. That is,

$$\text{local minimizer and MFCQ} \implies \text{KKT conditions (KKT point)}.$$

- The first bullet point follows from the *dual feasibility* condition of the approximating LP, and by defining $\lambda_i^* = 0$ for $i \notin \mathcal{I}(\boldsymbol{x}^*)$.
- The second bullet point follows from the definition of $\mathcal{I}(\boldsymbol{x}^*)$ and by defining $\lambda_i^* = 0$ for $i \notin \mathcal{I}(\boldsymbol{x}^*)$.

3.16.3 ★Slater's condition

Theorem (MFCQ from Slater)**:** Consider the set defined by

$$S := \{\boldsymbol{x} \in \mathbb{R}^n : g_i(\boldsymbol{x}) \leq 0 \quad \forall\, i \in \mathcal{I}\},$$

where g_i are convex C^1. Suppose that there exists $\bar{\boldsymbol{x}}$ satisfying

$$g_i(\bar{\boldsymbol{x}}) < 0 \quad \forall\, i \in \mathcal{I}.$$

Then MFCQ holds at every point in S.

Remarks:

- The set S in the above theorem is closed and convex.
- The condition that "there exists $\bar{\boldsymbol{x}}$ satisfying $g_i(\bar{\boldsymbol{x}}) < 0$ for all $i \in \mathcal{I}$" is called the Slater's condition. The $\bar{\boldsymbol{x}}$ is called a Slater point.
- One can indeed show that for the above S, the MFCQ holds at every point in S if and only if Slater's condition holds.

Theorem (MFCQ from generalized Slater)**:** Consider the set defined by

$$S := \{\boldsymbol{x} \in \mathbb{R}^n : g_i(\boldsymbol{x}) \leq 0 \quad \forall\, i \in \mathcal{I},\ \ \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}\},$$

where g_i are convex C^1 and $\boldsymbol{A} \in \mathbb{R}^{p\times n}$. Suppose that there exists $\bar{\boldsymbol{x}}$ satisfying

$$\boldsymbol{A}\bar{\boldsymbol{x}} = \boldsymbol{b}, \quad g_i(\bar{\boldsymbol{x}}) < 0 \quad \forall\, i \in \mathcal{I},$$

and $\boldsymbol{A}$ has full row rank. Then MFCQ holds at every point in S.

Remarks:

- The set S in the above theorem is **closed and convex**.
- The condition that "there exists $\bar{\boldsymbol{x}}$ satisfying $g_i(\bar{\boldsymbol{x}}) < 0$ for all $i \in \mathcal{I}$, $\boldsymbol{A}\bar{\boldsymbol{x}} = \boldsymbol{b}$ and $\boldsymbol{A}$ has full row rank" is called the generalized Slater's condition. The $\bar{\boldsymbol{x}}$ is called a generalized Slater point.
- One can indeed show that for the above S, MFCQ holds at every point in S if and only if generalized Slater's condition holds.

3.16.4 Role of convexity

Consider the following special instance of (4)

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{Minimize}} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \\ & g_i(\boldsymbol{x}) \leq 0, \quad i \in \mathcal{I}, \end{aligned} \tag{5}$$

here f and g_i are all convex C^1 functions, $\boldsymbol{A} \in \mathbb{R}^{p\times n}$.

Theorem (Sufficiency under convexity)**:** Consider (5). Suppose that there exist $\boldsymbol{x}^* \in \mathbb{R}^n$, $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ and $\boldsymbol{\mu}^* \in \mathbb{R}^p$ so that

- $\boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{b}$ and $g_i(\boldsymbol{x}^*) \leq 0$ for all $i \in \mathcal{I}$; and
- $\nabla f(\boldsymbol{x}^*) + \sum_{i\in\mathcal{I}} \lambda_i^* \nabla g_i(\boldsymbol{x}^*) + \boldsymbol{A}^\mathsf{T}\boldsymbol{\mu}^* = 0$; and
- $\lambda_i^* \geq 0$ and $\lambda_i^* g_i(\boldsymbol{x}^*) = 0$ for all $i \in \mathcal{I}$.

Then $\boldsymbol{x}^*$ is a global minimizer of (5).

3.17 Constrained Optimization (Penalty/Barrier methods)

Problem settings:

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{Minimize}} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & g_i(\boldsymbol{x}) \le 0, \quad i \in \mathcal{I} = \{1, \cdots, m\}. \end{aligned} \tag{6}$$

Here:

- f and g_i are all C^1 functions.
- Assume that the feasible set $\{\boldsymbol{x} : g_i(\boldsymbol{x}) \le 0 \quad \forall i \in \mathcal{I}\}$ is not empty.
- Two main classes:
  - ∗ Penalty method (*exterior type*);
  - ∗ Barrier method (*interior type*).
- Can be generalized to include equality constraints.

3.17.1 ★Penalty method

- Add a penalty if we get outside of the feasible set.
- If the penalty is sufficiently large, a minimizer is forced to be inside the feasible set.
- Initial idea: Define

$$P(\boldsymbol{x}) := \begin{cases} 0 & \text{if } g_i(\boldsymbol{x}) \le 0 \quad \forall i \in \mathcal{I}, \\ \infty & \text{otherwise.} \end{cases}$$

Then we can consider the unconstrained problem:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) + P(\boldsymbol{x}).$$

Except that the objective is highly discontinuous!

Definition (Penalty functions): A function $P : \mathbb{R}^n \to \mathbb{R}$ is a penalty function for the constraint set $\{\boldsymbol{x} : g_i(\boldsymbol{x}) \le 0 \, \forall i \in \mathcal{I}\}$ if

- $P(\boldsymbol{x}) \ge 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$; and
- $P(\boldsymbol{x}) = 0$ if and only if $g_i(\boldsymbol{x}) \le 0$ for all $i \in \mathcal{I}$.

Examples:

- $P(\boldsymbol{x}) = \sum_{i=1}^m \max\{g_i(\boldsymbol{x}), 0\}$.
- $P(\boldsymbol{x}) = \frac{1}{2} \sum_{i=1}^m \max\{g_i(\boldsymbol{x}), 0\}^2$. Courant-Beltrami penalty function: C^1 function

Idea: Solve $\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) + cP(\boldsymbol{x})$ for some very large $c > 0$. But...

- How large should c be?
- Can minimizers be found? (Existence? Just stationary points?)

Penalty method (basic version): Penalty method for (6): basic version Let $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$, $c > 0$ and $\eta > 1$. Set $c_1 = c$. For $k = 1, \cdots,$

- Find a global minimizer $\boldsymbol{x}^{(k)}$ of

$$q_{c_k}(\boldsymbol{x}) := f(\boldsymbol{x}) + \frac{c_k}{2} \sum_{i=1}^m \left(\max\{g_i(\boldsymbol{x}), 0\}\right)^2,$$

using $\boldsymbol{x}^{(k-1)}$ as the initial point for the iterative method.

- Update $c_{k+1} = \eta c_k$.

Remarks:

- As c increases, q_c becomes more *ill-conditioned*. The choice of $\boldsymbol{x}^{(k-1)}$ as a starting point for the iterative method helps alleviate the ill-conditioning.
- The above algorithm is only conceptual because finding global minimizers can be challenging if q_{c_k} is not convex. Global minimizers also may not exist! In general, only stationary points of q_{c_k} can be expected.

Theorem (Convergence of penalty method): Consider (6) and suppose that $\inf f > -\infty$. Let $\{\boldsymbol{x}^{(k)}\}$ be generated by the penalty method (basic version). Then any accumulation point $\boldsymbol{x}^*$ of $\{\boldsymbol{x}^{(k)}\}$ is a globally optimal solution of (6).

Proof sketch: Assume that $\{\boldsymbol{x}^{(k_i)}\}$ is a convergent subsequence with $\lim_{i\to\infty} \boldsymbol{x}^{(k_i)} = \boldsymbol{x}^*$.

- **Feasibility:** Fix any feasible $\boldsymbol{x}$. Then for each k_i, we have

$$\inf f + c_{k_i} P(\boldsymbol{x}^{(k_i)}) \le q_{c_{k_i}}(\boldsymbol{x}^{(k_i)}) \le q_{c_{k_i}}(\boldsymbol{x}) = f(\boldsymbol{x}),$$

  where the second inequality comes from the fact $\boldsymbol{x}^{(k_i)}$ is a global minimizer. Here P is the Courant-Beltrami penalty function.
  Then

$$P(\boldsymbol{x}^{(k_i)}) \le \frac{f(\boldsymbol{x}) - \inf f}{c_{k_i}}.$$

  Hence, $P(\boldsymbol{x}^*) = \lim_{i\to\infty} P(\boldsymbol{x}^{(k_i)}) = 0$, showing that $\boldsymbol{x}^*$ is feasible.

- **Optimality:** Fix any feasible $\boldsymbol{x}$. Then for each k_i, we have

$$f(\boldsymbol{x}^{(k_i)}) \le f(\boldsymbol{x}^{(k_i)}) + c_{k_i} P(\boldsymbol{x}^{(k_i)}) = q_{c_{k_i}}(\boldsymbol{x}^{(k_i)}) \le q_{c_{k_i}}(\boldsymbol{x}) = f(\boldsymbol{x}),$$

  where the second inequality comes from the fact $\boldsymbol{x}^{(k_i)}$ is a global minimizer. Then,

$$f(\boldsymbol{x}^*) = \lim_{i\to\infty} f(\boldsymbol{x}^{(k_i)}) \le f(\boldsymbol{x}),$$

  Since this is true for any feasible $\boldsymbol{x}$ and $\boldsymbol{x}^*$ is feasible, we conclude that $\boldsymbol{x}^*$ solves (6).

3.17.2 ★Barrier method

Recall that

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{Minimize}} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & g_i(\boldsymbol{x}) \le 0, \quad i \in \mathcal{I}. \end{aligned} \tag{7}$$

Here:

- f and g_i are all C^1 functions.
- For barrier methods, we assume in addition that

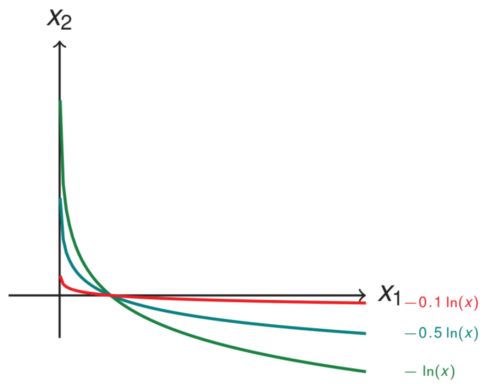
$$\mathcal{S}^0 := \{\boldsymbol{x} : g_i(\boldsymbol{x}) < 0 \quad \forall i \in \mathcal{I}\} \neq \varnothing.$$

- For simplicity, we focus on the case that all f and g_i are convex.
- In contrast to Penalty methods that are exterior methods, Barrier methods are interior methods: every iterate stays within the feasible region.
- One standard way is to make use of the log-barrier function:

$$B(\boldsymbol{x}) := -\sum_{i=1}^{m} \ln[-g_i(\boldsymbol{x})].$$

Then one minimizes

$$f_\mu(\boldsymbol{x}) := f(\boldsymbol{x}) + \mu B(\boldsymbol{x}) \quad \text{for some } \mu > 0.$$



Barrier method for (7) **(basic version)** Let $\boldsymbol{x}^{(0)} \in \mathcal{S}^0$, $\mu > 0$ and $\eta > 1$. Set $\mu_1 = \mu$. For $k = 1, \cdots,$

- Find a minimizer $\boldsymbol{x}^{(k)}$ of

$$f_{\mu_k}(\boldsymbol{x}) := f(\boldsymbol{x}) - \mu_k \sum_{i=1}^{m} \ln[-g_i(\boldsymbol{x})],$$

  using $\boldsymbol{x}^{(k-1)}$ as the initial point for the iterative method.
- Update $\mu_{k+1} = \mu_k/\eta$.

- Notice that μ_k is being **decreased** instead of being increased. Thus, at each $\boldsymbol{x}$, we have

$$\lim_{k\to\infty} \mu_k B(\boldsymbol{x}) = \begin{cases} 0 & \text{if } g_i(\boldsymbol{x}) < 0 \quad \forall\, i \in \mathcal{I}, \\ \infty & \text{otherwise.} \end{cases}$$

- In principle, one can still apply descent methods though the function $B(\boldsymbol{x})$ is not defined on the whole $\mathbb{R}^n$. This is because descent methods make sure f_μ value decreases; in particular, f_μ will remain finite, keeping $\boldsymbol{x}^{(k)}$ feasible.
- Unlike penalty function q_c, the function $-\ln(\cdot)$ is an analytic function on $\mathbb{R}_{++}$. Indeed, when $f, g_i \in C^2(\mathbb{R}^n)$, one typically uses Newton's method (or its variants) to minimize f_μ.
- The above algorithm is only conceptual. In practice, μ has to be decreased judiciously to avoid getting too close to the boundary. Moreover, minimizers of f_μ may not exist.