Summary of Matrix Computation

Jinhong Du

December 4, 2019

Contents

Chapter 1

Basics of Matrix

1.1 Norms

1.1.1 Definition

Definition. Norms

A norm is a real-valued function on $\mathbb{V}$

$$\|\cdot\| : \mathbb{V} \to \mathbb{R}$$

such that

(i) $\|\mathbf{v}\| \geq 0$ for all $\mathbf{v} \in \mathbb{V}$;

(ii) $\|\mathbf{v}\| = 0$ if and only if $\mathbf{v} = 0$;

(iii) $\|\alpha\mathbf{v}\| = |\alpha|\|\mathbf{v}\|$ for any $\alpha \in \mathbb{C}$ and $\mathbf{v} \in \mathbb{V}$;

(iv) $\|\mathbf{v} + \mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{w}\|$ for any $\mathbf{v}, \mathbf{w} \in \mathbb{V}$.

We will restrict our discussion to $\mathbf{V} = \mathbb{R}^n, \mathbb{C}^n, \mathbb{R}^{m\times n}$ and $\mathbb{C}^{m\times n}$ in the main content.

1.1.2 Properties

Continuity of Vector Norms

Theorem. Continuity of Vector Norms

The vector norm $\|\cdot\| : \mathbb{V} \to \mathbb{R}$ is a continuous function, i.e.

$$\lim_{n\to\infty} \|\mathbf{v}_n\| = \|\lim_{n\to\infty} \mathbf{v}_n\|.$$

Equivalence of Vector Norms

Definition. Equivalence of Norms

The vector norms $\|\cdot\|_a$ and $\|\cdot\|_b$ are equivalent if there exists $c_1, c_2 > 0$ such that $\forall\ \mathbf{v}$,

$$c_1\|\mathbf{v}\|_b \leq \|\mathbf{v}\|_a \leq c_2\|\mathbf{v}\|_b.$$

Theorem. Equivalence of Norms

All vector norms on $\mathbb{V}$ are equivalent if V is finite-dimensional.

1.1.3 Vector Norms

Definition. Vector Norms

A vector norm is a real-valued function on $\mathbb{V} = \mathbb{R}^n$ or $\mathbb{C}^n$

$$\|\cdot\| : \mathbb{V} \to \mathbb{R}$$

such that

(i) $\|\mathbf{v}\| \geq 0$ for all $\mathbf{v} \in \mathbb{V}$;
(ii) $\|\mathbf{v}\| = 0$ if and only if $\mathbf{v} = 0$;
(iii) $\|\alpha\mathbf{v}\| = |\alpha|\|\mathbf{v}\|$ for any $\alpha \in \mathbb{C}$ and $\mathbf{v} \in \mathbb{V}$;
(iv) $\|\mathbf{v} + \mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{w}\|$ for any $\mathbf{v}, \mathbf{w} \in \mathbb{V}$.

Some common vector norms are given by

- *l_1-norm* - $\|\mathbf{x}\|_1 = \sum_{k=1}^{n} |x_k|$;
- *l_2-norm/Euclidean norm* - $\|\mathbf{x}\|_2 = \sqrt{\sum_{k=1}^{n} |x_k|^2}$;
- *l_∞-norm/Chebyshev norm* - $\|\mathbf{x}\|_1 = \max\{|x_1|, \ldots, |x_n|\}$;
- *l_p-norm* - $\|\mathbf{x}\|_1 = \sqrt[p]{\sum_{k=1}^{n} |x_k|^p}$ for $p \in [1, \infty]$;
- *Mahalanobis norm* on $\mathbb{R}^n$ - $\|\mathbf{x}\|_A = \sqrt{\mathbf{x}^\top A\mathbf{x}}$ where $A \in \mathbb{R}^{n\times n}$ is positive-definite.

1.1.4 Matrix Norms

Definition. Matrix Norms

A matrix norm is a real-valued function on $\mathbb{V} = \mathbb{R}^{m\times n}$ or $\mathbb{C}^{m\times n}$

$$\|\cdot\| : \mathbb{V} \to \mathbb{R}$$

such that

(i) $\|A\| \geq 0$ for all $A \in \mathbb{V}$;

(ii) $\|A\| = 0$ if and only if $A = 0$;

(iii) $\|\alpha A\| = |\alpha|\|A\|$ for any $\alpha \in \mathbb{C}$ and $A \in \mathbb{V}$;

(iv) $\|A + B\| \leq \|A\| + \|B\|$ for any $A, B \in \mathbb{V}$.

Furthermore, if $\|\cdot\|$ also satisfies

(v) $\|AB\| \leq \|A\|\|B\|$ for any $A, B \in \mathbb{V}$. (sub-multiplicativity)

then $\|\cdot\|$ is a sub-multiplicative matrix norm.

Some common matrix norms are given by

- *Holder norm* - $\|A\|_{H,p} = \left(\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^p\right)^{\frac{1}{p}}$ for $p \in [1, \infty]$;
- *Frobenius norm* - $\|A\|_F = \|A\|_{H,2} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2}$;

1.1.5 Operation Norms

Definition. Operation Norms/Induced Norms

For $A \in \mathbb{V}^{m\times n}$ ($\mathbb{V} = \mathbb{R}$ or $\mathbb{C}$), if $\|\cdot\|_a$ and $\|\cdot\|_b$ are two vector norms on $\mathbb{V}^n$ and $\mathbb{V}^m$ respectively, then the matrix norm

$$\begin{aligned}\|A\|_{a,b} &= \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_a}{\|\mathbf{x}\|_b} \\ &= \max_{\|\mathbf{x}\|_b=1} \|A\mathbf{x}\| \\ &= \max_{\|\mathbf{x}\|_b\leq 1} \|A\mathbf{x}\|\end{aligned}$$

is called *operator norm / induced norm.*

Some common operator norms are given by

- *operator (p,q)-norm* - $\|A\|_{p,q}$;
- *operator p-norm* - $\|A\|_p = \|A\|_{p,q}$;
- *spectral p-norm* - $\|A\|_2 = \|A\|_{2,2}$;
- *spectral 1-norm* - $\|A\|_1 = \max_j \sum_{i=1}^{m}|a_{ij}|$;
- *spectral ∞-norm* - $\|A\|_\infty = \max_i \sum_{j=1}^{n}|a_{ij}|$.

Any operator norm is consistent in the sense that $\|A\mathbf{x}\|_a \leq \|A\|_{a,b}\|\mathbf{x}\|_b$. The operator may not be sub-multiplicative in general. However, in the special case, the operator p-norms are sub-multiplicative. More generally, given $A \in \mathbb{V}^{m\times n}$ and $B \in \mathbb{V}^{n\times p}$, it may always true that

$$\|AB\|_{\alpha,\gamma} \leq \|A\|_{\alpha,\beta}\|B\|_{\beta,\gamma}$$

for any vector norms $\|\cdot\|_\gamma$ on $\mathbb{V}^p$, $\|\cdot\|_\alpha$ on $\mathbb{V}^m$, $\|\cdot\|_\beta$ on $\mathbb{V}^n$.

1.2 Matrix Operations

1.2.1 Inner Product

1.2.2 Outer Product

1.2.3 Matrix Product

Chapter 2

Spectral theorem

2.1 Spectial Matrix

	Field		
	$\mathbb{R}$		$\mathbb{C}$
Symmetric	$A = A^\top$	Hermitian	$A = A^*$
Normal	$AA^\top = A^\top A$	Normal	$AA^* = A^*A$
Orthogonal	$AA^\top = A^\top A = I$	Unitary	$AA^* = A^*A = I$

Table 2.1: Special Matrices.

2.2 Eigenvalue Decomposition (EVD)

2.3 Jordan Canonical Form (JCF)

2.4 Schur Decomposition

2.5 Spectral Radius

Definition. **Spectral Radius**

The *spectral radius* is a function $C^{n\times n} \to \mathbb{R}$ such that

$$\rho(A) = \max\{|\lambda_1(A)|, \ldots, |\lambda_n(A)|\},$$

where $A \in \mathbb{R}^{n\times n}$.

Notice that the spectral radius is not a norm since for $J = \left[\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right]$, $\rho(J) = 0$.

The spectral radius is bounded by the spectral norm,

$$\rho(A) = |\lambda_1| = \frac{\|A\mathbf{x}_1\|_2}{\|\mathbf{x}_1\|_2} \leq \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \|A\|_2.$$

Furthermore, the spectral radius is bounded by any consistent norms as

$$\rho(A) = |\lambda_1| = \frac{\|A\mathbf{x}_1\|}{\|\mathbf{x}_1\|} \leq \|A\|.$$

Theorem. The Operator Norm Bounded by Spectral Radius

Given any $A \in \mathbb{C}^{n\times n}$, any $\epsilon > 0$, there exists an operator norm $\|A\|_\alpha = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_\alpha}{\|\mathbf{x}\|_\alpha}$ such that $\|A\|_\alpha \leq \rho(A) + \epsilon$.

Theorem. Limitation of Matrix Norms And Spectral Radius

Let $\|\cdot\|$ be any matrix norm, then $\forall\ A \in \mathbb{C}^{n\times n}$,

$$\lim_{k\to\infty} \|A^k\|^{\frac{1}{k}} = \rho(A).$$

Theorem. Limitation of Matrices And Spectral Radius

$\lim_{k\to\infty} A^k = \mathbf{0}$ if and only if $\rho(A) < 1$.

Chapter 3

Singular Value Decomposition (SVD)

3.1 Definition

3.1.1 Full SVD

3.1.2 Condensed SVD

3.1.3 Outer-Product SVD

3.2 Relation between SVD And Norms

$$\|A\|_2 = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \sigma_1.$$

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2} = \sqrt{\sum_{i=1}^{r} \sigma_i^2}, \qquad r = \operatorname{rank}(A)$$

3.3 Pseudoinverse And Condition Number

3.3.1 Pseudoinverse

Definition. Pseudoinverse (1)

Let $A \in \mathbb{C}^{m\times n}$. Then there exists the pseudoinverse of A, a unique $A^\dagger \in \mathbb{C}^{n\times m}$ such that

(i) $(A^\dagger A)^* = A^\dagger A$;

(ii) $(AA^\dagger)^* = AA^\dagger$;

(iii) $A^\dagger AA^\dagger = A$;

(iv) $AA^\dagger A = A^\dagger$.

Definition. Pseudoinverse (2)

Let $A \in \mathbb{C}^{m\times n}$, $A = U\Sigma V^*$ be the SVD of A and $r = \operatorname{rank}(A)$. Define $\Sigma^\dagger = \begin{bmatrix} \frac{1}{\sigma_1} & & & & \\ & \ddots & & & \\ & & \frac{1}{\sigma_r} & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{bmatrix} \in \mathbb{C}^{n\times m}$ as the pseudoinverse of diagonal matrix $\Sigma = \begin{bmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_r & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{bmatrix} \in \mathbb{C}^{m\times n}$. Then the pseudoinverse of A is given by $A^\dagger = V\Sigma^\dagger U^* \in \mathbb{C}^{n\times m}$.

Definition. Pseudoinverse (3)

Let $A \in \mathbb{C}^{m\times n}$. For any $\mathbf{x} \in \mathbb{C}^n$ and $\mathbf{b} \in \mathbb{C}^m$, the minimum norm least squares solution is given by $\mathbf{x} = A^\dagger \mathbf{b}$ uniquely. Here $A^\dagger \in \mathbb{C}^{n\times m}$ is called the pseudoinverse of A.

3.3.2 Condition Number

Definition. Condition Number

For $A \in \mathbb{C}^{n\times n}$, the *condition number* is given by

$$\kappa_p(A) = \|A\|_p \|A^{-1}\|_p,$$

for $p \in [1, \infty]$. When A is singular, $\kappa_p(A) = \infty$.
For $A \in \mathbb{C}^{m\times n}$, the *generalized condition number* is given by

$$\kappa_p(A) = \|A\|_p \|A^{\dagger}\|_p,$$

for $p \in [1, \infty]$.

Notice that the generalized condition number is well-defined even when A is nonsingular or A is not square. For the generalized condition number, we will usually use spectral norm, which yields that $\kappa_2(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)} \geq 1$. Also, for any unitary matrix $Q \in \mathbb{C}^{n\times n}$, $\kappa_2(Q) = 1$, that is why we always want to use unitarily transformations in the computation so that we can control the errors.

3.4 Best Rank-r Approximation Problem

Theorem. Eckart-Young

For $A \in \mathbb{C}^{m\times n}$, the solution to

$$\min_{\substack{X\in\mathbb{C}^{m\times n}\\ \operatorname{rank}(X)\le r}} \|A - X\|_2$$

or

$$\min_{\substack{X\in\mathbb{C}^{m\times n}\\ \operatorname{rank}(X)\le r}} \|A - X\|_F$$

is given by

$$X_r = U \begin{bmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_r & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{bmatrix} V^* = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^*,$$

where $A = U \begin{bmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_k & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{bmatrix} V^*$ is the SVD of A and $k = \operatorname{rank}(A) \ge r$.

Chapter 4

Matrix Subspaces

4.1 Four Foundamental Subspaces

For $A \in \mathbb{C}^{m\times n}$, the four foundamental subspaces are

1. $\ker(A) = \{\mathbf{x} \in \mathbb{C}^n : \mathbf{0} = A\mathbf{x}\}$;
2. $\mathrm{im}(A) = \{\mathbf{y} \in \mathbb{C}^m : \mathbf{y} = A\mathbf{x} \text{ for some } \mathbf{x} \in \mathbb{C}^n\}$;
3. $\ker(A^*) = \{\mathbf{x} \in \mathbb{C}^m : \mathbf{0} = A^*\mathbf{x}\}$;
4. $\mathrm{im}(A^*) = \{\mathbf{y} \in \mathbb{C}^n : \mathbf{y} = A^*\mathbf{x} \text{ for some } \mathbf{x} \in \mathbb{C}^m\}$.

Theorem. Fredholm Alternative

$$\mathbb{C}^m = \mathrm{im}(A) \oplus \ker(A^*)$$
$$\mathbb{C}^n = \mathrm{im}(A^*) \oplus \ker(A)$$

Let $A = U\Sigma V^*$ be the SVD decomposition of A and $r = \mathrm{rank}(A)$, then

1. $\ker(A) = \{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$;
2. $\mathrm{im}(A) = \{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$;
3. $\ker(A^*) = \{\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m\}$;
4. $\mathrm{im}(A^*) = \{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$.

4.2 Projection Matrix

Definition. Projection Matrix

The *projection matrix* is a matrix $P \in \mathbb{C}^{n\times n}$ such that $P^2 = P$.

The *orthogonal projection matrix* is a matrix $P \in \mathbb{C}^{n\times n}$ such that $P^2 = P$ and $P^* = P$.

For any subsapce $W \subseteq \mathbb{C}^n$, P_W means a projection matrix with $\mathrm{im}(P_W) = W$.

If P is an (orthogonal) projection matrix onto W, then $I_n - P$ is also an (orthogonal) projection matrix onto $W^\perp$.

For matrix $A \in \mathbb{C}^{m\times n}$,

1. $P_{\mathrm{im}(A)} = AA^\dagger$;
2. $P_{\ker(A^*)} = I_m - AA^\dagger$;
3. $P_{\mathrm{im}(A^*)} = A^\dagger A$;
4. $P_{\ker(A)} = I_n - A^\dagger A$;

Chapter 5

Error Analysis

5.1 Types of Errors

5.2 Backward Error Analysis

5.3 Forward-Backward Error Analysis

Chapter 6

Matrix Factorization

6.1 Rank-Retaining Decomposition

6.1.1 Definition

Definition. Rank-Retaining Decomposition

Rank-retaining decomposition of $A \in \mathbb{C}^{m\times n}$ is given by

$$A = GH$$

where $G \in \mathbb{C}^{m\times r}$, $H \in \mathbb{C}^{r\times n}$ and $r = \text{rank}(A) = \text{rank}(G) = \text{rank}(H)$.

Notice that any decomposition of A that satisfies the above condition can be a rank-retaining decomposition. So the rank-retaining decomposition is not unique.

6.1.2 Properties

(1) $G^*G \in \mathbb{C}^{r\times r}$ is invertible;
(2) $HH^* \in \mathbb{C}^{r\times r}$ is invertible;
(3) $\text{im}(A) = \text{im}(G)$;
(4) $\ker(A) = \ker(H)$;
(5) $\text{im}(A^*) = \text{im}(H^*)$;
(6) $\ker(A^*) = \ker(G^*)$.

Proof. (1) Since $\mathbb{C}^m = \text{im}(G) \oplus \ker(G^*)$ and $\text{rank}(G) = \text{rank}(G^*) = r$, we have $\forall\, \mathbf{x} \in \mathbb{C}^r$, $G^*G\mathbf{x} = \mathbf{0} \Longrightarrow G\mathbf{x} = \mathbf{0} \Longrightarrow \mathbf{x} = \mathbf{0}$. Then $\text{nullity}(G^*G) = 0$ and $\text{rank}(G^*G) = r - \text{nullity}(G^*G) = r$, i.e., G^*G is full rank and thus invertible.

(2) Since $\mathbb{C}^n = \text{im}(H^*) \oplus \ker(H)$ and $\text{rank}(H) = \text{rank}(H^*) = r$, we have $\forall\, \mathbf{x} \in \mathbb{C}^r$, $HH^*\mathbf{x} = \mathbf{0} \Longrightarrow H^*\mathbf{x} = \mathbf{0} \Longrightarrow \mathbf{x} = \mathbf{0}$. Then $\text{nullity}(HH^*) = 0$ and $\text{rank}(HH^*) = r - \text{nullity}(HH^*) = r$, i.e.,

HH^* is full rank and thus invertible.

(3) $\forall\ \mathbf{y} \in \text{im}(A)$, $\exists\ \mathbf{x} \in \mathbb{C}^n$, s.t. $\mathbf{y} = A\mathbf{z} = G(H\mathbf{x})$, i.e., $\mathbf{y} \in \text{im}(G)$.

$\forall\ \mathbf{y} \in \text{im}(G)$, $\exists\ \mathbf{z} \in \mathbb{C}^n$, s.t. $\mathbf{y} = G\mathbf{z}$. Since H has full row rank, there exists $\mathbf{x} \in \mathbb{C}^n$ s.t. $\mathbf{z} = H\mathbf{x}$. Then, $\mathbf{y} = G\mathbf{z} = GH\mathbf{x} = A\mathbf{x}$, i.e., $\mathbf{y} \in \text{im}(A)$.

Therefore, $\text{im}(A) = \text{im}(G)$.

(4) $\forall\ \mathbf{z} \in \ker(A)$, $A\mathbf{z} = \mathbf{0} \Longrightarrow GH\mathbf{z} = \mathbf{0}$. Since G has full column rank, $H\mathbf{z} = \mathbf{0}$, i.e., $\mathbf{z} \in \ker(H)$.

$\forall\ \mathbf{z} \in \ker(H)$, $H\mathbf{z} = \mathbf{0} \Longrightarrow A\mathbf{z} = GH\mathbf{z} = \mathbf{0}$, i.e., $\mathbf{x} \in \ker(A)$.

Therefore, $\ker(A) = \ker(H)$.

(5) Since $\mathbb{C}^m = \text{im}(A) \oplus \ker(A^*) = \text{im}(H^*) \oplus \ker(H)$, we have $\text{im}(A^*) = \text{im}(H^*)$.

(6) Since $\mathbb{C}^m = \text{im}(A^*) \oplus \ker(A) = \text{im}(H) \oplus \ker(H^*)$, we have $\text{im}(A^*) = \ker(H^*)$.

■

Theorem. Rank-retaining Representation of Pseudoinverse

Let $A = GH$ be rank-retaining decomposition, then

$$A^\dagger = H^*(HH^*)^{-1}(G^*G)^{-1}G^*.$$

6.2 QR Factorization

Without further explanation, we will assum $m \geq n$ to avoid some extra discussion, However the case $m<n$ is similar.

6.2.1 Overview

The QR decomposition for $A \in \mathbb{C}^{m\times n}$ is given by $A = QR$ where $Q \in \mathbb{C}^{m\times m}$ is an unitary matrix and $R \in \mathbb{C}^{m\times n}$ is an upper-triangular matrix.

There are three common ways to coompute the QR decomposition:

1. *Gram–Schmidt* or *modfied Gram–Schmidt orthogonalization.*
2. *Householder reflections*;
3. *Givens rotations*, or *Jacobi rotations*;

The Gram-Schmidt algorithm is numerical unstable and is not usually used. Gram–Schmidt applies a sequence of triangular matrices to orthogonalize A as follows

$$AR_1^{-1}\cdots R_n^{-1} = Q,$$

while Householder and Givens QR apply a sequence of orthogonal matrices to triangularize A as follows

$$Q_n^\top \cdots Q_1^\top A = R,$$

where the use of unitary matrices prevent errors to increase too fast.

6.2.2 Gram-Schmidt Algorithm

Assuming that $\operatorname{rank}(A) = n$. Write $A = [\,\mathbf{a}_1 \cdots \mathbf{a}_n\,]$ and $Q = [\,\mathbf{q}_1 \cdots \mathbf{q}_n\,]$

Suppose that we have orthogonalize the first $k-1$ columns of A and get the first $k-1$ columns of Q and R, denoted by $Q_{k-1} = \begin{bmatrix}\mathbf{q}_1 & \cdots & \mathbf{q}_{k-1}\end{bmatrix}$ and $R_{k-1} = \begin{bmatrix}\mathbf{r}_1 & \cdots & \mathbf{r}_{k-1}\end{bmatrix}$ where $\mathbf{r}_{i,(i+1):m}$, the subvector of $\mathbf{r}_i$, are $\mathbf{0}$. For $\mathbf{q}_k$ and $\mathbf{r}_k$, notice that

$$\begin{aligned} Q_{k-1}^\top \mathbf{q}_k &= \mathbf{0} \\ \mathbf{a}_k &= Q_k \mathbf{r}_{k,1:k} \\ &= Q_{k-1}\mathbf{r}_{k,1:(k-1)} + \mathbf{q}_k r_{kk} \end{aligned}$$

we have

$$Q_{k-1}^\top \mathbf{a}_k = Q_{k-1}^\top Q_{k-1}\mathbf{r}_{k,1:(k-1)} + Q_{k-1}^\top \mathbf{q}_k \mathbf{r}_k = \mathbf{r}_{k,1:(k-1)}$$

which can use to solve $\mathbf{r}_{k,1:(k-1)}$. Then $\mathbf{q}_k r_{kk} = \mathbf{a}_k - Q_{k-1}\mathbf{r}_{k,1:(k-1)}$. Finally, we have

$$\begin{aligned} \mathbf{r}_{k,1:(k-1)} &= Q_{k-1}^\top \mathbf{a}_k \\ r_{kk} &= \|\mathbf{a}_k - Q_{k-1}\mathbf{r}_{k,1:(k-1)}\|_2 \\ \mathbf{q}_k &= \frac{\mathbf{a}_k - Q_{k-1}\mathbf{r}_{k,1:(k-1)}}{r_{kk}}. \end{aligned}$$

Notes:

1. If $m = n$, then after n steps, we can get the full QR decomposition of A. If $m > n$, then we need to find the orthogonal complement of $\mathrm{im}(Q_n)$.
2. If $\mathrm{rank}(A) < n$, the algorithm cannot proceed when $r_{kk} = 0$ and column permutation is need.

6.2.3 Householder Algorithm

Assuming that $\mathrm{rank}(A) = n$. The Householder reflection matrix is $H = H_{\mathbf{u}} = I - 2\mathbf{u}\mathbf{u}^\top \in \mathbb{C}^{m\times m}$ such that

$$\begin{cases} H^\top H = I \\ H^\top = H \end{cases}$$

where $\mathbf{u} \in \mathbb{R}^m$ is a normal vector and $H\mathbf{a}$ is the reflection of $\mathbf{a} \in \mathbb{R}^m$ across the hyperplane that is normal to $\mathbf{u}$. Since $H_{\mathbf{u}}$ is invariant to $c\mathbf{u}$ for $c \neq 0$, we can restrict $\mathbf{u}$ to have unit 2-norm.

Suppose there is $H_1 = I - 2\mathbf{u}_1\mathbf{u}_1^\top$ such that $H_1\mathbf{a}_1 = \alpha\mathbf{e}_1$. From the relations $\|H_1\mathbf{a}_1\|_2 = \|\mathbf{a}_1\|_2$ and $\|\alpha\mathbf{e}_1\|_2 = |\alpha|\|\mathbf{e}_1\|_2 = |\alpha|$, we obtain $\alpha = \pm\|\mathbf{a}_1\|_2$. To avoid cancellation error in u_{11} below, we choose

$$\alpha = -\,\mathrm{sign}(a_{11})\|\mathbf{a}_1\|_2.$$

Since $\mathbf{a}_1 = \alpha H_1\mathbf{e}_1 = \alpha(\mathbf{e}_1 - 2\mathbf{u}_1\mathbf{u}_1^\top\mathbf{e}_1) = \alpha(\mathbf{e}_1 - 2\mathbf{u}_1 u_{11})$, we obtain

$$\begin{aligned} u_{11} &= \pm\sqrt{\frac{1}{2}\left(1 - \frac{a_{11}}{\alpha}\right)} = \pm\sqrt{\frac{1}{2\alpha}(\alpha - a_{11})} \\ u_{i1} &= -\frac{a_{i1}}{2\alpha u_{11}} \qquad\qquad i = 2, \ldots, m. \end{aligned}$$

Then, we can repeat this process for block matrix $A^{(2)}$,

$$A \quad\longrightarrow\quad \begin{bmatrix} * & * \\ 0 & A^{(2)} \end{bmatrix} \quad\longrightarrow\quad \cdots \quad\longrightarrow\quad R.$$

The corresponding orthonormal transform is given by

$$Q = H_1 \begin{bmatrix} I_1 & \\ & H_2 \end{bmatrix} \cdots \begin{bmatrix} I_{n-1} & \\ & H_n \end{bmatrix}$$

Also, $\mathbf{u}_k$ can be stored in the left lower-triangle of R.

6.2.4 Givens Algorithm

The Givens rotation is given by

$$\begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & \gamma & & & \sigma & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & -\sigma & & & & \gamma & \\ & & & & & & & & 1 \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{bmatrix} \begin{bmatrix} \vdots \\ a \\ \vdots \\ b \\ \vdots \end{bmatrix} = \begin{bmatrix} \vdots \\ r \\ \vdots \\ 0 \\ \vdots \end{bmatrix}$$

such that $r = \sqrt{a^2 + b^2}$.

We may only look at a 2-by-2 submatrix. We want

$$Q^\top A = R$$
$$\begin{bmatrix} \gamma & \sigma \\ -\sigma & \gamma \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix} = \begin{bmatrix} r_{11} \\ 0 \end{bmatrix}$$

where $Q^\top$ is a rotation matrix of the form $\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$ and $\gamma^2 + \sigma^2 = 1$. By solving the above equations, we obtain

$$\begin{cases} \gamma = \pm \frac{a_{11}}{\sqrt{a_{11}^2 + a_{21}^2}} \\ \sigma = \pm \frac{a_{21}}{\sqrt{a_{11}^2 + a_{21}^2}} \end{cases}$$

Since the product of two rotations is itself a rotation, we may use $O(n^2)$ rotations to perform Givens QR decomposition.

6.2.5 Variants

1. Full QR decomposition. $A = QR = Q \begin{bmatrix} R_1 \\ 0 \end{bmatrix}$, where $R \in \mathbb{C}^{n \times n}$ is upper-triangular.
2. Condensed QR decomposition. $A = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = Q_1 R_1$.
3. Rank-retaining QR decomposition. $A = Q \begin{bmatrix} R_1 & S \\ 0 & 0 \end{bmatrix} \Pi^\top$ where $\Pi \in \mathbb{C}^{n \times n}$ is a permutation matrix.
4. Complete orthogonal QR decomposition. $A = Q \begin{bmatrix} L & 0 \\ 0 & 0 \end{bmatrix} U^*$ coming from another QR decomposition $\begin{bmatrix} R_1^* \\ S^* \end{bmatrix} = Z \begin{bmatrix} L^* \\ 0 \end{bmatrix}$ and $U = \Pi Z$.

6.3 LU Factorization

6.3.1 Overview

The LU decomposition of $A \in \mathbb{C}^{m\times n}$ is given by $A = LU$ where $L \in \mathbb{R}^{m\times m}$ is lower-triangle and $U \in \mathbb{C}^{m\times n}$ is upper-triangle. If we use Gauss transformation to obtain LU decomposition, we can further requie that L is a unit lower-triangular matrix .

The LU decomposition is related to Gaussian elimination.

6.3.2 Plain LU Factorization

The existence of LU factorization (without pivoting) can be guaranteed by several conditions, one example is column diagonal dominance: if a nonsingular $A \in \mathbb{C}^{n\times n}$ satisfies $|a_{jj}| \geq \sum_{\substack{i=1\\ i\neq j}}^{n} |a_{ij}|$, $j = 1, \ldots, n$. There are necessary and sufficient conditions guaranteeing the existence of LU decomposition but those are difficult to check in practice.

Gauss transformation or *elimination matrices* is of the form $M = I - \mathbf{m}\mathbf{e}_i^\top \in \mathbb{C}^{m\times m}$ where $\mathbf{e}_i$ is the ith standard basis vector. The idea is to use a sequence of lower-triangular matrix $M_1, \ldots, M_n$ to transform A into an upper-triangular matrix U.

Suppose there is $M_1 = I - \mathbf{m}_1\mathbf{e}_1^\top$ such that $M_1\mathbf{a}_1 = \alpha\mathbf{e}_1$. If $a_{11} \neq 0$, we may set

$$\alpha = a_{11}, \qquad \mathbf{m}_1 = \begin{bmatrix} 0 \\ \frac{a_{21}}{a_{11}} \\ \vdots \\ \frac{a_{n1}}{a_{11}} \end{bmatrix}.$$

So we get

$$M_1 = \begin{bmatrix} 1 & & & \\ -\frac{a_{21}}{a_{11}} & 1 & & \\ \vdots & & \ddots & \\ -\frac{a_{n1}}{a_{11}} & & & 1 \end{bmatrix}.$$

Then, we can repeat this process for block matrix $A^{(2)}$,

$$A^{(1)} = A \quad \longrightarrow \quad A^{(2)} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a_{m2}^{(2)} & \cdots & a_{mn}^{(2)} \end{bmatrix} \quad \longrightarrow \quad \cdots \quad \longrightarrow \quad U.$$

The corresponding orthonormal transform is given by

$$L = M_1^{-1} \cdots M_n^{-1}$$

where the inverse matrix is easy to compute as

$$M_k^{-1} = I_m + \mathbf{m}_k \mathbf{e}_k^\top = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & \frac{a_{k+1,k}^{(k)}}{a_{kk}^{(k)}} & 1 & \\ & & \vdots & & \ddots \\ & & \frac{a_{m,k}^{(k)}}{a_{kk}^{(k)}} & & 1 \end{bmatrix}, \qquad \mathbf{m}_k = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \frac{a_{k+1,k}^{(k)}}{a_{kk}^{(k)}} \\ \vdots \\ \frac{a_{m,k}^{(k)}}{a_{kk}^{(k)}} \end{bmatrix}.$$

Notes: Although Gauss transformation is similar to Householder reflection in the sense that they are both rank-1 update to I and they zero out all the entries beneath the first in a vector $\mathbf{a}$, they have some differences, such as H is an unitary matrix, while M is a lower-triangular martix.

6.3.3 Gaussian Elimination with Partial Pivoting (GEPP)

For general nonsingular matrix A, i.e., $\text{rank}(A) = n$, some elements in each column must be nonzero. So we can use row permutation to deal with the case when $a_{kk}^{(k)} = 0$.

6.3.4 Gaussian Elimination with Complete Pivoting (GECP)

When $\text{rank}(A) < \min\{m, n\}$, the complete pivoting, which uses both row and column interchanges, is necessary.

6.3.5 LDU Decomposition

If $A \in \mathbb{R}^{n\times n}$ has nonsingular principal submatrices $A_{1:k,1:k}$ for $k = 1, \ldots, n$, then there exists a unit lower-triangular matrix $L \in \mathbb{R}^{n\times n}$, a unit upper-triangular matrix $U \in \mathbb{R}^{n\times n}$ and a diagonal matrix $D \in \mathbb{R}^{n\times n}$ such that $A = LDU$.

6.3.6 LDL^* Decomposition

For the LDU decomposition, if A is furthermore Hermitian, then $A = LDL^*$. It is also known as the square-root-free Cholesky factorization.

6.3.7 Cholesky factorizations

Both LDU decomposition and LDL^* decomposition are difficult to compute unless in the special case when A is symmetric positive definite. In such case, $A = LDL^*$ is related to the Cholesky decomposition $A = FF^*$ that $F = LD^{\frac{1}{2}}$.

In the field $\mathbb{R}$, the Cholesky decomposition is of the form $A = R^\top R = FF^\top$, where $R \in \mathbb{R}^{n\times n}$ is upper-triangular and $F \in \mathbb{R}^{n\times n}$ is lower-triangular. Both R and F have positive diagonal entries.

Writing

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{11} & \cdots & \cdots & a_{1n} \end{bmatrix} = \begin{bmatrix} f_{11} & & & \\ f_{21} & f_{22} & & \\ \vdots & \vdots & \ddots & \\ f_{n1} & f_{n2} & \cdots & f_{nn} \end{bmatrix} \begin{bmatrix} f_{11} & f_{21} & \cdots & f_{n1} \\ & f_{22} & \cdots & f_{n2} \\ & & \ddots & \vdots \\ & & & f_{nn} \end{bmatrix},$$

we have

$$\begin{aligned} a_{11} &= f_{11}^2 \\ a_{i1} &= f_{11} f_{i1} && i = 2, \ldots, n \\ &\vdots \\ a_{kk} &= \sum_{j=1}^{k} f_{kj}^2 \\ a_{ik} &= \sum_{j=1}^{k} f_{kk} f_{ij} && i = k+1, \ldots, n. \end{aligned}$$

Thus for $k = 1, \ldots, n$,

$$\begin{aligned} f_{kk} &= \sqrt{a_{kk} - \sum_{j=1}^{k-1} f_{kj}^2} \\ f_{ik} &= \frac{a_{ik} - \sum_{j=1}^{k-1} f_{kk} f_{ij}}{f_{kk}} && i = k+1, \ldots, n. \end{aligned}$$

We could use induction to show that the term in the square root is always positive.

Alternatively, write $A = FF^\top = \sum_{k=1}^{n} \mathbf{f}_k \mathbf{f}_k^\top$. By comparing the first column, we have

$$\mathbf{f}_1 = \frac{1}{\sqrt{a_{11}}} \mathbf{a}_1.$$

Let $A^{(1)} = A$ and $A^{(2)} = A^{(1)} - \mathbf{f}_1 \mathbf{f}_1^\top = \begin{bmatrix} 0 & 0 \\ 0 & A_2 \end{bmatrix}$. We can prove that $A^{(2)}$ is still positive definite, then

$$A^{(1)} \quad \longrightarrow \quad A^{(2)} \quad \longrightarrow \quad \cdots \quad \longrightarrow \quad 0.$$

and

$$\mathbf{f}_k = \frac{1}{\sqrt{a_{kk}^{(k)}}} \begin{bmatrix} \mathbf{0}_{n-k} \\ \mathbf{a}_k^{(k)} \end{bmatrix}$$

6.3.8 Variants

1. Plain LU decomposition. $A = LU$.
2. Gaussian Elimination with Partial Pivoting (GEPP). $A = \Pi_1^\top LU$.
3. Gaussian Elimination with Complete Pivoting (GECP). $A = \Pi_1^\top LU \Pi_2^\top$.
4. Condensed LU decomposition. For $r = \operatorname{rank}(A)$,

$$A = \Pi_1^\top LU \Pi_2^\top$$

$$
\begin{aligned}
&= \Pi_1^\top \begin{bmatrix} L_{11} & 0 \\ L_{21} & I_{m-r} \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \\ 0 & 0 \end{bmatrix} \Pi_2^\top \\
&= \Pi_1^\top \begin{bmatrix} L_{11} \\ L_{21} \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} \end{bmatrix} \Pi_2^\top \\
&= \Pi_1^\top \tilde{L} \tilde{U} \Pi_2^\top
\end{aligned}
$$

where $L_{11} \in \mathbb{R}^{r \times r}$ is unit lower-triangular and $U_{11} \in \mathbb{R}^{r \times r}$ is upper-triangular. L_{11} and U_{11} are both nonsingular.

5. Rank-retaining LU decomposition. $A = (\Pi_1^\top \tilde{L})(\tilde{U} \Pi_2^\top)$.
6. LDU decomposition. $A = LDU$.
7. $LDL^\top$ decomposition. $A = LDL^\top$.
8. Cholesky factorizations.

6.3.9 Uniqueness of the LU Factorization

The LU decomposition of a nonsingular matrix, if it exists (i.e., without row or column permutations), is unique.

If A has two LU decomposition, $A = L_1 U_1 = L_2 U_2$, then $L_2^{-1} L_1 = U_2 U_1^{-1}$. The left hand size is a unit lower-triangular matrix while the right hand side is a upper-triangular matrix, which means $L_2^{-1} L_1 = U_2 U_1^{-1} = I$, $L_1 = L_2$ and $U_1 = U_2$.

Similarly, the LDU decomposition and $LDL^\top$ decomposition are unique.

The Cholesky decomposition is also unique. If A has two Cholesky decomposition, $A = F_1 F_1^\top = F_2 F_2^\top$, then $F_2^{-\top} F_1 = F_2 F_1^{-\top}$. The left hand size is a lower-triangular matrix while the right hand side is a upper-triangular matrix, which means $F_2^{-\top} F_1 = F_2 F_1^{-\top} = D$, for some diagonal matrix D with diagonal entries $\frac{f_{1,kk}}{f_{2,kk}} = \frac{f_{2,kk}}{f_{1,kk}}$. Since $D = D^\top = F_1^\top F_2^{-1}$, we have $D^2 = F_2 F_1^{-\top} F_1^\top F_2^{-1} = I$, which implies that the diagonal entries of D are in $\{\pm 1\}$. Since we require the diagonal entries of F to be positive, we have $D = I$ and $F_1 = F_2$.

6.4 Block Factorization

6.4.1 Overview

Block LU factorization can be viewed as the generalization of LU decomposition to the block matrix. This works for rectangular matrices too but we keep our discussion to square matrices $A \in \mathbb{R}^{n\times n}$ for simplicity. Consider a 2-by-2 block martix $A \in \mathbb{R}^{2\times 2}$,

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

where $A_{11} \in \mathbb{R}^{p\times p}$, $A_{12} \in \mathbb{R}^{p\times q}$, $A_{21} \in \mathbb{R}^{q\times p}$ and $A_{22} \in \mathbb{R}^{q\times q}$ $(p+q=n)$.

If A_{11} is nonsingular, let $M_1 = I - U_1 V_1^\top \in \mathbb{R}^{n\times n}$ where $U_1, V_1 \in \mathbb{R}^{n\times p}$. We want

$$M_1 A = \begin{bmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{bmatrix}.$$

It turns out that $U = \begin{bmatrix} \mathbf{0} \\ A_{21}A_{11}^{-1} \end{bmatrix}$ and $V = \begin{bmatrix} I_p \\ \mathbf{0} \end{bmatrix}$ satisfy the above condition. So

$$M_1 A = \begin{bmatrix} A_{11} & A_{12} \\ \mathbf{0} & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix}.$$

We may write

$$\begin{aligned} A = LU &= \begin{bmatrix} A_{11} & A_{12} \\ \mathbf{0} & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix} \begin{bmatrix} I_p & \mathbf{0} \\ A_{21}A_{11}^{-1} & I_q \end{bmatrix} \\ &= \begin{bmatrix} A_{11} & \mathbf{0} \\ A_{21} & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix} \begin{bmatrix} I_p & A_{11}^{-1}A_{12} \\ \mathbf{0} & I_q \end{bmatrix}. \end{aligned}$$

However, L_{11}, L_{22}, U_{11} and U_{22} are arbitrary matrices and need not to be triangular matrix. The term $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$ is called *Schur complement*.

Similar results apply to LDU, $LDL^\top$, and Cholesky factorizations.

If A is symmetric positive definite, then is A_{11}. The block Cholesky decomposition is given by

$$A = \begin{bmatrix} R_{11}^\top & \mathbf{0} \\ R_{12}^\top & R_{22}^\top \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ \mathbf{0} & R_{22} \end{bmatrix} = \begin{bmatrix} R_{11}^\top R_{11} & R_{11}^\top R_{12} \\ R_{12}^\top R_{11} & R_{12}^\top R_{12} + R_{22}^\top R_{22} \end{bmatrix}$$

then $S = A_{22} - A_{21}A_{11}^{-1}A_{12} = R_{22}^\top R_{22}$.

6.4.2 Solving Linear Systems via Block Factorization

To solve the linear equations

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix},$$

whihc equals to

$$\begin{bmatrix} A_{11} & A_{12} \\ \mathbf{0} & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} I_p & \mathbf{0} \\ -A_{21}A_{11}^{-1} & I_q \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}.$$

i.e.,

$$\begin{cases} \begin{bmatrix} A_{11} & A_{12} \\ \mathbf{0} & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} \\ \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} I_p & \mathbf{0} \\ -A_{21}A_{11}^{-1} & I_q \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \end{cases}$$

6.4.3 Determinants and Inverses

From

$$A = \begin{bmatrix} A_{11} & A_{12} \\ \mathbf{0} & A_{22}-A_{21}A_{11}^{-1}A_{12} \end{bmatrix} \begin{bmatrix} I_p & \mathbf{0} \\ A_{21}A_{11}^{-1} & I_q \end{bmatrix},$$

if A_{11} is invertible, we have

$$\det(A) = \det(A_{11})\det(S),$$

which gives a sufficient and necessary condition for A to be nonsingular. That is A is invertible if and only if A_{11} and S are invertible.

If A is invertible, to obtain A^{-1}, we consider

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix},$$

and try to express in

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}.$$

From the first equations set, we have

$$\begin{bmatrix} A_{11} & A_{12} \\ \mathbf{0} & A_{22}-A_{21}A_{11}^{-1}A_{12} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} I_p & \mathbf{0} \\ -A_{21}A_{11}^{-1} & I_q \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}.$$

and

$$\begin{cases} \mathbf{x}_2 = S^{-1}(\mathbf{b}_2 - A_{21}A_{11}^{-1}\mathbf{b}_1) = S^{-1}\mathbf{b}_2 - S^{-1}A_{21}A_{11}^{-1}\mathbf{b}_1 \\ \mathbf{x}_1 = A_{11}^{-1}(\mathbf{b}_1 - A_{12}\mathbf{x}_2) = (A_{11}^{-1} + A_{11}^{-1}A_{12}S^{-1}A_{21}A_{11}^{-1})\mathbf{b}_1 - A_{11}^{-1}A_{12}S^{-1}\mathbf{b}_2 \end{cases}$$

i.e.,

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} A_{11}^{-1}+A_{11}^{-1}A_{12}S^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}S^{-1} \\ -S^{-1}A_{21}A_{11}^{-1} & S^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}.$$

Therefore, the inverse of A is given by

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} A_{11}^{-1}+A_{11}^{-1}A_{12}S^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}S^{-1} \\ -S^{-1}A_{21}A_{11}^{-1} & S^{-1} \end{bmatrix}$$

6.5 Rank-r Perturbed Problems

6.5.1 Rank-1 Update

Suppose that we have solved the problem $A\mathbf{x} = \mathbf{b}$ where $A \in \mathbb{R}^{n\times n}$ and we wish to solve the perturbed problem $(A + \mathbf{u}\mathbf{v}^\top)\mathbf{y} = \mathbf{b}$.

Sherman–Morrison formula. If A is invertible, then $A + \mathbf{u}\mathbf{v}^\top$ is invertible if and only if $1 + \mathbf{v}^\top A^{-1}\mathbf{u} \neq 0$, and

$$(A + \mathbf{u}\mathbf{v}^\top)^{-1} = A^{-1} - \frac{1}{1 + \mathbf{v}^\top A^{-1}\mathbf{u}} A^{-1}\mathbf{u}\mathbf{v}^\top A^{-1}.$$

Then we can solve for the perturbed problem by

1. solve $A\mathbf{x} = \mathbf{b}$;
2. solve $A\mathbf{w} = \mathbf{u}$ ($\mathbf{w} = A^{-1}\mathbf{u}$);
3. compute $\sigma = -\frac{1}{1+\mathbf{v}^\top\mathbf{w}}$;
4. compute $\mathbf{y} = \mathbf{x} + \sigma(\mathbf{v}^\top\mathbf{x})\mathbf{w}$.

6.5.2 Rank-r Update

We wish to solve the perturbed problem $(A + UV^\top)\mathbf{y} = \mathbf{b}$, where $U = [\,\mathbf{u}_1 \cdots \mathbf{u}_r\,]$, $V = [\,\mathbf{v}_1 \cdots \mathbf{v}_r\,] \in \mathbb{R}^{n\times r}$.

Sherman–Woodbury–Morrison formula. If A is invertible, then $A + UV^\top$ is invertible if and only if $I + V^\top A^{-1}U$ is invertible, and

$$(A + UV^\top)^{-1} = A^{-1} - A^{-1}U(I + V^\top A^{-1}U)^{-1}V^\top A^{-1}.$$

Chapter 7

Least Squares Problems

7.1 General Linear System Problems

7.1.1 Overview

For $A \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$, we want to find the solution $\mathbf{x} \in \mathbb{R}^n$ of the linear system $A\mathbf{x} = \mathbf{b}$. Usually, due to measure error, such solution may not exist, so we instead seek for an approximate solution by using least squares. In this chapter, we will restrict our discussion to real field in most cases.

7.1.2 Existence

The linear system is consistent if $\mathbf{b} \in \text{im}(A)$, which means that the solution must exist.

The linear system is inconsistent if the solution does not exist, then we can seek for least squares solution of the optimization problem $\min_{\mathbf{x} \in \mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\|_2$.

7.1.3 Uniqueness

If $\text{rank}(A) = n$, then the solution to either consistent or inconsistent linear system is unique.

However, if $\text{rank}(A) < n$, then the solution is not unique and we want the unique minimum length solution instead. If the linear system is consistent, then we seek for solution to the optimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} \{\|\mathbf{x}\|_2 : A\mathbf{x} = \mathbf{b}\}.$$

If the linear system is inconsistent, then we seek for solution to the optimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left\{ \|\mathbf{x}\|_2 : \mathbf{x} \in \arg\min_{\mathbf{y}} \|A\mathbf{y} - \mathbf{b}\|_2 \right\}.$$

7.2 Full Rank Least Squares

7.2.1 Problem Description

For $A \in \mathbb{C}^{m\times n}$, when $\text{rank}(A) = n$, the full rank least squares is seeking for the solution to the optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\|_2.$$

This problem always has a unique solution. To see this, we can decomposition $\mathbf{b}$ as $\mathbf{b} = \mathbf{b}_0 + \mathbf{b}_1$ where $\mathbf{b}_0 \in \ker(A^*)$ and $\mathbf{b}_1 \in \text{im}(A)$. Then $\|A\mathbf{x} - \mathbf{b}\|_2^2 = \|A\mathbf{x} - \mathbf{b}_1\|_2^2 + \|\mathbf{b}_0\|_2^2$. To solve the original problem, we simply need to solve $\min \|A\mathbf{x} - \mathbf{b}_1\|_2$. While this problem has unique solution since $\text{rank}(A) = n$ and $\mathbf{b}_1 \in \text{im}(A)$. Next, we consider three methods which give the same solution mathematically but have different numerical properties.

7.2.2 Solution by Normal Equations

Define

$$\phi(x) = \frac{1}{2}\|A\mathbf{x} - \mathbf{b}\|_2^2,$$

and set the derivative with respect to $\mathbf{x}$ to zero, we get

$$A^*A\mathbf{x} = A^*\mathbf{b}.$$

By solving this linear system, we can get the solution of the original problem.

When $n \ll m$, $A^*A \in \mathbb{C}^{n\times n}$ is a much smaller system and requires less arithmetic. However, if n get larger, it will become inefficient.

7.2.3 Solution by Augmented Systems

Let $\mathbf{r} = \mathbf{b} - A\mathbf{x}$ be the residual. The normal equations can be expressed as

$$\begin{cases} \mathbf{r} + A\mathbf{x} = \mathbf{b} \\ A^*\mathbf{r} = \mathbf{0} \end{cases}$$

i.e.,

$$\begin{bmatrix} I & A \\ A^* & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{r} \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix}$$

This is often a large system since the coefficient matrix has dimension $(m+n)\times(m+n)$, but it preserves the sparsity of A.

7.2.4 Solution by QR Factorization

Let $A = Q \begin{bmatrix} R \\ 0 \end{bmatrix}$ be the QR decomposition of A. Since the 2-norm is invariant under orthogonal transformations, we have

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{C}^n} \|\mathbf{b} - A\mathbf{x}\|_2 &= \min_{\mathbf{x}\in\mathbb{C}^n} \|Q^*\mathbf{b} - Q^*A\mathbf{x}\|_2 \\ &= \min_{\mathbf{x}\in\mathbb{C}^n} \left\| Q^*\mathbf{b} - \begin{bmatrix} R\mathbf{x} \\ 0 \end{bmatrix} \right\|_2 . \end{aligned}$$

Let $Q^*\mathbf{b} = \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix}$, then

$$\min_{\mathbf{x}\in\mathbb{C}^n} \|\mathbf{b} - A\mathbf{x}\|_2^2 = \min_{\mathbf{x}\in\mathbb{C}^n} \|\mathbf{c} - R\mathbf{x}\|_2^2 + \|\mathbf{d}\|_2^2.$$

Therefore the minimum is achieved by the vector $\mathbf{x}$ such that $R\mathbf{x} = \mathbf{c}$ and $\min_{\mathbf{x}\in\mathbb{C}^n} \|\mathbf{b} - A\mathbf{x}\|_2 = \|\mathbf{d}\|_2$. Since $\operatorname{rank}(A) = n$, R is invertible, so the solution is given by $\mathbf{x} = R^{-1}\mathbf{c}$. However, in practice, we would really calculate the inverse. Instead, we will use back substitution to solve this linear system.

7.3 Minimum Norm Least Squares

7.3.1 Problem Description

We want to solve the optimization problem

$$\min_{\mathbf{x}}\{\|\mathbf{x}\|_2 : \mathbf{x} \in \arg\min_{\mathbf{y}\in\mathbb{C}} \|A\mathbf{y} - \mathbf{b}\|_2\}.$$

We will see that such solution exists and is unique. This problem may be consistent or inconsistent. We will view it as inconsistent since if we know how to solve the inconsistent case, then it is trivial for the consistent case.

7.3.2 Solution by SVD

Let $A = U\Sigma V^*$ be the SVD of A where $\Sigma \in \mathbb{C}^{m\times n}$ is a diagonal matrix with diagonal entries $\sigma_1 \geq \cdots \sigma_r > 0 = \cdots = 0$ and $r = \text{rank}(A)$. Then,

$$\begin{aligned}
\|\mathbf{b} - A\mathbf{x}\|_2^2 &= \|\mathbf{b} - U\Sigma V^*\mathbf{x}\|_2^2 \\
&= \|U^*\mathbf{b} - \Sigma V^*\mathbf{x}\|_2^2 \\
&\overset{\substack{\mathbf{c}=U^*\mathbf{b}\\ \mathbf{y}=V^*\mathbf{x}}}{=\!=\!=} \|\mathbf{c} - \Sigma\mathbf{y}\|_2^2 \\
&= \sum_{i=1}^{r} |c_i - \sigma_i y_i|^2 + \sum_{j=r+1}^{m} c_j^2 \\
&\geq \sum_{j=r+1}^{m} c_j^2
\end{aligned}$$

the equality holds when $y_i = \frac{c_i}{\sigma_i}$ for $i = 1, \ldots, r$. So the solution to $\min_{\mathbf{y}} \|\mathbf{c} - \Sigma\mathbf{y}\|_2$ is given by $\mathbf{y} = \begin{bmatrix} \frac{c_1}{\sigma_1} \\ \vdots \\ \frac{c_r}{\sigma_r} \\ y_{r+1} \\ \vdots \\ y_n \end{bmatrix}$ where $y_{r+1}, \ldots, y_n$ can be any values. Since we seek for the minimum norm solution $\min \|\mathbf{x}\|_2^2 = \min \|V\mathbf{y}\|_2^2$, we must have $y_{r+1} = \ldots = y_n = 0$ and therefore the minimum norm solution to $\min_{\mathbf{y}} \|\mathbf{c} - \Sigma\mathbf{y}\|_2$ is $y_* = \begin{bmatrix} \frac{c_1}{\sigma_1} \\ \vdots \\ \frac{c_r}{\sigma_r} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$. So $\mathbf{x}_* = V\mathbf{y}_* = V\Sigma^\dagger U^*\mathbf{b}$ is the unique solution to the original problem.

7.3.3 Solution by Rank-Retaining Factorization

Similarly, decompose $\mathbf{b}$ as $\mathbf{b} = \mathbf{b}_0 + \mathbf{b}_1$ where $\mathbf{b}_0 \in \ker(A^*)$ and $\mathbf{b}_1 \in \text{im}(A)$. Then

$$\|\mathbf{b} - A\mathbf{x}\|_2^2 = \|\mathbf{b}_0 + \mathbf{b}_1 - A\mathbf{x}\|_2^2 = \|\mathbf{b}_0\|_2^2 + \|\mathbf{b}_1 - A\mathbf{x}\|_2^2 \geq \|\mathbf{b}_0\|_2^2.$$

We just need to minimize $\|\mathbf{b}_1 - A\mathbf{x}\|_2^2$.

Since $\mathbb{C}^n = \text{im}(A^*) \oplus \ker(A)$, $\mathbf{x} \in \mathbb{C}^n$ can be written uniquely as $\mathbf{x} = \mathbf{x}_0 + \mathbf{x}_1$ where $\mathbf{x}_0 \in \ker(A)$ and $\mathbf{x}_1 \in \text{im}(A^*)$. Then $\mathbf{b}_1 = A\mathbf{x} = A\mathbf{x}_1$ and $\|\mathbf{x}\|_2^2 = \|\mathbf{x}_0\|_2^2 + \|\mathbf{x}_1\|_2^2 \geq \|\mathbf{x}_1\|_2^2$. We just need to find $\mathbf{x}_1$ and set $\mathbf{x}_0 = \mathbf{0}$ so that the solution has minimum length.

Let $A = GH$ be the rank-retaining decomposition of A. Since $\mathbf{x}_1 \in \text{im}(A^*) = \text{im}(H^*)$, for some $\mathbf{v} \in \mathbb{C}^r$, $\mathbf{x}_1 = H^*\mathbf{v}$. Since $\mathbf{b}_1 \in \text{im}(A) = \text{im}(G)$, $\mathbf{b} = G\mathbf{s}$ for some $\mathbf{s} \in \mathbb{C}^r$. Therefore, $GHH^* = A\mathbf{x}_1 = \mathbf{b}_1 = \mathbf{v} = G\mathbf{s}$.

Since $G^*\mathbf{b} = G^*G\mathbf{s}$ and G^*G is nonsingular, we have

$$\mathbf{s} = (G^*G)^{-1}G^*\mathbf{b}.$$

So $(G^*G)HH^*\mathbf{v} = (G^*G)\mathbf{s}$. Since G^*G is nonsingular, $HH^*\mathbf{v} = \mathbf{s}$. Since HH^* is nonsingular,

$$\mathbf{v} = (HH^*)^{-1}\mathbf{s}.$$

Then

$$\mathbf{x}_1 = H^*(HH^*)^{-1}\mathbf{s} = \underbrace{H^*(HH^*)^{-1}(G^*G)^{-1}G^*}_{A^\dagger}\mathbf{b}.$$

7.4 Least Squares with Linear Constraints

7.4.1 Problem Description

The least squares problem with linear constraints is of the form

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & \|\mathbf{b} - A\mathbf{x}\|_2 \\ \text{subject to} \quad & C^\top \mathbf{x} = \mathbf{d} \end{aligned}$$

for $A \in \mathbb{R}^{m\times n}$, $C \in \mathbb{R}^{n\times p}$, $\mathbf{b} \in \mathbb{R}^m$ and $\mathbf{d} \in \mathbb{R}^p$. Also, we assume that $\text{rank}(A) = n$.

We will describe three methods, mathematically equivalent but with different numerical properties.

7.4.2 Solution by Using Lagrange Multipliers

Define the *Lagrangian*

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \|\mathbf{b} - A\mathbf{x}\|_2^2 + 2\boldsymbol{\lambda}^\top (C^\top \mathbf{x} - \mathbf{d}),$$

where $\boldsymbol{\lambda} \in \mathbb{R}^n$ is the vector of *Lagrange multipliers*.

Setting derivative with respect to $\mathbf{x}$ and $\boldsymbol{\lambda}$ to zero yields the *KKT conditions*

$$\begin{cases} \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = 2(A^\top A\mathbf{x} - A^\top \mathbf{b} + C\boldsymbol{\lambda}) = 0 \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = C^\top \mathbf{x} - \mathbf{d} = 0 \end{cases}$$

which gives the system

$$\begin{bmatrix} A^\top A & C \\ C^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} A^\top \mathbf{b} \\ \mathbf{d} \end{bmatrix}.$$

By solving this system, we can find the solution to the original problem. However, the term $A^\top A$ induces extra errors just like the one of the normal equation.

7.4.3 Solution by QR Factorization of A

Since $\text{rank}(A) = n$, we can first solve the unconstrained least squares problem

$$\hat{\mathbf{x}} = (A^\top A)^{-1} A^\top \mathbf{b} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\|_2.$$

From $A^\top A\mathbf{x} - A^\top \mathbf{b} + C\boldsymbol{\lambda} = 0$, we have

$$\mathbf{x} = \hat{\mathbf{x}} - (A^\top A)^{-1} C\boldsymbol{\lambda}.$$

Substitue it into $C^\top \mathbf{x} = \mathbf{d}$, we have

$$C^\top \hat{\mathbf{x}} - C^\top (A^\top A)^{-1} C\boldsymbol{\lambda} = \mathbf{d},$$

which we can then solve for $\boldsymbol{\lambda}$. Finally, $\mathbf{x}$ can be solved.

To compute $(A^\top A)^{-1}$ we can use QR decomposition of $A = Q \begin{bmatrix} R \\ \mathbf{0} \end{bmatrix}$. Then $(A^\top A)^{-1} = R^{-1}R^{-\top}$. To compute $C^\top R^{-1}R^{-\top}C$, we again use QR decomposition of $R^{-\top}C = Q_1 R_1$ and then $C^\top (A^\top A)^{-1} C = R_1^{-1} R_1^{-\top}$.

1. Compute full-rank QR decomposition of A, $A = Q \begin{bmatrix} R \\ \mathbf{0} \end{bmatrix}$ where $R \in \mathbb{R}^{n\times n}$ is nonsingular;
2. Solve the unconstrained least squares problem $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\|_2$;
3. Form $W = R^{-\top}C$ with p back substitutions $R^\top \mathbf{w}_i = \mathbf{c}_i$, $i = 1, \ldots, p$;
4. Compute QR factorization of W, $W = Q_1 R_1$;
5. Set $\boldsymbol{\eta} = C^\top \hat{\mathbf{x}} - \mathbf{d}$;
6. Solve $R_1^\top R_1 \boldsymbol{\lambda} = \boldsymbol{\eta}$ for $\boldsymbol{\lambda}$ with 2 back substitutions;
7. Set $\mathbf{x} = \hat{\mathbf{x}} - (R^\top R)^{-1} C\boldsymbol{\lambda}$ where term $(R^\top R)^{-1} C\boldsymbol{\lambda}$ is computed with 2 back substitutions.

This method, however, still has more unknowns than the original problem.

7.4.4 Solution by QR Factorization of C

Suppose $p \le n$, the QR decomposition of C is given by $C = Q_2 \begin{bmatrix} R_2 \\ \mathbf{0} \end{bmatrix}$ where $R_2 \in \mathbb{R}^{p\times p}$ is upper-triangular. Then the constraint becomes

$$\begin{bmatrix} R_2^\top & \mathbf{0} \end{bmatrix} Q_2^\top \mathbf{x} = \mathbf{d}.$$

If we set $Q_2^\top \mathbf{x} = \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}$ where $\mathbf{u} \in \mathbb{R}^p$ and $\mathbf{v} \in \mathbb{R}^{n-p}$, then $R_2^\top \mathbf{u} = \mathbf{d}$. Then

$$\begin{aligned} \|\mathbf{b} - A\mathbf{x}\|_2 &= \|\mathbf{b} - AQ_2 Q_2^\top \mathbf{x}\|_2 \\ &= \|\mathbf{b} - AQ_2 \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}\|_2 \\ &= \|\mathbf{b} - A_1\mathbf{u} - A_2\mathbf{v}\|_2 \end{aligned}$$

where $\begin{bmatrix} A_1 & A_2 \end{bmatrix} = AQ_2$.

1. Compute the QR decomposition of C;
2. Compute $\begin{bmatrix} A_1 & A_2 \end{bmatrix} = AQ$;
3. Solve $R_2^\top \mathbf{u} = \mathbf{d}$ to obtain a solution $\mathbf{u}_*$ (may be more than one).
4. Solve least squares problem $\mathbf{v}_* = \arg\min \|(\mathbf{b} - A_1\mathbf{u}) - A_2\mathbf{v}\|_2$;
5. Compute $\mathbf{x} = Q_2 \begin{bmatrix} \mathbf{u}_* \\ \mathbf{v}_* \end{bmatrix}$.

This method has fewer unknowns but may loss sparsity or otehr structure in A.

7.5 Least Squares with Quadratic Constraints

7.5.1 Problems Description

The least squares problem with quadratic constraints is of the form

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & \|\mathbf{b} - A\mathbf{x}\|_2 \\ \text{subject to} \quad & \|\mathbf{x}\|_2 \le \alpha \end{aligned}$$

for $A \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$ and $\alpha > 0$. Also, we assume that $\text{rank}(A) = n$.

7.5.2 Solution by SVD

If $\alpha \ge \|A^\dagger \mathbf{b}\|_2$, then $A^\dagger \mathbf{b}$ is the solution.

If $\alpha < \|A^\dagger \mathbf{b}\|_2$, then the original problem is equivalent to $\min \|A\mathbf{x} - \mathbf{b}\|_2$ s.t. $\|\mathbf{x}\|_2 = \alpha$. Let $L(\mathbf{x}, \mu) = \|A\mathbf{x} - \mathbf{b}\|_2^2 + \mu(\|\mathbf{x}\|_2^2 - \alpha^2)$. Setting the derivative to zeros yields

$$\begin{aligned} \nabla_{\mathbf{x}} L(\mathbf{x}, \mu) &= 2A^\top A - 2A^\top \mathbf{b} + 2\mu\mathbf{x} = 0 \\ \nabla_{\mu} L(\mathbf{x}, \mu) &= \|\mathbf{x}\|_2^2 - \alpha^2 = 0. \end{aligned}$$

Then $(A^\top A + \mu I_n)\mathbf{x} = A^\top \mathbf{b}$ and $\mathbf{x} = (A^\top A + \mu I_n)^{-1} A^\top \mathbf{b}$. Let $A = U\Sigma V^\top$ be the SVD of A, then

$$\begin{aligned} \alpha^2 = \mathbf{x}^\top \mathbf{x} &= \mathbf{b}^\top A (A^\top A + \mu I_n)^{-\top} (A^\top A + \mu I_n)^{-1} A^\top \mathbf{b} \\ &= \mathbf{b}^\top U\Sigma V^\top (V\Sigma^\top \Sigma V^\top + \mu I_n)^{-1} V\Sigma^\top U\mathbf{b} \\ &\overset{\mathbf{c}=U\mathbf{b}}{=\!=\!=} \mathbf{c}^\top \Sigma (\Sigma^\top \Sigma + \mu I_n)^{-2} \Sigma^\top \mathbf{c} \\ &= \sum_{i=1}^{n} \frac{c_i^2 \sigma_i^2}{(\sigma_i^2 + \mu)^2} \\ &\triangleq f(\mu). \end{aligned}$$

$f(\mu) = \alpha^2$ has no closed form solution for μ, but we can use Newton-Rhapson method to solve $f(\mu) = \alpha^2$ for μ. After that, $\mathbf{x} = (A^\top A + \mu I_n)^{-1} A^\top \mathbf{b}$ can be obtained.

7.6 Total Least Squares

7.6.1 Problem Description

The total least squares problem is given by

$$\begin{aligned} \min_{\mathbf{x},E,\mathbf{r}} \quad & \| \begin{bmatrix} E & \mathbf{r} \end{bmatrix} \|_F \\ \text{s.t.} \quad & (A+E)\mathbf{x} = \mathbf{b} + \mathbf{r} \end{aligned}$$

for $A, E \in \mathbb{C}^{m\times n}$, $\mathbf{b}, \mathbf{r} \in \mathbb{C}^m$, $\mathbf{x} \in \mathbb{R}^n$ and $\operatorname{rank}(A) = n$.

7.6.2 Solution by SVD

If $\mathbf{b} \in \operatorname{im}(A)$, then there exists $\mathbf{x} \in \mathbb{C}^n$ such that $A\mathbf{x} = \mathbf{b}$, which implies $E = \mathbf{0}$ and $\mathbf{r} = \mathbf{0}$.

If $\mathbf{b} \notin \operatorname{im}(A)$, rewrite the equation as

$$\left(\begin{bmatrix} A & \mathbf{b} \end{bmatrix} + \begin{bmatrix} E & \mathbf{r} \end{bmatrix}\right) \begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix} = \mathbf{0}.$$

Let $C = \begin{bmatrix} A & \mathbf{b} \end{bmatrix}$, $F = \begin{bmatrix} E & \mathbf{r} \end{bmatrix}$ and $\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ -1 \end{bmatrix}$. Since $\mathbf{b} \notin \operatorname{im}(A)$, $\operatorname{rank}(C) = n+1$. Let $C = U\Sigma V^*$ be SVD of C with sigular values $\sigma_1 \geq \cdots \sigma_{n+1} > 0$. Since $\mathbf{z} \neq \mathbf{0}$ (its last entry is -1), $\operatorname{rank}(C+F) \leq n$. From Eckart-Young theorem, the best (minimum distance in F-norm) rank-n approximate of C is $U \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ & & 0 \end{bmatrix} V^*$. So when $F = U \begin{bmatrix} 0 & & \\ & \ddots & \\ & & 0 \\ & & -\sigma_{n+1} \end{bmatrix} V^*$, $\|F\|_F^2 = \sigma_{n+1}^2$ achieves the minimum F-norm. Let $\mathbf{v}_{n+1}$ be the $(n+1)$-th right singular vector of C, then $(C+F)\mathbf{v}_{n+1} = \mathbf{0}$. Thus if $v_{n+1,n+1} \neq 0$, then a solution is given by $\mathbf{z} = -\frac{1}{v_{n+1,n+1}}\mathbf{v}_{n+1}$. Such $\mathbf{z}$ is not unique.

7.7 Other Optimization Problems about Matrices

$$\begin{aligned} \min_{X \in \mathbb{C}^{m \times n}} \quad & \|A - X\|_F \\ \text{s.t.} \quad & X^* X = I_n \end{aligned}$$

$$\begin{aligned} \min_{X \in \mathbb{C}^{m \times n}} \quad & \|A - BX\|_F = \|B^* A - X\|_F \\ \text{s.t.} \quad & X^* X = I_n \end{aligned}$$

$$\begin{aligned} \min_{X \in \mathbb{C}^{m \times n}} \quad & \|A - X\|_F \\ \text{s.t.} \quad & X^* = X \end{aligned}$$

Chapter 8

Iteration Methods

8.1 Splitting Methods

8.1.1 Overview

To solve $A\mathbf{x} = \mathbf{b}$ where $A \in \mathbb{R}^{n\times n}$, decompose A into the sum of two matrices $A = M - N$ such that M can be easily inverted and then do

$$M\mathbf{x}^{(k)} = N\mathbf{x}^{(k-1)} + \mathbf{b}.$$

For the solution $\mathbf{x}$, we also have

$$M\mathbf{x} = N\mathbf{x} + \mathbf{b}.$$

Subtract the above two equations and consider the error $\mathbf{e}^{(k)} = \mathbf{x} - \mathbf{x}^{(k)}$, we have

$$\mathbf{e}^{(k)} = B\mathbf{e}^{(k-1)}$$

where $B = M^{-1}N$ is called the iteration matrix.

Theorem. Convergence of Splitting Methods

$\mathbf{e}^{(k)} \to \mathbf{0}$ as $k \to \infty$ for all $\mathbf{e}^{(0)}$ if and only if $\rho(B) < 1$.

Proof. Note that $\mathbf{e}^{(k)} = B\mathbf{e}^{(k-1)} \to \mathbf{0}$ for all $\mathbf{e}^{(0)}$ if and only if $\lim_{k\to\infty} \mathbf{e}^{(k)} = \mathbf{0}$, since we can choose $\mathbf{e}^{(0)}$ to be each of the standard basis vector $\mathbf{e}_1, \ldots, \mathbf{e}_n$ in turn and so we get

$$B^k = B^k \begin{bmatrix} \mathbf{e}_1 & \cdots & \mathbf{e}_n \end{bmatrix} = \begin{bmatrix} B^k\mathbf{e}_1 & \cdots & B^k\mathbf{e}_n \end{bmatrix} \to \mathbf{0}$$

as $k \to \infty$. Let $B = XJ^\top X^{-1}$ be the Jordan decomposition of B. For a Jordan block,

$$J_r^k = \begin{bmatrix} \lambda_r^k & \binom{k}{1}\lambda_r^k & \cdots & \binom{k}{n_r-1}\lambda_r^{k-(n_r-1)} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \binom{k}{1}\lambda_r^k \\ & & & \lambda_r^k \end{bmatrix} \to \mathbf{0},$$

as $k \to \infty$ if and only if $\lim_{k\to\infty} \lambda_r^k = 0$. Sicne $\lim_{k\to\infty} B^k = \mathbf{0}$ if and only if $\lim_{k\to\infty} J_r^k = \mathbf{0}$ for all r, if and onlt if $\lim_{k\to\infty} \lambda_r^k = 0$ for all r, if and only if $\rho(B) < 1$. ■

And $\rho(B)$ is the average rate of convergence.

8.1.2 Jacobi Method

Jacobi Iteration

If we write

$$L = -\begin{bmatrix} 0 & & & \\ a_{21} & \ddots & & \\ \vdots & \ddots & \ddots & \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{bmatrix} \qquad D = \begin{bmatrix} a_{11} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & a_{nn} \end{bmatrix}, \qquad N = -\begin{bmatrix} 0 & a_{11} & \cdots & a_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & a_{n-1,n} \\ & & & 0 \end{bmatrix}$$

Let

$$M = D = \begin{bmatrix} a_{11} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & a_{nn} \end{bmatrix}, \qquad N = L + U = -\begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ a_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{n-1,n} \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{bmatrix}$$

then the iteration is given by

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j\neq i} a_{ij} x_j^{(k)}\right),$$

equivalently,

$$D\mathbf{x}^{(k+1)} = -(L+U)\mathbf{x}^{(k)} + \mathbf{b}.$$

Convergence Analysis

The iteration matrix is given by

$$B_J = M^{-1}N = -D^{-1}(L+U) = I - D^{-1}A = -\begin{bmatrix} 0 & \frac{a_{12}}{a_{11}} & \cdots & \frac{a_{1n}}{a_{11}} \\ \frac{a_{21}}{a_{22}} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{a_{n-1,n}}{a_{n-1,n-1}} \\ \frac{a_{n1}}{a_{nn}} & \cdots & \frac{a_{n,n-1}}{a_{nn}} & 0 \end{bmatrix}.$$

Theorem. Convergence of Jacobi Iterations

If A is strictly diagonally dominant, i.e., $\|B_J\|_\infty < 1$, then $\rho(B_J) \leq \|B_J\|_\infty < 1$ and Jacobi iterations converge for any $\mathbf{x}^{(0)}$.

8.1.3 Gauss–Seidel Method

Gauss-Seidel Iteration

In Jacobi iteration, we compute $x_i^{(k+1)}$ using the elements of $\mathbf{x}^{(k)}$. However, if we use the latest information available - $x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}$ - we can have Gauss-Seidel iteration,

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right),$$

equivalently,

$$\begin{aligned} D\mathbf{x}^{(k+1)} &= -L\mathbf{x}^{(k+1)} - U\mathbf{x}^{(k)} + \mathbf{b} \\ (L+D)\mathbf{x}^{(k+1)} &= -U\mathbf{x}^{(k)} + \mathbf{b}, \end{aligned}$$

Convergence Analysis

The iteration matrix is given by

$$B_{GS} = M^{-1}N = -(L+D)^{-1}U.$$

Theorem. Convergence of Gaus-Seidel Iterations

If A is strictly diagonally dominant, then $\rho(B_{GS}) \leq \|B_{GS}\|_\infty < 1$ and Gauss-Seidel iterations converge for any $\mathbf{x}^{(0)}$.

Proof. Let $r_i = \sum_{j \neq i}^{n} \left| \frac{a_{ij}}{a_{ii}} \right|$, if A is strictly diagonally dominant, we have $r = \max\limits_{1 \leq i \leq n} r_i < 1$. For the errors,

$$(L+D)\mathbf{e}^{(k+1)} = -U\mathbf{e}^{(k)},$$

i.e.,

$$a_{ii} e_i^{(k+1)} = -\sum_{j=1}^{i-1} a_{ij} e_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} e_j^{(k)}.$$

For $i = 1$, we have

$$|e_1^{(k+1)}| \leq \sum_{j=2}^{n} \left| \frac{a_{1j}}{a_{11}} \right| |e_j^{(k)}| \leq r_i \|\mathbf{e}^{(k)}\|_\infty \leq r \|\mathbf{e}^{(k)}\|_\infty.$$

Assume that $|e_p^{(k+1)}| \leq r\|\mathbf{e}^{(k)}\|_\infty$ for $p = 1, \ldots, i-1$. Then

$$|e_i^{(k+1)}| \leq \sum_{j=1}^{i-1} \left| \frac{a_{1j}}{a_{11}} \right| |e_j^{(k+1)}| + \sum_{j=i+1}^{n} \left| \frac{a_{1j}}{a_{11}} \right| |e_j^{(k)}|$$

$$\begin{aligned}
&\leq r\|\mathbf{e}^{(k)}\|_\infty \sum_{j=1}^{i-1}\left|\frac{a_{1j}}{a_{11}}\right| + \|\mathbf{e}^{(k)}\|_\infty \sum_{j=i+1}^{n}\left|\frac{a_{1j}}{a_{11}}\right| \\
&\leq \|\mathbf{e}^{(k)}\|_\infty \sum_{j\neq i}^{n}\left|\frac{a_{1j}}{a_{11}}\right| \\
&\leq r\|\mathbf{e}^{(k)}\|_\infty.
\end{aligned}$$

Therefore,

$$\|\mathbf{e}^{(k+1)}\|_\infty \leq r\|\mathbf{e}^{(k)}\|_\infty \leq r^{k+1}\|\mathbf{e}^{(0)}\|_\infty,$$

from which it follows that $\lim_{k\to\infty} \|\mathbf{e}^{(k)}\|_\infty = 0$ since $r < 1$. ■

8.1.4 SOR Method

SOR Iteration

The *successive over relaxation (SOR)* is given by the iteration

$$x_i^{(k+1)} = \omega \cdot \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right) + (1-\omega)x_i^{(k)},$$

or in the matrix form,

$$\begin{aligned}
D\mathbf{x}^{(k+1)} &= \omega(-L\mathbf{x}^{(k+1)} - U\mathbf{x}^{(k)} + \mathbf{b}) + (1-\omega)D\mathbf{x}^{(k)} \\
(\omega L + D)\mathbf{x}^{(k+1)} &= [(1-\omega)D - \omega U]\mathbf{x}^{(k)} + \omega\mathbf{b},
\end{aligned}$$

where ω is called the relaxation parameter.

The iteration martix is

$$B_\omega = (\omega L + D)^{-1}\left[(1-\omega)D - \omega U\right]$$

Convergence Analysis

SOR is at least as fast as Gauss–Seidel and often faster.

8.2 Semi-Iterative Methods

8.2.1 Overview

Semi-iterative methods generate

$$\mathbf{y}^{(k)} = B\mathbf{y}^{(k-1)} + \mathbf{c}$$

for suitable B and $\mathbf{c}$ and then form

$$\mathbf{x}^{(k)} = \sum_{j=0}^{k} \alpha_{jk}\mathbf{y}^{(j)}.$$

8.2.2 Richardson Method

Richardson Iteration

The Richardson iteration is given by

$$\begin{aligned}
\mathbf{r}^{(k)} &= \mathbf{b} - A\mathbf{x}^{(k)} \\
\mathbf{x}^{(k+1)} &= (I - \alpha A)\mathbf{x}^{(k)} + \alpha\mathbf{b} \\
&= \mathbf{x}^{(k)} + \alpha\mathbf{r}^{(k)} \\
&= \sum_{j=0}^{k} \alpha^{k-j+1}\mathbf{r}^{(j)}
\end{aligned}$$

where $\mathbf{r}^{(k)}$ is the residual in the kth step, which is also the backward error.

Convergence Analysis

Consider the error $\mathbf{e}^{(k)} = \mathbf{x} - \mathbf{x}^{(k)}$, we have

$$\mathbf{e}^{(k+1)} = B_\alpha \mathbf{e}^{(k)}$$

where the iteration matrix is given by $B_\alpha = I - \alpha A$. We want to choose α such that $\rho(B_\alpha)$ can be minimized. Suppose that A is symmetric positive definite, with eigenvalues $\mu_1 \geq \cdots \geq \mu_n > 0$. Then teh eigenvalues of B_α is given by $\lambda_i = 1 - \mu_i$ for $i = 1, \ldots, n$.

$$\min_{\alpha} \max_{1 \leq i \leq n} |1 - \alpha\mu_i| = \min_{\alpha} \max\{|1 - \alpha\mu_1|, |1 - \alpha\mu_n|\}$$

The optimal parameter α_* is attained when $|1-\alpha\mu_1| = |1-\alpha\mu_n|$ and $\alpha > 0$, which yields $\alpha_* = \frac{2}{\mu_1+\mu_n}$. The method converges for $0 < \alpha < \frac{2}{\mu_1+\mu_n} < \frac{2}{\mu_1}$. The optimal convergence rate is given by

$$\rho(B_{\alpha_*}) = \frac{\kappa(A) - 1}{\kappa(A) + 1},$$

whihc is related to $\kappa(A)$.

8.2.3 Steepest Descent Method

Steepest Descent Iteration

To speed up Richardson method, we consider varying the parameter α_k from one iteration to the next, i.e.,

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \mathbf{r}^{(k)}$$

Assume that A is symmetric positive definite. We wish to choose α_k such that the Mahalanobis norm $\|\mathbf{r}^{(k+1)}\|_{A^{-1}}$ is minimized. Since

$$\mathbf{r}^{(k+1)} = \mathbf{b} - A\mathbf{x}^{(k+1)} = \mathbf{b} - A\mathbf{x}^{(k)} - \alpha_k A\mathbf{r}^{(k)} = (I - \alpha_k A)\mathbf{r}^{(k)},$$

we have

$$\begin{aligned}
\|\mathbf{r}^{(k+1)}\|_{A^{-1}}^2 &= \mathbf{r}^{(k)\top}(A^{-1} - \alpha_k)(I - \alpha_k A)\mathbf{r}^{(k)} \\
&= \mathbf{r}^{(k)\top}(A^{-1} - 2\alpha_k + \alpha_k^2 A)\mathbf{r}^{(k)}, \\
\frac{\partial}{\partial \alpha_k}\|\mathbf{r}^{(k+1)}\|_{A^{-1}}^2 &= 2\mathbf{r}^{(k)\top} A\mathbf{r}^{(k)}\alpha_k - 2\mathbf{r}^{(k)\top}\mathbf{r}^{(k)}.
\end{aligned}$$

which yields

$$\alpha_*^{(k)} = \frac{\mathbf{r}^{(k)\top}\mathbf{r}^{(k)}}{\mathbf{r}^{(k)\top} A\mathbf{r}^{(k)}},$$

and the minimum norm

$$\|\mathbf{r}^{(k+1)}\|_{A^{-1}}^2 = \mathbf{r}^{(k)\top} A^{-1}\mathbf{r}^{(k)} - \frac{(\mathbf{r}^{(k)\top}\mathbf{r}^{(k)})^2}{\mathbf{r}^{(k)\top} A\mathbf{r}^{(k)}}.$$

Note that for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{0}$,

$$\lambda_{\min}(A) \leq \frac{\mathbf{x}^\top A\mathbf{x}}{\mathbf{x}^\top\mathbf{x}} \leq \lambda_{\max}(A)$$

and therefore

$$\frac{1}{\lambda_{\max}(A)} \leq \alpha_*^{(k)} \leq \frac{1}{\lambda_{\min}(A)}.$$

Convergence Analysis

It follows from the Kantorovich inequality that

$$\frac{\|\mathbf{r}^{(k+1)}\|_{A^{-1}}^2}{\|\mathbf{r}^{(k)}\|_{A^{-1}}^2} \leq \left(\frac{\kappa(A) - 1}{\kappa(A) + 1}\right)^2$$

which is the same rate of convergence as the Richardson iterations. However, α_* can not be determined if we don't know $\mu_1(A)$ and $\mu_n(A)$.

8.2.4 Chebyshev Iteration

Chebyshev iteration

Similar to Steepest Descent method,

$$\mathbf{x}^{(k+1)} = (I - \alpha_k A)\mathbf{x}^{(k)} + \alpha_k \mathbf{b}$$

and $\alpha^{(k)}$ is chosen to minimize the forward error $\|\mathbf{e}^{(k)}\|_2$.

Since

$$\begin{aligned}
\mathbf{e}^{(k+1)} &= (I - \alpha_k A)\mathbf{e}^{(k)} \\
&= \left[\prod_{j=0}^{k}(I - \alpha_j A)\right]\mathbf{e}^{(0)} \\
&= P_k(A)\mathbf{e}^{(0)} \\
\|\mathbf{e}^{(k+1)}\|_2 &\le \|P_k(A)\|_2 \|\mathbf{e}^{(0)}\|_2,
\end{aligned}$$

we need to minimize $\|P_k(A)\|_2$, the 2-norm of a degree-k polynomial. If we know the characteristic polynomial or the minimal polynomial of A, then $\|P_k(A)\|_2$ can be zero. However, almost never know the eigenvalues of A and even if we do, the calculation of monomials $I - \frac{1}{\mu_i}A$ may be unstable.

If A is symmetric positive definite, then it has unitary eigen-decomposition $A = Q\Lambda Q^\top$ with eigenvalues $\mu_1 \ge \cdots \ge \mu_n > 0$. We need extra constraints that $P_k(1) = \sum_{j=0}^{k} \alpha_j = 1$. Then

$$\begin{aligned}
\min_{\substack{P_k \in \mathbb{C}[x] \\ \deg(P_k)=k \\ P_k(1)=1}} \|P_k(A)\|_2 &= \min_{\substack{P_k \in \mathbb{C}[x] \\ \deg(P_k)=k \\ P_k(1)=1}} \max_i |P_k(\mu_i)| \\
&\le \min_{\substack{P_k \in \mathbb{C}[x] \\ \deg(P_k)=k \\ P_k(1)=1}} \max_{\mu \in [\mu_n, \mu_1]} |P_k(\mu)| \\
&\le \min_{\substack{P_k \in \mathbb{C}[x] \\ \deg(P_k)=k \\ P_k(1)=1}} \max_{\mu \in [a, b]} |P_k(\mu)|,
\end{aligned}$$

where a and b are chosen so that $a \le \mu_n \le \cdots \le \mu_1 \le b$ since usually we only know the lower and upper bound of A's eigenvalues. The unique solution to this optimization problem is given by $C_k(x) = \cos(k \cos^{-1}(x))$, the Chebyshev polynomial if $a = -1$ and $b = 1$. In general, we can transform the interval $[a, b]$ to $[-1, 1]$ by $t \mapsto \frac{2t-(a+b)}{b-a}$.

8.3 Krylov Subspace Methods

8.3.1 Overview

To solve $A\mathbf{x} = \mathbf{b}$, many Krylov subspace methods converge in k steps where k = number of distinct nonzero eigenvalues of A.

1. conjugate gradient (cg) method for symmetric positive definite A;
2. minimal residual (minres) method for symmetric A;
3. general minimial resitual (gmres) method for general A.

It aims to find $\mathbf{x}^{(k)} \in \text{span}\{\mathbf{b}, A\mathbf{b}, \ldots, A^k\mathbf{b}\}$ to approximate the solution $\mathbf{x}$.