MATH 532: Linear Algebra

Chapter 7: Eigenvalues and Eigenvectors

Greg Fasshauer

Department of Applied Mathematics
Illinois Institute of Technology

Spring 2015

Outline

Outline

Motivation

Eigenvalues are important, e.g.,

- to decouple systems of ODEs,
- to study physical phenomena such as resonance,
- to tackle the same kind of applications as the SVD (whenever the matrix is symmetric).

Definition

Let A be an $n \times n$ matrix. The scalars λ and nonzero n-vectors $\boldsymbol{x}$ satisfying

$$A\boldsymbol{x} = \lambda \boldsymbol{x}$$

are called eigenvalues and eigenvectors of A. We call $(\lambda, \boldsymbol{x})$ an eigenpair of A.

Definition

Let A be an $n \times n$ matrix. The scalars λ and nonzero n-vectors $\boldsymbol{x}$ satisfying

$$A\boldsymbol{x} = \lambda \boldsymbol{x}$$

are called eigenvalues and eigenvectors of A. We call $(\lambda, \boldsymbol{x})$ an eigenpair of A.
The set of all eigenvalues of A is called the spectrum $\sigma(A)$, i.e.,

$$\sigma(A) = \{\lambda : \ \lambda \text{ is an eigenvalue of } A\}.$$

Definition

Let A be an $n \times n$ matrix. The scalars λ and nonzero n-vectors $\boldsymbol{x}$ satisfying

$$\mathsf{A}\boldsymbol{x} = \lambda \boldsymbol{x}$$

are called eigenvalues and eigenvectors of A. We call $(\lambda, \boldsymbol{x})$ an eigenpair of A.
The set of all eigenvalues of A is called the spectrum $\sigma(\mathsf{A})$, i.e.,

$$\sigma(\mathsf{A}) = \{\lambda : \ \lambda \text{ is an eigenvalue of A}\}.$$

The spectral radius of A is given by

$$\rho(\mathsf{A}) = \max_{\lambda \in \sigma(\mathsf{A})} |\lambda|.$$

Theorem

The following are equivalent:

1. λ *is a eigenvalue of* A.
2. $\mathsf{A} - \lambda \mathsf{I}$ *is singular.*
3. $\det(\mathsf{A} - \lambda \mathsf{I}) = 0$.

Proof.

By definition, λ satisfies $A\boldsymbol{x} = \lambda\boldsymbol{x}$. This can be written as

$$(A - \lambda I)\boldsymbol{x} = \boldsymbol{0}.$$

We get a nontrivial solution (recall that eigenvectors are always nonzero) if and only if

$$A - \lambda I \quad \text{is singular.}$$

$\square$

Proof.

By definition, λ satisfies $\mathsf{A}\boldsymbol{x} = \lambda\boldsymbol{x}$. This can be written as

$$(\mathsf{A} - \lambda\mathsf{I})\boldsymbol{x} = \mathbf{0}.$$

We get a nontrivial solution (recall that eigenvectors are always nonzero) if and only if

$$\mathsf{A} - \lambda\mathsf{I} \quad \text{is singular.}$$

□

Remark

This proof shows that the eigenvector $\boldsymbol{x} \in N(\mathsf{A} - \lambda\mathsf{I})$.

Remark

- *In fact, any vector in* $\mathrm{N}(\mathsf{A} - \lambda \mathsf{I})$ *is an eigenvector of* A *associated with* λ*, i.e., eigenvectors are not unique.*

Remark

- *In fact, any vector in* $\mathsf{N}(\mathsf{A} - \lambda\mathsf{I})$ *is an eigenvector of* A *associated with* λ*, i.e., eigenvectors are not unique.*
- *Terminology:* $\mathsf{N}(\mathsf{A} - \lambda\mathsf{I})$ *is called the eigenspace of* A *associated with* λ*.*

Remark

- *In fact, any vector in* $\mathsf{N}(\mathsf{A} - \lambda\mathsf{I})$ *is an eigenvector of* A *associated with* λ*, i.e., eigenvectors are not unique.*
- *Terminology:* $\mathsf{N}(\mathsf{A} - \lambda\mathsf{I})$ *is called the eigenspace of* A *associated with* λ*.*
- *Geometric interpretation: For eigenpairs, matrix multiplication by* A *acts just like scalar multiplication, i.e.,* $\mathsf{A}\boldsymbol{x}$ *differs from* $\boldsymbol{x}$ *only by a stretch factor or a change in orientation (if* $\lambda < 0$*).*

Definition

Let A be an $n \times n$ matrix. The characteristic polynomial of A is given by

$$p(\lambda) = \det(\mathsf{A} - \lambda \mathsf{I}),$$

and $p(\lambda) = 0$ is called the characteristic equation.

Definition

Let A be an $n \times n$ matrix. The characteristic polynomial of A is given by

$$p(\lambda) = \det(A - \lambda I),$$

and $p(\lambda) = 0$ is called the characteristic equation.

Remark

The basic properties of determinant show that

- *degree*$(p) = n$,
- *the leading coefficient, i.e., the coefficient of* λ^n *is* $(-1)^n$.

Immediate consequences

1. The eigenvalues of A are roots of the characteristic polynomial.

Immediate consequences

1. The eigenvalues of A are roots of the characteristic polynomial.
2. A has n (possibly complex, but necessarily distinct) eigenvalues.

Immediate consequences

1. The eigenvalues of A are roots of the characteristic polynomial.
2. A has n (possibly complex, but necessarily distinct) eigenvalues.
3. If A is real, then complex eigenvalues appear in conjugate pairs, i.e., $\lambda \in \sigma(\mathsf{A}) \quad \Longrightarrow \quad \overline{\lambda} \in \sigma(\mathsf{A})$.

Immediate consequences

1. The eigenvalues of A are roots of the characteristic polynomial.
2. A has n (possibly complex, but necessarily distinct) eigenvalues.
3. If A is real, then complex eigenvalues appear in conjugate pairs, i.e., $\lambda \in \sigma(\mathsf{A}) \quad \Longrightarrow \quad \overline{\lambda} \in \sigma(\mathsf{A})$.
4. In particular, simple real (even integer) matrices can have complex eigenvalues and eigenvectors.

Example

Find the eigenvalues and eigenvectors of $A = \begin{pmatrix} 1 & 2 \\ -1 & 1 \end{pmatrix}$.

Example

Find the eigenvalues and eigenvectors of $A = \begin{pmatrix} 1 & 2 \\ -1 & 1 \end{pmatrix}$.

We need to solve

$$p(\lambda) = \det(A - \lambda I) = (1 - \lambda)^2 + 2 = 0$$

Example

Find the eigenvalues and eigenvectors of $A = \begin{pmatrix} 1 & 2 \\ -1 & 1 \end{pmatrix}$.

We need to solve

$$\begin{aligned} & p(\lambda) = \det(A - \lambda I) = (1-\lambda)^2 + 2 = 0 \\ \iff \quad & \lambda^2 - 2\lambda + 3 = 0 \end{aligned}$$

Example

Find the eigenvalues and eigenvectors of $A = \begin{pmatrix} 1 & 2 \\ -1 & 1 \end{pmatrix}$.

We need to solve

$$\begin{aligned} & p(\lambda) = \det(A - \lambda I) = (1-\lambda)^2 + 2 = 0 \\ \iff \quad & \lambda^2 - 2\lambda + 3 = 0 \\ \implies \quad & \lambda = \frac{2 \pm \sqrt{4-12}}{2} = 1 \pm \sqrt{2}i. \end{aligned}$$

Example

Find the eigenvalues and eigenvectors of $A = \begin{pmatrix} 1 & 2 \\ -1 & 1 \end{pmatrix}$.

We need to solve

$$\begin{aligned} & p(\lambda) = \det(A - \lambda I) = (1-\lambda)^2 + 2 = 0 \\ \iff \quad & \lambda^2 - 2\lambda + 3 = 0 \\ \implies \quad & \lambda = \frac{2 \pm \sqrt{4-12}}{2} = 1 \pm \sqrt{2}i. \end{aligned}$$

Therefore, $\sigma(A) = \{1 + i\sqrt{2}, 1 - i\sqrt{2}\}$.

Example (cont.)

Now, compute the eigenvectors for
$\lambda_1 = 1 + i\sqrt{2}$:

$$A - \lambda_1 I = \begin{pmatrix} -i\sqrt{2} & 2 \\ -1 & -i\sqrt{2} \end{pmatrix} \longrightarrow \begin{pmatrix} 0 & 0 \\ -1 & -i\sqrt{2} \end{pmatrix}$$

so that $N(A - \lambda_1 I) = \text{span}\{(i\sqrt{2}, -1)^T\}$.

Example (cont.)

Now, compute the eigenvectors for

$\lambda_1 = 1 + i\sqrt{2}$:

$$A - \lambda_1 I = \begin{pmatrix} -i\sqrt{2} & 2 \\ -1 & -i\sqrt{2} \end{pmatrix} \longrightarrow \begin{pmatrix} 0 & 0 \\ -1 & -i\sqrt{2} \end{pmatrix}$$

so that $N(A - \lambda_1 I) = \text{span}\{(i\sqrt{2},\ -1)^T\}$.

$\lambda_1 = 1 - i\sqrt{2}$:

$$A - \lambda_2 I = \begin{pmatrix} i\sqrt{2} & 2 \\ -1 & i\sqrt{2} \end{pmatrix} \longrightarrow \begin{pmatrix} 0 & 0 \\ -1 & i\sqrt{2} \end{pmatrix}$$

so that $N(A - \lambda_2 I) = \text{span}\{(i\sqrt{2},\ 1)^T\}$.

Remark

Since eigenvalues are the solution of polynomial equations and we know due to Abel's theorem that there is no closed form expression for roots of polynomials of degree five or greater, general methods for finding eigenvalues necessarily have to be iterative (and numerical).

Formulas for coefficients of characteristic polynomial

If we write

$$(-1)^n p(\lambda) = \lambda^n + c_1\lambda^{n-1} + c_2\lambda^{n-2} + \ldots + c_{n-1}\lambda + c_n$$

then without proof/derivation (see [Mey00] for details)

$$c_k = (-1)^k s_k, \qquad c_0 = 1,$$

where

$$\begin{aligned} s_k &= \sum(\text{all } k \times k \text{ determinant of principal submatrices}) \\ &= \sum(\text{all products of subsets of } k \text{ eigenvalues}) \end{aligned}$$

Special cases

$$\begin{aligned} \text{trace}(\mathsf{A}) &= \lambda_1 + \lambda_2 + \ldots + \lambda_n = -c_1, \\ \det(\mathsf{A}) &= \lambda_1\lambda_2\ldots\lambda_n = (-1)^n c_n. \end{aligned}$$

Example

Compute the characteristic polynomial for

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

We first compute

$$\begin{aligned} (-1)^3 p(\lambda) = -\det(A - \lambda I) &= (1-\lambda)^2(1+\lambda) \\ &= (\lambda^2 - 2\lambda + 1)(1+\lambda) \\ &= \lambda^3 - \lambda^2 - \lambda + 1. \end{aligned}$$

Example (cont.)

On the other hand (using the above formulas)

$$
\begin{aligned}
c_0 &= 1,\\
s_1 &= \det(1) = 1 \quad \Longrightarrow \quad c_1 = -s_1 = -1,\\
s_2 &= \det\begin{pmatrix} 1 & 2 \\ 0 & -1 \end{pmatrix} + \det\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} + \det\begin{pmatrix} -1 & 1 \\ 0 & 1 \end{pmatrix}\\
&= -1 + 1 - 1 = -1 \quad \Longrightarrow \quad c_2 = s_2 = -1,\\
s_3 &= \det(A) = -1 \quad \Longrightarrow \quad c_3 = -s_3 = 1.
\end{aligned}
$$

Example (cont.)

The corresponding eigenvectors are

$$\lambda = -1: \ \boldsymbol{x} = (1,\ -1,\ 0)^T$$
$$\lambda = 1: \ \boldsymbol{x} = (1,\ 0,\ 0)^T$$

Note that $\lambda = 1$ is a double eigenvalue, but the eigenspace is only one-dimensional, i.e., there is a deficiency (see algebraic vs. geometric multiplicities later).

Example

The trace and determinant combination is particularly applicable to 2×2 problems. Consider

$$A = \begin{pmatrix} 1 & 2 \\ -1 & 1 \end{pmatrix}$$

then

$$\begin{aligned} \mathrm{trace}(A) &= 2 = \lambda_1 + \lambda_2 \\ \det(A) &= 3 = \lambda_1 \lambda_2 \end{aligned}$$

so that $\lambda_1 = 2 - \lambda_2$ implies

$$(2 - \lambda_2)\lambda_2 = 3 \quad \Longrightarrow \quad \lambda_2^2 - 2\lambda_2 + 3 = 0$$

as earlier.

Often, the largest eigenvalue is especially important.
Recall spectral radius: $\rho(\mathsf{A}) = \max_{\lambda \in \sigma(\mathsf{A})} |\lambda|$.
A simple upper bound is, using any matrix norm,

$$\rho(\mathsf{A}) \leq \|\mathsf{A}\|.$$

We now prove this.

Proof.

First, we remember submultiplicativity of matrix norms, i.e.,

$$\|AX\| \leq \|A\|\|X\| \qquad \text{for any } X. \tag{1}$$

Now, take $X = \begin{pmatrix} \boldsymbol{x} & \mathbf{0} & \cdots \mathbf{0} \end{pmatrix}$ with $(\lambda, \boldsymbol{x})$ and eigenpair of A.
Then $AX = \lambda X$ and

$$\|AX\| = \|\lambda X\| = |\lambda|\|X\|. \tag{2}$$

Combine (1) and (2):

$$\begin{aligned} & |\lambda|\|X\| = \|AX\| \leq \|A\|\|X\| \\ \overset{\|X\|\neq 0}{\Longrightarrow} \quad & |\lambda| \leq \|A\| \\ \overset{\lambda \text{ arb.}}{\Longrightarrow} \quad & \rho(A) \leq \|A\| \end{aligned}$$

$\square$

More precise estimates of eigenvalues can be obtained with Gerschgorin circles.

Definition

Let $A \in \mathbb{C}^{n \times n}$. The Gerschgorin circles $\mathcal{G}_i$ of A are defined by

$$\mathcal{G}_i = \{z \in \mathbb{C} : \; |z - a_{ii}| \le r_i\}, \quad i = 1, \ldots, n$$

with $r_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|$, the (off-diagonal) row sums of A.

Remark

Analogous (but not the same) circles can be defined via column sums.

Theorem

Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ *and* $\mathcal{G}_i$, $i = 1, \ldots, n$, *be its Gerschgorin circles. Then*

$$\sigma(\mathsf{A}) \subseteq \bigcup_{i=1}^{n} \mathcal{G}_i.$$

Remark

If we use two sets of Gerschgorin circles, $\mathcal{G}_r$ *and* $\mathcal{G}_c$ *(defined via rows sums and via column sums, respectively), then we get a better estimate:*

$$\sigma(\mathsf{A}) \subseteq \mathcal{G}_r \cap \mathcal{G}_c.$$

Before we prove the theorem we illustrate with an example.

Example

Consider

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 2 & -1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

with rough estimate $\rho(A) \leq \|A\|_\infty = 3$.
The Gerschgorin circles are

$$\mathcal{G}_1 = \{z : \ |z - 1| \leq 1\}$$
$$\mathcal{G}_2 = \{z : \ |z + 1| \leq 2\}$$
$$\mathcal{G}_1 = \{z : \ |z - 1| \leq 1\}$$

Proof

Assume $(\lambda, \boldsymbol{x})$ us an eigenpair with $\boldsymbol{x}$ normalized, i.e., $\|\boldsymbol{x}\|_\infty = 1$. Consider i such that $|x_i| = \|\boldsymbol{x}\|_\infty = 1$. Then

$$\lambda x_i = (\lambda \boldsymbol{x})_i = (\mathsf{A}\boldsymbol{x})_i = \sum_{j=1}^{n} a_{ij} x_j = a_{ii} x_i + \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j$$

so that

$$(\lambda - a_{ii}) x_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j.$$

Proof (cont.)

Then

$$|\lambda - a_{ii}| = |\lambda - a_{ii}| \underbrace{|x_i|}_{=1} = \left| \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j \right|$$

$$\overset{\Delta \text{ ineq.}}{\leq} \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \underbrace{|x_j|}_{\leq \|\boldsymbol{x}\|_\infty = 1}$$

$$\leq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| = r_i.$$

Therefore $\lambda \in \mathcal{G}_i$ and each λ will lie in some $\mathcal{G}_i$, i.e.,

$$\sigma(\mathsf{A}) \subseteq \bigcup_{i=1}^{n} \mathcal{G}_i.$$

□

Remark

There is no reason to believe that every Gerschgorin circle contains an eigenvalue.

Example

The eigenvalues of $A = \begin{pmatrix} 0 & 1 \\ 4 & 0 \end{pmatrix}$ are $\lambda_{1,2} = \pm 2$.

But we have

$$\mathcal{G}_1 = \{z : \ |z| \le 1\}$$
$$\mathcal{G}_2 = \{z : \ |z| \le 4\}$$

and $\mathcal{G}_1$ does not contain an eigenvalue.

Remark

Recall that a diagonally dominant matrix satisfies

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \quad i = 1, \ldots, n.$$

However, then the proof above shows that $\lambda = 0$ cannot be an eigenvalue of a diagonally dominant matrix.
Therefore, diagonally dominant matrices are nonsingular (cf. HW).

Outline

Recall: Equivalence

$$A \sim B \quad \text{if and only if there exist } P, Q \text{ nonsingular s.t. } PAQ = B.$$

Recall: Equivalence

$$A \sim B \quad \text{if and only if there exist } P, Q \text{ nonsingular s.t. } PAQ = B.$$

Now

Definition

Two $n \times n$ matrices A and B are called similar if there exists a nonsingular P such that

$$P^{-1}AP = B.$$

Recall: Equivalence

$$A \sim B \quad \text{if and only if there exist } P, Q \text{ nonsingular s.t. } PAQ = B.$$

Now

Definition

Two $n \times n$ matrices A and B are called similar if there exists a nonsingular P such that

$$P^{-1}AP = B.$$

Definition

An $n \times n$ matrix A is called diagonalizable if A is similar to a diagonal matrix, i.e., if

$$P^{-1}AP = D$$

for some nonsingular matrix P.

Remark

We already know the SVD, i.e.,

$$\mathsf{A} = \mathsf{U}\mathsf{D}\mathsf{V}^T \quad \Longleftrightarrow \quad \mathsf{U}^T\mathsf{A}\mathsf{V} = \mathsf{D}, \quad \mathsf{U}, \mathsf{V} \text{ unitary,}$$

where D *contains the singular values of* A*.*

Remark

We already know the SVD, i.e.,

$$\mathsf{A} = \mathsf{U}\mathsf{D}\mathsf{V}^T \quad \Longleftrightarrow \quad \mathsf{U}^T\mathsf{A}\mathsf{V} = \mathsf{D}, \quad \mathsf{U}, \mathsf{V} \textit{ unitary},$$

where D *contains the singular values of* A*.*

Now we use a single transformation matrix, and D *will contain the eigenvalues of* A*.*

Remark

We already know the SVD, i.e.,

$$\mathsf{A} = \mathsf{U}\mathsf{D}\mathsf{V}^T \quad \Longleftrightarrow \quad \mathsf{U}^T\mathsf{A}\mathsf{V} = \mathsf{D}, \quad \mathsf{U}, \mathsf{V} \textit{ unitary,}$$

where D *contains the singular values of* A*.*

Now we use a single transformation matrix, and D *will contain the eigenvalues of* A*.*

However, every matrix A has an SVD. Not so now...

Theorem

An $n \times n$ matrix A is diagonalizable if and only if A possesses a complete set of eigenvectors (i.e., it has n linearly independent eigenvectors). Moreover,

$$\mathsf{P}^{-1}\mathsf{A}\mathsf{P} = \mathsf{D} = diag(\lambda_1, \ldots, \lambda_n)$$

if and only if $(\lambda_j, \mathsf{P}_{*j})$, $j = 1, \ldots, n$, are eigenpairs of A.

Theorem

An $n \times n$ matrix A is diagonalizable if and only if A possesses a complete set of eigenvectors (i.e., it has n linearly independent eigenvectors). Moreover,

$$\mathsf{P}^{-1}\mathsf{A}\mathsf{P} = \mathsf{D} = diag(\lambda_1, \ldots, \lambda_n)$$

if and only if $(\lambda_j, \mathsf{P}_{*j})$, $j = 1, \ldots, n$, are eigenpairs of A.

Remark

If A possesses a complete set of eigenvectors it is called nondefective (or nondeficient).

Proof.

$$
\begin{aligned}
& \mathsf{P}^{-1}\mathsf{A}\mathsf{P} = \mathsf{D} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \\
\iff \quad & \mathsf{A}\mathsf{P} = \mathsf{P}\mathsf{D} \\
\iff \quad & \mathsf{A}\begin{pmatrix} \mathsf{P}_{*1} & \mathsf{P}_{*2} & \cdots & \mathsf{P}_{*n} \end{pmatrix} = \begin{pmatrix} \mathsf{P}_{*1} & \mathsf{P}_{*2} & \cdots & \mathsf{P}_{*n} \end{pmatrix} \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \\
\iff \quad & \begin{pmatrix} \mathsf{A}\mathsf{P}_{*1} & \mathsf{A}\mathsf{P}_{*2} & \cdots & \mathsf{A}\mathsf{P}_{*n} \end{pmatrix} = \begin{pmatrix} \lambda_1\mathsf{P}_{*1} & \lambda_2\mathsf{P}_{*2} & \cdots & \lambda_n\mathsf{P}_{*n} \end{pmatrix} \\
\iff \quad & (\lambda_j, \mathsf{P}_{*j}) \text{ is an eigenpair of } \mathsf{A}
\end{aligned}
$$

Note that P is invertible if and only if the columns of P are linearly independent. $\square$

Example

Consider

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

with

$$\lambda_1 = 1, \quad N(A - I) = \text{span}\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \}$$

and

$$\lambda_2 = -1, \quad N(A + I) = \text{span}\{ \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \}$$

is not diagonalizable since the set of eigenvectors in not complete.

Example

Consider

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 2 & -1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

with characteristic polynomial

$$p(\lambda) = (1-\lambda)^2(1+\lambda) + 1 = \lambda^2 - \lambda^2 - 2\lambda = \lambda(\lambda+1)(\lambda-2)$$

and spectrum

$$\sigma(A) = \{-1, 0, 2\}.$$

Example (cont.)

Also, $N(\mathsf{A}+\mathsf{I})$:

$$\begin{pmatrix} 2 & 0 & 1 \\ 2 & 0 & 0 \\ 1 & 0 & 2 \end{pmatrix} \longrightarrow \begin{pmatrix} 2 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{3}{2} \end{pmatrix}$$

so that $N(\mathsf{A}+\mathsf{I}) = \operatorname{span}\{(0,1,0)^T\}$ (first eigenvector).

Example (cont.)

Also, $N(A + I)$:

$$\begin{pmatrix} 2 & 0 & 1 \\ 2 & 0 & 0 \\ 1 & 0 & 2 \end{pmatrix} \longrightarrow \begin{pmatrix} 2 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{3}{2} \end{pmatrix}$$

so that $N(A + I) = \text{span}\{(0, 1, 0)^T\}$ (first eigenvector).

$N(A)$:

$$\begin{pmatrix} 1 & 0 & 1 \\ 2 & -1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & 0 \end{pmatrix}$$

so that $N(A) = \text{span}\{(-1, -2, 1)^T\}$.

Example (cont.)

Also, $N(\mathsf{A}+\mathsf{I})$:

$$\begin{pmatrix} 2 & 0 & 1 \\ 2 & 0 & 0 \\ 1 & 0 & 2 \end{pmatrix} \longrightarrow \begin{pmatrix} 2 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{3}{2} \end{pmatrix}$$

so that $N(\mathsf{A}+\mathsf{I}) = \operatorname{span}\{(0,1,0)^T\}$ (first eigenvector).

$N(\mathsf{A})$:

$$\begin{pmatrix} 1 & 0 & 1 \\ 2 & -1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & 0 \end{pmatrix}$$

so that $N(\mathsf{A}) = \operatorname{span}\{(-1,-2,1)^T\}$.

$N(\mathsf{A}-2\mathsf{I})$:

$$\begin{pmatrix} -1 & 0 & 1 \\ 2 & -3 & 0 \\ 1 & 0 & -11 \end{pmatrix} \longrightarrow \begin{pmatrix} -1 & 0 & 1 \\ 0 & -3 & 2 \\ 0 & 0 & 0 \end{pmatrix}$$

so that $N(\mathsf{A}-2\mathsf{I}) = \operatorname{span}\{(1,\frac{2}{3},1)^T\}$.

Example (cont.)

Therefore

$$P = \begin{pmatrix} 0 & -1 & 1 \\ 1 & -2 & \frac{2}{3} \\ 0 & 1 & 1 \end{pmatrix}, \quad \text{so that} \quad P^{-1} = \begin{pmatrix} -\frac{4}{3} & 1 & \frac{2}{3} \\ -\frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}$$

and

$$P^{-1}AP = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Theorem

If A, B *are similar, then* $\sigma(\mathsf{A}) = \sigma(\mathsf{B})$.

Theorem

If A, B *are similar, then* $\sigma(\mathsf{A}) = \sigma(\mathsf{B})$.

Proof.

We show $\det(\mathsf{A} - \lambda\mathsf{I}) = \det(\mathsf{B} - \lambda\mathsf{I})$, i.e., A and B have the same characteristic polynomials.

Theorem

If A, B *are similar, then* $\sigma(\mathsf{A}) = \sigma(\mathsf{B})$*.*

Proof.

We show $\det(\mathsf{A} - \lambda\mathsf{I}) = \det(\mathsf{B} - \lambda\mathsf{I})$, i.e., A and B have the same characteristic polynomials.

Since A, B are similar there exists a nonsingular P such that $\mathsf{P}^{-1}\mathsf{A}\mathsf{P} = \mathsf{B}$.

Theorem

If A, B *are similar, then* $\sigma(\mathsf{A}) = \sigma(\mathsf{B})$*.*

Proof.

We show $\det(\mathsf{A} - \lambda\mathsf{I}) = \det(\mathsf{B} - \lambda\mathsf{I})$, i.e., A and B have the same characteristic polynomials.

Since A, B are similar there exists a nonsingular P such that $\mathsf{P}^{-1}\mathsf{A}\mathsf{P} = \mathsf{B}$. Now,

$$\det(\mathsf{B} - \lambda\mathsf{I}) = \det(\mathsf{P}^{-1}\mathsf{A}\mathsf{P} - \lambda\mathsf{I})$$

$\square$

Theorem

If A, B *are similar, then* $\sigma(A) = \sigma(B)$*.*

Proof.

We show $\det(A - \lambda I) = \det(B - \lambda I)$, i.e., A and B have the same characteristic polynomials.
Since A, B are similar there exists a nonsingular P such that $P^{-1}AP = B$. Now,

$$\begin{aligned}\det(B - \lambda I) &= \det(P^{-1}AP - \lambda I) \\ &= \det(P^{-1}AP - \lambda P^{-1}IP)\end{aligned}$$

□

Theorem

If A, B *are similar, then* $\sigma(A) = \sigma(B)$.

Proof.

We show $\det(A - \lambda I) = \det(B - \lambda I)$, i.e., A and B have the same characteristic polynomials.

Since A, B are similar there exists a nonsingular P such that $P^{-1}AP = B$. Now,

$$\begin{aligned}
\det(B - \lambda I) &= \det(P^{-1}AP - \lambda I) \\
&= \det(P^{-1}AP - \lambda P^{-1}IP) \\
&= \det\left(P^{-1}(A - \lambda I)P\right)
\end{aligned}$$

□

Theorem

If A, B *are similar, then* $\sigma(A) = \sigma(B)$.

Proof.

We show $\det(A - \lambda I) = \det(B - \lambda I)$, i.e., A and B have the same characteristic polynomials.

Since A, B are similar there exists a nonsingular P such that $P^{-1}AP = B$. Now,

$$\begin{aligned}
\det(B - \lambda I) &= \det(P^{-1}AP - \lambda I) \\
&= \det(P^{-1}AP - \lambda P^{-1}IP) \\
&= \det\left(P^{-1}(A - \lambda I)P\right) \\
&= \det(P^{-1})\det(A - \lambda I)\det(P) = \det(A - \lambda I)
\end{aligned}$$

$\square$

Theorem

If A, B *are similar, then* $\sigma(A) = \sigma(B)$.

Proof.

We show $\det(A - \lambda I) = \det(B - \lambda I)$, i.e., A and B have the same characteristic polynomials.

Since A, B are similar there exists a nonsingular P such that $P^{-1}AP = B$. Now,

$$\begin{aligned}
\det(B - \lambda I) &= \det(P^{-1}AP - \lambda I) \\
&= \det(P^{-1}AP - \lambda P^{-1}IP) \\
&= \det\left(P^{-1}(A - \lambda I)P\right) \\
&= \det(P^{-1})\det(A - \lambda I)\det(P) = \det(A - \lambda I)
\end{aligned}$$

since $\det(P^{-1}) = \frac{1}{\det(P)}$. $\square$

Remark

We saw above that there exist matrices that are not diagonalizable, i.e., are not similar to a diagonal matrix (of its eigenvalues).

Remark

We saw above that there exist matrices that are not diagonalizable, i.e., are not similar to a diagonal matrix (of its eigenvalues).

However, every square matrix A *is similar to a triangular matrix whose diagonal elements are the eigenvalues of* A

$\longrightarrow$ *Schur factorization (next).*

Theorem (Schur factorization)

For every $n \times n$ matrix A there exists a unitary matrix U (which is not unique) and an upper triangular matrix T (which is also not unique) such that

$$\mathsf{U}^*\mathsf{A}\mathsf{U} = \mathsf{T},$$

and the diagonal entries of T are the eigenvalues of A.

Proof

By induction. $n = 1$ is easy: $\mathsf{A} = a = \lambda$, $\mathsf{U} = 1$, $\mathsf{T} = \lambda$.
Assume the statement is true for $n - 1$, and show it also holds for n:
Take $(\lambda, \boldsymbol{x})$, an eigenpair of A with $\|\boldsymbol{x}\|_2 = 1$ and construct a Householder reflector R whose first column is $\boldsymbol{x}$ (see Sect. 5.6), i.e.,

$$\boldsymbol{x} = \mathsf{R}\boldsymbol{e}_1 \quad \overset{\mathsf{R}^{-1}=\mathsf{R}}{\Longleftrightarrow} \quad \mathsf{R}\boldsymbol{x} = \boldsymbol{e}_1.$$

Thus

$$\mathsf{R} = \begin{pmatrix} \boldsymbol{x} & \mathsf{V} \end{pmatrix}$$

for some V.

Proof (cont.)

Now

$$\begin{aligned}
R^*AR &\overset{R=R^*}{=} RAR = RA\begin{pmatrix}\boldsymbol{x} & V\end{pmatrix} \\
&= R\begin{pmatrix}A\boldsymbol{x} & AV\end{pmatrix} = R\begin{pmatrix}\lambda\boldsymbol{x} & AV\end{pmatrix} \\
&= \begin{pmatrix}\lambda \underbrace{R\boldsymbol{x}}_{=\boldsymbol{e}_1} & RAV\end{pmatrix} = \begin{pmatrix}\lambda\boldsymbol{e}_1 & R^*AV\end{pmatrix} \\
&= \begin{pmatrix}\lambda & \boldsymbol{x}^*AV \\ \mathbf{0} & V^*AV\end{pmatrix}
\end{aligned}$$

By the induction hypothesis V^*AV is similar to an upper triangular matrix, i.e., there exists a unitary Q such that

$$Q^*(V^*AV)Q = \widehat{T}.$$

Proof (cont.)

Finally, let $U = R\begin{pmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & Q \end{pmatrix}$ so that

$$
\begin{aligned}
U^*AU &= \begin{pmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & Q^* \end{pmatrix} \underbrace{R^*AR}_{=\begin{pmatrix} \lambda & \boldsymbol{x}^*AV \\ \mathbf{0} & V^*AV \end{pmatrix}} \begin{pmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & Q \end{pmatrix} \\
&= \begin{pmatrix} 1 & \mathbf{0}^* \\ \mathbf{0} & Q^* \end{pmatrix} \begin{pmatrix} \lambda & \boldsymbol{x}^*AVQ \\ \mathbf{0} & V^*AVQ \end{pmatrix} \\
&= \begin{pmatrix} \lambda & \boldsymbol{x}^*AVQ \\ \mathbf{0} & \underbrace{Q^*V^*AVQ}_{=\widehat{T}} \end{pmatrix} \\
&= T \qquad \text{upper triangular}
\end{aligned}
$$

Proof (cont.)

The diagonal entries of T are the eigenvalues of A since

- the similarity transformation preserves eigenvalues, and
- the eigenvalues of a triangular matrix are its diagonal elements.

$\square$

Theorem (Cayley–Hamilton Theorem)

Let $A \in \mathbb{C}^{n\times n}$ *and let* $p(\lambda) = 0$ *be its characteristic equation. Then*

$$p(A) = 0,$$

i.e., every square matrix satisfies its characteristic equation.

Theorem (Cayley–Hamilton Theorem)

Let $\mathsf{A} \in \mathbb{C}^{n \times n}$ *and let* $p(\lambda) = 0$ *be its characteristic equation. Then*

$$p(\mathsf{A}) = 0,$$

i.e., every square matrix satisfies its characteristic equation.

Proof.

There exist many different proofs. One possibility is via the Schur factorization theorem (see [Mey00, Ex. 7.2.2]). □

Multiplicities

Definition

Let $\lambda \in \sigma(\mathsf{A}) = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$.

1. The algebraic multiplicity of λ, $\text{algmult}_{\mathsf{A}}(\lambda)$, is its multiplicity as a root of the characteristic equation $p(\lambda) = 0$.
2. If $\text{algmult}_{\mathsf{A}}(\lambda) = 1$, then λ is called simple.
3. The geometric multiplicity of λ, $\text{geomult}_{\mathsf{A}}(\lambda)$, is $\dim N(\mathsf{A} - \lambda\mathsf{I})$, the dimension of the eigenspace of λ, i.e., the number of linearly independent eigenvectors associated with λ.
4. If $\text{algmult}_{\mathsf{A}}(\lambda) = \text{geomult}_{\mathsf{A}}(\lambda)$, then λ is called semi-simple.

Example

Consider

$$A = \begin{pmatrix} -1 & -1 & -2 \\ 8 & -11 & -8 \\ -10 & 11 & 7 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & -4 & -4 \\ 8 & -11 & -8 \\ -8 & 8 & 5 \end{pmatrix}$$

with

$$p_A(\lambda) = p_B(\lambda) = \lambda^3 + 5\lambda^2 + 3\lambda - 9 = (\lambda - 1)(\lambda + 3)^2$$

so that the eigenvalues are

$\lambda = 1$: simple,

$\lambda = -3$: with $\text{algmult}_A(-3) = \text{algmult}_B(-3) = 2$.

Example ((cont.))

Eigenvectors for $\lambda = -3$, A:

$$\mathsf{A} + 3\mathsf{I} = \begin{pmatrix} 2 & -1 & -2 \\ 8 & -8 & -8 \\ -10 & 11 & 10 \end{pmatrix} \longrightarrow \begin{pmatrix} 2 & -1 & -2 \\ 0 & -4 & 0 \\ 0 & 6 & 0 \end{pmatrix}$$

$$\Longrightarrow \quad N(\mathsf{A} + 3\mathsf{I}) = \text{span}\left\{ \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} \right\}$$

$$\Longrightarrow \quad 1 = \text{geomult}_{\mathsf{A}}(-3) < \text{algmult}_{\mathsf{A}}(-3) = 2.$$

Example ((cont.))

Eigenvectors for $\lambda = -3$, B:

$$\mathsf{B} + 3\mathsf{I} = \begin{pmatrix} 4 & -4 & -4 \\ 8 & -8 & -8 \\ -8 & 8 & 8 \end{pmatrix}$$

$$\implies \quad N(\mathsf{B} + 3\mathsf{I}) = \operatorname{span}\left\{ \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \right\}$$

$$\implies \quad \operatorname{geomult}_{\mathsf{B}}(-3) = 2 = \operatorname{algmult}_{\mathsf{B}}(-3).$$

In general we can say

Theorem

Let $\mathsf{A} \in \mathbb{C}^{n \times n}$ *and* $\lambda \in \sigma(\mathsf{A})$. *Then*

$$\mathit{geomult}_{\mathsf{A}}(\lambda) \leq \mathit{algmult}_{\mathsf{A}}(\lambda).$$

In general we can say

Theorem

Let $\mathsf{A} \in \mathbb{C}^{n \times n}$ *and* $\lambda \in \sigma(\mathsf{A})$. *Then*

$$\textit{geomult}_{\mathsf{A}}(\lambda) \leq \textit{algmult}_{\mathsf{A}}(\lambda).$$

Proof

Let's assume that $\text{algmult}_{\mathsf{A}}(\lambda) = k$. If we apply the Schur factorization to A we get

$$\mathsf{U}^*\mathsf{A}\mathsf{U} = \begin{pmatrix} \mathsf{T}_{11} & \mathsf{T}_{12} \\ \mathsf{O} & \mathsf{T}_{22} \end{pmatrix},$$

where T_{11} is $k \times k$ upper triangular with $\text{diag}(\mathsf{T}_{11}) = (\lambda, \ldots, \lambda)$.

Proof (cont.)

Also, $\lambda \notin \operatorname{diag}(\mathsf{T}_{22})$ (where T_{22} is also upper triangular).
Thus $\lambda \notin \sigma(\mathsf{T}_{22})$ and

$$\mathsf{T}_{22} - \lambda\mathsf{I} \quad \text{is nonsingular,}$$

i.e., $\operatorname{rank}(\mathsf{T}_{22} - \lambda\mathsf{I}) = n - k$.
Now,

$$\operatorname{geomult}_{\mathsf{A}}(\lambda) = \dim N(\mathsf{A} - \lambda\mathsf{I}) = n - \operatorname{rank}(\mathsf{A} - \lambda\mathsf{I}).$$

But, using a unitary (and therefore nonsingular) U,

$$\begin{aligned} \operatorname{rank}(\mathsf{A} - \lambda\mathsf{I}) &= \operatorname{rank}(\mathsf{U}^*(\mathsf{A} - \lambda\mathsf{I})\mathsf{U}) \\ &= \operatorname{rank}\begin{pmatrix} \mathsf{T}_{11} - \lambda\mathsf{I} & \mathsf{T}_{12} \\ \mathsf{O} & \mathsf{T}_{22} - \lambda\mathsf{I} \end{pmatrix} \\ &\geq \operatorname{rank}(\mathsf{T}_{22} - \lambda\mathsf{I}) = n - k. \end{aligned}$$

Therefore

$$\operatorname{geomult}_{\mathsf{A}}(\lambda) \leq n - (n - k) = k = \operatorname{algmult}_{\mathsf{A}}(\lambda). \qquad \square$$

Diagonalizability

Theorem

A matrix $\mathsf{A} \in \mathbb{C}^{n \times n}$ *is diagonalizable if and only if*

$$\textit{geomult}_{\mathsf{A}}(\lambda) = \textit{algmult}_{\mathsf{A}}(\lambda) \quad \textit{for all } \lambda \in \sigma(\mathsf{A}),$$

i.e., if and only if every eigenvalue is semi-simple.

Diagonalizability

Theorem

A matrix $\mathsf{A} \in \mathbb{C}^{n \times n}$ *is diagonalizable if and only if*

$$\textit{geomult}_{\mathsf{A}}(\lambda) = \textit{algmult}_{\mathsf{A}}(\lambda) \quad \textit{for all } \lambda \in \sigma(\mathsf{A}),$$

i.e., if and only if every eigenvalue is semi-simple.

Remark

This provides another interpretation for defective matrices, i.e., a matrix is diagonalizable if and only if it is not defective.

Proof

"$\Longleftarrow$": Assume $\text{geomult}_A(\lambda_i) = \text{algmult}_A(\lambda_i) = a_i$ for all i.
Furthermore, assume we have k distinct eigenvalues, i.e.,

$$\sigma(A) = \{\lambda_1, \ldots, \lambda_k\}.$$

Take $\mathcal{B}_i$ as a basis for $N(A - \lambda_i I)$, then

$$\mathcal{B} = \bigcup_{i=1}^{k} \mathcal{B}_i$$

consists of $\sum_{i=1}^{k} a_i = n$ vectors.
Moreover, $\mathcal{B}$ is linearly independent (see HW), and it forms a complete set of eigenvectors so that A is diagonalizable.

Proof (cont.)

"$\Longrightarrow$": Assume A is diagonalizable with λ such that $\text{algmult}_A(\lambda) = a$. Then

$$P^{-1}AP = D = \begin{pmatrix} \lambda I_{a\times a} & O \\ O & B \end{pmatrix},$$

where P is nonsingular and B is diagonal with $\lambda \notin B$.

As above,

$$\text{geomult}_A(\lambda) = \dim N(A - \lambda I) = n - \text{rank}(A - \lambda I).$$

However,

$$\begin{aligned} \text{rank}(A - \lambda I) &= \text{rank}(P(D - \lambda I)P{-}1) \\ &= \text{rank}\begin{pmatrix} O & O \\ O & B - \lambda I \end{pmatrix} = n - a. \end{aligned}$$

Together,

$$\text{geomult}_A(\lambda) = n - (n - a) = \text{algmult}_A(\lambda).$$

$\square$

Corollary

If all eigenvalues of A *are simple, then* A *is diagonalizable.*

Corollary

If all eigenvalues of A *are simple, then* A *is diagonalizable.*

Remark

The converse is not true. Our earlier example showed that B *is diagonalizable since* $\sigma(\mathsf{B}) = \{-3, 1\}$ *with*

$$\begin{aligned} geomult_{\mathsf{B}}(-3) &= algmult_{\mathsf{B}}(-3) = 2 \\ geomult_{\mathsf{B}}(1) &= algmult_{\mathsf{B}}(1) = 1, \end{aligned}$$

but $\lambda = -3$ *is a double eigenvalue.*

Spectral Theorem

Theorem

A matrix $\mathrm{A} \in \mathbb{C}^{n\times n}$ *with* $\sigma(\mathrm{A}) = \{\lambda_1, \ldots, \lambda_k\}$ *is diagonalizable if and only if there exist spectral projectors* G_i, $i = 1, \ldots, k$ *such that we have the spectral decomposition*

$$\mathrm{A} = \lambda_1 \mathrm{G}_1 + \lambda_2 \mathrm{G}_2 + \ldots + \lambda_k \mathrm{G}_k,$$

where the G_i *satisfy*

1. $\mathrm{G}_1 + \mathrm{G}_2 + \ldots + \mathrm{G}_k = \mathrm{I}$,
2. $\mathrm{G}_i \mathrm{G}_j = \mathrm{O}$, $i \neq j$,
3. G_i *is a projector onto* $N(\mathrm{A} - \lambda_i \mathrm{I})$ *along* $R(\mathrm{A} - \lambda_i \mathrm{I})$.

Proof

We discuss only "$\Longrightarrow$" for (1) and (2).

Assume A is diagonalizable, i.e., $A = PDP^{-1}$ with

$$P = \begin{pmatrix} X_1 & X_2 & \cdots & X_k \end{pmatrix},$$

where the columns of X_i form a basis for $N(A - \lambda_i I)$, i.e.,

$$A = \begin{pmatrix} X_1 & X_2 & \cdots & X_k \end{pmatrix} \begin{pmatrix} \lambda_1 I & & & O \\ & \lambda_2 I & & \\ & & \ddots & \\ O & & & \lambda_k I \end{pmatrix} \underbrace{\begin{pmatrix} Y_1^T \\ Y_2^T \\ \vdots \\ Y_k^T \end{pmatrix}}_{=P^{-1}}$$

Proof

We discuss only "$\Longrightarrow$" for (1) and (2).

Assume A is diagonalizable, i.e., $A = PDP^{-1}$ with

$$P = \begin{pmatrix} X_1 & X_2 & \cdots & X_k \end{pmatrix},$$

where the columns of X_i form a basis for $N(A - \lambda_i I)$, i.e.,

$$A = \begin{pmatrix} X_1 & X_2 & \cdots & X_k \end{pmatrix} \begin{pmatrix} \lambda_1 I & & & O \\ & \lambda_2 I & & \\ & & \ddots & \\ O & & & \lambda_k I \end{pmatrix} \underbrace{\begin{pmatrix} Y_1^T \\ Y_2^T \\ \vdots \\ Y_k^T \end{pmatrix}}_{=P^{-1}}$$

$$= \lambda_1 \underbrace{X_1 Y_1^T}_{=G_1} + \lambda_2 \underbrace{X_2 Y_2^T}_{=G_2} + \ldots + \lambda_k \underbrace{X_k Y_k^T}_{=G_k}.$$

Proof (cont.)

The identity

$$A = \lambda_1 G_1 + \lambda_2 G_2 + \ldots + \lambda_k G_k$$

is the spectral decomposition of A.

Proof (cont.)

The identity

$$A = \lambda_1 G_1 + \lambda_2 G_2 + \ldots + \lambda_k G_k$$

is the spectral decomposition of A.

If $\lambda_1 = \lambda_2 = \ldots = \lambda_k = 1$ then

$$PIP^{-1} = I = G_1 + G_2 + \ldots + G_k$$

and we have established (1).

Proof (cont.)

Moreover,

$$P^{-1}P = I \quad \Longleftrightarrow \quad \begin{pmatrix} Y_1^T X_1 & Y_1^T X_2 & \cdots & Y_1^T X_k \\ Y_2^T X_1 & Y_2^T X_2 & & \\ & & \ddots & \\ Y_k^T X_1 & & \cdots & Y_k^T X_k \end{pmatrix} = I$$

so that $Y_i^T X_j = \begin{cases} I, & i = j, \\ O, & i \neq j, \end{cases}$ and

Proof (cont.)

Moreover,

$$P^{-1}P = I \quad \Longleftrightarrow \quad \begin{pmatrix} Y_1^T X_1 & Y_1^T X_2 & \cdots & Y_1^T X_k \\ Y_2^T X_1 & Y_2^T X_2 & & \\ & & \ddots & \\ Y_k^T X_1 & & \cdots & Y_k^T X_k \end{pmatrix} = I$$

so that $Y_i^T X_j = \begin{cases} I, & i = j, \\ O, & i \neq j, \end{cases}$ and

$$G_i G_j$$

Proof (cont.)

Moreover,

$$P^{-1}P = I \iff \begin{pmatrix} Y_1^T X_1 & Y_1^T X_2 & \cdots & Y_1^T X_k \\ Y_2^T X_1 & Y_2^T X_2 & & \\ & & \ddots & \\ Y_k^T X_1 & & \cdots & Y_k^T X_k \end{pmatrix} = I$$

so that $Y_i^T X_j = \begin{cases} I, & i = j, \\ O, & i \neq j, \end{cases}$ and

$$G_i G_j = X_i \underbrace{Y_i^T X_j} Y_j^T$$

Proof (cont.)

Moreover,

$$\mathsf{P}^{-1}\mathsf{P} = \mathsf{I} \quad \Longleftrightarrow \quad \begin{pmatrix} \mathsf{Y}_1^T\mathsf{X}_1 & \mathsf{Y}_1^T\mathsf{X}_2 & \cdots & \mathsf{Y}_1^T\mathsf{X}_k \\ \mathsf{Y}_2^T\mathsf{X}_1 & \mathsf{Y}_2^T\mathsf{X}_2 & & \\ & & \ddots & \\ \mathsf{Y}_k^T\mathsf{X}_1 & & \cdots & \mathsf{Y}_k^T\mathsf{X}_k \end{pmatrix} = \mathsf{I}$$

so that $\mathsf{Y}_i^T\mathsf{X}_j = \begin{cases} \mathsf{I}, & i = j, \\ \mathsf{O}, & i \neq j, \end{cases}$ and

$$\mathsf{G}_i\mathsf{G}_j = \mathsf{X}_i \underbrace{\mathsf{Y}_i^T\mathsf{X}_j}_{=\delta_{ij}\mathsf{I}} \mathsf{Y}_j^T = \begin{cases} \mathsf{X}_i\mathsf{Y}_j^T, & i = j, \\ \mathsf{O}, & i \neq j. \end{cases}$$

Proof (cont.)

Moreover,

$$P^{-1}P = I \quad \Longleftrightarrow \quad \begin{pmatrix} Y_1^T X_1 & Y_1^T X_2 & \cdots & Y_1^T X_k \\ Y_2^T X_1 & Y_2^T X_2 & & \\ & & \ddots & \\ Y_k^T X_1 & & \cdots & Y_k^T X_k \end{pmatrix} = I$$

so that $Y_i^T X_j = \begin{cases} I, & i = j, \\ O, & i \neq j, \end{cases}$ and

$$G_i G_j = X_i \underbrace{Y_i^T X_j}_{=\delta_{ij} I} Y_j^T = \begin{cases} X_i Y_j^T, & i = j, \\ O, & i \neq j. \end{cases}$$

Thus $G_i^2 = G_i$ are projectors and we have established (2). $\square$

Remark

If λ_i is simple, then

$$\mathsf{G}_i = \frac{\boldsymbol{x}\boldsymbol{y}^*}{\boldsymbol{y}^*\boldsymbol{x}},$$

where $\boldsymbol{x}$, $\boldsymbol{y}^*$, respectively, are the right and left eigenvectors of A associated with λ_i.

Outline

Functions of Diagonalizable Matrices

We want to give meaning to

$$f(\mathsf{A}),$$

where

A: a square $n \times n$ matrix (below also diagonalizable),

f: a continuous function.

Functions of Diagonalizable Matrices

We want to give meaning to

$$f(\mathsf{A}),$$

where

A: a square $n \times n$ matrix (below also diagonalizable),

f: a continuous function.

Functions of matrices play an important role, e.g., in solving systems of ODEs.

Functions of Diagonalizable Matrices

We want to give meaning to

$$f(\mathsf{A}),$$

where

A: a square $n \times n$ matrix (below also diagonalizable),

f: a continuous function.

Functions of matrices play an important role, e.g., in solving systems of ODEs.

One possible approach is to use infinite series, such as

$$e^{\mathsf{A}} = \sum_{k=0}^{\infty} \frac{\mathsf{A}^k}{k!}.$$

Functions of Diagonalizable Matrices

We want to give meaning to

$$f(\mathsf{A}),$$

where

- A: a square $n \times n$ matrix (below also diagonalizable),
- f: a continuous function.

Functions of matrices play an important role, e.g., in solving systems of ODEs.

One possible approach is to use infinite series, such as

$$e^{\mathsf{A}} = \sum_{k=0}^{\infty} \frac{\mathsf{A}^k}{k!}.$$

However, it is not so easy to compute this series in practice (see, e.g., [MVL78, MVL03]) or to analyze the convergence of such types of series.

If A is diagonalizable then the series are easier to analyze:

If A is diagonalizable then the series are easier to analyze:

Recall: A diagonalizable means that there exists a nonsingular P such that

$$\mathsf{P}^{-1}\mathsf{A}\mathsf{P} = \mathsf{D} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n),$$

where the eigenvalues $\lambda_1, \ldots, \lambda_n$ need not be distinct.

If A is diagonalizable then the series are easier to analyze:

Recall: A diagonalizable means that there exists a nonsingular P such that

$$\mathsf{P}^{-1}\mathsf{A}\mathsf{P} = \mathsf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n),$$

where the eigenvalues $\lambda_1, \ldots, \lambda_n$ need not be distinct.

Moreover, from HW 11 we know that

$$\mathsf{P}^{-1}\mathsf{A}^k\mathsf{P} = \operatorname{diag}(\lambda_1^k, \ldots, \lambda_n^k) = \mathsf{D}^k.$$

If A is diagonalizable then the series are easier to analyze:

Recall: A diagonalizable means that there exists a nonsingular P such that

$$\mathsf{P}^{-1}\mathsf{A}\mathsf{P} = \mathsf{D} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n),$$

where the eigenvalues $\lambda_1, \ldots, \lambda_n$ need not be distinct.

Moreover, from HW 11 we know that

$$\mathsf{P}^{-1}\mathsf{A}^k\mathsf{P} = \mathrm{diag}(\lambda_1^k, \ldots, \lambda_n^k) = \mathsf{D}^k.$$

With this setup we can represent $f(\mathsf{A})$ as a power series in A.

$$f(A) = \sum_{k=0}^{\infty} c_k A^k$$

$$f(\mathsf{A}) = \sum_{k=0}^{\infty} c_k \mathsf{A}^k$$
$$= \sum_{k=0}^{\infty} c_k \mathsf{P}\mathsf{D}^k\mathsf{P}^{-1}$$

$$
\begin{aligned}
f(\mathsf{A}) &= \sum_{k=0}^{\infty} c_k \mathsf{A}^k \\
&= \sum_{k=0}^{\infty} c_k \mathsf{P}\mathsf{D}^k\mathsf{P}^{-1} = \mathsf{P}\left(\sum_{k=0}^{\infty} c_k \mathsf{D}^k\right)\mathsf{P}^{-1}
\end{aligned}
$$

$$
\begin{aligned}
f(\mathsf{A}) &= \sum_{k=0}^{\infty} c_k \mathsf{A}^k \\
&= \sum_{k=0}^{\infty} c_k \mathsf{P}\mathsf{D}^k\mathsf{P}^{-1} = \mathsf{P}\left(\sum_{k=0}^{\infty} c_k \mathsf{D}^k\right)\mathsf{P}^{-1} \\
&= \mathsf{P}\left(\sum_{k=0}^{\infty} c_k \operatorname{diag}(\lambda_1^k, \ldots, \lambda_n^k)\right)\mathsf{P}^{-1}
\end{aligned}
$$

$$
\begin{aligned}
f(\mathsf{A}) &= \sum_{k=0}^{\infty} c_k \mathsf{A}^k \\
&= \sum_{k=0}^{\infty} c_k \mathsf{P}\mathsf{D}^k\mathsf{P}^{-1} = \mathsf{P}\left(\sum_{k=0}^{\infty} c_k \mathsf{D}^k\right)\mathsf{P}^{-1} \\
&= \mathsf{P}\left(\sum_{k=0}^{\infty} c_k \operatorname{diag}(\lambda_1^k, \ldots, \lambda_n^k)\right)\mathsf{P}^{-1} \\
&= \mathsf{P}\operatorname{diag}\left(\sum_{k=0}^{\infty} c_k \lambda_1^k, \ldots, \sum_{k=0}^{\infty} c_k \lambda_n^k\right)\mathsf{P}^{-1}
\end{aligned}
$$

$$
\begin{aligned}
f(\mathsf{A}) &= \sum_{k=0}^{\infty} c_k \mathsf{A}^k \\
&= \sum_{k=0}^{\infty} c_k \mathsf{P}\mathsf{D}^k\mathsf{P}^{-1} = \mathsf{P}\left(\sum_{k=0}^{\infty} c_k \mathsf{D}^k\right)\mathsf{P}^{-1} \\
&= \mathsf{P}\left(\sum_{k=0}^{\infty} c_k \operatorname{diag}(\lambda_1^k, \ldots, \lambda_n^k)\right)\mathsf{P}^{-1} \\
&= \mathsf{P}\operatorname{diag}\left(\sum_{k=0}^{\infty} c_k \lambda_1^k, \ldots, \sum_{k=0}^{\infty} c_k \lambda_n^k\right)\mathsf{P}^{-1} \\
&= \mathsf{P}\operatorname{diag}\left(f(\lambda_1), \ldots, f(\lambda_n)\right)\mathsf{P}^{-1}
\end{aligned}
$$

$$
\begin{aligned}
f(\mathsf{A}) &= \sum_{k=0}^{\infty} c_k \mathsf{A}^k \\
&= \sum_{k=0}^{\infty} c_k \mathsf{P}\mathsf{D}^k\mathsf{P}^{-1} = \mathsf{P}\left(\sum_{k=0}^{\infty} c_k \mathsf{D}^k\right)\mathsf{P}^{-1} \\
&= \mathsf{P}\left(\sum_{k=0}^{\infty} c_k \operatorname{diag}(\lambda_1^k, \ldots, \lambda_n^k)\right)\mathsf{P}^{-1} \\
&= \mathsf{P}\operatorname{diag}\left(\sum_{k=0}^{\infty} c_k \lambda_1^k, \ldots, \sum_{k=0}^{\infty} c_k \lambda_n^k\right)\mathsf{P}^{-1} \\
&= \mathsf{P}\operatorname{diag}\left(f(\lambda_1), \ldots, f(\lambda_n)\right)\mathsf{P}^{-1} \\
&= \mathsf{P}f(\mathsf{D})\mathsf{P}^{-1}
\end{aligned}
$$

$$
\begin{aligned}
f(\mathsf{A}) &= \sum_{k=0}^{\infty} c_k \mathsf{A}^k \\
&= \sum_{k=0}^{\infty} c_k \mathsf{P}\mathsf{D}^k\mathsf{P}^{-1} = \mathsf{P}\left(\sum_{k=0}^{\infty} c_k \mathsf{D}^k\right)\mathsf{P}^{-1} \\
&= \mathsf{P}\left(\sum_{k=0}^{\infty} c_k \operatorname{diag}(\lambda_1^k, \ldots, \lambda_n^k)\right)\mathsf{P}^{-1} \\
&= \mathsf{P}\operatorname{diag}\left(\sum_{k=0}^{\infty} c_k \lambda_1^k, \ldots, \sum_{k=0}^{\infty} c_k \lambda_n^k\right)\mathsf{P}^{-1} \\
&= \mathsf{P}\operatorname{diag}\left(f(\lambda_1), \ldots, f(\lambda_n)\right)\mathsf{P}^{-1} \\
&= \mathsf{P} f(\mathsf{D})\mathsf{P}^{-1}
\end{aligned}
$$

Note how the matrix power series now has become a diagonal matrix of regular (scalar) power series in the eigenvalues of A.

Thus we can now define $f(\mathsf{A})$, A diagonalizable, as

$$\begin{aligned} f(\mathsf{A}) &= \mathsf{P} f(\mathsf{D}) \mathsf{P}^{-1} \\ &= \mathsf{P}\,\mathrm{diag}(f(\lambda_1), \ldots, f(\lambda_n)) \mathsf{P}^{-1}. \end{aligned}$$

Thus we can now define $f(\mathsf{A})$, A diagonalizable, as

$$\begin{aligned} f(\mathsf{A}) &= \mathsf{P} f(\mathsf{D}) \mathsf{P}^{-1} \\ &= \mathsf{P}\,\mathrm{diag}(f(\lambda_1), \ldots, f(\lambda_n))\mathsf{P}^{-1}. \end{aligned}$$

The advantage of this approach is that we have no problems analyzing convergence of the series (this is now standard calculus).

Thus we can now define $f(A)$, A diagonalizable, as

$$\begin{aligned} f(A) &= Pf(D)P^{-1} \\ &= P\,\mathrm{diag}(f(\lambda_1), \ldots, f(\lambda_n))P^{-1}. \end{aligned}$$

The advantage of this approach is that we have no problems analyzing convergence of the series (this is now standard calculus).

However, now there is a potential problem with uniqueness since P is not unique.

To understand the uniqueness issue we look more carefully and write

$$f(\mathsf{A}) = \mathsf{P}f(\mathsf{D})\mathsf{P}^{-1}$$

To understand the uniqueness issue we look more carefully and write

$$\begin{aligned} f(\mathsf{A}) &= \mathsf{P} f(\mathsf{D}) \mathsf{P}^{-1} \\ &= \begin{pmatrix} \mathsf{X}_1 & \cdots & \mathsf{X}_n \end{pmatrix} \begin{pmatrix} f(\lambda_1)\mathsf{I} & & \\ & \ddots & \\ & & f(\lambda_n)\mathsf{I} \end{pmatrix} \begin{pmatrix} \mathsf{Y}_1^T \\ \vdots \\ \mathsf{Y}_n^T \end{pmatrix} \end{aligned}$$

To understand the uniqueness issue we look more carefully and write

$$\begin{aligned} f(\mathsf{A}) &= \mathsf{P}f(\mathsf{D})\mathsf{P}^{-1} \\ &= \begin{pmatrix} \mathsf{X}_1 & \cdots & \mathsf{X}_n \end{pmatrix} \begin{pmatrix} f(\lambda_1)\mathsf{I} & & \\ & \ddots & \\ & & f(\lambda_n)\mathsf{I} \end{pmatrix} \begin{pmatrix} \mathsf{Y}_1^T \\ \vdots \\ \mathsf{Y}_n^T \end{pmatrix} \\ &= \sum_{i=1}^{n} f(\lambda_i)\mathsf{X}_i\mathsf{Y}_i^T \end{aligned}$$

To understand the uniqueness issue we look more carefully and write

$$
\begin{aligned}
f(\mathsf{A}) &= \mathsf{P} f(\mathsf{D}) \mathsf{P}^{-1} \\
&= \begin{pmatrix} \mathsf{X}_1 & \cdots & \mathsf{X}_n \end{pmatrix} \begin{pmatrix} f(\lambda_1)\mathsf{I} & & \\ & \ddots & \\ & & f(\lambda_n)\mathsf{I} \end{pmatrix} \begin{pmatrix} \mathsf{Y}_1^T \\ \vdots \\ \mathsf{Y}_n^T \end{pmatrix} \\
&= \sum_{i=1}^{n} f(\lambda_i) \mathsf{X}_i \mathsf{Y}_i^T \\
&= \sum_{i=1}^{n} f(\lambda_i) \mathsf{G}_i,
\end{aligned}
$$

To understand the uniqueness issue we look more carefully and write

$$
\begin{aligned}
f(\mathsf{A}) &= \mathsf{P} f(\mathsf{D}) \mathsf{P}^{-1} \\
&= \begin{pmatrix} \mathsf{X}_1 & \cdots & \mathsf{X}_n \end{pmatrix} \begin{pmatrix} f(\lambda_1)\mathsf{I} & & \\ & \ddots & \\ & & f(\lambda_n)\mathsf{I} \end{pmatrix} \begin{pmatrix} \mathsf{Y}_1^T \\ \vdots \\ \mathsf{Y}_n^T \end{pmatrix} \\
&= \sum_{i=1}^{n} f(\lambda_i) \mathsf{X}_i \mathsf{Y}_i^T \\
&= \sum_{i=1}^{n} f(\lambda_i) \mathsf{G}_i,
\end{aligned}
$$

where the spectral projectors G_i are unique.

To understand the uniqueness issue we look more carefully and write

$$
\begin{aligned}
f(\mathsf{A}) &= \mathsf{P} f(\mathsf{D}) \mathsf{P}^{-1} \\
&= \begin{pmatrix} \mathsf{X}_1 & \cdots & \mathsf{X}_n \end{pmatrix} \begin{pmatrix} f(\lambda_1)\mathsf{I} & & \\ & \ddots & \\ & & f(\lambda_n)\mathsf{I} \end{pmatrix} \begin{pmatrix} \mathsf{Y}_1^T \\ \vdots \\ \mathsf{Y}_n^T \end{pmatrix} \\
&= \sum_{i=1}^n f(\lambda_i) \mathsf{X}_i \mathsf{Y}_i^T \\
&= \sum_{i=1}^n f(\lambda_i) \mathsf{G}_i,
\end{aligned}
$$

where the spectral projectors G_i are unique.

Remark

Note how the spectral theorem helps us convert the problem from one with an infinite series to a single finite sum of length n.

The representation

$$f(\mathsf{A}) = \sum_{i=1}^{n} f(\lambda_i)\mathsf{G}_i$$

implies that any function of a diagonalizable matrix A is a polynomial in A.

The representation

$$f(\mathsf{A}) = \sum_{i=1}^{n} f(\lambda_i)\mathsf{G}_i$$

implies that any function of a diagonalizable matrix A is a polynomial in A.

To see this, we construct $p(\lambda_i) = f(\lambda_i)$, i.e., we construct a Lagrange interpolating polynomial to f at the eigenvalues of A:

The representation

$$f(\mathsf{A}) = \sum_{i=1}^{n} f(\lambda_i)\mathsf{G}_i$$

implies that any function of a diagonalizable matrix A is a polynomial in A.

To see this, we construct $p(\lambda_i) = f(\lambda_i)$, i.e., we construct a Lagrange interpolating polynomial to f at the eigenvalues of A:

$$p(z) = \sum_{i=1}^{n} f(\lambda_i)L_i(z)$$

$$\text{with } L_i(z) = \prod_{\substack{j=1\\ j\neq i}}^{n}(z-\lambda_j) \Big/ \prod_{\substack{j=1\\ j\neq i}}^{n}(\lambda_i-\lambda_j).$$

Thus,

$$f(\mathsf{A}) = \sum_{i=1}^{n} f(\lambda_i)\mathsf{G}_i$$

Thus,

$$f(\mathsf{A}) = \sum_{i=1}^{n} f(\lambda_i)\mathsf{G}_i$$
$$= \sum_{i=1}^{n} p(\lambda_i)\mathsf{G}_i = p(\mathsf{A}).$$

Thus,

$$\begin{aligned} f(\mathsf{A}) &= \sum_{i=1}^{n} f(\lambda_i)\mathsf{G}_i \\ &= \sum_{i=1}^{n} p(\lambda_i)\mathsf{G}_i = p(\mathsf{A}). \end{aligned}$$

On the other hand,

$$p(\mathsf{A}) = \sum_{i=1}^{n} f(\lambda_i) L_i(\mathsf{A})$$

Thus,

$$\begin{aligned} f(\mathsf{A}) &= \sum_{i=1}^{n} f(\lambda_i)\mathsf{G}_i \\ &= \sum_{i=1}^{n} p(\lambda_i)\mathsf{G}_i = p(\mathsf{A}). \end{aligned}$$

On the other hand,

$$p(\mathsf{A}) = \sum_{i=1}^{n} f(\lambda_i) L_i(\mathsf{A})$$

and we see that

$$\mathsf{G}_i = L_i(\mathsf{A}) = \prod_{\substack{j=1 \\ j\neq i}}^{n} (\mathsf{A} - \lambda_j \mathsf{I}) / \prod_{\substack{j=1 \\ j\neq i}}^{n} (\lambda_i - \lambda_j).$$

Remark

- *In fact, $f(\mathsf{A})$ is a polynomial in A for any square A (see HW — uses Cayley–Hamilton theorem).*

Remark

- *In fact, $f(\mathsf{A})$ is a polynomial in A for any square A (see HW — uses Cayley–Hamilton theorem).*
- *Moreover, for general (square) A we can always define $f(\mathsf{A})$ via an infinite series. Then one can prove*

Theorem

If $f(z) = \sum_{k=1}^{\infty} c_k (z - z_0)^k$ converges for $|z - z_0| < r$ and $|\lambda_i - z_0| < r$ for all $\lambda_i \in \sigma(\mathsf{A})$, then

$$f(\mathsf{A}) = \sum_{k=0}^{\infty} c_k (\mathsf{A} - z_0 \mathsf{I})^k.$$

The power method to compute the largest eigenvalue of A

Consider a matrix $A \in \mathbb{C}^{n \times n}$ with eigenvalues

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|,$$

i.e., A has a dominant (real) eigenvalue.

The power method to compute the largest eigenvalue of A

Consider a matrix $A \in \mathbb{C}^{n \times n}$ with eigenvalues

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|,$$

i.e., A has a dominant (real) eigenvalue.

Note that λ_1 is real since if it were complex, then we would also have $\overline{\lambda_1}$ with $|\overline{\lambda_1}| = |\lambda_1|$, so not dominant.

The power method to compute the largest eigenvalue of A

Consider a matrix $A \in \mathbb{C}^{n \times n}$ with eigenvalues

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|,$$

i.e., A has a dominant (real) eigenvalue.

Note that λ_1 is real since if it were complex, then we would also have $\overline{\lambda_1}$ with $|\overline{\lambda_1}| = |\lambda_1|$, so not dominant.

We now describe a numerical method to find λ_1 and explain how it can be viewed in the framework of this section.

Consider $f(z) = \left(\frac{z}{\lambda_1}\right)^k$. Then

$$f(\mathsf{A}) = \left(\frac{\mathsf{A}}{\lambda_1}\right)^k$$

Consider $f(z) = \left(\frac{z}{\lambda_1}\right)^k$. Then

$$\begin{aligned} f(\mathsf{A}) &= \left(\frac{\mathsf{A}}{\lambda_1}\right)^k \\ &= \sum_{i=1}^{n} f(\lambda_i)\mathsf{G}_i \end{aligned}$$

Consider $f(z) = \left(\frac{z}{\lambda_1}\right)^k$. Then

$$\begin{aligned} f(\mathsf{A}) &= \left(\frac{\mathsf{A}}{\lambda_1}\right)^k \\ &= \sum_{i=1}^{n} f(\lambda_i)\mathsf{G}_i \\ &= \sum_{i=1}^{n} \left(\frac{\lambda_i}{\lambda_1}\right)^k \mathsf{G}_i \end{aligned}$$

Consider $f(z) = \left(\frac{z}{\lambda_1}\right)^k$. Then

$$
\begin{aligned}
f(\mathsf{A}) &= \left(\frac{\mathsf{A}}{\lambda_1}\right)^k \\
&= \sum_{i=1}^n f(\lambda_i)\mathsf{G}_i \\
&= \sum_{i=1}^n \left(\frac{\lambda_i}{\lambda_1}\right)^k \mathsf{G}_i \\
&= \mathsf{G}_1 + \underbrace{\left(\frac{\lambda_2}{\lambda_1}\right)^k}_{\to 0}\mathsf{G}_2 + \ldots + \underbrace{\left(\frac{\lambda_n}{\lambda_1}\right)^k}_{\to 0}\mathsf{G}_n \to \mathsf{G}_1 \quad \text{for } k \to \infty.
\end{aligned}
$$

Consider $f(z) = \left(\frac{z}{\lambda_1}\right)^k$. Then

$$
\begin{aligned}
f(\mathsf{A}) &= \left(\frac{\mathsf{A}}{\lambda_1}\right)^k \\
&= \sum_{i=1}^{n} f(\lambda_i)\mathsf{G}_i \\
&= \sum_{i=1}^{n} \left(\frac{\lambda_i}{\lambda_1}\right)^k \mathsf{G}_i \\
&= \mathsf{G}_1 + \underbrace{\left(\frac{\lambda_2}{\lambda_1}\right)^k}_{\to 0} \mathsf{G}_2 + \ldots + \underbrace{\left(\frac{\lambda_n}{\lambda_1}\right)^k}_{\to 0} \mathsf{G}_n \to \mathsf{G}_1 \quad \text{for } k \to \infty.
\end{aligned}
$$

Therefore

$$
\left(\frac{\mathsf{A}}{\lambda_1}\right)^k \boldsymbol{x}_0 \to \mathsf{G}_1 \boldsymbol{x}_0 \in N(\mathsf{A} - \lambda_1 \mathsf{I})
$$

since G_1 is a projector onto $N(\mathsf{A} - \lambda_1 \mathsf{I})$.

Thus any initial vector $\boldsymbol{x}_0$ such that $\mathsf{G}_1\boldsymbol{x}_0 \neq \mathbf{0}$ (i.e., $\boldsymbol{x}_0 \notin R(\mathsf{A} - \lambda_1 \mathsf{I})$) will converge to an eigenvector of A associated with λ_1 via the iteration

$$\frac{\mathsf{A}^k \boldsymbol{x}_0}{\lambda_1^k}, \quad k = 1, 2, \ldots$$

In fact, $\mathsf{A}^k \boldsymbol{x}_0$ converges to the first eigenvector, as does any scalar multiple.

Thus any initial vector $\boldsymbol{x}_0$ such that $G_1\boldsymbol{x}_0 \neq \mathbf{0}$ (i.e., $\boldsymbol{x}_0 \notin R(A - \lambda_1 I)$) will converge to an eigenvector of A associated with λ_1 via the iteration

$$\frac{A^k\boldsymbol{x}_0}{\lambda_1^k}, \quad k = 1, 2, \ldots$$

In fact, $A^k\boldsymbol{x}_0$ converges to the first eigenvector, as does any scalar multiple.

To find the eigenvalue λ_1 one iterates for $k = 0, 1, 2, \ldots$

$$\boldsymbol{y}^{(k)} = A\boldsymbol{x}^{(k)}, \quad \nu^{(k)} = \mathrm{maxcomp}(\boldsymbol{y}^{(k)}), \quad \boldsymbol{x}^{(k+1)} = \frac{\boldsymbol{y}^{(k)}}{\nu^{(k)}}.$$

Thus any initial vector $\boldsymbol{x}_0$ such that $G_1\boldsymbol{x}_0 \neq \boldsymbol{0}$ (i.e., $\boldsymbol{x}_0 \notin R(A - \lambda_1 I)$) will converge to an eigenvector of A associated with λ_1 via the iteration

$$\frac{A^k \boldsymbol{x}_0}{\lambda_1^k}, \quad k = 1, 2, \ldots$$

In fact, $A^k\boldsymbol{x}_0$ converges to the first eigenvector, as does any scalar multiple.
To find the eigenvalue λ_1 one iterates for $k = 0, 1, 2, \ldots$

$$\boldsymbol{y}^{(k)} = A\boldsymbol{x}^{(k)}, \quad \nu^{(k)} = \text{maxcomp}(\boldsymbol{y}^{(k)}), \quad \boldsymbol{x}^{(k+1)} = \frac{\boldsymbol{y}^{(k)}}{\nu^{(k)}}.$$

In fact, $\nu^{(k)} \to \lambda_1$ since

$$\underbrace{A\boldsymbol{x}^{(k+1)}}_{\to A\boldsymbol{x}_1 = \lambda_1\boldsymbol{x}_1} = A\frac{\boldsymbol{y}^{(k)}}{\nu^{(k)}} = \underbrace{A^2\boldsymbol{x}^{(k)}}_{\to A^2\boldsymbol{x}_1 = \lambda_1^2\boldsymbol{x}_1} / \nu^{(k)}.$$

Remark

More details of the power method — as well as several other methods for finding eigenvalues — are discussed in MATH 577.

Outline

Normal Matrices

Consider an $n \times n$ matrix A. We know that

- A is diagonalizable (in the sense of similarity) if and only if A is nondefective, and
- A is unitarily similar to a triangular matrix (Schur).

Normal Matrices

Consider an $n \times n$ matrix A. We know that

- A is diagonalizable (in the sense of similarity) if and only if A is nondefective, and
- A is unitarily similar to a triangular matrix (Schur).

Question: What are the conditions on A such that it is unitarily diagonalizable?

Definition

A matrix $A \in \mathbb{C}^{n \times n}$ is called normal if

$$A^*A = AA^*.$$

Definition

A matrix $A \in \mathbb{C}^{n \times n}$ is called normal if

$$A^*A = AA^*.$$

Theorem

The matrix $A \in \mathbb{C}^{n \times n}$ is unitarily diagonalizable if and only if it is normal.

Proof (only easy direction).

Assume A is unitarily diagonalizable, i.e., there exists a unitary U such that

$$\mathsf{U}^*\mathsf{A}\mathsf{U} = \mathsf{D} \quad \Longleftrightarrow \quad \mathsf{A} = \mathsf{U}\mathsf{D}\mathsf{U}^*, \; \mathsf{A}^* = \mathsf{U}\overline{\mathsf{D}}\mathsf{U}^*.$$

Proof (only easy direction).

Assume A is unitarily diagonalizable, i.e., there exists a unitary U such that

$$\mathsf{U}^*\mathsf{A}\mathsf{U} = \mathsf{D} \quad \iff \quad \mathsf{A} = \mathsf{U}\mathsf{D}\mathsf{U}^*, \; \mathsf{A}^* = \mathsf{U}\overline{\mathsf{D}}\mathsf{U}^*.$$

Then

$$\mathsf{A}^*\mathsf{A} = \mathsf{U}\overline{\mathsf{D}}\underbrace{\mathsf{U}^*\mathsf{U}}_{=\mathsf{I}}\mathsf{D}\mathsf{U}^*,$$

$\square$

Proof (only easy direction).

Assume A is unitarily diagonalizable, i.e., there exists a unitary U such that

$$\mathsf{U}^*\mathsf{A}\mathsf{U} = \mathsf{D} \quad \iff \quad \mathsf{A} = \mathsf{U}\mathsf{D}\mathsf{U}^*, \; \mathsf{A}^* = \mathsf{U}\overline{\mathsf{D}}\mathsf{U}^*.$$

Then

$$\begin{aligned} \mathsf{A}^*\mathsf{A} &= \mathsf{U}\overline{\mathsf{D}}\underbrace{\mathsf{U}^*\mathsf{U}}_{=\mathsf{I}}\mathsf{D}\mathsf{U}^*, \\ \mathsf{A}\mathsf{A}^* &= \mathsf{U}\mathsf{D}\underbrace{\mathsf{U}^*\mathsf{U}}_{=\mathsf{I}}\overline{\mathsf{D}}\mathsf{U}^*. \end{aligned}$$

□

Proof (only easy direction).

Assume A is unitarily diagonalizable, i.e., there exists a unitary U such that

$$U^*AU = D \quad \iff \quad A = UDU^*, \ A^* = U\overline{D}U^*.$$

Then

$$A^*A = U\overline{D}\underbrace{U^*U}_{=I}DU^*,$$

$$AA^* = UD\underbrace{U^*U}_{=I}\overline{D}U^*.$$

Since

$$\overline{D}D = \sum_{i=1}^{n} |d_i|^2 = D\overline{D}$$

we have $A^*A = AA^*$ and A is normal. $\square$

Remark

- *Normal matrices are unitarily diagonalizable, i.e., they have an associated complete set of orthogonal eigenvectors.*

Remark

- *Normal matrices are unitarily diagonalizable, i.e., they have an associated complete set of orthogonal eigenvectors.*
- *However, not all complete sets of eigenvectors of normal matrices are orthogonal (see HW).*

Theorem

Let A *be normal with* $\sigma(\mathsf{A}) = \{\lambda_1, \ldots, \lambda_k\}$. *Then*

1. $R(\mathsf{A}) \perp N(\mathsf{A})$.
2. *Eigenvectors to distinct eigenvalues are orthogonal, i.e.,*

$$N(\mathsf{A} - \lambda_i \mathsf{I}) \perp N(\mathsf{A} - \lambda_j \mathsf{I}), \quad \lambda_i \neq \lambda_j.$$

3. *The spectral projectors* G_i *are orthogonal projectors.*

Proof

1. We know

$$\begin{aligned} N(\mathsf{A}^*\mathsf{A}) &= N(\mathsf{A}), \quad N(\mathsf{A}\mathsf{A}^*) = N(\mathsf{A}^*), \\ R(\mathsf{A})^\perp &= N(\mathsf{A}^*). \end{aligned}$$

Proof

1. We know

$$N(A^*A) = N(A), \quad N(AA^*) = N(A^*),$$
$$R(A)^\perp = N(A^*).$$

Since A is normal we know $N(A^*) = N(A)$ and the statement follows.

Proof

1. We know

$$N(A^*A) = N(A), \quad N(AA^*) = N(A^*),$$
$$R(A)^\perp = N(A^*).$$

   Since A is normal we know $N(A^*) = N(A)$ and the statement follows.

2. From above we know that $R(A)^\perp = N(A^*) = N(A)$ whenever A is normal.

Proof

1. We know

$$N(\mathsf{A}^*\mathsf{A}) = N(\mathsf{A}), \quad N(\mathsf{A}\mathsf{A}^*) = N(\mathsf{A}^*),$$
$$R(\mathsf{A})^\perp = N(\mathsf{A}^*).$$

Since A is normal we know $N(\mathsf{A}^*) = N(\mathsf{A})$ and the statement follows.

2. From above we know that $R(\mathsf{A})^\perp = N(\mathsf{A}^*) = N(\mathsf{A})$ whenever A is normal.
Moreover, $\mathsf{A} - \lambda\mathsf{I}$ is also normal since

$$(\mathsf{A} - \lambda\mathsf{I})^*(\mathsf{A} - \lambda\mathsf{I}) = \mathsf{A}^*\mathsf{A} - \lambda\mathsf{A}^* - \overline{\lambda}\mathsf{A} + |\lambda|^2\mathsf{I},$$
$$(\mathsf{A} - \lambda\mathsf{I})(\mathsf{A} - \lambda\mathsf{I})^* = \mathsf{A}\mathsf{A}^* - \overline{\lambda}\mathsf{A} - \lambda\mathsf{A}^* + |\lambda|^2\mathsf{I}.$$

Therefore,

$$N(\mathsf{A} - \lambda\mathsf{I}) = N\left((\mathsf{A} - \lambda\mathsf{I})^*\right) = N\left(\mathsf{A}^* - \overline{\lambda}\mathsf{I}\right).$$

Proof (cont.)

We also have

$$\lambda \in \sigma(A) \quad \Longleftrightarrow \quad \overline{\lambda} \in \sigma(A^*)$$

since

$$\begin{aligned} \det(A - \lambda I) = 0 \quad &\Longleftrightarrow \quad \overline{\det(A - \lambda I)} = 0 \\ &\overset{\overline{\det(A)} = \det(A^*)}{\Longleftrightarrow} \quad \det\left((A - \lambda I)^*\right) = 0 \\ &\Longleftrightarrow \quad \det(A^* - \overline{\lambda} I) = 0. \end{aligned}$$

Proof (cont.)

So we can consider two eigenpairs $(\lambda_i, \boldsymbol{x}_i)$ and $(\lambda_j, \boldsymbol{x}_j)$ of A.

Proof (cont.)

So we can consider two eigenpairs $(\lambda_i, \boldsymbol{x}_i)$ and $(\lambda_j, \boldsymbol{x}_j)$ of A. Conjugate transposition yields

$$\mathsf{A}\boldsymbol{x}_j = \lambda_j \boldsymbol{x}_j \quad \Longleftrightarrow \quad \boldsymbol{x}_j^* \mathsf{A}^* = \overline{\lambda_j} \boldsymbol{x}_j^*,$$

Proof (cont.)

So we can consider two eigenpairs $(\lambda_i, \boldsymbol{x}_i)$ and $(\lambda_j, \boldsymbol{x}_j)$ of A.
Conjugate transposition yields

$$\mathsf{A}\boldsymbol{x}_j = \lambda_j \boldsymbol{x}_j \quad \Longleftrightarrow \quad \boldsymbol{x}_j^* \mathsf{A}^* = \overline{\lambda_j} \boldsymbol{x}_j^*,$$

and from above this is equivalent to

$$\boldsymbol{x}_j^* \mathsf{A} = \lambda_j \boldsymbol{x}_j^*.$$

Proof (cont.)

So we can consider two eigenpairs $(\lambda_i, \boldsymbol{x}_i)$ and $(\lambda_j, \boldsymbol{x}_j)$ of A.
Conjugate transposition yields

$$\mathsf{A}\boldsymbol{x}_j = \lambda_j \boldsymbol{x}_j \quad \Longleftrightarrow \quad \boldsymbol{x}_j^* \mathsf{A}^* = \overline{\lambda_j} \boldsymbol{x}_j^*,$$

and from above this is equivalent to

$$\boldsymbol{x}_j^* \mathsf{A} = \lambda_j \boldsymbol{x}_j^*.$$

Now we multiply by $\boldsymbol{x}_i$

$$\boldsymbol{x}_j^* \underbrace{\mathsf{A}\boldsymbol{x}_i}_{=\lambda_i \boldsymbol{x}_i} = \lambda_j \boldsymbol{x}_j^* \boldsymbol{x}_i \quad \Longleftrightarrow \quad \lambda_i \boldsymbol{x}_j^* \boldsymbol{x}_i = \lambda_j \boldsymbol{x}_j^* \boldsymbol{x}_i$$

Proof (cont.)

So we can consider two eigenpairs $(\lambda_i, \boldsymbol{x}_i)$ and $(\lambda_j, \boldsymbol{x}_j)$ of A. Conjugate transposition yields

$$\mathsf{A}\boldsymbol{x}_j = \lambda_j \boldsymbol{x}_j \quad \Longleftrightarrow \quad \boldsymbol{x}_j^* \mathsf{A}^* = \overline{\lambda_j} \boldsymbol{x}_j^*,$$

and from above this is equivalent to

$$\boldsymbol{x}_j^* \mathsf{A} = \lambda_j \boldsymbol{x}_j^*.$$

Now we multiply by $\boldsymbol{x}_i$

$$\boldsymbol{x}_j^* \underbrace{\mathsf{A}\boldsymbol{x}_i}_{=\lambda_i \boldsymbol{x}_i} = \lambda_j \boldsymbol{x}_j^* \boldsymbol{x}_i \quad \Longleftrightarrow \quad \lambda_i \boldsymbol{x}_j^* \boldsymbol{x}_i = \lambda_j \boldsymbol{x}_j^* \boldsymbol{x}_i$$

$$\overset{\lambda_i \neq \lambda_j}{\Longleftrightarrow} \quad \boldsymbol{x}_j^* \boldsymbol{x}_i = 0.$$

Proof (cont.)

3. The spectral theorem states that the G_i are projectors onto $N(A - \lambda_i I)$ along $R(A - \lambda_i I)$.

Proof (cont.)

3. The spectral theorem states that the G_i are projectors onto $N(A - \lambda_i I)$ along $R(A - \lambda_i I)$.

   Above we showed that
   - $A - \lambda_i I$ is normal provided A is normal, and
   - $R(A)^{\perp} = N(A)$ whenever A is normal.

Proof (cont.)

3. The spectral theorem states that the G_i are projectors onto $N(A - \lambda_i I)$ along $R(A - \lambda_i I)$.

   Above we showed that

   - $A - \lambda_i I$ is normal provided A is normal, and
   - $R(A)^{\perp} = N(A)$ whenever A is normal.

   Therefore

$$R(A - \lambda_i I)^{\perp} = N(A - \lambda_i I)$$

   and G_i are orthogonal projectors. $\square$

Remark

- *Normal matrices include*
    - *real symmetric, Hermitian, skew-symmetric, skew-Hermitian, orthogonal, and unitary matrices.*

Remark

- *Normal matrices include*
  - *real symmetric, Hermitian, skew-symmetric, skew-Hermitian, orthogonal, and unitary matrices.*
- *All eigenvalues of Hermitian (or real symmetric) matrices are real:*

Remark

- *Normal matrices include*
  - *real symmetric, Hermitian, skew-symmetric, skew-Hermitian, orthogonal, and unitary matrices.*
- *All eigenvalues of Hermitian (or real symmetric) matrices are real:* *First,*

$$\mathsf{A}\boldsymbol{x} = \lambda\boldsymbol{x} \quad \Longleftrightarrow \quad \boldsymbol{x}^*\mathsf{A}^* = \overline{\lambda}\boldsymbol{x}^*.$$

Remark

- *Normal matrices include*
  - *real symmetric, Hermitian, skew-symmetric, skew-Hermitian, orthogonal, and unitary matrices.*
- *All eigenvalues of Hermitian (or real symmetric) matrices are real:* *First,*

$$\mathsf{A}\boldsymbol{x} = \lambda\boldsymbol{x} \quad \Longleftrightarrow \quad \boldsymbol{x}^*\mathsf{A}^* = \overline{\lambda}\boldsymbol{x}^*.$$

Multiply by $\boldsymbol{x}^*$ *and* $\boldsymbol{x}$*, respectively:*

$$\boldsymbol{x}^*\mathsf{A}\boldsymbol{x} = \lambda\boldsymbol{x}^*\boldsymbol{x} \quad \Longleftrightarrow \quad \boldsymbol{x}^*\mathsf{A}^*\boldsymbol{x} = \overline{\lambda}\boldsymbol{x}^*\boldsymbol{x}.$$

Remark

- *Normal matrices include*
  - *real symmetric, Hermitian, skew-symmetric, skew-Hermitian, orthogonal, and unitary matrices.*
- *All eigenvalues of Hermitian (or real symmetric) matrices are real: First,*

$$\mathsf{A}\boldsymbol{x} = \lambda \boldsymbol{x} \quad \Longleftrightarrow \quad \boldsymbol{x}^* \mathsf{A}^* = \overline{\lambda} \boldsymbol{x}^*.$$

Multiply by $\boldsymbol{x}^*$ *and* $\boldsymbol{x}$*, respectively:*

$$\boldsymbol{x}^* \mathsf{A}\boldsymbol{x} = \lambda \boldsymbol{x}^* \boldsymbol{x} \quad \Longleftrightarrow \quad \boldsymbol{x}^* \mathsf{A}^* \boldsymbol{x} = \overline{\lambda} \boldsymbol{x}^* \boldsymbol{x}.$$

Then, since $\mathsf{A}^* = \mathsf{A}$*,*

$$\lambda \boldsymbol{x}^* \boldsymbol{x} = \overline{\lambda} \boldsymbol{x}^* \boldsymbol{x} \quad \overset{\boldsymbol{x} \neq 0}{\Longleftrightarrow} \quad \lambda = \overline{\lambda}.$$

Moreover, one can show

Theorem

A *is real symmetric if and only if* A *is orthogonally diagonalizable, i.e.,*

$$P^T A P = D,$$

where P *is orthogonal and* D *is real.*

Rayleigh quotient

Definition

Let $A \in \mathbb{C}^{n\times n}$ and $\boldsymbol{x} \in \mathbb{C}^n$. Then

$$r(\boldsymbol{x}) = \frac{\boldsymbol{x}^* A \boldsymbol{x}}{\boldsymbol{x}^* \boldsymbol{x}}$$

is called the Rayleigh quotient of A associated with $\boldsymbol{x}$.

Rayleigh quotient

Definition

Let $A \in \mathbb{C}^{n \times n}$ and $\boldsymbol{x} \in \mathbb{C}^n$. Then

$$r(\boldsymbol{x}) = \frac{\boldsymbol{x}^* A \boldsymbol{x}}{\boldsymbol{x}^* \boldsymbol{x}}$$

is called the Rayleigh quotient of A associated with $\boldsymbol{x}$.

Remark

If $\boldsymbol{x}$ is an eigenvector of A then $r(\boldsymbol{x}) = \lambda$, the associated eigenvalue, i.e.,

$$A\boldsymbol{x} = \lambda \boldsymbol{x} \quad \Longrightarrow \quad \boldsymbol{x}^* A \boldsymbol{x} = \lambda \boldsymbol{x}^* \boldsymbol{x} \quad \Longleftrightarrow \quad r(\boldsymbol{x}) = \lambda.$$

Theorem

Let $\mathsf{A} \in \mathbb{C}^{n \times n}$ *be Hermitian. Then*

$$\lambda_{max} = \max_{\boldsymbol{x} \neq 0} r(\boldsymbol{x}), \qquad \lambda_{min} = \min_{\boldsymbol{x} \neq 0} r(\boldsymbol{x}).$$

Theorem

Let $A \in \mathbb{C}^{n \times n}$ *be Hermitian. Then*

$$\lambda_{max} = \max_{\boldsymbol{x} \neq 0} r(\boldsymbol{x}), \qquad \lambda_{min} = \min_{\boldsymbol{x} \neq 0} r(\boldsymbol{x}).$$

Remark

Since the eigenvalues of a Hermitian matrix are real they can indeed be ordered.

Theorem

Let $\mathsf{A} \in \mathbb{C}^{n \times n}$ *be Hermitian. Then*

$$\lambda_{max} = \max_{\boldsymbol{x} \neq 0} r(\boldsymbol{x}), \qquad \lambda_{min} = \min_{\boldsymbol{x} \neq 0} r(\boldsymbol{x}).$$

Remark

Since the eigenvalues of a Hermitian matrix are real they can indeed be ordered.

Proof (Only for the maximum eigenvalue).

First, we consider an equivalent formulation:

$$\lambda_{\max} = \max_{\|\boldsymbol{x}\|_2 = 1} \boldsymbol{x}^* \mathsf{A} \boldsymbol{x}.$$

Proof (cont.)

Now, since A is Hermitian, A is normal and therefore unitarily diagonalizable so that

$$\max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^* \mathsf{A} \boldsymbol{x} = \max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^* \mathsf{U}\mathsf{D}\mathsf{U}^* \boldsymbol{x}.$$

Proof (cont.)

Now, since A is Hermitian, A is normal and therefore unitarily diagonalizable so that

$$\max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^* \mathsf{A} \boldsymbol{x} = \max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^* \mathsf{U}\mathsf{D}\mathsf{U}^* \boldsymbol{x}.$$

Let $\boldsymbol{y} = \mathsf{U}^* \boldsymbol{x}$. Then

$$\|\boldsymbol{y}\|_2 = \|\mathsf{U}^* \boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2$$

and

Proof (cont.)

Now, since A is Hermitian, A is normal and therefore unitarily diagonalizable so that

$$\max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^*\mathsf{A}\boldsymbol{x} = \max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^*\mathsf{UDU}^*\boldsymbol{x}.$$

Let $\boldsymbol{y} = \mathsf{U}^*\boldsymbol{x}$. Then

$$\|\boldsymbol{y}\|_2 = \|\mathsf{U}^*\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2$$

and

$$\max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^*\mathsf{A}\boldsymbol{x} = \max_{\|\boldsymbol{y}\|_2=1} \boldsymbol{y}^*\mathsf{D}\boldsymbol{y}$$

Proof (cont.)

Now, since A is Hermitian, A is normal and therefore unitarily diagonalizable so that

$$\max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^* \mathsf{A} \boldsymbol{x} = \max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^* \mathsf{U} \mathsf{D} \mathsf{U}^* \boldsymbol{x}.$$

Let $\boldsymbol{y} = \mathsf{U}^* \boldsymbol{x}$. Then

$$\|\boldsymbol{y}\|_2 = \|\mathsf{U}^* \boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2$$

and

$$\begin{aligned} \max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^* \mathsf{A} \boldsymbol{x} &= \max_{\|\boldsymbol{y}\|_2=1} \boldsymbol{y}^* \mathsf{D} \boldsymbol{y} \\ &= \max_{\|\boldsymbol{y}\|_2=1} \sum_{i=1}^{n} \lambda_i |y_i|^2 \end{aligned}$$

Proof (cont.)

Now, since A is Hermitian, A is normal and therefore unitarily diagonalizable so that

$$\max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^* \mathsf{A} \boldsymbol{x} = \max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^* \mathsf{U}\mathsf{D}\mathsf{U}^* \boldsymbol{x}.$$

Let $\boldsymbol{y} = \mathsf{U}^* \boldsymbol{x}$. Then

$$\|\boldsymbol{y}\|_2 = \|\mathsf{U}^* \boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2$$

and

$$\begin{aligned}
\max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^* \mathsf{A} \boldsymbol{x} &= \max_{\|\boldsymbol{y}\|_2=1} \boldsymbol{y}^* \mathsf{D} \boldsymbol{y} \\
&= \max_{\|\boldsymbol{y}\|_2=1} \sum_{i=1}^{n} \lambda_i |y_i|^2 \\
&\le \lambda_{\max} \underbrace{\max_{\|\boldsymbol{y}\|_2=1} \sum_{i=1}^{n} |y_i|^2}_{=\|\boldsymbol{y}\|_2^2}
\end{aligned}$$

Proof (cont.)

Now, since A is Hermitian, A is normal and therefore unitarily diagonalizable so that

$$\max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^*\mathsf{A}\boldsymbol{x} = \max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^*\mathsf{UDU}^*\boldsymbol{x}.$$

Let $\boldsymbol{y} = \mathsf{U}^*\boldsymbol{x}$. Then

$$\|\boldsymbol{y}\|_2 = \|\mathsf{U}^*\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2$$

and

$$\begin{aligned}\max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^*\mathsf{A}\boldsymbol{x} &= \max_{\|\boldsymbol{y}\|_2=1} \boldsymbol{y}^*\mathsf{D}\boldsymbol{y} \\ &= \max_{\|\boldsymbol{y}\|_2=1} \sum_{i=1}^{n} \lambda_i |y_i|^2 \\ &\le \lambda_{\max} \underbrace{\max_{\|\boldsymbol{y}\|_2=1} \sum_{i=1}^{n} |y_i|^2}_{=\|\boldsymbol{y}\|_2^2} = \lambda_{\max}.\end{aligned}$$

Proof (cont.)

However, the upper bound can be achieved by making $\boldsymbol{x}$ a normalized eigenvector for $\lambda_{\max}$. Then

$$\boldsymbol{x}^* A \boldsymbol{x} = \boldsymbol{x}^* \lambda_{\max} \boldsymbol{x} = \lambda_{\max} \underbrace{\|\boldsymbol{x}\|_2^2}_{=1} = \lambda_{\max}.$$

So the claim is true. $\square$

As a generalization one can prove

Theorem (Courant–Fischer Theorem)

Let A *be an* $n \times n$ *Hermitian matrix. Its eigenvalues* $\lambda_{max} = \lambda_1 \geq \lambda_2 \leq \ldots \leq \lambda_n = \lambda_{min}$ *are given by*

$$\lambda_i = \max_{\dim \mathcal{V} = i} \min_{\substack{\boldsymbol{x} \in \mathcal{V} \\ \|\boldsymbol{x}\|_2 = 1}} \boldsymbol{x}^* \mathsf{A} \boldsymbol{x}$$

or

$$\lambda_i = \min_{\dim \mathcal{V} = n-i+1} \max_{\substack{\boldsymbol{x} \in \mathcal{V} \\ \|\boldsymbol{x}\|_2 = 1}} \boldsymbol{x}^* \mathsf{A} \boldsymbol{x}.$$

As a generalization one can prove

Theorem (Courant–Fischer Theorem)

Let A *be an* $n \times n$ *Hermitian matrix. Its eigenvalues* $\lambda_{max} = \lambda_1 \geq \lambda_2 \leq \ldots \leq \lambda_n = \lambda_{min}$ *are given by*

$$\lambda_i = \max_{\dim \mathcal{V} = i} \min_{\substack{\boldsymbol{x} \in \mathcal{V} \\ \|\boldsymbol{x}\|_2 = 1}} \boldsymbol{x}^* A \boldsymbol{x}$$

or

$$\lambda_i = \min_{\dim \mathcal{V} = n-i+1} \max_{\substack{\boldsymbol{x} \in \mathcal{V} \\ \|\boldsymbol{x}\|_2 = 1}} \boldsymbol{x}^* A \boldsymbol{x}.$$

Remark

- *Here* $\mathcal{V}$ *is a subspace of* $\mathbb{C}^n$.
- $i = n$ *in the max-min characterization leads to* $\mathcal{V} = \mathbb{C}^n$ *and* λ_{min}.
- $i = 1$ *in the min-max characterization leads to* $\mathcal{V} = \mathbb{C}^n$ *and* λ_{min}.

Remark

Since the singular values of A *are the square roots of the eigenvalues of* $\mathsf{A}^*\mathsf{A}$ *an analogous theorem holds for the singular values of* A *(see [Mey00, p. 555] for more details).*

Remark

Since the singular values of A *are the square roots of the eigenvalues of* A^*A *an analogous theorem holds for the singular values of* A *(see [Mey00, p. 555] for more details).*

In particular,

$$\sigma_{max} = \max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^* A^* A \boldsymbol{x} = \max_{\|\boldsymbol{x}\|_2=1} \|A\boldsymbol{x}\|_2 = \|A\|_2.$$

Outline

Positive Definite Matrices

Earlier we saw that if $\mathsf{A} \in \mathbb{R}^{n\times n}$ is symmetric, then

$$\mathsf{P}^T\mathsf{A}\mathsf{P} = \mathsf{D},$$

where P is an orthogonal matrix of eigenvectors and D is a real diagonal matrix of eigenvalues.

Positive Definite Matrices

Earlier we saw that if $\mathsf{A} \in \mathbb{R}^{n\times n}$ is symmetric, then

$$\mathsf{P}^T \mathsf{A} \mathsf{P} = \mathsf{D},$$

where P is an orthogonal matrix of eigenvectors and D is a real diagonal matrix of eigenvalues.

Question: What additional properties of A will ensure that its eigenvalues are all positive (nonnegative)?

A necessary condition

Let's assume that $\lambda_i \geq 0$, $i = 1, \ldots, n$. Then

$$\begin{aligned} \mathsf{D} &= \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \\ &= \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}) \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}) = \mathsf{D}^{1/2}\mathsf{D}^{1/2}. \end{aligned}$$

A necessary condition

Let's assume that $\lambda_i \geq 0$, $i = 1, \ldots, n$. Then

$$\begin{aligned} \mathsf{D} &= \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \\ &= \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}) \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}) = \mathsf{D}^{1/2}\mathsf{D}^{1/2}. \end{aligned}$$

So

$$\mathsf{A} = \mathsf{P}\mathsf{D}\mathsf{P}^T = \mathsf{P}\mathsf{D}^{1/2}\mathsf{D}^{1/2}\mathsf{P}^T =$$

A necessary condition

Let's assume that $\lambda_i \geq 0$, $i = 1, \ldots, n$. Then

$$\begin{aligned} \mathsf{D} &= \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \\ &= \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}) \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}) = \mathsf{D}^{1/2}\mathsf{D}^{1/2}. \end{aligned}$$

So

$$\mathsf{A} = \mathsf{P}\mathsf{D}\mathsf{P}^T = \mathsf{P}\mathsf{D}^{1/2}\mathsf{D}^{1/2}\mathsf{P}^T = \mathsf{B}^T\mathsf{B},$$

where $\mathsf{B} = \mathsf{D}^{1/2}\mathsf{P}^T$.

A necessary condition

Let's assume that $\lambda_i \geq 0$, $i = 1, \ldots, n$. Then

$$
\begin{aligned}
\mathsf{D} &= \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \\
&= \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}) \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}) = \mathsf{D}^{1/2}\mathsf{D}^{1/2}.
\end{aligned}
$$

So

$$
\mathsf{A} = \mathsf{PDP}^T = \mathsf{PD}^{1/2}\mathsf{D}^{1/2}\mathsf{P}^T = \mathsf{B}^T\mathsf{B},
$$

where $\mathsf{B} = \mathsf{D}^{1/2}\mathsf{P}^T$.

Moreover, $\lambda_i > 0$, $i = 1, \ldots, n$, implies D is nonsingular, and therefore B is nonsingular.

A necessary condition

Let's assume that $\lambda_i \geq 0$, $i = 1, \ldots, n$. Then

$$\begin{aligned} D &= \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \\ &= \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}) \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}) = D^{1/2} D^{1/2}. \end{aligned}$$

So

$$A = PDP^T = PD^{1/2}D^{1/2}P^T = B^T B,$$

where $B = D^{1/2}P^T$.

Moreover, $\lambda_i > 0$, $i = 1, \ldots, n$, implies D is nonsingular, and therefore B is nonsingular.

The converse is also true, i.e., if B nonsingular, then $\lambda_i > 0$ (since $D^{1/2} = BP$ and P orthogonal).

A sufficient condition

Having a factorization

$$A = B^T B$$

is also sufficient:

A sufficient condition

Having a factorization

$$A = B^T B$$

is also sufficient:

Assume $(\lambda, \boldsymbol{x})$ is an eigenpair of A. Then the Rayleigh quotient shows

$$\lambda = \frac{\boldsymbol{x}^T A \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}}$$

A sufficient condition

Having a factorization

$$\mathsf{A} = \mathsf{B}^T \mathsf{B}$$

is also sufficient:

Assume $(\lambda, \boldsymbol{x})$ is an eigenpair of A. Then the Rayleigh quotient shows

$$\lambda = \frac{\boldsymbol{x}^T \mathsf{A} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} = \frac{\boldsymbol{x}^T \mathsf{B}^T \mathsf{B} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}}$$

A sufficient condition

Having a factorization

$$A = B^T B$$

is also sufficient:

Assume $(\lambda, \boldsymbol{x})$ is an eigenpair of A. Then the Rayleigh quotient shows

$$\lambda = \frac{\boldsymbol{x}^T A \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} = \frac{\boldsymbol{x}^T B^T B \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} = \frac{\|B\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2} \geq 0.$$

A sufficient condition

Having a factorization

$$A = B^T B$$

is also sufficient:

Assume $(\lambda, \boldsymbol{x})$ is an eigenpair of A. Then the Rayleigh quotient shows

$$\lambda = \frac{\boldsymbol{x}^T A \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} = \frac{\boldsymbol{x}^T B^T B \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}} = \frac{\|B\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2} \geq 0.$$

Moreover, if B is nonsingular, then $N(B) = \{\boldsymbol{0}\}$ so that $B\boldsymbol{x} \neq \boldsymbol{0}$ and $\lambda > 0$.

A sufficient condition

Having a factorization

$$\mathsf{A} = \mathsf{B}^T\mathsf{B}$$

is also sufficient:

Assume $(\lambda, \boldsymbol{x})$ is an eigenpair of A. Then the Rayleigh quotient shows

$$\lambda = \frac{\boldsymbol{x}^T\mathsf{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} = \frac{\boldsymbol{x}^T\mathsf{B}^T\mathsf{B}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} = \frac{\|\mathsf{B}\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2} \geq 0.$$

Moreover, if B is nonsingular, then $N(\mathsf{B}) = \{\mathbf{0}\}$ so that $\mathsf{B}\boldsymbol{x} \neq \mathbf{0}$ and $\lambda > 0$.
Conversely, if $\lambda > 0$, then $\mathsf{B}\boldsymbol{x} \neq \mathbf{0}$, and — if $\boldsymbol{x} \neq \mathbf{0}$ —then B is nonsingular.

Remark

On slide #78 of Chapter 3 we defined:
A symmetric matrix A *is positive definite if it has an LU decomposition with positive pivots, i.e.,*

$$A = LDL^T = R^T R,$$

where $R = D^{1/2}L^T$ *is the upper triangular Cholesky factor of* A.

Remark

On slide #78 of Chapter 3 we defined:
A symmetric matrix A *is positive definite if it has an LU decomposition with positive pivots, i.e.,*

$$\mathsf{A} = \mathsf{L}\mathsf{D}\mathsf{L}^T = \mathsf{R}^T\mathsf{R},$$

where $\mathsf{R} = \mathsf{D}^{1/2}\mathsf{L}^T$ *is the upper triangular Cholesky factor of* A*.*

This agrees with our discussion above.

Theorem

A real symmetric matrix A *is* positive definite *if and only if any of the following equivalent conditions hold:*

1. A *has an LU factorization with positive pivots, or* A *has a Cholesky factorization* $A = R^T R$ *with upper triangular matrix* R *with positive diagonal entries.*
2. *All eigenvalues of* A *are positive.*
3. $\boldsymbol{x}^T A \boldsymbol{x} > 0$ *for all nonzero* $\boldsymbol{x} \in \mathbb{R}^n$.

Remark

- *Earlier we used (1) as the definition of positive definiteness. Often positive definiteness is defined via (3).*

Remark

- *Earlier we used (1) as the definition of positive definiteness. Often positive definiteness is defined via (3).*
- *For a Hermitian matrix* A *we replace the transpose* ^{T} *by conjugate transpose* * *and "real"" by "complex".*

Remark

- *Earlier we used (1) as the definition of positive definiteness. Often positive definiteness is defined via (3).*
- *For a Hermitian matrix* A *we replace the transpose* ^{T} *by conjugate transpose* * *and "real"" by "complex".*
- *A few more criteria are listed in [Mey00]. In particular, all principal minors of* A *must be positive. Therefore, if* A *has a nonpositive diagonal entry, then it can't be positive definite.*

Finally,

Definition

Let A be a real symmetric matrix. If

$$\boldsymbol{x}^T \mathsf{A} \boldsymbol{x} \geq 0$$

for all $\boldsymbol{x} \in \mathbb{R}^n$, then A is called positive semidefinite.

Finally,

Definition

Let A be a real symmetric matrix. If

$$\boldsymbol{x}^T \mathsf{A}\boldsymbol{x} \geq 0$$

for all $\boldsymbol{x} \in \mathbb{R}^n$, then A is called positive semidefinite.

Theorem

A is positive semidefinite if and only if all eigenvalues of A are nonnegative.

Finally,

Definition

Let A be a real symmetric matrix. If

$$\boldsymbol{x}^T A \boldsymbol{x} \geq 0$$

for all $\boldsymbol{x} \in \mathbb{R}^n$, then A is called positive semidefinite.

Theorem

A is positive semidefinite if and only if all eigenvalues of A are nonnegative.

Remark

A few more criteria are listed in [Mey00].

Positive definite matrices in applications

- Gram matrix in interpolation/least squares approximation:

$$A_{ij} = \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$$

where $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\} \subseteq \mathcal{V}$, $\mathcal{V}$ some inner product space.

Positive definite matrices in applications

- Gram matrix in interpolation/least squares approximation:

$$\mathsf{A}_{ij} = \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$$

where $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\} \subseteq \mathcal{V}$, $\mathcal{V}$ some inner product space.

If the $\boldsymbol{v}_i$ are linearly independent, then A is positive definite; otherwise positive semidefinite.

Positive definite matrices in applications

- Gram matrix in interpolation/least squares approximation:

$$\mathsf{A}_{ij} = \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$$

where $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\} \subseteq \mathcal{V}$, $\mathcal{V}$ some inner product space.

If the $\boldsymbol{v}_i$ are linearly independent, then A is positive definite; otherwise positive semidefinite.

  - If $\boldsymbol{v}_i$ are the columns of some matrix V, then $\mathsf{A} = \mathsf{V}^T\mathsf{V}$ is the matrix of the normal equations $\mathsf{V}^T\mathsf{V}\boldsymbol{x} = \mathsf{V}^T\boldsymbol{b}$.

Positive definite matrices in applications

- Gram matrix in interpolation/least squares approximation:

$$A_{ij} = \langle \boldsymbol{v}_i, \boldsymbol{v}_j \rangle$$

where $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\} \subseteq \mathcal{V}$, $\mathcal{V}$ some inner product space.

If the $\boldsymbol{v}_i$ are linearly independent, then A is positive definite; otherwise positive semidefinite.

  - If $\boldsymbol{v}_i$ are the columns of some matrix V, then $A = V^T V$ is the matrix of the normal equations $V^T V \boldsymbol{x} = V^T \boldsymbol{b}$.
  - If $\boldsymbol{v}_i = K(\cdot, \boldsymbol{x}_i)$ is a (reproducing) kernel function centered at $\boldsymbol{x}_i$, then $A_{ij} = \langle K(\cdot, \boldsymbol{x}_i), K(\cdot, \boldsymbol{x}_j) \rangle_{\mathcal{H}_K} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$. This is the matrix that appears in kriging and RBF interpolation.

- Hessian matrix in optimization:

- Hessian matrix in optimization: Start with *n*-dimensional Taylor theorem:

$$f(\boldsymbol{x}) = f(\boldsymbol{z}) + \sum_{i=1}^{n}(x_i - z_i)\frac{\partial f}{\partial x_i}(\boldsymbol{z}) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - z_i)(x_j - z_j)\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{z}) + \ldots$$

- Hessian matrix in optimization: Start with *n*-dimensional Taylor theorem:

$$
\begin{aligned}
f(\boldsymbol{x}) &= f(\boldsymbol{z}) + \sum_{i=1}^{n}(x_i - z_i)\frac{\partial f}{\partial x_i}(\boldsymbol{z}) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - z_i)(x_j - z_j)\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{z}) + \ldots \\
&= f(\boldsymbol{z}) + (\boldsymbol{x} - \boldsymbol{z})^T \nabla_f(\boldsymbol{z}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{z})^T \mathsf{H}_f(\boldsymbol{z})(\boldsymbol{x} - \boldsymbol{z}) + \ldots,
\end{aligned}
$$

where ∇_f is the gradient of f and H_f is its Hessian matrix.

- Hessian matrix in optimization: Start with n-dimensional Taylor theorem:

$$f(\boldsymbol{x}) = f(\boldsymbol{z}) + \sum_{i=1}^{n}(x_i - z_i)\frac{\partial f}{\partial x_i}(\boldsymbol{z}) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - z_i)(x_j - z_j)\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{z}) + \ldots$$
$$= f(\boldsymbol{z}) + (\boldsymbol{x} - \boldsymbol{z})^T \nabla_f(\boldsymbol{z}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{z})^T \mathrm{H}_f(\boldsymbol{z})(\boldsymbol{x} - \boldsymbol{z}) + \ldots,$$

where ∇_f is the gradient of f and H_f is its Hessian matrix.

From calculus it is known that convexity/concavity at a critical point $\boldsymbol{z}$, i.e., $\nabla_f(\boldsymbol{z}) = \boldsymbol{0}$, can be determined by the Hessian matrix. In fact,

- **Hessian matrix** in optimization: Start with *n-dimensional Taylor theorem*:

$$f(\boldsymbol{x}) = f(\boldsymbol{z}) + \sum_{i=1}^{n}(x_i - z_i)\frac{\partial f}{\partial x_i}(\boldsymbol{z}) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - z_i)(x_j - z_j)\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{z}) + \ldots$$

$$= f(\boldsymbol{z}) + (\boldsymbol{x} - \boldsymbol{z})^T \nabla_f(\boldsymbol{z}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{z})^T \mathrm{H}_f(\boldsymbol{z})(\boldsymbol{x} - \boldsymbol{z}) + \ldots,$$

where ∇_f is the gradient of f and H_f is its Hessian matrix.

From calculus it is known that convexity/concavity at a critical point $\boldsymbol{z}$, i.e., $\nabla_f(\boldsymbol{z}) = \mathbf{0}$, can be determined by the Hessian matrix. In fact,

  - If $\mathrm{H}_f(\boldsymbol{z})$ is positive definite, then f has a minimum at $\boldsymbol{z}$.

- Hessian matrix in optimization: Start with n-dimensional Taylor theorem:

$$f(\boldsymbol{x}) = f(\boldsymbol{z}) + \sum_{i=1}^{n}(x_i - z_i)\frac{\partial f}{\partial x_i}(\boldsymbol{z}) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - z_i)(x_j - z_j)\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{z}) + \ldots$$

$$= f(\boldsymbol{z}) + (\boldsymbol{x} - \boldsymbol{z})^T \nabla_f(\boldsymbol{z}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{z})^T \mathrm{H}_f(\boldsymbol{z})(\boldsymbol{x} - \boldsymbol{z}) + \ldots,$$

where ∇_f is the gradient of f and H_f is its Hessian matrix.

From calculus it is known that convexity/concavity at a critical point $\boldsymbol{z}$, i.e., $\nabla_f(\boldsymbol{z}) = \boldsymbol{0}$, can be determined by the Hessian matrix. In fact,

  - If $\mathrm{H}_f(\boldsymbol{z})$ is positive definite, then f has a minimum at $\boldsymbol{z}$.
  - If $\mathrm{H}_f(\boldsymbol{z})$ is negative definite, then f has a maximum at $\boldsymbol{z}$.

- Hessian matrix in optimization: Start with n-dimensional Taylor theorem:

$$f(\boldsymbol{x}) = f(\boldsymbol{z}) + \sum_{i=1}^{n}(x_i - z_i)\frac{\partial f}{\partial x_i}(\boldsymbol{z}) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - z_i)(x_j - z_j)\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{z}) + \ldots$$

$$= f(\boldsymbol{z}) + (\boldsymbol{x} - \boldsymbol{z})^T \nabla_f(\boldsymbol{z}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{z})^T \mathsf{H}_f(\boldsymbol{z})(\boldsymbol{x} - \boldsymbol{z}) + \ldots,$$

where ∇_f is the gradient of f and H_f is its Hessian matrix.

From calculus it is known that convexity/concavity at a critical point $\boldsymbol{z}$, i.e., $\nabla_f(\boldsymbol{z}) = \boldsymbol{0}$, can be determined by the Hessian matrix. In fact,

  - If $\mathsf{H}_f(\boldsymbol{z})$ is positive definite, then f has a minimum at $\boldsymbol{z}$.
  - If $\mathsf{H}_f(\boldsymbol{z})$ is negative definite, then f has a maximum at $\boldsymbol{z}$.

Moreover, if $\mathsf{H}_f(\boldsymbol{z})$ is positive semidefinite for all points in the domain of f, then f is a convex function.

- Covariance matrix in probability/statistics:

- Covariance matrix in probability/statistics: Let $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ be a vector of random variables with mean $\mu_i = \mathbb{E}[X_i]$, $i = 1, \ldots, n$. Then the covariance matrix of $\boldsymbol{X}$ is given by

$$A_{ij} = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$$

- Covariance matrix in probability/statistics: Let $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ be a vector of random variables with mean $\mu_i = \mathbb{E}[X_i]$, $i = 1, \ldots, n$. Then the covariance matrix of $\boldsymbol{X}$ is given by

$$A_{ij} = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$$

We can see that A is positive semidefinite:

$$\boldsymbol{z}^T \mathsf{A} \boldsymbol{z} = \mathbb{E}\left[\sum_{i=1}^{n} \sum_{j=1}^{n} z_i (X_i - \mu_i)(X_j - \mu_j) z_j\right]$$

- Covariance matrix in probability/statistics: Let $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ be a vector of random variables with mean $\mu_i = \mathbb{E}[X_i]$, $i = 1, \ldots, n$. Then the covariance matrix of $\boldsymbol{X}$ is given by

$$A_{ij} = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$$

We can see that A is positive semidefinite:

$$\begin{aligned} \boldsymbol{z}^T \mathsf{A} \boldsymbol{z} &= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} z_i(X_i - \mu_i)(X_j - \mu_j) z_j\right] \\ &= \mathbb{E}\left[\left(\sum_{i=1}^{n} z_i(X_i - \mu_i)\right)^2\right] \end{aligned}$$

- Covariance matrix in probability/statistics: Let $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ be a vector of random variables with mean $\mu_i = \mathbb{E}[X_i]$, $i = 1, \ldots, n$. Then the covariance matrix of $\boldsymbol{X}$ is given by

$$A_{ij} = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$$

We can see that A is positive semidefinite:

$$\begin{aligned} \boldsymbol{z}^T \mathsf{A} \boldsymbol{z} &= \mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^n z_i (X_i - \mu_i)(X_j - \mu_j) z_j\right] \\ &= \mathbb{E}\left[\left(\sum_{i=1}^n z_i (X_i - \mu_i)\right)^2\right] \geq 0. \end{aligned}$$

- Finite difference matrices: See, e.g., [Mey00, Example 7.6.2].

- Finite difference matrices: See, e.g., [Mey00, Example 7.6.2].
- "Stiffness" matrices: in finite element formulations, based on the interpretation of energy of some state $\boldsymbol{x}$ as a quadratic form $\boldsymbol{x}^T \mathsf{A}\boldsymbol{x}$. Positive energy (a fundamental physical assumption) means positive definite A.

  More details in MATH 581.

Quadratic forms

Definition

Let $\mathsf{A} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{x} \in \mathbb{R}^n$. The scalar function

$$f(\boldsymbol{x}) = \boldsymbol{x}^T \mathsf{A} \boldsymbol{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

is called a quadratic form.

The quadratic form $\boldsymbol{x}^T \mathsf{A} \boldsymbol{x}$ is called positive definite if the matrix A is positive definite.

Remark

We always assume that the matrix of a quadratic form is symmetric:

Remark

We always assume that the matrix of a quadratic form is symmetric:

Even if A *is not symmetric,* $\frac{\mathsf{A}+\mathsf{A}^T}{2}$ *always is symmetric.*

Remark

We always assume that the matrix of a quadratic form is symmetric:

Even if A *is not symmetric,* $\frac{A+A^T}{2}$ *always is symmetric.*

And we have for the quadratic form

$$\boldsymbol{x}^T \left(\frac{A + A^T}{2}\right) \boldsymbol{x} = \frac{1}{2}\boldsymbol{x}^T A \boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^T A^T \boldsymbol{x}$$

Remark

We always assume that the matrix of a quadratic form is symmetric:

Even if A *is not symmetric,* $\frac{A+A^T}{2}$ *always is symmetric.*

And we have for the quadratic form

$$\boldsymbol{x}^T\left(\frac{A+A^T}{2}\right)\boldsymbol{x} = \frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^T A^T\boldsymbol{x}$$
$$= \boldsymbol{x}^T A\boldsymbol{x}$$

Remark

We always assume that the matrix of a quadratic form is symmetric:

Even if A *is not symmetric,* $\frac{A+A^T}{2}$ *always is symmetric.*

And we have for the quadratic form

$$\begin{aligned}\boldsymbol{x}^T\left(\frac{A+A^T}{2}\right)\boldsymbol{x} &= \frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^T A^T\boldsymbol{x} \\ &= \boldsymbol{x}^T A\boldsymbol{x}\end{aligned}$$

because $\boldsymbol{x}^T A^T \boldsymbol{x} = \boldsymbol{x}^T A \boldsymbol{x}$ *is a scalar.*

Every quadratic form can be written in standard (i.e., diagonal) form since every real symmetric matrix is orthogonally similar to a diagonal matrix.

Every quadratic form can be written in standard (i.e., diagonal) form since every real symmetric matrix is orthogonally similar to a diagonal matrix.

Example

Take

$$\begin{aligned} f(\boldsymbol{x}) = x_1 x_2 &= \boldsymbol{x}^T \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \boldsymbol{x} \\ &= \boldsymbol{x}^T \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix} \boldsymbol{x} = \boldsymbol{x}^T \mathsf{A} \boldsymbol{x}. \end{aligned}$$

Every quadratic form can be written in standard (i.e., diagonal) form since every real symmetric matrix is orthogonally similar to a diagonal matrix.

Example

Take

$$f(\boldsymbol{x}) = x_1 x_2 = \boldsymbol{x}^T \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \boldsymbol{x}$$
$$= \boldsymbol{x}^T \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix} \boldsymbol{x} = \boldsymbol{x}^T \mathsf{A} \boldsymbol{x}.$$

We want to find the standard form $f(\boldsymbol{y}) = \boldsymbol{y}^T \mathsf{D} \boldsymbol{y}$, where D is diagonal and $\boldsymbol{y}$ are transformed coordinates.

Example (cont.)

We can compute the eigenvalues and (orthogonal) eigenvectors of A, i.e.,

$$
\begin{aligned}
& A = QDQ^T \\
\iff \quad & \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}
\end{aligned}
$$

Example (cont.)

We can compute the eigenvalues and (orthogonal) eigenvectors of A, i.e.,

$$
\begin{aligned}
& \mathsf{A} = \mathsf{QDQ}^T \\
\iff \quad & \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}
\end{aligned}
$$

so that

$$
f(\boldsymbol{x}) = \boldsymbol{x}^T\mathsf{A}\boldsymbol{x} = \underbrace{\boldsymbol{x}^T\mathsf{Q}}_{=\boldsymbol{y}^T}\mathsf{DQ}^T\boldsymbol{x} = \boldsymbol{y}^T\mathsf{D}\boldsymbol{y}
$$

and the standard form is

Example (cont.)

We can compute the eigenvalues and (orthogonal) eigenvectors of A, i.e.,

$$
\begin{aligned}
& \mathsf{A} = \mathsf{Q}\mathsf{D}\mathsf{Q}^T \\
\iff \quad & \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}
\end{aligned}
$$

so that

$$
f(\boldsymbol{x}) = \boldsymbol{x}^T \mathsf{A}\boldsymbol{x} = \underbrace{\boldsymbol{x}^T \mathsf{Q}}_{=\boldsymbol{y}^T}\mathsf{D}\mathsf{Q}^T\boldsymbol{x} = \boldsymbol{y}^T\mathsf{D}\boldsymbol{y}
$$

and the standard form is

$$
\boldsymbol{y}^T\begin{pmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{pmatrix}\boldsymbol{y} = \frac{1}{2}\left(y_1^2 - y_2^2\right).
$$

Remark

Instead of computing the eigenvalues and eigenvectors of A *in the example, we can also consider the factorization*

$$\mathsf{A} = \mathsf{LDL}^T.$$

Remark

Instead of computing the eigenvalues and eigenvectors of A *in the example, we can also consider the factorization*

$$\mathsf{A} = \mathsf{LDL}^T.$$

For a positive definite A *this is the Cholesky factorization, and it is cheaper to compute than eigenvalues and eigenvectors.*

Remark

Instead of computing the eigenvalues and eigenvectors of A *in the example, we can also consider the factorization*

$$\mathsf{A} = \mathsf{L}\mathsf{D}\mathsf{L}^T.$$

For a positive definite A *this is the Cholesky factorization, and it is cheaper to compute than eigenvalues and eigenvectors.*

Then

$$\boldsymbol{x}^T\mathsf{A}\boldsymbol{x} = \underbrace{\boldsymbol{x}^T\mathsf{L}}_{=\boldsymbol{y}^T}\mathsf{D}\mathsf{L}^T\boldsymbol{x} = \boldsymbol{y}^T\mathsf{D}\boldsymbol{y} = \sum_{i=1}^{n} p_i y_i^2,$$

where $\mathsf{D} = \operatorname{diag}(p_1, \ldots, p_n)$ *contains the pivots used in Gaussian elimination.*

Congruence transformations

Formally, the preceding argument uses a congruence transformation.

Definition

Two matrices $A, B \in \mathbb{R}^{n\times n}$ are called congruent if

$$B = C^T A C$$

for some nonsingular matrix C. Commonly used notation: $A \simeq B$.

Recall: A and B are similar if $\mathsf{B} = \mathsf{P}^{-1}\mathsf{AP}$, and similar matrices have the same eigenvalues.

Recall: A and B are similar if $B = P^{-1}AP$, and similar matrices have the same eigenvalues.

Now,

Definition

Let A be a real symmetric matrix. The triple (ρ, ν, ζ), where ρ, ν, and ζ, respectively, denote the number of positive, negative, and zero eigenvalues of A is called the inertia of A.

Theorem (Sylvester's Law of Inertia)

Let $\mathsf{A}, \mathsf{B} \in \mathbb{R}^{n \times n}$ *be symmetric. Then* A *and* B *are congruent, i.e.,* $\mathsf{A} \simeq \mathsf{B}$*, if and only if* A *and* B *have the same inertias.*

Theorem (Sylvester's Law of Inertia)

Let $\mathsf{A}, \mathsf{B} \in \mathbb{R}^{n \times n}$ *be symmetric. Then* A *and* B *are congruent, i.e.,* $\mathsf{A} \simeq \mathsf{B}$*, if and only if* A *and* B *have the same inertias.*

Proof.

See [Mey00]. □

Outline

Iterative Solvers

Consider the linear system

$$\mathsf{A}\boldsymbol{x} = \boldsymbol{b},$$

where A has many zero entries, i.e., A is sparse.

Iterative Solvers

Consider the linear system

$$\mathsf{A}\boldsymbol{x} = \boldsymbol{b},$$

where A has many zero entries, i.e., A is sparse.

In this case, direct factorization methods (such as LU, QR, SVD) are very inefficient to solve $\mathsf{A}\boldsymbol{x} = \boldsymbol{b}$.

Iterative Solvers

Consider the linear system

$$\mathsf{A}\boldsymbol{x} = \boldsymbol{b},$$

where A has many zero entries, i.e., A is sparse.

In this case, direct factorization methods (such as LU, QR, SVD) are very inefficient to solve $\mathsf{A}\boldsymbol{x} = \boldsymbol{b}$.

Instead, one uses iterative solvers.

The general framework for classical iterative solvers is as follows:

The general framework for classical iterative solvers is as follows:

- We split A into
$$A = M - N,$$
where M^{-1} exists and — ideally — is easy to compute.

The general framework for classical iterative solvers is as follows:

- We split A into

$$\mathsf{A} = \mathsf{M} - \mathsf{N},$$

  where M^{-1} exists and — ideally — is easy to compute.
- Then

$$\mathsf{A}\boldsymbol{x} = \boldsymbol{b} \quad \Longleftrightarrow \quad (\mathsf{M} - \mathsf{N})\boldsymbol{x} = \boldsymbol{b}$$

The general framework for classical iterative solvers is as follows:

- We split A into
$$\mathsf{A} = \mathsf{M} - \mathsf{N},$$
where M^{-1} exists and — ideally — is easy to compute.
- Then
$$\mathsf{A}\boldsymbol{x} = \boldsymbol{b} \iff (\mathsf{M} - \mathsf{N})\boldsymbol{x} = \boldsymbol{b} \iff \mathsf{M}\boldsymbol{x} = \mathsf{N}\boldsymbol{x} + \boldsymbol{b}$$

The general framework for classical iterative solvers is as follows:

- We split A into
$$\mathsf{A} = \mathsf{M} - \mathsf{N},$$
where M^{-1} exists and — ideally — is easy to compute.
- Then
$$\mathsf{A}\boldsymbol{x} = \boldsymbol{b} \iff (\mathsf{M} - \mathsf{N})\boldsymbol{x} = \boldsymbol{b} \iff \mathsf{M}\boldsymbol{x} = \mathsf{N}\boldsymbol{x} + \boldsymbol{b}$$
- and we iterate
$$\mathsf{M}\boldsymbol{x}^{(k)} = \mathsf{N}\boldsymbol{x}^{(k-1)} + \boldsymbol{b}$$

The general framework for classical iterative solvers is as follows:

- We split A into
$$\mathsf{A} = \mathsf{M} - \mathsf{N},$$
where M^{-1} exists and — ideally — is easy to compute.
- Then
$$\mathsf{A}\boldsymbol{x} = \boldsymbol{b} \iff (\mathsf{M} - \mathsf{N})\boldsymbol{x} = \boldsymbol{b} \iff \mathsf{M}\boldsymbol{x} = \mathsf{N}\boldsymbol{x} + \boldsymbol{b}$$
- and we iterate
$$\begin{aligned} & \mathsf{M}\boldsymbol{x}^{(k)} = \mathsf{N}\boldsymbol{x}^{(k-1)} + \boldsymbol{b} \\ \iff \quad & \boldsymbol{x}^{(k)} = \underbrace{\mathsf{M}^{-1}\mathsf{N}}_{=\mathsf{H}}\boldsymbol{x}^{(k-1)} + \underbrace{\mathsf{M}^{-1}\boldsymbol{b}}_{=\boldsymbol{d}}, \qquad k = 1, 2, 3, \ldots, \end{aligned}$$
where $\boldsymbol{x}^{(0)}$ is some initial guess and $\mathsf{H} = \mathsf{M}^{-1}\mathsf{N}$ is called the iteration matrix.

Theorem

Let M *and* N *be two matrices such that* $\mathsf{A} = \mathsf{M} - \mathsf{N}$ *and* $\mathsf{H} = \mathsf{M}^{-1}\mathsf{N}$*. If* $\rho(\mathsf{H}) < 1$ *then* A *is nonsingular and* $\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x} = \mathsf{A}^{-1}\boldsymbol{b}$*, i.e., the iterative method with iteration matrix* H*, converges for any initial guess* $\boldsymbol{x}^{(0)}$ *to the solution of* $\mathsf{A}\boldsymbol{x} = \boldsymbol{b}$*.*

Proof

First we show that A is nonsingular.

Proof

First we show that A is nonsingular.

Since $H = M^{-1}N$ (invertibility of M is an assumption) we have

$$\begin{aligned} A &= M - N \\ &= M - MH \\ &= M(I - H). \end{aligned} \tag{3}$$

Proof

First we show that A is nonsingular.

Since $H = M^{-1}N$ (invertibility of M is an assumption) we have

$$\begin{aligned} A &= M - N \\ &= M - MH \\ &= M(I - H). \end{aligned} \tag{3}$$

Now, since $\rho(H) < 1$ we know that $I - H$ is invertible via its Neumann series, and therefore A is invertible.

Proof (cont.)

Now we show that $\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x} = \mathsf{A}^{-1}\boldsymbol{b}$:

Proof (cont.)

Now we show that $\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x} = \mathsf{A}^{-1}\boldsymbol{b}$:

$$\boldsymbol{x}^{(k)} = \mathsf{H}\boldsymbol{x}^{(k-1)} + \boldsymbol{d}$$

Proof (cont.)

Now we show that $\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x} = \mathsf{A}^{-1}\boldsymbol{b}$:

$$\begin{aligned}
\boldsymbol{x}^{(k)} &= \mathsf{H}\boldsymbol{x}^{(k-1)} + \boldsymbol{d} \\
&= \mathsf{H}\left(\mathsf{H}\boldsymbol{x}^{(k-2)} + \boldsymbol{d}\right) + \boldsymbol{d} = \mathsf{H}^2\boldsymbol{x}^{(k-2)} + (\mathsf{I}+\mathsf{H})\boldsymbol{d}
\end{aligned}$$

Proof (cont.)

Now we show that $\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x} = \mathsf{A}^{-1}\boldsymbol{b}$:

$$
\begin{aligned}
\boldsymbol{x}^{(k)} &= \mathsf{H}\boldsymbol{x}^{(k-1)} + \boldsymbol{d} \\
&= \mathsf{H}\left(\mathsf{H}\boldsymbol{x}^{(k-2)} + \boldsymbol{d}\right) + \boldsymbol{d} = \mathsf{H}^2\boldsymbol{x}^{(k-2)} + (\mathsf{I} + \mathsf{H})\boldsymbol{d} \\
&\vdots \\
&= \mathsf{H}^k\boldsymbol{x}^{(0)} + (\mathsf{I} + \mathsf{H} + \ldots + \mathsf{H}^{k-1})\boldsymbol{d},
\end{aligned}
$$

where

Proof (cont.)

Now we show that $\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x} = \mathsf{A}^{-1}\boldsymbol{b}$:

$$
\begin{aligned}
\boldsymbol{x}^{(k)} &= \mathsf{H}\boldsymbol{x}^{(k-1)} + \boldsymbol{d} \\
&= \mathsf{H}\left(\mathsf{H}\boldsymbol{x}^{(k-2)} + \boldsymbol{d}\right) + \boldsymbol{d} = \mathsf{H}^2\boldsymbol{x}^{(k-2)} + (\mathsf{I} + \mathsf{H})\boldsymbol{d} \\
&\vdots \\
&= \mathsf{H}^k\boldsymbol{x}^{(0)} + (\mathsf{I} + \mathsf{H} + \ldots + \mathsf{H}^{k-1})\boldsymbol{d},
\end{aligned}
$$

where

$$\mathsf{H}^k \to \mathsf{O}$$

Proof (cont.)

Now we show that $\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x} = \mathsf{A}^{-1}\boldsymbol{b}$:

$$\begin{aligned}\boldsymbol{x}^{(k)} &= \mathsf{H}\boldsymbol{x}^{(k-1)} + \boldsymbol{d} \\ &= \mathsf{H}\left(\mathsf{H}\boldsymbol{x}^{(k-2)} + \boldsymbol{d}\right) + \boldsymbol{d} = \mathsf{H}^2\boldsymbol{x}^{(k-2)} + (\mathsf{I}+\mathsf{H})\boldsymbol{d} \\ &\vdots \\ &= \mathsf{H}^k\boldsymbol{x}^{(0)} + (\mathsf{I}+\mathsf{H}+\ldots+\mathsf{H}^{k-1})\boldsymbol{d},\end{aligned}$$

where

$$\mathsf{H}^k \to \mathsf{O} \quad \text{and} \quad (\mathsf{I}+\mathsf{H}+\ldots+\mathsf{H}^{k-1}) \to (\mathsf{I}-\mathsf{H})^{-1} \qquad \text{for } k\to\infty$$

Proof (cont.)

Now we show that $\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x} = \mathsf{A}^{-1}\boldsymbol{b}$:

$$\begin{aligned}
\boldsymbol{x}^{(k)} &= \mathsf{H}\boldsymbol{x}^{(k-1)} + \boldsymbol{d} \\
&= \mathsf{H}\left(\mathsf{H}\boldsymbol{x}^{(k-2)} + \boldsymbol{d}\right) + \boldsymbol{d} = \mathsf{H}^2\boldsymbol{x}^{(k-2)} + (\mathsf{I}+\mathsf{H})\boldsymbol{d} \\
&\vdots \\
&= \mathsf{H}^k\boldsymbol{x}^{(0)} + (\mathsf{I}+\mathsf{H}+\ldots+\mathsf{H}^{k-1})\boldsymbol{d},
\end{aligned}$$

where

$$\mathsf{H}^k \to \mathsf{O} \quad \text{and} \quad (\mathsf{I}+\mathsf{H}+\ldots+\mathsf{H}^{k-1}) \to (\mathsf{I}-\mathsf{H})^{-1} \qquad \text{for } k\to\infty$$

so that — using (3), i.e., $(\mathsf{I}-\mathsf{H})^{-1} = \mathsf{A}^{-1}\mathsf{M}$,

$$\lim_{k\to\infty} \boldsymbol{x}^{(k)} = (\mathsf{I}-\mathsf{H})^{-1}\boldsymbol{d}$$

Proof (cont.)

Now we show that $\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x} = \mathsf{A}^{-1}\boldsymbol{b}$:

$$\begin{aligned}
\boldsymbol{x}^{(k)} &= \mathsf{H}\boldsymbol{x}^{(k-1)} + \boldsymbol{d} \\
&= \mathsf{H}\left(\mathsf{H}\boldsymbol{x}^{(k-2)} + \boldsymbol{d}\right) + \boldsymbol{d} = \mathsf{H}^2\boldsymbol{x}^{(k-2)} + (\mathsf{I}+\mathsf{H})\boldsymbol{d} \\
&\ \vdots \\
&= \mathsf{H}^k\boldsymbol{x}^{(0)} + (\mathsf{I}+\mathsf{H}+\ldots+\mathsf{H}^{k-1})\boldsymbol{d},
\end{aligned}$$

where

$$\mathsf{H}^k \to \mathsf{O} \quad \text{and} \quad (\mathsf{I}+\mathsf{H}+\ldots+\mathsf{H}^{k-1}) \to (\mathsf{I}-\mathsf{H})^{-1} \qquad \text{for } k\to\infty$$

so that — using (3), i.e., $(\mathsf{I}-\mathsf{H})^{-1} = \mathsf{A}^{-1}\mathsf{M}$,

$$\begin{aligned}
\lim_{k\to\infty} \boldsymbol{x}^{(k)} &= (\mathsf{I}-\mathsf{H})^{-1}\boldsymbol{d} \\
&= (\mathsf{I}-\mathsf{H})^{-1}\mathsf{M}^{-1}\boldsymbol{b}
\end{aligned}$$

Proof (cont.)

Now we show that $\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x} = \mathsf{A}^{-1}\boldsymbol{b}$:

$$\begin{aligned}
\boldsymbol{x}^{(k)} &= \mathsf{H}\boldsymbol{x}^{(k-1)} + \boldsymbol{d} \\
&= \mathsf{H}\left(\mathsf{H}\boldsymbol{x}^{(k-2)} + \boldsymbol{d}\right) + \boldsymbol{d} = \mathsf{H}^2\boldsymbol{x}^{(k-2)} + (\mathsf{I}+\mathsf{H})\boldsymbol{d} \\
&\vdots \\
&= \mathsf{H}^k\boldsymbol{x}^{(0)} + (\mathsf{I}+\mathsf{H}+\ldots+\mathsf{H}^{k-1})\boldsymbol{d},
\end{aligned}$$

where

$$\mathsf{H}^k \to \mathsf{O} \quad \text{and} \quad (\mathsf{I}+\mathsf{H}+\ldots+\mathsf{H}^{k-1}) \to (\mathsf{I}-\mathsf{H})^{-1} \qquad \text{for } k\to\infty$$

so that — using (3), i.e., $(\mathsf{I}-\mathsf{H})^{-1} = \mathsf{A}^{-1}\mathsf{M}$,

$$\begin{aligned}
\lim_{k\to\infty} \boldsymbol{x}^{(k)} &= (\mathsf{I}-\mathsf{H})^{-1}\boldsymbol{d} \\
&= (\mathsf{I}-\mathsf{H})^{-1}\mathsf{M}^{-1}\boldsymbol{b} = \mathsf{A}^{-1}\boldsymbol{b}
\end{aligned}$$

Proof (cont.)

Now we show that $\lim_{k\to\infty} \boldsymbol{x}^{(k)} = \boldsymbol{x} = \mathsf{A}^{-1}\boldsymbol{b}$:

$$\begin{aligned}
\boldsymbol{x}^{(k)} &= \mathsf{H}\boldsymbol{x}^{(k-1)} + \boldsymbol{d} \\
&= \mathsf{H}\left(\mathsf{H}\boldsymbol{x}^{(k-2)} + \boldsymbol{d}\right) + \boldsymbol{d} = \mathsf{H}^2\boldsymbol{x}^{(k-2)} + (\mathsf{I}+\mathsf{H})\boldsymbol{d} \\
&\vdots \\
&= \mathsf{H}^k\boldsymbol{x}^{(0)} + (\mathsf{I}+\mathsf{H}+\ldots+\mathsf{H}^{k-1})\boldsymbol{d},
\end{aligned}$$

where

$$\mathsf{H}^k \to \mathsf{O} \quad \text{and} \quad (\mathsf{I}+\mathsf{H}+\ldots+\mathsf{H}^{k-1}) \to (\mathsf{I}-\mathsf{H})^{-1} \qquad \text{for } k\to\infty$$

so that — using (3), i.e., $(\mathsf{I}-\mathsf{H})^{-1} = \mathsf{A}^{-1}\mathsf{M}$,

$$\begin{aligned}
\lim_{k\to\infty} \boldsymbol{x}^{(k)} &= (\mathsf{I}-\mathsf{H})^{-1}\boldsymbol{d} \\
&= (\mathsf{I}-\mathsf{H})^{-1}\mathsf{M}^{-1}\boldsymbol{b} = \mathsf{A}^{-1}\boldsymbol{b} = \boldsymbol{x}.
\end{aligned}$$

Remark

In order to have a "good" iterative solver we will want

Remark

In order to have a "good" iterative solver we will want

- *fast convergence — ensured by $\rho(\mathsf{H}) \ll 1$,*

Remark

In order to have a "good" iterative solver we will want

- *fast convergence — ensured by* $\rho(\mathrm{H}) \ll 1$,
- *simple computation — ensured by easy computation of* M^{-1} *(or* $\mathrm{H} = \mathrm{M}^{-1}\mathrm{N}$ *and* $\boldsymbol{d} = \mathrm{M}^{-1}\boldsymbol{b}$*).*

Remark

In order to have a "good" iterative solver we will want

- *fast convergence — ensured by* $\rho(\mathsf{H}) \ll 1$,
- *simple computation — ensured by easy computation of* M^{-1} *(or* $\mathsf{H} = \mathsf{M}^{-1}\mathsf{N}$ *and* $\boldsymbol{d} = \mathsf{M}^{-1}\boldsymbol{b}$*).*

We conclude by presenting two standard examples:

- Jacobi iteration,
- Gauss-Seidel iteration.

Jacobi iteration

We take $M = D = \text{diag}(A)$, which is easy to invert.

Jacobi iteration

We take $M = D = \operatorname{diag}(A)$, which is easy to invert.

Then
$$A = M - N = D - N,$$
i.e., $N = -(A - D)$ or, if $A = L + D + U$, $N = -(L + U)$.

Jacobi iteration

We take $M = D = \mathrm{diag}(A)$, which is easy to invert.

Then

$$A = M - N = D - N,$$

i.e., $N = -(A - D)$ or, if $A = L + D + U$, $N = -(L + U)$.
Therefore $A\boldsymbol{x} = \boldsymbol{b}$ is solved via

$$D\boldsymbol{x}^{(k)} = N\boldsymbol{x}^{(k-1)} + \boldsymbol{b}, \qquad k = 1, 2, 3, \ldots,$$

Jacobi iteration

We take $\mathsf{M} = \mathsf{D} = \mathrm{diag}(\mathsf{A})$, which is easy to invert.

Then

$$\mathsf{A} = \mathsf{M} - \mathsf{N} = \mathsf{D} - \mathsf{N},$$

i.e., $\mathsf{N} = -(\mathsf{A} - \mathsf{D})$ or, if $\mathsf{A} = \mathsf{L} + \mathsf{D} + \mathsf{U}$, $\mathsf{N} = -(\mathsf{L} + \mathsf{U})$.
Therefore $\mathsf{A}\boldsymbol{x} = \boldsymbol{b}$ is solved via

$$\mathsf{D}\boldsymbol{x}^{(k)} = \mathsf{N}\boldsymbol{x}^{(k-1)} + \boldsymbol{b}, \qquad k = 1, 2, 3, \ldots,$$

or componentwise

$$\boldsymbol{x}_i^{(k)} = \frac{1}{a_{ii}} \left(b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} \boldsymbol{x}_j^{(k-1)} \right), \quad i = 1, 2, \ldots, n.$$

Remark

- *Jacobi iteration is embarrassingly parallel, i.e., the above componentwise equations can be directly implemented on n parallel processors.*

Remark

- *Jacobi iteration is embarrassingly parallel, i.e., the above componentwise equations can be directly implemented on n parallel processors.*
- *Also, only entries from the i^{th} row of the matrix are needed to update the i^{th} component of $\boldsymbol{x}$.*

Remark

- *Jacobi iteration is embarrassingly parallel, i.e., the above componentwise equations can be directly implemented on n parallel processors.*
- *Also, only entries from the i^{th} row of the matrix are needed to update the i^{th} component of $\boldsymbol{x}$.*
- *Jacobi iteration had long been considered as too simple (and too slow) to be useful. However, a recent modification [YM14] using relaxation has changed that. This modification was customized to solve elliptic PDEs via a finite difference discretization.*

Theorem

If A *is diagonally dominant, then Jacobi iteration converges for any initial guess.*

Proof.

Diagonal dominance says

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

Proof.

Diagonal dominance says

$$|a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|, \quad i = 1, \ldots, n \quad \Longleftrightarrow \quad \sum_{\substack{j=1\\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| < 1.$$

Proof.

Diagonal dominance says

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \quad i = 1, \ldots, n \quad \iff \quad \sum_{\substack{j=1 \\ j \neq i}}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| < 1.$$

Now

$$\|\mathsf{H}\|_\infty =$$

$\square$

Proof.

Diagonal dominance says

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \quad i = 1, \ldots, n \quad \Longleftrightarrow \quad \sum_{\substack{j=1 \\ j \neq i}}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| < 1.$$

Now

$$\|\mathsf{H}\|_\infty = \|\mathsf{D}^{-1}\mathsf{N}\|_\infty$$

□

Proof.

Diagonal dominance says

$$|a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|, \quad i = 1, \ldots, n \quad \Longleftrightarrow \quad \sum_{\substack{j=1\\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| < 1.$$

Now

$$\|\mathsf{H}\|_\infty = \|\mathsf{D}^{-1}\mathsf{N}\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right|$$

□

Proof.

Diagonal dominance says

$$|a_{ii}| > \sum_{\substack{j=1\\j\neq i}}^{n} |a_{ij}|, \quad i = 1, \ldots, n \quad \Longleftrightarrow \quad \sum_{\substack{j=1\\j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| < 1.$$

Now

$$\|\mathsf{H}\|_\infty = \|\mathsf{D}^{-1}\mathsf{N}\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} \left|\frac{a_{ij}}{a_{ii}}\right|$$

$$\stackrel{\text{diag}(\mathsf{N})=\mathbf{0}}{=} \max_{1\le i\le n} \sum_{\substack{j=1\\j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| < 1.$$

$\square$

Remark

Since $\rho(\mathsf{H}) < \|\mathsf{H}\|$, *diagonal dominance (or* $\|\mathsf{H}\|_\infty < 1$*) is a weaker condition than* $\rho(\mathsf{H}) < 1$*.*

Gauss–Seidel iteration

Let's again decompose $A = L + D + U$, but now take

$$M = D + L, \quad N = -U.$$

Gauss–Seidel iteration

Let's again decompose $A = L + D + U$, but now take

$$M = D + L, \quad N = -U.$$

Then

$$H = M^{-1}N = -(D + L)^{-1}U$$
$$\boldsymbol{d} = M^{-1}\boldsymbol{b} = (D + L)^{-1}\boldsymbol{b}.$$

Gauss–Seidel iteration

Let's again decompose $\mathsf{A} = \mathsf{L} + \mathsf{D} + \mathsf{U}$, but now take

$$\mathsf{M} = \mathsf{D} + \mathsf{L}, \quad \mathsf{N} = -\mathsf{U}.$$

Then

$$\mathsf{H} = \mathsf{M}^{-1}\mathsf{N} = -(\mathsf{D} + \mathsf{L})^{-1}\mathsf{U}$$
$$\boldsymbol{d} = \mathsf{M}^{-1}\boldsymbol{b} = (\mathsf{D} + \mathsf{L})^{-1}\boldsymbol{b}.$$

The iteration formula is

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

Gauss–Seidel iteration

Let's again decompose $\mathsf{A} = \mathsf{L} + \mathsf{D} + \mathsf{U}$, but now take

$$\mathsf{M} = \mathsf{D} + \mathsf{L}, \quad \mathsf{N} = -\mathsf{U}.$$

Then

$$\begin{aligned} \mathsf{H} &= \mathsf{M}^{-1}\mathsf{N} = -(\mathsf{D} + \mathsf{L})^{-1}\mathsf{U} \\ \boldsymbol{d} &= \mathsf{M}^{-1}\boldsymbol{b} = (\mathsf{D} + \mathsf{L})^{-1}\boldsymbol{b}. \end{aligned}$$

The iteration formula is

$$\begin{aligned} & \boldsymbol{x}^{(k)} = -(\mathsf{D} + \mathsf{L})^{-1}\mathsf{U}\boldsymbol{x}^{(k-1)} + (\mathsf{D} + \mathsf{L})^{-1}\boldsymbol{b} \\ \iff \quad & (\mathsf{D} + \mathsf{L})\boldsymbol{x}^{(k)} = \boldsymbol{b} - \mathsf{U}\boldsymbol{x}^{(k-1)}. \end{aligned}$$

Componentwise we get

$$\sum_{j=1}^{i-1} a_{ij}\boldsymbol{x}_j^{(k)} + a_{ii}\boldsymbol{x}_i^{(k)} = b_i - \sum_{j=i+1}^{n} a_{ij}\boldsymbol{x}_j^{(k-1)}.$$

Componentwise we get

$$\sum_{j=1}^{i-1} a_{ij}\boldsymbol{x}_j^{(k)} + a_{ii}\boldsymbol{x}_i^{(k)} = b_i - \sum_{j=i+1}^{n} a_{ij}\boldsymbol{x}_j^{(k-1)}.$$

Since when we work on the i^{th} component the components $\boldsymbol{x}_j^{(k)}$, $j < i$, have already been updated we can write

Componentwise we get

$$\sum_{j=1}^{i-1} a_{ij}\boldsymbol{x}_j^{(k)} + a_{ii}\boldsymbol{x}_i^{(k)} = b_i - \sum_{j=i+1}^{n} a_{ij}\boldsymbol{x}_j^{(k-1)}.$$

Since when we work on the i^{th} component the components $\boldsymbol{x}_j^{(k)}$, $j < i$, have already been updated we can write

$$\boldsymbol{x}_i^{(k)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}\boldsymbol{x}_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}\boldsymbol{x}_j^{(k-1)}\right), \quad i = 1, 2, \ldots, n.$$

Componentwise we get

$$\sum_{j=1}^{i-1} a_{ij}\boldsymbol{x}_j^{(k)} + a_{ii}\boldsymbol{x}_i^{(k)} = b_i - \sum_{j=i+1}^{n} a_{ij}\boldsymbol{x}_j^{(k-1)}.$$

Since when we work on the i^{th} component the components $\boldsymbol{x}_j^{(k)}$, $j < i$, have already been updated we can write

$$\boldsymbol{x}_i^{(k)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}\boldsymbol{x}_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}\boldsymbol{x}_j^{(k-1)}\right), \quad i = 1, 2, \ldots, n.$$

Remark

Gauss–Seidel iteration is similar to Jacobi iteration, but it uses the most recently computed information as soon as it becomes available (instead of waiting until the next iteration, as Jacobi does).

Convergence of Gauss-Seidel iteration

Theorem

Gauss–Seidel iteration converges for any initial guess if

1. A *is diagonally dominant, or*
2. A *is symmetric positive definite.*

Convergence of Gauss-Seidel iteration

Theorem

Gauss–Seidel iteration converges for any initial guess if

1. A *is diagonally dominant, or*
2. A *is symmetric positive definite.*

Proof.

1. In [Mey00],
2. on next few slides.

□

Convergence of Gauss-Seidel iteration

Theorem

Gauss–Seidel iteration converges for any initial guess if

1. A *is diagonally dominant, or*
2. A *is symmetric positive definite.*

Proof.

1. In [Mey00],
2. on next few slides.

□

Remark

Usually Gauss–Seidel converges faster than Jacobi. However, there are exceptions.

Proof (convergence for positive definite A)

Since A is symmetric, we can decompose

$$\mathsf{A} = \mathsf{L} + \mathsf{D} + \mathsf{L}^T, \qquad \mathsf{H} = -(\mathsf{D} + \mathsf{L})^{-1}\mathsf{L}^T.$$

Proof (convergence for positive definite A)

Since A is symmetric, we can decompose

$$A = L + D + L^T, \qquad H = -(D+L)^{-1}L^T.$$

Convergence will be ensured if we can show that $\rho(H) < 1$, i.e.,

$$\rho(-(D+L)^{-1}L^T) < 1.$$

Proof (convergence for positive definite A)

Since A is symmetric, we can decompose

$$\mathsf{A} = \mathsf{L} + \mathsf{D} + \mathsf{L}^T, \qquad \mathsf{H} = -(\mathsf{D} + \mathsf{L})^{-1}\mathsf{L}^T.$$

Convergence will be ensured if we can show that $\rho(\mathsf{H}) < 1$, i.e.,

$$\rho(-(\mathsf{D} + \mathsf{L})^{-1}\mathsf{L}^T) < 1.$$

Since D has positive entries (otherwise A couldn't be positive definite), D is positive definite (and therefore nonsingular) so that

$$\widetilde{\mathsf{H}} = \mathsf{D}^{1/2}\mathsf{H}\mathsf{D}^{-1/2}$$

has the same eigenvalues as H.

Proof (convergence for positive definite A)

Since A is symmetric, we can decompose

$$A = L + D + L^T, \qquad H = -(D + L)^{-1} L^T.$$

Convergence will be ensured if we can show that $\rho(H) < 1$, i.e.,

$$\rho(-(D + L)^{-1} L^T) < 1.$$

Since D has positive entries (otherwise A couldn't be positive definite), D is positive definite (and therefore nonsingular) so that

$$\widetilde{H} = D^{1/2} H D^{-1/2}$$

has the same eigenvalues as H.
Therefore, we now show that

$$\rho(\widetilde{H}) < 1.$$

Proof (cont.)

First, we rewrite $\widetilde{H}$. For this we require a push-through identity for the matrix inverse ([Ber09], similar to what we had in Chapter 3):

$$(I + AB)^{-1}A = A(I + BA)^{-1}. \tag{4}$$

Proof (cont.)

First, we rewrite $\widetilde{\mathsf{H}}$. For this we require a push-through identity for the matrix inverse ([Ber09], similar to what we had in Chapter 3):

$$(\mathsf{I} + \mathsf{AB})^{-1}\mathsf{A} = \mathsf{A}(\mathsf{I} + \mathsf{BA})^{-1}. \tag{4}$$

If we let $\mathsf{A} = \mathsf{D}^{-1/2}$ and $\mathsf{B} = \mathsf{LD}^{-1/2}$, then we get

$$(\mathsf{I} + \mathsf{D}^{-1/2}\mathsf{LD}^{-1/2})^{-1}\mathsf{D}^{-1/2} \stackrel{(4)}{=} \mathsf{D}^{-1/2}(\mathsf{I} + \mathsf{LD}^{-1/2}\mathsf{D}^{-1/2})^{-1} \tag{5}$$

Proof (cont.)

First, we rewrite $\widetilde{H}$. For this we require a push-through identity for the matrix inverse ([Ber09], similar to what we had in Chapter 3):

$$(I + AB)^{-1}A = A(I + BA)^{-1}. \tag{4}$$

If we let $A = D^{-1/2}$ and $B = LD^{-1/2}$, then we get

$$\begin{aligned}(I + D^{-1/2}LD^{-1/2})^{-1}D^{-1/2} &\stackrel{(4)}{=} D^{-1/2}(I + LD^{-1/2}D^{-1/2})^{-1} \\ &= D^{-1/2}(I + LD^{-1})^{-1}\end{aligned} \tag{5}$$

Proof (cont.)

First, we rewrite $\widetilde{H}$. For this we require a push-through identity for the matrix inverse ([Ber09], similar to what we had in Chapter 3):

$$(I + AB)^{-1}A = A(I + BA)^{-1}. \tag{4}$$

If we let $A = D^{-1/2}$ and $B = LD^{-1/2}$, then we get

$$\begin{aligned}(I + D^{-1/2}LD^{-1/2})^{-1}D^{-1/2} &\stackrel{(4)}{=} D^{-1/2}(I + LD^{-1/2}D^{-1/2})^{-1} \\ &= D^{-1/2}(I + LD^{-1})^{-1} \\ &= D^{-1/2}(DD^{-1} + LD^{-1})^{-1}\end{aligned}$$

(5)

Proof (cont.)

First, we rewrite $\widetilde{H}$. For this we require a push-through identity for the matrix inverse ([Ber09], similar to what we had in Chapter 3):

$$(I + AB)^{-1}A = A(I + BA)^{-1}. \tag{4}$$

If we let $A = D^{-1/2}$ and $B = LD^{-1/2}$, then we get

$$\begin{aligned}(I + D^{-1/2}LD^{-1/2})^{-1}D^{-1/2} &\overset{(4)}{=} D^{-1/2}(I + LD^{-1/2}D^{-1/2})^{-1} \\ &= D^{-1/2}(I + LD^{-1})^{-1} \\ &= D^{-1/2}(DD^{-1} + LD^{-1})^{-1} \\ &= D^{-1/2}\left((D + L)D^{-1}\right)^{-1}\end{aligned} \tag{5}$$

Proof (cont.)

First, we rewrite $\widetilde{\mathsf{H}}$. For this we require a push-through identity for the matrix inverse ([Ber09], similar to what we had in Chapter 3):

$$(\mathsf{I}+\mathsf{AB})^{-1}\mathsf{A} = \mathsf{A}(\mathsf{I}+\mathsf{BA})^{-1}. \tag{4}$$

If we let $\mathsf{A} = \mathsf{D}^{-1/2}$ and $\mathsf{B} = \mathsf{L}\mathsf{D}^{-1/2}$, then we get

$$\begin{aligned}
(\mathsf{I}+\mathsf{D}^{-1/2}\mathsf{L}\mathsf{D}^{-1/2})^{-1}\mathsf{D}^{-1/2} &\overset{(4)}{=} \mathsf{D}^{-1/2}(\mathsf{I}+\mathsf{L}\mathsf{D}^{-1/2}\mathsf{D}^{-1/2})^{-1} \\
&= \mathsf{D}^{-1/2}(\mathsf{I}+\mathsf{L}\mathsf{D}^{-1})^{-1} \\
&= \mathsf{D}^{-1/2}(\mathsf{D}\mathsf{D}^{-1}+\mathsf{L}\mathsf{D}^{-1})^{-1} \\
&= \mathsf{D}^{-1/2}\left((\mathsf{D}+\mathsf{L})\mathsf{D}^{-1}\right)^{-1} \\
&= \mathsf{D}^{-1/2}\mathsf{D}(\mathsf{D}+\mathsf{L})^{-1} = \mathsf{D}^{1/2}(\mathsf{D}+\mathsf{L})^{-1}. \qquad (5)
\end{aligned}$$

Proof (cont.)

Therefore

$$\begin{aligned}\widetilde{H} &= D^{1/2} H D^{-1/2} \\ &= -D^{1/2}(D+L)^{-1} L^T D^{-1/2}\end{aligned}$$

Proof (cont.)

Therefore

$$\begin{aligned}
\widetilde{H} &= D^{1/2} H D^{-1/2} \\
&= -D^{1/2}(D+L)^{-1} L^T D^{-1/2} \\
&\overset{(5)}{=} -(I + D^{-1/2} L D^{-1/2})^{-1} D^{-1/2} L^T D^{-1/2}
\end{aligned}$$

Proof (cont.)

Therefore

$$\begin{aligned}
\widetilde{H} &= D^{1/2} H D^{-1/2} \\
&= -D^{1/2}(D+L)^{-1} L^T D^{-1/2} \\
&\overset{(5)}{=} -(I + D^{-1/2} L D^{-1/2})^{-1} D^{-1/2} L^T D^{-1/2} \\
&= -(I + \widetilde{L})^{-1} \widetilde{L}^T,
\end{aligned}$$

where $\widetilde{L} = D^{-1/2} L D^{-1/2}$.

Proof (cont.)

Therefore

$$\begin{aligned}
\widetilde{\mathsf{H}} &= \mathsf{D}^{1/2}\mathsf{H}\mathsf{D}^{-1/2} \\
&= -\mathsf{D}^{1/2}(\mathsf{D}+\mathsf{L})^{-1}\mathsf{L}^T\mathsf{D}^{-1/2} \\
&\overset{(5)}{=} -(\mathsf{I}+\mathsf{D}^{-1/2}\mathsf{L}\mathsf{D}^{-1/2})^{-1}\mathsf{D}^{-1/2}\mathsf{L}^T\mathsf{D}^{-1/2} \\
&= -(\mathsf{I}+\widetilde{\mathsf{L}})^{-1}\widetilde{\mathsf{L}}^T,
\end{aligned}$$

where $\widetilde{\mathsf{L}} = \mathsf{D}^{-1/2}\mathsf{L}\mathsf{D}^{-1/2}$.

Now consider an eigenpair $(\lambda, \boldsymbol{x})$ of $\widetilde{\mathsf{H}}$ with $\boldsymbol{x}^*\boldsymbol{x} = 1$.

Proof (cont.)

Therefore

$$\begin{aligned}
\widetilde{\mathsf{H}} &= \mathsf{D}^{1/2}\mathsf{H}\mathsf{D}^{-1/2} \\
&= -\mathsf{D}^{1/2}(\mathsf{D}+\mathsf{L})^{-1}\mathsf{L}^T\mathsf{D}^{-1/2} \\
&\overset{(5)}{=} -(\mathsf{I}+\mathsf{D}^{-1/2}\mathsf{L}\mathsf{D}^{-1/2})^{-1}\mathsf{D}^{-1/2}\mathsf{L}^T\mathsf{D}^{-1/2} \\
&= -(\mathsf{I}+\widetilde{\mathsf{L}})^{-1}\widetilde{\mathsf{L}}^T,
\end{aligned}$$

where $\widetilde{\mathsf{L}} = \mathsf{D}^{-1/2}\mathsf{L}\mathsf{D}^{-1/2}$.
Now consider an eigenpair $(\lambda, \boldsymbol{x})$ of $\widetilde{\mathsf{H}}$ with $\boldsymbol{x}^*\boldsymbol{x} = 1$. Then

$$\widetilde{\mathsf{H}}\boldsymbol{x} = \lambda\boldsymbol{x} \quad \Longleftrightarrow \quad -\widetilde{\mathsf{L}}^T\boldsymbol{x} = \lambda(\mathsf{I}+\widetilde{\mathsf{L}})\boldsymbol{x}.$$

Proof (cont.)

Therefore

$$\begin{aligned}
\widetilde{\mathsf{H}} &= \mathsf{D}^{1/2}\mathsf{H}\mathsf{D}^{-1/2} \\
&= -\mathsf{D}^{1/2}(\mathsf{D}+\mathsf{L})^{-1}\mathsf{L}^T\mathsf{D}^{-1/2} \\
&\overset{(5)}{=} -(\mathsf{I}+\mathsf{D}^{-1/2}\mathsf{L}\mathsf{D}^{-1/2})^{-1}\mathsf{D}^{-1/2}\mathsf{L}^T\mathsf{D}^{-1/2} \\
&= -(\mathsf{I}+\widetilde{\mathsf{L}})^{-1}\widetilde{\mathsf{L}}^T,
\end{aligned}$$

where $\widetilde{\mathsf{L}} = \mathsf{D}^{-1/2}\mathsf{L}\mathsf{D}^{-1/2}$.

Now consider an eigenpair $(\lambda, \boldsymbol{x})$ of $\widetilde{\mathsf{H}}$ with $\boldsymbol{x}^*\boldsymbol{x} = 1$. Then

$$\widetilde{\mathsf{H}}\boldsymbol{x} = \lambda\boldsymbol{x} \quad \Longleftrightarrow \quad -\widetilde{\mathsf{L}}^T\boldsymbol{x} = \lambda(\mathsf{I}+\widetilde{\mathsf{L}})\boldsymbol{x}.$$

Multiplying by $\boldsymbol{x}^*$ yields

$$-\boldsymbol{x}^*\widetilde{\mathsf{L}}^T\boldsymbol{x} = \lambda(\underbrace{\boldsymbol{x}^*\boldsymbol{x}}_{=1}+\boldsymbol{x}^*\widetilde{\mathsf{L}}\boldsymbol{x}) \quad \Longleftrightarrow \quad \lambda = \frac{-\boldsymbol{x}^*\widetilde{\mathsf{L}}^T\boldsymbol{x}}{1+\boldsymbol{x}^*\widetilde{\mathsf{L}}\boldsymbol{x}}.$$

Proof (cont.)

Finally, we let $\boldsymbol{x}^* \widetilde{\mathsf{L}} \boldsymbol{x} = a + b\mathrm{i}$.

Proof (cont.)

Finally, we let $\boldsymbol{x}^*\widetilde{\mathsf{L}}\boldsymbol{x} = a + bi$. Then we have $\boldsymbol{x}^*\widetilde{\mathsf{L}}^T\boldsymbol{x} = a - bi$ so that

$$|\lambda|^2 = \left|\frac{-a+bi}{1+a+bi}\right|^2$$

Proof (cont.)

Finally, we let $\boldsymbol{x}^*\widetilde{\mathsf{L}}\boldsymbol{x} = a + bi$. Then we have $\boldsymbol{x}^*\widetilde{\mathsf{L}}^T\boldsymbol{x} = a - bi$ so that

$$|\lambda|^2 = \left|\frac{-a+bi}{1+a+bi}\right|^2 = \frac{a^2+b^2}{1+2a+a^2+b^2}$$

Proof (cont.)

Finally, we let $\boldsymbol{x}^*\widetilde{\mathsf{L}}\boldsymbol{x} = a + bi$. Then we have $\boldsymbol{x}^*\widetilde{\mathsf{L}}^T\boldsymbol{x} = a - bi$ so that

$$|\lambda|^2 = \left|\frac{-a+bi}{1+a+bi}\right|^2 = \frac{a^2+b^2}{1+2a+a^2+b^2} < 1$$

since $1 + 2a > 0$, as we now show:

Proof (cont.)

Finally, we let $\boldsymbol{x}^*\widetilde{\mathsf{L}}\boldsymbol{x} = a + bi$. Then we have $\boldsymbol{x}^*\widetilde{\mathsf{L}}^T\boldsymbol{x} = a - bi$ so that

$$|\lambda|^2 = \left|\frac{-a+bi}{1+a+bi}\right|^2 = \frac{a^2+b^2}{1+2a+a^2+b^2} < 1$$

since $1 + 2a > 0$, as we now show:

The matrix $\mathsf{D}^{-1/2}\mathsf{A}\mathsf{D}^{-1/2} = \widetilde{\mathsf{L}} + \mathsf{I} + \widetilde{\mathsf{L}}^T$ is positive definite, and therefore its quadratic form is positive.

Proof (cont.)

Finally, we let $\boldsymbol{x}^* \widetilde{\mathsf{L}} \boldsymbol{x} = a + bi$. Then we have $\boldsymbol{x}^* \widetilde{\mathsf{L}}^T \boldsymbol{x} = a - bi$ so that

$$|\lambda|^2 = \left| \frac{-a + bi}{1 + a + bi} \right|^2 = \frac{a^2 + b^2}{1 + 2a + a^2 + b^2} < 1$$

since $1 + 2a > 0$, as we now show:

The matrix $\mathsf{D}^{-1/2} \mathsf{A} \mathsf{D}^{-1/2} = \widetilde{\mathsf{L}} + \mathsf{I} + \widetilde{\mathsf{L}}^T$ is positive definite, and therefore its quadratic form is positive.

In particular, using the eigenvector $\boldsymbol{x}$ we have

$$0 < \boldsymbol{x}^* \mathsf{D}^{-1/2} \mathsf{A} \mathsf{D}^{-1/2} \boldsymbol{x}$$

Proof (cont.)

Finally, we let $\boldsymbol{x}^*\widetilde{\mathsf{L}}\boldsymbol{x} = a + bi$. Then we have $\boldsymbol{x}^*\widetilde{\mathsf{L}}^T\boldsymbol{x} = a - bi$ so that

$$|\lambda|^2 = \left|\frac{-a+bi}{1+a+bi}\right|^2 = \frac{a^2+b^2}{1+2a+a^2+b^2} < 1$$

since $1 + 2a > 0$, as we now show:

The matrix $\mathsf{D}^{-1/2}\mathsf{A}\mathsf{D}^{-1/2} = \widetilde{\mathsf{L}} + \mathsf{I} + \widetilde{\mathsf{L}}^T$ is positive definite, and therefore its quadratic form is positive.
In particular, using the eigenvector $\boldsymbol{x}$ we have

$$0 < \boldsymbol{x}^*\mathsf{D}^{-1/2}\mathsf{A}\mathsf{D}^{-1/2}\boldsymbol{x} = \underbrace{\boldsymbol{x}^*\widetilde{\mathsf{L}}\boldsymbol{x}}_{=a+bi} + \underbrace{\boldsymbol{x}^*\boldsymbol{x}}_{=1} + \underbrace{\boldsymbol{x}^*\widetilde{\mathsf{L}}^T\boldsymbol{x}}_{=a-bi}$$

Proof (cont.)

Finally, we let $\boldsymbol{x}^*\widetilde{\mathsf{L}}\boldsymbol{x} = a + bi$. Then we have $\boldsymbol{x}^*\widetilde{\mathsf{L}}^T\boldsymbol{x} = a - bi$ so that

$$|\lambda|^2 = \left|\frac{-a+bi}{1+a+bi}\right|^2 = \frac{a^2+b^2}{1+2a+a^2+b^2} < 1$$

since $1 + 2a > 0$, as we now show:

The matrix $\mathsf{D}^{-1/2}\mathsf{A}\mathsf{D}^{-1/2} = \widetilde{\mathsf{L}} + \mathsf{I} + \widetilde{\mathsf{L}}^T$ is positive definite, and therefore its quadratic form is positive.

In particular, using the eigenvector $\boldsymbol{x}$ we have

$$0 < \boldsymbol{x}^*\mathsf{D}^{-1/2}\mathsf{A}\mathsf{D}^{-1/2}\boldsymbol{x} = \underbrace{\boldsymbol{x}^*\widetilde{\mathsf{L}}\boldsymbol{x}}_{=a+bi} + \underbrace{\boldsymbol{x}^*\boldsymbol{x}}_{=1} + \underbrace{\boldsymbol{x}^*\widetilde{\mathsf{L}}^T\boldsymbol{x}}_{=a-bi} = 1 + 2a.$$

$\square$

Outline

Krylov Methods

We end with a very brief overview of Krylov methods.

This class of methods includes many of the state-of-the-art numerical methods for solving

$$A\boldsymbol{x} = \boldsymbol{b} \quad \text{or} \quad A\boldsymbol{x} = \lambda\boldsymbol{x}.$$

Some examples include:

- Linear system solvers:
  - conjugate gradient (CG), biconjugate gradient (BiCG), biconjugate gradient stabilized (BiCGSTAB), minimal residual (MINRES), generalized minimum residual (GMRES)
- Eigensolvers:
  - Lanczos iteration, Arnoldi iteration

The basic building blocks for all these methods are

Definition

For an $n \times n$ matrix A and nonzero n-vector $\boldsymbol{b}$ we define

Krylov sequence: $\{\boldsymbol{b}, \mathsf{A}\boldsymbol{b}, \mathsf{A}^2\boldsymbol{b}, \ldots\}$,

Krylov subspace: $\mathcal{K}_j = \text{span}\{\boldsymbol{b}, \mathsf{A}\boldsymbol{b}, \ldots, \mathsf{A}^{j-1}\boldsymbol{b}\}$,

Krylov matrix: $\mathsf{K} = \begin{pmatrix} \boldsymbol{b} & \mathsf{A}\boldsymbol{b} & \cdots & \mathsf{A}^{j-1}\boldsymbol{b} \end{pmatrix}$.

Consider

$$\begin{aligned} \mathsf{AK} &= \begin{pmatrix} \mathsf{A}\boldsymbol{b} & \mathsf{A}^2\boldsymbol{b} & \cdots & \mathsf{A}^j\boldsymbol{b} \end{pmatrix} \\ &= \mathsf{K}\begin{pmatrix} \boldsymbol{e}_2 & \boldsymbol{e}_3 & \cdots & \boldsymbol{e}_j & -\boldsymbol{c} \end{pmatrix}, \end{aligned}$$

where $\boldsymbol{c} = -\mathsf{K}^{-1}\mathsf{A}^j\boldsymbol{b}$.

Note that the first $j-1$ columns of AK coincide with columns 2 to j of K.

Consider

$$\begin{aligned} \mathsf{AK} &= \begin{pmatrix} \mathsf{A}\boldsymbol{b} & \mathsf{A}^2\boldsymbol{b} & \cdots & \mathsf{A}^j\boldsymbol{b} \end{pmatrix} \\ &= \mathsf{K}\begin{pmatrix} \boldsymbol{e}_2 & \boldsymbol{e}_3 & \cdots & \boldsymbol{e}_j & -\boldsymbol{c} \end{pmatrix}, \end{aligned}$$

where $\boldsymbol{c} = -\mathsf{K}^{-1}\mathsf{A}^j\boldsymbol{b}$.
Note that the first $j-1$ columns of AK coincide with columns 2 to j of K.
Letting $\mathsf{C} = \begin{pmatrix} \boldsymbol{e}_2 & \boldsymbol{e}_3 & \cdots & \boldsymbol{e}_j & -\boldsymbol{c} \end{pmatrix}$ we therefore have

$$\mathsf{AK} = \mathsf{KC} \quad \Longleftrightarrow \quad \mathsf{K}^{-1}\mathsf{AK} = \mathsf{C},$$

i.e., A and C are similar and have the same eigenvalues.

Consider

$$\begin{aligned}\mathsf{AK} &= \begin{pmatrix}\mathsf{A}\boldsymbol{b} & \mathsf{A}^2\boldsymbol{b} & \cdots & \mathsf{A}^j\boldsymbol{b}\end{pmatrix}\\ &= \mathsf{K}\begin{pmatrix}\boldsymbol{e}_2 & \boldsymbol{e}_3 & \cdots & \boldsymbol{e}_j & -\boldsymbol{c}\end{pmatrix},\end{aligned}$$

where $\boldsymbol{c} = -\mathsf{K}^{-1}\mathsf{A}^j\boldsymbol{b}$.
Note that the first $j-1$ columns of AK coincide with columns 2 to j of K.
Letting $\mathsf{C} = \begin{pmatrix}\boldsymbol{e}_2 & \boldsymbol{e}_3 & \cdots & \boldsymbol{e}_j & -\boldsymbol{c}\end{pmatrix}$ we therefore have

$$\mathsf{AK} = \mathsf{KC} \quad \Longleftrightarrow \quad \mathsf{K}^{-1}\mathsf{AK} = \mathsf{C},$$

i.e., A and C are similar and have the same eigenvalues.

Remark

The matrix C is called a companion matrix. It is upper Hessenberg, i.e., upper triangular with an additional nonzero subdiagonal. Computation with such matrices can be performed quite efficiently.

If $j = n$ and we use exact arithmetic then $\mathcal{K}_n = R(\mathsf{A})$.

If $j = n$ and we use exact arithmetic then $\mathcal{K}_n = R(\mathsf{A})$.

Since we know that $\boldsymbol{x} \in R(\mathsf{A})$, the fundamental idea of a Krylov method is to

- iteratively produce approximate solutions $\boldsymbol{x}_j$ that are projections into $\mathcal{K}_j$
- with the hope that low-dimensional Krylov subspaces already contain most of the essential information about $R(\mathsf{A})$.

If $j = n$ and we use exact arithmetic then $\mathcal{K}_n = R(\mathsf{A})$.

Since we know that $\boldsymbol{x} \in R(\mathsf{A})$, the fundamental idea of a Krylov method is to

- iteratively produce approximate solutions $\boldsymbol{x}_j$ that are projections into $\mathcal{K}_j$
- with the hope that low-dimensional Krylov subspaces already contain most of the essential information about $R(\mathsf{A})$.

The main practical problem with Krylov subspaces is that the vectors $\mathsf{A}^j\boldsymbol{b}$ all approach the dominant eigenvector of A (cf. power method), and so the Krylov matrix K becomes ill-conditioned.

The goal of all Krylov methods now is to find better bases for the Krylov subspaces $\mathcal{K}_j$.

The goal of all Krylov methods now is to find better bases for the Krylov subspaces $\mathcal{K}_j$.

This is essentially done via QR factorization, i.e., $\mathsf{K} = \mathsf{QR}$ leads to

$$\begin{aligned} \mathsf{AK} = \mathsf{KC} \quad &\Longleftrightarrow \quad \mathsf{AQR} = \mathsf{QRC} \\ &\Longleftrightarrow \quad \mathsf{Q}^T\mathsf{AQ} = \mathsf{RCR}^{-1} = \mathsf{H}, \end{aligned}$$

where H is another upper Hessenberg matrix.

Arnoldi iteration

Arnoldi iteration is the standard algorithm used to find the matrices Q and H.
At the j^{th} iteration it will produce matrices

- Q_j, $n \times j$ with orthogonal columns that form a basis for $\mathcal{K}_j$;
- Q_{j+1}, $n \times j+1$ with orthogonal columns that form a basis for $\mathcal{K}_{j+1}$;
- $\widetilde{\mathsf{H}}_j$, upper Hessenberg.

These matrices satisfy

$$\mathsf{A}\mathsf{Q}_j = \mathsf{Q}_{j+1}\widetilde{\mathsf{H}}_j.$$

GMRES

The GMRES methods attempts to solve $A\boldsymbol{x} = \boldsymbol{b}$ by minimizing the residual $\|\boldsymbol{b} - A\boldsymbol{x}_j\|_2$ at each iteration.

GMRES

The GMRES methods attempts to solve $A\boldsymbol{x} = \boldsymbol{b}$ by minimizing the residual $\|\boldsymbol{b} - A\boldsymbol{x}_j\|_2$ at each iteration.

Since the approximate solution $\boldsymbol{x}_j \in \mathcal{K}_j$ we can express it using an orthogonal basis, i.e.,

$$\boldsymbol{x}_j = Q_j \boldsymbol{z},$$

for an appropriate $\boldsymbol{z}$.

GMRES

The GMRES methods attempts to solve $A\boldsymbol{x} = \boldsymbol{b}$ by minimizing the residual $\|\boldsymbol{b} - A\boldsymbol{x}_j\|_2$ at each iteration.

Since the approximate solution $\boldsymbol{x}_j \in \mathcal{K}_j$ we can express it using an orthogonal basis, i.e.,

$$\boldsymbol{x}_j = Q_j\boldsymbol{z},$$

for an appropriate $\boldsymbol{z}$.
Then

$$\|\boldsymbol{b} - A\boldsymbol{x}_j\|_2 = \|\boldsymbol{b} - AQ_j\boldsymbol{z}\|_2 = \|\boldsymbol{b} - Q_{j+1}\widetilde{H}_j\boldsymbol{z}\|_2.$$

GMRES

The GMRES methods attempts to solve $A\boldsymbol{x} = \boldsymbol{b}$ by minimizing the residual $\|\boldsymbol{b} - A\boldsymbol{x}_j\|_2$ at each iteration.

Since the approximate solution $\boldsymbol{x}_j \in \mathcal{K}_j$ we can express it using an orthogonal basis, i.e.,

$$\boldsymbol{x}_j = Q_j \boldsymbol{z},$$

for an appropriate $\boldsymbol{z}$.
Then

$$\|\boldsymbol{b} - A\boldsymbol{x}_j\|_2 = \|\boldsymbol{b} - AQ_j\boldsymbol{z}\|_2 = \|\boldsymbol{b} - Q_{j+1}\widetilde{H}_j\boldsymbol{z}\|_2.$$

Multiplication by an orthogonal matrix does not change the 2-norm, so

$$\|\boldsymbol{b} - A\boldsymbol{x}_j\|_2 = \|Q_{j+1}^T\boldsymbol{b} - \underbrace{Q_{j+1}^T Q_{j+1}}_{=I}\widetilde{H}_j\boldsymbol{z}\|_2.$$

GMRES

The GMRES methods attempts to solve $A\boldsymbol{x} = \boldsymbol{b}$ by minimizing the residual $\|\boldsymbol{b} - A\boldsymbol{x}_j\|_2$ at each iteration.

Since the approximate solution $\boldsymbol{x}_j \in \mathcal{K}_j$ we can express it using an orthogonal basis, i.e.,

$$\boldsymbol{x}_j = Q_j \boldsymbol{z},$$

for an appropriate $\boldsymbol{z}$.
Then

$$\|\boldsymbol{b} - A\boldsymbol{x}_j\|_2 = \|\boldsymbol{b} - AQ_j\boldsymbol{z}\|_2 = \|\boldsymbol{b} - Q_{j+1}\widetilde{H}_j\boldsymbol{z}\|_2.$$

Multiplication by an orthogonal matrix does not change the 2-norm, so

$$\|\boldsymbol{b} - A\boldsymbol{x}_j\|_2 = \|Q_{j+1}^T\boldsymbol{b} - \underbrace{Q_{j+1}^T Q_{j+1}}_{=I}\widetilde{H}_j\boldsymbol{z}\|_2.$$

The minimizer $\boldsymbol{z}$ of the 2-norm on the right can be computed efficiently, and $\boldsymbol{x}_j = Q_j\boldsymbol{z}$.
More details are provided, e.g., in [Mey00].

References I

[Ber09] Dennis S. Bernstein, *Matrix Mathematics: Theory, Facts, and Formulas*, 2nd ed., Princeton University Press, Princeton, N.J., July 2009.

[Mey00] Carl D. Meyer, *Matrix Analysis and Applied Linear Algebra*, SIAM, Philadelphia, PA, 2000.

[MVL78] C. Moler and C. Van Loan, *Nineteen Dubious Ways to Compute the Exponential of a Matrix*, SIAM Rev. **20** (1978), no. 4, 801–836.

[MVL03] ______, *Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later*, SIAM Rev. **45** (2003), no. 1, 3–49.

[YM14] Xiyang I. A. Yang and Rajat Mittal, *Acceleration of the Jacobi iterative method by factors exceeding 100 using scheduled relaxation*, Journal of Computational Physics **274** (2014), 695–708.