EECS 559 Optimization Methods for SIPML

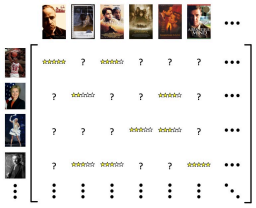
Lecture 3 – Basic Tools

Instructor: Prof. Qing Qu (qingqu@umich.edu)

Lecture Agenda

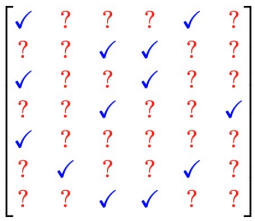
- **Basics of Matrix Analysis**
- Taylor Expansion & Lipschitz Function
- Optimality Conditions
- Rate of Convergence

Example: Matrix Completion



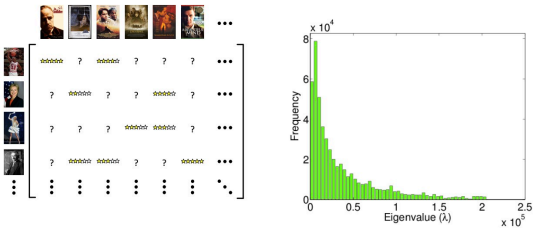
- **Netflix Challenge**: Netflix provides highly incomplete ratings from 0.5 million users for $\approx 17,770$ movies
- How to predict unseen user ratings for movies?

Example: Matrix Completion



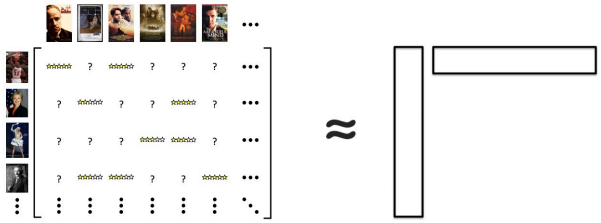
More unknowns than observations (under-determined system)

Example: Matrix Completion



A few factors explain most of the data: **low-rank** approximation

Example: Matrix Completion



How to exploit **low-rank structures** in predictions?

Example: Matrix Completion

Given observed entries $M_{i,j}$ with $(i,j) \in \Omega$, complete the matrix via rank minimization

$$\min_{\boldsymbol{X}} \text{rank}(\boldsymbol{X}), \quad \text{s.t.} \quad X_{i,j} = M_{i,j}, \quad (i,j) \in \Omega,$$

or, equivalently, we can rewrite it as

$$\min_{\boldsymbol{X}} \text{rank}(\boldsymbol{X}), \quad \text{s.t.} \quad \mathcal{P}_\Omega(\boldsymbol{X}) = \mathcal{P}_\Omega(\boldsymbol{M}).$$

Example: Matrix Completion

$$\min_{\boldsymbol{X}} \text{rank}(\boldsymbol{X}), \quad \text{s.t.} \quad \mathcal{P}_\Omega(\boldsymbol{X}) = \mathcal{P}_\Omega(\boldsymbol{M}),$$

where $\mathcal{P}_\Omega(\cdot)$ is the orthogonal projection onto the subspace of matrices support on Ω.

$$\mathcal{P}_\Omega(\boldsymbol{X}) := \begin{cases} X_{ij}, & \text{if } (i,j) \in \Omega, \\ 0, & \text{if } (i,j) \notin \Omega. \end{cases}$$

Example: Matrix Completion

$$\min_{\boldsymbol{X}} \ \mathrm{rank}(\boldsymbol{X}), \quad \text{s.t.} \quad \mathcal{P}_\Omega(\boldsymbol{X}) \ = \ \mathcal{P}_\Omega(\boldsymbol{M}),$$

- Similar to ℓ_0-norm minimization, the rank minimization is also **NP-hard**
- Remedy: convex relaxation via *nuclear norm*:

$$\mathrm{rank}(\boldsymbol{X}) \ = \ \|\boldsymbol{\sigma}(\boldsymbol{X})\|_0\,, \quad \|\boldsymbol{X}\|_* \ = \ \|\boldsymbol{\sigma}(\boldsymbol{X})\|_1\,,$$

where $\|\boldsymbol{X}\|_* = \sum_{i=1}^n \sigma_i(\boldsymbol{X})$ is a *convex surrogate* of $\mathrm{rank}(\boldsymbol{X})$

Matrix Analysis Basics

- **Matrix (Frobenius) inner product:**

$$\langle \boldsymbol{x}, \boldsymbol{z} \rangle \ = \ \sum_{i=1}^n x_i z_i \quad \Longrightarrow \quad \langle \boldsymbol{X}, \boldsymbol{Z} \rangle \ := \ \sum_{i=1}^m \sum_{j=1}^n X_{ij} Z_{ij}$$

- **Matrix trace:** for $\boldsymbol{M} \in \mathbb{R}^{n\times n}$, $\mathrm{tr}(\boldsymbol{M}) := \sum_{i=1}^n M_{ii}$

$$\langle \boldsymbol{X}, \boldsymbol{Z} \rangle \ = \ \mathrm{tr}\left(\boldsymbol{X}^\top \boldsymbol{Z}\right) \ = \ \mathrm{tr}\left(\boldsymbol{X}\boldsymbol{Z}^\top\right)$$

Matrix Analysis Basics

- **Matrix (Frobenius) inner product:**

$$\langle \boldsymbol{x}, \boldsymbol{z} \rangle \ = \ \sum_{i=1}^n x_i z_i \quad \Longrightarrow \quad \langle \boldsymbol{X}, \boldsymbol{Z} \rangle \ := \ \sum_{i=1}^m \sum_{j=1}^n X_{ij} Z_{ij}$$

- **Matrix trace:** for $\boldsymbol{M} \in \mathbb{R}^{n\times n}$, $\mathrm{tr}(\boldsymbol{M}) := \sum_{i=1}^n M_{ii}$

$$\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA})$$

$$\mathrm{tr}(\boldsymbol{A}_1\boldsymbol{A}_2\cdots\boldsymbol{A}_n) = \mathrm{tr}(\boldsymbol{A}_{\pi(1)}\boldsymbol{A}_{\pi(2)}\cdots\boldsymbol{A}_{\pi(n)})$$

Where π is a cyclic permutation on $\{1, 2, \cdots, n\}$

Matrix Analysis Basics

- **Matrix (Frobenius) inner product:**

$$\langle \boldsymbol{x}, \boldsymbol{z} \rangle \ = \ \sum_{i=1}^n x_i z_i \quad \Longrightarrow \quad \langle \boldsymbol{X}, \boldsymbol{Z} \rangle \ := \ \sum_{i=1}^m \sum_{j=1}^n X_{ij} Z_{ij}$$

- **Matrix trace:** for $\boldsymbol{M} \in \mathbb{R}^{n\times n}$, $\mathrm{tr}(\boldsymbol{M}) := \sum_{i=1}^n M_{ii}$

$$\langle \boldsymbol{X}, \boldsymbol{Z} \rangle \ = \ \mathrm{tr}\left(\boldsymbol{X}^\top \boldsymbol{Z}\right) \ = \ \mathrm{tr}\left(\boldsymbol{X}\boldsymbol{Z}^\top\right)$$

- **Frobenius norm:**

$$\|\boldsymbol{X}\|_F = \sqrt{\mathrm{tr}(\boldsymbol{X}^*\boldsymbol{X})} = \sqrt{\textstyle\sum_{i=1}^m \sum_{j=1}^n |X_{ij}|^2}$$

Eigenvectors & Eigenvalues

Let $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. We say that $\lambda \in \mathbb{C}$ is an *eigenvalue* of $\boldsymbol{A}$ if there exists some $\boldsymbol{v} \in \mathbb{C}^n \setminus \{\boldsymbol{0}\}$, such that

$$\boldsymbol{A}\cdot\boldsymbol{v} \ = \ \underbrace{\lambda}_{\textbf{eigenvalue}} \cdot \underbrace{\boldsymbol{v}}_{\textbf{eigenvector}}.$$

Fact. $\lambda \in \mathbb{C}$ is an eigenvalue of $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *iff* it is a root of the *characteristic polynomial* (i.e., $\xi(\lambda) = 0$)

$$\xi(\lambda) := \det(\boldsymbol{A} - \lambda\boldsymbol{I}).$$

Note: even for real $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, its eigenvalue λ is *not* necessarily real.

Eigen Decomposition of Symmetric Matrices

Fact. Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be *symmetric*. Then there exist *orthonormal* vectors $\{\boldsymbol{v}_i\}_{i=1}^n \in \mathbb{R}^n$ and *real* scalars $\lambda_1 \geq \cdots \geq \lambda_n$,

$$\boldsymbol{V} \ = \ \begin{bmatrix} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_n \end{bmatrix} \in O(n), \quad \boldsymbol{\Lambda} \ = \ \mathrm{diag}\left(\lambda_1, \lambda_2, \cdots, \lambda_n\right).$$

such that

$$\boldsymbol{A} \ = \ \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top \ = \ \sum_{i=1}^n \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top.$$

Matrix Analysis Basics

- **Positive (semi)definiteness**

Definition. A *symmetric* matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is *positive definite* if $\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} > 0$ for all nonzero $\boldsymbol{x} \in \mathbb{R}^n$.

It is *positive semidefinite* (*p.s.d.*) if $\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$.

$\boldsymbol{A} \in \mathbb{R}^{n\times n}$ positive definite: $\boldsymbol{A} \succ \boldsymbol{0}$

$\boldsymbol{A} \in \mathbb{R}^{n\times n}$ positive semidefinite: $\boldsymbol{A} \succeq \boldsymbol{0}$

Positive (Semi)definiteness

Definition. A *symmetric* matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is *positive definite* if $\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} > 0$ for all nonzero $\boldsymbol{x} \in \mathbb{R}^n$.

It is *positive semidefinite* (*p.s.d.*) if $\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$.

$\boldsymbol{A} \in \mathbb{R}^{n\times n}$ positive definite: $\boldsymbol{A} \succ \boldsymbol{0}$

$\boldsymbol{A} \in \mathbb{R}^{n\times n}$ positive semidefinite: $\boldsymbol{A} \succeq \boldsymbol{0}$

We also write $\boldsymbol{A} \succeq \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B}$ is *p.s.d.*, i.e., $\boldsymbol{A} - \boldsymbol{B} \succeq \boldsymbol{0}$.

Positive (Semi)definiteness

$\boldsymbol{A} \in \mathbb{R}^{n\times n}$ positive definite: $\quad \boldsymbol{A} \succ \boldsymbol{0}$

$\boldsymbol{A} \in \mathbb{R}^{n\times n}$ positive semidefinite: $\quad \boldsymbol{A} \succeq \boldsymbol{0}$

Fact. A symmetric matrix $\boldsymbol{A}$ is positive definite (resp. semidefinite) *iff all* of its eigenvalues $\{\lambda_i\}_{i=1}^n$ are positive (resp. nonnegative).

Singular Value Decomposition (SVD)

Fact (Compact SVD). Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with $\mathrm{rank}(\boldsymbol{A}) = r$. There exist $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$, and matrices

$$\boldsymbol{U} \in \mathbb{R}^{m\times r},\ \boldsymbol{V} \in \mathbb{R}^{n\times r},$$

with *orthonormal* columns ($\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}$, $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}$)such that

$$\boldsymbol{A} \;=\; \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \;=\; \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top, \quad \boldsymbol{\Sigma} \;=\; \mathrm{diag}\left(\sigma_1, \cdots, \sigma_r\right).$$

The $\{\sigma_i\}_{i=1}^r$ are called the *singular values* of $\boldsymbol{A}$;

The columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are called the (left & right) *singular vectors.*

Singular Value Decomposition (SVD)

Fact (properties of compact SVD).

Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with $\mathrm{rank}(\boldsymbol{A}) = r$ and compact SVD $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$

Let us denote:

$$\mathrm{range}(\boldsymbol{A}) \;:=\; \{\boldsymbol{A}\boldsymbol{x} \;\mid\; \boldsymbol{x} \in \mathbb{R}^n\} \;=\; \mathrm{col}(\boldsymbol{A}).$$

Then we have

- $\mathrm{range}(\boldsymbol{A}) \;=\; \mathrm{range}(\boldsymbol{U})$;
- $\mathrm{range}(\boldsymbol{A}^\top) \;=\; \mathrm{range}(\boldsymbol{V})$.

Singular Value Decomposition (SVD)

Relationship between eigen decomposition and SVD:

- The columns of $\boldsymbol{V}$ are eigenvectors of $\boldsymbol{A}^\top\boldsymbol{A}$;
- The columns of $\boldsymbol{U}$ are eigenvectors of $\boldsymbol{A}\boldsymbol{A}^\top$.

$$\sigma_i(\boldsymbol{A}) = \sqrt{\lambda_i(\boldsymbol{A}\boldsymbol{A}^\top)} = \sqrt{\lambda_i(\boldsymbol{A}^\top\boldsymbol{A})}$$

Singular Value Decomposition (SVD)

Theorem (best rank-r approximation).

Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with SVD, $\boldsymbol{A} \;=\; \sum_{i=1}^{\min\{m,n\}} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$. Then an optimal solution to the best rank-r approximation problem

$$\min_{\boldsymbol{X}} \left\|\boldsymbol{X} - \boldsymbol{A}\right\|_2, \quad \text{s.t.} \quad \mathrm{rank}(\boldsymbol{X}) \;\leq\; r$$

is the truncated SVD of the following form

$$\boxed{\widehat{\boldsymbol{A}}_r \;=\; \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top}, \quad \text{if } \sigma_r > \sigma_{r+1}, \text{ it is unique.}$$

Singular Value Decomposition (SVD)

Proof. Suppose $m \geq n$, then we observe that

$$\left\|\widehat{\boldsymbol{A}}_r - \boldsymbol{A}\right\|_2 \;=\; \left\|\sum_{i=r+1}^{n} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top\right\|_2 = \sigma_{r+1}$$

Second, to prove our result, we want to show that

$$\left\|\boldsymbol{B}_r - \boldsymbol{A}\right\|_2 \;\geq\; \sigma_{r+1}, \quad \forall \boldsymbol{B}_r \text{ with } \mathrm{rank}(\boldsymbol{B}_r) = r.$$

We suppose $\boldsymbol{B}_r = \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{V}}^\top$, $\widetilde{\boldsymbol{U}} \in \mathbb{R}^{m\times r}$, $\widetilde{\boldsymbol{V}} \in \mathbb{R}^{n\times r}$.

$$\begin{aligned}\left\|\boldsymbol{B}_r - \boldsymbol{A}\right\|_2^2 \;&=\; \sup_{\|\boldsymbol{z}\|_2 \leq 1} \left\|\left(\boldsymbol{B}_r - \boldsymbol{A}\right)\boldsymbol{z}\right\|_2^2 \\ &\geq\; \left\|\left(\boldsymbol{B}_r - \boldsymbol{A}\right)\boldsymbol{w}\right\|_2^2\end{aligned}$$

Singular Value Decomposition (SVD)

Since $\widetilde{\boldsymbol{V}}$ has r columns, then there must be a nontrivial linear combination of the first $r{+}1$ columns of $\boldsymbol{V}$, i.e.,

$$\boldsymbol{w} = \gamma_1\boldsymbol{v}_1 + \cdots + \gamma_{r+1}\boldsymbol{v}_{r+1}$$

such that $\widetilde{\boldsymbol{V}}^\top\boldsymbol{w} = \boldsymbol{0}$. Wlog, we can scale $\|\boldsymbol{w}\|_2 = 1$ and $\sum_{i=1}^{r+1}\gamma_i^2 = 1$

- Therefore, we have

$$\left\|\boldsymbol{A} - \boldsymbol{B}_k\right\|_2^2 \geq \left\|\left(\boldsymbol{A} - \boldsymbol{B}_k\right)\boldsymbol{w}\right\|_2^2 = \left\|\boldsymbol{A}\boldsymbol{w}\right\|_2^2 = \gamma_1^2\sigma_1^2 + \cdots + \gamma_{r+1}^2\sigma_{r+1}^2 \geq \sigma_{r+1}^2.$$

Matrix & Operator Norms

Definition. A vector norm is any real-valued function $\|\cdot\|$ that satisfies the following properties

- if $\boldsymbol{x} \neq \boldsymbol{0}$, then $\|\boldsymbol{x}\| > 0$
- for any $\alpha \neq 0$, then $\|\alpha\boldsymbol{x}\| = |\alpha|\,\|\boldsymbol{x}\|$
- triangle inequality: $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$.

There are many norms, for example

$$\|\boldsymbol{x}\|_p = \left(\sum_i |x_i|^p\right)^{1/p}, \text{ for any } p \geq 1.$$

Matrix & Operator Norms

Definition (matrix operator norm). Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. If $\|\cdot\|_a$ and $\|\cdot\|_b$ are norms on $\mathbb{R}^n$ and $\mathbb{R}^m$, we have

$$\boxed{\|\boldsymbol{A}\|_{a\to b} := \sup_{\|\boldsymbol{x}\|_a \le 1} \|\boldsymbol{A}\boldsymbol{x}\|_b .}$$

It satisfies the *three criteria* for norm, and is *submultiplicative*

$$\|\boldsymbol{A}\boldsymbol{B}\|_{a\to b} \le \|\boldsymbol{A}\|_{a\to b} \cdot \|\boldsymbol{B}\|_{a\to b} .$$

Matrix & Operator Norms

Definition (matrix operator norm). Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. If $\|\cdot\|_a$ and $\|\cdot\|_b$ are norms on $\mathbb{R}^n$ and $\mathbb{R}^m$, we have

$$\boxed{\|\boldsymbol{A}\|_{a\to b} := \sup_{\|\boldsymbol{x}\|_a \le 1} \|\boldsymbol{A}\boldsymbol{x}\|_b .}$$

- $\|\boldsymbol{A}\|_{2\to 2} = \sigma_1(\boldsymbol{A})$ (**spectral norm**, write as $\|\boldsymbol{A}\|$);
- $\|\boldsymbol{A}\|_{1\to b} = \max_{j=1,\cdots,n} \|\boldsymbol{A}\boldsymbol{e}_j\|_b$
- $\|\boldsymbol{A}\|_{a\to\infty} = \max_{i=1,\cdots,m} \|\boldsymbol{e}_i^*\boldsymbol{A}\|_b^*$, $\|\boldsymbol{v}\|_b^* := \sup_{\|\boldsymbol{u}\|_b \le 1} \langle \boldsymbol{u}, \boldsymbol{v}\rangle$.

Matrix & Operator Norms

Definition (unitary invariant matrix norms). Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. We say the matrix norm is unitary invariant if

$$\|\boldsymbol{A}\|_\sharp = \|\boldsymbol{P}\boldsymbol{A}\boldsymbol{Q}\|_\sharp , \quad \forall\, \boldsymbol{P} \in O(m),\ \boldsymbol{Q} \in O(n).$$

For example:

- **Spectral norm.** $\|\boldsymbol{A}\|_{2\to 2} = \sigma_1(\boldsymbol{A}) = \|\boldsymbol{\sigma}(\boldsymbol{A})\|_\infty$
- **Frobenius norm.**

$$\|\boldsymbol{A}\|_F = \sqrt{\textstyle\sum_{i=1}^{\min\{m,n\}} \sigma_i^2(\boldsymbol{A})} = \|\boldsymbol{\sigma}(\boldsymbol{A})\|_2$$

Matrix & Operator Norms

Definition (Schatten p-norm).
Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. For any $p \in [1, +\infty)$, the function

$$\boxed{\|\boldsymbol{A}\|_{\mathrm{S}_p} := \|\boldsymbol{\sigma}(\boldsymbol{A})\|_p}$$

is a norm on $\mathbb{R}^{m\times n}$.

Special case of great interest: **nuclear/trace norm**

$$\|\boldsymbol{A}\|_* := \|\boldsymbol{A}\|_{\mathrm{S}_1} = \textstyle\sum_i \sigma_i(\boldsymbol{A}) = \|\boldsymbol{\sigma}(\boldsymbol{A})\|_1$$

Lecture Agenda

- Basics of Matrix Analysis
- **Taylor Expansion & Lipschitz Function**
- Optimality Conditions
- Rate of Convergence

Lipschitz Continuity

Suppose that

- $f : \mathcal{X} \mapsto \mathcal{Y}$ with $\mathcal{X}$ and $\mathcal{Y}$ being open sets;
- $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$ are norms on $\mathcal{X}$ and $\mathcal{Y}$, respectively.

Definition. $f(\cdot)$ is Lipschitz continuous over $\mathcal{X}$ if $\exists L < \infty$, such that

$$\|f(\boldsymbol{y}) - f(\boldsymbol{x})\|_{\mathcal{Y}} \le L(\boldsymbol{x}) \|\boldsymbol{y} - \boldsymbol{x}\|_{\mathcal{X}} , \quad \forall \boldsymbol{x},\ \boldsymbol{y} \in \mathcal{X}.$$

Example I: ℓ_1-norm

ℓ^1-norm: $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \le \|\boldsymbol{x} - \boldsymbol{y}\|_1 \le \sqrt{n}\, \|\boldsymbol{x} - \boldsymbol{y}\|_2$$

- A subgradient of $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$ is not Lipschitz;
- The function $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$ is *not* continuously *differentiable*, but Lipschitz continuous

Lipschitz Continuity

Let f and g be Lipschitz continuous functions with best Lipschitz constants L_f and L_g, respectively.

$h(\boldsymbol{x})$	L_h	
$\alpha f(\boldsymbol{x}) + \beta$	$\lvert\alpha\rvert\, L_f$	scale/shift
$f(\boldsymbol{x} - \boldsymbol{x}_0)$	L_f	translate
$f(\boldsymbol{x}) + g(\boldsymbol{x})$	$\le L_f + L_g$	add
$f(g(\boldsymbol{x}))$	$\le L_f L_g$	compose (HW)
$\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$	$\|\boldsymbol{A}\|$	affine (for same norm on $\mathbb{F}^M$ and $\mathbb{F}^N$)
$f(\boldsymbol{x})g(\boldsymbol{x})$	?	multiply

Smooth Functions

Definition. A differentiable function $f(\boldsymbol{x})$ is called **smooth** *iff* it has a *Lipschitz continuous gradient*, i.e., *iff* $L < +\infty$ such that

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{z})\|_2 \leq L \|\boldsymbol{x} - \boldsymbol{z}\|_2, \quad \forall \boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^n$$

Note: Lipschitz continuity of ∇f is a stronger condition than mere continuity, so any differentiable function whose gradient is Lipschitz continuous is a **continuously differentiable** function.

Mean Value Theorem I

Theorem. Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be continuously differentiable. For any fixed $\boldsymbol{x}$ and $\boldsymbol{y}$, we have

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{z}(t_L)), \boldsymbol{y} - \boldsymbol{x} \rangle$$

for some $\boldsymbol{z}(t_L) = (1 - t_L) \cdot \boldsymbol{x} + t_L \cdot \boldsymbol{y}$ with $t_L \in (0, 1)$.

Mean Value Theorem and Lipschitzness

- If $f : \mathbb{R}^n \mapsto \mathbb{R}$ is continuously differentiable,

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq \sup_{\boldsymbol{z} \in \mathbb{R}^n} \|\nabla f(\boldsymbol{z})\|_2 \|\boldsymbol{x} - \boldsymbol{y}\|_2$$

so that the Lipschitz constant of f is $\sup_{\boldsymbol{z} \in \mathbb{R}^n} \|\nabla f(\boldsymbol{z})\|$

- If $f : \mathbb{R}^n \mapsto \mathbb{R}$ is twice continuously differentiable,

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq \sup_{\boldsymbol{z} \in \mathbb{R}^n} \left\|\nabla^2 f(\boldsymbol{z})\right\|_2 \|\boldsymbol{x} - \boldsymbol{y}\|_2$$

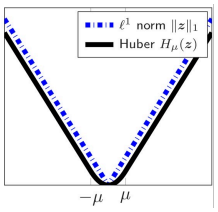
so that the Lipschitz constant of ∇f is $\sup_{\boldsymbol{z} \in \mathbb{R}^n} \left\|\nabla^2 f(\boldsymbol{z})\right\|_2$

Example II: Huber Function

Huber function: $H_\mu(\boldsymbol{x}) = \sum_{i=1}^n h_\mu(x_i)$, $h_\mu(x) = \begin{cases} |x| & |x| \geq \mu \\ \frac{x^2}{2\mu} + \frac{\mu}{2} & |x| < \mu \end{cases}$

By mean value theory,

$$\begin{aligned} |h_\mu(x) - h_\mu(y)| &\leq \sup_z \left|h'_\mu(z)\right| |x - y| \\ &\leq |x - y| \\ \left|h'_\mu(x) - h'_\mu(y)\right| &\leq \sup_z \left|\frac{\partial^2}{\partial z^2} h_\mu(z)\right| |x - y| \\ &\leq \frac{1}{\mu} |x - y| \end{aligned}$$



Mean Value Theorem II

Theorem. Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be twice continuously differentiable. For any fixed $\boldsymbol{x}$ and $\boldsymbol{y}$, we have

$$\nabla f(\boldsymbol{y}) = \nabla f(\boldsymbol{x}) + \int_0^1 \nabla^2 f(\boldsymbol{z}(t)) \cdot (\boldsymbol{y} - \boldsymbol{x}) dt$$

for some $\boldsymbol{z}(t) = (1 - t)\boldsymbol{x} + t\boldsymbol{y}$ with $t \in (0, 1)$, and that

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{1}{2} (\boldsymbol{y} - \boldsymbol{x})^\top \nabla^2 f(\boldsymbol{z}(t_Q)) (\boldsymbol{y} - \boldsymbol{x})$$

for some $t_Q \in (0, 1)$.

Taylor Approximation Theory I

Theorem (first-order approximation).

Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be continuously differentiable, and $\nabla f(\boldsymbol{x})$ is $\gamma_L(\boldsymbol{x}_0)$-Lipschitz continuous at $\boldsymbol{x}_0$, then

$$\left|f(\boldsymbol{x}) - \widehat{f}_L(\boldsymbol{x}; \boldsymbol{x}_0)\right| \leq \frac{\gamma_L(\boldsymbol{x}_0)}{2} \|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2,$$

where we define

$$\widehat{f}_L(\boldsymbol{x}; \boldsymbol{x}_0) := f(\boldsymbol{x}_0) + \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{x} - \boldsymbol{x}_0 \rangle.$$

Taylor Approximation Theory II

Theorem (second-order approximation).

Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be twice continuously differentiable, and suppose $\nabla^2 f(\boldsymbol{x})$ is $\gamma_Q(\boldsymbol{x}_0)$-Lipschitz continuous at $\boldsymbol{x}_0$, then

$$\left|f(\boldsymbol{x}) - \widehat{f}_Q(\boldsymbol{x}; \boldsymbol{x}_0)\right| \leq \frac{\gamma_Q(\boldsymbol{x}_0)}{6} \|\boldsymbol{x} - \boldsymbol{x}_0\|_2^3,$$

where we define

$$\begin{aligned} \widehat{f}_Q(\boldsymbol{x}; \boldsymbol{x}_0) &:= f(\boldsymbol{x}_0) + \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{x} - \boldsymbol{x}_0 \rangle \\ &\quad + \frac{1}{2} (\boldsymbol{x} - \boldsymbol{x}_0)^\top \nabla^2 f(\boldsymbol{x}_0) (\boldsymbol{x} - \boldsymbol{x}_0). \end{aligned}$$

Lecture Agenda

- Basics of Matrix Analysis
- Taylor Expansion & Lipschitz Function
- **Optimality Conditions**
- Rate of Convergence

Optimality: Unconstrained Problem

Consider a continuously differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \; f(\boldsymbol{x})$$

Definition. A point $\boldsymbol{x}_\star$ is a *stationary point* of $f(\cdot)$, if

$$\nabla f(\boldsymbol{x}_\star) \; = \; \mathbf{0}.$$

Here, $\boldsymbol{x}_\star$ is also known as a *critical point*.

Optimality: Unconstrained Problem

- **First-order *necessary* condition for optimality**

If $\boldsymbol{x}_\star$ is a local minimizer of $f : \mathbb{R}^n \mapsto \mathbb{R}$, and f is continuously differentiable in an open neighborhood of $\boldsymbol{x}_\star$, then we must have

$$\boxed{\nabla f(\boldsymbol{x}_\star) \; = \; \mathbf{0}.}$$

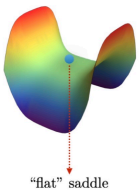
Optimality: Unconstrained Problem

- **Second-order *necessary* condition for optimality**

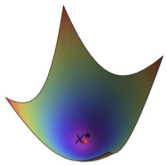
If $\boldsymbol{x}_\star$ is a local minimizer of $f : \mathbb{R}^n \mapsto \mathbb{R}$, and f is *twice* continuously differentiable in an open neighborhood around $\boldsymbol{x}_\star$, then we must have

$$(i) \; \nabla f(\boldsymbol{x}_\star) \; = \; \mathbf{0};$$
$$(ii) \; \nabla^2 f(\boldsymbol{x}_\star) \succeq \mathbf{0}.$$

Optimality: Unconstrained Problem



$f(\boldsymbol{x}) \; = \; x_1^3 - x_2^3$



$f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}, \;\; \boldsymbol{A} \succ \mathbf{0}.$

Optimality: Unconstraint Problem

- **Second-order *sufficient* condition for optimality**

If $f : \mathbb{R}^n \mapsto \mathbb{R}$ is *twice* continuously differentiable in an open neighborhood around $\boldsymbol{x}_\star$ and that

$$(i) \; \nabla f(\boldsymbol{x}_\star) \; = \; \mathbf{0};$$
$$(ii) \; \nabla^2 f(\boldsymbol{x}_\star) \succ \mathbf{0},$$

Global Optimality of Convex Functions

Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a *convex* function, then

- A local minimizer of f is also its global minimizer. If f is *strictly* convex, the global minimizer is *unique*.
- A point is a global minimizer of f *iff*

$$\mathbf{0} \; \in \; \partial f(\boldsymbol{x}_\star).$$

If $f \in \mathcal{C}^1$, then $\nabla f(\boldsymbol{x}_\star) = \mathbf{0}$ implies that $\boldsymbol{x}_\star$ is a *global* minimizer.

Optimality: Constrained Problem

Consider a smooth contrained problem with

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}), \quad \text{s.t.} \quad r_i(\boldsymbol{x}) \; = \; 0, \; 1 \le i \le p,$$
$$h_j(\boldsymbol{x}) \; \le \; 0, \; 1 \le j \le q.$$

Consider its *Lagrangian* function

$$\mathscr{L}(\boldsymbol{x}, \boldsymbol{u}, \boldsymbol{v}) \; = \; f(\boldsymbol{x}) + \sum_{i=1}^{p} u_i \cdot r_i(\boldsymbol{x}) + \sum_{j=1}^{q} v_j \cdot h_j(\boldsymbol{x}).$$

Optimality: Constraint Problem

First-order necessary condition (aka KKT condition)

- **Stationary**

$$\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}_\star, \boldsymbol{u}_\star, \boldsymbol{v}_\star) \; = \; \mathbf{0}$$

- **Feasibility**

(**Primal**) $r_i(\boldsymbol{x}_\star) \; = \; 0, \; 1 \le i \le p,$
$h_j(\boldsymbol{x}_\star) \; \le \; 0, \; 1 \le j \le q.$

(**Dual**) $\boldsymbol{v}_\star \ge \mathbf{0}.$

- **Complimentary Slackness**

$$h_j(\boldsymbol{x}_\star) \cdot v_{\star j} \; = \; 0, \quad \forall \, 1 \le j \le q.$$

Optimality: Constraint Problem

- KKT condition is only a **necessary condition** for general constraint nonlinear problems;
- For *convex problems* under certain constraint qualifications, it is also a *sufficient* condition.

Lecture Agenda

- Basics of Matrix Analysis
- Taylor Expansion & Lipschitz Function
- Optimality Conditions
- **Rate of Convergence**

Algorithmic Convergence

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}), \quad \text{s.t.} \quad \boldsymbol{x} \in \mathcal{C}.$$

Solve the problem via iterative methods of optimization, which produce a sequence of points

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_k, \cdots$$

starting from an initialization $\boldsymbol{x}_0$.

Algorithmic Convergence

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}), \quad \text{s.t.} \quad \boldsymbol{x} \in \mathcal{C}.$$

- $\mathcal{X}$ is the *set of all optimal solutions* $\boldsymbol{x}_\star$;
- $f_\star = f(\boldsymbol{x}_\star)$ is the *optimal function value*;
- ε is the user-defined *precision value*.

Algorithmic Convergence

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}), \quad \text{s.t.} \quad \boldsymbol{x} \in \mathcal{C}.$$

- **Total computation time:**

$$\boxed{\text{complexity} = \text{per iter cost} \times \#\text{of iterations.}}$$

- **Per iteration cost:** how much computation it takes to generate the next point $\boldsymbol{x}_{k+1}$ from $\boldsymbol{x}_k$.

Algorithmic Convergence

$$\boxed{\text{complexity} = \text{per iter cost} \times \#\text{of iterations.}}$$

Convergence rate. How quickly the sequence $\{\boldsymbol{x}_k\}_{k\geq 1}$ converges, measured by

Distance to a minimizer	$\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2 \leq \varepsilon$
Sub-optimality in objective	$\|f(\boldsymbol{x}_k) - f(\boldsymbol{x}_\star)\| \leq \varepsilon$
Gradient	$\|\nabla f(\boldsymbol{x}_k)\|_2 \leq \varepsilon$

Definitions of Convergence

- **Iterate convergence to the (set) minimizer**

$$\mathrm{d}(\boldsymbol{x}_k, \mathcal{X}) = \min_{\boldsymbol{x}_\star \in \mathcal{X}} \|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2 \leq \varepsilon.$$

- **Zero-th order function value convergence**

$$|f(\boldsymbol{x}_k) - f_\star| \leq \varepsilon$$

Function value convergence does *not* imply iterate convergence.

Definitions of Convergence

- **First-order function value convergence**

$$\|\nabla f(\boldsymbol{x}_k)\|_2 \leq \varepsilon.$$

  - For *convex functions*, this also means convergence to *the global minimizer*
  - For *nonconvex functions*, this only means convergence to a *stationary point* (e.g., a local minimizer or a saddle point).

Rate of Convergence

Definition. We say the Q-convergence is of order $p\ (\geq 1)$ and with factor $\gamma\ (> 0)$, if $\exists k_0$, such that $\forall k \geq k_0$:

$$\min_{\boldsymbol{x}_\star \in \mathcal{X}} \|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2 \ \leq\ \gamma \cdot \left(\min_{\boldsymbol{x}_\star \in \mathcal{X}} \|\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star\|_2 \right)^p.$$

- The larger p is, the faster the convergence.
- The smaller γ is, the faster convergence (with p fixed).
- Typically look for the largest p and the smallest γ such that the inequality holds.

Q-Convergence

The following sequence

$$\{a_k\}_{k\geq 1} \ =\ \left\{2, 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \cdots, \frac{1}{2^k}, \cdots\right\}$$

- **Q-Linear Convergence** $(p = 1,\ \gamma < 1)$

$$\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2 \ \leq\ \gamma \|\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star\|_2$$

Q-Convergence

The following sequence is Q-linearly convergent

$$\{a_k\}_{k\geq 1} \ =\ \left\{2, 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \cdots, \frac{1}{2^k}, \cdots\right\}$$

with

$$\gamma \ =\ \lim_{k\to\infty} \frac{|1/2^{k+1} - 0|}{|1/2^k - 0|} \ =\ \lim_{k\to\infty} \frac{2^k}{2^{k+1}} \ =\ \frac{1}{2}.$$

R-Convergence

Q-convergence does not include some sequences that converges reasonably fast:

$$\{b_k\}_{k\geq 1} \ =\ \left\{1, 1, \frac{1}{4}, \frac{1}{4}, \frac{1}{16}, \frac{1}{16}, \cdots, \frac{1}{4^{\lfloor k/2 \rfloor}}, \cdots\right\}$$

Definition. We say the R-convergence is of order $p\ (\geq 1)$,

$$\min_{\boldsymbol{x}_\star \in \mathcal{X}} \|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2 \ \leq\ \rho_k,$$

if the sequence $\{\rho_k\}_{k\geq 1}$ is Q-convergence of the order p.

Rate of Convergence

- **Q-Sublinear Convergence** $(p = 1,\ \gamma = 1)$

$$\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2 \ \leq\ k^{-1/r} \|\boldsymbol{x}_0 - \boldsymbol{x}_\star\|_2$$

  - ➢ implies k is order $\mathcal{O}(\frac{1}{\varepsilon^r})$.

Rate of Convergence

- **Q-Linear Convergence** $(p = 1,\ \gamma < 1)$

$$\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2 \ \leq\ \gamma \|\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star\|_2$$

  - ➢ $\gamma < 1$, implies k is order $\mathcal{O}(\log \frac{1}{\varepsilon})$

Rate of Convergence

- **Q-Superlinear Convergence** $(1 < p\ (\leq 2))$

$$\frac{\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2}{\|\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star\|_2} \ \to\ 0$$

Rate of Convergence

- **Quadratic Convergence** $(p = 2)$

$$\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2 \ \leq\ \gamma \|\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star\|_2^2$$

  - implies k is order $\mathcal{O}(\log\log \frac{1}{\varepsilon})$;
  - $\gamma > 0$ here does not need to be smaller than 1;
  - implies superlinear convergence;

Rate of Convergence

Methods	Convex Nonsmooth	Convex smooth	Strongly Convex	Strict Saddle
Subgradient Method	$O(1/\sqrt{k})$			
Gradient Descent		$O(1/k^2)$	Linear	$O(1/k^2)$
Proximal Method	$O(1/k^2)$		Linear	
Quasi-Newton Method			Superlinear	
Newton Method			Quadratic	Quadratic

References & Readings

- *Numerical Optimization*, Jorge Nocedal, and Stephen Wright, Springer. (**Chapter 2**)
- *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. John Wright, Yi Ma. (**Appendix A-C**)