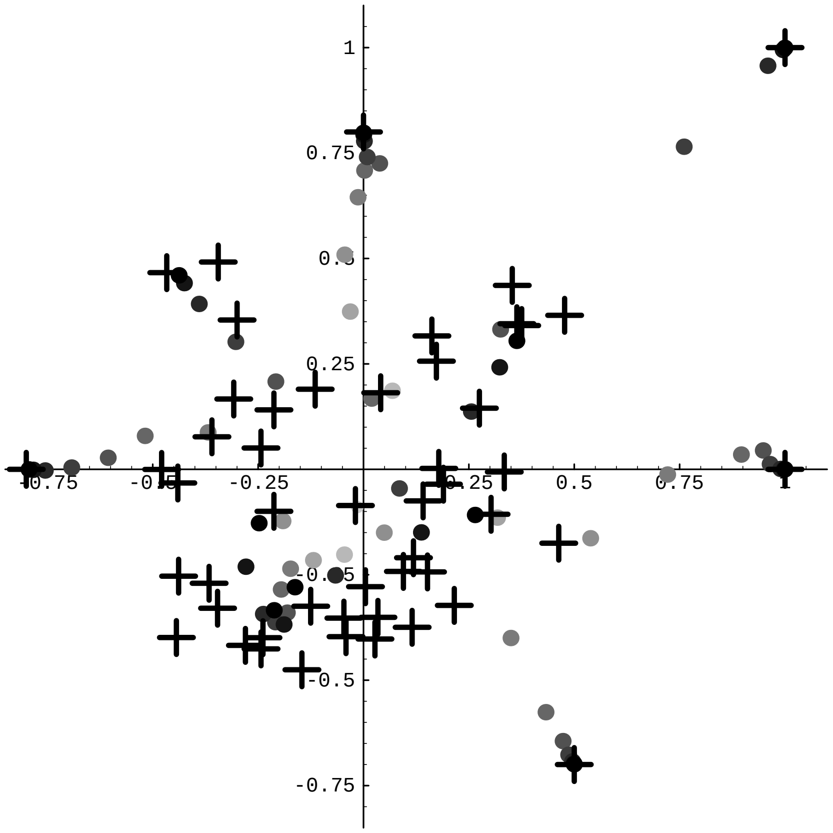
NUMERICAL LINEAR ALGEBRA; ITERATIVE METHODS

TIMO EIROLA AND OLAVI NEVANLINNA

Mat-1.175 Lecture notes
Spring 2003



Institute of Mathematics
Helsinki University of Technology
timo.eirola@hut.fi
olavi.nevanlinna@hut.fi

[0] Version: April 9, 2003

Contents

1. Introduction

These lecture notes consider techniques that are used for large systems of linear equations

$$\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b} \tag{1.1}$$

and eigenvalue problems of large matrices

$$\boldsymbol{A}\,\boldsymbol{x} = \lambda\,\boldsymbol{x}\ . \tag{1.2}$$

We will consider the following issues:

- sparsity: the number of nonzero entries in the matrix; how algorithms get faster, when this is small
- linearity: these techniques make essential use of the linearity of the system
- complexity: the number of floating point operations needed to obtain the solution
- round-off errors: how they affect the solution
- a priori / a posteriori error analysis

The methods can be divided in two classes which have the following characteristics:

Direct methods	**Krylov methods**
matrix $\boldsymbol{A}$ is given	only the operator $\boldsymbol{v} \to \boldsymbol{A}\boldsymbol{v}$ is available
manipulate the matrix	work in the subspace $\mathrm{span}\{\boldsymbol{b}, \boldsymbol{A}\boldsymbol{b}, \ldots, \boldsymbol{A}^{k-1}\boldsymbol{b}\}$
work in $\mathbb{C}^n$ or $\mathbb{R}^n$	work in $\mathbb{C}^n$, $\mathbb{R}^n$, or Hilbert space
dimension n is visible	dimension not visible
"exact" solution	approximate solution
all eigenvalues	only interesting eigenvalues

In applications large matrices are typically sparse so that multiplication of vectors with them is computationally a cheap operation. The iterative methods for linear equations and eigenvalue problems exploit this property. In these methods the main computational units are matrix–vector products, inner products of vectors, and formation of linear combinations of vectors. These are very different from the techniques of "massaging the matrix" that classical direct methods for these problems do (though iteratively for eigenvalue problems).

In these notes we first give a short review of direct techniques – *LU*– and *QR*–decompositions – for linear equations in chapters 2 and 4, respectively. Then in chapter 5 the basic eigendecompositions are reviewed. Chapter 6 is devoted to sensitivity issues of eigenvalues and the standard *QR*–iteration for computing them. In

[0] Version: April 9, 2003

chapter 7 we consider the theoretical properties of eigenvalues of Hermitian matrices and in chapter 8 some classical iterations to compute them.

Chapter 9 starts the main topic of these notes the *Krylov subspace* iterations. In these methods approximations for (1.1) and (1.2) are searched in the increasing sequence of subspaces spanned by vectors

$$\boldsymbol{b},\ \boldsymbol{A}\boldsymbol{b},\ \boldsymbol{A}^2\boldsymbol{b},\ \boldsymbol{A}^3\boldsymbol{b}, \dots\ . \tag{1.3}$$

First we study the Krylov subspace iterations for eigenvalue problems. Then in chapter 10 we consider the classical iterations for linear systems. The powerful conjugate gradient iteration for Hermitian systems is studied in chapter 11 and finally starting in chapter 12 the modern Krylov subspace iterations for unsymmetric linear problems are derived and analyzed.

1.1. Notation and basic properties.

The entries of a matrix $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ are denoted by $a_{ij} = (\boldsymbol{A})_{ij}$ and if the vectors $\boldsymbol{a}_j \in \mathbb{C}^m$ form the columns of A we have

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} = \boldsymbol{a}_1\boldsymbol{e}_1^* + \boldsymbol{a}_2\boldsymbol{e}_2^* + \cdots + \boldsymbol{a}_n\boldsymbol{e}_n^* \ ,$$

where the standard basis vectors $\boldsymbol{e}_j$ are the columns of the identity matrix $\boldsymbol{I}$. Matrix $\boldsymbol{A}$ defines a linear operator mapping $\boldsymbol{v} \in \mathbb{C}^n$ to $\boldsymbol{Av} \in \mathbb{C}^m$.

(1) Matrix $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ is a *square matrix*, if $m = n$.

(2) Matrix $\boldsymbol{A}$ on *upper(lower)triangular*, if $a_{ij} = 0$ for $i > j$ ($i < j$).

(3) $\boldsymbol{A}$ is a *diagonal matrix*, if it is both upper and lower triangular, i.e., if $a_{ij} = 0$, for $i \neq j$.

(4) The *product* $\boldsymbol{AB} \in \mathbb{C}^{m\times n}$ of matrices $\boldsymbol{A} \in \mathbb{C}^{m\times p}$ and $\boldsymbol{B} \in \mathbb{C}^{p\times n}$ is defined by

$$(\boldsymbol{AB})_{ij} = \textstyle\sum_{k=1}^{p} a_{ik}\, b_{kj} \ .$$

Matrix product is associative: $\boldsymbol{A}(\boldsymbol{BC}) = (\boldsymbol{AB})\boldsymbol{C}$, but not commutative: generally $\boldsymbol{AB} \neq \boldsymbol{BA}$.

(5) The determinant of a square matrix is denoted by $\det(\boldsymbol{A})$. The determinant of a triangular matrix is the product of its diagonal elements and

$$\det(\boldsymbol{AB}) = \det(\boldsymbol{A})\det(\boldsymbol{B}) \ .$$

(6) Identity matrix $\boldsymbol{I}$ satisfies $\boldsymbol{AI} = \boldsymbol{IA} = \boldsymbol{A}$ for all matrices $\boldsymbol{A}$.

(7) $\boldsymbol{A}$ is called *nonsingular*, if $\det(\boldsymbol{A}) \neq 0$. Then it has an *inverse* $\boldsymbol{A}^{-1}$. This satisfies

$$\boldsymbol{AA}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I} \ .$$

(8) The product of two lower triangular matrices is lower triangular. The inverse of a nonsingular lower triangular matrix is lower triangular. Similarly for upper triangular matrices.

(9) The transpose of $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ is denoted by $\boldsymbol{A}^T \in \mathbb{C}^{n\times m}$. It is defined by $(\boldsymbol{A}^T)_{ij} = (\boldsymbol{A})_{ji}$. Matrix $\boldsymbol{A}^* = \overline{\boldsymbol{A}^T}$ is called the *adjoint* of $\boldsymbol{A}$.

(10) $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is *symmetric*, if $\boldsymbol{A}^T = \boldsymbol{A}$.

(11) $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is *skew symmetric*, if $\boldsymbol{A}^T = -\boldsymbol{A}$.

(12) $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is *Hermitian*, if $\boldsymbol{A}^* = \boldsymbol{A}$.

(13) $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is *orthogonal*, if $\boldsymbol{A}^T = \boldsymbol{A}^{-1}$.

(14) $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is *unitary*, if $\boldsymbol{A}^* = \boldsymbol{A}^{-1}$.

(15) The matrix product satisfies $(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T$ and for invertible matrices $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.

(16) Let $\boldsymbol{A} \in \mathbb{C}^{m\times n}$. The *nullspace* (or *kernel*) of $\boldsymbol{A}$ is

$$N(\boldsymbol{A}) = \{\boldsymbol{x} \in \mathbb{C}^n \mid \boldsymbol{Ax} = 0\} ,$$

and the *range* is

$$R(\boldsymbol{A}) = \{\boldsymbol{Ax} \in \mathbb{C}^m \mid \boldsymbol{x} \in \mathbb{C}^n\} .$$

The following contains the main properties of invertible matrices

Theorem 1.1. *Let $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. Then the following are equivalent*

(1) $\boldsymbol{A}^{-1}$ *exists.*

(2) $\det(\boldsymbol{A}) \neq 0$.

(3) *The system $\boldsymbol{Ax} = \boldsymbol{b}$ has a unique solution for all $\boldsymbol{b}$.*

(4) *The columns of $\boldsymbol{A}$ are linearly independent.*

(5) *The rows of $\boldsymbol{A}$ are linearly independent.*

(6) $N(\boldsymbol{A}) = \{0\}$.

(7) $R(\boldsymbol{A}) = \mathbb{C}^n$.

The *dimension* of a space is the maximum number of linearly idependent vectors in it. The fundamental theorem of linear algebra (the dimension theorem) says that for every matrix $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ holds

$$\dim(R(\boldsymbol{A})) + \dim(N(\boldsymbol{A})) = n .$$

Let V and W be subspaces of a finite dimensional vector space U. By $V + W$ we mean the set $\big\{\boldsymbol{v} + \boldsymbol{w} \,\big|\, \boldsymbol{v} \in V ,\ \boldsymbol{w} \in W\big\}$. This is also a subspace and we have

$$\dim(V + W) = \dim(V) + \dim(W) - \dim(V \cap W) .$$

We write $U = V \oplus W$ if every vector $\boldsymbol{u} \in U$ can be written in a unique way as $\boldsymbol{u} = \boldsymbol{v} + \boldsymbol{w}$, where $\boldsymbol{v} \in V$ and $\boldsymbol{w} \in W$. This is equivalent to

$$V + W = U \qquad \text{and} \qquad V \cap W = \{0\} \ .$$

In this case $\dim(U) = \dim(V) + \dim(W)$.

The set of eigenvalues of $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is denoted by $\Lambda(\boldsymbol{A})$. It is called the *spectrum* of $\boldsymbol{A}$ and is defined by

$$\Lambda(\boldsymbol{A}) = \big\{\lambda \in \mathbb{C} \,\big|\, N(\lambda \boldsymbol{I} - \boldsymbol{A}) \neq \{0\}\big\} \ .$$

The p–norm of a vector $\boldsymbol{x} \in \mathbb{C}^n$ ($1 \leq p < \infty$) is defined by

$$\|\boldsymbol{x}\|_p = \Big(\textstyle\sum_{j=1}^n |x_j|^p\Big)^{\frac{1}{p}}$$

and $\|\boldsymbol{x}\|_\infty = \max_j |x_j|$. The 2–norm is called Euclidean and is denoted $\|\boldsymbol{x}\| = \|\boldsymbol{x}\|_2$. For a matrix $\boldsymbol{A} \in C^{m\times n}$ we set

$$\|\boldsymbol{A}\|_{p,q} = \max_{\|\boldsymbol{x}\|_p = 1} \|\boldsymbol{A}\boldsymbol{x}\|_q \ ,$$

$\|\boldsymbol{A}\|_p = \|\boldsymbol{A}\|_{p,p}$ and $\|\boldsymbol{A}\| = \|\boldsymbol{A}\|_2$. Further, the Frobenius-norm is defined as

$$\|\boldsymbol{A}\|_F = (\mathrm{tr}(\boldsymbol{A}^*\boldsymbol{A}))^{\frac{1}{2}} = \Big(\textstyle\sum_{i=1}^m \sum_{j=1}^n |a_{i,j}|^2\Big)^{\frac{1}{2}} = \Big[\displaystyle\sum_j \sigma_j(\boldsymbol{A})^2\Big]^{\frac{1}{2}}$$

and the trace–norm

$$\|\boldsymbol{A}\|_\sigma = \sum_j \sigma_j(\boldsymbol{A}) \ ,$$

where $\sigma_j(\boldsymbol{A})$, $j = 1, \ldots, n$ are the singular values of $\boldsymbol{A}$, i.e., the square roots of the eigenvalues of the matrix $\boldsymbol{A}^*\boldsymbol{A}$. We have: $\|\boldsymbol{A}\|_2 = \sigma_1(\boldsymbol{A})$ (the largest singular value).

The standard inner product of vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^n$ is denoted by

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \sum_j x_j\, \bar{y}_j \, .$$

The following *Hölder inequality* is left as an exercise (consulting literature is allowed).

Lemma 1.2. *If* $\frac{1}{p} + \frac{1}{q} = 1$, *then* $|\langle \boldsymbol{x}, \boldsymbol{y}\rangle| \leq \|\boldsymbol{x}\|_p \, \|\boldsymbol{y}\|_q$.

The special case $p = q = 2$ of this is the Cauchy–Schwartz–Bunyakovsky inequality: $|\langle \boldsymbol{x}, \boldsymbol{y}\rangle| \le \|\boldsymbol{x}\| \, \|\boldsymbol{y}\|$.

For $S \subset \mathbb{C}^n$ we denote by $S^\perp$ the set of vectors orthogonal to all vectors of S :

$$S^\perp = \left\{ \boldsymbol{v} \in \mathbb{C}^n \;\middle|\; \langle \boldsymbol{s}, \boldsymbol{v}\rangle = 0 \ \forall \boldsymbol{s} \in S \right\} .$$

The geometric form of the fundamental theorem of linear algebra is that any matrix $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ satisfies

$$N(\boldsymbol{A}) = R(\boldsymbol{A}^*)^\perp .$$

If V is a subspace of $\mathbb{C}^n$, then $\mathbb{C}^n = V \oplus V^\perp$.

If $(\boldsymbol{v}_1, \boldsymbol{v}_2, \dots)$ is a sequence of linearly independent vectors then the *Gram–Schmidt process* constructs an orthonormal sequence $(\boldsymbol{q}_1, \boldsymbol{q}_2, \dots)$ in the following way:

$$\begin{aligned} \boldsymbol{q}_1 &= \boldsymbol{v}_1 / \|\boldsymbol{v}_1\| \ , \\ \left.\begin{aligned} \boldsymbol{w}_k &= \boldsymbol{v}_k - \textstyle\sum_{j=1}^{k-1} \langle \boldsymbol{v}_k, \boldsymbol{q}_j\rangle \, \boldsymbol{q}_j \ , \\ \boldsymbol{q}_k &= \boldsymbol{w}_k / \|\boldsymbol{w}_k\| \ . \end{aligned}\right\} & \quad k = 2, 3, \dots \end{aligned} \tag{1.4}$$

This sequence satisfies

a) $(\boldsymbol{q}_1, \boldsymbol{q}_2, \dots)$ is orthonormal.

b) $\operatorname{span}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_k) = \operatorname{span}(\boldsymbol{v}_1, \dots, \boldsymbol{v}_k)$ for all $k \ge 1$.

By: $|\boldsymbol{x}|$ (respectively $|\boldsymbol{A}|$) we mean the vector (matrix) formed from the absolute values:

$$|\boldsymbol{x}| = \begin{bmatrix} |x_1| \\ |x_2| \\ \vdots \\ |x_n| \end{bmatrix} , \qquad |\boldsymbol{A}| = \begin{bmatrix} |A_{1,1}| & |A_{1,2}| & \dots & |A_{1,n}| \\ |A_{2,1}| & |A_{2,2}| & \dots & |A_{2,n}| \\ \vdots & \vdots & & \vdots \\ |A_{m,1}| & |A_{m,2}| & \dots & |A_{m,n}| \end{bmatrix} .$$

Notation $\boldsymbol{x} \preceq \boldsymbol{y}$ means $x_j \le y_j$ for all j ($\boldsymbol{A} \preceq \boldsymbol{B} \iff A_{i,j} \le B_{i,j} \ \forall i, j$).

We use the `Matlab`–notation for parts of matrices:

$$\boldsymbol{A}(i:j, k:l) = \begin{bmatrix} A_{i,k} & \dots & A_{i,l} \\ \vdots & & \vdots \\ A_{j,k} & \dots & A_{j,l} \end{bmatrix} .$$

The rounding operation of floating point numbers is denoted by fl . We assume constantly that the computer performs the operation $z = x \circ y$, where $\circ \in \{+, -, *, /\}$ such that first it computes the exact value, and then the result is rounded to the closest floating point number. We idealise the floating point arithmetic further by assuming that the exponential part of the floating point numbers is unbounded, i.e., under– or overflows do not happen. Thus we have the following set of floating point numbers in binary form with some fixed $s \in \mathbb{N}$:

$$\mathbb{F} = \left\{ \pm 0.d_1, d_2 \dots d_s \cdot 2^k \,\middle|\, d_i \in \{0,1\}\,,\ k \in \mathbb{Z} \right\} \ .$$

Then $fl \ : \ \mathbb{R} \to \mathbb{F}$ and we have: the result of a floating point operation on the computer satisfies

$$\widehat{z} = fl(x \circ y) = x \circ y + e \qquad \text{where} \qquad |e| \leq \mu \, |x \circ y| \ .$$

Here $\mu = 2^{-s}$ is the machine constant, typically $\approx 10^{-16}$.

Problem 1.1. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are $n \times n$– floating point matrices, then which inequality of the form $|\boldsymbol{E}| \preceq \ ?$ you get for the matrix product:

$$\widehat{\boldsymbol{AB}} = \boldsymbol{AB} + \boldsymbol{E} \ ?$$

The elementary permutation matrix is denoted by $\boldsymbol{\Pi}(j,k)$. Multiplication by this changes the j :th and the k :th component: $\boldsymbol{y} = \boldsymbol{\Pi}(j,k)\,\boldsymbol{x}$ has $y_k = x_j$, $y_j = x_k$, $y_i = x_i$, $i \neq j,k$.

2. Linear system $\boldsymbol{Ax}=\boldsymbol{b}$, elimination, $\boldsymbol{LU}$–decomposition.

Since the linear system $\boldsymbol{Ly}=\boldsymbol{b}$, where $\boldsymbol{L}$ is a lower triangular matrix, is easy to solve with recursive substitution as well as a system with an upper triangular matrix, the basic strategy for solving the linear system $\boldsymbol{Ax}=\boldsymbol{b}$ starts with the attempt to write $\boldsymbol{A}$ as a product $\boldsymbol{A}=\boldsymbol{LU}$ of a lower triangular matrix $\boldsymbol{L}$ and an upper triangular matrix $\boldsymbol{U}$. Then our system is equivalent to two easy ones:

$$\boldsymbol{Ax}=\boldsymbol{b} \iff \boldsymbol{LUx}=\boldsymbol{b} \iff \begin{cases} \boldsymbol{Ly}=\boldsymbol{b} \\ \boldsymbol{Ux}=\boldsymbol{y} \end{cases} .$$

2.1. $\boldsymbol{LU}$–decomposition.

This is obtained using *elimination* as follows:

Assume, that $a_{1,1}\neq 0$. Then

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ -\frac{a_{2,1}}{a_{1,1}} & 1 & 0 & \dots & 0 \\ -\frac{a_{3,1}}{a_{1,1}} & 0 & 1 & \dots & 0 \\ \vdots & & & \ddots & \\ -\frac{a_{n,1}}{a_{1,1}} & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & a_{2,3} & \dots & a_{2,n} \\ a_{3,1} & a_{3,2} & a_{3,3} & \dots & a_{3,n} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & a_{n,3} & \dots & a_{n,n} \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \dots & a_{1,n} \\ 0 & & & & \\ 0 & & \widetilde{\boldsymbol{A}} & & \\ \vdots & & & & \\ 0 & & & & \end{bmatrix} ,$$

i.e., $\boldsymbol{M}_1\boldsymbol{A}=\boldsymbol{A}_1$, where the part below the diagonal of the first column of $\boldsymbol{A}_1$ is zero. If (the pivot element) $\widetilde{a}_{1,1}\neq 0$, then we can continue:

$$\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & -\frac{\widetilde{a}_{2,1}}{\widetilde{a}_{1,1}} & 1 & \dots & 0 \\ \vdots & & & \ddots & \\ 0 & -\frac{\widetilde{a}_{n-1,1}}{\widetilde{a}_{1,1}} & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ 0 & \widetilde{a}_{1,1} & \dots & \widetilde{a}_{1,n-1} \\ 0 & \widetilde{a}_{2,1} & \dots & \widetilde{a}_{2,n-1} \\ \vdots & \vdots & & \vdots \\ 0 & \widetilde{a}_{n-1,1} & \dots & \widetilde{a}_{n-1,n-1} \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \dots & a_{1,n} \\ 0 & \widetilde{a}_{1,1} & \widetilde{a}_{1,2} & \dots & \widetilde{a}_{1,n-1} \\ 0 & 0 & & & \\ & \vdots & & \widetilde{\widetilde{\boldsymbol{A}}} & \\ 0 & 0 & & & \end{bmatrix} ,$$

i.e., $\boldsymbol{M}_2\boldsymbol{A}_1=\boldsymbol{M}_2\boldsymbol{M}_1\boldsymbol{A}=\boldsymbol{A}_2$. If the resulting square matrix with decreasing dimension never has zero pivot element (upper left corner) we get:

$$\boldsymbol{M}_{n-1}\dots\boldsymbol{M}_2\boldsymbol{M}_1\boldsymbol{A}=\boldsymbol{U} ,$$

where $\boldsymbol{U}$ is an upper triangular matrix. The matrices $\boldsymbol{M}_k$ are lower triangular, so that their product is also such. In order to form the decomposition $\boldsymbol{A}=\boldsymbol{LU}$ we need to compute the product

$$\boldsymbol{L}=\boldsymbol{M}_1^{-1}\boldsymbol{M}_2^{-1}\dots\boldsymbol{M}_{n-1}^{-1} .$$

[0] Version: April 9, 2003

Now each of the *Gauss transforms* $\boldsymbol{M}_k$ is of the form

$$\boldsymbol{M}_k = \boldsymbol{I} - \boldsymbol{t}_k \boldsymbol{e}_k^T \ , \quad \boldsymbol{t}_k = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \tau_{k+1}^k \\ \vdots \\ \tau_n^k \end{bmatrix} \ , \quad \boldsymbol{e}_k = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \ .$$

Lemma 2.1. *If* $\boldsymbol{v}^T\boldsymbol{u} \neq 1$ *then* $\boldsymbol{I} - \boldsymbol{u}\boldsymbol{v}^T$ *is invertible and* $(\boldsymbol{I} - \boldsymbol{u}\boldsymbol{v}^T)^{-1} = \boldsymbol{I} + \frac{\boldsymbol{u}\boldsymbol{v}^T}{1-\boldsymbol{v}^T\boldsymbol{u}}$.

Proof.

$$(\boldsymbol{I} + \tfrac{\boldsymbol{u}\boldsymbol{v}^T}{1-\boldsymbol{v}^T\boldsymbol{u}})(\boldsymbol{I} - \boldsymbol{u}\boldsymbol{v}^T) = \boldsymbol{I} - \boldsymbol{u}\boldsymbol{v}^T + \tfrac{1}{1-\boldsymbol{v}^T\boldsymbol{u}}\boldsymbol{u}\boldsymbol{v}^T - \tfrac{1}{1-\boldsymbol{v}^T\boldsymbol{u}}\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{u}\boldsymbol{v}^T = \boldsymbol{I} \ .$$

□

Since $\boldsymbol{e}_k^T\boldsymbol{t}_k = 0$, we now get $(\boldsymbol{I} - \boldsymbol{t}_k\boldsymbol{e}_k^T)^{-1} = \boldsymbol{I} + \boldsymbol{t}_k\boldsymbol{e}_k^T$. Generally: $j \leq k \implies \boldsymbol{e}_j^T\boldsymbol{t}_k = 0$, so that

$$\begin{aligned} \boldsymbol{L} &= (\boldsymbol{I} + \boldsymbol{t}_1\boldsymbol{e}_1^T)(\boldsymbol{I} + \boldsymbol{t}_2\boldsymbol{e}_2^T)\dots(\boldsymbol{I} + \boldsymbol{t}_{n-1}\boldsymbol{e}_{n-1}^T) \\ &= (\boldsymbol{I} + \boldsymbol{t}_1\boldsymbol{e}_1^T + \boldsymbol{t}_2\boldsymbol{e}_2^T)(\boldsymbol{I} + \boldsymbol{t}_3\boldsymbol{e}_3^T)\dots(\boldsymbol{I} + \boldsymbol{t}_{n-1}\boldsymbol{e}_{n-1}^T) \\ &= \boldsymbol{I} + \sum_k \boldsymbol{t}_k\boldsymbol{e}_k^T \\ &= \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ \frac{a_{2,1}}{a_{1,1}} & 1 & 0 & \dots & 0 \\ \frac{a_{3,1}}{a_{1,1}} & \frac{\tilde{a}_{2,1}}{\tilde{a}_{1,1}} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ \frac{a_{n,1}}{a_{1,1}} & \frac{\tilde{a}_{n-1,1}}{\tilde{a}_{1,1}} & \frac{\tilde{\tilde{a}}_{n-2,1}}{\tilde{\tilde{a}}_{1,1}} & \dots & 1 \end{bmatrix} \ . \end{aligned}$$

In other words, $\boldsymbol{L}$ is formed easily by collecting the nonzero parts of the $\boldsymbol{t}_k$–vectors of the Gauss transforms.

Problem 2.1. What is the *complexity* of the $\boldsymbol{LU}$–decomposition, i.e., how many floating point operations are needed? (Answer: $\approx \frac{2}{3}n^3$.)

Problem 2.2. Show, that $\det(\boldsymbol{A}_k(1:k, 1:k)) = \det(\boldsymbol{A}(1:k, 1:k))$. In other words the pivot elements never become zero, if and only if all the upper left determinants of $\boldsymbol{A}$ are $\neq 0$.

The following "programmed" algorithm starts from $fl(\boldsymbol{A})$, stores L (except the ones on the diagonal) below the diagonal and $\boldsymbol{U}$ on the diagonal and above it:

```
for $k$ =$1 : n-1$ do
        $t = \frac{1}{A(k,k)}\,A(k+1:n,k)$
        $A(k+1:n,k+1:n) = A(k+1:n,k+1:n) - t\;A(k,k+1:n)$
        $A(k+1:n,k) = t$
end
```

Theorem 2.2. *Let $\widehat{\boldsymbol{A}} = fl(\boldsymbol{A})$. If the pivot elements are nonzero, then the computer produces $\widehat{\boldsymbol{L}}, \widehat{\boldsymbol{U}}$ such that $\widehat{\boldsymbol{L}}\,\widehat{\boldsymbol{U}} = \widehat{\boldsymbol{A}} + \boldsymbol{H}$, where*

$$|\boldsymbol{H}| \preceq 2(n-1)\mu\,(|\widehat{\boldsymbol{A}}| + |\widehat{\boldsymbol{L}}|\,|\widehat{\boldsymbol{U}}|) + O(\mu^2)\ .$$

Proof. By induction: $n=1$: $\widehat{\boldsymbol{U}} = \widehat{\boldsymbol{A}}$, $\boldsymbol{L} = 1$, $\boldsymbol{H} = 0 \implies$ true.

Assume, that the theorem holds for all $(n-1)\times(n-1)$–matrices. Let

$$\widehat{\boldsymbol{A}} = \begin{bmatrix} \alpha & \boldsymbol{w}^T \\ \boldsymbol{v} & \boldsymbol{B} \end{bmatrix} \in \mathbb{R}^{n\times n}\ , \quad \boldsymbol{B} \in \mathbb{R}^{(n-1)\times(n-1)}\ .$$

At the first step of the algorithm:

$$\widehat{\boldsymbol{t}} = fl(\tfrac{1}{\alpha}\,\boldsymbol{v}) = \tfrac{1}{\alpha}\,\boldsymbol{v} + \boldsymbol{f} \quad , \qquad |\boldsymbol{f}| \preceq \tfrac{\mu}{\alpha}\,|\boldsymbol{v}|\ ,$$

$$\widehat{\boldsymbol{A}}^1 := \widehat{\boldsymbol{A}}_1(2:n,2:n) = \boldsymbol{B} - \widehat{\boldsymbol{t}}\boldsymbol{w}^T + \boldsymbol{F} \quad , \qquad |\boldsymbol{F}| \preceq 2\mu(|\boldsymbol{B}| + |\widehat{\boldsymbol{t}}||\boldsymbol{w}|^T) + O(\mu^2)\ .$$

After this the algorithm continues with the matrix $\widehat{\boldsymbol{A}}^1$. By induction hypotheses the result is $\widehat{\boldsymbol{L}}_1\widehat{\boldsymbol{U}}_1 = \widehat{\boldsymbol{A}}^1 + \boldsymbol{H}_1$, where

$$|\boldsymbol{H}_1| \preceq 2(n-2)\mu(|\widehat{\boldsymbol{A}}^1| + |\widehat{\boldsymbol{L}}_1||\widehat{\boldsymbol{U}}_1|) + O(\mu^2)\ ,$$

so that

$$\widehat{\boldsymbol{L}}\widehat{\boldsymbol{U}} = \begin{bmatrix} 1 & 0 \\ \widehat{\boldsymbol{t}} & \widehat{\boldsymbol{L}}_1 \end{bmatrix} \begin{bmatrix} \alpha & \boldsymbol{w}^T \\ 0 & \widehat{\boldsymbol{U}}_1 \end{bmatrix} = \widehat{\boldsymbol{A}} + \begin{bmatrix} 0 & 0 \\ \alpha\boldsymbol{f} & \boldsymbol{H}_1 + \boldsymbol{F} \end{bmatrix} = \widehat{\boldsymbol{A}} + \boldsymbol{H}\ .$$

We get

$$\begin{aligned}
|\widehat{\boldsymbol{A}}^1| &\preceq (1+2\mu)(|\boldsymbol{B}| + |\widehat{\boldsymbol{t}}||\boldsymbol{w}|^T) + O(\mu^2) \\
|\boldsymbol{H}_1 + \boldsymbol{F}| &\preceq 2(n-2)\mu(|\boldsymbol{B}| + |\widehat{\boldsymbol{t}}||\boldsymbol{w}|^T + |\widehat{\boldsymbol{L}}_1||\widehat{\boldsymbol{U}}_1|) + 2\mu(|\boldsymbol{B}| + |\widehat{\boldsymbol{t}}||\boldsymbol{w}|^T) + O(\mu^2) \\
&\preceq 2(n-1)\mu(|\boldsymbol{B}| + |\widehat{\boldsymbol{t}}||\boldsymbol{w}|^T + |\widehat{\boldsymbol{L}}_1||\widehat{\boldsymbol{U}}_1|) + O(\mu^2) \\
|\alpha \boldsymbol{f}| &\preceq \mu |\boldsymbol{v}| \\
|\boldsymbol{H}| &\preceq 2(n-1)\mu \left(\begin{bmatrix} |\alpha| & |\boldsymbol{w}|^T \\ |\boldsymbol{v}| & |\boldsymbol{B}| \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ |\widehat{\boldsymbol{t}}| & |\widehat{\boldsymbol{L}}_1| \end{bmatrix} \begin{bmatrix} |\alpha| & |\boldsymbol{w}|^T \\ 0 & |\widehat{\boldsymbol{U}}_1| \end{bmatrix} \right) + O(\mu^2) \\
&= 2(n-1)\mu(|\widehat{\boldsymbol{A}}| + |\widehat{\boldsymbol{L}}||\widehat{\boldsymbol{U}}|) + O(\mu^2) \ ,
\end{aligned}$$

i.e., the theorem holds for $n \times n$–matrices. □

2.2. **Partial pivoting.** If $\widehat{\boldsymbol{L}}$ or $\widehat{\boldsymbol{U}}$ above contain large elements, then the error $\boldsymbol{H}$ is not necessarily small. A better result is received by *partial pivoting*: find the element with largest absolute value of the first column of $\boldsymbol{A}$. Let it be $a_{k,1}$. Using the permutation matrix $\boldsymbol{\Pi}_1 = \boldsymbol{\Pi}(1,k)$ we change it to the place $1,1$ and eliminate:

$$\boldsymbol{M}_1 \boldsymbol{\Pi}_1 \boldsymbol{A} = \boldsymbol{M}_1 \begin{bmatrix} a_{k,1} & a_{k,2} & \dots & a_{k,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{bmatrix} = \begin{bmatrix} a_{k,1} & a_{k,2} & \dots & a_{k,n} \\ 0 & & & \\ \vdots & & \widetilde{\boldsymbol{A}} & \\ 0 & & & \end{bmatrix} .$$

Next we find the element with largest absolute value in the first column of $\widetilde{\boldsymbol{A}}$, etc. We get

$$\boldsymbol{M}_{n-1}\boldsymbol{\Pi}_{n-1} \dots \boldsymbol{M}_2 \boldsymbol{\Pi}_2 \boldsymbol{M}_1 \boldsymbol{\Pi}_1 \boldsymbol{A} = \boldsymbol{U} \ ,$$

where $\boldsymbol{M}_k = \boldsymbol{I} - \boldsymbol{t}_k \boldsymbol{e}_k^T$ and $|\boldsymbol{t}_k|^T \preceq [0 \dots 0 \ 1 \dots 1]$. Now $j > k \implies \boldsymbol{e}_k^T \boldsymbol{\Pi}_j = \boldsymbol{e}_k^T$, so that

$$\boldsymbol{\Pi}_j \boldsymbol{M}_k = \boldsymbol{\Pi}_j(\boldsymbol{I} - \boldsymbol{t}_k \boldsymbol{e}_k^T) = \boldsymbol{\Pi}_j - \boldsymbol{\Pi}_j \boldsymbol{t}_k \boldsymbol{e}_k^T \boldsymbol{\Pi}_j = (\boldsymbol{I} - \boldsymbol{\Pi}_j \boldsymbol{t}_k \boldsymbol{e}_k^T) \boldsymbol{\Pi}_j \ .$$

Denote: $\widetilde{\boldsymbol{t}}_k = \boldsymbol{\Pi}_{n-1} \dots \boldsymbol{\Pi}_{k+1} \boldsymbol{t}_k$, $\widetilde{\boldsymbol{M}_k} = \boldsymbol{I} - \widetilde{\boldsymbol{t}}_k \boldsymbol{e}_k^T$. Then

$$\boldsymbol{\Pi}_{n-1} \dots \boldsymbol{\Pi}_{k+1} \boldsymbol{M}_k = \widetilde{\boldsymbol{M}_k} \boldsymbol{\Pi}_{n-1} \dots \boldsymbol{\Pi}_{k+1}$$

and

$$\begin{aligned} \boldsymbol{U} &= \boldsymbol{M}_{n-1}\,\boldsymbol{\Pi}_{n-1}\,\boldsymbol{M}_{n-2}\,\boldsymbol{\Pi}_{n-2}\ldots\boldsymbol{M}_1\,\boldsymbol{\Pi}_1\,\boldsymbol{A} = \\ &= \widetilde{\boldsymbol{M}_{n-1}}\,\widetilde{\boldsymbol{M}_{n-2}}\,\boldsymbol{\Pi}_{n-1}\,\boldsymbol{\Pi}_{n-2}\,\boldsymbol{M}_{n-3}\ldots\boldsymbol{M}_1\,\boldsymbol{\Pi}_1\,\boldsymbol{A} = \\ &= \widetilde{\boldsymbol{M}_{n-1}}\,\widetilde{\boldsymbol{M}_{n-2}}\,\widetilde{\boldsymbol{M}_{n-3}}\ldots\widetilde{\boldsymbol{M}_2}\,\boldsymbol{\Pi}_{n-1}\,\boldsymbol{\Pi}_{n-2}\ldots\boldsymbol{\Pi}_2\,\boldsymbol{M}_1\,\boldsymbol{\Pi}_1\,\boldsymbol{A} = \\ &= \widetilde{\boldsymbol{M}_{n-1}}\ldots\widetilde{\boldsymbol{M}_2}\,\widetilde{\boldsymbol{M}_1}\,\boldsymbol{\Pi}_{n-1}\ldots\boldsymbol{\Pi}_2\,\boldsymbol{\Pi}_1\,\boldsymbol{A} \end{aligned}$$

We got: $\boldsymbol{\Pi}\,\boldsymbol{A} = \boldsymbol{L}\,\boldsymbol{U}$, where $\boldsymbol{\Pi} = \boldsymbol{\Pi}_{n-1}\ldots\boldsymbol{\Pi}_2\,\boldsymbol{\Pi}_1$ and as before

$$\boldsymbol{L} = \widetilde{\boldsymbol{M}_1}^{-1}\,\widetilde{\boldsymbol{M}_2}^{-1}\ldots\widetilde{\boldsymbol{M}_{n-1}}^{-1} = \boldsymbol{I} + \widetilde{\boldsymbol{t}}_1\,\boldsymbol{e}_1^T + \cdots + \widetilde{\boldsymbol{t}}_{n-1}\,\boldsymbol{e}_{n-1}^T\ .$$

Now all the elements of $\boldsymbol{L}$ have absolute values at most one.

Problem 2.3. "Program" the algorithm solving the system $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$ using partial pivoting.

Problem 2.4. Show: if $\det(\boldsymbol{A}) \neq 0$, then, using partial pivoting, in exact arithmetic all pivots are nonzero.

Problem 2.5. Prove a version of Theorem 2.2 now for the $\boldsymbol{LU}$–decomposition with partial pivoting.

Problem 2.6. $\boldsymbol{A}$ is *strictly diagonal dominant*, if $|a_{i,i}| > \sum_{j\neq i}|a_{i,j}|$ for all $i = 1,\ldots,n$. Show: if $\boldsymbol{A}^T$ is strictly diagonal dominant, then above $\boldsymbol{\Pi} = \boldsymbol{I}$, i.e., pivoting is not needed.

Problem 2.7. Using partial pivoting we get $\|\boldsymbol{L}\|_\infty \leq n$, but the worst case may produce $\|\boldsymbol{U}\|_\infty \sim 2^n\|\boldsymbol{A}\|_\infty$. Consider the following example:

$$a_{i,j} = \begin{cases} 1 & \text{, when } i = j \ \text{ or } j = n, \\ -1 & \text{, when } i > j, \\ 0 & \text{, else} \end{cases} \qquad .$$

2.3. **Full pivoting.** A reasonable growth rate can be guaranteed only for *full pivoting.* There one finds in all of $\boldsymbol{A}$ the element $a_{j,k}$ with largest absolute value. Then this is brought to place $1,1$ by multiplication with permutations from the left and the right: $\boldsymbol{\Pi}_1\,\boldsymbol{A}\,\boldsymbol{Q}_1 = \boldsymbol{\Pi}(1,j)\,\boldsymbol{A}\,\boldsymbol{\Pi}(1,k)$. Then eliminate: $\boldsymbol{M}_1\,\boldsymbol{\Pi}_1\,\boldsymbol{A}\,\boldsymbol{Q}_1 = \boldsymbol{A}_1$ and continuing this way we get:

$$\boldsymbol{M}_{n-1}\,\boldsymbol{\Pi}_{n-1}\ldots\boldsymbol{M}_2\,\boldsymbol{\Pi}_2\,\boldsymbol{M}_1\,\boldsymbol{\Pi}_1\,\boldsymbol{A}\,\boldsymbol{Q}_1\,\boldsymbol{Q}_2\ldots\boldsymbol{Q}_{n-1} = \boldsymbol{U}\ ,$$

from which, as before $\boldsymbol{\Pi}\,\boldsymbol{A}\,\boldsymbol{Q} = \boldsymbol{L}\,\boldsymbol{U}$, where $\boldsymbol{\Pi}$ and $\boldsymbol{Q}$ are permutation matrices. One can show the following

Theorem 2.3 (Wilkinson)**.** *With exact arithmetic the full pivoting algorithm produces:*

$$\|\boldsymbol{U}\|_\infty \leq \sqrt{n}(2\cdot 3^{\frac{1}{2}}\cdot 4^{\frac{1}{3}}\ldots n^{\frac{1}{n-1}})^{\frac{1}{2}}\|\boldsymbol{A}\|_\infty\ .$$

Problem 2.8. What is the complexity of full pivoting if a comparison is as expensive as a floating point operation?

2.4. The $\boldsymbol{LDL}^*$–decomposition and the Cholesky–decomposition.

Let $\boldsymbol{A}$ be a regular Hermitian matrix having a decomposition $\boldsymbol{A} = \boldsymbol{LU}$. Write $\boldsymbol{U} = \boldsymbol{D}\widetilde{\boldsymbol{L}}^*$, where $\boldsymbol{D}$ is a diagonal matrix and $\widetilde{\boldsymbol{L}}$ a lower triangular matrix, with ones on the diagonal. Then

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\widetilde{\boldsymbol{L}}^* = \widetilde{\boldsymbol{L}}\bar{\boldsymbol{D}}\boldsymbol{L}^* \qquad \text{, i.e.,} \qquad \boldsymbol{D}(\boldsymbol{L}^{-1}\widetilde{\boldsymbol{L}})^* = \boldsymbol{L}^{-1}\widetilde{\boldsymbol{L}}\bar{\boldsymbol{D}} \ .$$

The left hand side of the latter equation is an upper triangular matrix and on the right there is a lower triangular matrix. This is possible only, if $\boldsymbol{L}^{-1}\widetilde{\boldsymbol{L}}\bar{\boldsymbol{D}}$ is a diagonal matrix, and then $\boldsymbol{L}^{-1}\widetilde{\boldsymbol{L}}$ also has to be diagonal. Both $\boldsymbol{L}^{-1}$ and $\widetilde{\boldsymbol{L}}$ have ones on the diagonal (for $\boldsymbol{L}^{-1}$ consider forming the inverse by determinants), so that

$$\boldsymbol{L}^{-1}\widetilde{\boldsymbol{L}} = \boldsymbol{I} \qquad \text{, i.e.,} \qquad \widetilde{\boldsymbol{L}} = \boldsymbol{L} \qquad \text{and} \qquad \bar{\boldsymbol{D}} = \boldsymbol{D} \ .$$

We obtained the $\boldsymbol{LDL}^*$–decomposition of a Hermitian matrix:

$$\boldsymbol{A} = \boldsymbol{LDL}^* \ .$$

Further, if $\boldsymbol{A}$ positive definite the diagonal of $\boldsymbol{D}$ is positive, so that it can be written: $\boldsymbol{D} = (\boldsymbol{D}^{\frac{1}{2}})^2 = \text{diag}(\sqrt{d_1}, \dots, \sqrt{d_n})^2$. Then we get the Cholesky–decomposition of $\boldsymbol{A}$:

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{D}^{\frac{1}{2}}\boldsymbol{L}^* = \boldsymbol{G}\boldsymbol{G}^* \ ,$$

where $\boldsymbol{G} = \boldsymbol{L}\boldsymbol{D}^{\frac{1}{2}}$ is a lower triangular matrix with positive diagonal. This decomposition can be computed directly (with $\sim n^3/3$ fl–operations) as follows:

```
for $k = 1 : n$ do
    $A(k,k) = \sqrt{A(k,k)}$
    $A(k+1:n,k) = A(k+1:n,k)/A(k,k)$
    for $j = k+1 : n$ do
        $A(j:n,j) = A(j:n,j) - A(j:n,k)A(j,k)$
    end
end
```

Here we use only the lower triangular part of $\boldsymbol{A}$. The upper part is not needed and the result is stored in the place of the lower triangular part.

With Hermitian matrices we want to preserve Hermitianity in pivoting. This is done using symmetric pivoting: $\boldsymbol{\Pi A \Pi}^T$. With such we get the *diagonal element* with largest absolute value to the place $_{1,1}$.

3. Projections

A nonzero linear map $\boldsymbol{P}$ is a *projection* if $\boldsymbol{P}^2 = \boldsymbol{P}$. Then (assuming $\boldsymbol{P} \neq \boldsymbol{I}$) $\boldsymbol{I} - \boldsymbol{P}$ is also a projection:

$$(\boldsymbol{I}-\boldsymbol{P})^2 = (\boldsymbol{I}-\boldsymbol{P}) - (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{P} = \boldsymbol{I}-\boldsymbol{P}-\boldsymbol{P}+\boldsymbol{P}^2 = \boldsymbol{I}-\boldsymbol{P}\ .$$

If $\boldsymbol{V} \in \mathbb{C}^{n\times k}$ has linearly independent columns, then there exists $\boldsymbol{W} \in \mathbb{C}^{n\times k}$ such that $\boldsymbol{W}^*\boldsymbol{V} = \boldsymbol{I}$. For example[1], $\boldsymbol{W} = (\boldsymbol{V}^\dagger)^* = \boldsymbol{V}(\boldsymbol{V}^*\boldsymbol{V})^{-1}$ will do.

Lemma 3.1. *If $\boldsymbol{W}^*\boldsymbol{V} = \boldsymbol{I}$ then $\boldsymbol{P} = \boldsymbol{V}\boldsymbol{W}^*$ is a projection.*

Proof. $\boldsymbol{P}^2 = \boldsymbol{V}\boldsymbol{W}^*\boldsymbol{V}\boldsymbol{W}^* = \boldsymbol{V}\boldsymbol{W}^* = \boldsymbol{P}$. □

A projection $\boldsymbol{P}$ is called *orthogonal*, if $(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{x} \perp R(\boldsymbol{P})$ for all $\boldsymbol{x}$.

Lemma 3.2. *If $\boldsymbol{P}$ is a projection, then*

$$\boldsymbol{P} \textit{ is orthogonal} \iff \boldsymbol{P} = \boldsymbol{P}^*\ .$$

Proof. Assume $\boldsymbol{P}$ is orthogonal. Then, for all $\boldsymbol{x}, \boldsymbol{y}$,

$$0 = \langle \boldsymbol{x} - \boldsymbol{P}\boldsymbol{x}, \boldsymbol{P}\boldsymbol{y}\rangle = \langle \boldsymbol{P}^*(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{x}, \boldsymbol{y}\rangle\ .$$

Hence $\boldsymbol{P}^*(\boldsymbol{I}-\boldsymbol{P}) = 0$, i.e., $\boldsymbol{P}^* = \boldsymbol{P}^*\boldsymbol{P}$ and $\boldsymbol{P} = (\boldsymbol{P}^*\boldsymbol{P})^* = \boldsymbol{P}^*\boldsymbol{P} = \boldsymbol{P}^*$.

If $\boldsymbol{P} = \boldsymbol{P}^*$, then, for all $\boldsymbol{x}, \boldsymbol{y}$,

$$\langle \boldsymbol{x} - \boldsymbol{P}\boldsymbol{x}, \boldsymbol{P}\boldsymbol{y}\rangle = \langle \boldsymbol{P}^*(\boldsymbol{I}-\boldsymbol{P})\,\boldsymbol{x}, \boldsymbol{y}\rangle = \left\langle (\boldsymbol{P}-\boldsymbol{P}^2)\,\boldsymbol{x}, \boldsymbol{y}\right\rangle = 0\ ,$$

i.e., $\boldsymbol{x} - \boldsymbol{P}\boldsymbol{x} \perp R(\boldsymbol{P})$. □

Theorem 3.3. *For a projection $\boldsymbol{P}$ the following are equivalent*

(i) *$\boldsymbol{P}$ is orthogonal*
(ii) $\|\boldsymbol{x}\|^2 = \|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{x}\|^2 + \|\boldsymbol{P}\boldsymbol{x}\|^2$
(iii) $\|\boldsymbol{P}\|_2 = 1$
(iv) $R(\boldsymbol{P}) \perp N(\boldsymbol{P})$.

Proof. (i) $\Longrightarrow$ (ii) : If $\boldsymbol{P}$ is orthogonal, then

$$\begin{aligned}\|\boldsymbol{x}\|^2 &= \langle (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{x} + \boldsymbol{P}\boldsymbol{x}, (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{x} + \boldsymbol{P}\boldsymbol{x}\rangle \\ &= \langle (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{x}, (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{x}\rangle + \langle (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{x}, \boldsymbol{P}\boldsymbol{x}\rangle + \langle \boldsymbol{P}\boldsymbol{x}, (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{x}\rangle + \langle \boldsymbol{P}\boldsymbol{x}, \boldsymbol{P}\boldsymbol{x}\rangle \\ &= \|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{x}\|^2 + \|\boldsymbol{P}\boldsymbol{x}\|^2\ .\end{aligned}$$

[1] For given $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ the matrix $\boldsymbol{A}^\dagger = \lim_{\varepsilon\to 0}(\boldsymbol{A}^*\boldsymbol{A} + \varepsilon\boldsymbol{I})^{-1}\,\boldsymbol{A}^*$ is called the *pseudoinverse* (see section 4.4) of $\boldsymbol{A}$. If $\boldsymbol{A}$ has full column rank, then $\boldsymbol{A}^\dagger = (\boldsymbol{A}^*\boldsymbol{A})^{-1}\,\boldsymbol{A}^*$. In particular, if $\boldsymbol{A}$ is invertible, then $\boldsymbol{A}^\dagger = \boldsymbol{A}^{-1}$.

(ii) $\Longrightarrow$ (iii) is clear.

(iii) $\Longrightarrow$ (iv) : Assume $\|\boldsymbol{P}\|_2 = 1$. If $\boldsymbol{x} \in R(\boldsymbol{P})$ and $\boldsymbol{y} \in N(\boldsymbol{P})$ are non-orthogonal unit vectors, let $\boldsymbol{u} = \boldsymbol{x} - \langle \boldsymbol{x}, \boldsymbol{y} \rangle\ \boldsymbol{y}$. Then $\|\boldsymbol{P}\boldsymbol{u}\| = \|\boldsymbol{P}\boldsymbol{x}\| = \|\boldsymbol{x}\| = 1$, but

$$\begin{aligned}\|\boldsymbol{u}\|^2 &= \langle \boldsymbol{x} - \langle \boldsymbol{x}, \boldsymbol{y} \rangle\ \boldsymbol{y}\ , \boldsymbol{x} - \langle \boldsymbol{x}, \boldsymbol{y} \rangle\ \boldsymbol{y} \rangle \\ &= \langle \boldsymbol{x}, \boldsymbol{x} \rangle - \langle \boldsymbol{x}, \boldsymbol{y} \rangle \langle \boldsymbol{y}, \boldsymbol{x} \rangle - \overline{\langle \boldsymbol{x}, \boldsymbol{y} \rangle} \langle \boldsymbol{x}, \boldsymbol{y} \rangle + |\langle \boldsymbol{x}, \boldsymbol{y} \rangle|^2 \langle \boldsymbol{y}, \boldsymbol{y} \rangle = 1 - |\langle \boldsymbol{x}, \boldsymbol{y} \rangle|^2 < 1\ ,\end{aligned}$$

i.e., $\|\boldsymbol{P}\|_2 \geq 1/\|\boldsymbol{u}\| > 1$, a contradiction. Hence $R(\boldsymbol{P}) \perp N(\boldsymbol{P})$.

(iv) $\Longrightarrow$ (i) : Clear from $R(\boldsymbol{I} - \boldsymbol{P}) \subset N(\boldsymbol{P})$. □

Example 3.1. If $\boldsymbol{V} \in \mathbb{C}^{n \times k}$ has orthonormal columns, then $\boldsymbol{V}^*\boldsymbol{V} = \boldsymbol{I}$ and $\boldsymbol{P} = \boldsymbol{V}\boldsymbol{V}^*$ is Hermitian projection, in particular, orthogonal.

3.1. **Oblique projections.**

Problem 3.1. Show that for any projection

$$N(\boldsymbol{P}) = R(\boldsymbol{I} - \boldsymbol{P}) \qquad \text{and} \qquad \mathbb{C}^n = N(\boldsymbol{P}) \oplus R(\boldsymbol{P})\ .$$

For a subspace E denote by $\boldsymbol{P}_E^\perp$ the orthogonal projection onto E .

Theorem 3.4. *Let E and$F$ be subspaces such that*

$$\|\boldsymbol{P}_E^\perp - \boldsymbol{P}_F^\perp\|_2 < 1\ . \tag{3.1}$$

Then $\dim E = \dim F$, and for any basis $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$ of E there exists a basis $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d$ of F such that

$$\langle \boldsymbol{v}_i, \boldsymbol{w}_j \rangle = \delta_{i,j}\ , \quad i, j = 1, \ldots, d\ . \tag{3.2}$$

Such basis are called *biorthogonal.*

Proof. First note that $E \cap F^\perp = \{0\}$ since if $0 \neq \boldsymbol{y} \in E \cap F^\perp$ then $\boldsymbol{P}_F^\perp \boldsymbol{y} = 0$ and we get the following contradiction:

$$\|\boldsymbol{y}\| = \|\boldsymbol{P}_E^\perp \boldsymbol{y}\| = \|\boldsymbol{P}_E^\perp \boldsymbol{y} - \boldsymbol{P}_F^\perp \boldsymbol{y}\| \leq \|\boldsymbol{P}_E^\perp - \boldsymbol{P}_F^\perp\|_2 \|\boldsymbol{y}\| < \|\boldsymbol{y}\|\ .$$

$E \cap F^\perp = \{0\}$ further implies

$$\dim E \leq n - \dim F^\perp = n - (n - \dim F) = \dim F\ .$$

Similarly we obtain $\dim F \leq \dim E$.

Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$ be a basis of E . Set $E_j = \operatorname{span}(\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d\} \backslash \{\boldsymbol{v}_j\})$. Since $\dim E_j^\perp = n - d + 1$ and $\dim F = d$ we have $F \cap E_j^\perp \neq \{0\}$. So, for each j we can take $0 \neq \boldsymbol{y}_j \in F \cap E_j^\perp$. If $\langle \boldsymbol{v}_j, \boldsymbol{y}_j \rangle = 0$ then $\boldsymbol{y}_j \perp E$ which is impossible. Hence, for every j we can set $\boldsymbol{w}_j = \boldsymbol{y}_j / \langle \boldsymbol{v}_j, \boldsymbol{y}_j \rangle$. Clearly these satisfy (3.2).

If $\sum_{j=1}^{d} \alpha_j \boldsymbol{w}_j = 0$, then taking inner product with $\boldsymbol{v}_i$ implies $\alpha_i = 0$. Thus $\boldsymbol{w}_j$'s are linearly independent and there are $d = \dim F$ of them, hence they form a basis for F. □

Problem 3.2. Let $\boldsymbol{V}, \boldsymbol{W} \in \mathbb{C}^{n\times d}$ be such that $\boldsymbol{W}^*\boldsymbol{V} \in \mathbb{R}^{d\times d}$ is invertible. Show that $E = R(\boldsymbol{V})$ and $F = R(\boldsymbol{W})$ satisfy the assumption of the previous theorem.

Theorem 3.5. (3.1) *is equivalent to* $E \cap F^\perp = F \cap E^\perp = \{0\}$.

Proof. It was already shown in the proof of the previous theorem that (3.1) implies $E \cap F^\perp = F \cap E^\perp = \{0\}$.

Assume now that (3.1) does not hold. We will show that there exists a nonzero vector either in $E \cap F^\perp$ or in $F \cap E^\perp$. Let $\boldsymbol{x}$ be a unit vector such that $\|\boldsymbol{u}-\boldsymbol{v}\|_2 \geq 1$, where $\boldsymbol{u} = \boldsymbol{P}_E^\perp \boldsymbol{x}$, $\boldsymbol{v} = \boldsymbol{P}_F^\perp \boldsymbol{x}$.

Since $\langle \boldsymbol{u}, \boldsymbol{u}-\boldsymbol{x}\rangle = 0$ and $\boldsymbol{x} = \boldsymbol{u} + (\boldsymbol{x}-\boldsymbol{u})$ we get

$$\|2\boldsymbol{u}-\boldsymbol{x}\|^2 = \langle \boldsymbol{u}+\boldsymbol{u}-\boldsymbol{x}\,,\, \boldsymbol{u}+\boldsymbol{u}-\boldsymbol{x}\rangle = \|\boldsymbol{u}\|^2 + \|\boldsymbol{u}-\boldsymbol{x}\|^2 = \|\boldsymbol{x}\|^2 = 1\ .$$

Hence $\left\|\boldsymbol{u} - \frac{1}{2}\boldsymbol{x}\right\| = \frac{1}{2}$. Similarly, $\left\|\boldsymbol{v} - \frac{1}{2}\boldsymbol{x}\right\| = \frac{1}{2}$. Set $\boldsymbol{a} = \boldsymbol{u} - \frac{1}{2}\boldsymbol{x}$, $\boldsymbol{b} = \frac{1}{2}\boldsymbol{x} - \boldsymbol{v}$. Then

$$1 = \|\boldsymbol{a}\| + \|\boldsymbol{b}\| \geq \|\boldsymbol{a}+\boldsymbol{b}\| = \|\boldsymbol{u}-\boldsymbol{v}\| \geq 1\ .$$

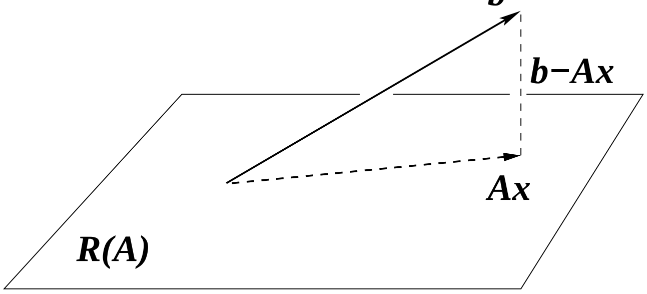
Hence $\|\boldsymbol{a}\| + \|\boldsymbol{b}\| = \|\boldsymbol{a}+\boldsymbol{b}\|$ and consequently $\boldsymbol{a} = \alpha\,\boldsymbol{b}$ with $\alpha > 0$. Since $\boldsymbol{a}$ and $\boldsymbol{b}$ have same length we get $\alpha = 1$, i.e. $\boldsymbol{u} - \frac{1}{2}\boldsymbol{x} = \frac{1}{2}\boldsymbol{x} - \boldsymbol{v}$. Thus $\boldsymbol{u} = \boldsymbol{x} - \boldsymbol{v}$ and $\boldsymbol{v} = \boldsymbol{x} - \boldsymbol{u}$. But $\boldsymbol{u} \in E$, $\boldsymbol{x}-\boldsymbol{v} \in F^\perp$, $\boldsymbol{v} \in F$, $\boldsymbol{x}-\boldsymbol{u} \in E^\perp$. Since both $\boldsymbol{u}$ and $\boldsymbol{v}$ cannot be zero, either $E \cap F^\perp$ or $F \cap E^\perp$ is different from $\{0\}$. □

Problem 3.3. Let E and H be subspaces of $\mathbb{C}^n$ such that $\mathbb{C}^n = E \oplus H$. Construct a projection $\boldsymbol{P}$ such that $R(\boldsymbol{P}) = E$ and $R(\boldsymbol{I}-\boldsymbol{P}) = H$. Hint: set $F = H^\perp$, take a basis $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$ of E, and use theorems 3.5 and 3.4.

Problem 3.4. Is $\boldsymbol{P}$ unique in the previous problem?

4. Least squares problems, $\boldsymbol{QR}$–decomposition

When the $m\times n$–system $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$ has more equations than unknowns ($m > n$), it usually does not have a solution. A least squares solution means a vector $\boldsymbol{x}$, for which $\|\boldsymbol{A}\,\boldsymbol{x} - \boldsymbol{b}\|_2$ is smallest, i.e., we look for the orthogonal projection of $\boldsymbol{b}$ onto $R(\boldsymbol{A})$.



In other words we look for $\boldsymbol{x}$ such that $\boldsymbol{A}\,\boldsymbol{x} - \boldsymbol{b} \in R(\boldsymbol{A})^{\perp}$. Since $R(\boldsymbol{A})^{\perp} = N(\boldsymbol{A}^*)$, we see that $\boldsymbol{x}$ has to solve $\boldsymbol{A}^*(\boldsymbol{A}\,\boldsymbol{x} - \boldsymbol{b}) = 0$, i.e.,

$$\boldsymbol{A}^*\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{A}^*\,\boldsymbol{b}\,. \tag{4.1}$$

The numerical solution of (4.1) is usually found by forming the $\boldsymbol{QR}$–decomposition of $\boldsymbol{A}$, i.e., writing $\boldsymbol{A} = \boldsymbol{QR}$, where the matrix $\boldsymbol{Q} \in \mathbb{C}^{m\times n}$ has orthonormal columns and $\boldsymbol{R}$ is an upper triangular matrix. Then $\boldsymbol{Q}^*\boldsymbol{Q} = \boldsymbol{I}$ and if $\boldsymbol{R}$ is invertible (the columns of $\boldsymbol{A}$ are linearly independent), then (4.1) becomes

$$\boldsymbol{R}^*\boldsymbol{R}\boldsymbol{x} = \boldsymbol{R}^*\boldsymbol{Q}^*\,\boldsymbol{b}\,, \qquad \text{i.e.,} \qquad \boldsymbol{R}\boldsymbol{x} = \boldsymbol{Q}^*\,\boldsymbol{b}\,,$$

which is easy to solve with backward substitutions.

Problem 4.1. Show that $\boldsymbol{R}$ is invertible if and only if the columns of $\boldsymbol{A}$ are linearly independent.

$\boldsymbol{QR}$–decomposition can be computed by performing the Gram–Schmidt process on the columns of $\boldsymbol{A}$. A better (more stable) way is to apply certain simple unitary transforms to $\boldsymbol{A}$. The following result gives the justification for this:

Problem 4.2. If $\boldsymbol{U} \in \mathbb{C}^{n\times n}$ is unitary: $\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}$, then

$$\|\boldsymbol{U}\boldsymbol{x}\|_2 = \|\boldsymbol{x}\|_2\ , \quad \|\boldsymbol{U}\boldsymbol{A}\|_2 = \|\boldsymbol{A}\|_2\ , \quad \|\boldsymbol{U}\boldsymbol{A}\|_F = \|\boldsymbol{A}\|_F\ , \quad \|\boldsymbol{U}\boldsymbol{A}\|_\sigma = \|\boldsymbol{A}\|_\sigma\ .$$

When computing $\boldsymbol{U}\boldsymbol{A}$, where $\boldsymbol{U} \in \mathbb{C}^{n\times n}$ is unitary, one has in the computer $\widehat{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{E}$ and $\widehat{\boldsymbol{U}} = \boldsymbol{U} + \boldsymbol{F}$, where usually $\|\boldsymbol{E}\|_2 \le C_n\mu\,\|\boldsymbol{A}\|_2$. $\|\boldsymbol{F}\|_2 \le c_n\mu$, and $c_n \ll C_n$. In the computer the result then is:

$$\widehat{\widehat{\boldsymbol{U}}\widehat{\boldsymbol{A}}} = \widehat{\boldsymbol{U}}\widehat{\boldsymbol{A}} + \boldsymbol{H} = (\boldsymbol{U}+\boldsymbol{F})(\boldsymbol{A}+\boldsymbol{E}) + \boldsymbol{H} = \boldsymbol{U}\boldsymbol{A} + \boldsymbol{F}\boldsymbol{A} + \boldsymbol{U}\boldsymbol{E} + \boldsymbol{F}\boldsymbol{E} + \boldsymbol{H} = \boldsymbol{U}\boldsymbol{A} + \widetilde{\boldsymbol{E}}\,,$$

where the multiplication error satisfies $\|\boldsymbol{H}\|_2 \leq \widetilde{c}_n \mu \, \|\boldsymbol{A}\|_2 + O(\mu^2)$ and $\widetilde{c}_n \ll C_n$. Then

$$\left\|\widetilde{\boldsymbol{E}}\right\|_2 \leq \|\boldsymbol{F}\|_2 \, \|\boldsymbol{A}\|_2 + \|\boldsymbol{E}\|_2 + \|\boldsymbol{H}\|_2 + O(\mu^2) \leq (C_n + c_n + \widetilde{c}_n) \, \mu \, \|\boldsymbol{A}\|_2 + O(\mu^2) \, .$$

Since the coefficient of $\|\boldsymbol{E}\|_2$ is one, the errors accumulate only additively, they are not amplified.

4.1. **Householder transformation.** Let $\boldsymbol{v} \in \mathbb{C}^n \setminus \{0\}$ be arbitrary. Then the matrix

$$\boldsymbol{H} = \boldsymbol{I} - 2 \, \frac{\boldsymbol{v} \, \boldsymbol{v}^*}{\boldsymbol{v}^* \boldsymbol{v}}$$

is Hermitian and unitary: $\boldsymbol{H}^* = \boldsymbol{H}$,

$$\boldsymbol{H}^* \boldsymbol{H} = \boldsymbol{H}^2 = \boldsymbol{I} - 4 \, \frac{\boldsymbol{v} \, \boldsymbol{v}^*}{\boldsymbol{v}^* \boldsymbol{v}} + 4 \, \frac{\boldsymbol{v} \, \boldsymbol{v}^* \boldsymbol{v} \, \boldsymbol{v}^*}{(\boldsymbol{v}^* \boldsymbol{v})^2} = \boldsymbol{I} \ .$$

It is called a Householder transformation. Often we need a transformation $\boldsymbol{H}$ such that a given vector $\boldsymbol{x}$ is turned, by the multiplication with $\boldsymbol{H}$ in the direction of $\boldsymbol{e}_1$. Since

$$\boldsymbol{H}\boldsymbol{x} = \boldsymbol{x} - 2 \, \frac{\langle \boldsymbol{x}, \boldsymbol{v} \rangle}{\langle \boldsymbol{v}, \boldsymbol{v} \rangle} \, \boldsymbol{v} \ ,$$

we have $\boldsymbol{H}\boldsymbol{x} \in \operatorname{span}(\boldsymbol{e}_1) \implies \boldsymbol{v} \in \operatorname{span}(\boldsymbol{x}, \boldsymbol{e}_1)$ (unless $\boldsymbol{x} \in \operatorname{span}(\boldsymbol{e}_1)$). The trial $\boldsymbol{v} = \boldsymbol{x} + \alpha \, \boldsymbol{e}_1$ gives

$$\begin{aligned} \langle \boldsymbol{x}, \boldsymbol{v} \rangle &= \langle \boldsymbol{x}, \boldsymbol{x} \rangle + \bar{\alpha} \, \langle \boldsymbol{x}, \boldsymbol{e}_1 \rangle \qquad \text{and} \\ \langle \boldsymbol{v}, \boldsymbol{v} \rangle &= \langle \boldsymbol{x}, \boldsymbol{x} \rangle + \bar{\alpha} \, \langle \boldsymbol{x}, \boldsymbol{e}_1 \rangle + \alpha \overline{\langle \boldsymbol{x}, \boldsymbol{e}_1 \rangle} + |\alpha|^2 \ , \end{aligned}$$

so that

$$\boldsymbol{H}\boldsymbol{x} = (1 - 2 \, \frac{\langle \boldsymbol{x}, \boldsymbol{x} \rangle + \bar{\alpha} \, \langle \boldsymbol{x}, \boldsymbol{e}_1 \rangle}{\langle \boldsymbol{x}, \boldsymbol{x} \rangle + \bar{\alpha} \, \langle \boldsymbol{x}, \boldsymbol{e}_1 \rangle + \alpha \overline{\langle \boldsymbol{x}, \boldsymbol{e}_1 \rangle} + |\alpha|^2}) \boldsymbol{x} - 2\alpha \, \frac{\langle \boldsymbol{x}, \boldsymbol{v} \rangle}{\langle \boldsymbol{v}, \boldsymbol{v} \rangle} \, \boldsymbol{e}_1 \ .$$

Choosing $\alpha = \frac{\langle \boldsymbol{x}, \boldsymbol{e}_1 \rangle}{|\langle \boldsymbol{x}, \boldsymbol{e}_1 \rangle|} \, \|\boldsymbol{x}\|_2$ ($\alpha = \|\boldsymbol{x}\|_2$, if $\langle \boldsymbol{x}, \boldsymbol{e}_1 \rangle = 0$) we get: $\bar{\alpha} \, \langle \boldsymbol{x}, \boldsymbol{e}_1 \rangle = |\langle \boldsymbol{x}, \boldsymbol{e}_1 \rangle| \, \|\boldsymbol{x}\|_2$, giving $\langle \boldsymbol{x}, \boldsymbol{v} \rangle = \frac{1}{2} \, \langle \boldsymbol{v}, \boldsymbol{v} \rangle$ and $\boldsymbol{H} \, \boldsymbol{x} = -\alpha \, \boldsymbol{e}_1$.

In practice one never forms the matrix $\boldsymbol{H}$, only the vector $\boldsymbol{v}$ is stored and for example the multiplication $\boldsymbol{H}\boldsymbol{A}$ is done as:

$$\boldsymbol{H}\boldsymbol{A} = \boldsymbol{A} - \tfrac{2}{\langle \boldsymbol{v}, \boldsymbol{v} \rangle} \, \boldsymbol{v} \, \boldsymbol{v}^* \, \boldsymbol{A} \ ,$$

i.e., a rank–one matrix is added to $\boldsymbol{A}$.

***QR*–decomposition with Householder transformations.** Assume $\boldsymbol{A} \in \mathbb{C}^{m \times n}$ and let $\boldsymbol{a}_1$ be the first column of $\boldsymbol{A}$. Find a vector $\boldsymbol{v}_1$ such that $\boldsymbol{H}_1 \, \boldsymbol{a}_1 = r_{1,1} \, \boldsymbol{e}_1$, where $\boldsymbol{H}_1 = \boldsymbol{I} - 2\,\dfrac{\boldsymbol{v}_1\,\boldsymbol{v}_1^*}{\boldsymbol{v}_1^*,\boldsymbol{v}_1}$. Then

$$\boldsymbol{H}_1\,\boldsymbol{A} = \begin{bmatrix} r_{1,1} & r_{1,2} & \dots r_{1,n} \\ 0 & & \\ \vdots & \widetilde{\boldsymbol{A}}_1 & \\ 0 & & \end{bmatrix} .$$

Let $\widetilde{\boldsymbol{a}}_1$ be the first column of $\widetilde{\boldsymbol{A}}_1$ and let $\widetilde{\boldsymbol{H}}_2 = \boldsymbol{I} - 2\,\dfrac{\widetilde{\boldsymbol{v}}_2\,\widetilde{\boldsymbol{v}}_2^*}{\widetilde{\boldsymbol{v}}_2^*\widetilde{\boldsymbol{v}}_2}$ be such that $\widetilde{\boldsymbol{H}}_2\,\widetilde{\boldsymbol{a}}_1 = r_{2,2}\,\boldsymbol{e}_1 \in \mathbb{C}^{n-1}$. Set $\boldsymbol{v}_2^* = [0\ \widetilde{\boldsymbol{v}}_2^*]$ and $\boldsymbol{H}_2 = I - 2\,\dfrac{\boldsymbol{v}_2\,\boldsymbol{v}_2^*}{\boldsymbol{v}_2^*\boldsymbol{v}_2} = \begin{bmatrix} 1 & 0 \\ 0 & \widetilde{\boldsymbol{H}}_2 \end{bmatrix}$. Then

$$\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A} = \begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3} & \dots & r_{1,n} \\ 0 & r_{2,2} & r_{2,3} & \dots & r_{2,n} \\ 0 & 0 & & & \\ \vdots & \vdots & & \widetilde{\boldsymbol{A}}_2 & \\ 0 & 0 & & & \end{bmatrix} .$$

Continuing this way we get:

$$\boldsymbol{H}_n \dots \boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A} = \begin{bmatrix} r_{1,1} & r_{1,2} & \dots & r_{1,n} \\ 0 & r_{2,2} & \dots & r_{2,n} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & r_{n,n} \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{R} \\ 0 \end{bmatrix} .$$

When solving a least squares problem $\min_{\boldsymbol{x}} \|\boldsymbol{A}\,\boldsymbol{x} - \boldsymbol{b}\|$ one does not need to form $\boldsymbol{Q}$ or store the vectors $\boldsymbol{v}_k$ if at each step also the right hand side is multiplied by $\boldsymbol{H}_k$ resulting in $\widetilde{\boldsymbol{b}} = \boldsymbol{H}_n \dots \boldsymbol{H}_2\boldsymbol{H}_1\,\boldsymbol{b}$ since by unitariness of the $\boldsymbol{H}_k$'s

$$\|\boldsymbol{A}\,\boldsymbol{x} - \boldsymbol{b}\|_2 = \|\boldsymbol{H}_n \dots \boldsymbol{H}_2\boldsymbol{H}_1(\boldsymbol{A}\,\boldsymbol{x} - \boldsymbol{b})\|_2 = \left\| \begin{bmatrix} \boldsymbol{R}\,\boldsymbol{x} - \widetilde{\boldsymbol{b}}_1 \\ \widetilde{\boldsymbol{b}}_2 \end{bmatrix} \right\|_2 ,$$

so that the best $\boldsymbol{x}$ is given by the solution of the system $\boldsymbol{R}\,\boldsymbol{x} = \widetilde{\boldsymbol{b}}_1$,

If we want to form the product

$$\boldsymbol{Q}_n = \boldsymbol{H}_1\boldsymbol{H}_2 \dots \boldsymbol{H}_n = \Big(\boldsymbol{I} - 2\,\frac{\boldsymbol{v}_1\,\boldsymbol{v}1^*}{\boldsymbol{v}_1^*\boldsymbol{v}_1}\Big)\Big(\boldsymbol{I} - 2\,\frac{\boldsymbol{v}_2\,\boldsymbol{v}_2^*}{\boldsymbol{v}_2^*\boldsymbol{v}_2}\Big) \dots \Big(\boldsymbol{I} - 2\,\frac{\boldsymbol{v}_n\,\boldsymbol{v}_n^*}{\boldsymbol{v}_n^*\boldsymbol{v}_n}\Big)$$

it is better to make it in the form $\boldsymbol{Q}_k = \boldsymbol{I} + \boldsymbol{W}_k\boldsymbol{V}_k^*$, $k = 1, \dots, n$ as follows:

set: $\boldsymbol{W}_1 = \boldsymbol{w}_1 = -\frac{2}{\boldsymbol{v}_1^*\boldsymbol{v}_1}\boldsymbol{v}_1$, $\boldsymbol{V}_1 = \boldsymbol{v}_1$. Then $\boldsymbol{Q}_1 = \boldsymbol{H}_1 = \boldsymbol{I} + \boldsymbol{W}_1\boldsymbol{V}_1^*$ and

$$\begin{aligned}\boldsymbol{Q}_k &= \boldsymbol{Q}_{k-1}\boldsymbol{H}_k = (\boldsymbol{I} + \boldsymbol{W}_{k-1}\boldsymbol{V}_{k-1}^*)(\boldsymbol{I} - 2\frac{\boldsymbol{v}_k\boldsymbol{v}_k^*}{\boldsymbol{v}_k^*\boldsymbol{v}_k}) = \\ &= \boldsymbol{I} + \boldsymbol{W}_{k-1}\boldsymbol{V}_{k-1}^* + \boldsymbol{w}_k\boldsymbol{v}_k^* = \boldsymbol{I} + \boldsymbol{W}_k\boldsymbol{V}_k^* ,\end{aligned}$$

where $\boldsymbol{w}_k = -\frac{2}{\boldsymbol{v}_k^*\boldsymbol{v}_k}\boldsymbol{Q}_{k-1}\boldsymbol{v}_k$ and $\boldsymbol{W}_k = [\boldsymbol{W}_{k-1}\ \boldsymbol{w}_k]$, $\boldsymbol{V}_k = [\boldsymbol{V}_{k-1}\ \boldsymbol{v}_k]$.

Notice that also here $\boldsymbol{v}_k$ is multiplied by the unitary matrix $\boldsymbol{Q}_{k-1}$ and then scaled so that the computation is stable. Also this representation of $\boldsymbol{Q}_n$ is very economic when $n \ll m$.

Problem 4.3. Show how to use Householder transformations to form the decomposition $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{L}$, where $\boldsymbol{Q}$ has orthonormal columns and $\boldsymbol{L}$ is a lower triangular matrix?

Problem 4.4. Assume we need to solve the $n \times n$ system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. Compute the work load (n^3–terms) for decompositions $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ and $\boldsymbol{\Pi}\boldsymbol{A}\widetilde{\boldsymbol{\Pi}} = \boldsymbol{L}\boldsymbol{U}$ (with full pivoting) assuming in the latter, that

work(comparison) = $k\cdot$ work(floating point operation).

For which value of k the work loads are the same?

4.2. **Givens rotations.** Let $\begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{C}^2 \setminus \{0\}$ and

$$\boldsymbol{U} = \begin{bmatrix} \bar{\alpha} & \bar{\beta} \\ -\beta & \alpha \end{bmatrix} , \qquad \text{where} \qquad \alpha = \frac{a}{m},\ \beta = \frac{b}{m},\ m = \sqrt{|a|^2 + |b|^2} .$$

Then $\boldsymbol{U}$ is unitary, $\det(\boldsymbol{U}) = 1$ and $\boldsymbol{U}\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} m \\ 0 \end{bmatrix}$. In other words, $\boldsymbol{U}$ rotates the vector $\begin{bmatrix} a \\ b \end{bmatrix}$ in the direction of vector $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

From these we can build unitary transformations that zeroes suitable elements of a vector/matrix: if $\alpha = \dfrac{x_j}{m}$, $\beta = \dfrac{x_k}{m}$, $m = \sqrt{|x_j|^2 + |x_k|^2}$, then setting:

$$\boldsymbol{G}_{j,k} = \begin{array}{cc} \begin{array}{ccccccccccc} & & & j & & & & k & & & \end{array} & \\ \begin{bmatrix} 1 & & & & & & & & & & \\ & \ddots & & & & & & & & & \\ & & 1 & & & & & & & & \\ & & & \bar{\alpha} & & & & \bar{\beta} & & & \\ & & & & 1 & & & & & & \\ & & & & & \ddots & & & & & \\ & & & & & & 1 & & & & \\ & & & -\beta & & & & \alpha & & & \\ & & & & & & & & 1 & & \\ & & & & & & & & & \ddots & \\ & & & & & & & & & & 1 \end{bmatrix} & \begin{array}{c} \\ \\ \\ j \\ \\ \\ \\ k \\ \\ \\ \\ \end{array} \end{array}$$

we get a unitary transformation satisfying

$$(\boldsymbol{G}_{j,k}\,\boldsymbol{x})_j = m \ , \quad (\boldsymbol{G}_{j,k}\,\boldsymbol{x})_k = 0 \ , \quad (\boldsymbol{G}_{j,k}\,\boldsymbol{x})_l = x_l \ , \ l \neq j,k \ .$$

Naturally, in practice multiplication of a vector $\boldsymbol{v}$ with $\boldsymbol{G}_{j,k}$ is done as $\boldsymbol{w} = \boldsymbol{v}$ followed by

$$\begin{bmatrix} w_j \\ w_k \end{bmatrix} = \boldsymbol{U} \begin{bmatrix} v_j \\ v_k \end{bmatrix} \ .$$

Similarly, $\boldsymbol{B} = \boldsymbol{G}_{j,k}\boldsymbol{A}$ is obtained by copying the other rows to $\boldsymbol{B}$ and computing the j :th and k :th elements of each column as above.

$\boldsymbol{QR}$–decomposition using Givens rotations can be done as follows:

$$\boldsymbol{A}^{1,2} = \boldsymbol{G}_{1,2}\,\boldsymbol{A} = \begin{bmatrix} \# & \# & \# & \dots & \# \\ 0 & \# & \# & \dots & \# \\ \# & \# & \# & \dots & \# \\ \# & \# & \# & \dots & \# \\ \vdots & \vdots & \vdots & & \vdots \\ \# & \# & \# & \dots & \# \end{bmatrix} \qquad \boldsymbol{A}^{1,3} = \boldsymbol{G}_{1,3}\boldsymbol{A}^{1,2} = \begin{bmatrix} \# & \# & \# & \dots & \# \\ 0 & \# & \# & \dots & \# \\ 0 & \# & \# & \dots & \# \\ \# & \# & \# & \dots & \# \\ \vdots & \vdots & \vdots & & \vdots \\ \# & \# & \# & \dots & \# \end{bmatrix}$$

$$\boldsymbol{A}^{1,m} = \boldsymbol{G}_{1,m}\boldsymbol{A}^{1,m-1} = \begin{bmatrix} \# & \# & \# & \dots & \# \\ 0 & \# & \# & \dots & \# \\ 0 & \# & \# & \dots & \# \\ 0 & \# & \# & \dots & \# \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & \# & \# & \dots & \# \end{bmatrix} \qquad \boldsymbol{A}^{2,3} = \boldsymbol{G}_{2,3}\boldsymbol{A}^{1,m} = \begin{bmatrix} \# & \# & \# & \dots & \# \\ 0 & \# & \# & \dots & \# \\ 0 & 0 & \# & \dots & \# \\ 0 & \# & \# & \dots & \# \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & \# & \# & \dots & \# \end{bmatrix}$$

$$\boldsymbol{A}^{2,m} = \boldsymbol{G}_{2,m}\boldsymbol{A}^{2,m-1} = \begin{bmatrix} \# & \# & \# & \dots & \# \\ 0 & \# & \# & \dots & \# \\ 0 & 0 & \# & \dots & \# \\ 0 & 0 & \# & \dots & \# \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \# & \dots & \# \end{bmatrix} \qquad \boldsymbol{A}^{n,m} = \boldsymbol{G}_{n,m}\boldsymbol{A}^{n,m-1} = \begin{bmatrix} \# & \# & \# & \dots & \# \\ 0 & \# & \# & \dots & \# \\ 0 & 0 & \# & \dots & \# \\ 0 & 0 & 0 & \ddots & \# \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

where $\boldsymbol{A}^{n,m} = \begin{bmatrix} \boldsymbol{R} \\ 0 \end{bmatrix}$. Again, if we are solving a least squares problem $\boldsymbol{A}\,\boldsymbol{x} = \boldsymbol{b}$ we need not form $\boldsymbol{Q}$ or store the rotations if at each step we multiply also the right hand side by $\boldsymbol{G}_{j,k}$.

4.3. **Uniqueness of $\boldsymbol{Q}\,\boldsymbol{R}$.** If $\boldsymbol{A} \in \mathbb{C}^{m\times n}$, $\operatorname{rank}(\boldsymbol{A}) = n$, then the $\boldsymbol{Q}\,\boldsymbol{R}$ decomposition of $\boldsymbol{A}$, where the diagonal elements of $\boldsymbol{R}$ are positive is unique:
If $\boldsymbol{A} = \boldsymbol{Q}\,\boldsymbol{R} = \widetilde{\boldsymbol{Q}}\,\widetilde{\boldsymbol{R}}$ are two such decompositions, then $\boldsymbol{U} = \widetilde{\boldsymbol{Q}}^*\,\boldsymbol{Q} = \widetilde{\boldsymbol{R}}\,\boldsymbol{R}^{-1}$ is upper triangular and we get

$$\boldsymbol{U}^*\,\boldsymbol{U}\,\boldsymbol{R} = \boldsymbol{Q}^*\,\widetilde{\boldsymbol{Q}}\,\widetilde{\boldsymbol{Q}}^*\,\boldsymbol{Q}\,\boldsymbol{R} = \boldsymbol{Q}^*\,\widetilde{\boldsymbol{Q}}\,\widetilde{\boldsymbol{Q}}^*\,\widetilde{\boldsymbol{Q}}\,\widetilde{\boldsymbol{R}} = \boldsymbol{Q}^*\,\widetilde{\boldsymbol{Q}}\,\widetilde{\boldsymbol{R}} = \boldsymbol{Q}^*\,\boldsymbol{Q}\,\boldsymbol{R} = \boldsymbol{R}\ ,$$

so that $\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}$, i.e., $\boldsymbol{U}$ is a unitary upper triangular matrix. This is possible only if it is a diagonal matrix, and $|u_{j,j}| = 1\ \forall j$. Since $\widetilde{R}$ and $\boldsymbol{R}^{-1}$ have positive diagonal elements so does $\boldsymbol{U}$ and we get $u_{j,j} = 1\ \forall j$. Hence $\boldsymbol{U} = \boldsymbol{I}$, $\widetilde{\boldsymbol{R}} = \boldsymbol{R}$, and $\widetilde{\boldsymbol{Q}} = \boldsymbol{Q}$.

The use of Householder transformations does not produce positive diagonal for $\boldsymbol{R}$ but setting $\boldsymbol{D} = \operatorname{diag}\left(\frac{r_{1,1}}{|r_{1,1}|} \dots, \frac{r_{n,n}}{|r_{n,n}|}\right)$ we get unitary $\boldsymbol{D}$ and from $\boldsymbol{A} = \boldsymbol{Q}\,\boldsymbol{R}$ we get $\boldsymbol{A} = \widetilde{\boldsymbol{Q}}\,\widetilde{\boldsymbol{R}}$, where $\widetilde{\boldsymbol{Q}} = \boldsymbol{Q}\,\boldsymbol{D}$ and the matrix $\widetilde{\boldsymbol{R}} = \bar{\boldsymbol{D}}\,\boldsymbol{R}$ has positive diagonal.

4.4. **Pseudoinverse.**

The (Moore–Penrose) pseudoinverse of a matrix $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ is defined as

$$\boldsymbol{A}^\dagger = \lim_{\varepsilon\to 0^+} (\boldsymbol{A}^*\boldsymbol{A} + \varepsilon\boldsymbol{I})^{-1}\boldsymbol{A}^* .$$

Problem 4.5. Show, that the limit above always exists. Hint: it suffices to show that $\lim_{\varepsilon\to 0^+}(\boldsymbol{A}^*\boldsymbol{A} + \varepsilon\boldsymbol{I})^{-1}\boldsymbol{A}^*\,\boldsymbol{v}$ exists for every $\boldsymbol{v} \in \mathbb{C}^n$. Consider first the cases $\boldsymbol{v} \in N(\boldsymbol{A}^*)$ and $\boldsymbol{v} \in R(\boldsymbol{A})$.

Problem 4.6. If $\boldsymbol{A}$ has rank n, then the pseudoinverse becomes

$$\boldsymbol{A}^\dagger = (\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^* .$$

This has the properties

$$\boldsymbol{A}^\dagger\boldsymbol{A} = \boldsymbol{I} \qquad \text{and} \qquad \boldsymbol{A}\boldsymbol{A}^\dagger\,\boldsymbol{x} = \boldsymbol{x} \qquad \text{for all} \qquad \boldsymbol{x} \in R(\boldsymbol{A}) .$$

In this full rank case the pseudoinverse is obtained easily from the $\boldsymbol{QR}$–decomposition of $\boldsymbol{A}$:

$$\boldsymbol{A}^\dagger = (\boldsymbol{R}^*\boldsymbol{Q}^*\boldsymbol{Q}\boldsymbol{R})^{-1}\boldsymbol{R}^*\boldsymbol{Q}^* = \boldsymbol{R}^{-1}\boldsymbol{Q}^* .$$

Remark 4.1. In the general case the pseudoinverse can be computed from the singular value decomposition of $\boldsymbol{A}$ (see section 7.3).

Remark 4.2. Another way to define $\boldsymbol{A}^\dagger$ is to require

a) $\boldsymbol{A}^\dagger\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^\dagger$ are both Hermitian,
b) $\boldsymbol{A}\boldsymbol{A}^\dagger\boldsymbol{A} = \boldsymbol{A}$,
c) $\boldsymbol{A}^\dagger\boldsymbol{A}\boldsymbol{A}^\dagger = \boldsymbol{A}^\dagger$.

4.5. **Other unitary decompositions.**

Problem 4.7. Let $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. Show how using Householder transformations one finds unitary $\boldsymbol{U}$ and $\boldsymbol{V}$ such that $\boldsymbol{B} = \boldsymbol{U}^*\boldsymbol{A}\boldsymbol{V}$ is a *bidiagonal matrix*:

$$\boldsymbol{B} = \boldsymbol{U}^*\boldsymbol{A}\boldsymbol{V} = \begin{bmatrix} d_1 & f_1 & & & \\ & d_2 & f_2 & & \\ & & \ddots & \ddots & \\ & & & d_{n-1} & f_{n-1} \\ & & & & d_n \end{bmatrix} .$$

Problem 4.8. Using Householder transformations show, that arbitrary square matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is unitarily similar to a *Hessenberg–matrix* $\boldsymbol{M}$ (that is $m_{j,k} = 0$

for $j \geq k+2$), i.e.,

$$\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U} = \boldsymbol{M} = \begin{bmatrix} \# & \# & \# & \dots & \# & \# & \# \\ \# & \# & \# & \dots & \# & \# & \# \\ 0 & \# & \# & \dots & \# & \# & \# \\ 0 & 0 & \# & \dots & \# & \# & \# \\ & & & \ddots & & & \\ 0 & 0 & 0 & \dots & \# & \# & \# \\ 0 & 0 & 0 & \dots & 0 & \# & \# \end{bmatrix} .$$

5. Eigendecompositions

The eigenvalues of a square matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ are the complex numbers λ for which there exists a nonzero vector $\boldsymbol{x}$ such that $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$. Clearly, such a λ satisfies $\det(\lambda\boldsymbol{I} - \boldsymbol{A}) = 0$ and also the latter is sufficient for the existence of a nonzero solution for $\lambda\boldsymbol{x} - \boldsymbol{A}\boldsymbol{x} = 0$. Hence the eigenvalues are the roots $\lambda_1, \dots, \lambda_n$ of the *characteristic polynomial*:

$$\det(z\boldsymbol{I} - \boldsymbol{A}) = z^n - \operatorname{tr}(\boldsymbol{A})\, z^{n-1} + \cdots + (-1)^n \det(\boldsymbol{A}) = (z-\lambda_1)\dots(z-\lambda_n) \ ,$$

from which one finds

$$\lambda_1 + \lambda_2 + \cdots + \lambda_n = \operatorname{tr}(\boldsymbol{A}) = a_{1,1} + a_{2,2} + \cdots + a_{n,n}$$

$$\text{and} \qquad \lambda_1\lambda_2\dots\lambda_n = \det(\boldsymbol{A}) \ .$$

Assume $\boldsymbol{A}$ and $\boldsymbol{B}$ are similar, denoted by $\boldsymbol{A} \sim \boldsymbol{B}$, in other words there exists a regular $\boldsymbol{S}$ such that $\boldsymbol{B} = \boldsymbol{S}\boldsymbol{A}\boldsymbol{S}^{-1}$. Then

$$\det(z\boldsymbol{I} - \boldsymbol{B}) = \det(\boldsymbol{S}(z\boldsymbol{I} - \boldsymbol{A})\boldsymbol{S}^{-1}) = \det(\boldsymbol{S})\det(z\boldsymbol{I} - \boldsymbol{A})\det(\boldsymbol{S})^{-1} = \det(z\boldsymbol{I} - \boldsymbol{A}) \ ,$$

so that $\boldsymbol{A}$ and $\boldsymbol{B}$ have the same characteristic polynomials and tr and $\det$ are similarity invariants:

$$\operatorname{tr}(\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}) \ , \qquad \det(\boldsymbol{B}) = \det(\boldsymbol{A}) \ .$$

The corresponding eigenvectors satisfy:

$$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x} \iff \boldsymbol{S}\boldsymbol{A}\boldsymbol{S}^{-1}\boldsymbol{S}\boldsymbol{x} = \lambda\boldsymbol{S}\boldsymbol{x} \iff \boldsymbol{B}\boldsymbol{S}\boldsymbol{x} = \lambda\boldsymbol{S}\boldsymbol{x} \ .$$

Problem 5.1. Denote by $\boldsymbol{A}^{j,k}$ the matrix that is left when the j :th row and the k :th column are deleted from $\boldsymbol{A}$. Set: $F(\boldsymbol{A}) = \sum_{j=1}^{n} \det(\boldsymbol{A}^{j,j})$. Show: $F(\boldsymbol{S}\boldsymbol{A}\boldsymbol{S}^{-1}) = F(\boldsymbol{A})$.

If $\boldsymbol{B} = \boldsymbol{S}\boldsymbol{A}\boldsymbol{S}^{-1}$, then $\boldsymbol{B}^k = \boldsymbol{S}\boldsymbol{A}\boldsymbol{S}^{-1}\boldsymbol{S}\boldsymbol{A}\boldsymbol{S}^{-1}\dots\boldsymbol{S}\boldsymbol{A}\boldsymbol{S}^{-1} = \boldsymbol{S}\boldsymbol{A}^k\boldsymbol{S}^{-1}$ and for any polynomial $p(z) = \alpha_n z^n + \alpha_{n-1}z^{n-1} + \cdots + \alpha_1 z + \alpha_0$ holds

$$p(\boldsymbol{B}) = \alpha_n\boldsymbol{B}^n + \alpha_{n-1}\boldsymbol{B}^{n-1} + \cdots + \alpha_1\boldsymbol{B} + \alpha_0\boldsymbol{I} = \boldsymbol{S}\,p(\boldsymbol{A})\,\boldsymbol{S}^{-1} \ .$$

The same holds for converging power series of matrices. For example:

$$e^{\boldsymbol{B}} = \sum_{k=0}^{\infty} \tfrac{1}{k!}\,\boldsymbol{B}^k = \boldsymbol{S}\,e^{\boldsymbol{A}}\,\boldsymbol{S}^{-1} \ .$$

The set of eigenvalues $\Lambda(\boldsymbol{A}) = \big\{\lambda \in \mathbb{C} \,\big|\, \det(\lambda\boldsymbol{I} - \boldsymbol{A}) = 0\big\}$ is called the *spectrum* of $\boldsymbol{A}$.

[0] Version: April 9, 2003

Problem 5.2. Let $\boldsymbol{A}$ be a block upper triangular matrix: $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{1,1} & \boldsymbol{A}_{1,2} \\ 0 & \boldsymbol{A}_{2,2} \end{bmatrix}$, $\boldsymbol{A}_{1,1} \in C^{p\times p}$, $\boldsymbol{A}_{2,2} \in C^{(n-p)\times(n-p)}$. Show: $\Lambda(\boldsymbol{A}) = \Lambda(\boldsymbol{A}_{1,1}) \cup \Lambda(\boldsymbol{A}_{2,2})$.

5.1. **Schur decomposition.** Every matrix is unitarily similar to a triangular matrix. This is the content of the Schur decomposition theorem, which is an important analytic and numerical tool.

Let us show first a general lemma.

Lemma 5.1. *Assume the matrix* $\boldsymbol{X} \in \mathbb{C}^{n\times p}$ *has linearly independent columns and* $\boldsymbol{AX} = \boldsymbol{XB}$, $\boldsymbol{A} \in \mathbb{C}^{n\times n}$, $\boldsymbol{B} \in \mathbb{C}^{p\times p}$. *Then* $R(\boldsymbol{X})$ *is* $\boldsymbol{A}$ *–invariant (* $\boldsymbol{x} \in R(\boldsymbol{X}) \implies \boldsymbol{Ax} \in R(\boldsymbol{X})$ *) and* $\Lambda(\boldsymbol{B}) \subset \Lambda(\boldsymbol{A})$. *Further, there exists an invertible* $\boldsymbol{S}$ *and a unitary* $\boldsymbol{Q}$ *such that*

$$\boldsymbol{S}^{-1}\boldsymbol{AS} = \begin{bmatrix} \boldsymbol{B} & \widetilde{\boldsymbol{A}}_{1,2} \\ 0 & \widetilde{\boldsymbol{A}}_{2,2} \end{bmatrix} \qquad \textit{and} \qquad \boldsymbol{Q}^*\boldsymbol{AQ} = \begin{bmatrix} \boldsymbol{T}_{1,1} & \boldsymbol{T}_{1,2} \\ 0 & \boldsymbol{T}_{2,2} \end{bmatrix} ,$$

where $\boldsymbol{T}_{1,1} \sim \boldsymbol{B}$. *In particular,* $\Lambda(\boldsymbol{T}_{1,1}) = \Lambda(\boldsymbol{B})$.

Proof. Let $\boldsymbol{Y} \in \mathbb{C}^{n\times(n-p)}$ be such that $\boldsymbol{S} = \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Y} \end{bmatrix}$ is invertible. Then $\boldsymbol{AS} = \begin{bmatrix} \boldsymbol{AX} & \boldsymbol{AY} \end{bmatrix} = \begin{bmatrix} \boldsymbol{XB} & \boldsymbol{AY} \end{bmatrix}$, so that

$$\boldsymbol{S}^{-1}\boldsymbol{AS} = \begin{bmatrix} \begin{bmatrix} \boldsymbol{I} \\ 0 \end{bmatrix} \boldsymbol{B} & \boldsymbol{S}^{-1}\boldsymbol{AY} \end{bmatrix} = \begin{bmatrix} \boldsymbol{B} & \widetilde{\boldsymbol{A}}_{1,2} \\ 0 & \widetilde{\boldsymbol{A}}_{2,2} \end{bmatrix} .$$

Take Householder matrices $\boldsymbol{H}_1, \dots, \boldsymbol{H}_p$ such that $\boldsymbol{H}_p \dots \boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{X} = \begin{bmatrix} \boldsymbol{R} \\ 0 \end{bmatrix}$ where $\boldsymbol{R}$ is an upper triangular matrix. Then $\boldsymbol{Q} = \boldsymbol{H}_1\boldsymbol{H}_2 \dots \boldsymbol{H}_p = \begin{bmatrix} \boldsymbol{Q}_1 & \boldsymbol{Q}_2 \end{bmatrix}$ is unitary and $\boldsymbol{Q}_1 = \boldsymbol{XR}^{-1}$, $\boldsymbol{Q}_2^*\boldsymbol{X} = \boldsymbol{Q}_2^*\boldsymbol{Q}_1\boldsymbol{R} = 0$. Hence the matrix $\boldsymbol{T} = \boldsymbol{Q}^*\boldsymbol{AQ}$ satisfies:

$$\begin{aligned} \boldsymbol{T}_{1,1} &= \boldsymbol{Q}_1^*\boldsymbol{AQ}_1 = \boldsymbol{Q}_1^*\boldsymbol{AXR}^{-1} = \boldsymbol{Q}_1^*\boldsymbol{XBR}^{-1} = \boldsymbol{RBR}^{-1} \\ \boldsymbol{T}_{2,1} &= \boldsymbol{Q}_2^*\boldsymbol{AQ}_1 = \boldsymbol{Q}_2^*\boldsymbol{XBR}^{-1} = 0 . \end{aligned}$$

□

Theorem 5.2 (Schur decomposition)**.** *For every* $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *there exists a unitary* $\boldsymbol{Q}$ *such that* $\boldsymbol{T} = \boldsymbol{Q}^*\boldsymbol{AQ}$ *is upper triangular. Here the diagonal elements of* $\boldsymbol{T}$ *(the eigenvalues of* $\boldsymbol{A}$ *) can be in any order.*

Proof. By induction: the case $n = 1$ is clear. Let λ be an eigenvalue of the matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ and $\boldsymbol{x} \neq 0$ a correspoding eigenvector. By the previous lemma there exists a unitary $\boldsymbol{U}$ such that

$$\boldsymbol{U}^*\boldsymbol{AU} = \begin{bmatrix} \lambda & \boldsymbol{w}^T \\ 0 & \boldsymbol{C} \end{bmatrix} .$$

By the induction hypotheses there exists a unitary $\widetilde{\boldsymbol{U}}$ such that $\widetilde{\boldsymbol{U}}^*\boldsymbol{C}\widetilde{\boldsymbol{U}}$ is an upper triangular matrix. Set: $\boldsymbol{Q} = \boldsymbol{U}\begin{bmatrix} 1 & \\ & \widetilde{\boldsymbol{U}} \end{bmatrix}$. Then

$$\boldsymbol{T} = \boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q} = \begin{bmatrix} 1 & \\ & \widetilde{\boldsymbol{U}} \end{bmatrix}^* \boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U} \begin{bmatrix} 1 & \\ & \widetilde{\boldsymbol{U}} \end{bmatrix} = \begin{bmatrix} \lambda & \boldsymbol{w}^T\widetilde{\boldsymbol{U}} \\ 0 & \widetilde{\boldsymbol{U}}^*\boldsymbol{C}\widetilde{\boldsymbol{U}} \end{bmatrix}$$

is an upper triangular matrix. □

Problem 5.3. Let $p_{\boldsymbol{A}}(z) = \det(z\boldsymbol{I} - \boldsymbol{A}) = z^n + c_{n-1}z^{n-1} + \cdots + c_1 z + c_0$ be the characteristic polynomial of matrix $\boldsymbol{A}$. Prove the *Cayley–Hamilton theorem*:

$$p_{\boldsymbol{A}}(\boldsymbol{A}) = \boldsymbol{A}^n + c_{n-1}\boldsymbol{A}^{n-1} + \cdots + c_1\boldsymbol{A} + c_0\boldsymbol{I} = 0 \ .$$

Hint: For a triangular $\boldsymbol{A} = \left[\begin{smallmatrix} \lambda & \boldsymbol{w} \\ 0 & \boldsymbol{B} \end{smallmatrix}\right]$ use $p_{\boldsymbol{A}}(\boldsymbol{A}) = (\boldsymbol{A} - \lambda\boldsymbol{I})\, p_{\boldsymbol{B}}(\boldsymbol{A})$ and induction.

Problem 5.4 (Real Schur decomposition)**.** Show, that for every $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ there exists an orthogonal matrix $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ (orthogonal: $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$) such that $\boldsymbol{T} = \boldsymbol{V}^T\boldsymbol{A}\boldsymbol{V}$ is a block upper triangular matrix, where the diagonal blocks of $\boldsymbol{T}$ are either scalars corresponding to the real eigenvalues of $\boldsymbol{A}$, or 2×2–matrices, which have a pair of complex eigenvalues. Hint: if $\lambda = \alpha + i\beta$ is a complex eigenvalue of $\boldsymbol{A}$ then there exist vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ such that $\boldsymbol{A}\begin{bmatrix} \boldsymbol{u} & \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} \boldsymbol{u} & \boldsymbol{v} \end{bmatrix}\begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix}$. Orthonormalise (Gram-Schmidt): $\begin{bmatrix} \boldsymbol{u} & \boldsymbol{v} \end{bmatrix}\begin{bmatrix} a & b \\ 0 & c \end{bmatrix} = \begin{bmatrix} \boldsymbol{q}_1 & \boldsymbol{q}_2 \end{bmatrix} = \boldsymbol{Q}_1$, form an orthogonal $\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{Q}_1 & \boldsymbol{Q}_2 \end{bmatrix}$ and multiply: $\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}$.

Definition 5.1. Matrix $\boldsymbol{A}$ is *normal*, if $\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^*$, i.e., if $[\boldsymbol{A}, \boldsymbol{A}^*] = 0$, where the *commutator* is defined as $[\boldsymbol{A}, \boldsymbol{B}] = \boldsymbol{A}\boldsymbol{B} - \boldsymbol{B}\boldsymbol{A}$.

Theorem 5.3. $\boldsymbol{A}$ *is normal, if and only if it is unitarily similar to a diagonal matrix.*

Proof. If $\boldsymbol{D}$ is a diagonal matrix and $\boldsymbol{Q}$ unitary such that $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^*$, then

$$\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{Q}\bar{\boldsymbol{D}}\boldsymbol{Q}^*\boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^* = \boldsymbol{Q}\boldsymbol{D}\bar{\boldsymbol{D}}\boldsymbol{Q}^* = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^*\boldsymbol{Q}\bar{\boldsymbol{D}}\boldsymbol{Q}^* = \boldsymbol{A}\boldsymbol{A}^* \ .$$

On the other hand, if $\boldsymbol{A}$ is normal, then let $\boldsymbol{Q}$ be unitary such that $\boldsymbol{T} = \boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}$ is upper triangular. Then

$$\boldsymbol{T}^*\boldsymbol{T} = \boldsymbol{Q}^*\boldsymbol{A}^*\boldsymbol{Q}\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{A}^*\boldsymbol{Q} = \boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}\boldsymbol{Q}^*\boldsymbol{A}^*\boldsymbol{Q} = \boldsymbol{T}\boldsymbol{T}^* \ ,$$

so that $\boldsymbol{T}$ is normal. We claim that, that a normal upper triangular matrix is diagonal. By induction: the case $n = 1$ is clear. Let $\boldsymbol{T} \in \mathbb{C}^{n\times n}$ be a normal upper triangular matrix. Then

$$(\boldsymbol{T}^*\boldsymbol{T})_{1,1} = |t_{1,1}|^2 = (\boldsymbol{T}\boldsymbol{T}^*)_{1,1} = |t_{1,1}|^2 + |t_{1,2}|^2 + \cdots + |t_{1,n}|^2 \ ,$$

so that $t_{1,2} = \cdots = t_{1,n} = 0$. Hence $\boldsymbol{T} = \begin{bmatrix} t_{1,1} & 0 \\ 0 & \widetilde{\boldsymbol{T}} \end{bmatrix}$, where $\widetilde{\boldsymbol{T}}$ is a $(n-1)\times(n-1)$ normal upper triangular matrix. By the induction hypotheses $\widetilde{\boldsymbol{T}}$ and hence also $\boldsymbol{T}$ is a diagonal matrix. □

Problem 5.5. Show that $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is normal if and only if $\|\boldsymbol{A}\|_F^2 = \sum_{j=1}^n |\lambda_j|^2$.

Theorem 5.4. *For given* $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *there exist vectors* $\boldsymbol{u}_j, \boldsymbol{v}_j \in \mathbb{C}^n$, $j = 1, \ldots, n$, *such that*

$$\boldsymbol{I} - z\,\boldsymbol{A} = (\boldsymbol{I} - z\,\boldsymbol{u}_n\,\boldsymbol{v}_n^*) \ldots (\boldsymbol{I} - z\,\boldsymbol{u}_2\,\boldsymbol{v}_2^*)(\boldsymbol{I} - z\,\boldsymbol{u}_1\,\boldsymbol{v}_1^*)\ , \tag{5.1}$$

$z \in \mathbb{C}$. *Further, the numbers* $\langle \boldsymbol{u}_j, \boldsymbol{v}_j \rangle$ *are the eigenvalues of* $\boldsymbol{A}$.

Proof. Take the Schur decomposition $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{T}\boldsymbol{Q}^*$ and write the upper triangular matrix as

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{t}_1^* \\ \boldsymbol{t}_2^* \\ \vdots \\ \boldsymbol{t}_n^* \end{bmatrix} = \boldsymbol{e}_1\,\boldsymbol{t}_1^* + \boldsymbol{e}_2\,\boldsymbol{t}_2^* + \cdots + \boldsymbol{e}_n\,\boldsymbol{t}_n^*\ .$$

Since the first $j-1$ components of $\boldsymbol{t}_j$ are zero, we have $\boldsymbol{t}_j^*\boldsymbol{e}_{j-k} = 0$, $j \geq 2,\ k \geq 1$. Hence

$$\begin{aligned} \boldsymbol{I} - z\,\boldsymbol{T} &= \boldsymbol{I} - z\ (\boldsymbol{e}_1\,\boldsymbol{t}_1^* + \boldsymbol{e}_2\,\boldsymbol{t}_2^* + \cdots + \boldsymbol{e}_n\,\boldsymbol{t}_n^*) \\ &= (\boldsymbol{I} - z\,\boldsymbol{e}_n\,\boldsymbol{t}_n^*)(\boldsymbol{I} - z\,\boldsymbol{e}_{n-1}\,\boldsymbol{t}_{n-1}^*) \ldots (\boldsymbol{I} - z\,\boldsymbol{e}_1\,\boldsymbol{t}_1^*) \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{I} - z\,\boldsymbol{A} &= \boldsymbol{Q}(\boldsymbol{I} - z\,\boldsymbol{T})\boldsymbol{Q}^* \\ &= \boldsymbol{Q}(\boldsymbol{I} - z\,\boldsymbol{e}_n\,\boldsymbol{t}_n^*)\boldsymbol{Q}^*\boldsymbol{Q}(\boldsymbol{I} - z\,\boldsymbol{e}_{n-1}\,\boldsymbol{t}_{n-1}^*)\boldsymbol{Q}^* \ldots \boldsymbol{Q}(\boldsymbol{I} - z\,\boldsymbol{e}_1\,\boldsymbol{t}_1^*)\boldsymbol{Q}^* \\ &= (\boldsymbol{I} - z\,\boldsymbol{Q}\boldsymbol{e}_n\,\boldsymbol{t}_n^*\boldsymbol{Q}^*) \ldots (\boldsymbol{I} - z\,\boldsymbol{Q}\boldsymbol{e}_1\,\boldsymbol{t}_1^*\boldsymbol{Q}^*)\ . \end{aligned}$$

Setting $\boldsymbol{u}_j = \boldsymbol{Q}\boldsymbol{e}_j$, $\boldsymbol{v}_j = \boldsymbol{Q}\boldsymbol{t}_j$ we get (5.1). Since the numbers

$$\boldsymbol{v}_j^*\boldsymbol{u}_j = \boldsymbol{T}_j^*\boldsymbol{Q}^*\boldsymbol{Q}\boldsymbol{e}_j = \boldsymbol{t}_j^*\boldsymbol{e}_j$$

are the diagonal elements of $\boldsymbol{T}$, they are the eigenvalues of $\boldsymbol{A}$. □

Corollary 5.5. *From* (5.1) *we can further write*

$$(\boldsymbol{I} - z\,\boldsymbol{A})^{-1} = \left(\boldsymbol{I} + \tfrac{z}{1-z\,\boldsymbol{v}_1^*\boldsymbol{u}_1}\,\boldsymbol{u}_1\,\boldsymbol{v}_1^*\right) \ldots \left(\boldsymbol{I} + \tfrac{z}{1-z\,\boldsymbol{v}_n^*\boldsymbol{u}_n}\,\boldsymbol{u}_n\,\boldsymbol{v}_n^*\right)$$

whenever $z\,\boldsymbol{v}_j^*\boldsymbol{u}_j \neq 1$ *for all* j.

Proof. Use lemma 2.1. □

5.2. **The Sylvester equation.** Consider the *Sylvester equation:* for given $\boldsymbol{A} \in \mathbb{C}^{n\times n}$, $\boldsymbol{B} \in \mathbb{C}^{m\times m}$, and $\boldsymbol{C} \in \mathbb{C}^{n\times m}$ find an $\boldsymbol{X} \in \mathbb{C}^{n\times m}$ such that

$$\boldsymbol{AX} - \boldsymbol{XB} = \boldsymbol{C} \ . \tag{5.2}$$

The following theorem gives the sovability condition for (5.2).

Theorem 5.6. *The linear transformation* $\boldsymbol{\phi} \, : \, \mathbb{C}^{n\times m} \to \mathbb{C}^{n\times m}$

$$\boldsymbol{\phi}(\boldsymbol{X}) = \boldsymbol{AX} - \boldsymbol{XB}$$

is invertible if and only if $\Lambda(\boldsymbol{A}) \cap \Lambda(\boldsymbol{B}) = \emptyset$.

Proof. Let $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ be unitary such that $\boldsymbol{L}^* = \boldsymbol{Q}_1^*\boldsymbol{A}^*\boldsymbol{Q}_1$ and $\boldsymbol{R} = \boldsymbol{Q}_2^*\boldsymbol{B}\boldsymbol{Q}_2$ are upper triangular matrices. Then

$$\begin{aligned}
&\boldsymbol{AX} - \boldsymbol{XB} = 0 && \Longleftrightarrow\\
&\boldsymbol{Q}_1^*\boldsymbol{A}\boldsymbol{Q}_1\boldsymbol{Q}_1^*\boldsymbol{X}\boldsymbol{Q}_2 - \boldsymbol{Q}_1^*\boldsymbol{X}\boldsymbol{Q}_2\boldsymbol{Q}_2^*\boldsymbol{B}\boldsymbol{Q}_2 = 0 && \Longleftrightarrow\\
&\boldsymbol{LY} - \boldsymbol{YR} = 0 \ ,
\end{aligned}$$

where $\boldsymbol{Y} = \boldsymbol{Q}_1^*\boldsymbol{X}\boldsymbol{Q}_2$. Now $\Lambda(\boldsymbol{L}) \cap \Lambda(\boldsymbol{R}) = \emptyset$, so that $l_{j,j} \neq r_{k,k}$ $\forall j,k$ and equation $\boldsymbol{\phi}(\boldsymbol{X}) = 0$ implies

$$\begin{bmatrix} l_{1,1} & 0 & \dots & 0\\ l_{2,1} & l_{2,2} & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ l_{n,1} & l_{n,2} & \dots & l_{n,n}\end{bmatrix}\begin{bmatrix} y_{1,1} & y_{1,2} & \dots & y_{1,m}\\ y_{2,1} & y_{2,2} & \dots & y_{2,m}\\ \vdots & \vdots & \ddots & \vdots\\ y_{n,1} & y_{n,2} & \dots & y_{n,m}\end{bmatrix} -$$

$$\begin{bmatrix} y_{1,1} & y_{1,2} & \dots & y_{1,m}\\ y_{2,1} & y_{2,2} & \dots & y_{2,m}\\ \vdots & \vdots & \ddots & \vdots\\ y_{n,1} & y_{n,2} & \dots & y_{n,m}\end{bmatrix}\begin{bmatrix} r_{1,1} & r_{1,2} & \dots & r_{1,m}\\ 0 & r_{2,2} & \dots & r_{2,m}\\ \vdots & & \ddots & \vdots\\ 0 & 0 & \dots & r_{m,m}\end{bmatrix} = 0 \ ,$$

which recursively gives

$$\begin{array}{ccccc}
(l_{1,1}-r_{1,1})y_{1,1} = 0_{1,1} \Longrightarrow & y_{1,1}=0 & \Longrightarrow (l_{1,1}-r_{2,2})y_{1,2} = 0_{1,2} & \Longrightarrow & y_{1,2}=0\\
\Downarrow & \dots & \Longrightarrow (l_{1,1}-r_{m,m})y_{1,m} = 0_{1,m} & \Longrightarrow & y_{1,m}=0\\
(l_{2,2}-r_{1,1})y_{2,1} = 0_{2,1} \Longrightarrow & y_{2,1}=0 & \Longrightarrow (l_{2,2}-r_{2,2})y_{2,2} = 0_{2,2} & \Longrightarrow & y_{2,2}=0\\
\Downarrow & \dots & \Longrightarrow (l_{2,2}-r_{m,m})y_{2,m} = 0_{2,m} & \Longrightarrow & y_{2,m}=0\\
\vdots & \vdots & & & \\
(l_{n,n}-r_{1,1})y_{n,1} = 0_{n,1} \Longrightarrow & y_{n,1}=0 & \Longrightarrow (l_{n,n}-r_{2,2})y_{n,2} = 0_{n,2} & \Longrightarrow & y_{n,2}=0\\
 & \dots & \Longrightarrow (l_{n,n}-r_{m,m})y_{n,m} = 0_{n,m} & \Longrightarrow & y_{n,m}=0
\end{array}$$

i.e., $\boldsymbol{Y} = 0$ and $\boldsymbol{X} = \boldsymbol{Q}_1\boldsymbol{Y}\boldsymbol{Q}_2^* = 0$. Hence the kernel of $\boldsymbol{\phi}$ is $\{0\}$, so that $\boldsymbol{\phi}$ is invertible.

For necessity, see problem 5.7. □

Remark 5.1. The proof above shows how one can (and usually should) solve the Sylvester equation: obtain the Schur decompositions of the coefficient matrices and on the right hand side set $\boldsymbol{Q}_1^*\boldsymbol{C}\boldsymbol{Q}_2$. We get the matrix $\boldsymbol{Y}$ by the recursion above and then $\boldsymbol{X}$ we get by multiplication. This is much cheaper than solving an $nm \times nm$ system of linear equations.

Problem 5.6. A more elegant proof (but no algorithm) for the previous theorem can be obtained as follows. Show: $\boldsymbol{AX} = \boldsymbol{XB} \implies \boldsymbol{A}^k\boldsymbol{X} = \boldsymbol{XB}^k$ and $p(\boldsymbol{A})\,\boldsymbol{X} = \boldsymbol{X}\,p(\boldsymbol{B})$ for every polynomial p. Take p to be the characteristic polynomial of $\boldsymbol{A}$. Show, that $p_{\boldsymbol{A}}(\boldsymbol{B})$ is invertible.

Problem 5.7. Show, that the eigenvalues of $\boldsymbol{\phi}$ are:

$$\lambda_i - \mu_j \ , \qquad \lambda_i \in \Lambda(\boldsymbol{A}) \, , \ \mu_j \in \Lambda(\boldsymbol{B}) \ .$$

Problem 5.8. Assume $\boldsymbol{AX} = \boldsymbol{XM}$ and $\boldsymbol{A}^*\boldsymbol{Y} = \boldsymbol{YN}$, where $\Lambda(\boldsymbol{M}) \cap \Lambda(\boldsymbol{N}^*) = \emptyset$. Show that $\boldsymbol{Y}^*\boldsymbol{X} = 0$. In particular, if $\boldsymbol{Ax} = \lambda\boldsymbol{x}$, $\boldsymbol{A}^*\boldsymbol{y} = \mu\boldsymbol{y}$, and $\bar\mu \neq \lambda$, then $\boldsymbol{y} \perp \boldsymbol{x}$.

By the Schur decomposition we know that every matrix is unitarily similar to a triangular matrix. Now we want to make further similarity tranformations to get every matrix in a more simple form. We will need the following.

Lemma 5.7. *Let* $\boldsymbol{T} = \begin{array}{cc} \begin{bmatrix} \boldsymbol{T}_{1,1} & \boldsymbol{T}_{1,2} \\ 0 & \boldsymbol{T}_{2,2} \end{bmatrix} & \begin{array}{c} p \\ q \end{array} \\ \begin{array}{cc} p & \quad q \end{array} & \end{array}$ *be such that* $\Lambda(\boldsymbol{T}_{1,1}) \cap \Lambda(\boldsymbol{T}_{2,2}) = \emptyset$. *Then there exists* $\boldsymbol{Z} \in \mathbb{C}^{p\times q}$ *such that*

$$\begin{bmatrix} \boldsymbol{I} & \boldsymbol{Z} \\ 0 & \boldsymbol{I} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{T}_{1,1} & \boldsymbol{T}_{1,2} \\ 0 & \boldsymbol{T}_{2,2} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \boldsymbol{Z} \\ 0 & \boldsymbol{I} \end{bmatrix} = \begin{bmatrix} \boldsymbol{T}_{1,1} & 0 \\ 0 & \boldsymbol{T}_{2,2} \end{bmatrix} \ .$$

Proof. Since $\begin{bmatrix} \boldsymbol{I} & \boldsymbol{Z} \\ 0 & \boldsymbol{I} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{Z} \\ 0 & \boldsymbol{I} \end{bmatrix}$, the matrix product on the left hand side above is

$$\begin{bmatrix} \boldsymbol{T}_{1,1} & \boldsymbol{T}_{1,1}\boldsymbol{Z} - \boldsymbol{Z}\boldsymbol{T}_{2,2} + \boldsymbol{T}_{1,2} \\ 0 & \boldsymbol{T}_{2,2} \end{bmatrix} \ ,$$

so that a suitable $\boldsymbol{Z}$ is found as the solution of equation $\boldsymbol{T}_{1,1}\boldsymbol{Z} - \boldsymbol{Z}\boldsymbol{T}_{2,2} = -\boldsymbol{T}_{1,2}$, which exists by the previous theorem. □

Let $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ be a given matrix with spectrum $\Lambda(\boldsymbol{A}) = \{\lambda_1, \dots, \lambda_p\}$. Take the Schur decomposition

$$\boldsymbol{T} = \boldsymbol{Q}^* \boldsymbol{A} \boldsymbol{Q} = \begin{bmatrix} \boldsymbol{T}_{1,1} & \boldsymbol{T}_{1,2} & \dots & \boldsymbol{T}_{1,p} \\ 0 & \boldsymbol{T}_{2,2} & \dots & \boldsymbol{T}_{2,p} \\ & & \ddots & \vdots \\ 0 & 0 & & \boldsymbol{T}_{p,p} \end{bmatrix} ,$$

where the diagonal of each $T_{j,j}$ consists of λ_j 's. By the previous lemma we get:

Corollary 5.8. *Every $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is similar to a block diagonal matrix*

$$\operatorname{diag}(\boldsymbol{T}_{1,1}, \boldsymbol{T}_{2,2}, \dots \boldsymbol{T}_{p,p}) ,$$

where the upper triangular matrices $\boldsymbol{T}_{j,j}$ have constant diagonals.

If $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ has n distinct eigenvalues, this form is a diagonal matrix. Note that transforming a matrix to this form is a stable process if the eigenvalues are well separated (see problem 5.7).

Problem 5.9. Assume that an eigenvalue λ_j is a multiple root of the characteristic polynomial and that the number of linearly independent eigenvectors corresponding to λ_j equals to its multiplicity. Show that the corresponding $\boldsymbol{T}_{j,j}$ is diagonal.

5.3. **Jordan form.**

If λ_j is a multiple root of the characteristic polynomial, then the number of linearly independent eigenvectors may be less than the multiplicity of λ_j. Here we discuss, to how simple form the triangular block $\boldsymbol{T}_{j,j}$ can be transformed in this case.

A *Jordan block* is a matrix of the form

$$\boldsymbol{J}_k(\lambda) = \begin{bmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda & 1 \\ & & & & \lambda \end{bmatrix} \in \mathbb{C}^{k\times k} .$$

Block diagonal matrices consisting of such diagonal blocks

$$\boldsymbol{J} = \operatorname{diag}(\boldsymbol{J}_{k_1}(\lambda_1), \boldsymbol{J}_{k_2}(\lambda_2), \dots, \boldsymbol{J}_{k_m}(\lambda_m))$$

are called *Jordan matrices.* Here the λ_j 's need not be different.

Theorem 5.9 (Jordan form). *For every matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ there exists an invertible matrix $\boldsymbol{S} \in \mathbb{C}^{n\times n}$ such that $\boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S}$ is a Jordan matrix.*

Proof. Let $\boldsymbol{V}$ be an invertible matrix that transforms $\boldsymbol{A}$ to the block diagonal form of the previous corollary. Write the diagonal blocks in the form $\boldsymbol{T}_{j,j} = \lambda_j \boldsymbol{I} + \boldsymbol{N}_j$, when $\boldsymbol{N}_j$ is a strictly upper triangular matrix (the diagonal is also zero). By Proposition 5.10 below, for each $\boldsymbol{N}_j$ there exists an $\boldsymbol{S}_j$ such that $\boldsymbol{J}_j = \boldsymbol{S}_j^{-1}\boldsymbol{N}_j\boldsymbol{S}_j$ is a Jordan matrix, so that $\boldsymbol{A}$ gets the Jordan form:

$$[\boldsymbol{V}\operatorname{diag}(\boldsymbol{S}_1,\ldots,\boldsymbol{S}_p)]^{-1}\boldsymbol{A}\boldsymbol{V}\operatorname{diag}(\boldsymbol{S}_1,\ldots,\boldsymbol{S}_p) = \operatorname{diag}(\lambda_1\boldsymbol{I}+\boldsymbol{J}_1,\ldots,\lambda_p\boldsymbol{I}+\boldsymbol{J}_p)\ .$$

□

Problem 5.10. Show: $\boldsymbol{J}_k(0)^k = 0$, $\quad \boldsymbol{J}_k(0)^T\boldsymbol{J}_k(0) = \begin{bmatrix}0 & 0\\ 0 & \boldsymbol{I}_{k-1}\end{bmatrix}$,

$$\boldsymbol{J}_k(0)\,\boldsymbol{e}_j = \boldsymbol{e}_{j-1}\,,\ j=2,\ldots,k\ , \qquad [\boldsymbol{I}-\boldsymbol{J}_k(0)^T\boldsymbol{J}_k(0)]\,\boldsymbol{x} = (\boldsymbol{x}^T\boldsymbol{e}_1)\,\boldsymbol{e}_1\ .$$

Definition 5.2. Matrix $\boldsymbol{N}$ is said to be *nilpotent* if $\boldsymbol{N}^p = 0$ for some $p \geq 1$. For example, strictly upper triangular matrices are nilpotent.

Problem 5.11. Assume $\boldsymbol{N}^p = 0$. Show that $(\boldsymbol{I}-\boldsymbol{N})^{-1} = \boldsymbol{I}+\boldsymbol{N}+\boldsymbol{N}^2+\cdots+\boldsymbol{N}^{p-1}$.

Proposition 5.10. *Let* $\boldsymbol{N} \in \mathbb{C}^{n\times n}$ *be nilpotent. Then there exists an invertible* $\boldsymbol{S} \in \mathbb{C}^{n\times n}$ *and* $n_1 \geq n_2 \geq \ldots n_m \geq 1$, $\sum_j n_j = n$, *such that*

$$\boldsymbol{S}^{-1}\boldsymbol{N}\boldsymbol{S} = \operatorname{diag}(\boldsymbol{J}_{n_1}(0),\ldots,\boldsymbol{J}_{n_m}(0))\ .$$

Proof. By induction: for $\boldsymbol{N} = 0$ there is nothing to prove so the case $n=1$ is clear. For $0 \neq \boldsymbol{N} \in \mathbb{C}^{n\times n}$ let $p \in \mathbb{N}$, $\boldsymbol{u} \in \mathbb{C}^n$ be such that $\boldsymbol{N}^p = 0$, $\boldsymbol{N}^{p-1}\boldsymbol{u} \neq 0$. Also take $\boldsymbol{w}$ such that $\langle \boldsymbol{N}^{p-1}\boldsymbol{u}, \boldsymbol{w}\rangle \neq 0$. Set

$$E = \left\{\boldsymbol{x} \,\middle|\, \langle \boldsymbol{N}^j\boldsymbol{x}, \boldsymbol{w}\rangle = 0\,,\ \forall j \geq 0\right\}\ .$$

Clearly $\boldsymbol{N}E \subset E$. The set $W = \{\boldsymbol{w}, \boldsymbol{N}^*\boldsymbol{w}, \ldots, (\boldsymbol{N}^*)^{p-1}\boldsymbol{w}\}$ is linearly independent since if $\gamma_1\boldsymbol{w}+\gamma_2\boldsymbol{N}^*\boldsymbol{w}+\ldots\gamma_p(\boldsymbol{N}^*)^{p-1}\boldsymbol{w} = 0$ then multiplication by $(\boldsymbol{N}^*)^{p-1}$ gives $\gamma_1 = 0$ after which multiplication by $(\boldsymbol{N}^*)^{p-2}$ gives $\gamma_2 = 0$ and so on. Hence, $E = W^\perp$ and it has dimension $n-p$.

Set $\boldsymbol{U} = [\boldsymbol{N}^{p-1}\boldsymbol{u},\ldots,\boldsymbol{N}\boldsymbol{u},\boldsymbol{u}]$ and let the columns of $\boldsymbol{V} \in \mathbb{C}^{n\times(n-p)}$ span E. To show that $[\boldsymbol{U}\ \boldsymbol{V}]$ is nonsingular assume $[\boldsymbol{U}\ \boldsymbol{V}]\left[\begin{smallmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{smallmatrix}\right] = 0$. If $\boldsymbol{\alpha} = 0$ then also $\boldsymbol{\beta} = 0$, so let $j \in \{1,\ldots,p\}$ be largest such that $\alpha_j \neq 0$. We have $\boldsymbol{U}\boldsymbol{\alpha} = -\boldsymbol{V}\boldsymbol{\beta} \in E$ so that multiplying $\sum_{k=1}^{j}\alpha_k\boldsymbol{N}^{p-k}\boldsymbol{u} \in E$ by $\boldsymbol{N}^{j-1}$ and using $\boldsymbol{N}E \subset E$ we get $\alpha_j\boldsymbol{N}^{p-1}\boldsymbol{u} \in E$, i.e., $\boldsymbol{N}^{p-1}\boldsymbol{u} \perp \boldsymbol{w}$, a contradiction.

Now we have $\boldsymbol{N}\boldsymbol{U} = \boldsymbol{U}\boldsymbol{J}_p(0)$ and $\boldsymbol{N}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{M}$ for some $\boldsymbol{M} \in \mathbb{C}^{(n-p)\times(n-p)}$. Hence

$$\boldsymbol{N}\begin{bmatrix}\boldsymbol{U} & \boldsymbol{V}\end{bmatrix} = \begin{bmatrix}\boldsymbol{U} & \boldsymbol{V}\end{bmatrix}\begin{bmatrix}\boldsymbol{J}_p(0) & 0\\ 0 & \boldsymbol{M}\end{bmatrix}\ ,$$

i.e., $\boldsymbol{N} \sim \left[\begin{smallmatrix}\boldsymbol{J}_p(0) & 0\\ 0 & \boldsymbol{M}\end{smallmatrix}\right]$. By the induction hypotheses, $\boldsymbol{M}$ is similar to some Jordan matrix $\boldsymbol{J}_{\boldsymbol{M}}$ and hence $\boldsymbol{N} \sim \left[\begin{smallmatrix}\boldsymbol{J}_p(0) & 0\\ 0 & \boldsymbol{J}_{\boldsymbol{M}}\end{smallmatrix}\right]$. Clearly, the largest Jordan block of $\boldsymbol{J}_{\boldsymbol{M}}$ has size $\leq p$ □

Remark 5.2. The previous proof shows, that, if $\boldsymbol{N}$ is a real nilpotent matrix, then it can be put to a Jordan form with a real similarity transform.

Problem 5.12. *The minimal polynomial* $m_{\boldsymbol{A}}$ of a square matrix $\boldsymbol{A}$ is defined to be the monic (the coefficient of the highest degree term is one) polynomial of lowest degree that *annihilates* $\boldsymbol{A}$, i.e., satisfies $p(\boldsymbol{A}) = 0$. Show that this is unique. For each eigenvalue λ_j, $j = 1, \ldots, m$ of $\boldsymbol{A}$ let k_j be the dimension of the largest Jordan block correspoding to λ_j. Show:

$$m_{\boldsymbol{A}}(z) = (z - \lambda_1)^{k_1}(z - \lambda_2)^{k_2} \ldots (z - \lambda_m)^{k_m} \ .$$

Problem 5.13. Show, that a matrix is diagonalisable (similar to a diagonal matrix) if and only if the roots of the minimal polynomial are simple.

Problem 5.14. Show, that every square matrix $\boldsymbol{A}$ satisfies $\boldsymbol{A} \sim \boldsymbol{A}^T$. Hint: show that

$$\begin{bmatrix} & & & 1 \\ & & \iddots & \\ & 1 & & \\ 1 & & & \end{bmatrix} \boldsymbol{J}_k(\lambda) \begin{bmatrix} & & & 1 \\ & & \iddots & \\ & 1 & & \\ 1 & & & \end{bmatrix} = \boldsymbol{J}_k(\lambda)^T \ .$$

Problem 5.15. The Schur decomposition can be computed as stably as one computes eigenvectors since these are used to form the Householder transformations, which operate stably. Jordan form is much more difficult to compute. Generally the numerical computation of a Jordan form is an ill–posed task unless the eigenvalues are simple. This is because the Jordan form is not a continuous function of the matrix. Think and give examples of which steps in the derivation of the Jordan form are not continuous, in particular, in Lemma 5.7 and proposition 5.10.

Problem 5.16. Show: $(\boldsymbol{J}_k(\lambda)^p)_{j,j+l} = \binom{p}{l} \lambda^{p-l}$, when $0 \leq l \leq \min(p, k-j)$, otherwise $(\boldsymbol{J}_k(\lambda)^p)_{j,j+l} = 0$.

Problem 5.17. Show: if all eigenvalues of $\boldsymbol{A}$ satisfy $|\lambda| < 1$, then $\lim_{k\to\infty} \boldsymbol{A}^k = 0$.

Problem 5.18. Show that $\|\boldsymbol{J}_k(\lambda)\|_2 \leq |\lambda| + 1$ and if $\lambda \neq 0$, then

$$\left\|\boldsymbol{J}_k(\lambda)^{-1}\right\|_2 \leq \sum_{j=1}^{k} |\lambda|^{-j} \ .$$

6. Numerical computation of eigenvalues

6.1. **Sensitivity of eigenvalues.**

Consider the changes of eigenvalues when the matrix is slightly perturbed.

Theorem 6.1. *The eigenvalues depend continuously on the matrix.*

Proof. The coefficients of the characteristic polynomial $\det(\lambda \boldsymbol{I} - \boldsymbol{A})$ are sums of subdeterminants of $\boldsymbol{A}$. Hence they depend continuously on the elements of the matrix. On the other hand, the roots of a polynomial are continuous functions of the coefficients, which proves the claim. □

However, the eigenvalues are not necessarily differentiable at points where the matrix has multiple eigenvalues. When approaching such points the derivative may go to infinity, as the following example shows:

$$\boldsymbol{A}(t) = \begin{bmatrix} & 1 & & \\ & & 1 & \\ & & & 1 \\ t & & & \end{bmatrix} , \qquad \Lambda(\boldsymbol{A}(t)) = \begin{cases} t^{1/4}\{1, i, -1, -i\} , & \text{when } t > 0 \\ \dfrac{|t|^{1/4}(1+i)}{\sqrt{2}}\{1, i, -1, -i\} , & \text{when } t < 0 . \end{cases}$$

The simplest theorem concerning the location of eigenvalues is the following:

Theorem 6.2 (Gershgorin discs)**.** *For any $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ we have*

$$\Lambda(\boldsymbol{A}) \subset G(\boldsymbol{A}) = \bigcup_{i=1}^{n} D_i ,$$

where $D_i \subset \mathbb{C}$ is a disc of radius $r_i = \sum_{j\neq i} |a_{i,j}|$ and centered at $a_{i,i}$.

Proof. If $\lambda \in \Lambda(\boldsymbol{A})$, then $\lambda\boldsymbol{I} - \boldsymbol{A}$ is singular. Let $\boldsymbol{x} \neq 0$ be such that $(\lambda\boldsymbol{I} - \boldsymbol{A})\,\boldsymbol{x} = 0$ and $|x_i| = \max_j |x_j|$. Then

$$(\lambda - a_{i,i})\,x_i = \sum_{j\neq i} a_{i,j}\,x_j , \qquad \text{from which} \qquad |\lambda - a_{i,i}| \le \sum_{j\neq i} |a_{i,j}| , \quad \text{i.e., } \lambda \in D_i .$$

□

Problem 6.1. Show: $\Lambda(\boldsymbol{A}) = \bigcap_{\det(\boldsymbol{X})\neq 0} G(\boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X})$.

[0] Version: April 9, 2003

Problem 6.2. Using Gershgorin discs of matrices $\boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{D}$, $\boldsymbol{D}$ diagonal, try to find a small set containing $\Lambda(\boldsymbol{A})$ in the case

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

Consider also $\boldsymbol{D}^{-1}\boldsymbol{A}^T\boldsymbol{D}$.

Problem 6.3. Show that any Gershgorin disc that does not intersect others contains an eigenvalue of $\boldsymbol{A}$. Hint: Consider the Gershgorin discs of $\boldsymbol{A}_t = \boldsymbol{D} + t\,(\boldsymbol{A} - \boldsymbol{D})$, $t \in [0,1]$, where $\boldsymbol{D}$ is the diagonal of $\boldsymbol{A}$. Use continuity of eigenvalues.

Problem 6.4. Show: $|\det(\boldsymbol{A})| \le \prod_{i=1}^{n} \Big(\sum_{j=1}^{n} |a_{i,j}| \Big)$.

A norm $\|\cdot\|_*$, defined for matrices is called a *matrix norm*, if $\|\boldsymbol{I}\|_* = 1$ and it satisfies the inequality $\|\boldsymbol{A}\boldsymbol{B}\|_* \le \|\boldsymbol{A}\|_* \|\boldsymbol{B}\|_*$ for all square matrices $\boldsymbol{A}$, $\boldsymbol{B}$ of the same size. For example, the norms, which are induced by vector norms in the standard way, satisfy this.

Problem 6.5. Show that for any matrix norm holds:

$$\|\boldsymbol{A}\|_* \ge \rho(\boldsymbol{A}) = \max_{\lambda \in \Lambda(\boldsymbol{A})} |\lambda| = \lim_{k\to\infty} \left\|\boldsymbol{A}^k\right\|_*^{1/k} .$$

Hence for a diagonal matrix we have $\|\boldsymbol{D}\|_* \ge \max_j |d_{j,j}|$.

The following result is often needed

Lemma 6.3. *Assume $\|\boldsymbol{A}\|_* < 1$ for some matrix norm. Then $\boldsymbol{I} - \boldsymbol{A}$ is invertible.*

Proof. We will show that $(\boldsymbol{I} - \boldsymbol{A})^{-1} = \sum_{k=0}^{\infty} \boldsymbol{A}^k$. Set

$$\boldsymbol{S}_n = \boldsymbol{I} + \boldsymbol{A} + \boldsymbol{A}^2 + \cdots + \boldsymbol{A}^n .$$

Then for $m > n$:

$$\begin{aligned} \|\boldsymbol{S}_m - \boldsymbol{S}_n\|_* &= \left\|\boldsymbol{A}^{n+1} + \boldsymbol{A}^{n+2} + \cdots + \boldsymbol{A}^m\right\|_* \\ &\le \|\boldsymbol{A}\|_*^{n+1} \left(1 + \|\boldsymbol{A}\|_* + \cdots + \|\boldsymbol{A}\|_*^{m-n-1}\right) \\ &\le \|\boldsymbol{A}\|_*^{n+1} \sum_{k=0}^{\infty} \|\boldsymbol{A}\|_*^k = \frac{\|\boldsymbol{A}\|_*^{n+1}}{1 - \|\boldsymbol{A}\|_*} . \end{aligned}$$

Hence, $\|\boldsymbol{S}_m - \boldsymbol{S}_n\|_* < \varepsilon$ for m and n large enough so that $\{\boldsymbol{S}_n\}_{n\geq 0}$ is a Cauchy–sequence, i.e., it converges[1]. Further,

$$(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{S}_n = \boldsymbol{S}_n - \boldsymbol{S}_{n+1} + \boldsymbol{I} = \boldsymbol{I} - \boldsymbol{A}^{n+1} \ ,$$

so that taking the limits, we get $(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{S}_\infty = \boldsymbol{I}$. □

Remark 6.1. A shorter, but nonconstructive proof goes as follows: If $\boldsymbol{I} - \boldsymbol{A}$ is not invertible, then there exists a nonzero matrix $\boldsymbol{X}$ such that $(\boldsymbol{I} - \boldsymbol{A})\,\boldsymbol{X} = 0$. By scaling, we can assume $\|\boldsymbol{X}\|_* = 1$ so that $\|\boldsymbol{A}\|_* = \|\boldsymbol{A}\|_* \|\boldsymbol{X}\|_* \geq \|\boldsymbol{A}\boldsymbol{X}\|_* = \|\boldsymbol{X}\|_* = 1$. Hence $\|\boldsymbol{A}\|_* \geq 1$.

Denote the distance of a point $\boldsymbol{x}$ from a set S by $\delta(\boldsymbol{x}, S) = \inf_{\boldsymbol{y}\in S} \|\boldsymbol{x} - \boldsymbol{y}\|$, the distance of a set R from S by $\delta(R, S) = \sup_{\boldsymbol{x}\in R} \delta(\boldsymbol{x}, S)$ and the distance between sets R and S by $d(R, S) = \max\{\delta(R, S), \delta(S, R)\}$.

If a diagonalizable matrix is perturbed, then the sensitivity of eigenvalues depends on how well linearly independent the eigenvectors are:

Theorem 6.4 (Bauer–Fike). *If* $\boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X} = \boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, *then*

$$\delta(\Lambda(\boldsymbol{A}+\boldsymbol{E}), \Lambda(\boldsymbol{A})) \leq \kappa_p(\boldsymbol{X})\,\|\boldsymbol{E}\|_p \ ,$$

where $\kappa_p(\boldsymbol{X}) = \|\boldsymbol{X}\|_p \left\|\boldsymbol{X}^{-1}\right\|_p$ *is the condition number of the matrix* $\boldsymbol{X}$ *in the* p*–norm.*

Proof. Let μ be an eigenvalue of matrix $\boldsymbol{A}+\boldsymbol{E}$. The case $\mu \in \Lambda(\boldsymbol{A})$ needs nothing to prove. Hence assume $\mu \in \Lambda(\boldsymbol{A}+\boldsymbol{E}) \setminus \Lambda(\boldsymbol{A})$, so that $\mu\boldsymbol{I} - \boldsymbol{A}$ is regular and $\mu\boldsymbol{I} - \boldsymbol{A} - \boldsymbol{E}$ is singular, as well as the matrix

$$(\mu\boldsymbol{I} - \boldsymbol{\Lambda})^{-1}\boldsymbol{X}^{-1}(\mu\boldsymbol{I} - \boldsymbol{A} - \boldsymbol{E})\,\boldsymbol{X} = \boldsymbol{I} - (\mu\boldsymbol{I} - \boldsymbol{\Lambda})^{-1}\boldsymbol{X}^{-1}\boldsymbol{E}\boldsymbol{X} \ .$$

According to the previous lemma the distance of this from $\boldsymbol{I}$ is at least one, so that

$$1 \leq \left\|(\mu\boldsymbol{I} - \boldsymbol{\Lambda})^{-1}\boldsymbol{X}^{-1}\boldsymbol{E}\boldsymbol{X}\right\|_p \leq \left\|(\mu\boldsymbol{I} - \boldsymbol{\Lambda})^{-1}\right\|_p \left\|\boldsymbol{X}^{-1}\right\|_p \|\boldsymbol{E}\|_p \|\boldsymbol{X}\|_p \ .$$

Since $\mu\boldsymbol{I} - \boldsymbol{\Lambda}$ is diagonal, we get

$$\left\|(\mu\boldsymbol{I} - \boldsymbol{\Lambda})^{-1}\right\|_p = \max_i |\mu - \lambda_i|^{-1} = \frac{1}{\min_i |\mu - \lambda_i|} \ ,$$

from which the claim follows. □

In particular, for a normal matrix $\boldsymbol{A}$ we get (an exercise):

$$\delta(\Lambda(\boldsymbol{A}+\boldsymbol{E}), \Lambda(\boldsymbol{A})) \leq \|\boldsymbol{E}\|_2 \ .$$

If the matrix is not diagonalizable, then the Schur decomposition gives:

[1] The set of matrices is finite dimensional, hence it is complete with any norm.

Theorem 6.5. *Let $\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{\Lambda} + \boldsymbol{N}$ be the Schur decomposition of $\boldsymbol{A}$, where $\boldsymbol{N}$ is a strictly upper triangular matrix, hence $|\boldsymbol{N}|^k = 0$ for some $k \le n$. Then*

$$\delta(\Lambda(\boldsymbol{A}+\boldsymbol{E}), \Lambda(\boldsymbol{A})) \le \max(\theta, \theta^{1/k}) \ ,$$

where $\theta = \|\boldsymbol{E}\|_2 \sum_{j=0}^{k-1} \|\boldsymbol{N}\|_2^k$.

Proof. If μ is an eigenvalue of matrix $\boldsymbol{A}+\boldsymbol{E}$ and $d = \delta(\mu, \Lambda(\boldsymbol{A})) > 0$, then as above $\mu\boldsymbol{I} - \boldsymbol{A}$ is invertible and $\boldsymbol{I} - (\mu\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{E}$ is singular, so that

$$1 \le \left\|(\mu\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{E}\right\|_2 \le \left\|(\mu\boldsymbol{I} - \boldsymbol{A})^{-1}\right\|_2 \|\boldsymbol{E}\|_2 = \left\|(\mu\boldsymbol{I} - \boldsymbol{\Lambda} - \boldsymbol{N})^{-1}\right\|_2 \|\boldsymbol{E}\|_2 \ .$$

Since $\boldsymbol{D} = (\mu\boldsymbol{I} - \boldsymbol{\Lambda})^{-1}$ is diagonal, $\boldsymbol{D}\boldsymbol{N}$ is strictly upper triangular and $(\boldsymbol{N}\boldsymbol{D})^k = 0$, so that

$$\begin{aligned}
&(\boldsymbol{D}^{-1} - \boldsymbol{N})(\boldsymbol{I} + \boldsymbol{D}\boldsymbol{N} + (\boldsymbol{D}\boldsymbol{N})^2 + \cdots + (\boldsymbol{D}\boldsymbol{N})^{k-1})\boldsymbol{D} = \\
&\quad =(\boldsymbol{D}^{-1} - \boldsymbol{N} + \boldsymbol{N} - \boldsymbol{N}\boldsymbol{D}\boldsymbol{N} + \boldsymbol{N}\boldsymbol{D}\boldsymbol{N} - \cdots + \boldsymbol{N}(\boldsymbol{D}\boldsymbol{N})^{k-2} - \boldsymbol{N}(\boldsymbol{D}\boldsymbol{N})^{k-1})\boldsymbol{D} = \\
&\quad =\boldsymbol{I} - (\boldsymbol{N}\boldsymbol{D})^k = \boldsymbol{I} \ .
\end{aligned}$$

Since $\|\boldsymbol{D}\|_2 = \frac{1}{d}$, this implies

$$\left\|(\mu\boldsymbol{I} - \boldsymbol{\Lambda} - \boldsymbol{N})^{-1}\right\|_2 = \left\|(\boldsymbol{D}^{-1} - \boldsymbol{N})^{-1}\right\|_2 \le \frac{1}{d}\sum_{j=0}^{k-1}\left(\frac{\|\boldsymbol{N}\|_2}{d}\right)^j \ .$$

Now, if $d \ge 1$, then $1 \le \|(\mu\boldsymbol{I} - \boldsymbol{\Lambda} - \boldsymbol{N})^{-1}\|_2 \|\boldsymbol{E}\|_2 \le \frac{\theta}{d}$, so that $d \le \theta$.
On the other hand, if $d < 1$, then $1 \le \|(\mu\boldsymbol{I} - \boldsymbol{\Lambda} - \boldsymbol{N})^{-1}\|_2 \|\boldsymbol{E}\|_2 \le \frac{\theta}{d^k}$, and $d \le \theta^{1/k}$. □

Finally, we will look how both eigenvalues and eigenvectors change when we perturb a matrix that has simple eigenvalues. For this we need the following.

Theorem 6.6. *Assume $\lambda_1, \ldots, \lambda_k$ are distinct eigenvalues of $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. Then the corresponding eigenvectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are linearly independent.*

Proof. Let $\alpha_1\boldsymbol{x}_1 + \alpha_2\boldsymbol{x}_2 + \cdots + \alpha_k\boldsymbol{x}_k = 0$. Multiplication of this by $\lambda_1\boldsymbol{I} - \boldsymbol{A}$ gives

$$\alpha_1(\lambda_1\boldsymbol{x}_1 - \boldsymbol{A}\boldsymbol{x}_1) + \alpha_2(\lambda_1\boldsymbol{x}_2 - \boldsymbol{A}\boldsymbol{x}_2) + \cdots + \alpha_k(\lambda_1\boldsymbol{x}_k - \boldsymbol{A}\boldsymbol{x}_k) = 0 \ ,$$

i.e.,

$$\alpha_2(\lambda_1 - \lambda_2)\boldsymbol{x}_2 + \cdots + \alpha_k(\lambda_1 - \lambda_k)\boldsymbol{x}_k = 0 \ ,$$

Similarly, after multiplication with $\lambda_2\boldsymbol{I} - \boldsymbol{A}$, ... , $\lambda_{k-1}\boldsymbol{I} - \boldsymbol{A}$, we have

$$\alpha_k(\lambda_1 - \lambda_k)(\lambda_2 - \lambda_k)\ldots(\lambda_{k-1} - \lambda_k)\boldsymbol{x}_k = 0 \ ,$$

i.e., $\alpha_k = 0$. Going backwards the intermediate equations we recursively get $\alpha_{k-1} = 0, \ldots, \alpha_2 = 0, \alpha_1 = 0$. □

Now we show that simple eigenvalues are smooth functions.

Theorem 6.7. *Let $\boldsymbol{A}$ have simple eigenvalues. Then the eigenvalues of $\boldsymbol{A} + z\,\boldsymbol{E}$ are analytic functions of z near $0 \in \mathbb{C}$, and the eigenvectors can also be taken to be analytic.*

Proof. Denote the eigenvalues by $\lambda_1, \ldots, \lambda_n$ and corresponding unit eigenvectors by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Set $\boldsymbol{X}_0 = [\boldsymbol{x}_1 \ \ldots \ \boldsymbol{x}_n]$, $\boldsymbol{\Lambda}_0 = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$. Then $\boldsymbol{X}_0$ is inverible. For $\boldsymbol{X}(z)$, $\boldsymbol{\Lambda}(z) \in \mathbb{C}^{n\times n}$, where $\boldsymbol{\Lambda}$ is diagonal, consider equation $\boldsymbol{F}(z, \boldsymbol{X}(z), \boldsymbol{\Lambda}(z)) = 0$, where[2]

$$\boldsymbol{F}(z, \boldsymbol{X}, \Lambda) = \begin{bmatrix} (\boldsymbol{A} + z\,\boldsymbol{E})\,\boldsymbol{X} - \boldsymbol{X}\Lambda \\ \mathrm{diag}(\boldsymbol{X}_0^*\boldsymbol{X} - \boldsymbol{I}) \end{bmatrix} .$$

Clearly $\boldsymbol{F}$ is analytic in $z, \boldsymbol{X}, \boldsymbol{\Lambda}$. We will use the implicit function theorem. For this we need to show that the derivative of $\boldsymbol{F}$ with respect to $(\boldsymbol{X}, \boldsymbol{\Lambda})$ at $(0, \boldsymbol{X}_0, \boldsymbol{\Lambda}_0)$ is invertible. Since it is a linear mapping from $\mathbb{C}^{(n+1)\times n}$ into itself, it suffices to show that it is an injection. So, assume $\boldsymbol{Y}, \boldsymbol{D} \in \mathbb{C}^{n\times n}$, $\boldsymbol{D}$ diagonal, are such that

$$D_{(\boldsymbol{X},\boldsymbol{\Lambda})}\boldsymbol{F}(0, \boldsymbol{X}_0, \boldsymbol{\Lambda}_0)(\boldsymbol{Y}, \boldsymbol{D}) = \begin{bmatrix} \boldsymbol{A}\boldsymbol{Y} - \boldsymbol{Y}\boldsymbol{\Lambda}_0 - \boldsymbol{X}_0\boldsymbol{D} \\ \mathrm{diag}(\boldsymbol{X}_0^*\boldsymbol{Y}) \end{bmatrix} = 0 .$$

Then, since $\boldsymbol{A} = \boldsymbol{X}_0\boldsymbol{\Lambda}_0\boldsymbol{X}_0^{-1}$, we have

$$\boldsymbol{\Lambda}_0\boldsymbol{X}_0^{-1}\boldsymbol{Y} - \boldsymbol{X}_0^{-1}\boldsymbol{Y}\boldsymbol{\Lambda}_0 - \boldsymbol{D} = 0 .$$

Denote $\boldsymbol{W} = \boldsymbol{X}_0^{-1}\boldsymbol{Y}$. Since the diagonal of $\boldsymbol{\Lambda}_0\boldsymbol{W} - \boldsymbol{W}\boldsymbol{\Lambda}_0$ is zero, we necessarily have $\boldsymbol{D} = 0$. Then $\boldsymbol{\Lambda}_0\boldsymbol{W} - \boldsymbol{W}\boldsymbol{\Lambda}_0 = 0$ implies $\lambda_i w_{i,j} - w_{i,j}\lambda_j = 0$, i.e., $w_{i,j} = 0$ for $i \neq j$. Finally,

$$\mathrm{diag}(\boldsymbol{X}_0^*\boldsymbol{Y}) = \mathrm{diag}(\boldsymbol{X}_0^*\boldsymbol{X}_0\boldsymbol{W}) = \begin{bmatrix} \|\boldsymbol{x}_1\|^2\, w_{1,1} & \ldots & \|\boldsymbol{x}_n\|^2\, w_{n,n} \end{bmatrix} = 0 ,$$

i.e., $\boldsymbol{W} = 0$ and, consequently, $\boldsymbol{Y} = 0$.
So, $D_{(\boldsymbol{X},\boldsymbol{\Lambda})}\boldsymbol{F}(0, \boldsymbol{X}_0, \boldsymbol{\Lambda}_0)$ is invertible and the equation $\boldsymbol{F}(z, \boldsymbol{X}(z), \boldsymbol{\Lambda}(z)) = 0$ defines $\boldsymbol{X}$ and $\boldsymbol{\Lambda}$ as analytic functions of z near zero. □

Let us compute the derivatives of $\boldsymbol{X}$ and $\boldsymbol{\Lambda}$. Differentiation gives:

$$\boldsymbol{A}\boldsymbol{X}'(0) + \boldsymbol{E}\boldsymbol{X}_0 - \boldsymbol{X}'(0)\boldsymbol{\Lambda}_0 - \boldsymbol{X}_0\boldsymbol{\Lambda}'(0) = 0 .$$

Denoting $\boldsymbol{Z} = \boldsymbol{X}_0^{-1}\boldsymbol{X}'(0)$ we get as in the proof above:

$$\boldsymbol{\Lambda}_0\boldsymbol{Z} - \boldsymbol{Z}\boldsymbol{\Lambda}_0 + \boldsymbol{X}_0^{-1}\boldsymbol{E}\boldsymbol{X}_0 - \boldsymbol{\Lambda}'(0) = 0 .$$

[2] Here $\mathrm{diag}(\boldsymbol{M}) = [m_{1,1}, m_{2,2}, \ldots, m_{n,n}]$.

Hence,

$$(\boldsymbol{\Lambda}'(0))_{i,i} = (\boldsymbol{X}_0^{-1}\boldsymbol{E}\boldsymbol{X}_0)_{i,i}$$

and

$$\boldsymbol{Z}_{i,j} = (\lambda_j - \lambda_i)^{-1}(\boldsymbol{X}_0^{-1}\boldsymbol{E}\boldsymbol{X}_0)_{i,j} \ ,$$

$$\boldsymbol{Z}_{i,i} = \tfrac{-1}{\|\boldsymbol{x}_i\|^2} \sum_{j\neq i} \langle \boldsymbol{x}_j, \boldsymbol{x}_i\rangle \, \boldsymbol{Z}_{j,i} \ .$$

From these, $\boldsymbol{X}'(0) = \boldsymbol{X}_0\boldsymbol{Z}$.

So: *sensitivity of the eigenvalues* is $\quad \sim \|\boldsymbol{X}^{-1}\| \ \|\boldsymbol{X}\|$

sensitivity of the eigenvectors is $\quad \sim \|\boldsymbol{X}^{-1}\| \ \|\boldsymbol{X}\|^2 \dfrac{1}{\min_{i,j} |\lambda_i - \lambda_j|}$.

Almost eigenvectors.

Let $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ be diagonalizable: $\boldsymbol{A} = \boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{X}^{-1}$. Let λ_1 be an eigenvalue, $\widehat{\lambda}$ close to it, and $\widehat{\boldsymbol{x}}$ a unit vector such that $\boldsymbol{r} = \widehat{\lambda}\,\widehat{\boldsymbol{x}} - \boldsymbol{A}\,\widehat{\boldsymbol{x}}$ is small. We want to estimate how close $\widehat{\boldsymbol{x}}$ is to the eigenspace corresponding to λ_1 .

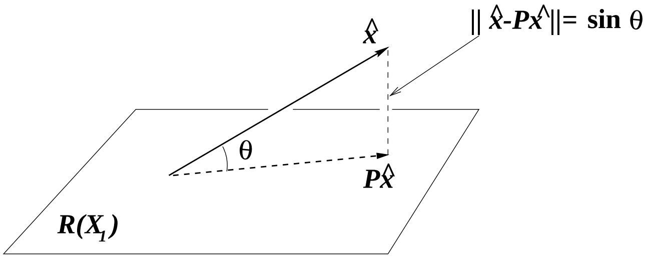
We may assume that $\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1\boldsymbol{I} & 0 \\ 0 & \boldsymbol{\Lambda}_2 \end{bmatrix}$, where λ_1 is not an eigenvalue of $\boldsymbol{\Lambda}_2$. Divide, correspondingly,

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \end{bmatrix} \ , \qquad \boldsymbol{X}^{-1} = \begin{bmatrix} \boldsymbol{Y}_1^* \\ \boldsymbol{Y}_2^* \end{bmatrix} \ .$$

The orthogonal projection onto $R(\boldsymbol{X}_1)$ — the eigenspace corresponding to λ_1 — is given by

$$\boldsymbol{P} = \boldsymbol{X}_1\boldsymbol{X}_1^\dagger = \boldsymbol{X}_1(\boldsymbol{X}_1^*\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^*$$

and the length of the orthogonal component is $\sin\theta$.



Lemma 6.8. $\boldsymbol{I} - \boldsymbol{P} = \boldsymbol{Y}_2\boldsymbol{Y}_2^\dagger$.

Proof. $\boldsymbol{I}-\boldsymbol{P}$ is the orthogonal projection onto $R(\boldsymbol{X}_1)^\perp$ and $\begin{bmatrix}\boldsymbol{Y}_1^*\\ \boldsymbol{Y}_2^*\end{bmatrix}\begin{bmatrix}\boldsymbol{X}_1 & \boldsymbol{X}_2\end{bmatrix}=\boldsymbol{I}$, so that $R(\boldsymbol{Y}_2)=R(\boldsymbol{X}_1)^\perp$. Hence, $\boldsymbol{I}-\boldsymbol{P}=\boldsymbol{Y}_2\boldsymbol{Y}_2^\dagger$. □

Recall that $\delta(\widehat{\lambda},\Lambda(\boldsymbol{\Lambda}_2))$ is the distance from $\widehat{\lambda}$ to the spectrum of $\boldsymbol{\Lambda}_2$.

Theorem 6.9. $\sin\theta \le \dfrac{\kappa_2(\boldsymbol{Y}_2)}{\delta(\widehat{\lambda},\Lambda(\boldsymbol{\Lambda}_2))}\,\|\boldsymbol{r}\|_2$, *where* $\kappa_2(\boldsymbol{Y}_2)=\|\boldsymbol{Y}_2\|_2\,\|\boldsymbol{Y}_2^\dagger\|_2\ge 1$.

Proof.

$$\begin{bmatrix}\boldsymbol{Y}_1^*\\ \boldsymbol{Y}_2^*\end{bmatrix}\boldsymbol{r}=\boldsymbol{X}^{-1}(\widehat{\lambda}\,\boldsymbol{I}-\boldsymbol{A})\,\widehat{\boldsymbol{x}}=\begin{bmatrix}(\widehat{\lambda}-\lambda_1)\,\boldsymbol{I} & 0\\ 0 & \widehat{\lambda}\boldsymbol{I}-\boldsymbol{\Lambda}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{Y}_1^*\\ \boldsymbol{Y}_2^*\end{bmatrix}\widehat{\boldsymbol{x}}\,.$$

Hence $\boldsymbol{Y}_2^*\widehat{\boldsymbol{x}}=(\widehat{\lambda}\boldsymbol{I}-\boldsymbol{\Lambda}_2)^{-1}\boldsymbol{Y}_2^*\boldsymbol{r}$ and, by the previous lemma,

$$\begin{aligned}\|(\boldsymbol{I}-\boldsymbol{P})\,\widehat{\boldsymbol{x}}\|_2&=\|\boldsymbol{Y}_2(\boldsymbol{Y}_2^*\boldsymbol{Y}_2)^{-1}(\widehat{\lambda}\boldsymbol{I}-\boldsymbol{\Lambda}_2)^{-1}\boldsymbol{Y}_2^*\boldsymbol{r}\|_2\\ &\le\|\boldsymbol{Y}_2(\boldsymbol{Y}_2^*\boldsymbol{Y}_2)^{-1}\|_2\|(\widehat{\lambda}\boldsymbol{I}-\boldsymbol{\Lambda}_2)^{-1}\|_2\,\|\boldsymbol{Y}_2^*\|_2\,\|\boldsymbol{r}\|_2\ ,\end{aligned}$$

which gives the result, since $\|(\widehat{\lambda}\boldsymbol{I}-\boldsymbol{\Lambda}_2)^{-1}\|=1/\delta(\widehat{\lambda},\Lambda(\boldsymbol{\Lambda}_2))$ and for any matrix $\|\boldsymbol{M}\|_2=\|\boldsymbol{M}^*\|_2$. □

Remark 6.2. If $\boldsymbol{Y}_2$ has orthonormal columns, then $\boldsymbol{Y}_2^\dagger=(\boldsymbol{Y}_2^*\boldsymbol{Y}_2)^{-1}\boldsymbol{Y}_2^*=\boldsymbol{Y}_2^*$ and $\|\boldsymbol{Y}_2\|=\|\boldsymbol{Y}_2^\dagger\|=1$, so that $\kappa_2(\boldsymbol{Y}_2)=1$. This is the case when $\boldsymbol{A}$ is normal.

Remark 6.3. If $\boldsymbol{A}$ is not diagonalizable but only the Jordan blocks corresponding to λ_1 are diagonal, then one gets a similar result, but $\delta(\widehat{\lambda},\Lambda(\boldsymbol{\Lambda}_2))$ is to be replaced by

$$\max\{\,\|(\widehat{\lambda}\boldsymbol{I}-\boldsymbol{J}(\lambda))^{-1}\|_2\,|\,\lambda\in\Lambda(\boldsymbol{A})\setminus\{\lambda_1\}\,\}\ ,$$

where $\boldsymbol{J}(\lambda)$ is the the largest Jordan block corresponding to λ. To get an estimate for this, see problem 5.18.

6.2. **Computation of eigenvalues, $\boldsymbol{QR}$–iteration.** Numerically it is not advisable to compute eigenvalues as the zeros of the characteristic polynomial since the roots of a polynomial are extremely sensitive functions of the coefficients[3]. The methods of practice are iterative.

[3] In fact, usually computation of the characteristic polynomial is done numerically via first computing the eigenvalues.

The simplest iteration for eigenvalues is the so called *power method*: pick an arbitrary $\boldsymbol{q}_0$ and iterate

```
for $k$ =1, 2, ...
        $\boldsymbol{z}_k = \boldsymbol{A}\boldsymbol{q}_{k-1}$
        $\boldsymbol{q}_k = \boldsymbol{z}_k / \|\boldsymbol{z}_k\|_2$
        $\mu_k = \langle \boldsymbol{A}\boldsymbol{q}_k, \boldsymbol{q}_k \rangle$
end
```

If λ is a simple eigenvalue of $\boldsymbol{A}$ such that others are less in absolute value, then $\boldsymbol{A}$ has the Jordan form:

$$\boldsymbol{J}_{\boldsymbol{A}} = \boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X} = \begin{bmatrix} \lambda & 0 \\ 0 & \boldsymbol{J} \end{bmatrix} ,$$

where $\mu \in \Lambda(\boldsymbol{J}) = \Lambda(\boldsymbol{A}) \setminus \{\lambda\} \implies |\mu| < |\lambda|$. Clearly: $\boldsymbol{q}_k = \alpha_k \boldsymbol{A}^k \boldsymbol{q}_0$, where $\alpha_k = \frac{1}{\|\boldsymbol{z}_1\|} \cdots \frac{1}{\|\boldsymbol{z}_k\|}$. If $\boldsymbol{v}^k = \boldsymbol{X}^{-1}\boldsymbol{q}_k$, then

$$\boldsymbol{v}^k = \alpha_k \lambda^k \begin{bmatrix} 1 & 0 \\ 0 & (\frac{1}{\lambda}\boldsymbol{J})^k \end{bmatrix} \boldsymbol{v}^0 .$$

Since $\lim_{k\to\infty}(\frac{1}{\lambda}\boldsymbol{J})^k = 0$, we see that, if $(\boldsymbol{v}^0)_1 \neq 0$, then

$$\langle \boldsymbol{A}\boldsymbol{q}_k, \boldsymbol{q}_k \rangle = \left\langle \boldsymbol{X}\boldsymbol{J}_{\boldsymbol{A}}\boldsymbol{v}^k, \boldsymbol{X}\boldsymbol{v}^k \right\rangle = (\boldsymbol{v}^k)^* \boldsymbol{X}^* \boldsymbol{X} \left(\lambda\, \boldsymbol{v}^k + \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{J} - \lambda \boldsymbol{I} \end{bmatrix} \boldsymbol{v}^k \right) ,$$

where $\lim_{k\to\infty} \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{J} - \lambda\, \boldsymbol{I} \end{bmatrix} \boldsymbol{v}^k = 0$, so that

$$\lim_{k\to\infty} \langle \boldsymbol{A}\boldsymbol{q}_k, \boldsymbol{q}_k \rangle = \lim_{k\to\infty} \lambda \left\| \boldsymbol{X}\boldsymbol{v}^k \right\|_2^2 = \lambda .$$

In practice the power method "converges"[4] to the eigenvector corresponding to the eigenvalue λ of largest absolute value. This happens with the speed η^k , where $\eta = \max_{\mu\in\Lambda(\boldsymbol{A})\setminus\{\lambda\}} |\mu| / |\lambda|$.

After the eigenvalue of largest absolute value has been found together with the corresponding eigenvectors of A and A^T , we can start looking for the next one by using the following.

Problem 6.6 (Deflation). Let $\lambda \neq 0$ be an eigenvalue of $\boldsymbol{A}$, $\boldsymbol{x} \neq 0$ a corresponding eigenvector, and $\boldsymbol{u}$ an eigenvector of $\boldsymbol{A}^T$ corresponding to λ scaled such that $\boldsymbol{u}^T\boldsymbol{x} = \lambda$. What is the connection between the eigenvalues and eigenvectors of matrices $\boldsymbol{A}$ and $\widetilde{\boldsymbol{A}} = \boldsymbol{A} - \boldsymbol{x}\,\boldsymbol{u}^T$? Hint: see problem 5.8.

[4] The vector does not really converge unless λ is real and nonnegative: it will tend to cycle $(\lambda/|\lambda|)^k \boldsymbol{q}$, where $\boldsymbol{q}$ is an eigenvector.

Often we want to compute an eigenpair $(\lambda_p, \boldsymbol{x}_p)$ of $\boldsymbol{A}$ such that λ_p is close to a given $\widehat{\lambda}$. Then we can use the power method with the matrix $(\boldsymbol{A} - \widehat{\lambda}\boldsymbol{I})^{-1}$. The eigenvalues of this matrix equal to $1/(\lambda_j - \widehat{\lambda})$, $j = 1, \ldots, n$, and the eigenvectors are the same. Thus we want that

$$|\lambda_p - \widehat{\lambda}| < |\lambda_i - \widehat{\lambda}| \,, \qquad i = 1, \ldots, n, i \neq p \,.$$

In practice the approximation $\widehat{\lambda}$ is also corrected at each step. This results to the following *inverse iteration*, which is quite powerful.

- Given $\widehat{\lambda}_0$, $\boldsymbol{u}_0$.
- Solve the system $(\boldsymbol{A} - \widehat{\lambda}_k\boldsymbol{I})\, \boldsymbol{w}_{k+1} = \boldsymbol{u}_k$.
- Set: $\boldsymbol{u}_{k+1} = \boldsymbol{w}_{k+1} / \|\boldsymbol{w}_{k+1}\|$, $\qquad \widehat{\lambda}_{k+1} = \langle \boldsymbol{A}\boldsymbol{u}_{k+1}, \boldsymbol{u}_{k+1} \rangle$.

Orthogonal iteration. The power method can be generalized to the following higher dimensional iteration. Let $\boldsymbol{Q}_0 \in \mathbb{C}^{n \times p}$ be such that $\boldsymbol{Q}_0^* \boldsymbol{Q}_0 = \boldsymbol{I}$. Iterate:

```
for $k = 1, 2, \ldots$
        $\boldsymbol{Z}_k = \boldsymbol{A}\,\boldsymbol{Q}_{k-1}$
        $\boldsymbol{Q}_k \boldsymbol{R}_k = \boldsymbol{Z}_k$        ($\boldsymbol{QR}$–decomposition of $\boldsymbol{Z}_k$)
end
```

Let the eigenvalues of $\boldsymbol{A}$ be ordered such that $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$ and let $\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U} = \boldsymbol{T} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) + \boldsymbol{N}$ and $\boldsymbol{V}^*\boldsymbol{A}^*\boldsymbol{V} = \widetilde{\boldsymbol{T}} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) + \widetilde{\boldsymbol{N}}$ be the Schur decompositions of $\boldsymbol{A}$ and $\boldsymbol{A}^*$. Then (Golub & van Loan, Theorem 7.3.1), if $|\lambda_p| > |\lambda_{p+1}|$ and $\|\boldsymbol{V}^*\boldsymbol{Q}_0\boldsymbol{Q}_0^*\boldsymbol{V} - \boldsymbol{P}_p\|_2 < 1$, where $\boldsymbol{P}_p = \begin{bmatrix} \boldsymbol{I}_p & 0 \\ 0 & 0 \end{bmatrix}$, then for every $\varepsilon > 0$ there exists a $C_\varepsilon > 0$ such that

$$\|\boldsymbol{U}^*\boldsymbol{Q}_k\boldsymbol{Q}_k^*\boldsymbol{U} - \boldsymbol{P}_p\|_2 < C_\varepsilon \left(\frac{|\lambda_{p+1}| + \varepsilon}{|\lambda_p| - \varepsilon} \right)^k \,,$$

i.e., the columns of $\boldsymbol{Q}_k$ span increasingly well the same subspace as the first p columns of the matrix $\boldsymbol{U}$. Moreover, a similar estimate holds for the distance $\delta(\Lambda(\boldsymbol{R}_k), \{\lambda_1, \ldots, \lambda_p\})$.

***QR*–iteration.** Consider more closely the case $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$, $\boldsymbol{Q}_k = [\boldsymbol{q}_1^k, \ldots, \boldsymbol{q}_n^k] \in \mathbb{C}^{n \times n}$. Then, assuming that $\left\|\boldsymbol{V}^*[\boldsymbol{q}_1^0, \ldots, \boldsymbol{q}_p^0][\boldsymbol{q}_1^0, \ldots, \boldsymbol{q}_p^0]^*\boldsymbol{V} - \boldsymbol{P}_p\right\|_2 < 1$, for all $p = 1, \ldots, n$, we get

$$\operatorname{span}(\boldsymbol{q}_1^k, \ldots, \boldsymbol{q}_p^k) \underset{k \to \infty}{\to} \operatorname{span}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p) \qquad \text{for all} \qquad p = 1, \ldots, n \,,$$

so that $\boldsymbol{T}_k = \boldsymbol{Q}_k^* \boldsymbol{A} \boldsymbol{Q}_k$ converges to an upper triangular matrix. Noticing that

$$\begin{aligned} \boldsymbol{T}_{k-1} &= \boldsymbol{Q}_{k-1}^* \boldsymbol{A} \boldsymbol{Q}_{k-1} = (\boldsymbol{Q}_{k-1}^* \boldsymbol{Q}_k) \boldsymbol{R}_k \\ \boldsymbol{T}_k &= \boldsymbol{Q}_k^* \boldsymbol{A} \boldsymbol{Q}_{k-1} \boldsymbol{Q}_{k-1}^* \boldsymbol{Q}_k = \boldsymbol{Q}_k^* \boldsymbol{Q}_k \boldsymbol{R}_k \boldsymbol{Q}_{k-1}^* \boldsymbol{Q}_k = \boldsymbol{R}_k (\boldsymbol{Q}_{k-1}^* \boldsymbol{Q}_k) \end{aligned}$$

and denoting $\widetilde{\boldsymbol{Q}}_k = \boldsymbol{Q}_{k-1}^* \boldsymbol{Q}_k$ we get:

$$\text{if} \quad \boldsymbol{T}_{k-1} = \widetilde{\boldsymbol{Q}}_k \boldsymbol{R}_k \,, \qquad \text{then} \qquad \boldsymbol{T}_k = \boldsymbol{R}_k \widetilde{\boldsymbol{Q}}_k \;.$$

in other words we get the iteration directly for matrices $\boldsymbol{T}_k$, where we form a $\boldsymbol{QR}$–decomposition of the previous one and the next we get, when these are multiplied in reverse order. This is called the $\boldsymbol{QR}$–iteration.

Problem 6.7. Consider the following $\boldsymbol{LU}$–iteration:

- Set $\boldsymbol{L}_0 \boldsymbol{U}_0 = \boldsymbol{A}$.
- Iterate: $\boldsymbol{L}_{k+1} \boldsymbol{U}_{k+1} = \boldsymbol{U}_k \boldsymbol{L}_k$, (new $\boldsymbol{LU}$–decomposition), $k = 0, 1, \dots$

Think what it means, if the matrices $\boldsymbol{S}_k = \boldsymbol{U}_k \boldsymbol{L}_k$ converge to an upper triangular matrix. When does this happen? Hint: consider $\widehat{\boldsymbol{L}}_k = \boldsymbol{L}_0 \boldsymbol{L}_1 \dots \boldsymbol{L}_k$ and show that $\widehat{\boldsymbol{L}}_k \boldsymbol{U}_k = \boldsymbol{A} \widehat{\boldsymbol{L}}_{k-1}$.

$\boldsymbol{QR}$–iteration in the Hessenberg form.

Comparing the orthogonal iteration and the $\boldsymbol{QR}$–iteration above, it seems that both require the same amount of floating point operations per step: one $\boldsymbol{QR}$–decomposition and one matrix multiplication, although in the $\boldsymbol{QR}$–iteration the multiplication includes a triangular matrix, so that it is slightly cheaper. $\boldsymbol{QR}$–iteration becomes much more economical, when the matrixes $\boldsymbol{T}_k$ are of the Hessenberg form:

1° Choose $\boldsymbol{Q}_0$ such that $\boldsymbol{H}_0 = \boldsymbol{Q}_0^* \boldsymbol{A} \boldsymbol{Q}_0$ is a Hessenberg matrix (Problem 4.8).

2° For $k = 1, 2, \dots$, we get the $\boldsymbol{QR}$–decomposition of the Hessenberg matrix $\boldsymbol{H}_{k-1}$ easily using the Givens rotations:

$$\begin{aligned}
&\boldsymbol{G}_{n-1,n} \dots \boldsymbol{G}_{2,3}\boldsymbol{G}_{1,2}\boldsymbol{H}_{k-1} \\
&= \boldsymbol{G}_{n-1,n} \dots \boldsymbol{G}_{2,3}
\begin{bmatrix}
\# & \# & \# & \dots & \# & \# & \# \\
0 & \# & \# & \dots & \# & \# & \# \\
0 & \# & \# & \dots & \# & \# & \# \\
0 & 0 & \# & \dots & \# & \# & \# \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \dots & \# & \# & \# \\
0 & 0 & 0 & \dots & 0 & \# & \#
\end{bmatrix} \\
&= \boldsymbol{G}_{n-1,n} \dots \boldsymbol{G}_{3,4}
\begin{bmatrix}
\# & \# & \# & \dots & \# & \# & \# \\
0 & \# & \# & \dots & \# & \# & \# \\
0 & 0 & \# & \dots & \# & \# & \# \\
0 & 0 & \# & \dots & \# & \# & \# \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \dots & \# & \# & \# \\
0 & 0 & 0 & \dots & 0 & \# & \#
\end{bmatrix} \\
&\dots = \boldsymbol{G}_{n-1,n}
\begin{bmatrix}
\# & \# & \# & \dots & \# & \# & \# \\
0 & \# & \# & \dots & \# & \# & \# \\
0 & 0 & \# & \dots & \# & \# & \# \\
0 & 0 & 0 & \ddots & \# & \# & \# \\
\vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \dots & 0 & \# & \# \\
0 & 0 & 0 & \dots & 0 & \# & \#
\end{bmatrix} \\
&=
\begin{bmatrix}
\# & \# & \# & \dots & \# & \# & \# \\
0 & \# & \# & \dots & \# & \# & \# \\
0 & 0 & \# & \dots & \# & \# & \# \\
0 & 0 & 0 & \ddots & \# & \# & \# \\
\vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \dots & 0 & \# & \# \\
0 & 0 & 0 & \dots & 0 & 0 & \#
\end{bmatrix} = \boldsymbol{R}_k \ ,
\end{aligned}$$

i.e., $\boldsymbol{H}_{k-1} = \boldsymbol{G}^*_{1,2} \dots \boldsymbol{G}^*_{n-1,n}\boldsymbol{R}_k$. This has the flop count $\approx 3n^2$.

3° We get $\boldsymbol{H}_k = \boldsymbol{R}_k\boldsymbol{G}^*_{1,2} \dots \boldsymbol{G}^*_{n-1,n}$ by applying the same Givens rotations now from the right. The result is again a Hessenberg matrix and work is also $\approx 3n^2$. Repeat from 2°.

The flop count of Hessenberg $\boldsymbol{QR}$–algorithm is only $6n^2$ per iteration.

Shifted $\boldsymbol{QR}$–iteration.

If we replace the matrix $\boldsymbol{A}$ with $\boldsymbol{A} - \mu\boldsymbol{I}$, in the $\boldsymbol{QR}$–iteration (in Hessenberg form), then the p:th subdiagonal element of the matrixes $\boldsymbol{H}_k$ converges to zero approximately with speed

$$\frac{|\lambda_{p+1} - \mu|^k}{|\lambda_p - \mu|^k} \ .$$

The shifted $\boldsymbol{QR}$–iteration tries to find a suitable μ and updates matrices $\boldsymbol{H}_k$, that are kept similar to $\boldsymbol{A}$ (and not to $\boldsymbol{A} - \mu\boldsymbol{I}$). The basic strategy is to use the

Hessenberg $\boldsymbol{QR}$–algorithm and choose μ to be the current element $h_{n,n}$:

```
$\boldsymbol{H}_0 = \boldsymbol{Q}_0^* \boldsymbol{A} \boldsymbol{Q}_0$        into the Hessenberg form
for $k =$ $1, 2, \ldots$
        $\mu = (\boldsymbol{H}_{k-1})_{n,n}$
        $\boldsymbol{Q}_k \boldsymbol{R}_k = \boldsymbol{H}_{k-1} - \mu \boldsymbol{I}$        (the $\boldsymbol{QR}$–decomposition of $\boldsymbol{H}_{k-1}$ using Givens)
        $\boldsymbol{H}_k = \boldsymbol{R}_k \boldsymbol{Q}_k + \mu \boldsymbol{I}$
end
```

The following shows the convergence speed of this. If

$$\boldsymbol{G}_{n-2,n-1} \ldots \boldsymbol{G}_{1,2}(\boldsymbol{H}_{k-1} - \mu \boldsymbol{I}) = \begin{bmatrix} \widetilde{\boldsymbol{R}}_k & \boldsymbol{X} \\ 0 & \begin{matrix} a & b \\ \varepsilon & 0 \end{matrix} \end{bmatrix}, \qquad \text{then} \qquad \boldsymbol{G}_{n-1,n} = \begin{bmatrix} \boldsymbol{I} & 0 \\ 0 & \begin{matrix} \bar\alpha & \bar\beta \\ -\beta & \alpha \end{matrix} \end{bmatrix},$$

where $\alpha = a/m$, $\beta = \varepsilon/m$, $m = \sqrt{|a|^2 + |\varepsilon|^2}$. Hence the lower–right $2{\times}2$ corner of $\boldsymbol{H}_k$ becomes

$$\begin{bmatrix} \bar\alpha & \bar\beta \\ -\beta & \alpha \end{bmatrix} \begin{bmatrix} a & b \\ \varepsilon & 0 \end{bmatrix} \begin{bmatrix} \alpha & -\bar\beta \\ \beta & \bar\alpha \end{bmatrix} = \begin{bmatrix} a\,|\alpha|^2 + b\bar\alpha\beta + \varepsilon\alpha\bar\beta & b\bar\alpha^2 - a\bar\alpha\bar\beta - \varepsilon\bar\beta^2 \\ -a\alpha\beta - b\beta^2 + \varepsilon\alpha^2 & a\,|\beta|^2 - b\bar\alpha\beta - \varepsilon\alpha\bar\beta \end{bmatrix} .$$

The $n,n-1$ element of $\boldsymbol{H}_k$ is then

$$(\boldsymbol{H}_k)_{n,n-1} = (-a^2\varepsilon - b\varepsilon^2 + \varepsilon a^2)/m^2 = -\frac{b\,\varepsilon^2}{|a|^2 + |\varepsilon|^2} .$$

If $|\varepsilon| \ll |a|$, then the corresponding element of $\boldsymbol{H}_k$ is already a very small number and if $(\boldsymbol{H}_k)_{n,n-1} \approx 0$, then $(\boldsymbol{H}_k)_{n,n}$ is very close to an eigenvalue of $\boldsymbol{A}$. The next ones we find by continuing similarly with the $(n-1){\times}(n-1)$ Hessenberg matrix $\boldsymbol{H}_k(1:n-1, 1:n-1)$.

Problem 6.8. Test the Hessenberg $\boldsymbol{QR}$–iteration with MATLAB:

```
H=hess(A), m=something;
for k=1:m, [Q,R]=qr(H); H=R*Q; end
```

(if $\boldsymbol{H}$ is not in the real Schur form, then iterate more). Test also the shifted $\boldsymbol{QR}$:

```
H=hess(A); nh=n;
for l=1:n-1
     while abs(H(nh,nh-1))>0.000001, mu=H(nh,nh);
          [Q,R]=qr(H-mu*eye(nh));
          H=R*Q+mu*eye(nh); iter=[l,H(nh,nh)], end
     H, nh=nh-1; H=H(1:nh,1:nh);
end
```

Try first the following matrices:

```
A=randn(4)+i*randn(4);   A=randn(7);   B=randn(7); A=B'*B;
A=[0,1,0,0,0;0,0,1,0,0;0,0,0,1,0;0,0,0,0,1;1,0,0,0,0];
```

Think what causes the difficulty in the last case and how to "wake up" the convergence.

If also the eigenvectors are desired, one can compute those first for the resulting upper triangular matrix, and from them for $\boldsymbol{A}$, but then we should keep track of the similarity transformation that, that brings $\boldsymbol{A}$ to $\boldsymbol{H}_k$. This increases considerably the work load of the iteration. A better way is to first compute (using shifted $\boldsymbol{QR}$) only the approximations $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_n$ of the eigenvalues and then apply the inverse iteration to improve the eigenvalues and compute the eigenvectors.

Correction with inverse iteration.

After the approximations $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_n$ of the eigenvalues of $\boldsymbol{A}$ have been found, their accuracy can be improved and the eigenvectors found by the following iteration:

```
for $j = 1, \ldots, n$
    $\mu_0 = \widehat{\lambda}_j$, $\boldsymbol{q}_0 \perp \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{j-1}\}$
    for $k = 1, 2, \ldots$
        $(\boldsymbol{A} - \mu_{k-1}\boldsymbol{I})\,\boldsymbol{z}_k = \boldsymbol{q}_{k-1}$
        $\boldsymbol{q}_k = \boldsymbol{z}_k / \|\boldsymbol{z}_k\|_2$
        $\mu_k = (\boldsymbol{q}_k)^* \boldsymbol{A} \boldsymbol{q}_k$
    end
    $\boldsymbol{x}_j = \boldsymbol{q}_k$, $\lambda_j = \mu_k$
end
```

For each j this converges as $\dfrac{|\lambda_j - \widehat{\lambda}_j|^k}{\delta(\widehat{\lambda}_j, \Lambda(\boldsymbol{A}) \setminus \{\lambda_j\})^k}$. An interesting phenomenon in the inverse iteration is that when $\mu \approx \lambda_j$, then $\boldsymbol{A} - \widehat{\lambda}_j \boldsymbol{I}$ is close to a singular matrix and numerical solution of the system $(\boldsymbol{A} - \mu\boldsymbol{I})\boldsymbol{z}_k = \boldsymbol{q}_{k-1}$ produces large errors. It turns out, however, that these errors tend to be in the direction of the eigenvector corresponding to the eigenvalue $\lambda_j - \widehat{\lambda}_j$, so that even the errors improve $\boldsymbol{z}^k$ as an approximative eigenvector.

Problem 6.9. Continue problem 6.8 by computing some of the eigenvectors using inverse iteration. Check your computations with the `Matlab` command `eig`.

7. Eigenvalues of a Hermitian matrix, singular value decomposition

7.1. **Eigenvalues of a Hermitian matrix.** If $\boldsymbol{A}$ is Hermitian: $\boldsymbol{A}^* = \boldsymbol{A}$, then it is normal and hence unitarily similar to a diagonal matrix

$$\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \ .$$

Then:

$$\operatorname{diag}(\bar{\lambda}_1, \ldots, \bar{\lambda}_n) = \boldsymbol{Q}^*\boldsymbol{A}^*\boldsymbol{Q} = \boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \ ,$$

so that $\bar{\lambda}_j = \lambda_j$, $j = 1, \ldots, n$, i.e., the eigenvalues are real. We assume, that they are numbered in the nondecreasing order: $\lambda_1 \le \cdots \le \lambda_n$.

The next theorem gives the largest and smallest eigenvalues of a Hermitian matrix as solutions to extremal problems.

Theorem 7.1 (Rayleigh-Ritz). *Let $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. Then the extremal values of the Rayleigh–Ritz quotient:*

$$\mu_1 = \min_{\boldsymbol{x}\neq 0} \operatorname{Re} \frac{\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}} \qquad \textit{and} \qquad \mu_n = \max_{\boldsymbol{x}\neq 0} \operatorname{Re} \frac{\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}}$$

are the smallest and largest eigenvalues of the matrix $\frac{1}{2}(\boldsymbol{A} + \boldsymbol{A}^*)$.

Proof. Matrix $\frac{1}{2}(\boldsymbol{A} + \boldsymbol{A}^*)$ is Hermitian, so that there exists a unitary U such that $U^*\frac{1}{2}(\boldsymbol{A} + \boldsymbol{A}^*)U = D = \operatorname{diag}(d_1, \ldots, d_n)$, where $d_1 \le \cdots \le d_n$. We get:

$$\begin{aligned}\mu_1 &= \min_{\boldsymbol{x}\neq 0} \operatorname{Re} \frac{\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}} = \min_{\|\boldsymbol{x}\|_2=1} \operatorname{Re} \boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} = \min_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^*\tfrac{1}{2}(\boldsymbol{A} + \boldsymbol{A}^*)\boldsymbol{x} \\ &= \min_{\|\boldsymbol{U}\boldsymbol{x}\|_2=1} \boldsymbol{x}^*\boldsymbol{U}^*\tfrac{1}{2}(\boldsymbol{A} + \boldsymbol{A}^*)\boldsymbol{U}\boldsymbol{x} = \min_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^*\boldsymbol{D}\boldsymbol{x} = d_1 \ .\end{aligned}$$

Similarly, $\mu_n = d_n$. □

In particular, when $\boldsymbol{A}$ is Hermitian, then μ_1 and μ_n are the smallest and largest eigenvalues of $\boldsymbol{A}$.

Corollary 7.2. *If $\boldsymbol{A}$ is Hermitian, $0 \neq \boldsymbol{x} \in \mathbb{C}^n$ and $\alpha = \frac{\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}}$, then $\boldsymbol{A}$ has an eigenvalue $\le \alpha$ and an eigenvalue $\ge \alpha$.*

Remark 7.1. Let $(\lambda, \boldsymbol{x})$ be an eigenvalue–eigenvector pair of Hermitian matrix $\boldsymbol{A}$. Then

$$\begin{aligned}\frac{(\boldsymbol{x}+\boldsymbol{v})^*\boldsymbol{A}(\boldsymbol{x}+\boldsymbol{v})}{(\boldsymbol{x}+\boldsymbol{v})^*(\boldsymbol{x}+\boldsymbol{v})} &= \frac{\lambda\boldsymbol{x}^*\boldsymbol{x} + \lambda\boldsymbol{v}^*\boldsymbol{x} + \lambda\boldsymbol{x}^*\boldsymbol{v} + \boldsymbol{v}^*\boldsymbol{A}\boldsymbol{v}}{\boldsymbol{x}^*\boldsymbol{x} + \boldsymbol{v}^*\boldsymbol{x} + \boldsymbol{x}^*\boldsymbol{v} + \boldsymbol{v}^*\boldsymbol{v}} \\ &= \lambda + \frac{\boldsymbol{v}^*(\boldsymbol{A} - \lambda\boldsymbol{I})\boldsymbol{v}}{\|\boldsymbol{x}+\boldsymbol{v}\|^2} = \lambda + O(\|\boldsymbol{v}\|^2) \ .\end{aligned}$$

[0] Version: April 9, 2003

This means that if a vector is ε–close to an eigenvector, then the Rayleigh–Ritz quotient gives the eigenvalue with accuracy $O(\varepsilon^2)$.

On the other hand, if for some $\boldsymbol{u} \neq 0$ we have $\frac{(\boldsymbol{u}+\boldsymbol{v})^*\boldsymbol{A}(\boldsymbol{u}+\boldsymbol{v})}{(\boldsymbol{u}+\boldsymbol{v})^*(\boldsymbol{u}+\boldsymbol{v})} = \alpha + O(\|\boldsymbol{v}\|^2)$, then for every $\boldsymbol{v} \perp \boldsymbol{u}$ we get

$$(\boldsymbol{u}+\boldsymbol{v})^*\boldsymbol{A}(\boldsymbol{u}+\boldsymbol{v}) = \alpha \|\boldsymbol{u}\|^2 + O(\|\boldsymbol{v}\|^2) ,$$

i.e., $\langle \boldsymbol{Au}, \boldsymbol{v}\rangle + \langle \boldsymbol{v}, \boldsymbol{Au}\rangle = 0$. Take $\boldsymbol{v} = \boldsymbol{Au} - \frac{\langle \boldsymbol{Au},\boldsymbol{u}\rangle}{\langle \boldsymbol{u},\boldsymbol{u}\rangle}\boldsymbol{u}$, which clearly is orthogonal to $\boldsymbol{u}$. Then

$$\left\langle \boldsymbol{Au}, \boldsymbol{Au} - \tfrac{\langle \boldsymbol{Au},\boldsymbol{u}\rangle}{\langle \boldsymbol{u},\boldsymbol{u}\rangle}\boldsymbol{u}\right\rangle + \left\langle \boldsymbol{Au} - \tfrac{\langle \boldsymbol{Au},\boldsymbol{u}\rangle}{\langle \boldsymbol{u},\boldsymbol{u}\rangle}\boldsymbol{u}, \boldsymbol{Au}\right\rangle = 0 ,$$

i.e., $2\|\boldsymbol{Au}\|^2 \|\boldsymbol{u}\|^2 = 2\langle \boldsymbol{Au}, \boldsymbol{u}\rangle^2$. Hence the Cauchy inequality for $\boldsymbol{Au}$ and $\boldsymbol{u}$ is an equality, which is possible only if $\boldsymbol{Au} = \lambda \boldsymbol{u}$ for some λ.

We have shown that the eigenvectors and eigenvalues of a Hermitian matrix $\boldsymbol{A}$ are, respectively, the critical points and the critical values of the real valued function $\boldsymbol{x} \to \frac{\boldsymbol{x}^*\boldsymbol{Ax}}{\boldsymbol{x}^*\boldsymbol{x}}$ (i.e., points where the derivative vanishes and the values at these points).

Matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is *positive definite*, if $\boldsymbol{x}^*\boldsymbol{Ax} > 0$ for all $0 \neq \boldsymbol{x} \in \mathbb{C}^n$ and positive *semidefinite* if $\boldsymbol{x}^*\boldsymbol{Ax} \geq 0$ for all $\boldsymbol{x} \in \mathbb{C}^n$. By Theorem 7.1 Hermitian $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is positive definite if and only if all its eigenvalues are positive.

Theorem 7.1 gives the smallest and largest eigenvalues of a Hermitian matrix as the extremal values of the Rayleigh–Ritz quotient and the remark shows that all eigenvalues are the critical values of it. Next we give an extremal property of all the eigenvalues.

Let $\mathcal{U}_{n,k}$ be the set of $n\times k$ matrices that have orthonormal columns.

Theorem 7.3 (Courant–Fisher min max theorem). *Let $\lambda_1 \leq \cdots \leq \lambda_n$ be the eigenvalues of a Hermitian matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$. Then for all $k = 1, \ldots, n$*

$$\begin{aligned}\lambda_k &= \min_{\boldsymbol{U}\in\mathcal{U}_{n,k}} \max_{0\neq\boldsymbol{v}\in\mathbb{C}^k} \frac{\boldsymbol{v}^*\boldsymbol{U}^*\boldsymbol{AUv}}{\boldsymbol{v}^*\boldsymbol{v}} \\ &= \max_{\boldsymbol{U}\in\mathcal{U}_{n,n-k+1}} \min_{0\neq\boldsymbol{v}\in\mathbb{C}^{n-k+1}} \frac{\boldsymbol{v}^*\boldsymbol{U}^*\boldsymbol{AUv}}{\boldsymbol{v}^*\boldsymbol{v}} .\end{aligned}$$

Proof. Let

$$\boldsymbol{Q}^*\boldsymbol{AQ} = \boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) .$$

Now $\boldsymbol{U} \in \mathcal{U}_{n,k} \iff \boldsymbol{Q}^*\boldsymbol{U} \in \mathcal{U}_{n,k}$, so that

$$\min_{\boldsymbol{U}\in\mathcal{U}_{n,k}} \max_{0\neq\boldsymbol{v}\in\mathbb{C}^k} \frac{\boldsymbol{v}^*\boldsymbol{U}^*\boldsymbol{AUv}}{\boldsymbol{v}^*\boldsymbol{v}} = \min_{\boldsymbol{U}\in\mathcal{U}_{n,k}} \max_{\|\boldsymbol{v}\|_2=1} \boldsymbol{v}^*\boldsymbol{U}^*\boldsymbol{Q\Lambda Q}^*\boldsymbol{Uv} = \min_{\boldsymbol{U}\in\mathcal{U}_{n,k}} \max_{\|\boldsymbol{v}\|_2=1} \boldsymbol{v}^*\boldsymbol{U}^*\boldsymbol{\Lambda Uv} .$$

If $\boldsymbol{U} \in \mathcal{U}_{n,k}$, then $\dim R(\boldsymbol{U}) = k$ and, since $\dim\{\boldsymbol{y} \,|\, y_1 = \cdots = y_{k-1} = 0\} = n-k+1$, there exists $\widetilde{\boldsymbol{v}} \in \mathbb{C}^k$, $\|\boldsymbol{v}\|_2 = 1$ such that vector $\widetilde{\boldsymbol{y}} = \boldsymbol{U}\widetilde{\boldsymbol{v}}$ satisfies: $\widetilde{y}_1 = \cdots = \widetilde{y}_{k-1} = 0$. Further, $\|\widetilde{\boldsymbol{y}}\|_2^2 = \widetilde{\boldsymbol{v}}^*\boldsymbol{U}^*\boldsymbol{U}\widetilde{\boldsymbol{v}} = 1$. Hence

$$\max_{\|\boldsymbol{v}\|_2=1} \boldsymbol{v}^*\boldsymbol{U}^*\boldsymbol{\Lambda}\boldsymbol{U}\boldsymbol{v} \geq \widetilde{\boldsymbol{v}}^*\boldsymbol{U}^*\boldsymbol{\Lambda}\boldsymbol{U}\widetilde{\boldsymbol{v}} = \widetilde{\boldsymbol{y}}^*\boldsymbol{\Lambda}\widetilde{\boldsymbol{y}} = \sum_{j=k}^{n} \lambda_j \left|\widetilde{y}_j\right|^2 \geq \lambda_k \ .$$

We obtained: for all $\boldsymbol{U} \in \mathcal{U}_{n,k}$ holds $\max_{\|\boldsymbol{v}\|_2=1} \boldsymbol{v}^*\boldsymbol{U}^*\boldsymbol{\Lambda}\boldsymbol{U}\boldsymbol{v} \geq \lambda_k$. On the other hand, if in the place of $\boldsymbol{U}$ we take the matrix $\boldsymbol{U}_0 = \begin{bmatrix}\boldsymbol{I}_k \\ 0\end{bmatrix}$, we get $\boldsymbol{U}_0^*\boldsymbol{\Lambda}\boldsymbol{U}_0 = \operatorname{diag}(\lambda_1, \ldots, \lambda_k)$, so that

$$\max_{\|\boldsymbol{v}\|_2=1} \boldsymbol{v}^*\boldsymbol{U}_0^*\boldsymbol{\Lambda}\boldsymbol{U}_0\boldsymbol{v} = \lambda_k \ .$$

This proves the first equation.

The second is obtained by noticing that that $\lambda_k(\boldsymbol{A}) = -\lambda_{n-k+1}(-\boldsymbol{A})$. Then by the first equation

$$\begin{aligned}\lambda_k &= -\min_{\boldsymbol{U}\in\mathcal{U}_{n,n-k+1}} \max_{0\neq\boldsymbol{v}\in\mathbb{C}^{n-k+1}} \frac{-\boldsymbol{v}^*\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}\boldsymbol{v}}{\boldsymbol{v}^*\boldsymbol{v}} \\ &= \max_{\boldsymbol{U}\in\mathcal{U}_{n,n-k+1}} \min_{0\neq\boldsymbol{v}\in\mathbb{C}^{n-k+1}} \frac{\boldsymbol{v}^*\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}\boldsymbol{v}}{\boldsymbol{v}^*\boldsymbol{v}} \ .\end{aligned}$$

□

Problem 7.1. Show, that the Courant–Fisher theorem can be rewritten in the form:

$$\lambda_k = \min_{\dim(V)=k} \max_{0\neq\boldsymbol{x}\in V} \frac{\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}} = \max_{\dim(V)=n-k+1} \min_{0\neq\boldsymbol{x}\in V} \frac{\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}} \ .$$

Here the first minimization (resp. maximization) is with respect to all subspaces of dimension k (resp. $n-k+1$).

The following theorem gives good bounds for the eigenvalues of a Hermitian perturbation of a Hermitian matrix.

Theorem 7.4 (Weyl)**.** *If $\boldsymbol{A}$, $\boldsymbol{B} \in \mathbb{C}^{n\times n}$ are Hermitian, then for all $k = 1, \ldots, n$ holds:*

$$\lambda_k(\boldsymbol{A}) + \lambda_1(\boldsymbol{B}) \leq \lambda_k(\boldsymbol{A}+\boldsymbol{B}) \leq \lambda_k(\boldsymbol{A}) + \lambda_n(\boldsymbol{B}) \ .$$

Proof. If $\|\boldsymbol{x}\|_2 = 1$, then $\lambda_1(\boldsymbol{B}) \leq \boldsymbol{x}^*\boldsymbol{B}\boldsymbol{x} \leq \lambda_n(\boldsymbol{B})$. Since $\boldsymbol{U} \in \mathcal{U}_{n,k} \implies \|\boldsymbol{U}\boldsymbol{v}\|_2 = \|\boldsymbol{v}\|_2$, we get

$$\lambda_k(\boldsymbol{A}{+}\boldsymbol{B}) = \min_{\boldsymbol{U}\in\mathcal{U}_{n,k}} \max_{\|\boldsymbol{v}\|_2=1} \boldsymbol{v}^*\boldsymbol{U}^*(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{U}\boldsymbol{v}$$

$$\begin{cases} \geq \min\limits_{\boldsymbol{U}\in\mathcal{U}_{n,k}} \max\limits_{\|\boldsymbol{v}\|_2=1} (\boldsymbol{v}^*\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}\boldsymbol{v} + \lambda_1(\boldsymbol{B})) = \lambda_k(\boldsymbol{A}) + \lambda_1(\boldsymbol{B}) \\ \leq \min\limits_{\boldsymbol{U}\in\mathcal{U}_{n,k}} \max\limits_{\|\boldsymbol{v}\|_2=1} (\boldsymbol{v}^*\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}\boldsymbol{v} + \lambda_n(\boldsymbol{B})) = \lambda_k(\boldsymbol{A}) + \lambda_n(\boldsymbol{B}) \ . \end{cases}$$

□

In the case of rank–one perturbation of $\boldsymbol{A}$ we get a sharper result.

Theorem 7.5. *Let* $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *be Hermitian,* $\boldsymbol{a} \in \mathbb{C}^n$ *and* $\widetilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{a}\boldsymbol{a}^*$. *Then*

$$\lambda_1(\boldsymbol{A}) \leq \lambda_1(\widetilde{\boldsymbol{A}}) \leq \lambda_2(\boldsymbol{A}) \leq \lambda_2(\widetilde{\boldsymbol{A}}) \leq \ldots \lambda_n(\boldsymbol{A}) \leq \lambda_n(\widetilde{\boldsymbol{A}}) \ .$$

Proof. Since $\boldsymbol{v}^*\boldsymbol{U}^*\boldsymbol{a}\boldsymbol{a}^*\boldsymbol{U}\boldsymbol{v} = |\boldsymbol{a}^*\boldsymbol{U}\boldsymbol{v}|^2$, we get:

$$\lambda_k(\widetilde{\boldsymbol{A}}) = \min_{\boldsymbol{U}\in\mathcal{U}_{n,k}} \max_{\|\boldsymbol{v}\|_2=1} \boldsymbol{v}^*\boldsymbol{U}^*(\boldsymbol{A}+\boldsymbol{a}\boldsymbol{a}^*)\boldsymbol{U}\boldsymbol{v} \geq \min_{\boldsymbol{U}\in\mathcal{U}_{n,k}} \max_{\|\boldsymbol{v}\|_2=1} \boldsymbol{v}^*\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}\boldsymbol{v} = \lambda_k(\boldsymbol{A}) \ .$$

In the other direction we use the form 7.1 of the Courant–Fisher theorem:

$$\begin{aligned} \lambda_{k+1}(\boldsymbol{A}) &= \max_{\dim(V)=n-k} \min_{0\neq\boldsymbol{x}\in V} \frac{\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}} \geq \max_{\substack{\dim(V)=n-k \\ V\perp\boldsymbol{a}}} \min_{0\neq\boldsymbol{x}\in V} \frac{\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}} \\ &= \max_{\dim(V)=n-k+1} \min_{\substack{0\neq\boldsymbol{x}\in V \\ \boldsymbol{x}\perp\boldsymbol{a}}} \frac{\boldsymbol{x}^*(\boldsymbol{A}+\boldsymbol{a}\boldsymbol{a}^*)\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}} \geq \max_{\dim(V)=n-k+1} \min_{0\neq\boldsymbol{x}\in V} \frac{\boldsymbol{x}^*(\boldsymbol{A}+\boldsymbol{a}\boldsymbol{a}^*)\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}} \\ &= \lambda_k(\widetilde{\boldsymbol{A}}) . \end{aligned}$$

□

Another application of the Courant–Fisher theorem gives inequalities between eigenvalues of a Hermitian matrix and its submatrix:

Theorem 7.6. *If* $\widehat{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{b} \\ \boldsymbol{b}^* & c \end{bmatrix} \in \mathbb{C}^{(n+1)\times(n+1)}$ *is Hermitian, then*

$$\lambda_1(\widehat{\boldsymbol{A}}) \leq \lambda_1(\boldsymbol{A}) \leq \lambda_2(\widehat{\boldsymbol{A}}) \leq \ldots \lambda_n(\boldsymbol{A}) \leq \lambda_{n+1}(\widehat{\boldsymbol{A}}) \ .$$

Proof.

$$\begin{aligned}
\lambda_k(\widehat{\boldsymbol{A}}) &= \min_{\boldsymbol{U}\in\mathcal{U}_{n+1,k}} \max_{0\neq v\in\mathbb{C}^k} \frac{\boldsymbol{v}^*\boldsymbol{U}^*\widehat{\boldsymbol{A}}\boldsymbol{U}\boldsymbol{v}}{\boldsymbol{v}^*\boldsymbol{v}} \\
&\le \min_{\boldsymbol{U}\in\mathcal{U}_{n,k}} \max_{0\neq v\in\mathbb{C}^k} \frac{\boldsymbol{v}^*\begin{bmatrix}\boldsymbol{U}^* & 0\end{bmatrix}\widehat{\boldsymbol{A}}\begin{bmatrix}\boldsymbol{U}\\ 0\end{bmatrix}\boldsymbol{v}}{\boldsymbol{v}^*\boldsymbol{v}} \\
&= \min_{U\in\mathcal{U}_{n,k}} \max_{0\neq v\in\mathbb{C}^k} \frac{\boldsymbol{v}^*\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}\boldsymbol{v}}{\boldsymbol{v}^*\boldsymbol{v}} = \lambda_k(\boldsymbol{A}) \\
\lambda_{k+1}(\widehat{\boldsymbol{A}}) &= \max_{\boldsymbol{U}\in\mathcal{U}_{n+1,n-k+1}} \min_{0\neq v\in\mathbb{C}^{n-k+1}} \frac{\boldsymbol{v}^*\boldsymbol{U}^*\widehat{\boldsymbol{A}}\boldsymbol{U}\boldsymbol{v}}{\boldsymbol{v}^*\boldsymbol{v}} \\
&\ge \max_{\boldsymbol{U}\in\mathcal{U}_{n,n-k+1}} \min_{0\neq v\in\mathbb{C}^{n-k+1}} \frac{\boldsymbol{v}^*\begin{bmatrix}\boldsymbol{U}^* & 0\end{bmatrix}\widehat{\boldsymbol{A}}\begin{bmatrix}\boldsymbol{U}\\ 0\end{bmatrix}\boldsymbol{v}}{\boldsymbol{v}^*\boldsymbol{v}} \\
&= \max_{\boldsymbol{U}\in\mathcal{U}_{n,n-k+1}} \min_{0\neq \boldsymbol{v}\in\mathbb{C}^{n-k+1}} \frac{\boldsymbol{v}^*\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}\boldsymbol{v}}{\boldsymbol{v}^*\boldsymbol{v}} = \lambda_k(\boldsymbol{A})
\end{aligned}$$

□

Problem 7.2 (Poincaré separation theorem)**.** Let $\boldsymbol{A}\in\mathbb{C}^{n\times n}$ be Hermitian, $\boldsymbol{U}\in\mathcal{U}_{n,m}$, and $\boldsymbol{B}=\boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}$. Show that for every $k=1,\ldots,m$ holds

$$\lambda_k(\boldsymbol{A})\le\lambda_k(\boldsymbol{B})\le\lambda_{k+n-m}(\boldsymbol{A})\,.$$

Hint: consider first the case $\boldsymbol{U}^*=\begin{bmatrix}\boldsymbol{I} & 0\end{bmatrix}$ and use the previous theorem. Then take $\boldsymbol{U}^*=\begin{bmatrix}\boldsymbol{I} & 0\end{bmatrix}\boldsymbol{Q}$.

Finally let us mention the following without a proof:

Theorem 7.7 (Hoffman-Wielandt)**.** *If* $\boldsymbol{A}$ *and* $\boldsymbol{E}$ *are Hermitian, then*

$$\sum_{j=1}^{n}(\lambda_j(\boldsymbol{A}+\boldsymbol{E})-\lambda_j(\boldsymbol{A}))^2\le\|\boldsymbol{E}\|_F^2\ .$$

7.2. **Inertia.** The *inertia* of a Hermitian matrix $\boldsymbol{A}$ is defined as:

$$i(\boldsymbol{A})=(i_-(\boldsymbol{A}),i_0(\boldsymbol{A}),i_+(\boldsymbol{A}))\ ,$$

where $i_-(\boldsymbol{A})/i_+(\boldsymbol{A})$ is the number of negative/positive eigenvalues (with multiplicities) and $i_0(\boldsymbol{A})=n-i_-(\boldsymbol{A})-i_+(\boldsymbol{A})$ is the multiplicity of eigenvalue 0.

Matrix $\boldsymbol{B}$ is *congruent* with matrix $\boldsymbol{A}$, if there exists a regular matrix $\boldsymbol{S}$ such that $\boldsymbol{B}=\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S}$.

Theorem 7.8 (Sylvester's law of inertia). *Congruent Hermitian matrices have the same inertia.*

Proof. Let $\boldsymbol{S}$ be regular, $\boldsymbol{B} = \boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S}$, $\lambda_1(\boldsymbol{A}) \le \cdots \le \lambda_p(\boldsymbol{A}) < 0$ the negative eigenvalues of $\boldsymbol{A}$, $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_p$ the correspoding eigenvectors, and $\boldsymbol{Q} = \begin{bmatrix}\boldsymbol{q}_1 & \ldots & \boldsymbol{q}_p\end{bmatrix}$. Set $V_0 = R(\boldsymbol{S}^{-1}\boldsymbol{Q})$. Then

$$\begin{aligned}\lambda_p(\boldsymbol{B}) &= \min_{\dim(V)=p} \max_{0\ne\boldsymbol{x}\in V} \frac{\boldsymbol{x}^*\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S}\boldsymbol{x}}{\|\boldsymbol{x}\|_2^2} \\ &\le \max_{0\ne\boldsymbol{x}\in V_0} \frac{\boldsymbol{x}^*\boldsymbol{S}^*\boldsymbol{A}\boldsymbol{S}\boldsymbol{x}}{\|\boldsymbol{x}\|_2^2} = \max_{0\ne\boldsymbol{y}\in\mathbb{C}^p} \frac{\boldsymbol{y}^*\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q}\boldsymbol{y}}{\left\|\boldsymbol{S}^{-1}\boldsymbol{Q}\boldsymbol{y}\right\|_2^2} .\end{aligned}$$

Since $\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q} = \operatorname{diag}(\lambda_1(\boldsymbol{A}), \ldots, \lambda_p(\boldsymbol{A})$ and

$$\left\|\boldsymbol{S}^{-1}\boldsymbol{Q}\boldsymbol{y}\right\|_2 \le \left\|\boldsymbol{S}^{-1}\right\|_2 \|\boldsymbol{Q}\boldsymbol{y}\|_2 = \left\|\boldsymbol{S}^{-1}\right\|_2 \|\boldsymbol{y}\|_2 ,$$

we get

$$\lambda_p(\boldsymbol{B}) \le \max_{0\ne\boldsymbol{y}\in\mathbb{C}^p} \frac{\sum_{j=1}^p \lambda_j(\boldsymbol{A})\,|y_j|^2}{\left\|\boldsymbol{S}^{-1}\right\|_2^2 \|\boldsymbol{y}\|_2^2} \le \frac{\lambda_p}{\left\|\boldsymbol{S}^{-1}\right\|_2^2} < 0 .$$

Hence $i_-(\boldsymbol{B}) \ge i_-(\boldsymbol{A}) = p$. Similarly, $i_-(\boldsymbol{B}) \le i_-(\boldsymbol{A})$, $i_+(\boldsymbol{B}) \le i_+(\boldsymbol{A})$, and $i_+(\boldsymbol{B}) \ge i_+(\boldsymbol{A})$, so that also $i_0(\boldsymbol{B}) = i_0(\boldsymbol{A})$. □

7.3. Singular value decomposition.

The following theorem is an important tool of analysis. Further it gives means to solve illconditioned systems of equations.

Theorem 7.9. *For every $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ there exist unitary $\boldsymbol{U} \in \mathbb{C}^{m\times m}$ and $\boldsymbol{V} \in \mathbb{C}^{n\times n}$ such that $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$, where $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m\times n}$, $p = \min(m, n)$ and $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_p$.*

Proof. Let $\sigma_1 = \|\boldsymbol{A}\|_2$, $\boldsymbol{x}$ be such that $\|\boldsymbol{x}\|_2 = 1$, $\|\boldsymbol{A}\boldsymbol{x}\|_2 = \sigma_1$. Set $\boldsymbol{y} = \frac{1}{\sigma_1}\boldsymbol{A}\boldsymbol{x}$, so that $\|\boldsymbol{y}\|_2 = 1$. Let $\boldsymbol{U}_1 = \begin{bmatrix}\boldsymbol{y} & \widetilde{\boldsymbol{U}}_1\end{bmatrix}$ and $\boldsymbol{V}_1 = \begin{bmatrix}\boldsymbol{x} & \widetilde{\boldsymbol{V}}_1\end{bmatrix}$ be unitary (we get them by completing $\boldsymbol{y}$ and $\boldsymbol{x}$ to orthonormal basis of $\mathbb{C}^m$ and $\mathbb{C}^n$, respectively). Then

$$\boldsymbol{U}_1^*\boldsymbol{A}\boldsymbol{V}_1 = \begin{bmatrix}\boldsymbol{y}^* \\ \widetilde{\boldsymbol{U}}_1^*\end{bmatrix} \begin{bmatrix}\boldsymbol{A}\boldsymbol{x} & \boldsymbol{A}\widetilde{\boldsymbol{V}}_1\end{bmatrix} = \begin{bmatrix}\sigma_1 & \boldsymbol{w}^* \\ 0 & \boldsymbol{A}_1\end{bmatrix} = \boldsymbol{B} .$$

Now $\boldsymbol{B}\begin{bmatrix}\sigma_1 \\ \boldsymbol{w}\end{bmatrix} = \begin{bmatrix}\sigma_1^2 + \|\boldsymbol{w}\|_2^2 \\ \boldsymbol{A}_1\boldsymbol{w}\end{bmatrix}$, so that

$$\|\boldsymbol{B}\|_2 \ge \frac{\sigma_1^2 + \|\boldsymbol{w}\|_2^2}{\sqrt{\sigma_1^2 + \|\boldsymbol{w}\|_2^2}} = \sqrt{\sigma_1^2 + \|\boldsymbol{w}\|_2^2} \ge \sigma_1 = \|\boldsymbol{A}\|_2 = \|\boldsymbol{B}\|_2 ,$$

which implies $\|\boldsymbol{w}\|_2 = 0$. Continue similarly with matrix $\boldsymbol{A}_1$. □

Numbers σ_j are called the *singular values* of $\boldsymbol{A}$ and the columns of matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are, respectively the left and right *singular vectors* of $\boldsymbol{A}$. From equation

$$\boldsymbol{A}^*\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}\boldsymbol{V}^* \qquad \text{ja} \qquad \boldsymbol{A}\boldsymbol{A}^* = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{U}^*$$

we see, that $\boldsymbol{A}^*\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^*$ are unitarily similar to diagonal matrices $\boldsymbol{\Sigma}^T\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T$, respectively, and that the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are eigenvectors of matrices $\boldsymbol{A}\boldsymbol{A}^*$ and $\boldsymbol{A}^*\boldsymbol{A}$, respectively.

Remark 7.2. If $\boldsymbol{A}$ above is a real matrix, then the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ can be chosen to be real orthogonal matrices.

Remark 7.3. The singular value decomposition can also be written in the form:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* = \sum_{j=1}^{p} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^* \ ,$$

where the vectors $\boldsymbol{u}_j$ and $\boldsymbol{v}_j$ are the first p columns of $\boldsymbol{U}$ and $\boldsymbol{V}$.

Problem 7.3 (Polar decomposition)**.** Show, that for every $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ there exists a Hermitian positive semidefinite $\boldsymbol{P}$ and $\boldsymbol{U} \in \mathbb{C}^{m\times n}$, with $\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}$ such that $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{U}$. Furthermore then: $\boldsymbol{P}^2 = \boldsymbol{A}\boldsymbol{A}^*$.

When $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ is invertible, then $\boldsymbol{A}^{-1} = \boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^*$ and the singular values of $\boldsymbol{A}^{-1}$ are $1/\sigma_j$, $j = n, \ldots, 1$. In particular,

$$\left\|\boldsymbol{A}^{-1}\right\|_2 = \sigma_n^{-1} \qquad \text{and} \qquad \kappa_2(\boldsymbol{A}) = \|\boldsymbol{A}\|_2 \left\|\boldsymbol{A}^{-1}\right\|_2 = \sigma_1/\sigma_n \ .$$

Problem 7.4. Show: $\prod_{j=1}^n \sigma_j = |\det \boldsymbol{A}|$, $\boldsymbol{A} \in \mathbb{C}^{n\times n}$.

Problem 7.5. Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0$ be the singular values of matrix $\boldsymbol{A} \in \mathbb{C}^{m\times n}$. Show that $r = \operatorname{rank}(\boldsymbol{A})$, $R(\boldsymbol{A}) = \operatorname{span}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r)$, and $N(\boldsymbol{A}) = \operatorname{span}(\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_n)$. Show also that the pseudoinverse of $\boldsymbol{A}$ is obtained as:

$$\boldsymbol{A}^\dagger = \lim_{\varepsilon\to 0^+} (\boldsymbol{A}^*\boldsymbol{A} + \varepsilon\boldsymbol{I})^{-1}\boldsymbol{A}^* = \boldsymbol{V} \begin{bmatrix} \sigma_1^{-1} & & & \\ & \ddots & & \\ & & \sigma_r^{-1} & \\ & & & \end{bmatrix} \boldsymbol{U}^*.$$

Problem 7.6. Show: every $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ satisfies:

$$\|\boldsymbol{A}\|_F^2 = \operatorname{tr}(\boldsymbol{A}^*\boldsymbol{A}) = \sum_{j=1}^{r} \sigma_j^2 \ .$$

Hint: tr is invariant under similarity transforms.

Problem 7.7. Show: if $\boldsymbol{A}$ is invertible, then the singular matrix $\boldsymbol{B}$ closest to $\boldsymbol{A}$ satisfies: $\|\boldsymbol{A}-\boldsymbol{B}\|_2=\sigma_n$.

The following more general result is also left as an exercise.

Theorem 7.10. *Define* $\boldsymbol{A}_j=\sum_{k=1}^{j}\sigma_k\,\boldsymbol{u}_k\boldsymbol{v}_k^*$, $j=1,\ldots,r$, *(*$r=\operatorname{rank}(\boldsymbol{A})$ *) so that* $\boldsymbol{A}_r=\boldsymbol{A}$ *. Then* $\operatorname{rank}(\boldsymbol{A}_j)=j$ *and*

$$\|\boldsymbol{A}-\boldsymbol{A}_j\|_2=\sigma_{j+1}=\inf_{\operatorname{rank}(\boldsymbol{B})=j}\|\boldsymbol{A}-\boldsymbol{B}\|_2 \ .$$

Sensitivity of solution of a linear system Let $\boldsymbol{A}$ be invertible and $\boldsymbol{Ax}=\boldsymbol{b}$. Consider the problems:

$$(\boldsymbol{A}+\varepsilon\boldsymbol{E})\boldsymbol{x}(\varepsilon)=\boldsymbol{b}+\varepsilon\boldsymbol{f} \ ,$$

so that $\boldsymbol{Ex}(\varepsilon)+(\boldsymbol{A}+\varepsilon\boldsymbol{E})\dot{\boldsymbol{x}}(\varepsilon)=\boldsymbol{f}$, from which $\dot{\boldsymbol{x}}(0)=\boldsymbol{A}^{-1}(\boldsymbol{f}-\boldsymbol{Ex}(0))$ and

$$\begin{aligned}\frac{\|\boldsymbol{x}(\varepsilon)-\boldsymbol{x}\|}{\|\boldsymbol{x}\|}&\le\varepsilon\left\|\boldsymbol{A}^{-1}\right\|(\|\boldsymbol{f}\|+\|\boldsymbol{E}\|\,\|\boldsymbol{x}\|)/\,\|\boldsymbol{x}\|+O(\varepsilon^2)\\&\le\|\boldsymbol{A}\|\left\|\boldsymbol{A}^{-1}\right\|(\frac{\varepsilon\,\|\boldsymbol{f}\|}{\|\boldsymbol{b}\|}+\frac{\varepsilon\,\|\boldsymbol{A}\|}{\|\boldsymbol{A}\|})+O(\varepsilon^2) \ .\end{aligned}$$

In other words:

$$\text{relative change in } \boldsymbol{x}\le\kappa(\boldsymbol{A})\big(\text{relative change in } \boldsymbol{b}+\text{relative change in } \boldsymbol{A}\big)+O(\varepsilon^2) \ .$$

When the norm is Euclidean, then $\kappa_2(\boldsymbol{A})=\|\boldsymbol{A}\|_2\left\|\boldsymbol{A}^{-1}\right\|_2=\sigma_1/\sigma_n$.

Problem 7.8. Show, that the k :th singular value of matrix $\boldsymbol{A}\in\mathbb{C}^{m\times n}$ is given by:

$$\sigma_k=\max_{\boldsymbol{U}\in\mathcal{U}_{n,k}}\ \min_{0\neq\boldsymbol{v}\in\mathbb{C}^k}\ \frac{\|\boldsymbol{AUv}\|_2}{\|\boldsymbol{v}\|_2} \ .$$

Let $\boldsymbol{A}=\boldsymbol{U\Sigma V}^*$ be the singular value decomposition of matrix $\boldsymbol{A}\in\mathbb{C}^{m\times n}$, $m\ge n$ where $\boldsymbol{U}=\begin{bmatrix}\boldsymbol{U}_1 & \boldsymbol{U}_2\end{bmatrix}$, $\boldsymbol{U}_1\in\mathbb{C}^{m\times n}$, and

$$\boldsymbol{U}_1^*\boldsymbol{AV}=\begin{bmatrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_n\end{bmatrix} \ .$$

Then

$$\boldsymbol{Q}=\frac{1}{\sqrt{2}}\begin{bmatrix}\boldsymbol{v} & \boldsymbol{V} & 0\\ \boldsymbol{U}_1 & -\boldsymbol{U}_1 & \sqrt{2}\boldsymbol{U}_2\end{bmatrix}\in\mathbb{C}^{(m+n)\times(m+n)}$$

is unitary and

$$\boldsymbol{Q}^* \begin{bmatrix} 0 & \boldsymbol{A}^* \\ \boldsymbol{A} & 0 \end{bmatrix} \boldsymbol{Q} = \frac{1}{2} \begin{bmatrix} \boldsymbol{v}^* & \boldsymbol{U}_1^* \\ \boldsymbol{V}^* & -\boldsymbol{U}_1^* \\ 0 & \sqrt{2}\boldsymbol{U}_2^* \end{bmatrix} \begin{bmatrix} \boldsymbol{A}^*\boldsymbol{U}_1 & -\boldsymbol{A}^*\boldsymbol{U}_1 & \sqrt{2}\boldsymbol{A}^*\boldsymbol{U}_2 \\ \boldsymbol{A}\boldsymbol{V} & \boldsymbol{A}\boldsymbol{V} & 0 \end{bmatrix}$$

$$= \operatorname{diag}(\sigma_1, \ldots, \sigma_n, -\sigma_1, \ldots, -\sigma_n, 0, \ldots, 0) \ .$$

Problem 7.9. Use the Weyl theorem 7.4 to matrices

$$\begin{bmatrix} 0 & \boldsymbol{A}^* \\ \boldsymbol{A} & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & \boldsymbol{A}^* + \boldsymbol{E}^* \\ \boldsymbol{A} + \boldsymbol{E} & 0 \end{bmatrix}$$

and show, that for all $k = 1, \ldots, n$ holds

$$|\sigma_k(\boldsymbol{A} + \boldsymbol{E}) - \sigma_k(\boldsymbol{A})| \leq \|\boldsymbol{E}\|_2 \ .$$

Problem 7.10. Similarly, using the Hoffman–Wielandt theorem 7.7 show:

$$\sum_{k=1}^{n} |\sigma_k(\boldsymbol{A} + \boldsymbol{E}) - \sigma_k(\boldsymbol{A})|^2 \leq \|\boldsymbol{E}\|_F^2 \ .$$

Problem 7.11. Let $\boldsymbol{A} \in \mathbb{C}^{m \times n}$, $m \geq n$ and $\widetilde{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{a} \end{bmatrix} \in \mathbb{C}^{m \times (n+1)}$. Show:

$$\sigma_1(\widetilde{\boldsymbol{A}}) \geq \sigma_1(\boldsymbol{A}) \geq \sigma_2(\widetilde{\boldsymbol{A}}) \geq \sigma_2(\boldsymbol{A}) \geq \ldots \sigma_n(\boldsymbol{A}) \geq \sigma_{n+1}(\widetilde{\boldsymbol{A}}) \ .$$

Hint: Theorem 7.6 for matrix $\widetilde{\boldsymbol{A}}^* \widetilde{\boldsymbol{A}}$.

Problem 7.12. Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}$. Find the connection between the singular values and singular vectors of matrices $\boldsymbol{A} + i\boldsymbol{B}$ and $\begin{bmatrix} \boldsymbol{A} & -\boldsymbol{B} \\ \boldsymbol{B} & \boldsymbol{A} \end{bmatrix}$.

8. Computation of eigenvalues of a Hermitian matrix

8.1. **Hermitian $\boldsymbol{QR}$ iteration.** When we apply the Hessenberg $\boldsymbol{QR}$ algorithm to a Hermitian matrix, we notice:

1) $\boldsymbol{H}_0 = \boldsymbol{U}_0^* \boldsymbol{A} \boldsymbol{U}_0$ is Hermitian and Hessenberg form, in partcular it is a *tridiagonal matrix*

$$\boldsymbol{H}_0 = \begin{bmatrix} h_1 & g_1 & & & \\ \bar{g}_1 & h_2 & g_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \bar{g}_{n-2} & h_{n-1} & g_{n-1} \\ & & & \bar{g}_{n-1} & h_n \end{bmatrix}.$$

2) Hermitianity and tridiagonality are preserved in the $\boldsymbol{QR}$ iteration.
3) Only real shifts are needed.

We get the Hermitian $\boldsymbol{QR}$ iteration:

```
$\boldsymbol{T}_0 = \boldsymbol{Q}_0^* \boldsymbol{A} \boldsymbol{Q}_0$        tridiagonal form
for $k = 1, 2, \ldots$
        choose a real $\mu$
        $\boldsymbol{Q}_k \boldsymbol{R}_k = \boldsymbol{T}_{k-1} - \mu \boldsymbol{I}$        (The $\boldsymbol{QR}$ decomposition of $\boldsymbol{T}_{k-1} - \mu \boldsymbol{I}$)
        $\boldsymbol{T}_k = \boldsymbol{R}_k \boldsymbol{Q}_k + \mu \boldsymbol{I}$
end
```

Problem 8.1. Count the floating point operations of this algorithm taking into account that $\boldsymbol{A}$ is Hermitian and $\boldsymbol{T}_k$'s are tridiagonal matrices.

Problem 8.2. Experiment this algorithm with different Hermitian matrices $\boldsymbol{A}$ using shift strategies

a) $\mu = h_n$
b) $\mu =$ the smaller eigenvalue of matrix $\begin{bmatrix} h_{n-1} & g_{n-1} \\ \bar{g}_{n-1} & h_n \end{bmatrix}$.

Problem 8.3. The *Rayleigh quotient iteration* is based on the following heuristics: if $\boldsymbol{x}$ is an approximate eigenvector of $\boldsymbol{A}$, then $\mu = \frac{\boldsymbol{x}^* \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^* \boldsymbol{x}}$ is close to an eigenvalue of $\boldsymbol{A}$, so that the solution of the system $(\boldsymbol{A} - \mu \boldsymbol{I})\boldsymbol{z} = \boldsymbol{x}$ should be a better approximation of the eigenvector.

[0] Version: April 9, 2003

Experiment the following algorithm for several Hermitian $\boldsymbol{A}$

```
$\|\boldsymbol{x}_0\|_2 = 1$
for $k =$ $0, 1, \ldots$
    $\mu_k = \boldsymbol{x}_k^* \boldsymbol{A} \boldsymbol{x}_k$
    solve $(\boldsymbol{A} - \mu \boldsymbol{I})\boldsymbol{z}_k = \boldsymbol{x}_k$
    $\boldsymbol{x}_{k+1} = \boldsymbol{z}_k / \|\boldsymbol{z}_k\|_2$
end
```

First the work load of this seems big, since at every step we need to solve a system of equations, but in many cases the matrix $\boldsymbol{A}$ is sparse and approximate solution of the system is cheap.

8.2. **Jacobi iteration.** The idea of Jacobi iteration is based on systematic decreasing of the off–diagonal elements of a Hermitian matrix. This is done with Givens rotations, which in this context are also called the Jacobi rotations.

Schematically the algorithm goes as:
Let $\boldsymbol{A}$ Hermitian. Set

$$\operatorname{off}(\boldsymbol{A}) = \sum_{j \neq k} |a_{j,k}|^2 \ .$$

Pick an $a_{p,q} \neq 0$ and take unitary $\begin{bmatrix} \bar{\alpha} & \bar{\beta} \\ -\beta & \alpha \end{bmatrix}$ such that

$$\begin{bmatrix} \bar{\alpha} & \bar{\beta} \\ -\beta & \alpha \end{bmatrix} \begin{bmatrix} a_{p,p} & a_{p,q} \\ a_{q,p} & a_{q,q} \end{bmatrix} \begin{bmatrix} \alpha & -\bar{\beta} \\ \beta & \bar{\alpha} \end{bmatrix} = \begin{bmatrix} b_{p,p} & 0 \\ 0 & b_{q,q} \end{bmatrix} \ .$$

Then the correspoding rotation $\boldsymbol{J}_{p,q}$ gives a unitary similarity transformation $\boldsymbol{B} = \boldsymbol{J}_{p,q} \boldsymbol{A} \boldsymbol{J}_{p,q}^*$. Since the Frobenius norm does not change under unitary transformations and since

$$|a_{p,p}|^2 + 2\,|a_{p,q}|^2 + |a_{q,q}|^2 = |b_{p,p}|^2 + |b_{q,q}|^2 \ ,$$

we get:

$$\operatorname{off}(\boldsymbol{B}) = \|\boldsymbol{B}\|_F^2 - \sum_{k=1}^{n} |b_{k,k}|^2 = \|\boldsymbol{A}\|_F^2 - \sum_{k=1}^{n} |a_{k,k}|^2 - 2\,|a_{p,q}|^2 = \operatorname{off}(\boldsymbol{A}) - 2\,|a_{p,q}|^2 \ .$$

In this method the zeros don't stay zeros, but off decreases at each step. Typically the $a_{p,q}$ with largest absolute value is chosen.

Problem 8.4. Experiment the Jacobi iteration with some Hermitian matrices.

Jacobi iteration is not very effective, but there exists versions of it that are easily parallelizable.

8.3. **Divide and conquer method.** If the aim is to exploit parallel computation, then the following approach is very natural. Consider computation of eigenvalues of a tridiagonal Hermitian matrix. Let us write it in the form: a block diagonal matrix plus a rank–one correction:

$$\boldsymbol{T} = \begin{bmatrix} h_1 & g_1 & & \\ \bar{g}_1 & h_2 & \ddots & \\ & \ddots & \ddots & g_{n-1} \\ & & \bar{g}_{n-1} & h_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{T}_1 & 0 \\ 0 & \boldsymbol{T}_2 \end{bmatrix} + \boldsymbol{v}\boldsymbol{v}^* ,$$

where $\boldsymbol{T}_1 \in \mathbb{C}^{m\times m}$ and $\boldsymbol{v} = \boldsymbol{e}_m + \theta\boldsymbol{e}_{m+1}$. Then in the center we have:

$$\begin{bmatrix} h_m & g_m \\ \bar{g}_m & h_{m+1} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{T}_1)_{m,m} & 0 \\ 0 & (\boldsymbol{T}_2)_{1,1} \end{bmatrix} + \begin{bmatrix} 1 & \bar{\theta} \\ \theta & |\theta|^2 \end{bmatrix} ,$$

so that we have to take $\theta = \bar{g}_m$.

In the divide and conquer method the eigendecompositions $\boldsymbol{T}_j = \boldsymbol{Q}_j\boldsymbol{D}_j\boldsymbol{Q}_j^*$ of matrices $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ are computed in parallel. Then setting $\boldsymbol{U} = \begin{bmatrix} \boldsymbol{Q}_1 & 0 \\ 0 & \boldsymbol{Q}_2 \end{bmatrix}$ we get

$$\boldsymbol{U}^*\boldsymbol{T}\boldsymbol{U} = \begin{bmatrix} \boldsymbol{D}_1 & 0 \\ 0 & \boldsymbol{D}_2 \end{bmatrix} + \boldsymbol{z}\boldsymbol{z}^* ,$$

where $\boldsymbol{z} = \boldsymbol{U}^*\boldsymbol{v}$. Hence we get the diagonalization of $\boldsymbol{T}$ from the eigendecompositions of the blocks, provided that we can easily compute the eigenvalues and eigenvectors of a matrix of the form: diagonal plus a rank–one matrix.

Eigenvalues and vectors of matrix $\boldsymbol{D} + \boldsymbol{z}\boldsymbol{z}^*$. Let us start with the case where the eigenvalues of $\boldsymbol{D}$ are simple and all the components of $\boldsymbol{z}$ are different from zero. Assume further, that the diagonal elements of $\boldsymbol{D}$ are in the increasing order (which is easily obtained by permutations).

Theorem 8.1. *Let $\boldsymbol{D} = \operatorname{diag}(d_1, \dots, d_n) \in \mathbb{R}^{n\times n}$, $d_1 < d_2 < \cdots < d_n$ and $\boldsymbol{z} \in \mathbb{C}^n$, $z_j \neq 0\ \forall j$. Then $\Lambda(\boldsymbol{D} + \boldsymbol{z}\boldsymbol{z}^*) \cap \Lambda(\boldsymbol{D}) = \emptyset$ and the eigenvalues $\lambda_1 < \cdots < \lambda_n$ of the matrix $\boldsymbol{D} + \boldsymbol{z}\boldsymbol{z}^*$ and correspoding eigenvectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_n$ satisfy:*

a) $\boldsymbol{v}_j = \alpha_j(\boldsymbol{D} - \lambda_j\boldsymbol{I})^{-1}\boldsymbol{z}$
b) $\Lambda(\boldsymbol{D} + \boldsymbol{z}\boldsymbol{z}^*) = \{\lambda \mid f(\lambda) = 0\}$, *where* $f(\lambda) = 1 + \boldsymbol{z}^*(\boldsymbol{D} - \lambda\boldsymbol{I})^{-1}\boldsymbol{z}$.
c) $d_1 < \lambda_1 < d_2 < \lambda_2 < \dots d_n < \lambda_n$.

Proof. Let $\lambda \in \Lambda(\boldsymbol{D} + \boldsymbol{z}\boldsymbol{z}^*)$ and $\boldsymbol{v} \neq 0$ be such that

$$(\boldsymbol{D} + \boldsymbol{z}\boldsymbol{z}^*)\boldsymbol{v} = \lambda\boldsymbol{v} . \tag{8.1}$$

We have $\lambda \notin \Lambda(\boldsymbol{D})$, since if $\lambda = d_j$, then

$$0 = \boldsymbol{e}_j^T((\boldsymbol{D} - d_j\boldsymbol{I})\boldsymbol{v} + \boldsymbol{z}\boldsymbol{z}^*\boldsymbol{v}) = z_j\boldsymbol{z}^*\boldsymbol{v} \ ,$$

from which $\boldsymbol{z}^*\boldsymbol{v} = 0$ and $\boldsymbol{D}\boldsymbol{v} = d_j\boldsymbol{v}$, so that $\boldsymbol{v} = \mu\boldsymbol{e}_j$ and $\boldsymbol{z}^*\boldsymbol{v} = \mu z_j \neq 0$, a contradiction. Hence $\Lambda(\boldsymbol{D} + \boldsymbol{z}\boldsymbol{z}^*) \cap \Lambda(\boldsymbol{D}) = \emptyset$.

From equation (8.1) we also see, that $\boldsymbol{z}^*\boldsymbol{v}_j \neq 0$ for all j, and that $(\boldsymbol{D} - \lambda_j\boldsymbol{I})\boldsymbol{v}_j = -\boldsymbol{z}^*\boldsymbol{v}_j\boldsymbol{z}$, which imply a).

c) is given directly by theorem 7.5, since we don't have equalities.

Multiplication of (8.1) by $\boldsymbol{z}^*(\boldsymbol{D} - \lambda_j\boldsymbol{I})^{-1}$ gives:

$$\boldsymbol{z}^*\boldsymbol{v}_j(1 + \boldsymbol{z}^*(\boldsymbol{D} - \lambda_j\boldsymbol{I})^{-1}\boldsymbol{z}) = 0 \ ,$$

so that the eigenvalues are the zeros of f. Other zeros do not exist since

$$f(\lambda) = 1 + \sum_{j=1}^{n} \frac{|z_j|^2}{d_j - \lambda} \qquad \text{and} \qquad f'(\lambda) = \sum_{j=1}^{n} \frac{|z_j|^2}{(d_j - \lambda)^2} > 0 \ ,$$

so that f is strictly increasing on each interval

$$(-\infty, d_1), (d_1, d_2), \ldots, (d_{n-1}, d_n), (d_n, \infty)$$

and $\lim_{|\lambda|\to\infty} f(\lambda) = 1$, i.e., f has exactly n zeros. □

Problem 8.5. Show, that the general case $\boldsymbol{D} = \operatorname{diag}(d_1, \ldots, d_n) \in \mathbb{R}^{n\times n}$, $\boldsymbol{z} \in \mathbb{C}^n$ can be brought by unitary transformation to the form

$$\boldsymbol{U}^*(\boldsymbol{D} + \boldsymbol{z}\boldsymbol{z}^*)\,\boldsymbol{U} = \begin{bmatrix} \widetilde{\boldsymbol{D}}_1 + \widetilde{\boldsymbol{z}}\widetilde{\boldsymbol{z}}^* & 0 \\ 0 & \widetilde{\boldsymbol{D}}_2 \end{bmatrix} \ ,$$

where $\widetilde{\boldsymbol{D}}_1 + \widetilde{\boldsymbol{z}}\widetilde{\boldsymbol{z}}^*$ satisfies the conditions of the previous theorem. Hint: if $d_j = d_k$, then take a rotation $\boldsymbol{G}_{j,k}$ such that $(\boldsymbol{G}_{j,k}\boldsymbol{z})_k = 0$. This does not change $\boldsymbol{D}$: $\boldsymbol{G}_{j,k}\boldsymbol{D}\boldsymbol{G}_{j,k}^* = \boldsymbol{D}$. Do this to all multiple eigenvalues and finally apply a permutation.

Problem 8.6. Estimate the flop count of the divide and conquer method, when $n = 2^k$ and all problems are divided into two of half the size until we have one–dimensional problems. Assuming we have n processors that can be run in parallel what is the computing time?

8.4. **Computation of the singular value decomposition.** In principle one can compute the singular values for example by applying the Hermitian $\boldsymbol{QR}$ iteration to the matrix $\boldsymbol{A}^*\boldsymbol{A}$. The following approach, however, turns out to be more efficient.

Problem 8.7. Given $\boldsymbol{A} \in \mathbb{C}^{m\times n}$, $m \geq n$, show, how to find unitary $\boldsymbol{U}_0$ and $\boldsymbol{V}_0$ such that

$$\boldsymbol{U}_0^*\boldsymbol{A}\boldsymbol{V}_0 = \begin{bmatrix}\boldsymbol{B}\\ 0\end{bmatrix} = \begin{bmatrix} b_1 & c_1 & & \\ & \ddots & \ddots & \\ & & b_{n-1} & c_{n-1} \\ & & & b_n \\ & & & \\ & & & \end{bmatrix} \in \mathbb{R}^{m\times n} \ ,$$

i.e., $\boldsymbol{B}$ is a real *bidiagonal matrix.*

After this the problem is to find the singular value decomposition of the bidiagonal matrix $\boldsymbol{B}$. This can be done using the following Golub–Kahan iteration, which essentially does the same as the Hermitian $\boldsymbol{QR}$ iteration for the matrix $\boldsymbol{B}^T\boldsymbol{B}$, but transforms only the matrix $\boldsymbol{B}$.

1) Choose a shift μ , for example, the smaller eigenvalue of the lower right 2×2 corner $\begin{bmatrix} b_{n-1}^2 + c_{n-2}^2 & b_{n-1}c_{n-1} \\ b_{n-1}c_{n-1} & b_n^2 + c_{n-1}^2 \end{bmatrix}$ of the matrix $\boldsymbol{B}^T\boldsymbol{B}$.
2) Take $\boldsymbol{G}_{1,2} = \begin{bmatrix} \alpha & \beta & \\ -\beta & \alpha & 0 \\ 0 & & \boldsymbol{I} \end{bmatrix}$ such that $\begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix} \begin{bmatrix} b_1^2 - \mu \\ b_1c_1 \end{bmatrix} = \begin{bmatrix} \# \\ 0 \end{bmatrix}$, i.e., as the first rotation for $\boldsymbol{B}^T\boldsymbol{B}$.
3) Update the matrix $\boldsymbol{B}\boldsymbol{G}_{1,2}^T$ with rotations in turn from the left and from the right such that the result is again a bidiagonal matrix:

$$\boldsymbol{B}\boldsymbol{G}_{1,2} = \begin{bmatrix} \# & \# & & & & \\ \# & \# & \# & & & \\ & & \# & \# & & \\ & & & \ddots & \ddots & \\ & & & & \# & \# \\ & & & & & \# \end{bmatrix} \xrightarrow{\boldsymbol{U}_1[\cdot]} \begin{bmatrix} \# & \# & \# & & & \\ 0 & \# & \# & & & \\ & & \# & \# & & \\ & & & \ddots & \ddots & \\ & & & & \# & \# \\ & & & & & \# \end{bmatrix} \xrightarrow{[\cdot]\boldsymbol{V}_2^T}$$

$$\begin{bmatrix} \# & \# & 0 & & & \\ 0 & \# & \# & & & \\ & \# & \# & \# & & \\ & & & \ddots & \ddots & \\ & & & & \# & \# \\ & & & & & \# \end{bmatrix} \xrightarrow{\boldsymbol{U}_2[\cdot]} \begin{bmatrix} \# & \# & 0 & & & \\ 0 & \# & \# & \# & & \\ & 0 & \# & \# & & \\ & & & \ddots & \ddots & \\ & & & & \# & \# \\ & & & & & \# \end{bmatrix} \xrightarrow{[\cdot]\boldsymbol{V}_3^T}$$

$$\dots \begin{bmatrix} \# & \# & 0 & & & \\ 0 & \# & \# & 0 & & \\ & 0 & \# & \# & 0 & \\ & & & \ddots & \ddots & 0 \\ & & & 0 & \# & \# \\ & & & & \# & \# \end{bmatrix} \xrightarrow{\boldsymbol{U}_{n-1}[\cdot]} \begin{bmatrix} \# & \# & 0 & & & \\ 0 & \# & \# & 0 & & \\ & 0 & \# & \# & 0 & \\ & & & \ddots & \ddots & 0 \\ & & & 0 & \# & \# \\ & & & & 0 & \# \end{bmatrix} = \widetilde{\boldsymbol{B}}$$

4) Set $\boldsymbol{B} = \widetilde{\boldsymbol{B}}$ and return to 1).

One iteration step is cheap: only $O(n)$ flops. The result is:

$$\begin{aligned} \widetilde{\boldsymbol{B}}^T \widetilde{\boldsymbol{B}} &= (\boldsymbol{U}_{n-1} \dots \boldsymbol{U}_1 \boldsymbol{B} \boldsymbol{G}_{1,2}^T \boldsymbol{V}_2^T \dots \boldsymbol{V}_{n-1}^T)^T \boldsymbol{U}_{n-1} \dots \boldsymbol{U}_1 \boldsymbol{B} \boldsymbol{G}_{1,2}^T \boldsymbol{V}_2^T \dots \boldsymbol{V}_{n-1}^T \\ &= \boldsymbol{V}_{n-1} \dots \boldsymbol{V}_2 \boldsymbol{G}_{1,2} \boldsymbol{B}^T \boldsymbol{B} \boldsymbol{G}_{1,2}^T \boldsymbol{V}_2^T \dots \boldsymbol{V}_{n-1}^T \ . \end{aligned}$$

It turns out, that this is the same result that we would have got by a shifted $\boldsymbol{QR}$ iteration step for the matrix $\boldsymbol{B}^T\boldsymbol{B}$ (see: Golub & Van Loan: Matrix Computations).

Problem 8.8. Compute the number of flops in the previous iteration.

If some c_j above becomes zero during the iteration, then the problem becomes block diagonal, i.e., divides into two smaller problems.

Similarly if some b_j becomes zero, we can divide the problem into two in the following way:

$$\begin{bmatrix} \boldsymbol{B}_1 & & & & & \\ & \# & & & & \\ & \# & \# & & & \\ & & \# & \# & & \\ & & & \ddots & \ddots & \\ & & & & \# & \# \\ & & & & & \# \end{bmatrix} \xrightarrow{\boldsymbol{G}_{j,j+1}} \begin{bmatrix} \boldsymbol{B}_1 & & & & & \\ & 0 & \# & & & \\ & \# & \# & & & \\ & & \# & \# & & \\ & & & \ddots & \ddots & \\ & & & & \# & \# \\ & & & & & \# \end{bmatrix}$$

$$\xrightarrow{\boldsymbol{G}_{j,j+2}} \begin{bmatrix} \boldsymbol{B}_1 & & & & & \\ & 0 & 0 & \# & & \\ & \# & \# & & & \\ & & \# & \# & & \\ & & & \ddots & \ddots & \\ & & & & \# & \# \\ & & & & & \# \end{bmatrix} \dots \xrightarrow{\boldsymbol{G}_{j,n}} \begin{bmatrix} \boldsymbol{B}_1 & & & & & \\ & 0 & 0 & 0 & \dots & 0 \\ & \# & \# & & & \\ & & \# & \# & & \\ & & & \ddots & \ddots & \\ & & & & \# & \# \\ & & & & & \# \end{bmatrix}$$

9. Krylov subspace iterations for eigenvalue problems

From now on we mostly consider methods that do not assume that when solving an eigenvalue problem we would have the matrix in the memory and we could manipulate it. We think having only a *linear operator*: $\boldsymbol{A} : \mathbb{C}^n \to \mathbb{C}^n$. Such situations we encounter for example, when $\boldsymbol{A}$ is a subroutine which for given vector $\boldsymbol{x} \in \mathbb{C}^n$ returns the vector $\boldsymbol{Ax} \in \mathbb{C}^n$. Often n is also so large, that storing an $n{\times}n$ matrix in the memory is not reasonable.

A subspace of the form

$$\mathcal{K}_j(\boldsymbol{A},\boldsymbol{b}) = \operatorname{span}(\boldsymbol{b}, \boldsymbol{Ab}, \boldsymbol{A}^2\boldsymbol{b}, \dots, \boldsymbol{A}^{j-1}\boldsymbol{b})$$

is called a *Krylov subspace.* The corresponding matrix

$$\boldsymbol{K}_j(\boldsymbol{A},\boldsymbol{b}) = \begin{bmatrix}\boldsymbol{b} & \boldsymbol{Ab} & \dots & \boldsymbol{A}^{j-1}\boldsymbol{b}\end{bmatrix}$$

is called a *Krylov matrix.*

One connection of such a matrix with the eigenvalues of $\boldsymbol{A}$ is seen in the following: Assume, that $\boldsymbol{X} = \boldsymbol{K}_n(\boldsymbol{A},\boldsymbol{b})$ is regular. Then $\boldsymbol{X} = \begin{bmatrix}\boldsymbol{x}_1 & \dots & \boldsymbol{x}_n\end{bmatrix}$ gives a similarity transformation to an interesting form:

$$\begin{aligned}\boldsymbol{X}^{-1}\boldsymbol{AX} =& \boldsymbol{X}^{-1}\begin{bmatrix}\boldsymbol{x}_2 & \boldsymbol{x}_3 & \dots & \boldsymbol{x}_n & \boldsymbol{Ax}_n\end{bmatrix} = \boldsymbol{X}^{-1}\boldsymbol{X}\begin{bmatrix}0 & \\ \boldsymbol{I} & -\boldsymbol{\alpha}\end{bmatrix} \\ =& \begin{bmatrix} & & & & -\alpha_0 \\ 1 & & & & -\alpha_1 \\ & 1 & & & -\alpha_2 \\ & & \ddots & & \vdots \\ & & & 1 & -\alpha_{n-1}\end{bmatrix} = \boldsymbol{B}\ ,\end{aligned}$$

where $\boldsymbol{\alpha} = -\boldsymbol{X}^{-1}\boldsymbol{A}^n\boldsymbol{b}$. This is called the *companion matrix* of $\boldsymbol{A}$. It satisfies:

$$\begin{aligned}\det(z\boldsymbol{I}-\boldsymbol{B}) =& \begin{vmatrix} z & & & & \alpha_0 \\ -1 & z & & & \alpha_1 \\ & -1 & z & & \alpha_2 \\ & & \ddots & \ddots & \vdots \\ & & & -1 & z+\alpha_{n-1}\end{vmatrix} = z\begin{vmatrix} z & & & \alpha_1 \\ -1 & z & & \alpha_2 \\ & \ddots & \ddots & \vdots \\ & & -1 & z+\alpha_{n-1}\end{vmatrix} + \alpha_0 = \\ =& z\det(z\boldsymbol{I}-\boldsymbol{B}_1) + \alpha_0 = z^n + \alpha_{n-1}z^{n-1} + \dots + \alpha_1 z + \alpha_0\ .\end{aligned}$$

in other words $-\boldsymbol{\alpha} = \boldsymbol{K}_n(\boldsymbol{A},\boldsymbol{b})^{-1}\boldsymbol{A}^n\boldsymbol{b}$ gives the coefficients of the characteristic polynomial of $\boldsymbol{B}$ and hence also of $\boldsymbol{A}$. Looking for the eigenvectors of $\boldsymbol{B}$ from the

[0] Version: April 9, 2003

system of equations

$$\begin{bmatrix} \lambda & & & & \alpha_0 \\ -1 & \lambda & & & \alpha_1 \\ & -1 & \lambda & & \alpha_2 \\ & & \ddots & \ddots & \vdots \\ & & & -1 & \lambda+\alpha_{n-1} \end{bmatrix} \boldsymbol{v} = 0$$

we see, that if $v_n = 0$, then $\boldsymbol{v} = 0$ and on the other hand, if $v_n = 1$, then other components are uniquely defined. Hence geometric multiplicity of each eigenvalue is one.

The following works also for nonregular $\boldsymbol{K}_n(\boldsymbol{A}, \boldsymbol{b})$. Let $p(z) = z^m + a_1 z^{m-1} + \cdots + a_{m-1}z + a_m$ be the minimal polynomial of $\boldsymbol{A}_{|\mathcal{K}_n(\boldsymbol{A},\boldsymbol{b})}$, i.e., the monic polynomial of lowest degree such that $p(\boldsymbol{A})\boldsymbol{b} = 0$. Set $\boldsymbol{c}_0 = \boldsymbol{b}$ and $\boldsymbol{c}_j = \boldsymbol{A}\boldsymbol{c}_{j-1} + a_{j-1}\boldsymbol{b}$, $j = 1, \ldots, m$. Then

$$\begin{aligned} \boldsymbol{c}_0 &= \boldsymbol{b} \\ \boldsymbol{c}_1 &= \boldsymbol{A}\,\boldsymbol{b} + a_1\boldsymbol{b} \\ \boldsymbol{c}_2 &= \boldsymbol{A}^2\boldsymbol{b} + a_1\boldsymbol{A}\boldsymbol{b} + a_2\boldsymbol{b} \\ &\vdots \\ \boldsymbol{c}_m &= \boldsymbol{A}^m\boldsymbol{b} + a_1\boldsymbol{A}^{m-1}\boldsymbol{b} + \cdots + a_{m-1}\boldsymbol{A}\,\boldsymbol{b} + a_m\boldsymbol{b} = 0 \end{aligned}$$

and $\boldsymbol{A}\boldsymbol{c}_j = \boldsymbol{c}_{j+1} - a_{j+1}\boldsymbol{c}_0$. Set $\boldsymbol{C} = [\boldsymbol{c}_0 \ \ldots \ \boldsymbol{c}_{m-1}]$. Then

$$\begin{aligned} \boldsymbol{A}\,\boldsymbol{C} &= [\boldsymbol{A}\,\boldsymbol{c}_0 \ \boldsymbol{A}\,\boldsymbol{c}_1 \ \ldots \ \boldsymbol{A}\,\boldsymbol{c}_{m-1}] \\ &= [\boldsymbol{c}_1 - a_1\boldsymbol{c}_0 \quad \boldsymbol{c}_2 - a_2\boldsymbol{c}_0 \quad \ldots \quad \boldsymbol{c}_m - a_m\boldsymbol{q}_0] \\ &= [\boldsymbol{c}_0 \ \boldsymbol{c}_1 \ \boldsymbol{c}_2 \ \ldots \ \boldsymbol{c}_{m-1}] \begin{bmatrix} -a_1 & -a_2 & & \ldots & -a_m \\ 1 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \end{bmatrix} \\ &= \boldsymbol{C}\,\boldsymbol{H}\ . \end{aligned}$$

Since $\boldsymbol{C}$ has linearly independent columns (why?) we get by lemma 5.1 that $\Lambda(\boldsymbol{H}) \subset \Lambda(\boldsymbol{A})$, i.e., we obtain m eigenvalues of $\boldsymbol{A}$ by computing the roots of p.

Another motivation for the use of Krylov subspaces we get from the following. Let $\boldsymbol{A}$ be real and symmetric. Then the largest and smallest eigenvalue of $\boldsymbol{A}$ are obtained from the extremal values of the function $r : \mathbb{R}^n \setminus \{0\} \to \mathbb{R} : r(\boldsymbol{x}) = \dfrac{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}}$.

Let $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots$ be orthonormal, $\boldsymbol{Q}_j = \begin{bmatrix} \boldsymbol{q}_1 & \ldots & \boldsymbol{q}_j \end{bmatrix}$ and $\boldsymbol{B}_j = \boldsymbol{Q}_j^T \boldsymbol{A} \boldsymbol{Q}_j$. Then the largest and smallest eigenvalue of $\boldsymbol{B}_j$ satisfy:

$$\begin{aligned} m_j =& \lambda_1(\boldsymbol{B}_j) = \min_{\|\boldsymbol{v}\|_2=1} \boldsymbol{v}^T \boldsymbol{Q}_j^T \boldsymbol{A} \boldsymbol{Q}_j \boldsymbol{v} \geq \min_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} = \lambda_1(\boldsymbol{A}) \\ M_j =& \lambda_j(\boldsymbol{B}_j) = \max_{\|\boldsymbol{v}\|_2=1} \boldsymbol{v}^T \boldsymbol{Q}_j^T \boldsymbol{A} \boldsymbol{Q}_j \boldsymbol{v} \leq \max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} = \lambda_n(\boldsymbol{A}) \ . \end{aligned}$$

How should we choose $\boldsymbol{q}_{j+1}$ such that m_{j+1} and M_{j+1} would be clearly better approximations for the smallest and largest eigenvalue of $\boldsymbol{A}$? Since

$$\nabla r(\boldsymbol{x}) = \frac{2\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} - 2\frac{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{(\boldsymbol{x}^T\boldsymbol{x})^2}\boldsymbol{x} = \frac{2}{\boldsymbol{x}^T\boldsymbol{x}}(\boldsymbol{A}\boldsymbol{x} - r(\boldsymbol{x})\boldsymbol{x}) \ ,$$

and if $\boldsymbol{u}_j, , \boldsymbol{v}_j \in \operatorname{span}(\boldsymbol{q}_1, \ldots \boldsymbol{q}_j)$ are such that $r(\boldsymbol{u}_j) = m_j$ and $r(\boldsymbol{v}_j) = M_j$, then we would like to have: $\nabla r(\boldsymbol{u}_j)$, $\nabla r(\boldsymbol{v}_j) \in \operatorname{span}(\boldsymbol{q}_1, \ldots \boldsymbol{q}_{j+1})$. This is possible, if

$$\boldsymbol{A} \operatorname{span}(\boldsymbol{q}_1, \ldots \boldsymbol{q}_j) \subset \operatorname{span}(\boldsymbol{q}_1, \ldots \boldsymbol{q}_{j+1}) \ .$$

This we have, if $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_j$ is an orthonormal base of some Krylov subspace:

$$\begin{aligned} \operatorname{span}(\boldsymbol{q}_1, \ldots \boldsymbol{q}_j) &= \operatorname{span}(\boldsymbol{b}, \boldsymbol{A}\boldsymbol{b}, \ldots, \boldsymbol{A}^{j-1}\boldsymbol{b}) \ , \\ \operatorname{span}(\boldsymbol{q}_1, \ldots \boldsymbol{q}_{j+1}) &= \operatorname{span}(\boldsymbol{b}, \boldsymbol{A}\boldsymbol{b}, \ldots, \boldsymbol{A}^{j}\boldsymbol{b}) \ . \end{aligned}$$

This real symmetric case[1] was just for the motivation for the use of the Krylov subspaces. Now we continue with a general Hermitian matrix $\boldsymbol{A} \in \mathbb{C}^{n\times n}$.

9.1. **Lanczos iteration for a Hermitian operator.** We can obtain an orthonormal basis from the Gram–Schmidt process, i.e., from the $\boldsymbol{QR}$ decomposition. So, let

$$\boldsymbol{K}_j(\boldsymbol{A}, \boldsymbol{b}) = \begin{bmatrix} \boldsymbol{b} & \boldsymbol{A}\boldsymbol{b} & \ldots & \boldsymbol{A}^{j-1}\boldsymbol{b} \end{bmatrix} = \boldsymbol{Q}_j \boldsymbol{R}_j \ .$$

Then

$$\boldsymbol{A}\boldsymbol{Q}_j\boldsymbol{R}_j = \boldsymbol{A}\boldsymbol{K}_j(\boldsymbol{A}, \boldsymbol{b}) = \begin{bmatrix} \boldsymbol{A}\boldsymbol{b} & \boldsymbol{A}^2\boldsymbol{b} & \ldots & \boldsymbol{A}^j\boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Q}_j\boldsymbol{R}_j \begin{bmatrix} 0 \\ \boldsymbol{I} \end{bmatrix} & \boldsymbol{A}^j\boldsymbol{b} \end{bmatrix} \ ,$$

which gives

$$\boldsymbol{Q}_j^*\boldsymbol{A}\boldsymbol{Q}_j = \begin{bmatrix} \boldsymbol{R}_j \begin{bmatrix} 0 \\ \boldsymbol{I} \end{bmatrix} & \boldsymbol{Q}_j^*\boldsymbol{A}^j\boldsymbol{b} \end{bmatrix} \boldsymbol{R}_j^{-1} = \boldsymbol{H}_j\boldsymbol{R}_j^{-1} \ ,$$

where $\boldsymbol{H}_j$ is a Hessenberg matrix. A product of such and an upper triangular matrix is again a Hessenberg matrix, so that $\boldsymbol{T}_j = \boldsymbol{Q}_j^*\boldsymbol{A}\boldsymbol{Q}_j$ is a Hermitian matrix and of the Hessenberg form, i.e., a tridiagonal matrix.

Problem 9.1. ("$\boldsymbol{A}$ is of degree m at $\boldsymbol{b}$.") Let $m = \dim(\mathcal{K}_n(\boldsymbol{A}, \boldsymbol{b}))$. Show:

[1] The complex case can be done similarly, but since $z \to \bar{z}$ is not differentiable, this would have caused extra confusing details.

a) $\boldsymbol{A}\mathcal{K}_m(\boldsymbol{A},\boldsymbol{b}) \subset \mathcal{K}_m(\boldsymbol{A},\boldsymbol{b})$ and $\mathcal{K}_m(\boldsymbol{A},\boldsymbol{b}) = \mathcal{K}_n(\boldsymbol{A},\boldsymbol{b})$
b) There exists a monic[2] polynomial p of degree m such that $p(\boldsymbol{A})\boldsymbol{b} = 0$.
c) $\boldsymbol{A}\boldsymbol{Q}_m = \boldsymbol{Q}_m\boldsymbol{T}_m$.

The Lanczos idea is to compute the vectors $\boldsymbol{q}_j$ and the matrix

$$\boldsymbol{T}_m = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \bar{\beta}_1 & \alpha_2 & \beta_2 & & \\ & \bar{\beta}_2 & \ddots & \ddots & \\ & & \ddots & \alpha_{m-1} & \beta_{m-1} \\ & & & \bar{\beta}_{m-1} & \alpha_m \end{bmatrix}$$

directly. Comparing the columns of the equation $\boldsymbol{A}\boldsymbol{Q}_m = \boldsymbol{Q}_m\boldsymbol{T}_m$ we get:

$$\begin{aligned} \boldsymbol{A}\boldsymbol{q}_1 &= \alpha_1\boldsymbol{q}_1 + \bar{\beta}_1\boldsymbol{q}_2 \\ \boldsymbol{A}\boldsymbol{q}_2 &= \beta_1\boldsymbol{q}_1 + \alpha_2\boldsymbol{q}_2 + \bar{\beta}_2\boldsymbol{q}_3 \\ &\vdots \\ \boldsymbol{A}\boldsymbol{q}_j &= \beta_{j-1}\boldsymbol{q}_{j-1} + \alpha_j\boldsymbol{q}_j + \bar{\beta}_j\boldsymbol{q}_{j+1} \end{aligned}$$

Orthonormality implies: $\alpha_j = \boldsymbol{q}_j^*\boldsymbol{A}\,\boldsymbol{q}_j$ and $\boldsymbol{q}_{j+1}$ has to be a unit vector in the direction of

$$\boldsymbol{r}_j = \boldsymbol{A}\boldsymbol{q}_j - \beta_{j-1}\boldsymbol{q}_{j-1} - \alpha_j\boldsymbol{q}_j \,.$$

β_j can be chosen to be real: $\beta_j = \|\boldsymbol{r}_j\|_2^{-1}$. Continue this way until $\boldsymbol{r}_j = 0$.

Problem 9.2. Show: $\boldsymbol{r}_j \neq 0$, when $j < m$ and $\boldsymbol{r}_m = 0$.

We get the Lanczos iteration:

```
$\beta_0\,\boldsymbol{q}_0 = 0$, $\boldsymbol{q}_1 = \boldsymbol{b}/\,\|\boldsymbol{b}\|_2$
for    $j = 1, 2, \ldots$
       $\alpha_j = \boldsymbol{q}_j^*\boldsymbol{A}\,\boldsymbol{q}_j$
       $\boldsymbol{r}_j = \boldsymbol{A}\boldsymbol{q}_j - \beta_{j-1}\boldsymbol{q}_{j-1} - \alpha_j\boldsymbol{q}_j$
       $\beta_j = \|\boldsymbol{r}_j\|_2$
       if $\beta_j = 0$ stop
           else  $\boldsymbol{q}_{j+1} = \boldsymbol{r}_j/\beta_j$
end
```

In the end of the iteration $(j = m)$ we have $\boldsymbol{A}\boldsymbol{Q}_m = \boldsymbol{Q}_m\boldsymbol{T}_m$, so that $\Lambda(\boldsymbol{T}_m) \subset \Lambda(\boldsymbol{A})$. To obtain the Schur decomposition of the real tridiagonal matrix $\boldsymbol{T}_m$ we can use the symmetric $\boldsymbol{QR}$ iteration or the divide and conquer method.

[2] Monic: the coefficient of the highest power is one.

The main advantage of the Lanczos iteration is that the eigenvalues of $\boldsymbol{T}_j$ are good approximations of the eigenvalues of $\boldsymbol{A}$ (especially of the largest and the smallest) often already for $j \ll m$. The next result shows that if β_j is small, then $\boldsymbol{A}\boldsymbol{Q}_j \approx \boldsymbol{Q}_j\boldsymbol{T}_j$,

Proposition 9.1. *Above we have:* $\boldsymbol{A}\boldsymbol{Q}_j = \boldsymbol{Q}_j\boldsymbol{T}_j + \boldsymbol{r}_j\boldsymbol{e}_j^T$.

Proof.

$$\boldsymbol{K}_{j+1}(\boldsymbol{A},\boldsymbol{b}) = \begin{bmatrix}\boldsymbol{b} & \boldsymbol{A}\boldsymbol{K}_j(\boldsymbol{A},\boldsymbol{b})\end{bmatrix} = \begin{bmatrix}\boldsymbol{b} & \boldsymbol{A}\boldsymbol{Q}_j\boldsymbol{R}_j\end{bmatrix} = \boldsymbol{Q}_{j+1}\boldsymbol{R}_{j+1} \ ,$$

where $\boldsymbol{Q}_{j+1} = \begin{bmatrix}\boldsymbol{Q}_j & \boldsymbol{q}_{j+1}\end{bmatrix}$, $\boldsymbol{R}_{j+1} = \begin{bmatrix}\boldsymbol{R}_j & \boldsymbol{w}_j \\ 0 & \rho_{j+1}\end{bmatrix}$. We get:

$$\boldsymbol{A}\boldsymbol{Q}_j = \boldsymbol{Q}_j\boldsymbol{H}_j\boldsymbol{R}_j^{-1} + \rho_{j+1}\boldsymbol{q}_{j+1}\boldsymbol{e}_j^T\boldsymbol{R}_j^{-1} = \boldsymbol{Q}_j\boldsymbol{T}_j + \frac{\rho_{j+1}}{\rho_j}\boldsymbol{q}_{j+1}\boldsymbol{e}_j^T \ .$$

Comparing the last columns of this to the equation

$$\boldsymbol{A}\boldsymbol{q}_j = \beta_{j-1}\boldsymbol{q}_{j-1} + \alpha_j\boldsymbol{q}_j + \beta_j\boldsymbol{q}_{j+1}$$

we see: $\frac{\rho_{j+1}}{\rho_j}\boldsymbol{q}_{j+1} = \beta_j\boldsymbol{q}_{j+1} = \boldsymbol{r}_j$, which shows the claim. □

From this we further get the error of the eigenvalue equation:

Proposition 9.2. *If* $\boldsymbol{T}_j = \boldsymbol{S}_j \operatorname{diag}(\mu_1,\ldots,\mu_j)\boldsymbol{S}_j^*$ *is the Schur decomposition of* $\boldsymbol{T}_j$, *and* $\boldsymbol{Y} = \begin{bmatrix}\boldsymbol{y}_1 & \ldots & \boldsymbol{y}_j\end{bmatrix} = \boldsymbol{Q}_j\boldsymbol{S}_j$, *then*

$$\|\boldsymbol{A}\boldsymbol{y}_k - \mu_k\boldsymbol{y}_k\|_2 = \|\boldsymbol{r}_j\|_2\,|s_{j,k}| \ .$$

Proof.

$$\boldsymbol{A}\boldsymbol{Y}_j = \boldsymbol{Y}_j \operatorname{diag}(\mu_1,\ldots,\mu_j) + \boldsymbol{r}_j\boldsymbol{e}_j^T\boldsymbol{S}_j \ ,$$

in particular, $\boldsymbol{A}\boldsymbol{y}_k = \mu_k\boldsymbol{y}_k + \boldsymbol{r}_j\boldsymbol{e}_j^T\boldsymbol{S}_j\boldsymbol{e}_k$. □

Problem 9.3. Show, that above we get: $\delta(\mu_k, \Lambda(\boldsymbol{A})) \le \beta_j\,|s_{j,k}|$.

The convergence theorems of the Lanczos iteration are collected under the title *Kaniel–Paige* theory. The following is a sample of them. It tells that the convergence of the extremal eigenvalues is fast.

Theorem 9.3. *Let* $\boldsymbol{v}_1$ *be the unit eigenvector of* $\boldsymbol{A}$ *correspoding to* λ_1 *and* $\phi_1 = \arccos(|\boldsymbol{q}_1^*\boldsymbol{v}_1|)$. *Then for the iteration above we have:*

$$0 \le \mu_1 - \lambda_1 \le \frac{(\lambda_n - \lambda_1)\tan(\phi_1)^2}{t_{j-1}(1+\delta_1)^2} \ ,$$

where $\delta_1 = 2\frac{\lambda_2-\lambda_1}{\lambda_n-\lambda_2}$ *and* t_{j-1} *is the Tshebyshev polynomial* [3] *of degree* $j-1$.

Proof.

$$\mu_1 = \min_{\boldsymbol{y}\neq 0} \frac{\boldsymbol{y}^*\boldsymbol{T}_j\boldsymbol{y}}{\boldsymbol{y}^*\boldsymbol{y}} = \min_{\boldsymbol{y}\neq 0} \frac{\boldsymbol{y}^*\boldsymbol{Q}_j^*\boldsymbol{A}\boldsymbol{Q}_j\boldsymbol{y}}{(\boldsymbol{Q}_j\boldsymbol{y})^*\boldsymbol{Q}_j\boldsymbol{y}} = \min_{0\neq\boldsymbol{w}\in\mathcal{K}_j(\boldsymbol{A},\boldsymbol{b})} \frac{\boldsymbol{w}^*\boldsymbol{A}\boldsymbol{w}}{\boldsymbol{w}^*\boldsymbol{w}} \geq \min_{0\neq\boldsymbol{w}} \frac{\boldsymbol{w}^*\boldsymbol{A}\boldsymbol{w}}{\boldsymbol{w}^*\boldsymbol{w}} = \lambda_1 \ .$$

Every $\boldsymbol{w} \in \mathcal{K}_j(\boldsymbol{A},\boldsymbol{b})$ is of the form $p(\boldsymbol{A})\boldsymbol{b}$, where $p(\boldsymbol{A})$ is a polynomial of $\boldsymbol{A}$ of degree at most $j-1$ (denote: $p \in \mathbb{P}_{j-1}$), so that

$$\mu_1 = \min_{0\neq p\in\boldsymbol{P}_{j-1}} \frac{\boldsymbol{b}^*p(\boldsymbol{A})^*\boldsymbol{A}p(\boldsymbol{A})\boldsymbol{b}}{\boldsymbol{b}^*p(\boldsymbol{A})^*p(\boldsymbol{A})\boldsymbol{b}} \ .$$

If $\boldsymbol{b} = \sum_{k=1}^n c_k\boldsymbol{v}_k$, where $\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n$ are the orthonormal eigenvectors of $\boldsymbol{A}$, then

$$p(\boldsymbol{A})\boldsymbol{q}_1 = \sum_{k=1}^n c_kp(\lambda_k)\,\boldsymbol{v}_k \ ,$$

so that for arbitrary $p \in \mathbb{P}_{j-1}\setminus\{0\}$ we get

$$\begin{aligned}\mu_1 &\leq \frac{\boldsymbol{q}_1^*p(\boldsymbol{A})^*\boldsymbol{A}p(\boldsymbol{A})\boldsymbol{q}_1}{\boldsymbol{q}_1^*p(\boldsymbol{A})^*p(\boldsymbol{A})\boldsymbol{q}_1} = \frac{\sum_{k=1}^n |c_kp(\lambda_k)|^2\,\lambda_k}{\sum_{k=1}^n |c_kp(\lambda_k)|^2} \\ &\leq \frac{|c_1p(\lambda_1)|^2\,\lambda_1 + \lambda_n\sum_{k=2}^n |c_kp(\lambda_k)|^2}{|c_1p(\lambda_1)|^2 + \sum_{k=2}^n |c_kp(\lambda_k)|^2} = \lambda_1 + (\lambda_n-\lambda_1)\frac{\sum_{k=2}^n |c_kp(\lambda_k)|^2}{\sum_{k=1}^n |c_kp(\lambda_k)|^2} \ .\end{aligned}$$

Now, choosing $p(x) = t_{j-1}(2\frac{\lambda_n-x}{\lambda_n-\lambda_2}-1)$, we have $|p(\lambda_k)| \leq 1$, $k = 2,\ldots,n$, and

$$\mu_1 \leq \lambda_1 + (\lambda_n-\lambda_1)\frac{1-|c_1|^2}{|c_1|^2\left|t_{j-1}(2\frac{\lambda_n-\lambda_1}{\lambda_n-\lambda_2}-1)\right|^2} = \lambda_1 + \frac{(\lambda_n-\lambda_1)\tan(\phi_1)^2}{t_{j-1}(1+\delta_1)^2} \ .$$

Here we used: $|c_1| = \cos(\phi_1)$. □

Similarly we can show that

$$0 \leq \lambda_n - \mu_j \leq \frac{(\lambda_n-\lambda_1)\tan(\phi_n)^2}{t_{j-1}(1+\delta_n)^2} \ ,$$

where $\phi_n = \arccos(|\boldsymbol{q}_1^*\boldsymbol{v}_n|)$ and $\delta_n = 2\frac{\lambda_n-\lambda_{n-1}}{\lambda_{n-1}-\lambda_1}$.

Problem 9.4. Compute how fast the numbers $t_j(1+\delta)$ grow with j. Answer: $\sim \frac{1}{2}(1+\delta+\sqrt{2\delta+\delta^2}\,)^j$. Hint: $t_{j+1}(x) = 2x\,t_j(x) - t_{j-1}(x)$.

[3] Tsebyshev polynomials are defined by $t_k(x) = \cos(k\arccos(x))$, (when $|x| \leq 1$). More about these later.

So, the Lanczos iteration seems to be an ideal method. The bad news then is that this is true only for exact arithmetic. Including the round–off errors changes the situation essentially. It is not difficult to show that for the computed vectors $\widehat{\boldsymbol{q}}_j$, $\widehat{\boldsymbol{r}}_,$ and for the tridiagonal $\widehat{\boldsymbol{T}}_j$ the error in the equation

$$\boldsymbol{A}\widehat{\boldsymbol{Q}}_j = \widehat{\boldsymbol{Q}}_j\widehat{\boldsymbol{T}}_j + \widehat{\boldsymbol{r}}_j\boldsymbol{e}_j^T + \boldsymbol{E}_j$$

usually is moderate: $\|\boldsymbol{E}\|_2 \approx \mu\,\|\boldsymbol{A}\|_2$. The problem is that due to the round–off errors, the computed $\widehat{\boldsymbol{Q}}_j$ loose the orthogonality of the columns quickly. This is because of the equation $\boldsymbol{r}_j$

$$\boldsymbol{r}_j = \boldsymbol{A}\boldsymbol{q}_j - \alpha_j\boldsymbol{q}_j - \beta_{j-1}\boldsymbol{q}_{j-1}\ ,$$

where in the computation the round–off errors are of the size $\mu\,\|\boldsymbol{A}\boldsymbol{q}_j\|_2$. When $\beta_j = \|\boldsymbol{r}_j\|_2$ itself is small, then, for example,

$$\widehat{\boldsymbol{q}}_j^*\,\widehat{\boldsymbol{q}}_{j+1} = \widehat{\widehat{\boldsymbol{q}}_j^*\,\widehat{\boldsymbol{r}}_j/\widehat{\beta}_j} \approx \frac{\mu\,\|\boldsymbol{A}\boldsymbol{q}_j\|_2}{\widehat{\beta}_j}\ .$$

This loss of orthogonality seems to happen hand in hand with the convergence.

As a remedy for this a reorthogonalization strategy is proposed, i.e., make sure that the computed $\widehat{\boldsymbol{q}}_{j+1}$ is orthogonal to the previous ones by subtracting from it the components in their directions. The algorithm obtained this way is equivalent to the Arnoldi iteration (introduced later), that works also for non–Hermitian matrices. But now we would loose the advantages brought by Hermitianity.

A more detailed analysis shows, that the errors in the vectors $\widehat{\boldsymbol{q}}_j$ are mainly in the direction of the already converged $\boldsymbol{A}$ eigenvectors. Then it suffices to remove these components, i.e.,ortogonalize only against all the converged eigenvectors. This carries the name *selective reorthogonalization* in the literature.

Problem 9.5. Test the Lanczos iteration for a real symmetric matrix starting with an almost eigenvector. Does the reorthogonalization help?

9.2. **Bi-orthogonal Lanczos.** The Lanczos iteration for a Hermitian matrix started with the attempt to compute a unitary similar transformation to a Hermitian tridiagonal form: $\boldsymbol{Q}^*\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{T}$. Let $\boldsymbol{A}$ now be a general square matrix and let us try to compute a similarity transformation to a tridiagonal form:

$$\boldsymbol{V}^{-1}\boldsymbol{A}\boldsymbol{V} = \boldsymbol{T} = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \gamma_1 & \alpha_2 & \beta_2 & & \\ & \gamma_2 & \ddots & \ddots & \\ & & \ddots & \alpha_{n-1} & \beta_{n-1} \\ & & & \gamma_{n-1} & \alpha_n \end{bmatrix}\ .$$

Denote $\boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1 & \dots & \boldsymbol{v}_n \end{bmatrix}$, $\boldsymbol{V}^{-1} = \begin{bmatrix} \boldsymbol{w}_1 & \dots & \boldsymbol{w}_n \end{bmatrix}^*$, so that $\boldsymbol{w}_j^* \boldsymbol{v}_k = \delta_{j,k}$. Rewrite the equations

$$\boldsymbol{AV} = \boldsymbol{VT} \qquad \text{and} \qquad \boldsymbol{A}^*\boldsymbol{W} = \boldsymbol{WT}^*$$

columnwise:

$$\begin{aligned}
\boldsymbol{Av}_1 &= \alpha_1\boldsymbol{v}_1 + \gamma_1\boldsymbol{v}_2 \\
\boldsymbol{A}^*\boldsymbol{w}_1 &= \bar{\alpha}_1\boldsymbol{w}_1 + \bar{\beta}_1\boldsymbol{w}_2 \\
\boldsymbol{Av}_2 &= \beta_1\boldsymbol{v}_1 + \alpha_2\boldsymbol{v}_2 + \gamma_2\boldsymbol{v}_3 \\
\boldsymbol{A}^*\boldsymbol{w}_2 &= \bar{\gamma}_1\boldsymbol{w}_1 + \bar{\alpha}_2\boldsymbol{w}_2 + \bar{\beta}_2\boldsymbol{w}_3 \\
&\vdots \\
\boldsymbol{Av}_j &= \beta_{j-1}\boldsymbol{v}_{j-1} + \alpha_j\boldsymbol{v}_j + \gamma_j\boldsymbol{v}_{j+1} \\
\boldsymbol{A}^*\boldsymbol{w}_j &= \bar{\gamma}_{j-1}\boldsymbol{w}_{j-1} + \bar{\alpha}_j\boldsymbol{w}_j + \bar{\beta}_j\boldsymbol{w}_{j+1}
\end{aligned}$$

Multiplying the $\boldsymbol{Av}_j$ –equation by $\boldsymbol{w}_j^*$ gives $\alpha_j = \boldsymbol{w}_j^*\boldsymbol{Av}_j$. Denote

$$\begin{aligned}
\boldsymbol{r}_j &= \boldsymbol{Av}_j - \alpha_j\boldsymbol{v}_j - \beta_{j-1}\boldsymbol{v}_{j-1} \\
\boldsymbol{s}_j &= \boldsymbol{A}^*\boldsymbol{w}_j - \bar{\alpha}_j\boldsymbol{w}_j - \bar{\gamma}_{j-1}\boldsymbol{w}_{j-1} \ ,
\end{aligned}$$

so that

$$\boldsymbol{v}_{j+1} = \boldsymbol{r}_j/\gamma_j \ , \qquad \boldsymbol{w}_{j+1} = \boldsymbol{s}_j/\bar{\beta}_j \ .$$

For γ_j and β_j we get the condition:

$$1 = \boldsymbol{w}_{j+1}^*\boldsymbol{v}_{j+1} = \frac{\boldsymbol{s}_j^*\boldsymbol{r}_j}{\beta_j\gamma_j} \ .$$

This does not determine γ_j and β_j uniquely. The usual choice is $\gamma_j = |\boldsymbol{s}_j^*\boldsymbol{r}_j|^{1/2}$, which causes β_j to have the same absolute value $\beta_j = \frac{\boldsymbol{s}_j^*\boldsymbol{r}_j}{\gamma_j}$.

The iteration can be continued until $\boxed{\boldsymbol{s}_j^*\boldsymbol{r}_j = 0}$. Then, if $\boldsymbol{r}_j = 0$, we are in a good situation, since then holds: $\boldsymbol{AV}_j = \boldsymbol{V}_j\boldsymbol{T}_j$, where $\boldsymbol{V}_j = \begin{bmatrix} \boldsymbol{v}_1 & \dots & \boldsymbol{v}_j \end{bmatrix}$ and $\boldsymbol{T}_j$ is the upper left $j \times j$ corner of $\boldsymbol{T}$ and $\Lambda(\boldsymbol{T}_j) \subset \Lambda(\boldsymbol{A})$. Further, from the eigenvectors $\boldsymbol{y}_k$ of $\boldsymbol{T}_j$ we get eigenvectors $\boldsymbol{V}_j\boldsymbol{y}_k$ of $\boldsymbol{A}$. Similarly for $\boldsymbol{A}^*$, if $\boldsymbol{s}_j = 0$.

The situation becomes more complicated if $\boldsymbol{s}_j^*\boldsymbol{r}_j = 0$ without either of them being nullvector. Then the strategy is to try a bigger step. Release the requirement that $\boldsymbol{T}$ should be a tridiagonal and let it be “thicker” at some places. This strategy is called the *Look ahead Lanczos* iteration and it is a subject of active current research.

The loss of orthogonality ($\boldsymbol{w}_j^*\boldsymbol{v}_k = \delta_{j,k}$) caused by the round-off errors is also here a problem and this is again tried to be remedied by (selective) reorthogonalization.

9.3. **Arnoldi iteration.** The most reliable Krylov subspace iteration is based on the computation of an orthonormal basis for $\mathcal{K}_j(\boldsymbol{A},\boldsymbol{b})$, similarly to the Lanczos iteration for Hermitian matrices.

Think again that we would perform the $\boldsymbol{QR}$ decompositions to the Krylov matrices

$$\boldsymbol{K}_j(\boldsymbol{A},\boldsymbol{b}) = \begin{bmatrix}\boldsymbol{b} & \boldsymbol{Ab} & \dots & \boldsymbol{A}^{j-1}\boldsymbol{b}\end{bmatrix} = \boldsymbol{Q}_j\boldsymbol{R}_j \ .$$

Then

$$\boldsymbol{K}_{j+1}(\boldsymbol{A},\boldsymbol{b}) = \begin{bmatrix}\boldsymbol{b} & \boldsymbol{AK}_j(\boldsymbol{A},\boldsymbol{b})\end{bmatrix} = \begin{bmatrix}\boldsymbol{b} & \boldsymbol{AQ}_j\boldsymbol{R}_j\end{bmatrix} = \boldsymbol{Q}_{j+1}\boldsymbol{R}_{j+1} \ ,$$

where

$$\boldsymbol{Q}_{j+1} = \begin{bmatrix}\boldsymbol{Q}_j & \boldsymbol{q}_{j+1}\end{bmatrix} \ , \qquad \boldsymbol{R}_{j+1} = \begin{bmatrix}\boldsymbol{R}_j & \boldsymbol{w}_j \\ 0 & \rho_{j+1}\end{bmatrix} = \left[\begin{array}{cc} r_{1,1} & \\ 0 & \widetilde{\boldsymbol{H}}_j \\ \vdots & \\ 0 & \\ 0 & 0 \dots 0 \ \rho_{j+1}\end{array}\right] .$$

We get:

$$\boldsymbol{AQ}_j\boldsymbol{R}_j = \boldsymbol{Q}_j\widetilde{\boldsymbol{H}}_j + \rho_{j+1}\boldsymbol{q}_{j+1}\boldsymbol{e}_j^T \ ,$$

i.e.,

$$\boldsymbol{AQ}_j = \boldsymbol{Q}_j\widetilde{\boldsymbol{H}}_j\boldsymbol{R}_j^{-1} + \rho_{j+1}\boldsymbol{q}_{j+1}\boldsymbol{e}_j^T/\rho_j = \boldsymbol{Q}_j\boldsymbol{H}_j + h_{j+1,j}\boldsymbol{q}_{j+1}\boldsymbol{e}_j^T \ ,$$

where $\boldsymbol{H}_j = \widetilde{\boldsymbol{H}}_j\boldsymbol{R}_j^{-1}$ is a Hessenberg matrix and $h_{j+1,j} = \rho_{j+1}/\rho_j$. As in the problem 9.1 we get again: if $m = \dim(\mathcal{K}_n(\boldsymbol{A},\boldsymbol{b}))$, then

a) $\boldsymbol{A}\mathcal{K}_m(\boldsymbol{A},\boldsymbol{b}) \subset \mathcal{K}_m(\boldsymbol{A},\boldsymbol{b})$ and $\mathcal{K}_m(\boldsymbol{A},\boldsymbol{b}) = \mathcal{K}_n(\boldsymbol{A},\boldsymbol{b})$

b) If $\boldsymbol{A}$ is invertible, then there exists a polynomial p of degree m such that $p(0) = 1$ and $p(\boldsymbol{A})\,\boldsymbol{b} = 0$.

c) $\boldsymbol{AQ}_m = \boldsymbol{Q}_m\boldsymbol{H}_m$.

In the Arnoldi iteration the vectors $\boldsymbol{q}_j$ and the Hessenberg matrix

$$\boldsymbol{H}_m = \begin{bmatrix} h_{1,1} & h_{1,2} & h_{1,3} & \dots & h_{1,m} \\ h_{2,1} & h_{2,2} & h_{2,3} & \dots & h_{2,m} \\ 0 & h_{3,2} & h_{3,3} & \dots & h_{3,m} \\ & & \ddots & & \vdots \\ 0 & 0 & 0 & \dots & h_{m,m} \end{bmatrix}$$

are computed directly, without performing the $\boldsymbol{QR}$ decomposition of $\boldsymbol{K}_j(\boldsymbol{A},\boldsymbol{b})$. From equation $\boldsymbol{Aq}_j = \boldsymbol{Q}_j\boldsymbol{w}_j + h_{j+1,j}\boldsymbol{q}_{j+1}$ we see, that $\boldsymbol{w}_j = \boldsymbol{Q}_j^*\boldsymbol{Aq}_j$.

We get the Arnoldi iteration (with modified Gram–Schmidt):

```
$\boldsymbol{q}_1 = \boldsymbol{b}/\left\|\boldsymbol{b}\right\|_2$ , $\boldsymbol{Q}_1 = \left[\boldsymbol{q}_1\right]$
for    $j = 1, 2, \ldots$
        $\boldsymbol{r}_j = \boldsymbol{A}\boldsymbol{q}_j$
        for  $k = 1, ..., j$ ,   $h_{k,j} = \boldsymbol{q}_k^* \boldsymbol{r}_j$ ,   $\boldsymbol{r}_j = \boldsymbol{r}_j - h_{k,j}\,\boldsymbol{q}_k$   end
        $h_{j+1,j} = \|\boldsymbol{r}_j\|_2$
        if $h_{j+1,j} = 0$ stop  else  $\boldsymbol{q}_{j+1} = \boldsymbol{r}_j / h_{j+1,j}$ ,      $\boldsymbol{Q}_{j+1} = \left[\boldsymbol{Q}_j \quad \boldsymbol{q}_{j+1}\right]$
end
```

In the end of the iteration $(j{=}m)$ we have $\boldsymbol{A}\boldsymbol{Q}_m = \boldsymbol{Q}_m\boldsymbol{H}_m$, so that $\Lambda(\boldsymbol{H}_m) \subset \Lambda(\boldsymbol{A})$.

Modified Gram–Schmidt means that the operation $\boldsymbol{r}_j = (\boldsymbol{I} - \boldsymbol{Q}_j\boldsymbol{Q}_j^*)\boldsymbol{A}\boldsymbol{q}_j$ is computed such that the components of $\boldsymbol{r}_j^0 = \boldsymbol{A}\boldsymbol{q}_j$ in the directions of each $\boldsymbol{q}_k$ are removed one at a time. This way the non-orthogonality that already exists in the vectors $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_j$ causes less errors.

Arnoldi iteration is commonly used in situations where we are not interested in all the eigenvalues of $\boldsymbol{A}$ but only the "outer" ones hoping that the eigenvalues of $\boldsymbol{H}_j$ would approximate these well already for small values of j like in the Lanczos iteration in the Hermitian case. If

$$\boldsymbol{H}_j\boldsymbol{y}_k^{(j)} = \mu_k^{(j)}\boldsymbol{y}_k^{(j)} \ , \qquad \|\boldsymbol{y}_k^{(j)}\|_2 = 1 \ ,$$

then the numbers $\mu_k^{(j)}$ are called the *Ritz values* and the vectors $\boldsymbol{u}_k^{(j)} = \boldsymbol{Q}_j\boldsymbol{y}_k^{(j)}$ the *Ritz vectors.* These are approximate eigenvalues and eigenvectors of $\boldsymbol{A}$. Equation

$$\boldsymbol{A}\boldsymbol{Q}_j = \boldsymbol{Q}_j\boldsymbol{H}_j + h_{j+1,j}\boldsymbol{q}_{j+1}\boldsymbol{e}_j^T$$

gives now

$$\left\|\boldsymbol{A}\boldsymbol{u}_k^{(j)} - \mu_k^{(j)}\boldsymbol{u}_k^{(j)}\right\|_2 = |h_{j+1,j}| \left|(\boldsymbol{y}_k^{(j)})_j\right|$$

and the Bauer–Fike theorem 6.4 gives an error bound for the Ritz values:

Theorem 9.4. *If* $\boldsymbol{A}$ *is diagonalizable :* $\boldsymbol{V}^{-1}\boldsymbol{A}\boldsymbol{V} = \boldsymbol{D}$, *then*

$$\delta(\mu_k^{(j)}, \Lambda(\boldsymbol{A})) \le \kappa_2(\boldsymbol{V})\, |h_{j+1,j}| \left|(\boldsymbol{y}_k^{(j)})_j\right| \ .$$

In the sequel denote by $\boldsymbol{P}_j = \boldsymbol{Q}_j\boldsymbol{Q}_j^*$ the orthogonal projection onto $\mathcal{K}_j(\boldsymbol{A}, \boldsymbol{b})$.

For every $\lambda_k \in \Lambda(\boldsymbol{A})$ define:

$$\varepsilon_k^{(j)} = \min_{\substack{p \in \mathbb{P}_{j-1} \\ p(\lambda_k)=1}} \ \max_{\lambda \in \Lambda(\boldsymbol{A}) \setminus \{\lambda_k\}} |p(\lambda)| \ .$$

So, $\varepsilon_k^{(j)}$ measures, how small a polynomial of degree ${\scriptstyle j-1}$ can become on the rest of the spectrum of $\boldsymbol{A}$ when at λ_k it has the value one.

Lemma 9.5. *Let* $\boldsymbol{A}$ *be diagonalizable :* $\boldsymbol{A v}_k - \lambda_k \boldsymbol{v}_k$, $\|\boldsymbol{v}_k\|_2 = 1$, $k = 1, \ldots, n$, $\det(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n) \neq 0$. *Let* $\boldsymbol{q}_1 = \sum_{k=1}^n c_k \boldsymbol{v}_k$. *Then*

$$\|\boldsymbol{v}_k - \boldsymbol{P}_j \boldsymbol{v}_k\|_2 \leq \frac{\varepsilon_k^{(j)}}{|c_k|} \sum_{\substack{i=1 \\ i \neq k}}^n |c_i| \ .$$

Proof.

$$\begin{aligned} \|(\boldsymbol{I} - \boldsymbol{P}_j) c_k \boldsymbol{v}_k\|_2 &= \min_{p \in \mathbb{P}_{j-1}} \|c_k \boldsymbol{v}_k - p(\boldsymbol{A}) \boldsymbol{q}_1\|_2 \\ &= \min_{p \in \mathbb{P}_{j-1}} \left\| c_k \boldsymbol{v}_k - \sum_{i=1}^n c_i p(\lambda_i) \boldsymbol{v}_i \right\|_2 \\ &\leq \min_{\substack{p \in \mathbb{P}_{j-1} \\ p(\lambda_k)=1}} \left\| \sum_{\substack{i=1 \\ i \neq k}}^n c_i p(\lambda_i) \boldsymbol{v}_i \right\|_2 \leq \varepsilon_k^{(j)} \sum_{\substack{i=1 \\ i \neq k}}^n |c_i| \end{aligned}$$

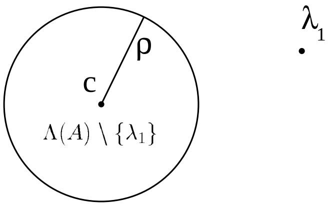
□

Each eigenvector $\boldsymbol{v}_k$ is approximately in $\mathcal{K}_j(\boldsymbol{A}, \boldsymbol{b})$ if the correspoding $\varepsilon_k^{(j)}$ is small. For example, if $\Lambda(\boldsymbol{A}) \setminus \{\lambda_1\}$ is in the disc of radius ρ and centered at c and if λ_1 outside the disc then using the polynomial

$$p(z) = (z - c)^j / (\lambda_1 - c)^j$$

we get

$$\varepsilon_1^{(j+1)} \leq \frac{\rho^j}{|\lambda_1 - c|^j} \ .$$



Denote: $\boldsymbol{A}_j = \boldsymbol{P}_j \boldsymbol{A}_{|\mathcal{K}_j(\boldsymbol{A}, \boldsymbol{b})}$.

Lemma 9.6. $p \in \mathbb{P}_{j-1} \implies p(\boldsymbol{A})\, \boldsymbol{b} = p(\boldsymbol{A}_j)\, \boldsymbol{b}$ *and* $p \in \mathbb{P}_j \implies \boldsymbol{P}_j p(\boldsymbol{A})\, \boldsymbol{b} = p(\boldsymbol{A}_j)\, \boldsymbol{b}$.

Proof. If $0 \leq i < j$, then $\boldsymbol{A}^i \boldsymbol{b} \in \mathcal{K}_j(\boldsymbol{A}, \boldsymbol{b})$, so that $\boldsymbol{P}_j \boldsymbol{A}^i \boldsymbol{b} = \boldsymbol{A}^i \boldsymbol{b}$. We get recursively:

$$\begin{aligned} \boldsymbol{A}_j^i \boldsymbol{b} &= \boldsymbol{P}_j \boldsymbol{A} \ldots \boldsymbol{P}_j \boldsymbol{A} \boldsymbol{P}_j \boldsymbol{A} \boldsymbol{P}_j \boldsymbol{A} \boldsymbol{b} = \boldsymbol{P}_j \boldsymbol{A} \ldots \boldsymbol{P}_j \boldsymbol{A} \boldsymbol{P}_j \boldsymbol{A}^2 \boldsymbol{b} \\ &= \boldsymbol{P}_j \boldsymbol{A} \ldots \boldsymbol{P}_j \boldsymbol{A}^3 \boldsymbol{b} = \boldsymbol{P}_j \boldsymbol{A}^i \boldsymbol{b} = \boldsymbol{A}^i \boldsymbol{b} \ . \end{aligned}$$

Finally: $\boldsymbol{A}_j^j \boldsymbol{b} = \boldsymbol{P}_j \boldsymbol{A}^j \boldsymbol{b}$. □

The following theorem tells something about, where the Ritz values tend to be.

Theorem 9.7. *The characteristic polynomial* $\widetilde{p}_j$ *of* $\boldsymbol{H}_j$ *solves the minimization problem*

$$\|\widetilde{p}_j(\boldsymbol{A})\boldsymbol{b}\|_2 = \min_{\substack{p\in\mathbb{P}_j \\ p\ monic}} \|p(\boldsymbol{A})\boldsymbol{b}\|_2 \ .$$

Moreover, this minimal value is $\left\|(\boldsymbol{I}-\boldsymbol{Q}_j\boldsymbol{Q}_j^*)\boldsymbol{A}^j\boldsymbol{b}\right\|_2 = |\rho_{j+1}| = \|\boldsymbol{b}\|_2 \prod_{k=1}^{j} |h_{k+1,k}|$.

Proof. Since

$$\boldsymbol{A}_j = \boldsymbol{P}_j\boldsymbol{A}_{|\mathcal{K}_j(\boldsymbol{A},\boldsymbol{b})} = \boldsymbol{P}_j\boldsymbol{A}\boldsymbol{P}_{j\,|\mathcal{K}_j(\boldsymbol{A},\boldsymbol{b})} = \boldsymbol{Q}_j\boldsymbol{Q}_j^*\boldsymbol{A}\boldsymbol{Q}_j\boldsymbol{Q}^*_{j\,|\mathcal{K}_j(\boldsymbol{A},\boldsymbol{b})} = \boldsymbol{Q}_j\boldsymbol{H}_j\boldsymbol{Q}^*_{j\,|\mathcal{K}_j(\boldsymbol{A},\boldsymbol{b})}$$

and because $\boldsymbol{Q}_j = \left(\boldsymbol{Q}^*_{j\,|\mathcal{K}_j(\boldsymbol{A},\boldsymbol{b})}\right)^{-1}$, the matrices $\boldsymbol{A}_j$ and $\boldsymbol{H}_j$ are similar. By the Cayley–Hamilton theorem we have: $\widetilde{p}_j(\boldsymbol{A}_j) = 0$. Hence for every $\boldsymbol{y} \in \mathcal{K}_j(\boldsymbol{A},\boldsymbol{b})$ we get using the lemma above:

$$0 = \langle \boldsymbol{y}, \widetilde{p}_j(\boldsymbol{A}_j)\boldsymbol{b}\rangle = \langle \boldsymbol{y}, \boldsymbol{P}_j\widetilde{p}_j(\boldsymbol{A})\boldsymbol{b}\rangle = \langle \boldsymbol{P}_j\boldsymbol{y}, \widetilde{p}_j(\boldsymbol{A})\boldsymbol{b}\rangle = \langle \boldsymbol{y}, \widetilde{p}_j(\boldsymbol{A})\boldsymbol{b}\rangle \ .$$

In other words $\widetilde{p}_j(\boldsymbol{A})\boldsymbol{b} \perp \mathcal{K}_j(\boldsymbol{A},\boldsymbol{b})$. If p is monic and of degree j , then $p - \widetilde{p}_j \in \mathbb{P}_{j-1}$, and $(p-\widetilde{p}_j)(\boldsymbol{A})\boldsymbol{b} \in \mathcal{K}_j(\boldsymbol{A},\boldsymbol{b})$, so that by the Pythagorean theorem:

$$\|p(\boldsymbol{A})\boldsymbol{b}\|_2^2 = \|\widetilde{p}_j(\boldsymbol{A})\boldsymbol{b} + (p-\widetilde{p}_j)(\boldsymbol{A})\boldsymbol{b}\|_2^2 = \|\widetilde{p}_j(\boldsymbol{A})\boldsymbol{b}\|_2^2 + \|(p-\widetilde{p}_j)(\boldsymbol{A})\boldsymbol{b}\|_2^2 \geq \|\widetilde{p}_j(\boldsymbol{A})\boldsymbol{b}\|_2^2 \ .$$

Hence $\widetilde{p}_j$ solves the minimization problem.

Since $\widetilde{p}_j(\boldsymbol{A})\boldsymbol{b} = \boldsymbol{A}^j\boldsymbol{b} - \widetilde{r}_{j-1}(\boldsymbol{A})\boldsymbol{b} \perp \mathcal{K}_j(\boldsymbol{A},\boldsymbol{b})$, where $\widetilde{r}_{j-1}(\boldsymbol{A})\boldsymbol{b} \in \mathcal{K}_j(\boldsymbol{A},\boldsymbol{b})$, we notice, that $\widetilde{r}_{j-1}(\boldsymbol{A})\boldsymbol{b}$ is the orthogonal projection of $\boldsymbol{A}^j\boldsymbol{b}$ onto $\mathcal{K}_j(\boldsymbol{A},\boldsymbol{b})$. In other words $\widetilde{r}_{j-1}(\boldsymbol{A})\boldsymbol{b} = \boldsymbol{Q}_j\boldsymbol{Q}_j^*\boldsymbol{A}^j\boldsymbol{b}$. Further, $\boldsymbol{A}^j\boldsymbol{b} = \boldsymbol{Q}_{j+1}\boldsymbol{R}_{j+1}\boldsymbol{e}_{j+1}$, so that

$$\|\widetilde{p}_j(\boldsymbol{A})\boldsymbol{b}\|_2 = \left\|(\boldsymbol{I}-\boldsymbol{Q}_j\boldsymbol{Q}_j^*)\begin{bmatrix}\boldsymbol{Q}_j & \boldsymbol{q}_{j+1}\end{bmatrix}\begin{bmatrix}\boldsymbol{R}_j & \boldsymbol{w}_j \\ 0 & \rho_{j+1}\end{bmatrix}\boldsymbol{e}_{j+1}\right\|_2 = |\rho_{j+1}| \ .$$

On the other hand $\boldsymbol{A}^j\boldsymbol{b} = \boldsymbol{A}\boldsymbol{Q}_j\boldsymbol{R}_j\boldsymbol{e}_j = (\boldsymbol{Q}_j\boldsymbol{H}_j + h_{j+1,j}\,\boldsymbol{q}_{j+1}\boldsymbol{e}_j^T)\boldsymbol{R}_j\boldsymbol{e}_j$, from which

$$\left\|(\boldsymbol{I}-\boldsymbol{Q}_j\boldsymbol{Q}_j^*)\boldsymbol{A}^j\boldsymbol{b}\right\|_2 = |h_{j+1,j}|\,|\rho_j| \ .$$

This gives:

$$\begin{aligned}\|\widetilde{p}_j(\boldsymbol{A})\boldsymbol{b}\|_2 &= |\rho_{j+1}| = |h_{j+1,j}|\,|\rho_j| = |h_{j+1,j}|\,|h_{j,j-1}|\,|\rho_{j-1}| \\ &= |h_{j+1,j}|\,|h_{j,j-1}|\ldots|h_{2,1}|\,|\rho_1| = \|\boldsymbol{b}\|_2\prod_{k=1}^{j}|h_{k+1,k}| \ .\end{aligned}$$

□

Problem 9.6. How $\widetilde{p}_j$ is seen in the matrix $\boldsymbol{K}_j(\boldsymbol{A},\boldsymbol{b})^\dagger\boldsymbol{A}\boldsymbol{K}_j(\boldsymbol{A},\boldsymbol{b})$?

Using the theorem above we finally get the result showing the speed of approximation of the eigenvalues of $\boldsymbol{A}$. Here the main role is played by the *geometric mean* of the subdiagonal entries of the Hessenberg matrix $\boldsymbol{H}_j$.

Theorem 9.8. *If* $\boldsymbol{A}$ *is diagonalizable :* $\boldsymbol{V}^{-1}\boldsymbol{A}\boldsymbol{V} = \boldsymbol{D}$, $\boldsymbol{V} = \left[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\right]$, *and* $\boldsymbol{b} = \sum_{i=1}^n c_i\, \boldsymbol{v}_i$, *then*

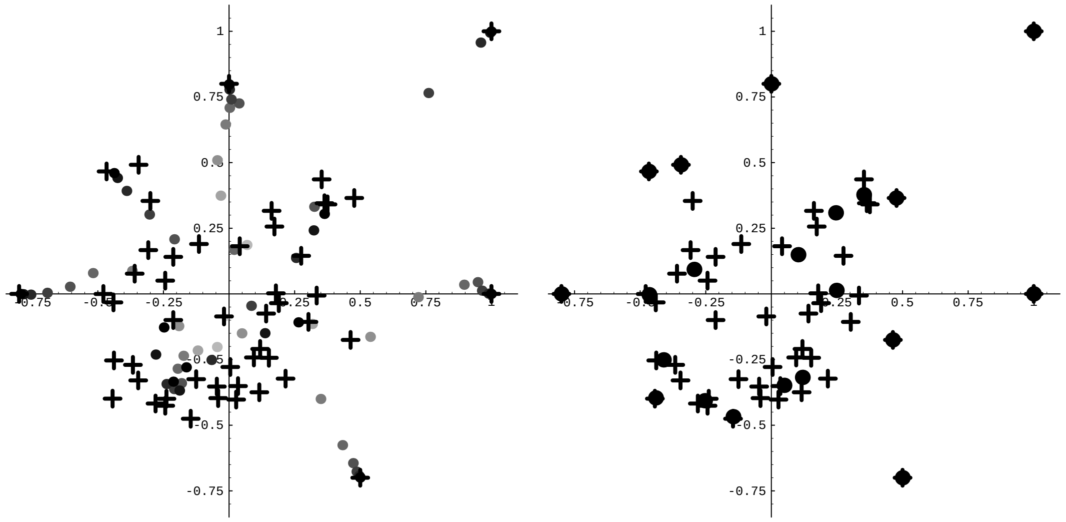
$$\delta(\lambda_k, \Lambda(\boldsymbol{H}_j)) \le \left(\frac{\|\boldsymbol{V}^{-1}\|_2}{|c_k|}\,|\rho_{j+1}|\right)^{1/j} = \left(\frac{\|\boldsymbol{V}^{-1}\|_2\,\|\boldsymbol{b}\|_2}{|c_k|}\prod_{k=1}^{j}|h_{k+1,k}|\right)^{1/j} .$$

Proof. Set $\varepsilon = \delta(\lambda_k, \Lambda(\boldsymbol{H}_j))$. Now $\widetilde{p}_j(z) = \prod_{l=1}^j (z - \mu_l^{(j)})$, so that $|\widetilde{p}_j(\lambda_k)| \ge \varepsilon^j$. From the previous theorem we get:

$$\begin{aligned}|\rho_{j+1}| &= \|\widetilde{p}_j(\boldsymbol{A})b\|_2 = \left\|\boldsymbol{V}\boldsymbol{V}^{-1}\sum_{i=1}^n c_i\widetilde{p}_j(\lambda_i)\boldsymbol{v}_i\right\|_2 = \left\|\boldsymbol{V}\sum_{i=1}^n c_i\widetilde{p}_j(\lambda_i)\boldsymbol{e}_i\right\|_2 \\ &\ge \frac{1}{\|\boldsymbol{V}^{-1}\|_2}\left(\textstyle\sum_{i=1}^n |c_i|^2\,|\widetilde{p}_j(\lambda_i)|^2\right)^{1/2} \ge \frac{|c_k|\,|\widetilde{p}_j(\lambda_k)|}{\|\boldsymbol{V}^{-1}\|_2} \ge \frac{|c_k|\,\varepsilon^j}{\|\boldsymbol{V}^{-1}\|_2} .\end{aligned}$$

□

In the following figures the eigenvalues of $\boldsymbol{A}$ are denoted by plus and the eigenvalues of $\boldsymbol{H}_j$ by dots with darkness increasing with the iteration index. On the left there are ten iterations, on the right $j = 20$. These show clearly, how the outer eigenvalues are found soon.



Problem 9.7. Let $\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ \boldsymbol{I}_{n-1} & 0 \end{bmatrix} \in \mathbb{C}^{n \times n}$ and $\boldsymbol{b} = \boldsymbol{e}_1$. Compute the Ritz values. This is a bad case for the Arnoldi iteration.

10. Classical iterations for linear systems

In the rest of these notes we consider iterative solution of an $n \times n$ system of linear equations $\boldsymbol{Ax} = \boldsymbol{b}$.

Assume $\boldsymbol{A}$ is regular, so that the equation has a unique solution $\boldsymbol{x}$. Many iterations are based on an invertible approximation $\boldsymbol{M}$ of $\boldsymbol{A}$, such that the system $\boldsymbol{My} = \boldsymbol{c}$ is easy to solve. Such $\boldsymbol{M}$ is called a *preconditioner*. Then the equation is written in the form $\boldsymbol{Mx} = \boldsymbol{Nx} + \boldsymbol{b}$, where $\boldsymbol{M} - \boldsymbol{N} = \boldsymbol{A}$. Starting with an initial guess $\boldsymbol{x}_0$ we iterate

$$\boldsymbol{M}\boldsymbol{x}_{k+1} = \boldsymbol{N}\boldsymbol{x}_k + \boldsymbol{b}\ , \qquad k = 0, 1, 2, \ldots \tag{10.1}$$

Taking the limits of both sides, we see that if this converges, then the limit is a solution. Iteration 10.1 is economical if it converges fast and if flop counts of the operations: multiplication by $\boldsymbol{N}$ and solving the system $\boldsymbol{My} = \boldsymbol{c}$ are small. This is often the case especially when $\boldsymbol{A}$ is *sparse*, i.e., when most of its elements are zero.

Denote the equation error, i.e., the *residual* by $\boldsymbol{r}_k = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k$. Then

$$\begin{aligned}\boldsymbol{r}_{k+1} =& \boldsymbol{b} - \boldsymbol{A}\boldsymbol{M}^{-1}(\boldsymbol{N}\boldsymbol{x}_k + \boldsymbol{b}) = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{M}^{-1}\boldsymbol{b} - \boldsymbol{A}\boldsymbol{M}^{-1}(\boldsymbol{M} - \boldsymbol{A})\boldsymbol{x}_k \\ =& (\boldsymbol{I} - \boldsymbol{A}\boldsymbol{M}^{-1})\boldsymbol{b} - (\boldsymbol{A} - \boldsymbol{A}\boldsymbol{M}^{-1}\boldsymbol{A})\boldsymbol{x}_k = (\boldsymbol{I} - \boldsymbol{A}\boldsymbol{M}^{-1})(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k) = \boldsymbol{N}\boldsymbol{M}^{-1}\boldsymbol{r}_k\ ,\end{aligned}$$

so that $\boldsymbol{r}_k = (\boldsymbol{N}\boldsymbol{M}^{-1})^k\boldsymbol{r}_0$. The error in the solution $\boldsymbol{e}_k = x^* - x_k = \boldsymbol{A}^{-1}\boldsymbol{r}_k$ satisfies, respectively,

$$\boldsymbol{e}_{k+1} = \boldsymbol{A}^{-1}\boldsymbol{r}_{k+1} = \boldsymbol{A}^{-1}(\boldsymbol{I} - \boldsymbol{A}\boldsymbol{M}^{-1})\boldsymbol{r}_k = (\boldsymbol{I} - \boldsymbol{M}^{-1}\boldsymbol{A})\boldsymbol{e}_k = \boldsymbol{M}^{-1}\boldsymbol{N}\boldsymbol{e}_k$$

and $\boldsymbol{e}_k = (\boldsymbol{M}^{-1}\boldsymbol{N})^k\boldsymbol{e}_0$.

The spectral radius is defined: $\rho(\boldsymbol{A}) = \max_{\lambda \in \Lambda(\boldsymbol{A})} |\lambda|$. By problem 5.17 we get:

Lemma 10.1. *Iteration 10.1 converges for all* $\boldsymbol{x}_0$, *if and only if* $\rho(\boldsymbol{M}^{-1}\boldsymbol{N}) < 1$.

The spectral radius can be characterized also by:

Lemma 10.2. $\rho(\boldsymbol{A}) = \lim\limits_{k\to\infty} \left\|\boldsymbol{A}^k\right\|^{\frac{1}{k}}$.

Here the norm can be any matrix norm. In the proof we need the following result.

Problem 10.1. Let $\|\cdot\|$ and $\|\cdot\|_*$ be two matrix norms in $\mathbb{C}^{m\times n}$. Show that there exist positive constants c, C such that

$$c\,\|\boldsymbol{M}\| \le \|\boldsymbol{M}\|_* \le C\,\|\boldsymbol{M}\|$$

for all $\boldsymbol{M} \in \mathbb{C}^{m\times n}$. This means that the norms are equivalent.

[0] Version: April 9, 2003

Proof. By the previous problem we have positive c, C such that for every matrix $\boldsymbol{M}$ holds

$$c\left\|\boldsymbol{M}\right\|_2 \leq \left\|\boldsymbol{M}\right\| \leq C\left\|\boldsymbol{M}\right\|_2 \ ,$$

from which

$$c^{\frac{1}{k}}\left\|\boldsymbol{A}^k\right\|_2^{\frac{1}{k}} \leq \left\|\boldsymbol{A}^k\right\|^{\frac{1}{k}} \leq C^{\frac{1}{k}}\left\|\boldsymbol{A}^k\right\|_2^{\frac{1}{k}} \ .$$

Since $c^{\frac{1}{k}}, C^{\frac{1}{k}} \underset{k\to\infty}{\to} 1$, it suffices to prove the claim for the 2–norm. Let $\boldsymbol{J} = \boldsymbol{X}^{-1}\boldsymbol{A}\boldsymbol{X}$ be the Jordan form of $\boldsymbol{A}$. Then $\boldsymbol{A}^k = \boldsymbol{X}\boldsymbol{J}^k\boldsymbol{X}^{-1}$ so that

$$(\left\|\boldsymbol{X}\right\|_2^{-1}\left\|\boldsymbol{X}^{-1}\right\|_2^{-1})^{\frac{1}{k}}\left\|\boldsymbol{J}^k\right\|_2^{\frac{1}{k}} \leq \left\|\boldsymbol{A}^k\right\|_2^{\frac{1}{k}} \leq (\left\|\boldsymbol{X}\right\|_2\left\|\boldsymbol{X}^{-1}\right\|_2)^{\frac{1}{k}}\left\|\boldsymbol{J}^k\right\|_2^{\frac{1}{k}} \ ,$$

so that showing $\rho(\boldsymbol{J}) = \lim_{k\to\infty}\left\|\boldsymbol{J}^k\right\|_2^{\frac{1}{k}}$ is enough. Since $\left\|\boldsymbol{J}^k\right\|_2 = \max_j\left\|\boldsymbol{J}(\lambda_j, r_j)^k\right\|_2$, where the matrices $J(\lambda_j, r_j)$ are the Jordan blocks of J and for these problem 5.16 gives[1]:

$$|\lambda|^k \leq \left\|\boldsymbol{J}(\lambda, r)_2^k\right\| \leq \sqrt{r}\max_{0\leq l<r}\binom{k}{l}|\lambda|^{k-l} \leq \sqrt{r}\,\max(1, \lambda^{-r})\,k^r\,|\lambda|^k \ .$$

The claim follows then from $\lim\limits_{k\to\infty}(k^r)^{\frac{1}{k}} = \lim\limits_{k\to\infty}e^{\frac{r}{k}\log(k)} = 1$. □

Lemma 10.3. *For every* $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ *and* $\varepsilon > 0$ *there exists a vector norm* $\left\|\cdot\right\|_\varepsilon$ *such that for the corresponding matrix norm we get*

$$\left\|\boldsymbol{A}\right\|_\varepsilon \leq \rho(\boldsymbol{A}) + \varepsilon \ .$$

Proof. Set $\boldsymbol{B} = \frac{1}{\rho(\boldsymbol{A})+\varepsilon}\boldsymbol{A}$ and $\left\|\boldsymbol{x}\right\|_\varepsilon = \sup_{k\geq 0}\left\|\boldsymbol{B}^k\boldsymbol{x}\right\|$. Since $\lim_{k\to\infty}\left\|\boldsymbol{A}^k\right\|^{1/k} = \rho(\boldsymbol{A})$ there exists m such that $\left\|\boldsymbol{A}^m\right\| < (\rho(\boldsymbol{A}) + \varepsilon)^m$. Hence $\left\|\boldsymbol{x}\right\|_\varepsilon$ is finite for every $\boldsymbol{x}$. Then we get:

1) $\left\|\boldsymbol{x}\right\|_\varepsilon = 0 \implies \left\|\boldsymbol{x}\right\| = 0 \implies \boldsymbol{x} = 0$

2) $\left\|\boldsymbol{x}+\boldsymbol{y}\right\|_\varepsilon = \sup_{k\geq 0}\left\|\boldsymbol{B}^k(\boldsymbol{x}+\boldsymbol{y})\right\| \leq \sup_{k\geq 0}\left\|\boldsymbol{B}^k\boldsymbol{x}\right\| + \sup_{k\geq 0}\left\|\boldsymbol{B}^k\boldsymbol{y}\right\| = \left\|\boldsymbol{x}\right\|_\varepsilon + \left\|\boldsymbol{y}\right\|_\varepsilon$

3) $\left\|\alpha\boldsymbol{x}\right\|_\varepsilon = \sup_{k\geq 0}\left\|\alpha\boldsymbol{B}^k\boldsymbol{x}\right\| = |\alpha|\left\|\boldsymbol{x}\right\|_\varepsilon$.

I.e., we really got a norm. Finally,

$$\begin{aligned}\left\|\boldsymbol{A}\boldsymbol{x}\right\|_\varepsilon &= \sup_{k\geq 0}\left\|\boldsymbol{B}^k\boldsymbol{A}\boldsymbol{x}\right\| = (\rho(\boldsymbol{A}) + \varepsilon)\sup_{k\geq 1}\left\|\boldsymbol{B}^k\boldsymbol{x}\right\| \\ &\leq (\rho(\boldsymbol{A}) + \varepsilon)\sup_{k\geq 0}\left\|\boldsymbol{B}^k\boldsymbol{x}\right\| = (\rho(\boldsymbol{A}) + \varepsilon)\left\|\boldsymbol{x}\right\|_\varepsilon \ .\end{aligned}$$

□

[1] using also: $\boldsymbol{A} \in \mathbb{C}^{n\times n} \implies \left\|\boldsymbol{A}\right\|_2 \leq \sqrt{n}\,\max_{i,j}|a_{i,j}|$

Remark 10.1. We have the following inequalities

$$\|\boldsymbol{x}\| \le \|\boldsymbol{x}\|_\epsilon \le C\,\|\boldsymbol{x}\| \ ,$$

where $C = \max_{k\ge 0} \left\|\boldsymbol{B}^k\right\|$.

Problem 10.2. Using a suitable norm for $\boldsymbol{M}^{-1}\boldsymbol{N}$ give another proof for lemma 10.1.

Problem 10.3. Assume you want to solve $\boldsymbol{x} = \boldsymbol{C}\boldsymbol{x} + \boldsymbol{c}$. If $\rho(\boldsymbol{C}) \ge 1$ the iteration $\boldsymbol{x}^{k+1} = \boldsymbol{C}\boldsymbol{x}^k + \boldsymbol{c}$ does not converge. Consider iterating

$$\boldsymbol{x}^{k+1} = \left((1-\omega)\,\boldsymbol{I} + \omega\,\boldsymbol{C}\right)\boldsymbol{x}^k + \omega\,\boldsymbol{c}$$

for $\omega \ne 0$. Show that if it converges, then the limit is a solution. Discuss the choice of ω. This technique is called *relaxation.* (Underrelaxation for $\omega < 1$ and overrelaxation in case $\omega > 1$.)

10.1. **Jacobi and Gauss–Seidel iterations.**

Write $\boldsymbol{A} = \boldsymbol{D} + \boldsymbol{L} + \boldsymbol{U}$, where $\boldsymbol{D} = \operatorname{diag}(a_{1,1}, \dots, a_{n,n})$, $\boldsymbol{L}$ consists of the subdiagonal elements of $\boldsymbol{A}$ and $\boldsymbol{U}$ is the part above the diagonal.

In the **Jacobi iteration** we choose $\boldsymbol{M}_J = \boldsymbol{D}$, so that the iteration becomes:

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

Now the convergence depends on the spectral radius of the matrix $\boldsymbol{E}_J = -\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{U})$. Typically it converges the better the more diagonally dominant $\boldsymbol{A}$ is.

Lemma 10.4. *Let $\boldsymbol{A}$ be strictly diagonally dominant, i.e., for all j holds $|a_{j,j}| > \sum_{k\ne j} |a_{j,k}|$. Then the Jacobi iteration converges.*

Proof. Let $\lambda \in \Lambda(\boldsymbol{E}_J)$, and let $\boldsymbol{x}$ be the corresponding eigenvector and $\boldsymbol{x}_j$ the component of $\boldsymbol{x}$ with the largest absolute value. From equation $(\boldsymbol{L}+\boldsymbol{U})\boldsymbol{x} = -\lambda\boldsymbol{D}\boldsymbol{x}$ we then get

$$|\lambda a_{j,j}\boldsymbol{x}_j| = \Big|\sum_{k\ne j} a_{j,k}\boldsymbol{x}_k\Big| \le \sum_{k\ne j} |a_{j,k}|\,|\boldsymbol{x}_k| \ ,$$

from which $$|\lambda| \le \sum_{k\ne j} \frac{|a_{j,k}|\,|\boldsymbol{x}_k|}{|a_{j,j}|\,|\boldsymbol{x}_j|} < 1 \ .$$ □

In the **Gauss–Seidel iteration** we take $\boldsymbol{M}_{GS} = \boldsymbol{D} + \boldsymbol{L}$. Then in each step we need to solve a lower triangular system (with forward substitutions): $(\boldsymbol{D}+\boldsymbol{L})\boldsymbol{x}_{k+1} = \boldsymbol{b} - \boldsymbol{U}\boldsymbol{x}_k$ and the iteration becomes:

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

The convergence of this is determined by the spectral radius of the matrix $\boldsymbol{E}_{GS} = -(\boldsymbol{D}+\boldsymbol{L})^{-1}\boldsymbol{U}$.

Example 10.1. For $\boldsymbol{A} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$ we get the iteration matrices for Jacobi and Gauss-Seidel:

$$\boldsymbol{E}_J = \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix} , \qquad \boldsymbol{E}_{GS} = \begin{bmatrix} 2 & 0 \\ -1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2} \\ 0 & \frac{1}{4} \end{bmatrix} .$$

Hence $\rho(\boldsymbol{E}_J) = \frac{1}{2}$, $\rho(\boldsymbol{E}_{GS}) = \frac{1}{4}$.

Problem 10.4. Show: if $\boldsymbol{A}$ is strictly diagonally dominant, then the Gauss–Seidel iteration converges.

For this we further have

Theorem 10.5. *If $\boldsymbol{A}$ is Hermitian and positive definite, then the Gauss–Seidel iteration converges.*

Proof. If $\boldsymbol{A}$ is Hermitian, then $\boldsymbol{U} = \boldsymbol{L}^*$ and we have to show that

$$\rho((\boldsymbol{D} + \boldsymbol{L})^{-1}\boldsymbol{L}^*) < 1 .$$

Now the entries of D are positive so that it has a positive square root $\boldsymbol{D}^{\frac{1}{2}}$ and $(\boldsymbol{D} + \boldsymbol{L})^{-1}\boldsymbol{L}^*$ is similar to matrix

$$\boldsymbol{D}^{\frac{1}{2}}(\boldsymbol{D} + \boldsymbol{L})^{-1}\boldsymbol{L}^*\boldsymbol{D}^{-\frac{1}{2}} = (\boldsymbol{I} + \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{L}\boldsymbol{D}^{-\frac{1}{2}})^{-1}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{L}^*\boldsymbol{D}^{-\frac{1}{2}} = (\boldsymbol{I} + \boldsymbol{L}_1)^{-1}\boldsymbol{L}_1^* ,$$

where $\boldsymbol{L}_1 = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{L}\boldsymbol{D}^{-\frac{1}{2}}$. If $(\boldsymbol{I} + \boldsymbol{L}_1)^{-1}\boldsymbol{L}_1^*\boldsymbol{x} = \lambda\,\boldsymbol{x}$, where $\|\boldsymbol{x}\|_2 = 1$, then $\boldsymbol{L}_1^*\boldsymbol{x} = \lambda\,\boldsymbol{x} + \lambda\boldsymbol{L}_1\,\boldsymbol{x}$, so that $\lambda = \bar{\alpha}/(1+\alpha)$, where $\alpha = \boldsymbol{x}^*\boldsymbol{L}_1\boldsymbol{x}$. Since

$$\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}} = \boldsymbol{I} + \boldsymbol{L}_1 + \boldsymbol{L}_1^*$$

is positive definite, we get $1 + \alpha + \bar{\alpha} > 0$, from which

$$|\lambda|^2 = \frac{\bar{\alpha}}{1+\alpha}\,\frac{\alpha}{1+\bar{\alpha}} = \frac{|\alpha|^2}{1 + \alpha + \bar{\alpha} + |\alpha|^2} < 1 . \qquad \square$$

10.2. **SOR iteration (overrelaxation).** The Gauss–Seidel iteration can be written as:

$$\boldsymbol{x}_{k+1} = \boldsymbol{D}^{-1}(\boldsymbol{b} - \boldsymbol{U}\boldsymbol{x}_k - \boldsymbol{L}\boldsymbol{x}_{k+1}) ,$$

where the components of $\boldsymbol{x}_{k+1}$ are already computed when they are needed. The **S**uccessive **O**ver **R**elaxation method starts from the idea that if this is a good step, then why not take it a little longer

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \omega\,[\boldsymbol{D}^{-1}(\,\boldsymbol{b} - \boldsymbol{U}\boldsymbol{x}_k - \boldsymbol{L}\boldsymbol{x}_{k+1}) - \boldsymbol{x}_k] , \tag{10.2}$$

where $\omega > 1$ (usually). Writing SOR in the standard form with $\boldsymbol{M}$ and $\boldsymbol{N}$ we get:

$$(\tfrac{1}{\omega}\boldsymbol{D} + \boldsymbol{L})\boldsymbol{x}_{k+1} = [(\tfrac{1}{\omega} - 1)\boldsymbol{D} - \boldsymbol{U}]\,\boldsymbol{x}_k + \boldsymbol{b} ,$$

i.e., $\boldsymbol{M}_\omega \boldsymbol{x}_{k+1} = \boldsymbol{N}_\omega \boldsymbol{x}_k + \boldsymbol{b}$, where

$$\boldsymbol{M}_\omega = \tfrac{1}{\omega}(\boldsymbol{D} + \omega\, \boldsymbol{L}) \ , \quad \boldsymbol{N}_\omega = \tfrac{1}{\omega}((1-\omega)\boldsymbol{D} - \omega\, \boldsymbol{U}) \ .$$

With the value $\omega = 1$ SOR is just the Gauss–Seidel iteration. The convergence of SOR is determined by the spectral radius of the matrix $\boldsymbol{E}_\omega = \boldsymbol{M}_\omega^{-1}\boldsymbol{N}_\omega$. First result concerning this is

Lemma 10.6. $\rho(\boldsymbol{E}_\omega) \geq |\omega - 1|$.

Hence it is necessary for the convergence that $\omega \in (0,2)$.

Proof. Since the determinant is the product of the eigenvalues and the determinant of a triangular matrix is the product of the diagonal elements we get:

$$\begin{aligned}\prod_j |\lambda_j| &= \left|\det(\boldsymbol{M}_\omega^{-1}\boldsymbol{N}_\omega)\right| = \left|\det((\boldsymbol{D}+\omega\boldsymbol{L})^{-1}((1-\omega)\boldsymbol{D} - \omega\boldsymbol{U}))\right| \\ &= \left|\det(\boldsymbol{D}+\omega\boldsymbol{L})^{-1}\det((1-\omega)\boldsymbol{D}-\omega\boldsymbol{U})\right| = |1-\omega|^n \ .\end{aligned}$$

Hence at least one of the eigenvalues has absolute value $\geq |1-\omega|$. □

Theorem 10.7. *If $\boldsymbol{A}$ is Hermitian and positive definite, then SOR converges for all $\omega \in (0,2)$.*

Proof. Fix $\omega \in (0,2)$. We have $\boldsymbol{A} = \boldsymbol{D} + \boldsymbol{L} + \boldsymbol{L}^*$ and $\boldsymbol{M}_\omega = \frac{1}{\omega}\boldsymbol{D} + \boldsymbol{L}$. Now the matrix

$$\boldsymbol{M}_\omega + \boldsymbol{M}_\omega^* - \boldsymbol{A} = \tfrac{2}{\omega}\,\boldsymbol{D} + \boldsymbol{L} + \boldsymbol{L}^* - \boldsymbol{D} - \boldsymbol{L} - \boldsymbol{L}^* = (\tfrac{2}{\omega} - 1)\,\boldsymbol{D}$$

is positive definite. Set $\boldsymbol{Q} := \boldsymbol{A}^{-1}(2\,\boldsymbol{M}_\omega - \boldsymbol{A})$. If $\lambda, \boldsymbol{x}$ is an eigenpair of $\boldsymbol{Q}$ then $(2\,\boldsymbol{M}_\omega - \boldsymbol{A})\,\boldsymbol{x} = \lambda\,\boldsymbol{A}\,\boldsymbol{x}$ and

$$\begin{aligned}\boldsymbol{x}^*(2\,\boldsymbol{M}_\omega - \boldsymbol{A})\,\boldsymbol{x} &= \lambda\,\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} \\ \boldsymbol{x}^*(2\,\boldsymbol{M}_\omega^* - \boldsymbol{A})\,\boldsymbol{x} &= \overline{\lambda}\,\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} \ ,\end{aligned}$$

i.e., $\boldsymbol{x}^*(\boldsymbol{M}_\omega + \boldsymbol{M}_\omega^* - \boldsymbol{A})\,\boldsymbol{x} = \operatorname{Re}\lambda\,\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}$. Since $\boldsymbol{M}_\omega + \boldsymbol{M}_\omega^* - \boldsymbol{A}$ and $\boldsymbol{A}$ are positive definite, we get $\operatorname{Re}\lambda > 0$. Now

$$(\boldsymbol{Q} - \boldsymbol{I})(\boldsymbol{Q} + \boldsymbol{I})^{-1} = (2\,\boldsymbol{A}^{-1}\boldsymbol{M}_\omega - 2\,\boldsymbol{I})\,\tfrac{1}{2}\,\boldsymbol{M}_\omega^{-1}\boldsymbol{A} = \boldsymbol{E}_\omega \ .$$

Hence the eigenvalues of $\boldsymbol{E}_\omega$ are of the form $\frac{\lambda-1}{\lambda+1}$, $\lambda \in \Lambda(\boldsymbol{Q})$ and for $\lambda = \alpha + i\beta$, $\alpha > 0$ we have

$$\left|\frac{\lambda-1}{\lambda+1}\right| = \sqrt{\frac{(1-\alpha)^2+\beta^2}{(1+\alpha)^2+\beta^2}} < 1 \ .$$

□

If SOR converges for some ω then, since spectral radius is a continuous function of the matrix, it has a minimum < 1 on the interval $(0, 2)$. This minimum point ω_{opt} is called the optimal relaxation parameter. Generally it is a difficult task to find it, but there are many cases, where it is well known. In particular, in some cases of discretized partial differential equations. There a thumb rule is that if the Jacobi iteration converges, then Gauss–Seidel converges twice that fast and the optimal SOR much faster.

Definition 10.1 (R. Varga)**.** Matrix $\boldsymbol{A} = \boldsymbol{L} + \boldsymbol{D} + \boldsymbol{U}$ is said to be *consistently ordered*, if the spectrum of

$$\boldsymbol{B}_\alpha = \alpha \boldsymbol{D}^{-1}\boldsymbol{L} + \tfrac{1}{\alpha}\,\boldsymbol{D}^{-1}\boldsymbol{U}$$

is independent of $\alpha \in \mathbb{C} \setminus \{0\}$.

An important set in this class is the following. We say that the matrix $\boldsymbol{A}$ has *property* A^P, if it can be similarity transformed with a permutation $\boldsymbol{\Pi}$ to the form:

$$\boldsymbol{\Pi}\boldsymbol{A}\boldsymbol{\Pi}^T = \begin{bmatrix} \boldsymbol{D}_1 & \boldsymbol{U}_1 & 0 & \dots & 0 \\ \boldsymbol{L}_1 & \boldsymbol{D}_2 & \boldsymbol{U}_2 & \ddots & \vdots \\ 0 & \boldsymbol{L}_2 & \boldsymbol{D}_3 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \boldsymbol{U}_{r-1} \\ 0 & \dots & 0 & \boldsymbol{L}_{r-1} & \boldsymbol{D}_r \end{bmatrix}, \tag{10.3}$$

where matrices $\boldsymbol{D}_j$ are invertible and diagonal.

Problem 10.5. Let $\boldsymbol{B} \in \mathbb{R}^{n \times n}$ be a tridiagonal matrix, whose diagonal is constant -4 and other elements are ones. Find a permutation that takes the $nm \times nm$ matrix

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{B} & \boldsymbol{I} & 0 & \dots & 0 \\ \boldsymbol{I} & \boldsymbol{B} & \boldsymbol{I} & \ddots & \vdots \\ 0 & \boldsymbol{I} & \boldsymbol{B} & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \boldsymbol{I} \\ 0 & \dots & 0 & \boldsymbol{I} & \boldsymbol{B} \end{bmatrix}$$

into the form above. This $\boldsymbol{A}$ comes from the uniform discretization of the Laplace operator $\Delta u = u_{xx} + u_{yy}$ in a rectangle of the plane.

Theorem 10.8. *Assume* $\boldsymbol{A}$ *has property* A^P. *Then it is consistently ordered.*

Proof. Let $\widetilde{\boldsymbol{A}} = \boldsymbol{\Pi}\boldsymbol{A}\boldsymbol{\Pi}^T = \widetilde{\boldsymbol{D}} + \widetilde{\boldsymbol{L}} + \widetilde{\boldsymbol{U}}$ be of the form of (10.3). Since a permutation similarity transformation maps the diagonal elements onto the diagonal and others outside it we have: $\widetilde{\boldsymbol{D}} = \boldsymbol{\Pi D \Pi}^T$ and $\widetilde{\boldsymbol{L}} + \widetilde{\boldsymbol{U}} = \boldsymbol{\Pi}(\boldsymbol{L}+\boldsymbol{U})\boldsymbol{\Pi}^T$. Now $\widetilde{\boldsymbol{E}}_J = -\widetilde{\boldsymbol{D}}^{-1}(\widetilde{\boldsymbol{L}} + \widetilde{\boldsymbol{U}})$ is similar to $\boldsymbol{E}_J$:

$$\widetilde{\boldsymbol{E}}_J = -\boldsymbol{\Pi D}^{-1}\boldsymbol{\Pi}^T\boldsymbol{\Pi}(\boldsymbol{L}+\boldsymbol{U})\boldsymbol{\Pi}^T = \boldsymbol{\Pi E}_J\boldsymbol{\Pi}^T \ ,$$

so that $\Lambda(\widetilde{\boldsymbol{E}}_J) = \Lambda(\boldsymbol{E}_J)$. Set $\boldsymbol{B}_\alpha = \alpha\boldsymbol{D}^{-1}\boldsymbol{L} + \alpha^{-1}\boldsymbol{D}^{-1}\boldsymbol{U}$. Then

$$\boldsymbol{\Pi B}_\alpha\boldsymbol{\Pi}^T = \widetilde{\boldsymbol{D}}^{-1}(\alpha\widetilde{\boldsymbol{L}} + \alpha^{-1}\widetilde{\boldsymbol{U}}) = \begin{bmatrix} \boldsymbol{I} & & & \\ & \alpha\boldsymbol{I} & & \\ & & \ddots & \\ & & & \alpha^{r-1}\boldsymbol{I} \end{bmatrix} \widetilde{\boldsymbol{D}}^{-1}(\widetilde{\boldsymbol{L}} + \widetilde{\boldsymbol{U}}) \begin{bmatrix} \boldsymbol{I} & & & \\ & \alpha^{-1}\boldsymbol{I} & & \\ & & \ddots & \\ & & & \alpha^{1-r}\boldsymbol{I} \end{bmatrix} \ ,$$

so that $\boldsymbol{B}_\alpha \sim -\widetilde{\boldsymbol{E}}_J \sim -\boldsymbol{E}_J$ for all $\alpha \in \mathbb{C}$. □

The following result reveals a connection between the convergence speeds of the Jacobi iteration and SOR for consistently ordered matrices

Theorem 10.9. *Assume $\boldsymbol{A}$ is consistently ordered. Then*

a) $\mu \in \Lambda(\boldsymbol{E}_J) \implies -\mu \in \Lambda(\boldsymbol{E}_J)$.

b) *If $\omega \neq 0$, then for any $\mu \in \Lambda(\boldsymbol{E}_J)$ there exists $\lambda \in \Lambda(\boldsymbol{E}_\omega)$ such that*

$$(\lambda + \omega - 1)^2 = \lambda\omega^2\mu^2$$

and for any $\lambda \in \Lambda(\boldsymbol{E}_\omega)$, $\lambda \neq 0$ there exists $\mu \in \Lambda(\boldsymbol{E}_J)$ such that the equation above holds.

Proof. a) $\boldsymbol{B}_1 = -\boldsymbol{E}_J$ and $\boldsymbol{B}_{-1} = \boldsymbol{E}_J$.

b) Write:

$$\boldsymbol{E}_\omega = (\tfrac{1}{\omega}\,\boldsymbol{D} + \boldsymbol{L})^{-1}\,(\tfrac{1}{\omega}\,(1-\omega)\,\boldsymbol{D} - \boldsymbol{U}) = (\boldsymbol{I} + \omega\boldsymbol{D}^{-1}\boldsymbol{L})^{-1}((1-\omega)\,\boldsymbol{I} - \omega\,\boldsymbol{D}^{-1}\boldsymbol{U})\,.$$

Since $\det(\boldsymbol{I} + \omega\boldsymbol{D}^{-1}\boldsymbol{L}) = 1$, we get

$$\begin{aligned} \det(\lambda\boldsymbol{I} - \boldsymbol{E}_\omega) &= \det(\boldsymbol{I} + \omega\boldsymbol{D}^{-1}\boldsymbol{L})\det(\lambda\boldsymbol{I} - \boldsymbol{E}_\omega) \\ &= \det(\lambda\boldsymbol{I} + \lambda\,\omega\boldsymbol{D}^{-1}\boldsymbol{L} - (1-\omega)\boldsymbol{I} + \omega\boldsymbol{D}^{-1}\boldsymbol{U}) \\ &= \det((\lambda + \omega - 1)\boldsymbol{I} + \lambda\,\omega\boldsymbol{D}^{-1}\boldsymbol{L} + \omega\boldsymbol{D}^{-1}\boldsymbol{U}) \ . \end{aligned}$$

Suppose $\mu \in \Lambda(\boldsymbol{E}_J)$ and λ satisfies $(\lambda + \omega - 1)^2 = \lambda\omega^2\mu^2$. By a) we can assume $\lambda + \omega - 1 = -\sqrt{\lambda}\,\omega\,\mu$.

Now, if $\lambda = 0$, then $\omega = 1$ and $\det(\boldsymbol{E}_1) = \det(-(\boldsymbol{D}+\boldsymbol{L})^{-1}\boldsymbol{U}) = 0$, since $\det(\boldsymbol{U}) = 0$. Hence $\lambda \in \Lambda(\boldsymbol{E}_1)$.

If $\lambda \neq 0$, then

$$\begin{aligned}\det(\lambda \boldsymbol{I} - \boldsymbol{E}_\omega) &= \det((\lambda + \omega - 1)\boldsymbol{I} + \omega\sqrt{\lambda}(\sqrt{\lambda}\boldsymbol{D}^{-1}\boldsymbol{L} + \tfrac{1}{\sqrt{\lambda}}\,\boldsymbol{D}^{-1}\boldsymbol{U}) \\ &= (\omega\sqrt{\lambda})^n \ \det(\tfrac{\lambda+\omega-1}{\omega\sqrt{\lambda}}\,\boldsymbol{I} + (\sqrt{\lambda}\,\boldsymbol{D}^{-1}\boldsymbol{L} + \tfrac{1}{\sqrt{\lambda}}\,\boldsymbol{D}^{-1}\boldsymbol{U})\ .\end{aligned}$$

But $\frac{\lambda+\omega-1}{\omega\sqrt{\lambda}} = -\mu \in \Lambda(\boldsymbol{E}_J) = \Lambda(\sqrt{\lambda}\,\boldsymbol{D}^{-1}\boldsymbol{L} + \frac{1}{\sqrt{\lambda}}\,\boldsymbol{D}^{-1}\boldsymbol{U})$. Hence $\det(\lambda\boldsymbol{I} - \boldsymbol{E}_\omega) = 0$.

For the reverse, read the proof above backwards. □

Problem 10.6. Show: if $\boldsymbol{A}$ is consistently ordered then the Gauss–Seidel iteration converges if and only if the Jacobi iteration converges and $\rho(\boldsymbol{E}_{GS}) = \rho(\boldsymbol{E}_J)^2$.

Problem 10.7. Show: if $\boldsymbol{A}$ is Hermitian and is consistently ordered, then the Jacobi iteration converges.

Further, if the eigenvalues of the Jacobi iteration are real, then the following gives the optimal value of the relaxation parameter.

Theorem 10.10. *Let $\boldsymbol{A}$ be consistently ordered and $\Lambda(\boldsymbol{E}_J) \subset (-1, 1)$. Then*

$$\omega_{\text{opt}} = \frac{2}{1 + \sqrt{1 - \rho(\boldsymbol{E}_J)^2}}$$

and

$$\rho(\boldsymbol{E}_{\omega_{\text{opt}}}) = \omega_{\text{opt}} - 1 = \frac{1 - \sqrt{1 - \rho(\boldsymbol{E}_J)^2}}{1 + \sqrt{1 - \rho(\boldsymbol{E}_J)^2}}\ .$$

Proof. Let $\rho = \rho(\boldsymbol{E}_J) < 1$. Then $\rho \in \Lambda(\boldsymbol{E}_J)$ and the corresponding eigenvalue λ of SOR satisfies

$$\lambda^2 - 2(1 - \omega + \tfrac{1}{2}\,\omega^2\rho^2)\,\lambda + (\omega - 1)^2 = 0\ ,$$

and a short computation gives

$$\lambda = \left(\tfrac{1}{2}\rho\omega \pm \sqrt{1 - \omega + \rho^2\omega^2/4}\right)^2\ . \tag{10.4}$$

Let $\omega_0 \in (0, 2)$ be the zero of the discriminant:

$$1 - \omega_0 + \rho^2\omega_0^2/4 = 0\ , \qquad \text{i.e.,} \qquad \omega_0 = \frac{2}{1 + \sqrt{1 - \rho^2}}\ .$$

When $\omega < \omega_0$, the function $f(\omega) = \frac{1}{2}\rho\omega + \sqrt{1-\omega+\rho^2\omega^2/4}$ is decreasing:

$$\begin{aligned} f'(\omega) =& \tfrac{1}{2}\rho + \tfrac{1}{2}\frac{-1+\frac{1}{2}\omega\rho^2}{\sqrt{1-\omega+\rho^2\omega^2/4}} \\ =& \frac{1}{2\sqrt{1-\omega+\rho^2\omega^2/4}}\left(\rho\sqrt{1-\omega+\rho^2\omega^2/4}-1+\tfrac{1}{2}\omega\rho^2\right) \\ <& \frac{1}{2\sqrt{1-\omega+\rho^2\omega^2/4}}\left(\sqrt{(1-\omega/2)^2}-1+\tfrac{1}{2}\omega\right) = 0\ . \end{aligned}$$

On the other hand, when $2 > \omega > \omega_0$, we get:

$$\begin{aligned} |\lambda| =& \left|\tfrac{1}{2}\rho\omega + i\sqrt{\omega-1-\rho^2\omega^2/4}\right|^2 \\ =& \rho^2\omega^2/4+\omega-1-\rho^2\omega^2/4 = \omega-1\ , \end{aligned}$$

so that $|\lambda|$ is increasing. Hence ω_0 gives the minimum for $|\lambda|$. Still we need to check the other eigenvalues of SOR. If $\lambda' \in \Lambda(\boldsymbol{E}_{\omega_0})$, then we get it from some eigenvalue $\mu \in [0,\rho]$ of $\boldsymbol{E}_J$ like the formula (10.4), where now $1-\omega_0+\mu^2\omega_0^2/4 < 0$, so that

$$|\lambda'| = \left|\tfrac{1}{2}\mu\omega_0 + i\sqrt{\omega_0-1-\mu^2\omega_0^2/4}\right|^2 = \omega_0 - 1\ .$$

Hence the optimal relaxation parameter shifts all the eigenvalues of the iteration matrix to the circle of radius $\omega_0 - 1$. □

Example 10.2. Let $\boldsymbol{A}$ $_{n\times n}$ be a tridiagonal matrix with diagonal constant 2 and ones on both sides of it so that it is already in the form (10.3):

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 & & & \\ 1 & 2 & 1 & & \\ & 1 & 2 & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & 2 \end{bmatrix}\ .$$

Since

$$\sin((k-1)\alpha) + 2\sin(k\alpha) + \sin((k+1)\alpha) = 2(\cos(\alpha)+1)\sin(k\alpha)\ ,$$

we see, that $v = (\sin(\alpha), \sin(2\alpha), \ldots, \sin(n\alpha))$ is an eigenvector of $\boldsymbol{A}$ provided that

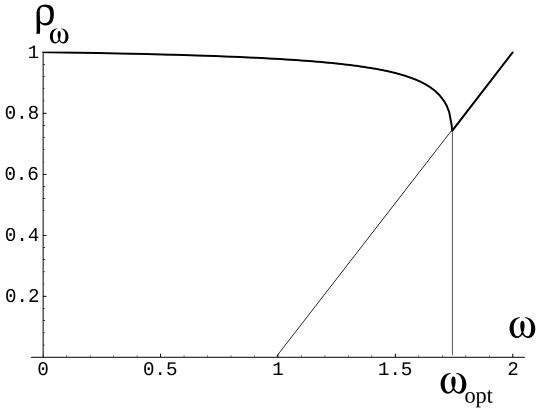
$$\sin((n+1)\alpha) = 0 \qquad \text{, i.e., if} \qquad \alpha = j\pi/(n+1)\ ,\ j = 1,\ldots,n\ .$$

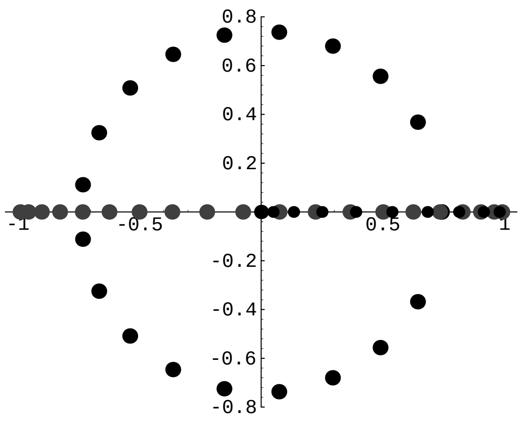
Hence $\boldsymbol{A}$ has eigenvalues and eigenvectors:

$$\lambda_j = 2(\cos(\tfrac{j\pi}{n+1})+1)\ , \qquad \boldsymbol{v}_j = (\sin(\tfrac{j\pi}{n+1}), \sin(2\tfrac{j\pi}{n+1}), \ldots, \sin(n\tfrac{j\pi}{n+1}))\ .$$

$j = 1,\ldots,n$. Now, since $\boldsymbol{E}_J = \frac{1}{2}(\boldsymbol{A} - 2\boldsymbol{I})$, we get $\rho(\boldsymbol{E}_J) = \cos(\pi/(n+1))$.

If $n = 20$, we get $\rho(\boldsymbol{E}_J) = 0.988831$ and the optimal relaxation parameter is $\omega_{\text{opt}} = 1.74058$ which gives the spectral radius $\rho(\boldsymbol{E}_{\omega_{\text{opt}}}) = 0.74058$ for SOR. In the following $\rho_\omega = \rho(\boldsymbol{E}_\omega)$. On the right eigenvalues of the Jacobi iteration are drawn grey, those of Gauss–Seidel are small black and of SOR larger black points.





Problem 10.8. In the problem 10.5 compute the optimal ω for SOR iteration. Hint: Look for the eigenvectors of $\boldsymbol{A}$ as “Cartesian products” of the eigenvectors of Example 10.2.

10.3. **SSOR (Symmetric SOR).** In the next section we will consider Tsebyshev–acceleration of an iteration. It can become very effective when the spectrum of the iteration matrix is real. Even if $\boldsymbol{A}$ is Hermitian, the iteration matrix

$$\boldsymbol{E}_\omega = (\boldsymbol{D} + \omega \boldsymbol{L})^{-1}((1-\omega)\boldsymbol{D} - \omega\, \boldsymbol{U})$$

of the SOR iteration

$$\tfrac{1}{\omega}\,(\boldsymbol{D} + \omega\, \boldsymbol{L})\, \boldsymbol{x}_{k+1} = \tfrac{1}{\omega}\,[(1-\omega)\boldsymbol{D} - \omega\, \boldsymbol{U}]\, \boldsymbol{x}_k + \boldsymbol{b}$$

is usually not Hermitian, and it does not have real spectrum. The idea of symmetric SOR (SSOR) is to change the roles of $\boldsymbol{L}$ and $\boldsymbol{U}$ after every step, i.e., every second iteration is performed using matrices

$$\widetilde{\boldsymbol{M}}_\omega = \tfrac{1}{\omega}\,(\boldsymbol{D} + \omega\, \boldsymbol{U}) \qquad \text{and} \qquad \widetilde{\boldsymbol{N}}_\omega = \tfrac{1}{\omega}\,[(1-\omega)\boldsymbol{D} - \omega\, \boldsymbol{L}]\ .$$

Usually SSOR is written in the form:

$$\begin{aligned}(\boldsymbol{D} + \omega \boldsymbol{L})\, \boldsymbol{x}_{k+\frac{1}{2}} &= [(1-\omega)\boldsymbol{D} - \omega\, \boldsymbol{U}]\, \boldsymbol{x}_k + \omega \boldsymbol{b} \\ (\boldsymbol{D} + \omega \boldsymbol{U})\, \boldsymbol{x}_{k+1} &= [(1-\omega)\boldsymbol{D} - \omega\, \boldsymbol{L}]\, \boldsymbol{x}_{k+\frac{1}{2}} + \omega \boldsymbol{b}\ .\end{aligned}$$

Then, if $\boldsymbol{A}$ is Hermitian, we have $\boldsymbol{U} = \boldsymbol{L}^*$, $\widetilde{\boldsymbol{M}}_\omega = \boldsymbol{M}_\omega^*$, $\widetilde{\boldsymbol{N}}_\omega = \boldsymbol{N}_\omega^*$ and the iteration matrix becomes

$$\boldsymbol{E}_\omega^{\text{SSOR}} = (\boldsymbol{M}_\omega^*)^{-1} \boldsymbol{N}_\omega^* \boldsymbol{M}_\omega^{-1} \boldsymbol{N}_\omega\ .$$

Neither this is usually Hermitian, but since[2]

$$\begin{aligned}\boldsymbol{D}^{-1}\boldsymbol{N}_\omega^*\boldsymbol{M}_\omega^{-1}\boldsymbol{D} &= \boldsymbol{D}^{-1}((1-\omega)\boldsymbol{D}-\omega\boldsymbol{L})\,(\boldsymbol{D}+\omega\boldsymbol{L})^{-1}\,\boldsymbol{D}\\ &= ((1-\omega)\boldsymbol{I}-\omega\boldsymbol{D}^{-1}\boldsymbol{L})\,(\boldsymbol{I}+\omega\boldsymbol{D}^{-1}\boldsymbol{L})^{-1}\\ &= (\boldsymbol{I}+\omega\boldsymbol{D}^{-1}\boldsymbol{L})^{-1}\,((1-\omega)\boldsymbol{I}-\omega\boldsymbol{D}^{-1}\,\boldsymbol{L})\\ &= (\boldsymbol{D}(\boldsymbol{I}+\omega\boldsymbol{D}^{-1}\boldsymbol{L}))^{-1}\,\boldsymbol{D}\,((1-\omega)\boldsymbol{I}-\omega\boldsymbol{D}^{-1}\boldsymbol{L}) = \boldsymbol{M}_\omega^{-1}\boldsymbol{N}_\omega^*\,,\end{aligned}$$

we get

$$\boldsymbol{E}_\omega^{\text{SSOR}} = (\boldsymbol{M}_\omega^*)^{-1}\boldsymbol{D}\boldsymbol{M}_\omega^{-1}\boldsymbol{N}_\omega^*\boldsymbol{D}^{-1}\boldsymbol{N}_\omega\,.$$

Now, if $\boldsymbol{D}$ has positive diagonal elements, then $\boldsymbol{N}_\omega^*\boldsymbol{D}^{-1}\boldsymbol{N}_\omega = \boldsymbol{K}^*\boldsymbol{K}$, where $\boldsymbol{K} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{N}_\omega$, and

$$\boldsymbol{K}\boldsymbol{E}_\omega^{\text{SSOR}}\boldsymbol{K}^{-1} = \boldsymbol{K}(\boldsymbol{M}_\omega\boldsymbol{D}^{-1}\boldsymbol{M}_\omega^*)^{-1}\boldsymbol{K}^*$$

is Hermitian and hence the iteration matrix $\boldsymbol{E}_\omega^{\text{SSOR}}$ has real eigenvalues.

We will not consider the selection of the relaxation parameter for symmetric SOR. Let us mention only that the ω of SOR is usually good also for SSOR and that SSOR is not very sensitive with respect to the parameter.

Problem 10.9. Experiment Jacobi, Gauss–Seidel, SOR and SSOR for the problem 10.5 with $\boldsymbol{b}$ a random vector and $m \approx n \approx 20$. Take the ω from problem 10.8.

Theorem 10.11. *Assume $\boldsymbol{A} = \boldsymbol{A}^*$ is positive definite. Then SSOR converges for every $\omega \in (0,2)$,*

Proof. Write

$$\begin{aligned}\boldsymbol{E}_\omega^{\text{SSOR}} &= (\tfrac{1}{\omega}\,(\boldsymbol{D}+\omega\boldsymbol{U}))^{-1}\,\tfrac{1}{\omega}((1-\omega)\boldsymbol{D}-\omega\boldsymbol{L})\,(\tfrac{1}{\omega}(\boldsymbol{D}+\omega\boldsymbol{L}))^{-1}\,\tfrac{1}{\omega}((1-\omega)\boldsymbol{D}-\omega\boldsymbol{U})\\ &= (\tfrac{1}{\omega}\,\boldsymbol{D}+\boldsymbol{L}^*)^{-1}\,(\tfrac{1-\omega}{\omega}\boldsymbol{D}-\boldsymbol{L})\,(\tfrac{1}{\omega}\boldsymbol{D}+\boldsymbol{L})^{-1}\,(\tfrac{1-\omega}{\omega}\boldsymbol{D}-\boldsymbol{L}^*)\\ &= \boldsymbol{D}^{-1/2}\,(\tfrac{1}{\omega}\boldsymbol{I}+\boldsymbol{K}^*)^{-1}\,(\tfrac{1-\omega}{\omega}\boldsymbol{I}-\boldsymbol{K})\;(\tfrac{1}{\omega}\boldsymbol{I}+\boldsymbol{K})^{-1}\,(\tfrac{1-\omega}{\omega}\boldsymbol{I}-\boldsymbol{K}^*)\,\boldsymbol{D}^{\frac{1}{2}}\,,\end{aligned}$$

where $\boldsymbol{K} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{L}\boldsymbol{D}^{-\frac{1}{2}}$. Hence $\boldsymbol{E}_\omega^{\text{SSOR}}$ is similar to

$$\boldsymbol{S}_\omega := (\tfrac{1}{\omega}\,\boldsymbol{I}+\boldsymbol{K}^*)^{-1}\,(\tfrac{1-\omega}{\omega}\boldsymbol{I}-\boldsymbol{K})\;(\tfrac{1}{\omega}\boldsymbol{I}+\boldsymbol{K})^{-1}\,(\tfrac{1-\omega}{\omega}\boldsymbol{I}-\boldsymbol{K}^*)\,,$$

which further is similar to

$$(\tfrac{1}{\omega}\boldsymbol{I}+\boldsymbol{K}^*)\,\boldsymbol{S}_\omega\,(\tfrac{1}{\omega}\boldsymbol{I}+\boldsymbol{K}^*)^{-1} = (\tfrac{1}{\omega}\boldsymbol{I}+\boldsymbol{K})^{-1}\,(\tfrac{1-\omega}{\omega}\boldsymbol{I}-\boldsymbol{K})\,(\tfrac{1-\omega}{\omega}\boldsymbol{I}-\boldsymbol{K}^*)\,(\tfrac{1}{\omega}\boldsymbol{I}+\boldsymbol{K}^*)^{-1}\,,$$

i.e., a positive semidefinite matrix. Hence the spectrum of $\boldsymbol{E}_\omega^{\text{SSOR}}$ is nonnegative. Set $\boldsymbol{B} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2} = \boldsymbol{K}+\boldsymbol{I}+\boldsymbol{K}^*$. Clearly $\boldsymbol{B}$ is positive definite. Since

$$(\tfrac{1}{\omega}\boldsymbol{I}+\boldsymbol{K}^*)^{-1}\,(\tfrac{1-\omega}{\omega}\boldsymbol{I}-\boldsymbol{K}) = (\tfrac{1}{\omega}\boldsymbol{I}+\boldsymbol{K}^*)^{-1}\,(\tfrac{1}{\omega}\boldsymbol{I}+\boldsymbol{K}^*-\boldsymbol{K}^*-\boldsymbol{I}-\boldsymbol{K}) = \boldsymbol{I}-(\tfrac{1}{\omega}\boldsymbol{I}+\boldsymbol{K}^*)^{-1}\boldsymbol{B}\,,$$

[2] Here we use the following result: if $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ p and q are polynomials, $\det(q(\boldsymbol{A})) \neq 0$, then $p(\boldsymbol{A})q(\boldsymbol{A}) = q(\boldsymbol{A})p(\boldsymbol{A})$ and $p(\boldsymbol{A})q(\boldsymbol{A})^{-1} = q(\boldsymbol{A})^{-1}p(\boldsymbol{A})$.

we get

$$\boldsymbol{S}_\omega = [\boldsymbol{I} - (\tfrac{1}{\omega}\boldsymbol{I} + \boldsymbol{K}^*)^{-1}\boldsymbol{B}]\,[\boldsymbol{I} - (\tfrac{1}{\omega}\boldsymbol{I} + \boldsymbol{K})^{-1}\boldsymbol{B}]\ .$$

Hence $\boldsymbol{S}_\omega$ is similar to

$$\begin{aligned}
&\boldsymbol{B}^{1/2}\boldsymbol{S}\boldsymbol{B}^{-1/2} = [\boldsymbol{I} - \boldsymbol{B}^{1/2}(\tfrac{1}{\omega}\boldsymbol{I} + \boldsymbol{K}^*)^{-1}\boldsymbol{B}^{1/2}]\,[\boldsymbol{I} - \boldsymbol{B}^{1/2}(\tfrac{1}{\omega}\boldsymbol{I} + \boldsymbol{K})^{-1}\boldsymbol{B}^{1/2}]\\
&= \boldsymbol{I} - \boldsymbol{B}^{1/2}[(\tfrac{1}{\omega}\boldsymbol{I} + \boldsymbol{K}^*)^{-1} + (\tfrac{1}{\omega}\boldsymbol{I} + \boldsymbol{K})^{-1} - (\tfrac{1}{\omega}\boldsymbol{I} + \boldsymbol{K}^*)^{-1}\boldsymbol{B}(\tfrac{1}{\omega}\boldsymbol{I} + \boldsymbol{K})^{-1}]\boldsymbol{B}^{1/2}\\
&= \boldsymbol{I} - \boldsymbol{B}^{1/2}(\tfrac{1}{\omega}\boldsymbol{I} + \boldsymbol{K}^*)^{-1}\,[\tfrac{1}{\omega}\boldsymbol{I} + \boldsymbol{K} + \tfrac{1}{\omega}\boldsymbol{I} + \boldsymbol{K}^* - \boldsymbol{I} - \boldsymbol{K} - \boldsymbol{K}^*]\,(\tfrac{1}{\omega}\boldsymbol{I} + \boldsymbol{K})^{-1}\boldsymbol{B}^{1/2}\\
&= \boldsymbol{I} - (\tfrac{2}{\omega} - 1)\boldsymbol{B}^{1/2}(\tfrac{1}{\omega}\boldsymbol{I} + \boldsymbol{K}^*)^{-1}\,(\tfrac{1}{\omega}\boldsymbol{I} + \boldsymbol{K})^{-1}\boldsymbol{B}^{1/2}\ .
\end{aligned}$$

Here a positive definite matrix is subtracted from $\boldsymbol{I}$. Thus the eigenvalues of $\boldsymbol{S}_\omega$ are less than one. □

10.4. **Tshebyshev–iteration.** Assume that for the problem $\boldsymbol{Ax} = \boldsymbol{b}$ we have obtained some splitting $\boldsymbol{A} = \boldsymbol{M} - \boldsymbol{N}$ and the corresponding iteration:

$$\boldsymbol{M}\boldsymbol{x}_{k+1} = \boldsymbol{N}\boldsymbol{x}_k + \boldsymbol{b} \quad , \qquad \boldsymbol{E} = \boldsymbol{M}^{-1}\boldsymbol{N}\ .$$

Let us try to accelerate the iteration in the form

$$\boldsymbol{y}_k = \sum_{j=0}^{k} \gamma_{k,j}\ \boldsymbol{x}_j$$

trying to choose the coefficients $\gamma_{k,j}$ such that $\boldsymbol{y}_k$ would be a better approximation to the solution. If $\boldsymbol{x}_0$ already happened to be the solution then $\boldsymbol{x}_k = \boldsymbol{x}$ for all k, so that naturally we then want also $\boldsymbol{y}_k = \boldsymbol{x}$. This we get by requiring $\sum_{j=0}^{k} \gamma_{k,j} = 1$ for all k. Then, since $\boldsymbol{x}_j - \boldsymbol{x} = \boldsymbol{E}^j(\boldsymbol{x}_0 - \boldsymbol{x}) = \boldsymbol{E}^j\boldsymbol{e}_0$, we obtain that the error of $\boldsymbol{y}_k$ satisfies

$$\boldsymbol{y}_k - \boldsymbol{x} = \sum_{j=0}^{k} \gamma_{k,j}\ (\boldsymbol{x}_j - \boldsymbol{x}) = \sum_{j=0}^{k} \gamma_{k,j}\ \boldsymbol{E}^j\boldsymbol{e}_0 = p_k(\boldsymbol{E})\boldsymbol{e}_0\ , \tag{10.5}$$

where p_k is a polynomial of degree k such that $p_k(1) = 1$. Naturally we would like to have p_k such that $\|p_k(\boldsymbol{E})\|_2$ is small. Consider here only the case where $\boldsymbol{E}$ is Hermitian and assume further that $\rho(\boldsymbol{E}) < 1$, i.e., that the basic iteration converges, and that we know some bounds for the spectrum of $\boldsymbol{E}$:

$$-1 < \alpha \le \lambda_1 \le \cdots \le \lambda_n \le \beta < 1\ .$$

Then $\boldsymbol{E}$ is unitarily similar to a diagonal matrix: $\boldsymbol{E} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^*$, so that $p_k(\boldsymbol{E}) = \boldsymbol{Q}p_k(\boldsymbol{\Lambda})\boldsymbol{Q}^*$. We obtain:

$$\|p_k(\boldsymbol{E})\|_2 = \max_{\lambda\in\Lambda(\boldsymbol{E})} |p_k(\lambda)| \le \max_{\lambda\in[\alpha,\beta]} |p_k(\lambda)|\ .$$

Problem 10.10. Corresponding to this what do you get for the case of SSOR when $\boldsymbol{A}$ is Hermitian. Then $\boldsymbol{E}$ is not symmetric, but it is diagonalizable and has real eigenvalues.

Now our task is to find a polynomial p_k such that $p_k(1) = 1$ and $\max_{\lambda\in[\alpha,\beta]} |p_k(\lambda)|$ is the smallest possible. The solution is found from the Tshebyshev polynomials:

$$t_k(\tau) = \cos(k \arccos(\tau)) \ .$$

Lemma 10.12. *If q_k is a polynomial of degree at most k and satisfies $|q_k(\tau)| \le 1$ for all $\tau \in [-1, 1]$, then*

$$q_k(s) \le t_k(s) \qquad \textit{for all} \qquad s > 1 \ .$$

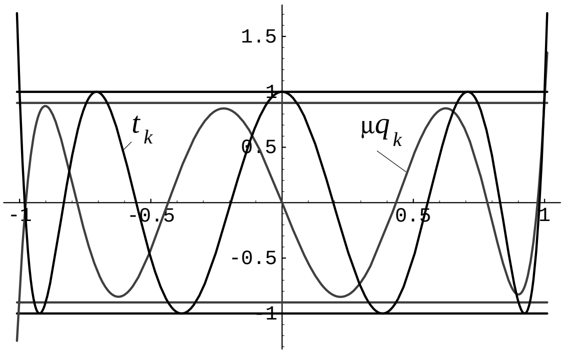
Proof. Assume the contrary: let q_k satisfy the assumptions but for some $s > 1$ holds $q_k(s) > t_k(s)$. Let $\mu < 1$ be such that $\mu q_k(s) > t_k(s)$. Now

$$t_k(\tau_j) = (-1)^j \ , \qquad \text{where} \qquad \tau_j = \cos(j\pi/k)\ ,\ j = 0, \dots, k$$

Since $|\mu q_k(\tau)| \le \mu$ for all $\tau \in [-1, 1]$, setting

$$r(\tau) = \mu q_k(\tau) - t_k(\tau)$$

we get a polynomial of degree k which changes its sign on each interval (τ_{j+1}, τ_j). Hence r has k zeros on the interval $[-1, 1]$ and $r(1) < 0$. Moreover, $r(s) > 0$, so that r has $k+1$ zeros, which is impossible.



□

By mapping the interval $[\alpha, \beta]$ linearly onto the interval $[-1, 1]$ and scaling gives us the best possible polynomials:

$$p_k(\tau) = t_k\left(\tfrac{2}{\beta-\alpha}\,\tau - \tfrac{\alpha+\beta}{\beta-\alpha}\right)/t_k\left(\tfrac{2}{\beta-\alpha} - \tfrac{\alpha+\beta}{\beta-\alpha}\right) = \tfrac{1}{c_k}\,t_k(\rho\,\tau - \sigma) \ ,$$

where $\rho = \frac{2}{\beta-\alpha}$, $\sigma = \frac{\alpha+\beta}{\beta-\alpha}$ and $c_k = t_k(\rho - \sigma) = t_k(\mu)$, $\mu = \rho - \sigma$.

Formulas

$$\begin{aligned}\cos((k+1)\theta) &= \cos\theta\cos k\theta - \sin\theta\sin k\theta\\ \cos((k-1)\theta) &= \cos\theta\cos k\theta + \sin\theta\sin k\theta\end{aligned}$$

imply

$$\cos((k+1)\theta) + \cos((k-1)\theta) = 2\cos\theta\cos k\theta \ ,$$

and substitution $\theta = \arccos\tau$ gives the following recursion for the Tshebyshev polynomials:

$$t_{k+1}(\tau) = 2\,\tau\, t_k(\tau) - t_{k-1}(\tau) \ , \qquad t_0(\tau) = 1 \ , \quad t_1(\tau) = \tau \ .$$

Using this we get the iteration directly for the $\boldsymbol{y}_k$ vectors without computing the basic iteration. From equation (10.5) we get

$$\boldsymbol{y}_k = p_k(\boldsymbol{E})\,\boldsymbol{e}_0 + \boldsymbol{x} = \tfrac{1}{c_k}\,t_k(\rho\,\boldsymbol{E} - \sigma\boldsymbol{I})\,\boldsymbol{e}_0 + \boldsymbol{x}$$

and from this $t_k(\rho\,\boldsymbol{E} - \sigma\boldsymbol{I})\boldsymbol{e}_0 = c_k(\boldsymbol{y}_k - \boldsymbol{x})$. This and the recursion for the Tshebyshev polynomials give

$$\begin{aligned}
\boldsymbol{y}_{k+1} =& \tfrac{1}{c_{k+1}} t_{k+1}(\rho\,\boldsymbol{E} - \sigma\boldsymbol{I})\boldsymbol{e}_0 + \boldsymbol{x} \\
=& \tfrac{1}{c_{k+1}}\big(2(\rho\,\boldsymbol{E} - \sigma\boldsymbol{I})\ t_k(\rho\,\boldsymbol{E} - \sigma\boldsymbol{I}) - t_{k-1}(\rho\,\boldsymbol{E} - \sigma\boldsymbol{I})\big)\ \boldsymbol{e}_0 + \boldsymbol{x} \\
=& \tfrac{1}{c_{k+1}}\big(2(\rho\,\boldsymbol{E} - \sigma)\ c_k(\boldsymbol{y}_k - \boldsymbol{x}) - c_{k-1}(\boldsymbol{y}_{k-1} - \boldsymbol{x})\big) + \boldsymbol{x} \\
=& \tfrac{2c_k}{c_{k+1}}(\rho\,\boldsymbol{E} - \sigma)\ \boldsymbol{y}_k - \tfrac{c_{k-1}}{c_{k+1}}\boldsymbol{y}_{k-1} + \big(1 + \tfrac{2\sigma c_k}{c_{k+1}} + \tfrac{c_{k-1}}{c_{k+1}}\big)\boldsymbol{x} - \tfrac{2\rho c_k}{c_{k+1}}\boldsymbol{E}\boldsymbol{x} \ .
\end{aligned}$$

Equations $\boldsymbol{E}\boldsymbol{x} = \boldsymbol{M}^{-1}\boldsymbol{N}\boldsymbol{x} = \boldsymbol{M}^{-1}(\boldsymbol{M} - \boldsymbol{A})\boldsymbol{x} = \boldsymbol{x} - \boldsymbol{M}^{-1}\boldsymbol{b}$ and

$$c_{k+1} + 2\sigma c_k + c_{k-1} - 2\rho c_k = c_{k+1} + c_{k-1} - 2\mu c_k = 0$$

imply

$$\boldsymbol{y}_{k+1} = \tfrac{2\rho c_k}{c_{k+1}}\ (\boldsymbol{E}\ \boldsymbol{y}_k + \boldsymbol{M}^{-1}\boldsymbol{b}) - \tfrac{2\sigma c_k}{c_{k+1}}\ \boldsymbol{y}_k - \tfrac{c_{k-1}}{c_{k+1}}\boldsymbol{y}_{k-1} \ .$$

Denote

$$\boldsymbol{z}_k = \boldsymbol{M}^{-1}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{y}_k) = \boldsymbol{M}^{-1}\boldsymbol{b} + \boldsymbol{E}\boldsymbol{y}_k - \boldsymbol{y}_k \ ,$$

$$\gamma = \tfrac{\rho}{\rho-\sigma} = \tfrac{2}{2-\alpha-\beta} \qquad \text{and} \qquad \omega_k = 2\mu\tfrac{c_k}{c_{k+1}} \ .$$

Then

$$\tfrac{c_{k-1}}{c_{k+1}} = \tfrac{2\mu c_k - c_{k+1}}{c_{k+1}} = \omega_k - 1$$

and the algorithm can be written:

Tshebyshev iteration:

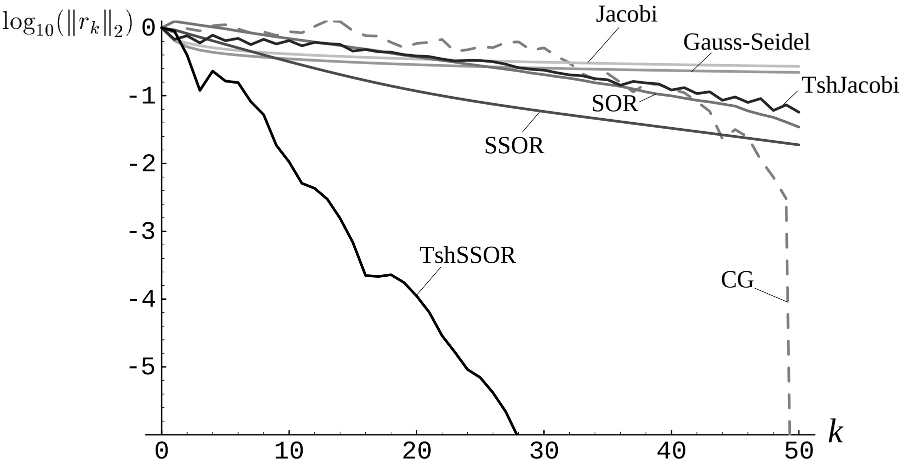
```
$\mu = (2-\alpha-\beta)/(\beta-\alpha)\,,\ \gamma = 2/(2-\alpha-\beta)$
$\boldsymbol{y}_0 = \boldsymbol{x}_0\,,\ \boldsymbol{y}_1 = \gamma\,\boldsymbol{x}_1$
for $k = 1, 2, \ldots$
    $\boldsymbol{M}\,\boldsymbol{z}_k = \boldsymbol{b} - \boldsymbol{A}\,\boldsymbol{y}_k$
    $\omega_k = 2\,\mu\,t_k(\mu)/t_{k+1}(\mu)$
    $\boldsymbol{y}_{k+1} = \omega_k\,(\boldsymbol{y}_k - \boldsymbol{y}_{k-1} + \gamma\,\boldsymbol{z}_k) + \boldsymbol{y}_{k-1}$
end
```

Problem 10.11. In fact we don't need the Tshebyshev polynomials here (except as a theoretical tool): show, that the numbers ω_k satisfy:

$$\omega_0 = 2 \quad , \qquad \omega_{k+1} = \frac{1}{1-\omega_k/(4\mu^2)} \qquad \text{and} \qquad \omega_{k+1} < \omega_k\,.$$

In the following picture we have solved the problem of example 10.2 in the case $n = 50$ with Jacobi, Gauss–Seidel, SOR, SSOR, and with Tshebyshev accelerations of Jacobi (TshJacobi) and SSOR (TshSSOR). It is clear from the picture which is the fastest. The dashed line shows the speed of the conjugate gradient method, which here in the unpreconditioned form does not reveal its power yet.

The picture is not completely fair, since workwise the Jacobi and TshJacobi are clearly the cheapest per step, Gauss–Seidel and SOR have about the same work per step, while the work of SSOR and TshSSOR is about twice that much.



Problem 10.12. The symmetric SOR is not directly in the form $\boldsymbol{M}\boldsymbol{x}_{k+1} = \boldsymbol{N}\boldsymbol{x}_k + \boldsymbol{b}$, so that Tshebyshev iteration cannot be immediately written in the form above.

How to modify it to become suitable? (A couple of rows in the place of $\boldsymbol{M}\boldsymbol{z}_k = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{y}_k\,.$)

Problem 10.13. Experiment the Tshebyshev acceleration of the Jacobi iteration and the symmetric SOR for problem 10.9.

11. The conjugate gradient method

In this section we consider iterative solution of an $n \times n$ linear system $\boldsymbol{Ax} = \boldsymbol{b}$ in the case where $\boldsymbol{A}$ is Hermitian and positive definite. This presentation derives the conjugate gradient method from the Lanczos algorithm, so that the properties of the latter become available immediately. The classical way of deriving the conjugate gradient method is outlined in a sequence of problems.

Consider the minimization of the function $f : \mathbb{C}^n \to \mathbb{R}$

$$f(\boldsymbol{x}) = \boldsymbol{x}^* \boldsymbol{A}\boldsymbol{x} - \boldsymbol{x}^*\boldsymbol{b} - \boldsymbol{b}^*\boldsymbol{x} \ . \tag{11.1}$$

Calculation

$$\begin{aligned} f(\boldsymbol{A}^{-1}\boldsymbol{b} + \boldsymbol{d}) &= (\boldsymbol{A}^{-1}\boldsymbol{b} + \boldsymbol{d})^* \boldsymbol{A}(\boldsymbol{A}^{-1}\boldsymbol{b} + \boldsymbol{d}) - (\boldsymbol{A}^{-1}\boldsymbol{b} + \boldsymbol{d})^*\boldsymbol{b} - \boldsymbol{b}^*(\boldsymbol{A}^{-1}\boldsymbol{b} + \boldsymbol{d}) \\ &= \boldsymbol{b}^*\boldsymbol{A}^{-1}(\boldsymbol{b} + \boldsymbol{A}\boldsymbol{d}) + \boldsymbol{d}^*(\boldsymbol{b} + \boldsymbol{A}\boldsymbol{d}) - 2\boldsymbol{b}^*\boldsymbol{A}^{-1}\boldsymbol{b} - \boldsymbol{d}^*\boldsymbol{b} - \boldsymbol{b}^*\boldsymbol{d} \\ &= \boldsymbol{d}^*\boldsymbol{A}\boldsymbol{d} - \boldsymbol{b}^*\boldsymbol{A}^{-1}\boldsymbol{b} \end{aligned}$$

shows that $\boldsymbol{x}_* = \boldsymbol{A}^{-1}\boldsymbol{b}$ is a strict minimum of f and the substitution $\boldsymbol{d} = -\boldsymbol{A}^{-1}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})$ gives

$$f(\boldsymbol{x}) = (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})^*\boldsymbol{A}^{-1}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}) - \boldsymbol{b}^*\boldsymbol{A}^{-1}\boldsymbol{b} \ . \tag{11.2}$$

11.1. **Minimization in the Krylov subspace.** Let us start searching the minimum of f in the Krylov subspace $\mathcal{K}_j = V(\boldsymbol{b}, \boldsymbol{A}\boldsymbol{b}, \ldots, \boldsymbol{A}^{j-1}\boldsymbol{b})$. Recall that the columns of the matrix

$$\boldsymbol{Q}_j = \begin{bmatrix} \boldsymbol{q}_1 & \cdots & \boldsymbol{q}_j \end{bmatrix} = \begin{bmatrix} \boldsymbol{b} & \boldsymbol{A}\boldsymbol{b} & \ldots & \boldsymbol{A}^{j-1}\boldsymbol{b} \end{bmatrix} \boldsymbol{R}^{-1}$$

form an orthonormal basis of $\mathcal{K}_j$ and that

$$\boldsymbol{A}\boldsymbol{Q}_j = \boldsymbol{Q}_j\boldsymbol{T}_j + \widetilde{\boldsymbol{r}}_j\boldsymbol{e}_j^T \ ,$$

where $\boldsymbol{T}_j$ is the Hermitian tridiagonal matrix produced by the Lanczos algorithm. We added the tilde: $\widetilde{\boldsymbol{r}}_j$ here, since $\boldsymbol{r}$ will be needed for other purposes.

So, the task is to minimize

$$f(\boldsymbol{Q}_j\boldsymbol{v}) = \boldsymbol{v}^*\boldsymbol{Q}_j^*\boldsymbol{A}\boldsymbol{Q}_j\boldsymbol{v} - \boldsymbol{v}^*\boldsymbol{Q}_j^*\boldsymbol{b} - \boldsymbol{b}^*\boldsymbol{Q}_j\boldsymbol{v}$$

with respect to $\boldsymbol{v}$. This has the solution: $\boldsymbol{v}_j = (\boldsymbol{Q}_j^*\boldsymbol{A}\boldsymbol{Q}_j)^{-1}\boldsymbol{Q}_j^*\boldsymbol{b} = \boldsymbol{T}_j^{-1}\boldsymbol{Q}_j^*\boldsymbol{b}$, $\boldsymbol{x}_j = \boldsymbol{Q}_j\boldsymbol{T}_j^{-1}\boldsymbol{Q}_j^*\boldsymbol{b}$ and the corresponding residual satisfies:

$$\boldsymbol{r}_j = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_j = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{Q}_j\boldsymbol{T}_j^{-1}\boldsymbol{Q}_j^*\boldsymbol{b} = (\boldsymbol{I} - \boldsymbol{Q}_j\boldsymbol{Q}_j^*)\boldsymbol{b} - \widetilde{\boldsymbol{r}}_j\boldsymbol{e}_j^T\boldsymbol{T}_j^{-1}\boldsymbol{Q}_j^*\boldsymbol{b} = -\boldsymbol{\eta}_j^*\boldsymbol{b}\widetilde{\boldsymbol{r}}_j \ ,$$

[0] Version: April 9, 2003

where $\boldsymbol{\eta}_j = \boldsymbol{Q}_j\boldsymbol{T}_j^{-1}\boldsymbol{e}_j$. In particular, the residuals are orthogonal. Since in the Lanczos $\beta_j = \|\widetilde{\boldsymbol{r}}_j\|$, we get

$$\boldsymbol{q}_{j+1} = \widetilde{\boldsymbol{r}}_j/\beta_j = \frac{-1}{\beta_j\boldsymbol{\eta}_j^*\boldsymbol{b}}\,\boldsymbol{r}_j \qquad \text{and} \qquad \beta_j = \frac{\|\boldsymbol{r}_j\|}{|\boldsymbol{\eta}_j^*\boldsymbol{b}|}\ .$$

The vectors $\boldsymbol{\eta}_j$ satisfy:

$$\boldsymbol{A}\boldsymbol{\eta}_j = \boldsymbol{A}\boldsymbol{Q}_j\boldsymbol{T}_j^{-1}\boldsymbol{e}_j = \boldsymbol{Q}_j\boldsymbol{e}_j + \boldsymbol{e}_j^T\boldsymbol{T}^{-1}\boldsymbol{e}_j\widetilde{\boldsymbol{r}}_j \in V(\boldsymbol{q}_j, \boldsymbol{q}_{j+1})\ ,$$

so that these are $\boldsymbol{A}$–orthogonal: $k < j \implies$

$$\boldsymbol{\eta}_k^*\boldsymbol{A}\boldsymbol{\eta}_j = \boldsymbol{e}_k^T\boldsymbol{T}_k^{-1}\boldsymbol{Q}_k^*[\boldsymbol{q}_j + \boldsymbol{e}_j^T\boldsymbol{T}^{-1}\boldsymbol{e}_j\beta_j\boldsymbol{q}_{j+1}] = 0\ .$$

Since

$$\begin{aligned}
\boldsymbol{T}_{j+1}\boldsymbol{Q}_{j+1}^*(\boldsymbol{x}_{j+1} - \boldsymbol{x}_j) &= \boldsymbol{T}_{j+1}\boldsymbol{Q}_{j+1}^*\boldsymbol{Q}_{j+1}\boldsymbol{T}_{j+1}^{-1}\boldsymbol{Q}_{j+1}^*\boldsymbol{b} - \boldsymbol{T}_{j+1}\begin{bmatrix}\boldsymbol{Q}_j^*\\ \boldsymbol{q}_{j+1}^*\end{bmatrix}\boldsymbol{Q}_j\boldsymbol{T}_j^{-1}\boldsymbol{Q}_j^*\boldsymbol{b}\\
&= \boldsymbol{Q}_{j+1}^*\boldsymbol{b} - \begin{bmatrix}\boldsymbol{T}_j & \beta_j\boldsymbol{e}_j\\ \beta_j\boldsymbol{e}_j^T & \alpha_{j+1}\end{bmatrix}\begin{bmatrix}\boldsymbol{T}_j^{-1}\boldsymbol{Q}_j^*\boldsymbol{b}\\ 0\end{bmatrix}\\
&= \begin{bmatrix}\boldsymbol{Q}_j^*\boldsymbol{b}\\ \boldsymbol{q}_{j+1}^*\boldsymbol{b}\end{bmatrix} - \begin{bmatrix}\boldsymbol{Q}_j^*\boldsymbol{b}\\ \beta_j\boldsymbol{\eta}_j^*\boldsymbol{b}\end{bmatrix} = (\boldsymbol{q}_{j+1}^*\boldsymbol{b} - \beta_j\boldsymbol{\eta}_j^*\boldsymbol{b})\boldsymbol{e}_{j+1} = \gamma_j\boldsymbol{e}_{j+1}
\end{aligned}$$

and since $\boldsymbol{x}_{j+1} - \boldsymbol{x}_j \in R(\boldsymbol{Q}_{j+1})$, we get

$$\boldsymbol{x}_{j+1} = \boldsymbol{x}_j + \gamma_j\boldsymbol{\eta}_{j+1}\ . \tag{11.3}$$

Let us derive a recursion for the vectors $\boldsymbol{\eta}_j$: if $\boldsymbol{T}_j\boldsymbol{w}_j = \boldsymbol{e}_j$, then

$$\begin{bmatrix}\boldsymbol{T}_j & \beta_j\boldsymbol{e}_j\\ \beta_j\boldsymbol{e}_j^T & \alpha_{j+1}\end{bmatrix}\begin{bmatrix}-\beta_j\boldsymbol{w}_j\\ 1\end{bmatrix} = \begin{bmatrix}0\\ \alpha_{j+1} - \beta_j^2\boldsymbol{e}_j^T\boldsymbol{w}_j\end{bmatrix} = d_j\boldsymbol{e}_{j+1}\ ,$$

so that $\boldsymbol{w}_{j+1} = \frac{1}{d_j}\begin{bmatrix}-\beta_j\boldsymbol{w}_j\\ 1\end{bmatrix}$, where $d_j = \alpha_{j+1} - \beta_j^2\boldsymbol{e}_j^T\boldsymbol{w}_j$. We get:

$$\boldsymbol{\eta}_{j+1} = \boldsymbol{Q}_{j+1}\boldsymbol{w}_{j+1} = \frac{1}{d_j}\begin{bmatrix}\boldsymbol{Q}_j & \boldsymbol{q}_{j+1}\end{bmatrix}\begin{bmatrix}-\beta_j\boldsymbol{w}_j\\ 1\end{bmatrix} = \frac{-\beta_j}{d_j}\boldsymbol{\eta}_j + \frac{1}{d_j}\boldsymbol{q}_{j+1}\ .$$

Since $\boldsymbol{q}_1 = \boldsymbol{b}/\|\boldsymbol{b}\|$, we also see that

$$\begin{aligned}
\boldsymbol{\eta}_j^*\boldsymbol{b} &= \boldsymbol{e}_j^T\boldsymbol{T}_j^{-1}\boldsymbol{Q}_j^*\boldsymbol{b} = \|\boldsymbol{b}\|\,\boldsymbol{w}_j^T\boldsymbol{e}_1\\
\boldsymbol{\eta}_{j+1}^*\boldsymbol{b} &= \|\boldsymbol{b}\|\,\boldsymbol{w}_{j+1}^T\boldsymbol{e}_1 = -\|\boldsymbol{b}\|\,\frac{\beta_j}{d_j}\boldsymbol{w}_j^T\boldsymbol{e}_1 = \frac{-\beta_j}{d_j}\boldsymbol{\eta}_j^*\boldsymbol{b}\ .
\end{aligned}$$

We can write a recursion in nicer form for the vectors

$$\begin{aligned}
\boldsymbol{p}_j :=& d_j^2\ \boldsymbol{\eta}_{j+1}^*\boldsymbol{b}\ \boldsymbol{\eta}_{j+1} = d_j\ \beta_j\ \boldsymbol{\eta}_j^*\boldsymbol{b}\ \Big[\ \frac{\beta_j}{d_j}\ \boldsymbol{\eta}_j + \frac{1}{d_j\beta_j\boldsymbol{\eta}_j^*\boldsymbol{b}}\ \boldsymbol{r}_j\ \Big]\\
=& \beta_j^2\ \boldsymbol{\eta}_j^*\boldsymbol{b}\ \boldsymbol{\eta}_j + \boldsymbol{r}_j = \frac{\beta_j^2}{d_{j-1}^2}\ \boldsymbol{p}_{j-1} + \boldsymbol{r}_j\ .
\end{aligned} \tag{11.4}$$

Then (11.3) implies:

$$\boldsymbol{x}_{j+1} = \boldsymbol{x}_j + \rho_j \boldsymbol{p}_j \qquad \text{and} \qquad \boldsymbol{r}_{j+1} = \boldsymbol{r}_j - \rho_j \boldsymbol{A}\boldsymbol{p}_j \ . \tag{11.5}$$

Since $\boldsymbol{\eta}_{j-1} \in R(\boldsymbol{Q}_j)$ and $\boldsymbol{r}^*_{j-1}\boldsymbol{Q}_{j-1} = 0$, we obtain

$$\boldsymbol{r}^*_{j-1}\boldsymbol{\eta}_j = \frac{1}{d_{j-1}}\,\boldsymbol{r}^*_{j-1}\boldsymbol{q}_j = \frac{-\left\|\boldsymbol{r}_{j-1}\right\|^2}{d_{j-1}\beta_{j-1}\boldsymbol{\eta}^*_{j-1}\boldsymbol{b}}\ ,$$

from which

$$d^2_{j-1} = \frac{\left\|\boldsymbol{r}_{j-1}\right\|^4}{\left|\beta_{j-1}\,\boldsymbol{r}^*_{j-1}\boldsymbol{\eta}_j\;\boldsymbol{\eta}^*_{j-1}\boldsymbol{b}\right|^2} = \frac{\left\|\boldsymbol{r}_{j-1}\right\|^2}{\left|\boldsymbol{r}^*_{j-1}\boldsymbol{\eta}_j\right|^2}$$

and the coefficients entering the $\boldsymbol{p}_j$ we get

$$\mu_j := \frac{\beta_j^2}{d^2_{j-1}} = \frac{\left\|\boldsymbol{r}_j\right\|^2\left|\boldsymbol{r}^*_{j-1}\boldsymbol{\eta}_j\right|^2}{\left\|\boldsymbol{r}_{j-1}\right\|^2\left|\boldsymbol{\eta}^*_j\boldsymbol{b}\right|^2} = \frac{\left\|\boldsymbol{r}_j\right\|^2}{\left\|\boldsymbol{r}_{j-1}\right\|^2}\ . \tag{11.6}$$

This is because $\boldsymbol{A}\boldsymbol{\eta}_j \in V(\boldsymbol{q}_j, \boldsymbol{q}_{j+1})$ implies

$$\boldsymbol{\eta}^*_j\boldsymbol{r}_{j-1} = \boldsymbol{\eta}^*_j(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{Q}_{j-1}\boldsymbol{v}_{j-1}) = \boldsymbol{\eta}^*_j\boldsymbol{b}\ .$$

The coefficient ρ_j of equation (11.5) can be computed directly from the condition that $f(\boldsymbol{x}_j + \rho\boldsymbol{p}_j)$ is minimized when ρ is ρ_j, so that

$$\rho_j = \frac{\boldsymbol{p}^*_j\boldsymbol{r}_j}{\boldsymbol{p}^*_j\boldsymbol{A}\boldsymbol{p}_j} = \frac{\left\|\boldsymbol{r}_j\right\|^2}{\boldsymbol{p}^*_j\boldsymbol{A}\boldsymbol{p}_j}\ . \tag{11.7}$$

Collecting the equations (11.4) – (11.7) we get

Conjugate Gradient method :

```
$\boldsymbol{x}_0 = 0,\ \boldsymbol{r}_{-1} = \boldsymbol{r}_0 = \boldsymbol{b},\ \boldsymbol{p}_{-1} = 0$
for   $j = 0, 1, 2, \ldots$
      $\mu_j = \left\|\boldsymbol{r}_j\right\|^2 / \left\|\boldsymbol{r}_{j-1}\right\|^2$
      $\boldsymbol{p}_j = \mu_j\boldsymbol{p}_{j-1} + \boldsymbol{r}_j$
      $\rho_j = \left\|\boldsymbol{r}_j\right\|^2 / \boldsymbol{p}^*_j\boldsymbol{A}\boldsymbol{p}_j$
      $\boldsymbol{x}_{j+1} = \boldsymbol{x}_j + \rho_j\boldsymbol{p}_j$
      $\boldsymbol{r}_{j+1} = \boldsymbol{r}_j - \rho_j\boldsymbol{A}\boldsymbol{p}_j$
end
```

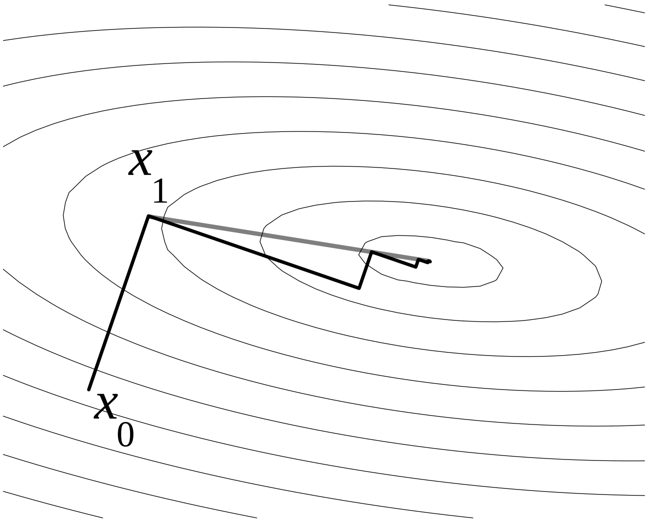
Usually the conjugate gradient method is derived in the following way:

Problem 11.1. The *gradient method* for minimizing the function $f : \mathbb{R}^n \to \mathbb{R}$ is the iteration:

$$\boldsymbol{x}_{j+1} = \boldsymbol{x}_j - \alpha_j \nabla f(\boldsymbol{x}_j) \ , \qquad \text{where} \qquad f(\boldsymbol{x}_{j+1}) = \min_{\alpha>0} f(\boldsymbol{x}_j - \alpha \nabla f(\boldsymbol{x}_j)) \ .$$

Here we go in the direction opposite to the gradient and minimize a function of one real variable. Which iteration this gives for minimizing the function $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{b}$?

In the following picture the steps of the gradient method are in black, those of the conjugate gradients in grey.



Problem 11.2. Let $\boldsymbol{x}_{j+1} = \boldsymbol{x}_j + \rho_j \boldsymbol{p}_j$ be some iteration for minimizing the function $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{b}$, where ρ_j minimizes $f(\boldsymbol{x}_j + \rho \boldsymbol{p}_j)$. Assume, that the directions $\boldsymbol{p}_j$ are $\boldsymbol{A}$–orthogonal: $\boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{p}_j = 0$ for all $k \neq j$. Show that

$$f(\boldsymbol{x}_{j+1}) = \min_{\boldsymbol{v} \in V(\boldsymbol{p}_0, \ldots, \boldsymbol{p}_j)} f(\boldsymbol{x}_0 + \boldsymbol{v}) \ .$$

Problem 11.3. Let $\boldsymbol{p}_0 = \boldsymbol{r}_0$ above and $\boldsymbol{p}_j = \boldsymbol{r}_j + \mu_j \boldsymbol{p}_{j-1}$, where μ_j is chosen such that $\boldsymbol{p}_{j-1}^T \boldsymbol{A} \boldsymbol{p}_j = 0$. Show, that the resulting algorithm is the same as the conjugate gradient method .

The use of the Lanczos process for deriving the conjugate gradient method gives now the following theorem free, otherwise it would require a lengthy proof.

Theorem 11.1. *The conjugate gradient method satisfies:*

a) *the residuals are orthogonal:* $\boldsymbol{r}_k^T \boldsymbol{r}_j = 0$ *for all* $k \neq j$

b) *vectors* $\boldsymbol{p}_j$ *t are* $\boldsymbol{A}$*–orthogonal:* $\boldsymbol{p}_k^T\boldsymbol{A}\boldsymbol{p}_j = 0$ *for all* $k \neq j$

c) $\boldsymbol{x}_j$ *minimizes* $(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})^*\boldsymbol{A}^{-1}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x})$ *in the subspace* $\mathcal{K}_j$

d) *the iteretion takes at most* n *steps:* $\boldsymbol{r}_j = 0$ *, i.e.,* $\boldsymbol{x}_j = \boldsymbol{x}$ *for some* $j \leq n$ *.*

11.2. **Convergence of the conjugate gradient method.** denote: $\|\boldsymbol{x}\|_{\boldsymbol{A}} = \sqrt{\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x}}$. Measuring in this "energy"– norm we get at least the following speed:

Theorem 11.2. $$\|\boldsymbol{x} - \boldsymbol{x}_j\|_{\boldsymbol{A}} \leq 2\,\|\boldsymbol{x}\|_{\boldsymbol{A}} \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^j ,$$
where $\kappa = \lambda_n/\lambda_1$ *is the condition number of* $\boldsymbol{A}$ *.*

Proof. Since $\boldsymbol{x}_j = p_{j-1}(\boldsymbol{A})\boldsymbol{b} \in \mathcal{K}_j$, where p_{j-1} is polynomial of degree at most $j-1$, we get:

$$\begin{aligned}\|\boldsymbol{x} - \boldsymbol{x}_j\|_{\boldsymbol{A}}^2 &= (\boldsymbol{A}^{-1}\boldsymbol{b} - p_{j-1}(\boldsymbol{A})\boldsymbol{b})^*\boldsymbol{A}(\boldsymbol{A}^{-1}\boldsymbol{b} - p_{j-1}(\boldsymbol{A})\boldsymbol{b})\\ &= (\boldsymbol{b} - \boldsymbol{A}p_{j-1}(\boldsymbol{A})\boldsymbol{b})^*\boldsymbol{A}^{-1}(\boldsymbol{b} - \boldsymbol{A}p_{j-1}(\boldsymbol{A})\boldsymbol{b})\\ &= (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_j)^*\boldsymbol{A}^{-1}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_j) = f(\boldsymbol{x}_j) - \boldsymbol{b}^*\boldsymbol{A}^{-1}\boldsymbol{b}.\end{aligned}$$

Hence $\boldsymbol{x}_j$ minimizes also $\|\boldsymbol{x} - \boldsymbol{x}_j\|_{\boldsymbol{A}}$ in the subspace $\mathcal{K}_j$, so that

$$\begin{aligned}\|\boldsymbol{x} - \boldsymbol{x}_j\|_{\boldsymbol{A}} &= \min_{p\in\mathbb{P}_{j-1}} \|\boldsymbol{x} - p(\boldsymbol{A})\boldsymbol{b}\|_{\boldsymbol{A}} = \min_{p\in\mathbb{P}_{j-1}} \left\|\boldsymbol{x} - \boldsymbol{A}p(\boldsymbol{A})\boldsymbol{A}^{-1}\boldsymbol{b}\right\|_{\boldsymbol{A}}\\ &= \min_{p\in\mathbb{P}_{j-1}} \|\boldsymbol{x} - \boldsymbol{A}p(\boldsymbol{A})\boldsymbol{x}\|_{\boldsymbol{A}} = \min_{\substack{p\in\mathbb{P}_j\\ p(0)=1}} \|p(\boldsymbol{A})\boldsymbol{x}\|_{\boldsymbol{A}} \ .\end{aligned}$$

Let $\boldsymbol{U}$ perform the unitary transformation of $\boldsymbol{A}$ into a diagonal matrix $\boldsymbol{\Lambda} = \boldsymbol{U}^*\boldsymbol{A}\boldsymbol{U}$. Then

$$\begin{aligned}\|p(\boldsymbol{A})\boldsymbol{x}\|_{\boldsymbol{A}}^2 &= \boldsymbol{x}^*\bar{p}(\boldsymbol{A})\boldsymbol{A}p(\boldsymbol{A})\boldsymbol{x} = \boldsymbol{x}^*\boldsymbol{U}\bar{p}(\boldsymbol{\Lambda})\boldsymbol{\Lambda}p(\boldsymbol{\Lambda})\boldsymbol{U}^*\boldsymbol{x}\\ &= \sum_k |p(\lambda_k)|^2\,\lambda_k\,|(\boldsymbol{U}^*\boldsymbol{x})_k|^2 \leq \max_i |p(\lambda_i)|^2 \sum_k \lambda_k\,|(\boldsymbol{U}^*\boldsymbol{x})_k|^2\\ &= \max_i |p(\lambda_i)|^2\,\|\boldsymbol{x}\|_{\boldsymbol{A}}^2 \leq \max_{\lambda\in[\lambda_1,\lambda_n]} |p(\lambda)|^2\,\|\boldsymbol{x}\|_{\boldsymbol{A}}^2 \ .\end{aligned}$$

Again the best polynomial , i.e., the solution of the problem

$$\min_{\substack{p\in\mathbb{P}_j\\ p(0)=1}} \max_{\lambda\in[\lambda_1,\lambda_n]} |p(\lambda)|$$

is found using the Tshebyshev polynomials with change of variables that maps the interval $[\lambda_1,\lambda_n]$ onto $[-1,1]$ (reversing the direction, so that zero is mapped to a number bigger than one)

$$p_j(\tau) = \frac{t_j\left(\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1} - \frac{2}{\lambda_n-\lambda_1}\,\tau\right)}{t_j\left(\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1}\right)} \ .$$

Hence the minimum value is:

$$\min_{\substack{p\in\mathbb{P}_j\\ p(0)=1}} \max_{\lambda\in[\lambda_1,\lambda_n]} |p(\lambda)| = t_j(\tfrac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1})^{-1} = t_j(\tfrac{\kappa+1}{\kappa-1})^{-1} \ ,$$

where $\kappa = \lambda_n/\lambda_1$. By the next problem $\tau > 1 \implies$
$t_j(\tau) > \frac{1}{2}(\tau + \sqrt{\tau^2-1})^j$. Since

$$\frac{\kappa+1}{\kappa-1} + \sqrt{\left(\frac{\kappa+1}{\kappa-1}\right)^2 - 1} = \frac{\kappa+1+2\sqrt{\kappa}}{\kappa-1} = \frac{(\sqrt{\kappa}+1)^2}{(\sqrt{\kappa}+1)(\sqrt{\kappa}-1)} = \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}$$

we get

$$t_j(\frac{\kappa+1}{\kappa-1})^{-1} < 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^j \ ,$$

which implies the claim. □

Problem 11.4. Show $\tau > 1 \implies t_j(\tau) = \dfrac{1}{2}\ [\ (\tau+\sqrt{\tau^2-1})^j + (\tau-\sqrt{\tau^2-1})^j\ ]$. Hint: the t_j recursion.

Problem 11.5. Show, that if $\boldsymbol{A}$ has p distinct eigenvalues, then the conjugate gradient method stops latest at step p. Hint: the previous proof. Since the full method is a continuous function of $\boldsymbol{A}$ this implies also, that the method is fast if the eigenvalues are in small "clusters".

11.3. **Preconditioned conjugate gradients.** Let $\boldsymbol{M}$ be Hermitian and positive definite – some approximation of $\boldsymbol{A}$ – such that the systems $\boldsymbol{My} = \boldsymbol{c}$ are easy to solve. Then $\boldsymbol{M}$ has a Hermitian and positive definite square root $\boldsymbol{S}$: $\boldsymbol{M} = \boldsymbol{S}^2 = \boldsymbol{S}^*\boldsymbol{S}$. Think of solving by the conjugate gradient method the following equation that is equivalent to $\boldsymbol{Ax} = \boldsymbol{b}$

$$\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{x}} = \widetilde{\boldsymbol{b}} \ , \qquad \text{where} \qquad \widetilde{\boldsymbol{A}} = \boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S}^{-1} \ , \ \widetilde{\boldsymbol{x}} = \boldsymbol{Sx} \text{ and } \widetilde{\boldsymbol{b}} = \boldsymbol{S}^{-1}\boldsymbol{b} \ .$$

The iteration for this is:

$$\begin{aligned}
&\widetilde{\boldsymbol{x}}_0 = 0\,,\ \widetilde{\boldsymbol{r}}_{-1} = \widetilde{\boldsymbol{r}}_0 = \widetilde{\boldsymbol{b}}\,,\ \widetilde{\boldsymbol{p}}_{-1} = 0\\
&\qquad \mu_j = \|\widetilde{\boldsymbol{r}}_j\|^2 \,/\, \|\widetilde{\boldsymbol{r}}_{j-1}\|^2 \qquad\qquad \widetilde{\boldsymbol{p}}_j = \mu_j\,\widetilde{\boldsymbol{p}}_{j-1} + \widetilde{\boldsymbol{r}}_j\\
&\qquad \rho_j = \|\widetilde{\boldsymbol{r}}_j\|^2 \,/\widetilde{\boldsymbol{p}}_j^*\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{p}}_j \qquad\quad \widetilde{\boldsymbol{x}}_{j+1} = \boldsymbol{x}_j + \rho_j\,\widetilde{\boldsymbol{p}}_j\\
&\qquad \widetilde{\boldsymbol{r}}_{j+1} = \widetilde{\boldsymbol{r}}_j - \rho_j\,\widetilde{\boldsymbol{A}}\,\widetilde{\boldsymbol{p}}_j
\end{aligned}$$

Now, set $\boldsymbol{x}_j = \boldsymbol{S}^{-1}\widetilde{\boldsymbol{x}}_j$, $\boldsymbol{p}_k = \boldsymbol{S}^{-1}\widetilde{\boldsymbol{p}}_j$. Then

$$\begin{aligned}
&\boldsymbol{r}_j = \boldsymbol{b} - \boldsymbol{Ax}_j = \boldsymbol{S}\widetilde{\boldsymbol{b}} - \boldsymbol{S}\widetilde{\boldsymbol{A}}\boldsymbol{S}\boldsymbol{x}_j = \boldsymbol{S}(\widetilde{\boldsymbol{b}} - \widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{x}}_j) = \boldsymbol{S}\widetilde{\boldsymbol{r}}_j\\
&\beta_j = \boldsymbol{r}_j^*\boldsymbol{S}^{-2}\boldsymbol{r}_j/\boldsymbol{r}_{j-1}^*\boldsymbol{S}^{-2}\boldsymbol{r}_{j-1} \qquad\qquad \boldsymbol{p}_k = \boldsymbol{S}^{-2}\boldsymbol{r}_j + \beta_j\boldsymbol{p}_{j-1}\\
&\rho_j = \boldsymbol{r}_j^*\boldsymbol{S}^{-2}\boldsymbol{r}_j/\boldsymbol{p}_j^*\boldsymbol{A}\boldsymbol{p}_j \qquad\qquad \boldsymbol{x}_{j+1} = \boldsymbol{x}_j + \rho_j\,\boldsymbol{p}_j
\end{aligned}$$

Denoting $\boldsymbol{z}_j = \boldsymbol{S}^{-2}\boldsymbol{r}_j$, i.e., $\boldsymbol{M}\boldsymbol{z}_j = \boldsymbol{r}_j$, we get

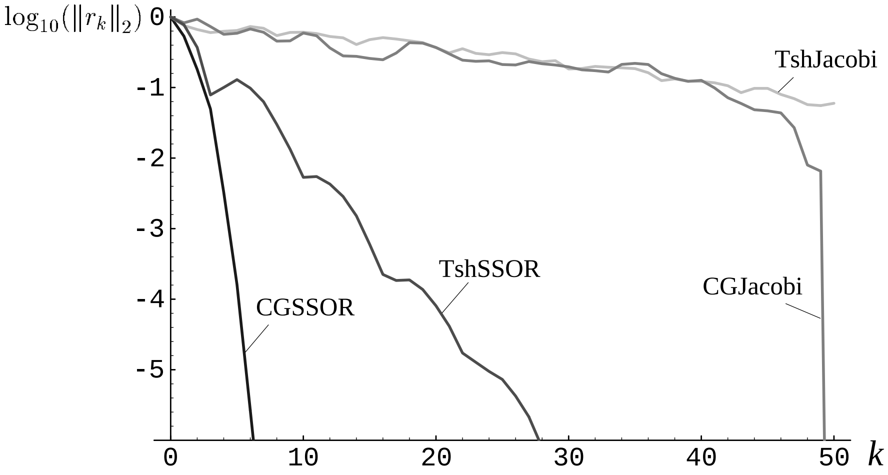
Preconditioned conjugate gradient method :

```
$\boldsymbol{x}_0 = 0\,,\ \boldsymbol{r}_{-1} = \boldsymbol{r}_0 = \boldsymbol{b}\,,\ \boldsymbol{p}_{-1} = 0$
for   $j = 0, 1, 2, \ldots$
      $\boldsymbol{M}\boldsymbol{z}_j = \boldsymbol{r}_j$
      $\mu_j = \boldsymbol{r}_j^*\boldsymbol{z}_j / \boldsymbol{r}_{j-1}^*\boldsymbol{z}_{j-1}$
      $\boldsymbol{p}_j = \mu_j \boldsymbol{p}_{j-1} + \boldsymbol{z}_j$
      $\rho_j = \boldsymbol{r}_j^*\boldsymbol{z}_j / \boldsymbol{p}_j^*\boldsymbol{A}\boldsymbol{p}_j$
      $\boldsymbol{x}_{j+1} = \boldsymbol{x}_j + \rho_j \boldsymbol{p}_j$
      $\boldsymbol{r}_{j+1} = \boldsymbol{r}_j - \rho_j \boldsymbol{A}\boldsymbol{p}_j$
end
```

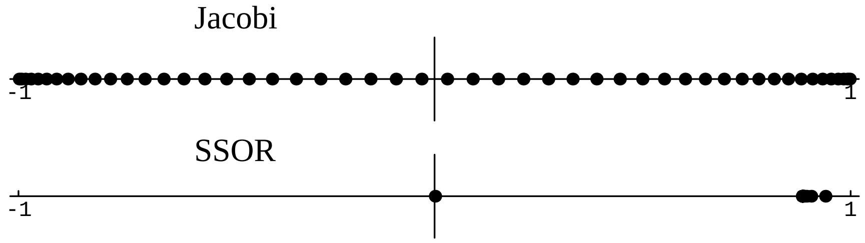
Problem 11.6. Derive the preconditioned conjugate gradient method for the case, where $M^{-1}\boldsymbol{A}$ is Hermitian and positive definite.

Problem 11.7. Prove a version of theorem 11.2 for these preconditioned methods.

In the following picture the Jacobi iteration and SSOR are accelerated by the Tshebyshev iteration and by the conjugate gradient method (i.e., the Jacobi and SSOR matrices M are used as preconditioners). The system is that of example 10.2 matrix.



For Jacobi both accelerators are about as good, while for SSOR the preconditioned conjugate gradient is here clearly better. Explain the reason of this from the spectra of the corresponding iteration matrices in the following figure.



When compared with the Tshebyshev iteration the conjugate gradient method has the advantage that no parameters are needed.

12. Krylov subspace iterations for nonsymmetric systems

In this section we consider approximation of the solution of the general system of equations $\boldsymbol{Ax}=\boldsymbol{b}$ by vectors in the Krylov subspace: $\boldsymbol{x}_j \in \mathcal{K}_j(\boldsymbol{A},\boldsymbol{b})$.

In section 11 the conjugate gradient method was derived from the Lanczos iteration of a Hermitian matrix. Similarly here we get from the Arnoldi iteration the so called *generalized minimal residual method* (GMRes), from the nonsymmetric (biorthogonal) Lanczos the *biconjugate gradient method* (BiCG) and from that, further, the *biconjugate gradient squared method* (BiCGS) and from the look ahead Lanczos the *quasiminimal residual method* (QMR). Hence the iterations for eigenvalues and for systems of equations have much in common:

Lanczos	∼	Conjugate gradient method (CG)
Arnoldi	∼	Generalized minimal residual method (GMRes)
Biorthogonal Lanczos	∼	Biconjugate gradient method (BiCG, BiCGS)
Look ahead Lanczos	∼	Quasimininimum residual method (QMR)

12.1. **Generalized minimal residual method (GMRes).** The GMRes method finds a vector $\boldsymbol{x}_j \in \boldsymbol{x}_0 + \mathcal{K}_j(\boldsymbol{A},\boldsymbol{r}_0)$, that minimizes the norm of the residual $\boldsymbol{r}_j = \boldsymbol{b} - \boldsymbol{Ax}_j$. This is obtained using the orthonormal basis of the Krylov subspace $\mathcal{K}_j(\boldsymbol{A},\boldsymbol{r}_0)$ produced by the Arnoldi process.

Let $\boldsymbol{x}_0$ be an initial approximation for the problem $\boldsymbol{Ax}=\boldsymbol{b}$. Set up the Arnoldi iteration (see section 9.3) with starting vector $\boldsymbol{q}_1 = \boldsymbol{r}_0/\left\|\boldsymbol{r}_0\right\|_2 = \boldsymbol{r}_0/\beta$, where $\boldsymbol{r}_0 = \boldsymbol{b} - \boldsymbol{Ax}_0$. Then we get:

$$\boldsymbol{AQ}_j = \boldsymbol{Q}_j\boldsymbol{H}_j + h_{j+1,j}\boldsymbol{q}_{j+1}\boldsymbol{e}_j^T = \boldsymbol{Q}_{j+1}\widehat{\boldsymbol{H}}_j\ , \quad \text{where} \quad \widehat{\boldsymbol{H}}_j = \begin{bmatrix} \boldsymbol{H}_j \\ {\scriptstyle 0\ldots 0\ h_{j+1,j}} \end{bmatrix} \in \mathbb{C}^{(j+1)\times j}\ .$$

Since the columns of the matrix $\boldsymbol{Q}_j = \begin{bmatrix}\boldsymbol{q}_1 \ldots \boldsymbol{q}_j\end{bmatrix}$ form an orthonormal basis of $\mathcal{K}_j(\boldsymbol{A},\boldsymbol{r}_0)$ we look for the vector $\boldsymbol{x}_j \in \boldsymbol{x}_0 + \mathcal{K}_j(\boldsymbol{A},\boldsymbol{r}_0)$ in the form $\boldsymbol{x}_j = \boldsymbol{x}_0 + \boldsymbol{Q}_j\boldsymbol{\xi}_j$. The solution that minimizes the norm of the residual is obtained from:

$$\begin{aligned}\left\|\boldsymbol{b} - \boldsymbol{A}(\boldsymbol{x}_0 + \boldsymbol{Q}_j\boldsymbol{\xi}_j)\right\|_2 &= \min_{\boldsymbol{\xi}} \left\|\boldsymbol{b} - \boldsymbol{A}(\boldsymbol{x}_0 + \boldsymbol{Q}_j\boldsymbol{\xi})\right\|_2 = \min_{\boldsymbol{\xi}} \left\|\boldsymbol{r}_0 - \boldsymbol{Q}_{j+1}\widehat{\boldsymbol{H}}_j\boldsymbol{\xi}\right\|_2 \\ &= \min_{\boldsymbol{\xi}} \left\|\boldsymbol{Q}_{j+1}(\beta\,\boldsymbol{e}_1 - \widehat{\boldsymbol{H}}_j\boldsymbol{\xi})\right\|_2 = \min_{\boldsymbol{\xi}} \left\|\beta\,\boldsymbol{e}_1 - \widehat{\boldsymbol{H}}_j\boldsymbol{\xi}\right\|_2\ ,\end{aligned}$$

since $\boldsymbol{Q}_{j+1}$ has orthonormal columns. This is a standard least squares problem that can be solved using the $\boldsymbol{QR}$ decomposition. So, let $\widehat{\boldsymbol{H}}_j = \boldsymbol{S}_j\boldsymbol{U}_j$, where the

[0] Version: April 9, 2003

columns of the matrix $\boldsymbol{S}_j \in \mathbb{C}^{(j+1)\times j}$ are orthonormal and $\boldsymbol{U}_j \in \mathbb{C}^{j\times j}$ is an upper triangular matrix. Then by the Pythagorean theorem:

$$\begin{aligned}\left\|\beta\, e_1 - \widehat{\boldsymbol{H}}_j\boldsymbol{\xi}\right\|_2^2 &= \left\|\boldsymbol{S}_j\,(\boldsymbol{S}_j^*\,\beta\,\boldsymbol{e}_1 - \boldsymbol{U}_j\boldsymbol{\xi}) + \beta\,(\boldsymbol{I} - \boldsymbol{S}_j\boldsymbol{S}_j^*)\boldsymbol{e}_1\right\|_2^2 \\ &= \left\|\boldsymbol{S}_j^*\beta\boldsymbol{e}_1 - \boldsymbol{U}_j\boldsymbol{\xi}\right\|_2^2 + \beta\,\left\|(\boldsymbol{I} - \boldsymbol{S}_j\boldsymbol{S}_j^*)\boldsymbol{e}_1\right\|_2^2\ ,\end{aligned}$$

from which $\boldsymbol{\xi}_j = \beta\,\boldsymbol{U}_j^{-1}\boldsymbol{S}_j^*\boldsymbol{e}_1$ and the residual $\boldsymbol{r}_j = \boldsymbol{b} - \boldsymbol{A}(\boldsymbol{x}_0 + \boldsymbol{Q}_j\boldsymbol{\xi}_j)$ satisfies:

$$\|\boldsymbol{r}_j\|_2 = \beta\,\left\|(\boldsymbol{I} - \boldsymbol{S}_j\boldsymbol{S}_j^*)\boldsymbol{e}_1\right\|_2\ .$$

In the GMRes method we need not form the intermediate vectors $\boldsymbol{\xi}_j$, $\boldsymbol{x}_j$ and $\boldsymbol{r}_j$. It suffices to run the Arnoldi process, perform the $\boldsymbol{QR}$ decompositions of the matrices $\widehat{\boldsymbol{H}}_j$, and monitor when $\left\|(\boldsymbol{I} - \boldsymbol{S}_j\boldsymbol{S}_j^*)\boldsymbol{e}_1\right\|_2$ becomes sufficiently small, and then compute the corresponding iterate.

We get the $\boldsymbol{QR}$ decompositions of the $\widehat{\boldsymbol{H}}_j$ matrices inexpensively using the following recursion. If $\widehat{\boldsymbol{H}}_{j-1}$ is brought by Givens rotations to the upper triangular form:

$$\boldsymbol{G}_{j-1,j}\ldots\boldsymbol{G}_{1,2}\widehat{\boldsymbol{H}}_{j-1} = \begin{bmatrix}\boldsymbol{U}_{j-1}\\ 0...0\end{bmatrix}\ ,$$

then for the the following we first get:

$$\boldsymbol{G}_{j-1,j}\ldots\boldsymbol{G}_{1,2}\widehat{\boldsymbol{H}}_j = \boldsymbol{G}_{j-1,j}\ldots\boldsymbol{G}_{1,2}\begin{bmatrix}\widehat{\boldsymbol{H}}_{j-1} & \boldsymbol{w}_j\\ 0...0 & h_{j+1,j}\end{bmatrix} = \begin{bmatrix}\boldsymbol{U}_{j-1} & \\ 0...0 & \widetilde{\boldsymbol{w}}_j\\ 0...0 & h_{j+1,j}\end{bmatrix}\ .$$

Now the next rotation is chosen such that

$$\boldsymbol{G}_{j,j+1}\begin{bmatrix}\widetilde{\boldsymbol{w}}_j\\ h_{j+1,j}\end{bmatrix} = \begin{bmatrix}\boldsymbol{u}_j\\ 0\end{bmatrix}\ .$$

and we set: $\boldsymbol{U}_j = \begin{bmatrix}\boldsymbol{U}_{j-1} & \boldsymbol{u}_j\\ 0...0 & \end{bmatrix}$. In each step we need to apply the rotations only to the new vector $\begin{bmatrix}\boldsymbol{w}_j\\ h_{j+1,j}\end{bmatrix}$.

Define the vectors $\boldsymbol{f}_j$ as follows $\boldsymbol{f}_0 = \beta$, $\boldsymbol{f}_j = \boldsymbol{G}_{j,j+1}\begin{bmatrix}\boldsymbol{f}_{j-1}\\ 0\end{bmatrix} \in \mathbb{C}^{j+1}$. Then $\boldsymbol{\xi}_j$ can finally be computed directly from the triangular system $\boldsymbol{U}_j\boldsymbol{\xi}_j = [\boldsymbol{I}\ 0]\,\boldsymbol{f}_j$.

The monitoring of the convergence is simple when the following result is used:

Problem 12.1. Show that $\left|(\boldsymbol{f}_j)_{j+1}\right| = \beta\ \left\|(\boldsymbol{I} - \boldsymbol{S}_j\boldsymbol{S}_j^*)\boldsymbol{e}_1\right\|_2$.

Hence we get[1]:

GMRes algorithm:

```
$\boldsymbol{r}_0 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0$ , $\beta = \|\boldsymbol{r}_0\|$ , $\boldsymbol{q}_1 = \boldsymbol{r}_0/\beta$ , $\boldsymbol{Q}_1 = [\boldsymbol{q}_1]$
        $\boldsymbol{f}_0 = \beta\boldsymbol{e}_1$ , $\boldsymbol{U}_0 = [\,]$ , ready = false , $j = 1$
  while not ready
        $\boldsymbol{p}_j = \boldsymbol{A}\boldsymbol{q}_j$
        for  $k = 1, ..., j$ ,   $h_{k,j} = \boldsymbol{q}_k^* \boldsymbol{p}_j$ ,   $\boldsymbol{p}_j = \boldsymbol{p}_j - h_{k,j}\boldsymbol{q}_k$   end
        $\boldsymbol{w}_j = h_{1:j,j}$ , $h_{j+1,j} = \|\boldsymbol{p}_j\|_2$
        if $h_{j+1,j} \neq 0$ then   $\boldsymbol{q}_{j+1} = \boldsymbol{p}_j / h_{j+1,j}$ , $\boldsymbol{Q}_{j+1} = [\boldsymbol{Q}_j \; \boldsymbol{q}_{j+1}]$
        $\widetilde{\boldsymbol{w}}_j = \boldsymbol{G}_{j-1,j} \ldots \boldsymbol{G}_{1,2}\boldsymbol{w}_j$
        $\begin{bmatrix} \boldsymbol{u}_j \\ 0 \end{bmatrix} = \boldsymbol{G}_{j,j+1} \begin{bmatrix} \widetilde{\boldsymbol{w}}_j \\ h_{j+1,j} \end{bmatrix}$ ,     $\boldsymbol{U}_j = \begin{bmatrix} \boldsymbol{U}_{j-1} & \boldsymbol{u}_j \\ 0...0 & \end{bmatrix}$
        $\boldsymbol{f}_j = \boldsymbol{G}_{j,j+1}\boldsymbol{f}_{j-1}$
        if  $(\boldsymbol{f}_j)_{j+1} = 0$  then  ready = true else  $j = j + 1$
  end
    $\boldsymbol{x}_j = \boldsymbol{x}_0 + \boldsymbol{Q}_j \, \boldsymbol{U}_j^{-1} \, [\boldsymbol{I} \; 0] \, \boldsymbol{f}_j$
```

Remark. If $h_{j+1,j} = 0$ above, then $\boldsymbol{G}_{j,j+1} = \boldsymbol{I}$ and $(\boldsymbol{f}_j)_{j+1} = 0$.

A *preconditioned* GMRes is obtained from this by simply replacing the multiplication $\boldsymbol{A}\boldsymbol{q}_j$ by $\boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{q}_j$.

The work load of the Arnoldi iteration increases for each step: at step j in addition to the matrix–vector multiplication $\boldsymbol{A}\boldsymbol{q}_j$ (or $\boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{q}_j$) we have to compute j inner products $\boldsymbol{q}_k^*\boldsymbol{p}_j$ of length n and j vector updates $\boldsymbol{p}_j = \boldsymbol{p}_j - h_{k,j}\boldsymbol{q}_k$. The inner products and updates require approximately $4jn$ floating point operations, i.e., in m steps this work is $\approx 2m^2n$. The flops required for the QR decompositions are neglible compared to this. Since the work/iteration grows easily too big, the common way is to stop after a fixed number m of steps, form the vector $\boldsymbol{x}_m$ and restart the iteration with this initial approximation. This is called the GMRes(m) algorithm.

On the other hand, since GMRes minimizes the norm of $\boldsymbol{r}_j$, no other Krylov subspace iteration can be faster, when measured in the dimension of the subspace, than the (full) GMRes. The other methods that we study next try to perform as well as GMRes, but using much less arithmetic operations.

[1] The $\boldsymbol{r}_j$ vector of the Arnoldi iteration is here $\boldsymbol{p}_j$.

12.2. **Biconjugate gradient method.** The biorthogonal Lanczos is much less expensive than the Arnoldi iteration. If we start the process with vector $\boldsymbol{r}_0 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0$, it proceeds as (see section 9.2):

$$\begin{aligned}
\boldsymbol{v}_1 &= \boldsymbol{r}_0 / \left\|\boldsymbol{r}_0\right\|_2 \ , \ \boldsymbol{w}_1 = \boldsymbol{v}_1 \, , \beta_0 = \gamma_0 = 0 \\
\alpha_j &= \boldsymbol{w}_j^* \boldsymbol{A} \boldsymbol{v}_j \\
\boldsymbol{r}_j &= \boldsymbol{A}\boldsymbol{v}_j - \alpha_j \boldsymbol{v}_j - \beta_{j-1} \boldsymbol{v}_{j-1} \\
\boldsymbol{s}_j &= \boldsymbol{A}^* \boldsymbol{w}_j - \bar{\alpha}_j \boldsymbol{w}_j - \bar{\gamma}_{j-1} \boldsymbol{w}_{j-1} \\
\gamma_j &= \left|\boldsymbol{s}_j^* \boldsymbol{r}_j\right|^{1/2} \ , \ \beta_j = \boldsymbol{s}_j^* \boldsymbol{r}_j / \gamma_j \\
\boldsymbol{v}_{j+1} &= \boldsymbol{r}_j / \gamma_j \, , \ \boldsymbol{w}_{j+1} = \boldsymbol{s}_j / \bar{\beta}_j \ .
\end{aligned}$$

Writing $\boldsymbol{V}_j = \left[\boldsymbol{v}_1 \dots \boldsymbol{v}_j\right]$, $\boldsymbol{W}_j = \left[\boldsymbol{w}_1 \dots \boldsymbol{w}_j\right]$ and

$$\boldsymbol{T}_j = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \gamma_1 & \alpha_2 & \beta_2 & & \\ & \gamma_2 & \ddots & \ddots & \\ & & \ddots & \alpha_{j-1} & \beta_{j-1} \\ & & & \gamma_{j-1} & \alpha_j \end{bmatrix} , \qquad \widehat{\boldsymbol{T}}_j = \begin{bmatrix} \boldsymbol{T}_j \\ 0 \dots 0 \ \gamma_j \end{bmatrix} \in \mathbb{C}^{(j+1)\boldsymbol{x}j}$$

we get:

$$\boldsymbol{A}\boldsymbol{V}_j = \boldsymbol{V}_{j+1} \widehat{\boldsymbol{T}}_j \ , \qquad \boldsymbol{W}_j^* \boldsymbol{V}_j = \boldsymbol{I} \ .$$

Since the basis $\boldsymbol{v}_1, \dots, \boldsymbol{v}_j$ of the Krylov subspace $\mathcal{K}_j(\boldsymbol{A}, \boldsymbol{r}_0)$ is not orthogonal, the true minimization of the residual can not be done easily here. The biconjugate gradient method chooses the vector $\boldsymbol{x}_j \in \boldsymbol{x}_0 + \mathcal{K}_j(\boldsymbol{A}, \boldsymbol{r}_0)$ such that $\boldsymbol{W}_j^*(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_j) = 0$. Writing $\boldsymbol{x}_j = \boldsymbol{x}_0 + \boldsymbol{V}_j \boldsymbol{\xi}_j$ we get:

$$0 = \boldsymbol{W}_j^*(\boldsymbol{b} - \boldsymbol{A}(\boldsymbol{x}_0 + \boldsymbol{V}_j \boldsymbol{\xi}_j)) = \left\|\boldsymbol{r}_0\right\|_2 \boldsymbol{e}_1 - \boldsymbol{W}_j^* \boldsymbol{V}_{j+1} \widehat{\boldsymbol{T}}_j \boldsymbol{\xi}_j = \left\|\boldsymbol{r}_0\right\|_2 \boldsymbol{e}_1 - \boldsymbol{T}_j \boldsymbol{\xi}_j \ .$$

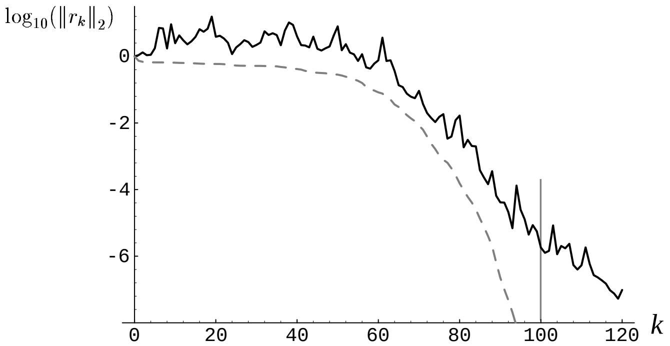
From this we solve $\boldsymbol{\xi}_j = \left\|\boldsymbol{r}_0\right\|_2 \boldsymbol{T}_j^{-1} \boldsymbol{e}_1$ and get $\boldsymbol{x}_j$. Similarly to the derivation of the conjugate gradient method from the Lanczos iteration (i.e., a long computation) also here we get a recursion directly for residuals $\boldsymbol{r}_j$ and the change directions $\boldsymbol{p}_j$. The result is:

Biconjugate gradient method (BiCG):

```
$\boldsymbol{r}_0 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0\,,\ \widetilde{\boldsymbol{r}}_0 = \boldsymbol{r}_0\,,\ \nu_{-1} = \widetilde{\boldsymbol{r}}_0^*\boldsymbol{r}_0\,,\ \boldsymbol{p}_{-1} = \widetilde{\boldsymbol{p}}_{-1} = 0$
  for  $j = 0, 1, 2, \ldots$
      $\nu_j = \widetilde{\boldsymbol{r}}_j^*\boldsymbol{r}_j\,,\ \mu_j = \nu_j/\nu_{j-1}$
      $\boldsymbol{p}_j = \boldsymbol{r}_j + \mu_j\boldsymbol{p}_{j-1}\,,\ \widetilde{\boldsymbol{p}}_j = \widetilde{\boldsymbol{r}}_j + \bar{\mu}_j\widetilde{\boldsymbol{p}}_{j-1}$
      $\rho_j = \nu_j/\widetilde{\boldsymbol{p}}_j^*\boldsymbol{A}\boldsymbol{p}_j$
      $\boldsymbol{x}_{j+1} = \boldsymbol{x}_j + \rho_j\boldsymbol{p}_j$
      $\boldsymbol{r}_{j+1} = \boldsymbol{r}_j - \rho_j\boldsymbol{A}\boldsymbol{p}_j\,,\ \widetilde{\boldsymbol{r}}_{j+1} = \widetilde{\boldsymbol{r}}_j - \bar{\rho}_j\boldsymbol{A}^*\widetilde{\boldsymbol{p}}_j$
  end
```

Above $\widetilde{\boldsymbol{r}}_0$ can be any vector satisfying $\widetilde{\boldsymbol{r}}_0^*\boldsymbol{r}_0 \neq 0$ (like the vector $\boldsymbol{w}_1$ in the biorthogonal Lanczos).

The biconjugate gradient method inherits the problems of the biorthogonal Lanczos. First the norms of the residuals behave usually very irregularly. Below is a typical case:



Here $\boldsymbol{A}$ is a random 100*x*100 matrix. The dashed line depicts the GMRes iteration. From the picture we also see that BiCG does not find the solution in $n = 100$ steps although theoretically it should. This is due to the loss of biorthogonality caused by round–off errors.

Second, there may occur divisions by small numbers (zeros) that lead to instabilities. This can happen in two places:

$$\text{either}\quad \widetilde{\boldsymbol{p}}_j^*\boldsymbol{A}\boldsymbol{p}_j = 0 \qquad \text{or} \qquad \widetilde{\boldsymbol{r}}_j^*\boldsymbol{r}_j = 0 \qquad \text{with} \qquad \boldsymbol{r}_j \neq 0\,.$$

Problem 12.2. Show that $\widetilde{\boldsymbol{p}}_j^*\boldsymbol{A}\boldsymbol{p}_j = 0$ means that $\boldsymbol{T}_j$ is not invertible.

12.3. **Quasiminimal residual method (QMR).** The first possible division by zero in the biconjugate gradient method $\widetilde{\boldsymbol{p}}_j^*\boldsymbol{A}\boldsymbol{p}_j = 0$ ($\boldsymbol{T}_j$ is not invertible) can be fixed as follows. Instead of the requirement $\boldsymbol{W}_j^*(\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}_j)=0$ notice that:

$$\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}_j=\boldsymbol{V}_{j+1}(\|\boldsymbol{r}_0\|_2\,\boldsymbol{e}_1-\widehat{\boldsymbol{T}}_j\boldsymbol{\xi}_j)$$

and choose $\boldsymbol{\xi}_j$ to be such that it minimizes the norm of $\|\boldsymbol{r}_0\|_2\,\boldsymbol{e}_1-\widehat{\boldsymbol{T}}_j\boldsymbol{\xi}_j$. This *quasiminimization* can be done as in the GMRes method: using the QR decomposition, now even more simply, since $\widehat{\boldsymbol{T}}_j$ is a tridiagonal matrix, so that in its $\boldsymbol{QR}$ decomposition $\widehat{\boldsymbol{T}}_j=\boldsymbol{S}_j\boldsymbol{U}_j$ the upper triangular matrix $\boldsymbol{U}_j$ has only three diagonals different from zero (why?). Direct application of this leads to the following preliminary version of the quasiminimal residual method

QMR version 0[2]:

```
$\boldsymbol{r}_0=\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}_0,\ \boldsymbol{v}_1=\boldsymbol{w}_1=\boldsymbol{r}_0/\|\boldsymbol{r}_0\|,\ \boldsymbol{G}_{0,1}=\boldsymbol{I}$
$\alpha_1=\boldsymbol{w}_1^*\boldsymbol{A}\boldsymbol{v}_1$
$\boldsymbol{r}_1=\boldsymbol{A}\boldsymbol{v}_1-\alpha_1\boldsymbol{v}_1,\ \boldsymbol{s}_1=\boldsymbol{A}^*\boldsymbol{w}_1-\bar\alpha_1\boldsymbol{w}_1$
$\gamma_1=\sqrt{|\boldsymbol{s}_1^*\boldsymbol{r}_1|},\ \beta_1=\boldsymbol{s}_1^*\boldsymbol{r}_1/\gamma_1$
$\boldsymbol{v}_2=\boldsymbol{r}_1/\gamma_1,\ \boldsymbol{w}_2=\boldsymbol{s}_1/\bar\beta_1$
$\begin{bmatrix}\boldsymbol{U}_1\\0\end{bmatrix}=\boldsymbol{G}_{1,2}\begin{bmatrix}\alpha_1\\\gamma_1\end{bmatrix},\quad \boldsymbol{f}_1=\boldsymbol{G}_{1,2}\begin{bmatrix}\|\boldsymbol{r}_0\|\\0\end{bmatrix}$
$\boldsymbol{x}_1=\boldsymbol{x}_0+\boldsymbol{v}_1\,\boldsymbol{U}_1^{-1}(\boldsymbol{f}_1)_1$
for $j=2,3,\ldots$
    $\alpha_j=\boldsymbol{w}_j^*\boldsymbol{A}\boldsymbol{v}_j$
    $\boldsymbol{r}_j=\boldsymbol{A}\boldsymbol{v}_j-\alpha_j\boldsymbol{v}_j-\beta_{j-1}\boldsymbol{v}_{j-1},\ \boldsymbol{s}_j=\boldsymbol{A}^*\boldsymbol{w}_j-\bar\alpha_j\boldsymbol{w}_j-\bar\gamma_{j-1}\boldsymbol{w}_{j-1}$
    $\boldsymbol{w}_j=\boldsymbol{G}_{j-1,j}\boldsymbol{G}_{j-2,j-1}\begin{bmatrix}0\\\beta_{j-1}\\\alpha_j\end{bmatrix}$
    $\gamma_j=\sqrt{|\boldsymbol{s}_j^*\boldsymbol{r}_j|},\ \beta_j=\boldsymbol{s}_j^*\boldsymbol{r}_j/\gamma_j$
    $\begin{bmatrix}\boldsymbol{u}_j\\0\end{bmatrix}=\boldsymbol{G}_{j,j+1}\begin{bmatrix}\boldsymbol{w}_j\\\gamma_j\end{bmatrix},\quad \boldsymbol{U}_j=\begin{bmatrix}\boldsymbol{U}_{j-1}&\boldsymbol{u}_j\\0\ldots0&\end{bmatrix},\quad \boldsymbol{f}_j=\boldsymbol{G}_{j,j+1}\begin{bmatrix}\boldsymbol{f}_{j-1}\\0\end{bmatrix}$
    $\boldsymbol{x}_j=\boldsymbol{x}_0+\begin{bmatrix}\boldsymbol{v}_1\ldots\boldsymbol{v}_j\end{bmatrix}\,\boldsymbol{U}_j^{-1}\,[\boldsymbol{I}\ 0]\,\boldsymbol{f}_j$
    $\boldsymbol{v}_{j+1}=\boldsymbol{r}_j/\gamma_j,\ \boldsymbol{w}_{j+1}=\boldsymbol{s}_j/\bar\beta_j$
end
```

In this version we need to store all the vectors $\boldsymbol{v}_1,\ldots,\boldsymbol{v}_j$, so that $\boldsymbol{x}_j$ could be computed.

[2] Remark: here $\boldsymbol{r}_j$ is as in the Lanczos algorithm and generally $\boldsymbol{r}_j\neq\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}_j$.

Problem 12.3. Denote: $\big[\boldsymbol{p}_1 \ldots \boldsymbol{p}_j\big] = \big[\boldsymbol{v}_1 \ldots \boldsymbol{v}_j\big]\ \boldsymbol{U}_j^{-1}$. Derive a recursion for the $\boldsymbol{p}_j$ vectors. Show that from these we get: $\boldsymbol{x}_j = \boldsymbol{x}_{j-1} + (\boldsymbol{f}_j)_j\, \boldsymbol{p}_j$.

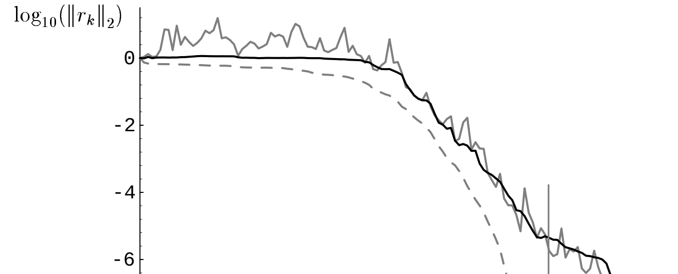
Using this we get the following:

Myopic **QMR:**

```
$\boldsymbol{r}_0 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0$, $\boldsymbol{v}_1 = \boldsymbol{w}_1 = \boldsymbol{r}_0/\|\boldsymbol{r}_0\|$, $\boldsymbol{p}_{-1} = \boldsymbol{p}_0 = \boldsymbol{v}_0 = \boldsymbol{w}_0 = 0$
$\beta_0 = \gamma_0 = 0$, $\boldsymbol{G}_{-1,0} = \boldsymbol{G}_{0,1} = \boldsymbol{I}$, $\boldsymbol{f}_0 = \|\boldsymbol{r}_0\|$
for $j = 1,2,\ldots$
    $\alpha_j = \boldsymbol{w}_j^*\boldsymbol{A}\boldsymbol{v}_j$, $\boldsymbol{r}_j = \boldsymbol{A}\boldsymbol{v}_j - \alpha_j\boldsymbol{v}_j - \beta_{j-1}\boldsymbol{v}_{j-1}$, $\boldsymbol{s}_j = \boldsymbol{A}^*\boldsymbol{w}_j - \bar{\alpha}_j\boldsymbol{w}_j - \bar{\gamma}_{j-1}\boldsymbol{w}_{j-1}$
    $\boldsymbol{w}_j = \boldsymbol{G}_{j-1,j}\boldsymbol{G}_{j-2,j-1}\begin{bmatrix}0\\ \beta_{j-1}\\ \alpha_j\end{bmatrix}$
    $\begin{bmatrix}\boldsymbol{u}_j\\ 0\end{bmatrix} = \boldsymbol{G}_{j,j+1}\begin{bmatrix}\boldsymbol{w}_j\\ \gamma_j\end{bmatrix}$, $\boldsymbol{f}_j = \boldsymbol{G}_{j,j+1}\begin{bmatrix}\boldsymbol{f}_{j-1}\\ 0\end{bmatrix}$
    $\boldsymbol{p}_j = [\boldsymbol{v}_j - (\boldsymbol{u}_j)_{j-1}\,\boldsymbol{p}_{j-1} - (\boldsymbol{u}_j)_{j-1}\,\boldsymbol{p}_{j-2}]/(\boldsymbol{u}_j)_j$, $\boldsymbol{x}_j = \boldsymbol{x}_{j-1} + (\boldsymbol{f}_j)_j\,\boldsymbol{p}_j$
    $\gamma_j = \sqrt{|\boldsymbol{s}_j^*\boldsymbol{r}_j|}$, $\beta_j = \boldsymbol{s}_j^*\boldsymbol{r}_j/\gamma_j$, $\boldsymbol{v}_{j+1} = \boldsymbol{r}_j/\gamma_j$, $\boldsymbol{w}_{j+1} = \boldsymbol{s}_j/\bar{\beta}_j$
end
```

Here the memory requirement is more reasonable. In step j we need to access only seven vectors from the previous steps $\boldsymbol{v}_{j-1}$, $\boldsymbol{v}_j$, $\boldsymbol{w}_{j-1}$, $\boldsymbol{w}_j$, $\boldsymbol{p}_{j-1}$, $\boldsymbol{p}_j$ and $\boldsymbol{x}_{j-1}$. From vector $\boldsymbol{f}_j$ it suffices to remember only the last component.

This quasiminimization strategy of the residual also fixes the nervous behavior of the norm of the residual. In the following picture we have added (to the previous picture) the results (in black) of this QMR version:



Notice that the slowing down of the convergence due to the loss of orthogonality is a problem also in this algorithm.

The other division–by–zero problem of the biconjugate gradient method:

$$\widetilde{\boldsymbol{r}}_j^*\boldsymbol{r}_j = 0 \qquad \text{with} \qquad \boldsymbol{r}_j \neq 0$$

is the basic problem of the biorthogonal Lanczos and it can be fixed only if we base the method on the look ahead Lanczos. This leads to the genuine QMR method, but we won't present it here.

12.4. **Biconjugate gradient squared.** Return to the biconjugate gradient method. Let us figure out for what the "dual vectors" $\widetilde{\boldsymbol{r}}_j$ and $\widetilde{\boldsymbol{p}}_j$ are needed: only to compute the numbers

$$\nu_j = \widetilde{\boldsymbol{r}}_j^*\boldsymbol{r}_j \qquad \text{and} \qquad \rho_j = \nu_j/\widetilde{\boldsymbol{p}}_j^*\boldsymbol{A}\boldsymbol{p}_j \ ,$$

i.e., in two inner products. It is easy to see, that the vectors of this method satisfy

$$\begin{aligned} \boldsymbol{p}_j &= \psi_j(\boldsymbol{A})\boldsymbol{r}_0 \ , \qquad & \widetilde{\boldsymbol{p}}_j &= \psi_j(\boldsymbol{A})^*\widetilde{\boldsymbol{r}}_0 \ , \\ \boldsymbol{r}_j &= \phi_j(\boldsymbol{A})\boldsymbol{r}_0 \ , \qquad & \widetilde{\boldsymbol{r}}_j &= \phi_j(\boldsymbol{A})^*\widetilde{\boldsymbol{r}}_0 \ , \end{aligned}$$

where ψ_j and ϕ_j are polynomials of degree j. In particular, we get $\boldsymbol{p}_j$ and $\widetilde{\boldsymbol{p}}_j$ (respectively $\boldsymbol{r}_j$ and $\widetilde{\boldsymbol{r}}_j$) using *the same* polynomial. Let us define an (indefinite) inner product for polynomials

$$\langle \phi, \psi \rangle = \widetilde{\boldsymbol{r}}_0^*\phi(\boldsymbol{A})^*\psi(\boldsymbol{A})\boldsymbol{r}_0$$

and the polynomials $\chi(t) = t$, $\mathbf{1}(t) = 1$. Then the biconjugate gradient method for the polynomials can be written as

Polynomial biconjugate gradient method:

```
$\phi_0 = \mathbf{1}$ , $\nu_{-1} = \langle \bar{\phi}_0, \phi_0 \rangle$ , $\psi_{-1} = 0$
for  $j = 0, 1, 2, \ldots$
    $\nu_j = \langle \bar{\phi}_j, \phi_j \rangle$ , $\mu_j = \nu_j/\nu_{j-1}$
    $\psi_j = \phi_j + \mu_j \ \psi_{j-1}$
    $\rho_j = \nu_j/ \langle \bar{\psi}_j, \chi\psi_j \rangle$
    $\phi_{j+1} = \phi_j - \rho_j \ \chi \ \psi_j$
end
```

The biconjugate gradient squared method (BiCGS) starts by noticing (notice) that the required inner products can be computed also as follows:

$$\nu_j = \langle \bar{\phi}_j, \phi_j \rangle = \langle \mathbf{1}, \phi_j^2 \rangle \qquad \text{and} \qquad \langle \bar{\psi}_j, \chi\psi_j \rangle = \langle \mathbf{1}, \chi\psi_j^2 \rangle \ .$$

Let us form the iteration for the polynomials

$$\eta_j = \phi_j^2 \ , \quad \zeta_j = \psi_j^2 \ , \quad \xi_j = \phi_j\psi_{j-1} \ , \quad \gamma_j = \phi_j\psi_j \ .$$

Problem 12.4 (Polynomial BiCGS)**.** Show that for these we get the iteration:

```
$\eta_0 = \mathbf{1}\,,\ \nu_{-1} = \langle \mathbf{1}, \eta_0 \rangle\ ,\ \xi_0 = \zeta_{-1} = 0$
for  $j = 0, 1, 2, \dots$
    $\nu_j = \langle \mathbf{1}, \eta_j \rangle\ ,\ \mu_j = \nu_j / \nu_{j-1}$
    $\gamma_j = \eta_j + \mu_j\, \xi_j\,, \quad \zeta_j = \gamma_j + \mu_j\, \xi_j + \mu_j^2\, \zeta_{j-1}$
    $\rho_j = \nu_j / \langle \mathbf{1}, \chi\, \zeta_j \rangle$
    $\xi_{j+1} = \gamma_j - \rho_j\, \chi\, \zeta_j\,, \quad \eta_{j+1} = \eta_j - \rho_j\, \chi\, (\gamma_j + \xi_{j+1})$
end
```

Returning back to the iteration of vectors we get an interesting method , where the transpose of $\boldsymbol{A}$ is not needed and which satisfies: if $\boldsymbol{r}_j^{\mathrm{BiCG}} = \phi_j(\boldsymbol{A})\boldsymbol{r}_0$, then $\boldsymbol{r}_j^{\mathrm{BiCGS}} = \phi_j(\boldsymbol{A})^2\,\boldsymbol{r}_0$. In particular, if BiCG converges nicely, then $\phi_j(\boldsymbol{A})$ has small norm and BiCGS converges almost twice that fast. Denote:

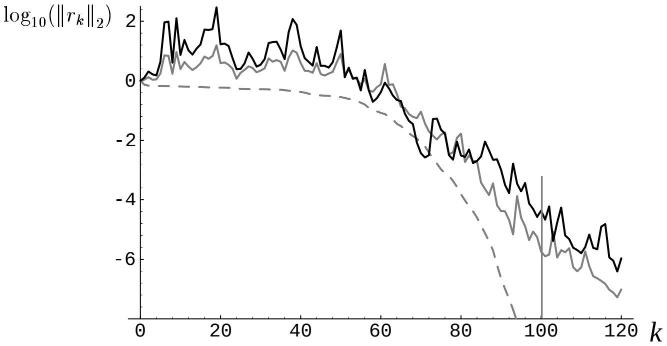
$$\begin{aligned} \boldsymbol{r}_j &= \phi_j(\boldsymbol{A})^2 \boldsymbol{r}_0 = \eta_j(\boldsymbol{A})\boldsymbol{r}_0\ , & \boldsymbol{p}_j &= \psi_j(\boldsymbol{A})^2 \boldsymbol{r}_0 = \zeta_j(\boldsymbol{A})\boldsymbol{r}_0\ , \\ \boldsymbol{q}_j &= \phi_j(\boldsymbol{A})\psi_{j-1}(\boldsymbol{A})\boldsymbol{r}_0 = \boldsymbol{\xi}_j(\boldsymbol{A})\boldsymbol{r}_0\ , & \boldsymbol{u}_j &= \phi_j(\boldsymbol{A})\psi_j(\boldsymbol{A})\boldsymbol{r}_0 = \gamma_j(\boldsymbol{A})\boldsymbol{r}_0\ . \end{aligned}$$

Problem 12.5. Show that then we get the following

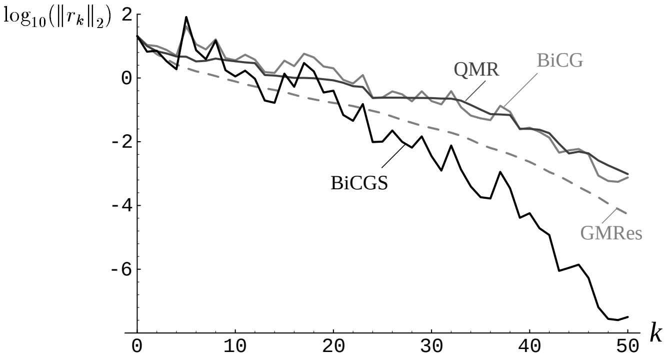
Biconjugate gradient squared method (BiCGS):

```
$\boldsymbol{r}_0 = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0\,,\ \widetilde{\boldsymbol{r}}_0 = \boldsymbol{r}_0\,,\ \nu_{-1} = \boldsymbol{r}_0^*\boldsymbol{r}_0\,,\ \boldsymbol{p}_{-1} = \boldsymbol{q}_0 = 0$
for  $j = 0, 1, 2, \dots$
    $\nu_j = \widetilde{\boldsymbol{r}}_0^*\boldsymbol{r}_j\,,\ \mu_j = \nu_j / \nu_{j-1}$
    $\boldsymbol{u}_j = \boldsymbol{r}_j + \mu_j\, \boldsymbol{q}_j\,, \quad \boldsymbol{p}_j = \boldsymbol{u}_j + \mu_j\, \boldsymbol{q}_j + \mu_j^2\, \boldsymbol{p}_{j-1}$
    $\boldsymbol{v}_j = \boldsymbol{A}\, \boldsymbol{p}_j\,, \quad \rho_j = \nu_j / \widetilde{\boldsymbol{r}}_0^*\boldsymbol{v}_j$
    $\boldsymbol{q}_{j+1} = \boldsymbol{u}_j - \rho_j\, \boldsymbol{v}_j\,, \quad \boldsymbol{r}_{j+1} = \boldsymbol{r}_j - \rho_j\, \boldsymbol{A}\, (\boldsymbol{u}_j + \boldsymbol{q}_{j+1})$
    $\boldsymbol{x}_{j+1} = \boldsymbol{x}_j + \rho_j(\boldsymbol{u}_j + \boldsymbol{q}_{j+1})$
end
```

In comparison with the BiCG method here the multiplication with $\boldsymbol{A}^*$ has been replaced by another multiplication with $\boldsymbol{A}$. The example of the previous pictures looks now as follows. Here the dashed line corresponds to GMRes, the grey to BiCG, and the black to BiCGS.



Let us take another 100x100 problem, where BiCG works fine. Then BiCGS is even better:



12.5. **On convergence.** This was supposed to fill an entire chapter, but let us just state the following

Theorem 12.1. *Let* $\boldsymbol{A}$ *be diagonalizable :* $\boldsymbol{A} = \boldsymbol{X}\Lambda\boldsymbol{X}^{-1}$. *Then the GMRes residuals satisfy:*

$$\frac{\|\boldsymbol{r}_j\|_2}{\|\boldsymbol{r}_0\|_2} \leq \kappa_2(\boldsymbol{X}) \min_{\substack{p\in\boldsymbol{P}_{j+1}\\ p(0)=1}} \max_{\lambda\in\Lambda(\boldsymbol{A})} |p(\lambda)| \ .$$

Problem 12.6. *Proof. ...* □

And mention that:
Factum: If $\boldsymbol{A} = \widetilde{\boldsymbol{A}} + \boldsymbol{E}$, where the matrix $\boldsymbol{E}$ has rank k , then the corresponding GMRes iterates satisfy:

$$\|\boldsymbol{r}_{j+k}\| \leq \|\widetilde{\boldsymbol{r}}_j\| \ .$$

This is good to keep in mind when looking for preconditioners: if a small change in rank gives a system that is easily solvable then this is sure to give a good preconditioner.

13. Preconditioning

In the previous chapter we saw that preconditioning may have a drastic effect on the convergence speed of an iterative solver. In this chapter we look at general strategies and ideas to build preconditioners.

General preconditioning. Assume we are solving the system of equations

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{13.1}$$

and let $\boldsymbol{K}$ and $\boldsymbol{M}$ be matrices somehow close to $\boldsymbol{A}$ and such that the systems $\boldsymbol{K}\boldsymbol{y} = \boldsymbol{c}$ and $\boldsymbol{M}\boldsymbol{z} = \boldsymbol{d}$ are easy to solve. Replace (13.1) by the equation

$$\boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{K}^{-1}\,\boldsymbol{y} = \boldsymbol{M}^{-1}\boldsymbol{b}\ . \tag{13.2}$$

Then $\boldsymbol{M}$ is called a left and $\boldsymbol{K}$ a right preconditioner. Form the iteration for the partition $\boldsymbol{A}\boldsymbol{K}^{-1} = \boldsymbol{M} - \boldsymbol{N}$:

$$\boldsymbol{M}\boldsymbol{y}_{k+1} = \boldsymbol{N}\boldsymbol{y}_k + \boldsymbol{b} = (\boldsymbol{M} - \boldsymbol{A}\boldsymbol{K}^{-1})\boldsymbol{y}_k + \boldsymbol{b}\ ,$$

i.e.,

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k + \boldsymbol{M}^{-1}(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{K}^{-1}\boldsymbol{y}_k)\ .$$

This converges, if the spectral radius of the matrix $\boldsymbol{I} - \boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{K}^{-1}$ is less than one. Notice, that the matrices $\boldsymbol{K}$ and $\boldsymbol{M}$ are not needed only some routines that perform the operations $\boldsymbol{v} \to \boldsymbol{K}^{-1}\boldsymbol{v}$ and $\boldsymbol{v} \to \boldsymbol{M}^{-1}\boldsymbol{v}$. The iteration can then be accelerated with the Tshebyshev iteration, in the Hermitian case with the preconditioned conjugate gradient method, or in the general case with methods like GMRes, QMR, CGS, etc.

Generally, the analysis of these iterations show that they are fast in cases where the iteration matrix is of the form

$$\boldsymbol{I} - \boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{K}^{-1} = \boldsymbol{R} + \boldsymbol{F}\ ,$$

where the norm of $\boldsymbol{R}$ is small and the rank of $\boldsymbol{F}$ is small.

In the sequel we restrict ourselves to consider only left preconditioning, i.e., we assume, that $\boldsymbol{K} = \boldsymbol{I}$.

The general preconditioning strategies can be devided in the following types:

1) Classical preconditioners: $\boldsymbol{M}$ is chosen as in the Jacobi, Gauss Seidel, SOR, or SSOR iterations.
2) ILU preconditioners (incomplete LU factorization): here $\boldsymbol{M} = \boldsymbol{L}\boldsymbol{U}$, where $\boldsymbol{L}$ and $\boldsymbol{U}$ are (sparse) lower and upper triangular matrices such that $\boldsymbol{L}\boldsymbol{U} \approx \boldsymbol{A}$.

[0] Version: April 9, 2003

3) Polynomial preconditioners: here a direct approximation for the inverese of $\boldsymbol{A}$ is looked for in the polynomial form: $\boldsymbol{H} = \boldsymbol{M}^{-1} = p(\boldsymbol{A}) \approx \boldsymbol{A}^{-1}$.
4) Direct sparse approximation for the inverse: look for $\boldsymbol{H} = \boldsymbol{M}^{-1}$ from some nice class such that for example $\|\boldsymbol{I} - \boldsymbol{H}\boldsymbol{A}\|_F$ is as small as possible.
5) The multigrid idea: an approximate inverse $\boldsymbol{H}$ is constructed from a solver of a smaller, but similar problem.
6) Other approaches (not considered here) we get, when we look for matrices somehow close to $\boldsymbol{A}$ and that have certain structure so that the corresponding systems can be easily solved with e.g. circulant or Toeplitz solvers or using fast Fourier transform.

If the preconditioner is directly a matrix $\boldsymbol{H} = \boldsymbol{M}^{-1}$ we talk about explicit preconditioning, otherwise implicit (i.e., when we solve systems $\boldsymbol{M}\boldsymbol{z} = \boldsymbol{d}$).

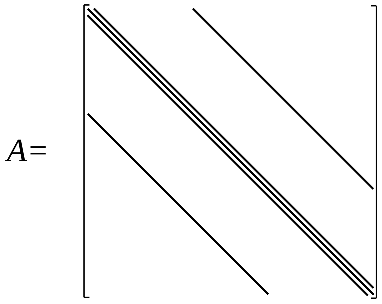
13.1. **ILU preconditioners.** Incomplete $\boldsymbol{LU}$ factorizations. Consider first the $\boldsymbol{LU}$ factorization of a *band matrix* $\boldsymbol{A}$ having the width u of upper band and l of the lower band, in other words, $a_{i,j} = 0$, for $j > i + u$ and $i > j + l$. The $\boldsymbol{LU}$ factorization of such a matrix is also banded (check):

$$\boldsymbol{A} = \boldsymbol{LU} =$$

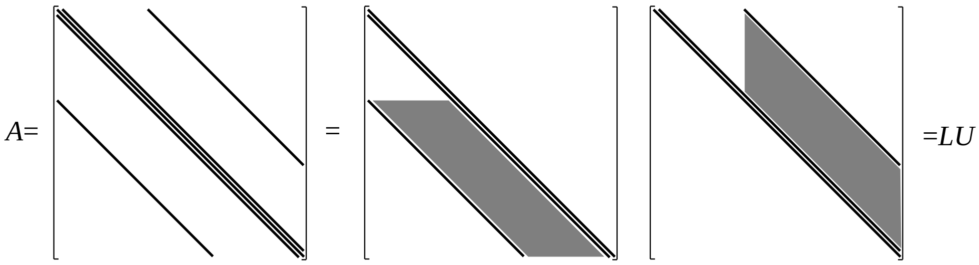
$$
\begin{array}{c}
 \\ l\left\{\begin{array}{c} \\ \\ \\ \end{array}\right. \\ \\ \\ \\ \\ \\
\end{array}
\begin{bmatrix}
\# & \overbrace{\# & \cdots & \#}^{u} & & & & & \\
\# & \# & \# & \cdots & \# & & & & \\
\vdots & \# & \# & \cdot & \cdot & \ddots & & & \\
\# & \vdots & \cdot & \cdot & \cdot & \cdot & \ddots & & \\
 & \# & \cdot & \cdot & \cdot & \cdot & \cdot & \# & \\
 & & \ddots & \cdot & \cdot & \cdot & \cdot & \vdots & \# \\
 & & & \ddots & \cdot & \cdot & \# & \# & \vdots \\
 & & & & \# & \cdots & \# & \# & \# \\
 & & & & & \# & \cdots & \# & \#
\end{bmatrix}
=
\begin{array}{c}
l\left\{\begin{array}{c} \\ \\ \\ \end{array}\right. \\ \\ \\ \\ \\ \\
\end{array}
\begin{bmatrix}
1 & & & & & & & & \\
\# & 1 & & & & & & & \\
\vdots & \# & 1 & & & & & & \\
\# & \vdots & \cdot & \cdot & & & & & \\
 & \# & \cdot & \cdot & \cdot & & & & \\
 & & \ddots & \cdot & \cdot & \cdot & & & \\
 & & & \ddots & \cdot & \cdot & 1 & & \\
 & & & & \# & \cdots & \# & 1 & \\
 & & & & & \# & \cdots & \# & 1
\end{bmatrix}
\begin{bmatrix}
\# & \overbrace{\# & \cdots & \#}^{u} & & & & & \\
 & \# & \# & \cdots & \# & & & & \\
 & & \# & \cdot & \cdot & \ddots & & & \\
 & & & \cdot & \cdot & \cdot & \ddots & & \\
 & & & & \cdot & \cdot & \cdot & \# & \\
 & & & & & \cdot & \cdot & \vdots & \# \\
 & & & & & & \# & \# & \vdots \\
 & & & & & & & \# & \# \\
 & & & & & & & & \#
\end{bmatrix}
$$

The work needed to form this factorization is $\approx 2nlu$. It is a branch of art, how to get small $l + u$ by permutations. This we however won't study here.

In many sparse cases $\boldsymbol{A}$ has nonzero elements only on some diagonals:



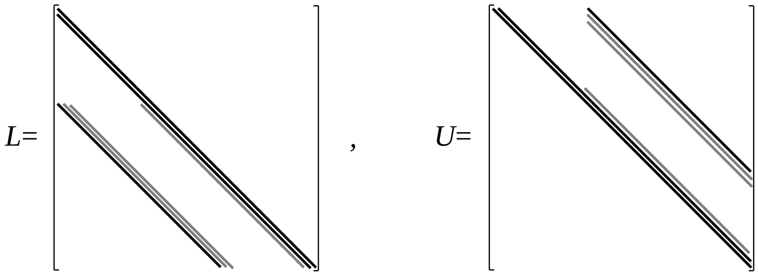
then, unfortunately, the LU factorization of $\boldsymbol{A}$ is not that sparse but all the grey areas of the following figure generally become nonzero



The idea of the incomplete $\boldsymbol{LU}$ factorization (ILU) is to fix some sparsity structure for $\boldsymbol{L}$ and $\boldsymbol{U}$ and require, that the equation $\boldsymbol{A} = \boldsymbol{LU}$ holds for the corresponding elements. Usually the structure is chosen to be the same as that of $\boldsymbol{A}$ or somewhat more dense. So, we choose a set $J \subset \{1, \ldots, n\}^2$ of index pairs such that $\{(i,j) \mid a_{i,j} \neq 0\} \subset J$. Denote $\mathcal{M}_J = \{\boldsymbol{M} \in \mathbb{C}^{n \times n} \mid m_{i,j} = 0\,,\ (i,j) \notin J\}$ and search for $\boldsymbol{L}, \boldsymbol{U} \in \mathcal{M}_J$ (lower and upper triangular matrix, respectively) such that

$$(\boldsymbol{LU})_{i,j} = a_{i,j} \quad \text{for all} \quad (i,j) \in J \ .$$

A typical choice in the case of the figure above is



The computation of the incomplete $\boldsymbol{LU}$ factorization is done in principle similarly to that of the full factorization but considering only the index set J. The existence of the factorization (pivot elements are nonzero) is generally difficult to guarantee. Also pivoting techniques become complicated. Some theorems exist for the so called M–matrices[1]

Often e.g. due to the lack of pivoting the computation of the ILU factorization can become very unstable. As a remedy, some modifications have been designed (MILU = modified ILU) for example such that we add a scalars on the diagonal of U such that the row sums of $\boldsymbol{LU}$ and $\boldsymbol{A}$ become the same. Another way to try to stabilize the computation is to perform the ILU for the matrix $\boldsymbol{A}+\boldsymbol{D}$, where $\boldsymbol{D}$ is a positive diagonal matrix.

[1] $\boldsymbol{A}$ is an M–matrix, if $a_{i,j} \leq 0$ for all $i \neq j$ and $(\boldsymbol{A}^{-1})_{i,j} \geq 0$ for all i, j.

When $\boldsymbol{A}$ is Hermitian and positive definite, then the ILU is searched in the form $\boldsymbol{U} = \boldsymbol{L}^*$. Then $\boldsymbol{L}\boldsymbol{L}^*$ is the incomplete *Cholesky factorization* of $\boldsymbol{A}$.

A drawback of the ILU preconditioning is that it is not very suitable for parallel processing.

Problem 13.1. Test the preconditioned conjugate gradient method in the Problem 10.9 using the incomplete Cholesky factorization as a preconditioner.

13.2. **Polynomial preconditioning.** In cases, where $\boldsymbol{A}$ (or already slightly preconditioned $\boldsymbol{A}$) is cheap to apply in a parallel architecture it often pays to try polynomial preconditioning.

Let be $\boldsymbol{A} = \boldsymbol{M}_0 - \boldsymbol{N}_0$ some converging decomposition, in other words,

$$\rho(\boldsymbol{E}_0) = \rho(\boldsymbol{M}_0^{-1}\boldsymbol{N}_0) < 1 \ .$$

Then

$$\sum_{j=0}^{\infty} \boldsymbol{E}_0^j = \boldsymbol{I} + \boldsymbol{E}_0 + \boldsymbol{E}_0^2 + \cdots = (\boldsymbol{I} - \boldsymbol{E}_0)^{-1} \ .$$

This series converges, since in a suitable norm $\|\boldsymbol{E}_0\|_* < 1$, and $\left\|\boldsymbol{E}_0^j\right\|_* \le \|\boldsymbol{E}_0\|_*^j$. We get:

$$\boldsymbol{A}^{-1} = [\boldsymbol{M}_0(\boldsymbol{I} - \boldsymbol{M}_0^{-1}\boldsymbol{N}_0)]^{-1} = (\boldsymbol{I} - \boldsymbol{E}_0)^{-1}\boldsymbol{M}_0^{-1} \ .$$

A simple approximation for $\boldsymbol{A}^{-1}$ we get from the trucated series:

$$\boldsymbol{H}_m = \sum_{j=0}^{m} \boldsymbol{E}_0^j \boldsymbol{M}_0^{-1} \ .$$

This proposes the preconditioner

$$\boldsymbol{H}_m = \sum_{j=0}^{m} (\boldsymbol{I} - \boldsymbol{M}_0^{-1}\boldsymbol{A})^j \boldsymbol{M}_0^{-1} = p_m(\boldsymbol{M}_0^{-1}\boldsymbol{A})\boldsymbol{M}_0^{-1}$$

where p_m is a polynomial of degree m.

Problem 13.2. Show, that above $p_m(x) = \sum_{k=0}^{m} \binom{m+1}{k+1} (-x)^k$. Hint: show first that $\sum_{j=0}^{m} \binom{j}{k} = \binom{m+1}{k+1}$.

Another way to construct polynomial preconditioners is to start with the ideal goal $p_m(\boldsymbol{A})\boldsymbol{A} = I$. Then $p_m(\lambda)\lambda = 1$ for every eigenvalue λ of $\boldsymbol{A}$. Now, if $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_m$ are some approximate eigenvalues of $\boldsymbol{A}$ then it sounds reasonable to choose p_m to be the polynomial that satisfies $p_m(\widehat{\lambda}_j)\widehat{\lambda}_j = 1$, $j = 1, \ldots, m$.

For example, if $\boldsymbol{A}$ is Hermitian and positive definite and we start with the conjugate gradient method, then the corresponding Lanczos tridiagonal matrix $\boldsymbol{T}_j$ is also found with little extra work. From this we can get approximations for the eigenvalues and from those a polynomial preconditioner, so that we can continue using a preconditioned method.

13.3. **Direct sparse approximations of the inverse.** Explicit well parallelizable preconditioners can also be looked for in the following way: choose some sparsity structure J and look for the matrix $\boldsymbol{H} \in \mathcal{M}_J$ (see subsection 13.1) for which

$$\|\boldsymbol{I} - \boldsymbol{H}\boldsymbol{A}\|_F$$

is the smallest possible. Since

$$\|\boldsymbol{I} - \boldsymbol{H}\boldsymbol{A}\|_F^2 = \|\boldsymbol{I} - \boldsymbol{A}^*\boldsymbol{H}^*\|_F^2 = \sum_{i,j=1}^{n} |(\boldsymbol{I} - \boldsymbol{A}^*\boldsymbol{H}^*)_{i,j}|^2 = \sum_{j=1}^{n} \left\|\boldsymbol{A}^*\boldsymbol{h}_j^* - \boldsymbol{e}_j\right\|^2 ,$$

where $\boldsymbol{h}_j$ is the j^{th} row of $\boldsymbol{H}$, we see that the minimization reduces to n sparate problems, each of which has its own sparsity requirement for $\boldsymbol{h}_j$. In the minimization we have the corresponding rows of $\boldsymbol{A}$ to which we perform for example the $\boldsymbol{QR}$ factorization. This seems to be expensive workwise, but it parallelizes completely.

13.4. **The multigrid idea.** This is to construct an approximation for the inverse from a solver $\widehat{\boldsymbol{H}}$ of a smaller but similar problem.

Here we consider only the cases of Example 10.2 and Problem 10.5, i.e., the Poisson equation in one and two space dimensions.

The simplest version of the one dimensional equation is

$$-u_{xx}(x) = f(x)\,, \quad x \in (0,1)\,,\ u(0) = u(1) = 0\,.$$

For this we get the discretization

$$-u_{xx}(ih) = \tfrac{1}{h^2}[-u(ih-h) + 2u(ih) - u(ih+h)]$$

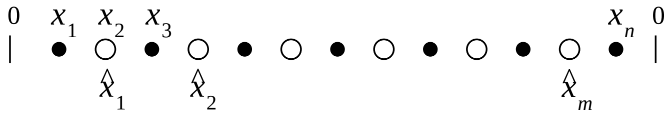
where $h = 1/(n+1)$. The matrix of the corresponding system $\boldsymbol{A}\boldsymbol{u} = b$, where

$$u = (u(h), u(2h), \ldots, u(nh))\,,\ b = (f(h), f(2h), \ldots, f(nh))\,,$$

becomes

$$\boldsymbol{A} = \tfrac{1}{(n+1)^2} \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{n\times n}\ .$$

Let $n = 2m+1$ be odd and assume, that we can easily solve the corresponding system $\widehat{\boldsymbol{A}}\widehat{\boldsymbol{u}} = \widehat{\boldsymbol{b}}$ (approximately) by $\widehat{\boldsymbol{u}} = \widehat{\boldsymbol{H}}\widehat{\boldsymbol{b}}$, where the coefficient matrix is similar but smaller $\widehat{\boldsymbol{A}} \in \mathbb{R}^{m\times m}$. The idea is now to use this solver as a preconditioner for $\boldsymbol{A}$. Then we have to transform the fine grid problems of size $n = 2m+1$, to problems on the coarser grid, i.e., of size m and the solutions of the latter back to n–vectors. In other words the fine grid functions have to be ”projected” on the coarse grid and functions defined on the latter ”extended” on the fine grid. The latter is done by interpolation. The grids are depicted in the following:



When $v_j \approx v(jh) = v(x_j)$ and $\widehat{v}_j \approx \widehat{v}(j2h) = \widehat{v}(\widehat{x}_j)$, we get using the linear interpolation:

$$v_{2j+1} = \tfrac{1}{2}(\widehat{v}_j + \widehat{v}_{j+1}) \ , \quad v_{2j} = \widehat{v}_j \ .$$

In the matrix language this is

$$\boldsymbol{v} = \boldsymbol{Q}_1 \widehat{\boldsymbol{v}} \ , \qquad \text{where} \qquad \boldsymbol{Q}_1 = \begin{bmatrix} \frac{1}{2} & & & \\ 1 & & & \\ \frac{1}{2} & \frac{1}{2} & & \\ & 1 & & \\ & \frac{1}{2} & \frac{1}{2} & \\ & & \vdots & \\ & & \frac{1}{2} & \frac{1}{2} \\ & & & 1 \\ & & & \frac{1}{2} \end{bmatrix} \in \mathbb{R}^{n\times m} \ .$$

The projection could be done simply $\widehat{u}_j = u_{2j}$, but a weighted mean turns out to be nicer

$$\widehat{u}_j = \tfrac{1}{4}[u_{2j-1} + 2u_{2j} + u_{2j+1}] \ ,$$

since then $\widehat{\boldsymbol{u}} = \frac{1}{2}\boldsymbol{Q}_1^T \boldsymbol{u}$, which brings more symmetry in the iteration.

A naive iteration based purely on the multigrid preconditioning would be

$$\boldsymbol{u}_{k+1} = \boldsymbol{u}_k + \tfrac{1}{2}\boldsymbol{Q}_1 \widehat{\boldsymbol{H}} \boldsymbol{Q}_1^T (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{u}_k) \ ,$$

where $\widehat{\boldsymbol{H}} \approx \widehat{\boldsymbol{A}}^{-1}$ (a good symmetric approximation). Here k is the iteration (not component) index. This however does not work, since in the iteration the preconditioner has to be invertible. Here the rank of the matrix $\boldsymbol{Q}_1 \widehat{\boldsymbol{H}} \boldsymbol{Q}_1^T$ is only m. A working multigrid iteration we get by combining this with a cheap basic iteration on the fine grid. A good choice turns out to be *underrelaxed* Jacobi: $\boldsymbol{H} = \frac{1}{2}\,\mathrm{diag}(\boldsymbol{A})^{-1}$. In order to get the full iteration to become based on a symmetric preconditioner, we perform the Jacobi twice, before and after the coarse grid step. Denoting $\boldsymbol{Q} = \frac{1}{\sqrt{2}}\boldsymbol{Q}_1$ we finally get:

Multigrid iteration:

$$\begin{aligned}\boldsymbol{v}_k &= \boldsymbol{u}_k + \boldsymbol{H}\ (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{u}_k) && \text{Jacobi step}\\ \boldsymbol{w}_k &= \boldsymbol{v}_k + \boldsymbol{Q}\ \widehat{\boldsymbol{H}}\boldsymbol{Q}^T\ (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{v}_k) && \text{coarse grid step}\\ \boldsymbol{u}_{k+1} &= \boldsymbol{w}_k + \boldsymbol{H}\ (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{w}_k) && \text{Jacobi step}\end{aligned}$$

This, in fact, is anly a 2-grid iteration, but if we use the same iteration to build the solver $\widehat{\boldsymbol{H}}$ we recursively get the genuine multigrid iteration.

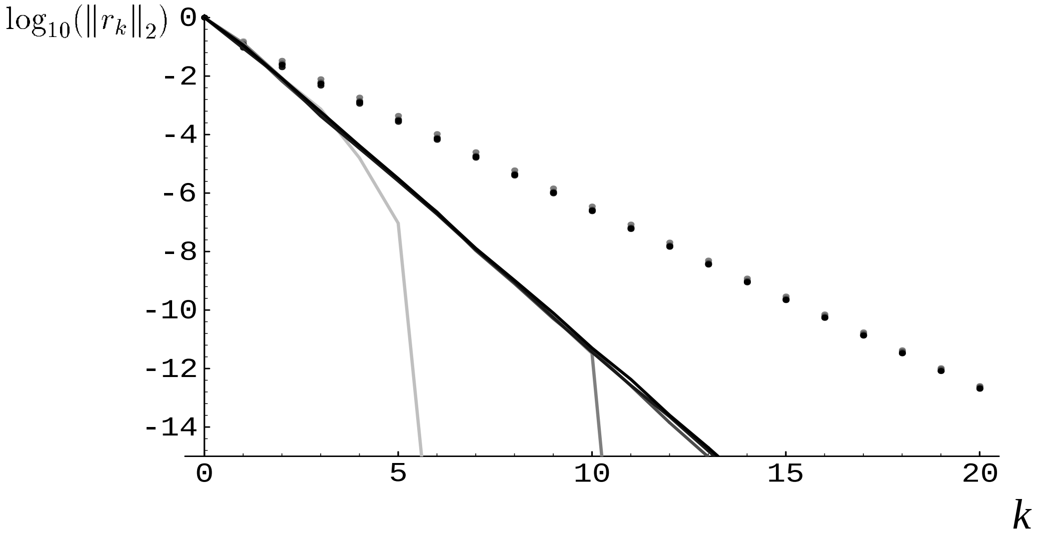
Problem 13.3. Show, that the previous iteration can be written $\boldsymbol{u}_{k+1} = \boldsymbol{u}_k + \widetilde{\boldsymbol{H}}\ (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{u}_k)$, where

$$\widetilde{\boldsymbol{H}} = 2\boldsymbol{H} + \boldsymbol{C} - \boldsymbol{HAH} - \boldsymbol{CAH} - \boldsymbol{HAC} + \boldsymbol{HACAH}$$

and $\boldsymbol{C} = \boldsymbol{Q}\ \widehat{\boldsymbol{H}}\boldsymbol{Q}^T$. Hence the preconditioner is symmetric. Show: if $\rho(\boldsymbol{I} - \boldsymbol{HA}) < 1$ and if $\widehat{\boldsymbol{H}}$ is positive definite, then also $\widetilde{\boldsymbol{H}}$ is positive definite.

Problem 13.4. When the conditions of the previous problem are valid, this works also as a preconditioner for the conjugate gradient method. How do you write the preconditioning step?

In the following figure the one dimensional Poisson equation has been solved with grid sizes $h = 0.1,\ 0.05,\ 0.025,\ 0.0125$ and 0.00625, i.e., for n values $9, 19, 39, 79$ and 159, darker ones corresponding to smaller h. Here $\widehat{\boldsymbol{H}} = \widehat{\boldsymbol{A}}^{-1}$. The direct multigrid iteration is drawn in dots and the preconditioned conjugate gradient method in solid lines.. The remarkable thing here is that the speed does not seem to depend on h. How is this with the classic iterations?



For the two dimensional Poisson equation

$$-u_{xx}(x,y) - u_{yy}(x,y) = f(x,y)\,,\ (x,y) \in (0,1)\times(0,1)$$

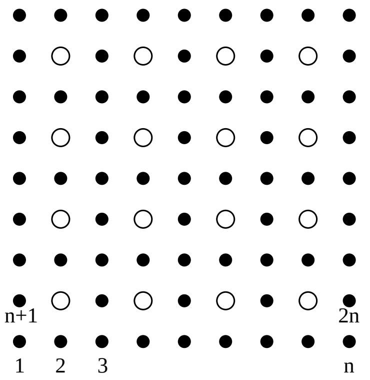
we get respectively the discretization:

$$\begin{aligned} -u_{xx}(ih,jh) &= \frac{1}{h^2}[-u(ih-h,jh) + 2u(ih,jh) - u(ih+h,jh)] \\ -u_{yy}(ih,jh) &= \frac{1}{h^2}[-u(ih,jh-h) + 2u(ih,jh) - u(ih,jh+h)]\ , \end{aligned}$$

and the corresponding matrix

$$\boldsymbol{A} = \tfrac{1}{(n+1)^2}\begin{bmatrix} 4 & -1 & & & -1 & & & & & & \\ -1 & 4 & \ddots & & & -1 & & & & & \\ & \ddots & \ddots & -1 & & & \ddots & & & & \\ & & -1 & 4 & & & & -1 & & & \\ -1 & & & & 4 & -1 & & & -1 & & \\ & -1 & & & -1 & 4 & \ddots & & & \ddots & \\ & & \ddots & & & \ddots & \ddots & & & & -1 \\ & & & -1 & & & & \ddots & \ddots & & \\ & & & & -1 & & & \ddots & \ddots & \ddots & \\ & & & & & \ddots & & & \ddots & 4 & -1 \\ & & & & & & -1 & & & -1 & 4 \end{bmatrix} \in \mathbb{R}^{n^2\times n^2}\ .$$

Let again $n = 2m+1$ and $\widehat{\boldsymbol{A}} \in \mathbb{R}^{m^2\times m^2}$ be the corresponding smaller matrix. The grids are depicted in the following

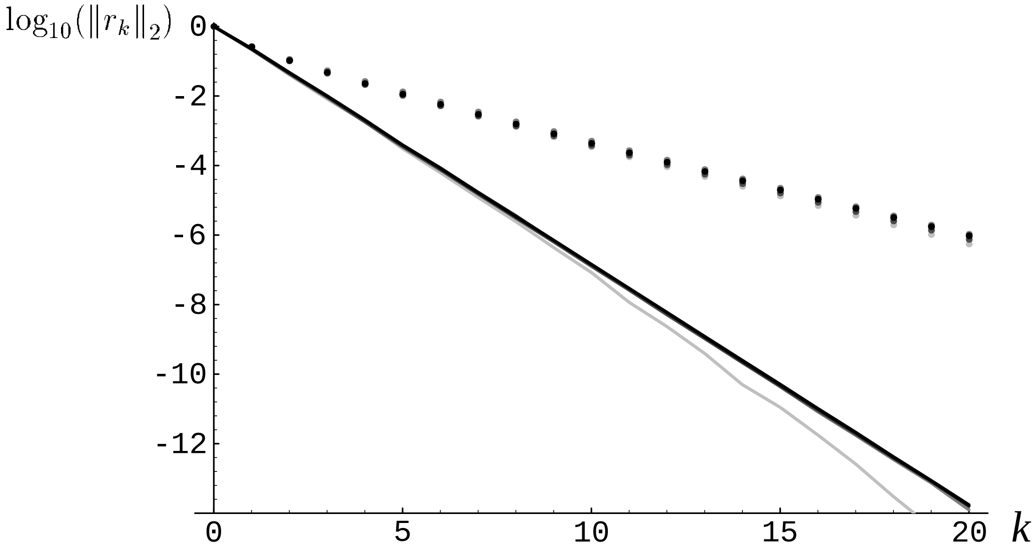


Set $v_{i+nj} \approx v(ih,jh) = v(x_i,y_j)$ and $\widehat{v}_{i+mj} \approx \widehat{v}(i2h,j2h) = \widehat{v}(\widehat{x}_i,\widehat{y}_j)$. If $\boldsymbol{Q}_1$ denotes the interpolation matrix of the 1-dim case, then for the present case the linear interpolation matrix becomes:

$$\boldsymbol{Q}_2 = \begin{bmatrix} \frac{1}{2}\boldsymbol{Q}_1 & & & & \\ \boldsymbol{Q}_1 & & & & \\ \frac{1}{2}\boldsymbol{Q}_1 & \frac{1}{2}\boldsymbol{Q}_1 & & & \\ & \boldsymbol{Q}_1 & & & \\ & \frac{1}{2}\boldsymbol{Q}_1 & \frac{1}{2}\boldsymbol{Q}_1 & & \\ & & \ddots & \ddots & \\ & & & \frac{1}{2}\boldsymbol{Q}_1 & \frac{1}{2}\boldsymbol{Q}_1 \\ & & & & \boldsymbol{Q}_1 \\ & & & & \frac{1}{2}\boldsymbol{Q}_1 \end{bmatrix} = \boldsymbol{Q}_1 \otimes \boldsymbol{Q}_1 \in \mathbb{R}^{n^2\times m^2}\ .$$

Problem 13.5. Show this[2].

The projection becomes respectively $\widehat{v} = \frac{1}{4}Q_2^T v$ and for Q we take $Q = \frac{1}{2}Q_2$. Otherwise we proceed similarly as before and the corresponding figure, again with grid sizes $h = 0.1, 0.05,\ 0.025,\ 0.0125$ and 0.00625 (in the last case $\boldsymbol{A}$ is a $25\,281 \times 25\,281$ matrix, but solving this on a PC is no problem) is the following. The speed is not much worse than in the one dimensional case and it is essentially independent on h.



[2] Above we used the *Kronecker product*: If $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{p\times q}$, then $\boldsymbol{A}\otimes\boldsymbol{B} \in \mathbb{R}^{mp\times nq}$: $(\boldsymbol{A}\otimes\boldsymbol{B})_{(i-1)p+k,(j-1)q+l} = a_{i,j}b_{k,l}$.

References

1. A. M. Bruaset, *A survey of preconditioned iterative methods*, Longman, London, 1995.
2. G. Golub and C. Van Loan, *Matrix computations*, John Hopkins Univerity Press, Baltimore, 1989.
3. G. Golub and G. Meurant, *Résolution numérique des grands systèmes linéaires*, Editions Eyrolles, Paris, 1983.
4. R.A. Horn and C.R. Johnson, *Matrix analysis*, Cambridge University Press, New York, 1985.
5. O. Nevanlinna, *Convergence of iterations for linear equations*, Birkhäuser, Basel, 1993.
6. J. H. Wilkinson, *The algebraic eigenvalue problem*, Clarendon Press, Oxford, 1965.
7. D. Young, *Iterative solution of large linear systems*, Academic Press, New York, 1971.