Numerical Linear Algebra
A Solution Manual

Georg Muntingh and Christian Schulz
Preface

This solution manual gradually appeared over the course of several years, when first Christian Schulz (2005–2008) and then I (2009–2013) were guiding the exercise sessions of a numerical linear algebra course at the University of Oslo.

We would like to thank Tom Lyche for providing solutions to some of the exercises. Several students contributed by pointing out mistakes and improvements to the exercises and solutions. Any remaining mistakes are, of course, our own.

Blindern, November 2013, Georg Muntingh.
## Contents

Preface i

Chapter 0. Preliminaries 1
  Exercise 0.24: The $\mathbf{A}^T\mathbf{A}$ inner product 1
  Exercise 0.25: Angle between vectors in complex case 1
  Exercise 0.41: The inverse of a general $2 \times 2$ matrix 1
  Exercise 0.42: The inverse of a $2 \times 2$ matrix 1
  Exercise 0.43: Sherman-Morrison formula 2
  Exercise 0.44: Cramer’s rule; special case 2
  Exercise 0.45: Adjoint matrix; special case 2
  Exercise 0.47: Determinant equation for a plane 2
  Exercise 0.48: Signed area of a triangle 3
  Exercise 0.49: Vandermonde matrix 3
  Exercise 0.50: Cauchy determinant 4
  Exercise 0.51: Inverse of the Hilbert matrix 5

Chapter 1. Examples of Linear Systems 7
  Exercise 1.2: Gaussian elimination example 7
  Exercise 1.8: Strict diagonal dominance 7
  Exercise 1.9: LU factorization of 2nd derivative matrix 7
  Exercise 1.10: Inverse of 2nd derivative matrix 8
  Exercise 1.11: Central difference approximation of 2nd derivative 8
  Exercise 1.12: Two point boundary value problem (TODO) 9
  Exercise 1.13: Two point boundary value problem; computation (TODO) 9
  Exercise 1.14: Matrix element as a quadratic form 9
  Exercise 1.15: Outer product expansion of a matrix 9
  Exercise 1.16: The product $\mathbf{A}^T\mathbf{A}$ 9
  Exercise 1.17: Outer product expansion 10
  Exercise 1.18: System with many right hand sides; compact form 10
  Exercise 1.19: Block multiplication example 10
  Exercise 1.20: Another block multiplication example 10

Chapter 2. LU Factorizations 11
  Exercise 2.3: Column oriented backsolve (TODO) 11
  Exercise 2.6: Computing the inverse of a triangular matrix 11
  Exercise 2.15: Row interchange 11
  Exercise 2.16: LU of singular matrix 12
  Exercise 2.17: LU and determinant 12
  Exercise 2.18: Diagonal elements in $\mathbf{U}$ 12
  Exercise 2.20: Finite sums of integers 13
  Exercise 2.21: Operations 14
Exercise 5.59: Biorthogonal expansion 32
Exercise 5.60: Generalized Rayleigh quotient 33

Chapter 6. The Singular Value Decomposition 34
Exercise 6.14: SVD examples 34
Exercise 6.15: More SVD examples 35
Exercise 6.17: Counting dimensions of fundamental subspaces 35
Exercise 6.18: Rank and nullity relations 36
Exercise 6.19: Orthonormal bases example 36
Exercise 6.20: Some spanning sets 37
Exercise 6.21: Singular values and eigenpair of composite matrix 37
Exercise 6.27: Rank example 37
Exercise 6.28: Another rank example 38

Chapter 7. Matrix Norms 40
Exercise 7.4: Consistency of sum norm? 40
Exercise 7.5: Consistency of max norm? 40
Exercise 7.6: Consistency of modified max norm? 40
Exercise 7.8: The sum norm is subordinate to? 41
Exercise 7.9: The max norm is subordinate to? 42
Exercise 7.16: Spectral norm 42
Exercise 7.17: Spectral norm of the inverse 42
Exercise 7.18: $p$-norm example 43
Exercise 7.21: Unitary invariance of the spectral norm 43
Exercise 7.22: $\|AU\|_2$ rectangular $A$ 43
Exercise 7.23: $p$-norm of diagonal matrix 43
Exercise 7.24: Spectral norm of a column vector 44
Exercise 7.25: Norm of absolute value matrix 44
Exercise 7.32: Sharpness of perturbation bounds 45
Exercise 7.33: Condition number of 2nd derivative matrix 45
Exercise 7.44: When is a complex norm an inner product norm? 47
Exercise 7.45: $p$-norm for $p = 1$ and $p = \infty$ 48
Exercise 7.46: The $p$-norm unit sphere 48
Exercise 7.47: Sharpness of $p$-norm inequality 48
Exercise 7.48: $p$-norm inequalities for arbitrary $p$ 49

Chapter 8. The Classical Iterative Methods 50
Exercise 8.2: Richardson and Jacobi 50
Exercise 8.13: Convergence of the R-method when eigenvalues have positive real part (TODO) 50
Exercise 8.16: Example: GS converges, J diverges 50
Exercise 8.17: Divergence example for J and GS 51
Exercise 8.18: Strictly diagonally dominance; The J method 51
Exercise 8.19: Strictly diagonally dominance; The GS method 51
Exercise 8.23: Convergence example for fix point iteration 52
Exercise 8.24: Estimate in Lemma 8.22 can be exact 52
Exercise 8.25: Slow spectral radius convergence 53
Exercise 8.31: A special norm (TODO) 55
Exercise 8.33: When is $A + E$ nonsingular? 55
Exercise 12.7: Gerschgorin, strictly diagonally dominant matrix  
Exercise 12.12: $\infty$-norm of a diagonal matrix  
Exercise 12.15: Number of arithmetic operations  
Exercise 12.17: Number of arithmetic operations  
Exercise 12.18: Tridiagonalize a symmetric matrix  
Exercise 12.22: Counting eigenvalues  
Exercise 12.23: Overflow in $LDL^T$ factorization  
Exercise 12.24: Simultaneous diagonalization  
Exercise 12.25: Program code for one eigenvalue  
Exercise 12.26: Determinant of upper Hessenberg matrix (TODO)  

Chapter 13. The QR Algorithm  
Exercise 13.4: Orthogonal vectors  
Exercise 13.14: QR convergence detail (TODO)
CHAPTER 0

Preliminaries

Exercise 0.24: The $A^T A$ inner product

Assume that $A \in \mathbb{R}^{m \times n}$ has linearly independent columns. We show that $\langle \cdot, \cdot \rangle_A : (x, y) \mapsto x^T A^T A y$ satisfies the axioms of an inner product on a real vector space $V$, as described in Definition 0.20. Let $x, y, z \in V$ and $a, b \in \mathbb{R}$, and let $\langle \cdot, \cdot \rangle$ be the standard inner product on $V$.

**Positivity.** One has $\langle x, x \rangle_A = x^T A^T A x = \langle Ax, Ax \rangle \geq 0$, with equality holding if and only if $Ax = 0$. Since $Ax$ is a linearly combination of the columns of $A$ with coefficients the entries of $x$, and since the columns of $A$ are assumed to be linearly independent, one has $Ax = 0$ if and only if $x = 0$.

**Symmetry.** One has $\langle x, y \rangle_A = x^T A^T A y = (x^T A^T A y)^T = y^T A^T A x = \langle y, x \rangle_A$.

**Linearity.** One has $\langle ax + by, z \rangle_A = (ax + by)^T A^T A z = ax^T A^T A z + by^T A^T A z = a \langle x, z \rangle_A + b \langle y, z \rangle_A$.

Exercise 0.25: Angle between vectors in complex case

By the Cauchy-Schwarz inequality for a complex inner product space,

$$0 \leq \frac{|\langle x, y \rangle|}{\|x\| \|y\|} \leq 1.$$ 

Note that taking $x$ and $y$ perpendicular yields zero, taking $x$ and $y$ equal yields one, and any value in between can be obtained by picking an appropriate affine combination of these two cases.

Since the cosine decreases monotonously from one to zero on the interval $[0, \pi/2]$, there is a unique argument $\theta \in [0, \pi/2]$ such that

$$\cos \theta = \frac{|\langle x, y \rangle|}{\|x\| \|y\|}.$$ 

Exercise 0.41: The inverse of a general $2 \times 2$ matrix

A straightforward computation yields

$$\frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} ad - bc & 0 \\ 0 & ad - bc \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

showing that the two matrices are inverse to each other.

Exercise 0.42: The inverse of a $2 \times 2$ matrix

By Exercise 0.41, and using that $\cos^2 \theta + \sin^2 \theta = 1$, the inverse is given by

$$\begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix}.$$
Exercise 0.43: Sherman-Morrison formula

A direct computation yields
\[
(A + BC)(A^{-1} - A^{-1}B(I + C^{T}A^{-1}B)^{-1}C^{T}A^{-1})
= I - B(I + C^{T}A^{-1}B)^{-1}C^{T}A^{-1} + BC^{T}A^{-1} - BC^{T}A^{-1}B(I + C^{T}A^{-1}B)^{-1}C^{T}A^{-1}
= I + BC^{T}A^{-1} - B(I + C^{T}A^{-1}B)(I + C^{T}A^{-1}B)^{-1}C^{T}A^{-1}
= I + BC^{T}A^{-1} - BC^{T}A^{-1}
= I,
\]
showing that the two matrices are inverse to each other.

Exercise 0.44: Cramer’s rule; special case

Cramer’s rule yields
\[
x_1 = \frac{\begin{vmatrix} 3 & 2 \\ 6 & 1 \end{vmatrix}}{\begin{vmatrix} 1 & 2 \\ 2 & 1 \end{vmatrix}} = 3, \quad x_2 = \frac{\begin{vmatrix} 1 & 3 \\ 2 & 1 \end{vmatrix}}{\begin{vmatrix} 1 & 2 \\ 2 & 1 \end{vmatrix}} = 0.
\]

Exercise 0.45: Adjoint matrix; special case

We are given the matrix
\[
A = \begin{bmatrix} 2 & -6 & 3 \\ 3 & -2 & -6 \\ 6 & 3 & 2 \end{bmatrix}.
\]
Computing the cofactors of A gives
\[
adj^{T}A = \begin{bmatrix} (-1)^{1+1} & -2 & -6 \\ 3 & 2 & 6 \\ (-1)^{1+2} & 3 & -6 \\ 6 & 2 & 3 \\ (-1)^{1+3} & 3 & -2 \\ 6 & 3 & 2 \end{bmatrix}
= \begin{bmatrix} 14 & 21 & 42 \\ -42 & -14 & 21 \\ 21 & -42 & 14 \end{bmatrix}^{T}.
\]

One checks directly that \(adj_{A}A = det(A)I\), with \(det(A) = 343\).

Exercise 0.47: Determinant equation for a plane

Let \(ax + by + cz + d = 0\) be an equation for a plane through the points \((x_i, y_i, z_i)\), with \(i = 1, 2, 3\). There is precisely one such plane if and only if the points are not colinear. Then \(ax_i + by_i + cz_i + d = 0\) for \(i = 1, 2, 3\), so that
\[
\begin{bmatrix} x & y & z & 1 \\ x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.
\]
Since the coordinates \(a, b, c, d\) of the plane are not all zero, the above matrix is singular, implying that its determinant is zero. Computing this determinant by cofactor expansion of the first row gives the equation

\[
\begin{vmatrix}
 y_1 & z_1 & 1 & x_1 & y_1 & z_1 \\
 y_2 & z_2 & 1 & x_2 & y_2 & z_2 \\
 y_3 & z_3 & 1 & x_3 & y_3 & z_3 \\
\end{vmatrix} + \begin{vmatrix}
 y_1 & z_1 & 1 & y \\
 y_2 & z_2 & 1 & y \\
 y_3 & z_3 & 1 & y \\
\end{vmatrix} + \begin{vmatrix}
 x_1 & y_1 & z_1 & x \\
 x_2 & y_2 & z_1 & x \\
 x_3 & y_3 & z_1 & x \\
\end{vmatrix} = 0
\]

of the plane.

**Exercise 0.48: Signed area of a triangle**

Let \(T\) denote the triangle with vertices \(P_1, P_2, P_3\). Since the area of a triangle is invariant under translation, we can assume \(P_1 = A = (0, 0), P_2 = (x_2, y_2), P_3 = (x_3, y_3), B = (x_3, 0),\) and \(C = (x_2, 0)\). As is clear from Figure 2, the area \(A(T)\) can be expressed as

\[
A(T) = A(ABP_3) + A(P_3BCP_2) - A(ACP_2)
\]

\[
= \frac{1}{2} x_3 y_3 + (x_2 - x_3) y_2 + \frac{1}{2} (x_2 - x_3)(y_3 - y_2) - \frac{1}{2} x_2 y_2
\]

\[
= \frac{1}{2} \begin{vmatrix}
 1 & 1 & 1 \\
 1 & x_2 & x_3 \\
 0 & y_2 & y_3
\end{vmatrix}
\]

which is what needed to be shown.

**Exercise 0.49: Vandermonde matrix**

For any \(n = 1, 2, \ldots,\) let

\[
D_n := \begin{vmatrix}
 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\
 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\
 1 & x_3 & x_3^2 & \cdots & x_3^{n-1} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
 1 & x_n & x_n^2 & \cdots & x_n^{n-1}
\end{vmatrix}
\]

be the determinant of the Vandermonde matrix in the Exercise. Clearly the formula

\[
(*) \quad D_N = \prod_{1 \leq j < i \leq N} (x_i - x_j)
\]

holds for \(N = 1\) (in which case the product is empty and defined to be 1) and \(N = 2\).

Let us assume \((*)\) holds for \(N = n - 1 > 2\). Since the determinant is an alternating multilinear form, adding a scalar multiple of one column to another does not change the value of the determinant. Subtracting \(x_k^k\) times column \(k\) from column \(k + 1\) for \(k = n - 1, n - 2, \ldots, 1\), we find

\[
D_n = \begin{vmatrix}
 1 & x_1 - x_n & x_1^2 - x_1 x_n & \cdots & x_1^{n-1} - x_1^{n-2} x_n \\
 1 & x_2 - x_n & x_2^2 - x_2 x_n & \cdots & x_2^{n-1} - x_2^{n-2} x_n \\
 1 & x_3 - x_n & x_3^2 - x_3 x_n & \cdots & x_3^{n-1} - x_3^{n-2} x_n \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
 1 & x_n - x_n & x_n^2 - x_n x_n & \cdots & x_n^{n-1} - x_n^{n-2} x_n
\end{vmatrix}
\]
Next, by cofactor expansion along the last row and by the multilinearity in the rows,

\[
D_n = (-1)^{n-1} \cdot 1 \cdot \begin{vmatrix}
    x_1 - x_n & x_1^2 - x_1x_n & \cdots & x_1^{n-1} - x_1^{n-2}x_n \\
    x_2 - x_n & x_2^2 - x_2x_n & \cdots & x_2^{n-1} - x_2^{n-2}x_n \\
    \vdots & \vdots & \ddots & \vdots \\
    x_{n-1} - x_n & x_{n-1}^2 - x_{n-1}x_n & \cdots & x_{n-1}^{n-1} - x_{n-1}^{n-2}x_n \\
\end{vmatrix}
\]

\[
= (-1)^{n-1}(x_1 - x_n)(x_2 - x_n) \cdots (x_{n-1} - x_n)D_{n-1}
\]

\[
= (x_n - x_1)(x_n - x_2) \cdots (x_n - x_{n-1}) \prod_{1 \leq j < i \leq n-1} (x_i - x_j).
\]

By induction, we conclude that (*) holds for any \( N = 1, 2, \ldots \)

**Exercise 0.50: Cauchy determinant**

(a) Let \([\alpha_1, \dotsc, \alpha_n]^T, [\beta_1, \dotsc, \beta_n]^T \in \mathbb{R}^n\) and let

\[
A = (a_{i,j})_{i,j} = \left( \frac{1}{\alpha_i + \beta_j} \right)_{i,j}.
\]

Multiplying the \(i\)th row of \(A\) by \(\prod_{k=1}^n (\alpha_i + \beta_k)\) for \(i = 1, 2, \ldots, n\) gives a matrix

\[
C = (c_{i,j})_{i,j}, \quad c_{i,j} = \prod_{k=1}^n (\alpha_i + \beta_k).
\]

The determinant of an \(n \times n\) matrix is a homogeneous polynomial of degree \(n\) in the entries of the matrix. Since each entry of \(C\) is a polynomial of degree \(n-1\) in the variables \(\alpha_i, \beta_j\), the determinant of \(C\) must be a homogeneous polynomial of degree \(n(n-1)\) in \(\alpha_i, \beta_j\).

By the multilinearity of the determinant, \(\det C = \prod_{i,j=1}^n (\alpha_i + \beta_j) \det A\). Since \(A\) vanishes whenever \(\alpha_i = \alpha_j\) or \(\beta_i = \beta_j\) for \(i \neq j\), the homogeneous polynomial \(\det C\) contains factors \((\alpha_i - \alpha_j)\) and \((\beta_i - \beta_j)\) for \(1 \leq i < j \leq n\). As there are precisely \(2 \cdot \binom{n}{2} = (n-1)n\) such factors, necessarily

\[
(*) \quad \det C = k \prod_{1 \leq i < j \leq n} (\alpha_i - \alpha_j) \prod_{1 \leq i < j \leq n} (\beta_i - \beta_j)
\]

for some constant \(k\). To determine \(k\), we can evaluate \(\det C\) at a particular value, for instance any \(\{\alpha_i, \beta_j\}_{i,j}\) satisfying \(\alpha_1 + \beta_1 = \cdots = \alpha_n + \beta_n = 0\). In that case \(C\) becomes a diagonal matrix with determinant

\[
\det C = \prod_{i=1}^n \prod_{k=1}^n (\alpha_i + \beta_k) = \prod_{i=1}^n \prod_{k=1}^n (\alpha_i - \alpha_k) = \prod_{1 \leq i < k \leq n} (\alpha_i - \alpha_k) \prod_{1 \leq i < k \leq n} (\alpha_k - \alpha_i).
\]


Comparing with (⋆) shows that $k = 1$. We conclude that
\[
(\star\star) \quad \det A = \frac{\prod_{1 \leq i < j \leq n}(\alpha_i - \alpha_j) \prod_{1 \leq i < j \leq n}(\beta_i - \beta_j)}{\prod_{i,j=1}^n(\alpha_i + \beta_j)}.
\]

(b) Deleting row $l$ and column $k$ from $A$, results in the matrix $A_{l,k}$ associated to the vectors $[\alpha_1, \ldots, \alpha_{l-1}, \alpha_{l+1}, \ldots, \alpha_n]$ and $[\beta_1, \ldots, \beta_{k-1}, \beta_{k+1}, \ldots, \beta_n]$. By the adjoint formula for the inverse $A^{-1} = (b_{k,l})$ and by (⋆\star),
\[
b_{k,l} := (-1)^{k+l} \frac{\det A_{l,k}}{\det A}
= (-1)^{k+l} \frac{\prod_{i,j=1}^n(\alpha_i + \beta_j) \prod_{1 \leq i < j \leq n}(\alpha_i - \alpha_j) \prod_{1 \leq i < j \leq n}(\beta_i - \beta_j)}{\prod_{i,j=1}^n(\alpha_i + \beta_j) \prod_{i \neq l} \prod_{j \neq k}(\alpha_i - \alpha_j) \prod_{1 \leq i < j \leq n}(\beta_i - \beta_j)}
= (\alpha_l + \beta_k) \prod_{s=1}^n (\alpha_s + \beta_k) \prod_{s=1}^n (\beta_s + \alpha_l)
\prod_{s=1}^n (\alpha_s - \alpha_l) \prod_{s=1}^n (\beta_s - \beta_k)
= (\alpha_l + \beta_k) \prod_{s=1}^n \frac{\alpha_s + \beta_k}{\alpha_s - \alpha_l} \prod_{s=1}^n \frac{\beta_s + \alpha_l}{\beta_s - \beta_k},
\]
which is what needed to be shown.

**Exercise 0.51: Inverse of the Hilbert matrix**

If we write
\[
\alpha = [\alpha_1, \ldots, \alpha_n] = [1, 2, \ldots, n], \quad \beta = [\beta_1, \ldots, \beta_n] = [0, 1, \ldots, n - 1],
\]
then the Hilbert matrix matrix is of the form $H_n = (h_{i,j}) = (1/(\alpha_i + \beta_j))$. By Exercise 0.50.(b), its inverse $T_n = (t_{i,j}^n) := H_n^{-1}$ is given by
\[
t_{i,j}^n = (i + j - 1) \prod_{s=1}^{i+j-1} s + \frac{i - 1}{s - j} \prod_{s=1}^{i+j-1} s + \frac{j - 1}{s - i}, \quad 1 \leq i, j \leq n.
\]

We wish to show that
\[
(\star) \quad t_{i,j}^n = \frac{f(i)f(j)}{i + j - 1}, \quad 1 \leq i, j \leq n,
\]
where $f : \mathbb{N} \rightarrow \mathbb{Q}$ is the sequence defined by
\[
f(1) = -n, \quad f(i + 1) = \left(\frac{i^2 - n^2}{i^2}\right) f(i), \quad \text{for } i = 1, 2, \ldots.
\]
Clearly (⋆) holds when $i = j = 1$. Suppose that (⋆) holds for some $(i, j)$. Then

$$t_{i+1,j}^n = (i + j) \prod_{s=1, s \neq j}^{n} \frac{s + 1 + i - 1}{s - j} \prod_{s=1, s \neq i+1}^{n} \frac{s + j - 1}{s - 1 - i}$$

$$= (i + j) \frac{1}{(i + j)^2} \prod_{s=2, s \neq j}^{n+1} (s + i - 1) \prod_{s=1, s \neq i+1}^{n} (s + j - 1)$$

$$= (i + j - 1)^2(n + i)(n - i) \prod_{s=1, s \neq j}^{n} (s + i - 1) \prod_{s=1, s \neq i+1}^{n} (s + j - 1)$$

$$= \frac{1}{i + j} \frac{i^2 - n^2}{i^2} (i + j - 1) \prod_{s=1, s \neq j}^{n} \frac{s + i - 1}{s - j} (i + j - 1) \prod_{s=1, s \neq i+1}^{n} \frac{s + j - 1}{s - i}$$

$$= \frac{1}{i + j} \frac{i^2 - n^2}{i^2} f(i) f(j)$$

$$= \frac{f(i + 1)f(j)}{(i + 1) + j - 1},$$

so that (⋆) holds for $(i + 1, j)$. Carrying out a similar calculation for $(i, j + 1)$, or using the symmetry of $T_n$, we conclude by induction that (⋆) holds for any $i, j$. 
CHAPTER 1

Examples of Linear Systems

Exercise 1.2: Gaussian elimination example

Applying the recipe of Example 1.1 to the augmented matrix $[A|b]$, yields

$$\begin{bmatrix} 1 & 1 & -1 & 1 \\ -1 & 1 & 3 & 1 \\ 2 & 8 & 3 & 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 1 & -1 & 1 \\ 0 & 2 & 2 & 2 \\ 0 & 6 & 5 & -1 \end{bmatrix} \sim \begin{bmatrix} 1 & 1 & -1 & 1 \\ 0 & 2 & 2 & 2 \\ 0 & 0 & -1 & -7 \end{bmatrix}.$$  

Using back substitution we find that $x = [14, -6, 7]^T$.

Exercise 1.8: Strict diagonal dominance

Suppose that the matrix $A = \text{tridiag}(a_i, d_i, c_i) \in \mathbb{C}^{n \times n}$ as in (1.3) is strictly diagonally dominant. To show that an LU factorization of $A$ exists, we again need to show that $u_1, \ldots, u_{n-1}$ in (1.4) are nonzero. It therefore suffices to show by induction that

$$|u_k| > |c_k|, \quad k = 1, \ldots, n - 1. \quad (\star)$$

This clearly holds for $k = 1$, since $A$ is strictly diagonally dominant. Assume $(\star)$ holds for some $k$ satisfying $1 \leq k \leq n - 2$. Using (1.4), the reverse triangle inequality, the induction hypothesis, and strict diagonal dominance of $A$, we have

$$|u_{k+1}| = \left| \frac{d_{k+1} - a_k c_k}{u_k} \right| \geq |d_{k+1}| - \left| \frac{c_k}{u_k} |a_k| \right| \geq |d_{k+1}| - |a_k| > |c_{k+1}|,$$

and it follows by induction that an LU factorization exists. Moreover, since the LU factorization is completely determined by (1.4), it is unique.

Exercise 1.9: LU factorization of 2nd derivative matrix

Let $L = (l_{ij})_{ij}, U = (r_{ij})_{ij}$ and $T$ be as in the exercise. Clearly $L$ is unit lower triangular and $U$ is upper triangular. We compute the product $LU$ by separating cases for its entries. There are several ways to carry out and write down this computation, some more precise than others. For instance,

$$(LU)_{11} = 1 \cdot 2 = 2;$$

$$(LU)_{ii} = -\frac{i - 1}{i} \cdot -1 + 1 \cdot \frac{i + 1}{i} = 2, \quad \text{for } i = 2, \ldots, m;$$

$$(LU)_{i,i-1} = -\frac{i - 1}{i} \cdot \frac{i}{i - 1} = -1, \quad \text{for } i = 2, \ldots, m;$$

$$(LU)_{i-1,i} = 1 \cdot -1 = -1, \quad \text{for } i = 2, \ldots, m;$$

$$(LU)_{ij} = 0, \quad \text{for } |i - j| \geq 2.$$  

It follows that $T = LU$ is an LU factorization.
Another way to show that $T = LU$ is by induction. For $m = 1$, one has $L_1U_1 = 1 \cdot 2 = T_1$. Now let $m > 1$ be arbitrary and assume that $L_mU_m = T_m$. With
\[
a := [0, \ldots, 0, -\frac{m}{m+1}]^T, \quad b := [0, \ldots, 0, -1]^T,
\]
block multiplication yields
\[
L_{m+1}U_{m+1} = \begin{bmatrix} L_m & 0 \\ a^T & 1 \end{bmatrix} \begin{bmatrix} U_m & b \\ \frac{m+2}{m+1} \end{bmatrix} = \begin{bmatrix} T_m & \frac{L_mb}{m+2} \\ \frac{a^TU_m}{m+2} & \frac{b}{m+2} \end{bmatrix} = T_{m+1}.
\]
By induction, we can then conclude that $T_m = L_mU_m$ for all $m \geq 1$.

**Exercise 1.10: Inverse of 2nd derivative matrix**

Let $S = (s_{ij})_{ij}$ be defined by
\[s_{ij} = s_{ji} = \left(1 - \frac{i}{m+1}\right)j, \quad \text{for } 1 \leq j \leq i \leq m.
\]
In order to show that $S = T^{-1}$, we multiply $S$ by $T$ and show that the result is the identity matrix. To simplify notation we define $s_{ij} := 0$ whenever $i = 0$, $i = m+1$, $j = 0$, or $j = m+1$. With $1 \leq j < i \leq m$, we find
\[
(ST)_{i,j} = \sum_{k=1}^{m} s_{i,k}T_{k,j} = -s_{i,j-1} + 2s_{i,j} - s_{i,j+1}
\]
\[= \left(1 - \frac{i}{m+1}\right)(-j + 1 + 2j - j - 1) = 0,
\]
\[
(ST)_{j,i} = \sum_{k=1}^{m} s_{j,k}T_{k,i} = -s_{j,i-1} + 2s_{j,i} - s_{j,i+1}
\]
\[= -\left(1 - \frac{i-1}{m+1}\right)j + 2\left(1 - \frac{i}{m+1}\right)j - \left(1 - \frac{i+1}{m+1}\right)j
\]
\[= -j + 2j - j + j \cdot \frac{i-1-2i+i+1}{m+1} = 0,
\]
\[
(ST)_{i,i} = \sum_{k=1}^{m} s_{i,k}T_{k,i} = -s_{i,i-1} + 2s_{i,i} - s_{i,i+1}
\]
\[= -\left(1 - \frac{i}{m+1}\right)(i-1) + 2\left(1 - \frac{i}{m+1}\right)i - \left(1 - \frac{i+1}{m+1}\right)i = 1
\]
which means that $ST = I$. Moreover, since $S$, $T$, and $I$ are symmetric, transposing this equation yields $TS = I$. We conclude that $S = T^{-1}$.

**Exercise 1.11: Central difference approximation of 2nd derivative**

If all $h_i$ equal to the same number $h$, then
\[
\lambda_i = \mu_i = \frac{2h}{h+h} = 1, \quad \delta_i = \frac{y_{i+1} - y_i}{h}, \quad \beta_i = 3(\delta_{i-1} + \delta_i) = 3\frac{y_{i+1} - y_{i-1}}{h},
\]
which is what needed to be shown.
Exercise 1.12: Two point boundary value problem (TODO)

Exercise 1.13: Two point boundary value problem; computation (TODO)

Exercise 1.14: Matrix element as a quadratic form

Write $A = (a_{ij})_{ij}$ and $e_i = (\delta_{ik})_k$, where

$$\delta_{ik} = \begin{cases} 1 & \text{if } i = k, \\ 0 & \text{otherwise}, \end{cases}$$

is the Kronecker delta. Then, by the definition of the matrix product,

$$e_i^T A e_j = e_i^T (A e_j) = e_i^T \left( \sum_k a_{ik} \delta_{jk} \right)_l = e_i^T (a_{lj}) = \sum_l \delta_{il} a_{lj} = a_{ij}.$$ 

Exercise 1.15: Outer product expansion of a matrix

Let $\delta_{ij}$ denote the Kronecker delta. For any indices $1 \leq k \leq m$ and $1 \leq l \leq n$, the $(k,l)$-th entry of the matrix $e_i e_j^T$ satisfies

$$(e_i e_j^T)_{kl} = \sum_o (e_i)_{ko} (e_j^T)_{ol} = (e_i)_{k1} (e_j^T)_{1l} = \delta_{ik} \delta_{jl}.$$ 

It follows that

$$\left( \sum_i \sum_j a_{ij} e_i e_j^T \right)_{kl} = \sum_i \sum_j a_{ij} (e_i e_j^T)_{kl} = \sum_i \sum_j a_{ij} \delta_{ik} \delta_{jl} = a_{kl}$$ 

for any indices $k,l$, implying the statement of the Exercise.

Exercise 1.16: The product $A^T A$

A matrix product is defined as long as the dimensions of the matrices are compatible. More precisely, for the matrix product $AB$ to be defined, the number of columns in $A$ must equal the number of rows in $B$.

Let now $A$ be an $n \times m$ matrix. Then $A^T$ is an $m \times n$ matrix, and as a consequence the product $B := A^T A$ is well defined. Moreover, the $(i,j)$-th entry of $B$ is given by

$$(B)_{ij} = (A^T A)_{ij} = \sum_{k=1}^n a_{ki} a_{kj} = a_j^T a_i = \langle a_i, a_j \rangle,$$

which is what needed to be shown.
Exercise 1.17: Outer product expansion

Recall that the matrix product of $A \in \mathbb{C}^{m,n}$ and $B^T = C \in \mathbb{C}^{n,p}$ is defined by

$$(AC)_{ij} = \sum_{k=1}^{n} a_{ik} c_{kj} = \sum_{k=1}^{n} a_{ik} b_{kj}.$$ 

For the outer product expansion of the columns of $A$ and $B$, on the other hand, we find $(a_k b_k^T)_{ij} = a_{ik} b_{jk}$. It follows that

$$(AB^T)_{ij} = \sum_{k=1}^{n} a_{ik} b_{jk} = \sum_{k=1}^{n} (a_k b_k^T)_{ij}.$$ 

Exercise 1.18: System with many right hand sides; compact form

Let $A$, $B$, and $X$ be as in the Exercise.

$(\implies)$: Suppose $AX = B$. Multiplying this equation from the right by $e_j$ yields $Ax_j = b_j$ for $j = 1, \ldots, p$.

$(\impliedby)$: Suppose $Ax_j = b_j$ for $j = 1, \ldots, p$. Let $I = I_p$ denote the identity matrix. Then

$$AX = AXI = AX[e_1, \ldots, e_p] = [AXe_1, \ldots, AXe_p] = [Ax_1, \ldots, Ax_p] = [b_1, \ldots, b_p] = B.$$ 

Exercise 1.19: Block multiplication example

The product $AB$ of two matrices $A$ and $B$ is defined precisely when the number of columns of $A$ is equal to the number of rows of $B$. For both sides in the equation $AB = A_1 B_1$ to make sense, both pairs $(A, B)$ and $(A_1, B_1)$ need to be compatible in this way. Conversely, if the number of columns of $A$ equals the number of rows of $B$ and the number of columns of $A_1$ equals the number of rows of $B_1$, then there exist integers $m, p, n,$ and $s$ with $1 \leq s \leq p$ such that

$$A \in \mathbb{C}^{m,p}, \quad B \in \mathbb{C}^{p,n}, \quad A_1 \in \mathbb{C}^{m,s}, \quad A_2 \in \mathbb{C}^{m,p-s}, \quad B_1 \in \mathbb{C}^{s,n}.$$ 

Then

$$(AB)_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj} = \sum_{k=1}^{s} a_{ik} b_{kj} + \sum_{k=s+1}^{p} a_{ik} \cdot 0 = (A_1 B_1)_{ij}.$$ 

Exercise 1.20: Another block multiplication example

Since the matrices have compatible dimensions, a direct computation gives

$$CAB = \begin{bmatrix} 1 & 0^T & \lambda & a^T & 1 & 0^T \\ 0 & C_1 & 0 & A_1 & 0 & B_1 \end{bmatrix} = \begin{bmatrix} \lambda & a^T \\ 0 & C_1 A_1 \end{bmatrix} \begin{bmatrix} 1 & 0^T \\ 0 & C_1 B_1 \end{bmatrix} = \begin{bmatrix} \lambda & a^T B_1 \\ 0 & C_1 A_1 B_1 \end{bmatrix}.$$
Exercise 2.3: Column oriented backsolve (TODO)

Exercise 2.6: Computing the inverse of a triangular matrix

This exercise introduces an efficient method for computing the inverse $B$ of a triangular matrix $A$.

Let us solve the problem for an upper triangular matrix (the lower triangular case is similar). By the rules of block multiplication,

$$[Ab_1, \ldots, Ab_n] = A[b_1, \ldots, b_n] = AB = I = [e_1, \ldots, e_n].$$

By Lemma 1.22, the matrix $B$ is upper triangular, implying that the other entries $b_{k+1,k}$, $\ldots, b_{n,k}$ in $b_k$ are zero. The $k$th column in this matrix equation can be partitioned into blocks, as

$$
\begin{bmatrix}
a_{11} & \cdots & a_{1,k} & a_{1,k+1} & \cdots & a_{1,n} \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
a_{k,k} & a_{k,k+1} & \cdots & a_{k,n} \\
0 & a_{k+1,k} & \cdots & a_{k+1,n} \\
& & \vdots & \ddots & \vdots & \vdots \\
& & & \vdots & \ddots & \ddots \\
0 & & & & \vdots & \ddots \\
& & & & & \ddots \\
0 & & & & & \ddots \\
\end{bmatrix}
\begin{bmatrix}
b_{1k} \\
\vdots \\
b_{kk} \\
0 \\
\vdots \\
0 \\
0 \\
\end{bmatrix}
= 
\begin{bmatrix}
0 \\
\vdots \\
0 \\
1 \\
\vdots \\
0 \\
0 \\
\end{bmatrix}
.$$  

Evaluating the upper block matrix multiplication then yields (2.4). Solving the above system thus yields the $k$th column of $B$.

Performing this block multiplication for $k = n, n - 1, \ldots, 1$, we see that the computations after step $k$ only use the first $k - 1$ leading principal submatrices of $A$. It follows that the column $b_k$ computed at step $k$ can be stored in row (or column) $k$ of $A$ without altering the remaining computations.

Exercise 2.15: Row interchange

Suppose we are given an LU factorization

$$
\begin{bmatrix}
1 & 1 \\
0 & 1 \\
\end{bmatrix}
= 
\begin{bmatrix}
1 & 0 \\
l_{21} & 1 \\
\end{bmatrix}
\begin{bmatrix}
u_{11} & u_{12} \\
0 & u_{22} \\
\end{bmatrix}.
$$

Carrying out the matrix multiplication on the right hand side, one finds that

$$
\begin{bmatrix}
1 & 1 \\
0 & 1 \\
\end{bmatrix}
= 
\begin{bmatrix}
u_{11} & u_{12} \\
l_{21}u_{11} & l_{21}u_{12} + u_{22} \\
\end{bmatrix}.
$$
implying that \( u_{11} = u_{12} = 1 \). It follows that necessarily \( l_{21} = 0 \) and \( u_{22} = 1 \), and the pair

\[
L = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad U = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}
\]

is the only possible LU factorization of the matrix \( \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \). One directly checks that this is indeed an LU factorization.

**Exercise 2.16: LU of singular matrix**

Suppose we are given an LU factorization

\[
\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ l_{21} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix}.
\]

Carrying out the matrix multiplication on the right hand side, one finds that

\[
\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} \\ l_{21}u_{11} & l_{21}u_{12} + u_{22} \end{bmatrix},
\]

implying that \( u_{11} = u_{12} = 1 \). It follows that necessarily \( l_{21} = 1/u_{11} = 1 \) and \( u_{22} = 1 - l_{21}u_{12} = 0 \), and the pair

\[
L = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \quad U = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}
\]

is the only possible LU factorization of the matrix \( \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \). One directly checks that this is indeed an LU factorization.

**Exercise 2.17: LU and determinant**

Suppose \( A \) has an LU factorization \( A = LU \). Then, by Lemma 2.11, \( A_{[k]} = L_{[k]}U_{[k]} \) is an LU factorization for \( k = 1, \ldots, n \). By induction, the cofactor expansion of the determinant yields that the determinant of a triangular matrix is the product of its diagonal entries. One therefore finds that \( \det(L_{[k]}) = 1, \det(U_{[k]}) = u_{11} \cdots u_{kk} \) and

\[
\det(A_{[k]}) = \det(L_{[k]}U_{[k]}) = \det(L_{[k]}) \det(U_{[k]}) = u_{11} \cdots u_{kk}
\]

for \( k = 1, \ldots, n \).

**Exercise 2.18: Diagonal elements in U**

From Exercise 2.17, we know that \( \det(A_{[k]}) = u_{11} \cdots u_{kk} \) for \( k = 1, \ldots, n \). Since \( A \) is nonsingular, its determinant \( \det(A) = u_{11} \cdots u_{nn} \) is nonzero. This implies that \( \det(A_{[k]}) = u_{11} \cdots u_{kk} \neq 0 \) for \( k = 1, \ldots, n \), yielding \( a_{11} = u_{11} \) for \( k = 1 \) and a well-defined quotient

\[
\frac{\det(A_{[k]})}{\det(A_{[k-1]})} = \frac{u_{1,1} \cdots u_{k-1,k-1} u_{k,k}}{u_{1,1} \cdots u_{k-1,k-1}} = u_{k,k}
\]

for \( k = 2, \ldots, n \).
Exercise 2.20: Finite sums of integers

There are many ways to prove these identities. While the quickest way to prove these identities is by induction, we choose a generating function approach because it is a powerful method that works in a wide range of circumstances.

It is easily checked that the identities hold for \( m = 1, 2, 3 \). So let \( m \geq 4 \) and let

\[
P_m := 1 + x + \cdots + x^m = \frac{1 - x^{m+1}}{1 - x}.
\]

Then

\[
P'_m = \frac{1 - (m + 1)x^m + mx^{m+1}}{(x - 1)^2},
\]

\[
P''_m = \frac{-2 + (m^2 + m)x^{m-1} + 2(1 - m^2)x^m + (m^2 - m)x^{m+1}}{(x - 1)^3}.
\]

Applying l’Hôpital’s rule twice, we find

\[
1 + 2 + \cdots + m = P'_m(1)
\]

\[
eq \lim_{x \to 1} \frac{1 - (m + 1)x^m + mx^{m+1}}{(x - 1)^2}
\]

\[
eq \lim_{x \to 1} \frac{-m(m + 1)x^{m-1} + m(m + 1)x^m}{2(x - 1)}
\]

\[
= \frac{1}{2} m(m + 1),
\]

establishing (2.10). In addition it follows that

\[
1 + 3 + \cdots + 2m - 1 = \sum_{k=1}^{m}(2k - 1) = -m + 2 \sum_{k=1}^{m} k = -m + m(m + 1) = m^2,
\]

which establishes (2.12). Next, applying l’Hôpital’s rule three times, we find that

\[
1 \cdot 2 + 2 \cdot 3 + \cdots + (m - 1) \cdot m = P''_m(1)
\]

is equal to

\[
= \lim_{x \to 1} \frac{-2 + (m^2 + m)x^{m-1} + 2(1 - m^2)x^m + (m^2 - m)x^{m+1}}{(x - 1)^3}
\]

\[
= \lim_{x \to 1} \frac{(m - 1)(m^2 + m)x^{m-2} + 2m(1 - m^2)x^{m-1} + (m + 1)(m^2 - m)x^m}{3(x - 1)^2}
\]

\[
= \lim_{x \to 1} \frac{(m - 2)(m - 1)(m^2 + m)x^{m-3} + 2(m - 1)m(1 - m^2)x^{m-2} + m(m + 1)(m^2 - m)x^{m-1}}{6(x - 1)}
\]

\[
= \frac{1}{3} (m - 1)m(m + 1),
\]

establishing (2.13). Finally,

\[
1^2 + 2^2 + \cdots + m^2 = \sum_{k=1}^{m} k^2 = \sum_{k=1}^{m} ((k - 1)k + k) = \sum_{k=1}^{m} (k - 1)k + \sum_{k=1}^{m} k
\]

\[
= \frac{1}{3} (m - 1)m(m + 1) + \frac{1}{2} m(m + 1) = \frac{1}{3} (m + 1)(m + \frac{1}{2})m,
\]

which establishes (2.11).
Exercise 2.21: Operations

Solving a $k \times k$ upper triangular system for $k = n, n - 1, \ldots, 1$, takes

$$\sum_{k=1}^{n} k^2 = \frac{1}{3} n(n + 1)(n + 1) \approx \frac{1}{2} G_n$$

arithmetic operations.

Exercise 2.22: Multiplying triangular matrices

Computing the $(i, j)$-th entry of the matrix $AB$ amounts to computing the inner product of the $i$th row $a^T_i$ of $A$ and the $j$th column $b_j$ of $B$. Because of the triangular nature of $A$ and $B$, only the first $i$ entries of $a^T_i$ can be nonzero and only the first $j$ entries of $b_j$ can be nonzero. The computation $a^T_i b_j$ therefore involves $\min\{i, j\}$ multiplications and $\min\{i, j\} - 1$ additions. Carrying out this calculation for all $i$ and $j$, amounts to a total number of

$$\sum_{i=1}^{n} \sum_{j=1}^{n} (2 \min\{i, j\} - 1) = \sum_{i=1}^{n} \left( i(2j - 1) + \sum_{j=i}^{n} (2i - 1) \right)$$

$$= \sum_{i=1}^{n} \left( -i + i(i + 1) + (n - i)(2i - 1) \right) = \sum_{i=1}^{n} (i^2 + 2ni - n + i)$$

$$= -n^2 + (2n + 1) \sum_{i=1}^{n} i - \sum_{i=1}^{n} i^2$$

$$= -n^2 + \frac{1}{2} n(n + 1)(2n + 1) - \frac{1}{6} n(n + 1)(2n + 1)$$

$$= -n^2 + \frac{1}{3} n(n + 1)(2n + 1) = \frac{2}{3} n^3 + \frac{1}{3} n = \frac{1}{3} n(2n^2 + 1)$$

arithmetic operations. A similar calculation gives the same result for the product $BA$.

Exercise 2.30: Making a block LU into an LU (TODO)

Exercise 2.39: Positive definite characterizations

We check the equivalent statements of Theorem 2.38 for the matrix

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

1. Obviously $A$ is symmetric. In addition $A$ is positive definite, because

$$\begin{bmatrix} x \\ y \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 2x^2 - 2xy + 2y^2 = (x - y)^2 + x^2 + y^2 > 0$$

for any nonzero vector $[x, y]^T \in \mathbb{R}^2$.

2. The eigenvalues of $A$ are the roots of the characteristic equation

$$0 = \det(A - \lambda I) = (2 - \lambda)^2 - 1 = (\lambda - 1)(\lambda - 3).$$

Hence the eigenvalues are $\lambda = 1$ and $\lambda = 3$, which are both positive.

3. The leading principal submatrices of $A$ are $[2]$ and $A$ itself, which both have positive determinants.
4. One checks that $A = B B^T$, for the nonsingular matrix

$$B = \begin{bmatrix} \sqrt{2} & 0 \\ -1/\sqrt{2} & \sqrt{3}/2 \end{bmatrix}.$$ 

Exercise 2.61: Using PLU of $A$ to solve $A^T x = b$

If $A = PLR$, then $A^T = R^T L^T P^T$. The matrix $L^T$ is upper triangular and the matrix $R^T$ is lower triangular, implying that $R^T L^T$ is an LU factorization of $A^T P$. Since $A$ is nonsingular, the matrix $R^T$ must be nonsingular, and we can apply Algorithms 2.1 and 2.2 to economically solve the systems $R^T z = b$, $L^T y = z$, and $P^T x = y$, to find a solution $x$ to the system $R^T L^T P^T x = A^T x = b$.

Exercise 2.62: Using PLU to compute the determinant

If $A = PLU$, then

$$\det(A) = \det(PLU) = \det(P) \det(L) \det(U)$$

and the determinant of $A$ can be computed from the determinants of $P$, $L$, and $U$. Since the latter two matrices are triangular, their determinants are simply the products of their diagonal entries. The matrix $P$, on the other hand, is a permutation matrix, so that every row and column is everywhere 0, except for a single entry (where it is 1). Its determinant is therefore quickly computed by cofactor expansion.

Exercise 2.63: Using PLU to compute the inverse (TODO)

Exercise 2.67: Direct proof of Theorem 2.64 (TODO)
CHAPTER 3

The Kronecker Product

Exercise 3.2: $2 \times 2$ Poisson matrix

For $m = 2$, the Poisson matrix $A$ is the $2^2 \times 2^2$ matrix given by

$$
\begin{bmatrix}
4 & -1 & -1 & 0 \\
-1 & 4 & 0 & -1 \\
-1 & 0 & 4 & -1 \\
0 & -1 & -1 & 4 \\
\end{bmatrix}.
$$

In every row $i$, one has $|a_{ii}| = 4 > 2 = |−1| + |−1| + |0| = \sum_{j \neq i} |a_{ij}|$. In other words, $A$ is strictly diagonally dominant.

Exercise 3.5: Properties of Kronecker products

Let be given matrices $A, A_1, A_2 \in \mathbb{R}^{p \times q}$, $B, B_1, B_2 \in \mathbb{R}^{r \times s}$, and $C \in \mathbb{R}^{t \times u}$. Then $(\lambda A) \otimes (\mu B) = \lambda \mu (A \otimes B)$ by definition of the Kronecker product and since

$$(\lambda A) \mu b_{11} \quad (\lambda A) \mu b_{12} \quad \cdots \quad (\lambda A) \mu b_{1s}$$

$$(\lambda A) \mu b_{21} \quad (\lambda A) \mu b_{22} \quad \cdots \quad (\lambda A) \mu b_{2s}$$

$$\vdots$$

$$(\lambda A) \mu b_{r1} \quad (\lambda A) \mu b_{r2} \quad \cdots \quad (\lambda A) \mu b_{rs}$$

$$= \lambda \mu \begin{bmatrix}
A_{b_{11}} & A_{b_{12}} & \cdots & A_{b_{1s}} \\
A_{b_{21}} & A_{b_{22}} & \cdots & A_{b_{2s}} \\
\vdots & \vdots & \ddots & \vdots \\
A_{b_{r1}} & A_{b_{r2}} & \cdots & A_{b_{rs}} \\
\end{bmatrix}.$$

The identity $(A_1 + A_2) \otimes B = (A_1 \otimes B) + (A_2 \otimes B)$ follows from

$$(A_1 + A_2) b_{11} \quad (A_1 + A_2) b_{12} \quad \cdots \quad (A_1 + A_2) b_{1s}$$

$$(A_1 + A_2) b_{21} \quad (A_1 + A_2) b_{22} \quad \cdots \quad (A_1 + A_2) b_{2s}$$

$$\vdots$$

$$(A_1 + A_2) b_{r1} \quad (A_1 + A_2) b_{r2} \quad \cdots \quad (A_1 + A_2) b_{rs}$$

$$= \begin{bmatrix}
A_{1b_{11}} + A_{2b_{11}} & A_{1b_{12}} + A_{2b_{12}} & \cdots & A_{1b_{1s}} + A_{2b_{1s}} \\
A_{1b_{21}} + A_{2b_{21}} & A_{1b_{22}} + A_{2b_{22}} & \cdots & A_{1b_{2s}} + A_{2b_{2s}} \\
\vdots & \vdots & \ddots & \vdots \\
A_{1b_{r1}} + A_{2b_{r1}} & A_{1b_{r2}} + A_{2b_{r2}} & \cdots & A_{1b_{rs}} + A_{2b_{rs}} \\
\end{bmatrix}.$$

A similar argument proves $A \otimes (B_1 + B_2) = (A \otimes B_1) + (A \otimes B_2)$, and therefore the bilinearity of the Kronecker product. The associativity $(A \otimes B) \otimes C = A \otimes (B \otimes C)$
follows from
\[
\begin{bmatrix}
A_{b_{11}} & \cdots & A_{b_{1s}} \\
\vdots & \ddots & \vdots \\
A_{br_{1}} & \cdots & A_{br_{s}}
\end{bmatrix} \otimes C
= \begin{bmatrix}
A_{b_{11}c_{11}} & \cdots & A_{b_{1s}c_{11}} & A_{b_{11}c_{1u}} & \cdots & A_{b_{1s}c_{1u}} \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
A_{br_{1}c_{11}} & \cdots & A_{br_{s}c_{11}} & A_{br_{1}c_{1u}} & \cdots & A_{br_{s}c_{1u}} \\
A_{b_{11}c_{tu}} & \cdots & A_{b_{1s}c_{tu}} & A_{b_{11}c_{tu}} & \cdots & A_{b_{1s}c_{tu}} \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
A_{br_{1}c_{tu}} & \cdots & A_{br_{s}c_{tu}} & A_{br_{1}c_{tu}} & \cdots & A_{br_{s}c_{tu}}
\end{bmatrix}
= \begin{bmatrix}
B_{c_{11}} & \cdots & B_{c_{1u}} \\
\vdots & \ddots & \vdots \\
B_{c_{tu}} & \cdots & B_{c_{tu}}
\end{bmatrix}.
\]

Exercise 3.11: 2nd derivative matrix is positive definite

Applying Lemma 3.8 to the case that \(a = -1\) and \(d = 2\), one finds that the eigenvalues \(\lambda_j\) of the matrix \(\text{tridiag}(-1, 2, -1) \in \mathbb{R}^{m \times m}\) are

\[
\lambda_j = d + 2a \cos \left( \frac{j\pi}{m+1} \right) = 2 \left( 1 - \cos \left( \frac{j\pi}{m+1} \right) \right),
\]

for \(j = 1, \ldots, m\). Moreover, as \(|\cos(x)| < 1\) for any \(x \in (0, \pi)\), it follows that \(\lambda_j > 0\) for \(j = 1, \ldots, m\). Since, in addition, \(\text{tridiag}(-1, 2, -1)\) is symmetric, Lemma 2.41 implies that the matrix \(\text{tridiag}(-1, 2, -1)\) is symmetric positive definite.

Exercise 3.12: 1D test matrix is positive definite?

The statement of this exercise is a generalization of the statement of Exercise 3.11. Consider a matrix \(M = \text{tridiag}(a, d, a) \in \mathbb{R}^{m \times m}\) for which \(d > 0\) and \(d \geq 2|a|\). By Lemma 3.8, the eigenvalues \(\lambda_j\), with \(j = 1, \ldots, m\), of the matrix \(M\) are

\[
\lambda_j = d + 2a \cos \left( \frac{j\pi}{m+1} \right).
\]

If \(a = 0\), then all these eigenvalues are equal to \(d\) and therefore positive. If \(a \neq 0\), write \(\text{sgn}(a)\) for the sign of \(a\). Then

\[
\lambda_j \geq 2|a| \left( 1 + \frac{a}{|a|} \cos \left( \frac{j\pi}{m+1} \right) \right) = 2|a| \left[ 1 + \text{sgn}(a) \cos \left( \frac{j\pi}{m+1} \right) \right] > 0,
\]
again because $|\cos(x)| < 1$ for any $x \in (0, \pi)$. Since, in addition, $M$ is symmetric, Lemma 2.41 implies that $M$ is symmetric positive definite.

**Exercise 3.13: Eigenvalues $2 \times 2$ for 2D test matrix**

One has

$$A\mathbf{x} = \begin{bmatrix} 2d & a & a & 0 \\ a & 2d & 0 & a \\ a & 0 & 2d & a \\ 0 & a & a & 2d \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2d + 2a \\ 2d + 2a \\ 2d + 2a \end{bmatrix} = (2d + 2a) \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \lambda \mathbf{x},$$

which means that $(\lambda, \mathbf{x})$ is an eigenpair of $A$. For $j = k = 1$ and $m = 2$, Theorem 3.10.1 implies that

$$\mathbf{x}_{1,1} = s_1 \otimes s_1 = \begin{bmatrix} \sqrt{3}/2 \\ \sqrt{3}/2 \end{bmatrix} \otimes \begin{bmatrix} \sqrt{3}/2 \\ \sqrt{3}/2 \end{bmatrix} = \begin{bmatrix} 3/4 \\ 3/4 \\ 3/4 \end{bmatrix} \propto \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \mathbf{x}.$$ 

Equation (3.20), on the other hand, implies that

$$\lambda_{1,1} = 2d + 4a \cos \left(\frac{\pi}{3}\right) = 2d + 2a = \lambda.$$ 

We conclude that the eigenpair $(\lambda, \mathbf{x})$ agrees with the eigenpair $(\lambda_{1,1}, \mathbf{x}_{1,1})$.

**Exercise 3.14: Nine point scheme for Poisson problem**

(a) If $m = 2$, the boundary condition yields

$$\begin{bmatrix} v_{00} & v_{01} & v_{02} & v_{03} \\ v_{10} & v_{13} & 0 & 0 \\ v_{20} & v_{23} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

leaving four equations to determine the interior points $v_{11}, v_{12}, v_{21}, v_{22}$. As $6h^2/12 = 1/(2(m + 1)^2) = 1/18$ for $m = 2$, we obtain

$$20v_{11} - 4v_{01} - 4v_{10} - 4v_{21} - 4v_{12} - v_{00} - v_{02} - v_{20} - v_{02} - v_{22} = \frac{1}{18}(8f_{11} + f_{01} + f_{10} + f_{21} + f_{12}),$$

$$20v_{21} - 4v_{11} - 4v_{20} - 4v_{31} - 4v_{22} - v_{10} - v_{30} - v_{12} - v_{32} = \frac{1}{18}(8f_{21} + f_{11} + f_{20} + f_{31} + f_{22}),$$

$$20v_{12} - 4v_{02} - 4v_{11} - 4v_{22} - 4v_{13} - v_{01} - v_{21} - v_{03} - v_{23} = \frac{1}{18}(8f_{12} + f_{02} + f_{11} + f_{22} + f_{13}),$$

$$20v_{22} - 4v_{12} - 4v_{21} - 4v_{32} - 4v_{23} - v_{11} - v_{31} - v_{13} - v_{33} = \frac{1}{18}(8f_{22} + f_{12} + f_{21} + f_{32} + f_{23}),$$

Using the values known from the boundary condition, these equations can be simplified to

$$20v_{11} - 4v_{21} - 4v_{12} - v_{22} = \frac{1}{18}(8f_{11} + f_{01} + f_{10} + f_{21} + f_{12}),$$
20v_{21} - 4v_{11} - 4v_{22} - v_{12} = \frac{1}{18}(8f_{21} + f_{11} + f_{20} + f_{31} + f_{22}),

20v_{12} - 4v_{11} - 4v_{22} - v_{21} = \frac{1}{18}(8f_{12} + f_{02} + f_{11} + f_{22} + f_{13}),

20v_{22} - 4v_{12} - 4v_{21} - v_{11} = \frac{1}{18}(8f_{22} + f_{12} + f_{21} + f_{32} + f_{23}).

Substituting these values in our linear system, we obtain

\begin{equation}
\begin{bmatrix}
  f_{00} & f_{01} & f_{02} & f_{03} \\
  f_{10} & f_{11} & f_{12} & f_{13} \\
  f_{20} & f_{21} & f_{22} & f_{23} \\
  f_{30} & f_{31} & f_{32} & f_{33}
\end{bmatrix}
\begin{bmatrix}
  v_{11} \\
  v_{21} \\
  v_{12} \\
  v_{22}
\end{bmatrix}
= \begin{bmatrix}
  0 \\
  0 \\
  3\pi^2/2 \\
  3\pi^2/2 \\
\end{bmatrix}
\end{equation}

Substituting these values in our linear system, we obtain

\begin{equation}
\begin{bmatrix}
  20 & -4 & -4 & -1 \\
  -4 & 20 & -1 & -4 \\
  -4 & -1 & 20 & -4 \\
  -1 & -4 & -4 & 20
\end{bmatrix}
\begin{bmatrix}
  v_{11} \\
  v_{21} \\
  v_{12} \\
  v_{22}
\end{bmatrix}
= \frac{8 + 1 + 13\pi^2}{18} \begin{bmatrix}
  1 \\
  1 \\
  1 \\
  1
\end{bmatrix}
= \begin{bmatrix}
  5\pi^2/6 \\
  5\pi^2/6 \\
  5\pi^2/6 \\
  5\pi^2/6
\end{bmatrix}
\end{equation}

Solving this system we find that $v_{11} = v_{12} = v_{21} = v_{22} = \frac{5\pi^2}{66}$.

**Exercise 3.15:** Matrix equation for nine point scheme

(a) Let

\[ T = \begin{bmatrix}
  2 & -1 & 0 \\
  -1 & 2 & -1 \\
  0 & \cdots & \cdots & \cdots \\
  -1 & 2 & -1 \\
  0 & -1 & 2
\end{bmatrix}, \quad V = \begin{bmatrix}
  v_{11} & \cdots & v_{1m} \\
  \vdots & \ddots & \vdots \\
  v_{m1} & \cdots & v_{mm}
\end{bmatrix} \]

be of equal dimensions. Implicitly assuming the boundary condition

\( v_{0,k} = v_{m+1,k} = v_{j,0} = v_{j,m+1} = 0, \) for $j, k = 0, \ldots, m+1$, the $(j,k)$-th entry of $TV + VT$ can be written as

\[ 4v_{j,k} - 4v_{j-1,k} - 4v_{j+1,k} - 4v_{j,k-1} - 4v_{j,k+1}. \]

(Compare Equations (3.4) – (3.5).) Similarly, writing out two matrix products, the $(j,k)$-th entry of $TVT = T(VT)$ is found to be

\[ -1(-1v_{j-1,k} + 2v_{j,k} - 1v_{j,k+1}) + v_{j,k-1} - 2v_{j,k} + v_{j,k+1} \]

Together, these observations yield that the System (3.22) is equivalent to (*) and

\[ TV + VT - \frac{1}{6}TVT = h^2\mu F. \]

(b) It is a direct consequence of properties 7 and 8 of Theorem 3.7 that this equation can be rewritten to one of the form $Ax = b$, where

\[ A = T \otimes I + I \otimes T - \frac{1}{6}T \otimes T, \quad x = \text{vec}(V), \quad b = h^2\text{vec}(\mu F). \]
Exercise 3.16: Biharmonic equation

(a) Writing \( v = -\nabla^2 u \), the second line in Equation (3.24) is equivalent to
\[
u(s, t) = v(s, t) = 0, \quad \text{for} \ (s, t) \in \partial \Omega,
\]
while the first line is equivalent to
\[
f(s, t) = \nabla^2 u(s, t) = \nabla^2 (\nabla^2 u(s, t)) = -\nabla^2 v(s, t), \quad \text{for} \ (s, t) \in \Omega.
\]

(b) By property 8 of Theorem 3.7,
\[
(A \otimes I + I \otimes B)\text{vec}(V) = \text{vec}(F) \iff AV + VB^T = F,
\]
whenever \( A \in \mathbb{R}^{r \times r}, B \in \mathbb{R}^{s \times s}, F, V \in \mathbb{R}^{r \times s} \) (the identity matrices are assumed to be of the appropriate dimensions). Using \( T = T^T \), this equation implies
\[
TV + VT = h^2 F \iff (T \otimes I + I \otimes T)\text{vec}(V) = h^2 \text{vec}(F),
\]
\[
TU + UT = h^2 V \iff (T \otimes I + I \otimes T)\text{vec}(U) = h^2 \text{vec}(V).
\]
Substituting the equation for \( \text{vec}(V) \) into the equation for \( \text{vec}(F) \), one obtains the equation
\[
A \text{vec}(U) = h^4 \text{vec}(F),
\]
where \( A := (T \otimes I + I \otimes T)^2 \), which is a linear system of \( m^2 \) equations.

(c) The equations \( h^2 V = (TU + UT) \) and \( TV + VT = h^2 F \) together yield the normal form
\[
T(TU + UT) + (TU + UT)T = T^2 U + 2TUT + UT^2 = h^4 F.
\]
The vector form is given in (b). Using the distributive property of matrix multiplication and the mixed product rule of Lemma 3.6, the matrix \( A = (T \otimes I + I \otimes T)^2 \) can be rewritten as
\[
A = (T \otimes I)(T \otimes I) + (T \otimes I)(I \otimes T) + (I \otimes T)(T \otimes I) + (I \otimes T)(I \otimes T)
\]
\[
= T^2 \otimes I + 2T \otimes T + I \otimes T^2.
\]
Writing \( x := \text{vec}(U) \) and \( b := h^4 \text{vec}(F) \), the linear system of (b) can be written as
\( Ax = b \).

(d) Since \( T \) and \( I \) are symmetric positive definite, property 6 of Theorem 3.7 implies that \( M := T \otimes I + I \otimes T \) is symmetric positive definite as well. The square of any symmetric positive definite matrix is symmetric positive definite as well, implying that \( A = M^2 \) is symmetric positive definite. Let us now show this more directly by calculating the eigenvalues of \( A \).

By Lemma 3.8, we know the eigenpairs \( (\lambda_i, s_i) \), where \( i = 1, \ldots, m \), of the matrix \( T \). By property 5 of Theorem 3.7, it follows that the eigenpairs of \( M \) are \( (\lambda_i + \lambda_j, s_i \otimes s_j) \), for \( i, j = 1, \ldots, m \). If \( B \) is any matrix with eigenpairs \( (\mu_i, v_i) \), where \( i = 1, \ldots, m \), then \( B^2 \) has eigenpairs \( (\mu_i^2, v_i) \), as
\[
B^2 v_i = B(B v_i) = B(\mu_i v_i) = \mu_i (B v_i) = \mu_i^2 v_i, \quad \text{for} \ i = 1, \ldots, m.
\]
It follows that \( A = M^2 \) has eigenpairs \( ((\lambda_i + \lambda_j)^2, s_i \otimes s_j) \), for \( i, j = 1, \ldots, m \). (Note that we can verify this directly by multiplying \( A \) by \( s_i \otimes s_j \) and using the mixed product rule.) Since the \( \lambda_i \) are positive, the eigenvalues of \( A \) are positive. We conclude that \( A \) is symmetric positive definite.
Writing $A = T^2 \otimes I + 2T \otimes T + I \otimes T^2$ and computing the block structure of each of these terms, one finds that $A$ has bandwidth $2m$, in the sense that any row has at most $4m + 1$ nonzero elements.

(e) One can expect to solve the system of (b) faster, as it is typically quicker to solve two simple systems instead of one complex system.
CHAPTER 4

Fast Direct Solution of a Large Linear System

Exercise 4.5: Fourier matrix

The Fourier matrix $F_N$ has entries

$$(F_N)_{j,k} = \omega_{N}^{(j-1)(k-1)}, \quad \omega_N := e^{-\frac{2\pi i}{N}} = \cos\left(\frac{2\pi}{N}\right) - i \sin\left(\frac{2\pi}{N}\right).$$

In particular for $N = 4$, this implies that $\omega_4 = -i$ and

$$F_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{bmatrix}.$$

Computing the transpose and Hermitian transpose gives

$$F_4^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{bmatrix} = F_4, \quad F_4^H = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{bmatrix} \neq F_4,$$

which is what needed to be shown.

Exercise 4.6: Sine transform as Fourier transform

According to Lemma 4.2, the Discrete Sine Transform can be computed from the Discrete Fourier Transform by

$$S_m x_k = i \left(\frac{1}{2}(F_{2m+2}z)_{k+1}\right),$$

where $z = [0, x_1, \ldots, x_m, 0, -x_m, \ldots, -x_1]^T$.

For $m = 1$ this means that $z = [0, x_1, 0, -x_1]^T$ and $S_1 x_1 = \frac{i}{2}(F_4z)_2$.

Since $h = \frac{1}{m+1} = \frac{1}{2}$ for $m = 1$, computing the DST directly gives

$$S_1 x_1 = \sin(\pi h) x_1 = \sin\left(\frac{\pi}{2}\right) x_1 = x_1,$$

while computing the Fourier transform gives

$$F_4z = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{bmatrix} \begin{bmatrix} 0 \\ x_1 \\ 0 \\ -x_1 \end{bmatrix} = \begin{bmatrix} 0 \\ -2ix_1 \\ 0 \\ 2ix_1 \end{bmatrix}.$$

Multiplying the Fourier transform with $\frac{i}{2}$, one finds

$$S_1 x_1 = x_1 = \frac{i}{2}(F_4z)_2,$$
which is what needed to be shown.

**Exercise 4.7: Explicit solution of the discrete Poisson equation**

For any integer \( m \geq 1 \), let \( h = 1/(m+1) \). For \( j = 1, \ldots, m \), let \( \lambda_j = 4 \sin^2 (j \pi h/2) \), \( D = \text{diag}(\lambda_1, \ldots, \lambda_m) \), and \( S = (s_{jk})_{jk} = (\sin(jk\pi h))_{jk} \). By Section 4.2, the solution to the discrete Poisson equation is \( V = SXS \), where \( X \) is found by solving \( DX + XD = 4h^2FS \). Since \( D \) is diagonal, one has

\[
x_{pr} = 4h^4 (\text{SFS})_{pr} = 4h^4 \sum_{k=1}^m \sum_{l=1}^m s_{pk}f_{kl}s_{lr}
\]

\[
\frac{\lambda_p + \lambda_r}{\lambda_p + \lambda_r}
\]

so that

\[
v_{ij} = \sum_{p=1}^m \sum_{r=1}^m s_{ip}x_{pr}s_{rj} = 4h^4 \sum_{p=1}^m \sum_{r=1}^m \sum_{k=1}^m \sum_{l=1}^m s_{ip}s_{pk}s_{lr}s_{rj} \frac{\lambda_p + \lambda_r}{\lambda_p + \lambda_r} f_{kl}
\]

\[
= h^4 \sum_{p=1}^m \sum_{r=1}^m \sum_{k=1}^m \sum_{l=1}^m \sin \left( \frac{ip\pi}{m+1} \right) \sin \left( \frac{pk\pi}{m+1} \right) \sin \left( \frac{lr\pi}{m+1} \right) \sin \left( \frac{rj\pi}{m+1} \right) f_{kl}
\]

which is what needed to be shown.

**Exercise 4.8: Improved version of Algorithm 4.1**

Given is that

\[
(*) \quad TV + VT = h^2F.
\]

Let \( T = SDS^{-1} \) be the orthogonal diagonalization of \( T \) from Equation (4.4), and write \( X = VS \) and \( C = h^2FS \).

(a) Multiplying Equation (\*) from the right by \( S \), one obtains

\[
TX + XD = TVS + VSD = TVS + VT = h^2FS = C.
\]

(b) Writing \( C = [c_1, \ldots, c_m] \), \( X = [x_1, \ldots, x_m] \) and applying the rules of block multiplication, we find

\[
[c_1, \ldots, c_m] = C
= TX + XD
= T[x_1, \ldots, x_m] + X[\lambda_1e_1, \ldots, \lambda_me_m]
= [Tx_1 + \lambda_1xe_1, \ldots, Tx_m + \lambda_me_m]
= [Tx_1 + \lambda_1x_1, \ldots, Tx_m + \lambda_mx_m]
= [(T + \lambda_1I)x_1, \ldots, (T + \lambda_mI)x_m],
\]

which is equivalent to System (4.9). To find \( X \), we therefore need to solve the \( m \) tridiagonal linear systems of (4.9). Since the eigenvalues \( \lambda_1, \ldots, \lambda_m \) are positive, each matrix \( T + \lambda_jI \) is diagonally dominant. By Theorem 1.7, every such matrix is nonsingular and has a unique LU factorization. Algorithms 1.3 and 1.4 then solve the corresponding system \((T + \lambda_jI)x_j = c_j \) in \( O(\delta m) \) operations for some constant \( \delta \). Doing this for all \( m \) columns \( x_1, \ldots, x_m \), one finds the matrix \( X \) in \( O(\delta m^2) \) operations.

(c) To find \( V \), we first find \( C = h^2FS \) by performing \( O(2m^3) \) operations. Next we find \( X \) as in step b) by performing \( O(\delta m^2) \) operations. Finally we compute \( V = 2hXS \) by performing \( O(2m^3) \) operations. In total, this amounts to \( O(4m^3) \) operations.
(d) As explained in Section 4.3, multiplying by the matrix $S$ can be done in $O(2m^2 \log_2 m)$ operations by using the Fourier transform. The two matrix multiplications in c) can therefore be carried out in

$$O(4\gamma m^2 \log_2 m) = O(4\gamma n \log_2 n^{1/2}) = O(2\gamma n \log_2 n)$$

operations.

**Exercise 4.9: Fast solution of 9 point scheme**

Analogously to Section 4.2, we use the relations between the matrices $T, S, X, D$ to rewrite Equation (3.23).

$$TV + VT - \frac{1}{6}TVT = h^2 \mu F$$

$$\iff TSXS + SXST - \frac{1}{6}TSXST = h^2 \mu F$$

$$\iff STXS^2 + S^2XS - \frac{1}{6}STSXS = h^2 \mu SFS$$

$$\iff S^2DXS^2 + S^2XS^2D - \frac{1}{6}S^2DXS^2D = h^2 \mu SFS$$

$$\iff DX + XD - \frac{1}{6}DXD = 4h^4 \mu SFS = 4h^4 G$$

Writing $D = \text{diag}(\lambda_1, \ldots, \lambda_m)$, the $(j,k)$-th entry of $DX + XD - \frac{1}{6}DXD$ is equal to $\lambda_j x_{jk} + x_{jk} \lambda_k - \frac{1}{6} \lambda_j x_{jk} \lambda_k$. Isolating $x_{jk}$ and writing $\lambda_j = 4\sigma_j = 4 \sin^2(j\pi h/2)$ then yields

$$x_{jk} = \frac{4h^4 g_{jk}}{\lambda_j + \lambda_k - \frac{1}{6} \lambda_j \lambda_k} = \frac{h^4 g_{jk}}{\sigma_j + \sigma_k - \frac{2}{3} \sigma_j \sigma_k}, \quad \sigma_j = \sin^2 \left( \frac{j\pi h}{2} \right).$$

Defining $\alpha := j\pi h/2$ and $\beta = k\pi h/2$, one has $0 < \alpha, \beta < \pi/2$. Note that

$$\sigma_j + \sigma_k - \frac{2}{3} \sigma_j \sigma_k > \sigma_j + \sigma_k - \sigma_j \sigma_k$$

$$= 2 - \cos^2 \alpha - \cos^2 \beta - (1 - \cos^2 \alpha)(1 - \cos^2 \beta)$$

$$= 1 - \cos^2 \alpha \cos^2 \beta$$

$$\geq 1 - \cos^2 \beta$$

$$\geq 0.$$

Let $A = T \otimes I + I \otimes T - \frac{1}{6}T \otimes T$ be as in Exercise 3.15.(b) and $s_i$ as in Section 4.2. Applying the mixed-product rule, one obtains

$$A(s_i \otimes s_j) = (T \otimes I + I \otimes T)(s_i \otimes s_j) - \frac{1}{6}(T \otimes T)(s_i \otimes s_j) =$$

$$(\lambda_i + \lambda_j)(s_i \otimes s_j) - \frac{1}{6} \lambda_i \lambda_j (s_i \otimes s_j) = (\lambda_i + \lambda_j - \frac{1}{6} \lambda_i \lambda_j)(s_i \otimes s_j).$$

The matrix $A$ therefore has eigen vectors $s_i \otimes s_j$, and counting them shows that these must be all of them. As shown above, the corresponding eigen values $\lambda_i + \lambda_j - \frac{1}{6} \lambda_i \lambda_j$ are positive, implying that the matrix $A$ is positive definite. It follows that the System (3.22) always has a (unique) solution.
Exercise 4.10: Algorithm for fast solution of 9 point scheme

The following describes an algorithm for solving System (3.22).

**Algorithm 1** A method for solving the discrete Poisson problem (3.22)

**Require:** An integer \( m \) denoting the grid size, a matrix \( \mu F \in \mathbb{R}^{m \times m} \) of function values.

**Ensure:** The solution \( V \) to the discrete Poisson problem (3.22).

1: \( h \leftarrow \frac{1}{m+1} \)
2: \( S \leftarrow (\sin(jk\pi h))_{j,k=1}^{m} \)
3: \( \sigma \leftarrow (\sin^2(\frac{j\pi h}{2}))_{j=1}^{m} \)
4: \( G \leftarrow S \mu F S \)
5: \( X \leftarrow \left(\frac{h^4 u_{jk}}{\sigma_j + \sigma_k - \frac{3}{2} \sigma_j \sigma_k}\right)_{j,k=1}^{m} \)
6: \( V \leftarrow S X S \)

For the individual steps in this algorithm, the time complexities are shown in the following table.

<table>
<thead>
<tr>
<th>step</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>complexity</td>
<td>( O(1) )</td>
<td>( O(m^2) )</td>
<td>( O(m) )</td>
<td>( O(m^3) )</td>
<td>( O(m^2) )</td>
<td>( O(m^3) )</td>
</tr>
</tbody>
</table>

Hence the overall complexity is determined by the four matrix multiplications and given by \( O(m^3) \).

Exercise 4.11: Fast solution of biharmonic equation

From Exercise 3.16 we know that \( T \in \mathbb{R}^{m \times m} \) is the second derivative matrix. According to Lemma 3.8, the eigenpairs \((\lambda_j, s_j)\), with \( j = 1, \ldots, m \), of \( T \) are given by

\[
\begin{align*}
    s_j &= [\sin(j\pi h), \sin(2j\pi h), \ldots, \sin(mj\pi h)]^T, \\
    \lambda_j &= 2 - 2\cos(j\pi h) = 4\sin^2(\frac{j\pi h}{2}),
\end{align*}
\]

and satisfy \( s_j^T s_k = \delta_{j,k}/(2h) \) for all \( j, k \), where \( h := 1/(m+1) \). Using, in order, that \( U = S X S \), \( T S = S D \), and \( S^2 = I/(2h) \), one finds that

\[
\begin{align*}
    h^4 F &= T^2 U + 2 T U T + U T^2 \\
    \iff h^4 F &= T^2 S X S + 2 T S X S T + S X S T^2 \\
    \iff h^4 S F S &= S T^2 S X S^2 + 2 S T S X S T S + S^2 X S T^2 S \\
    \iff h^4 S F S &= S^2 D^2 X S^2 + 2 S^2 D X S D + S^2 X S D^2 \\
    \iff h^4 S F S &= I D^2 X I/(4h^2) + 2 I D X I D/(4h^2) + I X I D^2/(4h^2) \\
    \iff 4 h^6 G &= D^2 X + 2 D X D + X D^2,
\end{align*}
\]

where \( G := S F S \). The \((j,k)\)-th entry of the latter matrix equation is

\[
4 h^6 g_{jk} = \lambda_j^2 x_{jk} + 2 \lambda_j x_{jk} \lambda_k + x_{jk} \lambda_k^2 = x_{jk} (\lambda_j + \lambda_k)^2.
\]

Writing \( \sigma_j := \sin^2(j\pi h/2) = \lambda_j / 4 \), one obtains

\[
x_{jk} = \frac{4 h^6 g_{jk}}{(\lambda_j + \lambda_k)^2} = \frac{4 h^6 g_{jk}}{(4 \sin^2(j\pi h/2) + 4 \sin^2(k\pi h/2))} = \frac{h^6 g_{jk}}{4(\sigma_j + \sigma_k)^2}.
\]
Exercise 4.12: Algorithm for fast solution of biharmonic equation

In order to derive an algorithm that computes $U$ in Problem 3.16, we can adjust Algorithm 4.1 by replacing the computation of the matrix $X$ by the formula from Exercise 4.11. This adjustment does not change the complexity of Algorithm 4.1, which therefore remains $O(\delta n^{3/2})$. The new algorithm can be implemented in Matlab as in Listing 4.1.

Listing 4.1. A simple fast solution to the biharmonic equation

```matlab
function U = simplefastbiharmonic(F)
    m = length(F);
    h = 1/(m+1);
    hv = pi*h*(1:m)';
    sigma = sin(hv/2).^2;
    S = sin(hv*(1:m));
    G = S*F*S;
    X = (h^6)*G./(4*(sigma*ones(1,m)+ones(m,1)*sigma').^2);
    U = zeros(m+2,m+2);
    U(2:m+1,2:m+1) = S*X*S;
end
```

Exercise 4.13: Check algorithm for fast solution of biharmonic equation

The Matlab function from Listing 4.2 directly solves the standard form $Ax = b$ of Equation (3.26), making sure to return a matrix of the same dimension as the implementation from Listing 4.1.

Listing 4.2. A direct solution to the biharmonic equation

```matlab
function V = standardbiharmonic(F)
    m = length(F);
    h = 1/(m+1);
    T = gallery('tridiag', m, -1, 2, -1);
    A = kron(T^2, eye(m)) + 2*kron(T,T) + kron(eye(m),T^2);
    b = h.^4*F(:);
    x = A\b;
    V = zeros(m+2, m+2);
    V(2:m+1,2:m+1) = reshape(x,m,m);
end
```

After specifying $m = 4$ by issuing the command $F = ones(4,4)$, the commands `simplefastbiharmonic(F)` and `standardbiharmonic(F)` both return the matrix

$$
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0.0015 & 0.0024 & 0.0024 & 0.0015 & 0 \\
0 & 0.0024 & 0.0037 & 0.0037 & 0.0024 & 0 \\
0 & 0.0024 & 0.0037 & 0.0037 & 0.0024 & 0 \\
0 & 0.0015 & 0.0024 & 0.0024 & 0.0015 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
$$

For large $m$, it is more insightful to plot the data returned by our Matlab functions. For $m = 50$, we solve and plot our system with the commands in Listing 4.3.
Listing 4.3. Solving the biharmonic equation and plotting the result

1 \begin{verbatim}
\texttt{F = ones(50, 50);
U = simplefastbiharmonic(F);
V = standardbiharmonic(F);
surf(U);
surf(V);
\end{verbatim}

On the face of it, these plots seem to be virtually identical. But exactly how close are they? We investigate this by plotting the difference with the command \texttt{surf(U-V)}, which gives

We conclude that their maximal difference is of the order of $10^{-14}$, which makes them indeed very similar.

Exercise 4.14: Fast solution of biharmonic equation using 9 point rule

(TODO)
CHAPTER 5

Matrix Reduction by Similarity Transformations

Exercise 5.3: Idempotent matrix
Suppose that \((\lambda, x)\) is an eigenpair of a matrix \(A\) satisfying \(A^2 = A\). Then
\[
\lambda x = Ax = A^2 x = \lambda A x = \lambda^2 x.
\]
Since any eigenvector is nonzero, one has \(\lambda = \lambda^2\), from which it follows that either \(\lambda = 0\) or \(\lambda = 1\). We conclude that the eigenvalues of any idempotent matrix can only be zero or one.

Exercise 5.4: Nilpotent matrix
Suppose that \((\lambda, x)\) is an eigenpair of a matrix \(A\) satisfying \(A^k = 0\) for some natural number \(k\). Then
\[
0 = A^k x = \lambda A^{k-1} x = \lambda^2 A^{k-2} x = \cdots = \lambda^k x.
\]
Since any eigenvector is nonzero, one has \(\lambda^k = 0\), from which it follows that \(\lambda = 0\). We conclude that any eigenvalue of a nilpotent matrix is zero.

Exercise 5.5: Eigenvalues of a unitary matrix
Let \(x\) be an eigenvector corresponding to \(\lambda\). Then \(Ax = \lambda x\) and, as a consequence, \(x^* A^* = x^* \lambda\). To use that \(A^* A = I\), it is tempting to multiply the left hand sides of these equations, yielding
\[
|\lambda|^2 \|x\|^2 = x^* A x = x^* A^* Ax = x^* I x = \|x\|^2.
\]
Since \(x\) is an eigenvector, it must be nonzero. Nonzero vectors have nonzero norms, and we can therefore divide the above equation by \(\|x\|^2\), which results in \(|\lambda|^2 = 1\). Taking square roots we find that \(|\lambda| = 1\), which is what needed to be shown. Apparently the eigenvalues of any unitary matrix reside on the unit circle in the complex plane.

Exercise 5.6: Nonsingular approximation of a singular matrix
Let \(\lambda_1, \ldots, \lambda_n\) be the eigenvalues of the matrix \(A\). As the matrix \(A\) is singular, its determinant \(\det(A) = \lambda_1 \cdots \lambda_n\) is zero, implying that one of its eigenvalues is zero. If all the eigenvalues of \(A\) are zero let \(\varepsilon_0 := 1\). Otherwise, let \(\varepsilon_0 := \min_{\lambda_i \neq 0} |\lambda_i|\) be the absolute value of the eigenvalue closest to zero. By definition of the eigenvalues, \(\det(A - \lambda I)\) is zero for \(\lambda = \lambda_1, \ldots, \lambda_n\), and nonzero otherwise. In particular \(\det(A - \varepsilon I)\) is nonzero for any \(\varepsilon \in (0, \varepsilon_0)\), and \(A - \varepsilon I\) will be nonsingular in this interval. This is what we needed to prove.
Exercise 5.7: Companion matrix

(a) To show that \((-1)^n f\) is the characteristic polynomial \(\pi_A\) of the matrix \(A\), we need to compute

\[
\pi_A(\lambda) = \det(A - \lambda I) = \det \begin{bmatrix}
-q_{n-1} - \lambda & -q_{n-2} & \cdots & -q_1 & -q_0 \\
1 & -\lambda & \cdots & 0 & 0 \\
0 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & 1 & -\lambda
\end{bmatrix}.
\]

By the rules of determinant evaluation, we can subtract from any column a linear combination of the other columns without changing the value of the determinant. Multiply columns 1, 2, \ldots, \(n-1\) by \(\lambda^{n-1}, \lambda^{n-2}, \ldots, \lambda\) and adding the corresponding linear combination to the final column, we find

\[
\pi_A(\lambda) = \det \begin{bmatrix}
-q_{n-1} - \lambda & -q_{n-2} & \cdots & -q_1 & -f(\lambda) \\
1 & -\lambda & \cdots & 0 & 0 \\
0 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & 1 & 0
\end{bmatrix} = (-1)^n f(\lambda),
\]

where the second equality follows from cofactor expansion along the final column. Multiplying this equation by \((-1)^n\) yields the statement of the Exercise.

(b) Similar to (a), by multiplying rows 2, 3, \ldots, \(n\) by \(\lambda, \lambda^2, \ldots, \lambda^{n-1}\) and adding the corresponding linear combination to the first row.

Exercise 5.16: Schur decomposition example

The matrix \(U\) is unitary, as \(U^*U = U^TU = I\). One directly verifies that

\[
R := U^TAU = \begin{bmatrix}
-1 & -1 \\
0 & 4
\end{bmatrix}.
\]

Since this matrix is upper triangular, \(A = URU^T\) is a Schur decomposition of \(A\).

Exercise 5.19: Skew-Hermitian matrix

By definition, a matrix \(C\) is skew-Hermitian if \(C^* = -C\).

\(\Rightarrow\): Suppose that \(C = A + iB\), with \(A, B \in \mathbb{R}^{m \times m}\), is skew-Hermitian. Then

\[
-A - iB = -C = C^* = (A + iB)^* = A^T - iB^T,
\]

which implies that \(A^T = -A\) and \(B = B^T\) (use that two complex numbers coincide if and only if their real parts coincide and their imaginary parts coincide). In other words, \(A\) is skew-Hermitian and \(B\) is real symmetric.

\(\Leftarrow\): Suppose that we are given matrices \(A, B \in \mathbb{R}^{m \times m}\) such that \(A\) is skew-Hermitian and \(B\) is real symmetric. Let \(C = A + iB\). Then

\[
C^* = (A + iB)^* = A^T - iB^T = -A - iB = -(A + iB) = -C,
\]

meaning that \(C\) is skew-Hermitian.
Exercise 5.20: Eigenvalues of a skew-Hermitian matrix

Let \( A \) be a skew-Hermitian matrix and consider a Schur triangularization \( A = U R U^* \) of \( A \). Then
\[
R = U^* A U = U^* (-A^*) U = -U^* A^* U = -(U^* A U)^* = -R^*.
\]
Since \( R \) differs from \( A \) by a similar transformation, their eigenvalues coincide (use the multiplicative property of the determinant to show that
\[
\det(A - \lambda I) = \det(U^*) \det(URU^* - \lambda I) \det(U) = \det(R - \lambda I).
\]
As \( R \) is a triangular matrix, its eigenvalues \( \lambda_i \) appear on its diagonal. From the equation \( R = -R^* \) it then follows that \( \lambda_i = -\overline{\lambda_i} \), implying that each \( \lambda_i \) is purely imaginary.

Exercise 5.31: Eigenvalue perturbation for Hermitian matrices

Since a positive semidefinite matrix has no negative eigenvalues, one has \( \beta_n \geq 0 \). It immediately follows from \( \alpha_i + \beta_n \leq \gamma_i \) that in this case \( \gamma_i \geq \alpha_i \).

Exercise 5.33: Hoffman-Wielandt

The matrix \( A \) has eigenvalues 0 and 4, and the matrix \( B \) has eigenvalue 0 with algebraic multiplicity two. Independently of the choice of the permutation \( i_1, \ldots, i_n \), the Hoffman-Wielandt Theorem would yield
\[
16 = \sum_{j=1}^n |\mu_{ij} - \lambda_j|^2 \leq \sum_{i=1}^n \sum_{j=1}^n |a_{ij} - b_{ij}|^2 = 12,
\]
which clearly cannot be valid. The Hoffman-Wielandt Theorem cannot be applied to these matrices, because \( B \) is not normal,
\[
B^H B = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \neq \begin{bmatrix} 2 & -2 \\ -2 & 2 \end{bmatrix} = B B^H.
\]

Exercise 5.42: Find eigenpair example

As \( A \) is a triangular matrix, its eigenvalues correspond to the diagonal entries. One finds two eigenvalues \( \lambda_1 = 1 \) and \( \lambda_2 = 2 \), the latter with algebraic multiplicity two. Solving \( A x_1 = \lambda_1 x_1 \) and \( A x_2 = \lambda_2 x_2 \), one finds (valid choices of) eigenpairs, for instance
\[
(\lambda_1, x_1) = (1, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}), \quad (\lambda_2, x_2) = (2, \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}).
\]

Exercise 5.45: Jordan example

Given matrices
\[
A = \begin{bmatrix} 3 & 0 & 1 \\ -4 & 1 & -2 \\ -4 & 0 & -1 \end{bmatrix}, \quad J = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},
\]
we are asked to find a matrix \( S = [s_1, s_2, s_3] \) satisfying
\[
(\star) \quad [A s_1, A s_2, A s_3] = A S = S J = [s_1, s_1 + s_2, s_3].
\]
Let \( s_1 = (A - I)s_2 \), with \( s_2 \not\in \text{ker}(A - I) \). Then \( s_1, s_2 \neq 0 \), and \( As_2 = s_1 + s_2 \) as required by the second column in (\( \star \)). Moreover, since \( (A - I)^2 = 0 \), the vector \( s_1 \in \text{ker}(A - I) \) is an eigenvector of \( A \) and satisfies the first column in (\( \star \)). Any eigenvector \( s_3 \in \text{ker}(A - I) \) will satisfy the third column in (\( \star \)), and choosing \( s_3 \) to be linearly independent of \( s_1, s_2 \) will guarantee that the matrix \( S \) is invertible. For instance, we could choose

\[
S = \begin{bmatrix} 2 & 1 & 1 \\ -4 & 0 & 0 \\ -4 & 0 & -2 \end{bmatrix}.
\]

Exercise 5.46: Big Jordan example

The matrix \( A \) has Jordan form \( A = SJS^{-1} \), with

\[
J = \begin{bmatrix} 3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}, \quad S = \frac{1}{9} \begin{bmatrix} -14 & 9 & -5 & 6 & 0 & -8 & 9 & 9 \\ -28 & 18 & -10 & 12 & 0 & -7 & 0 & 0 \\ -42 & 27 & -15 & 18 & 0 & -6 & 0 & -9 \\ -56 & 36 & -20 & 24 & 0 & -5 & 0 & 0 \\ -70 & 45 & -16 & 24 & 9 & -4 & 0 & 0 \\ -84 & 54 & -12 & 24 & 9 & 0 & -3 & 0 \\ -98 & 63 & -12 & 24 & 9 & 0 & -2 & 0 \\ -112 & 63 & -12 & 24 & 9 & 0 & -1 & 0 \end{bmatrix}.
\]

Exercise 5.49: Properties of the Jordan form

Let \( J = S^{-1}AS \) be the Jordan form of the matrix \( A \) as in Theorem 5.44. Items 1. – 3. are easily shown by induction, making use of the rules of block multiplication in 2. and 3. For Item 4., write \( E_m := J_m(\lambda) - I_m \), with \( J_m(\lambda) \) the Jordan block of order \( m \).

By the binomial theorem,

\[
J_m(\lambda)^r = (E_m + \lambda I_m)^r = \sum_{k=0}^{r} \binom{r}{k} E_m^k (\lambda I_m)^{r-k} = \sum_{k=0}^{r} \binom{r}{k} \lambda^{r-k} E_m^k.
\]

Since \( E_m^k = 0 \) for any \( k \geq m \), we obtain

\[
J_m(\lambda)^r = \sum_{k=0}^{\min\{r, m-1\}} \binom{r}{k} \lambda^{r-k} E_m^k.
\]

Exercise 5.50: Powers of a Jordan block

Let \( S \) be as in Exercise 5.45. We show by induction on \( n \) that

\[
(*) \quad J^n = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}^n = \begin{bmatrix} 1 & n & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.
\]

Clearly (\( \star \)) holds for \( n = 1 \). Suppose (\( \star \)) holds for some \( n \geq 1 \). Then

\[
J^{n+1} = J^n J = \begin{bmatrix} 1 & n & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & n+1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.
\]
implying that \( (*) \) holds for \( n + 1 \). In particular we find \( J^{100} \). It follows that

\[
A^{100} = (SJS^{-1})^{100} = SJ^{100}S^{-1} = \begin{bmatrix} 2 & 1 & 1 \\ -4 & 0 & 0 \\ -4 & 0 & -2 \end{bmatrix} \begin{bmatrix} 1 & 100 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 2 & 1 & 1 \\ -4 & 0 & 0 \\ -4 & 0 & -2 \end{bmatrix}
\]

\[
= \begin{bmatrix} 2 & 1 & 1 \\ -4 & 0 & 0 \\ -4 & 0 & -2 \end{bmatrix} \begin{bmatrix} 1 & 100 & 0 \\ 0 & -1/4 & 0 \\ 0 & 1/2 & -1/2 \end{bmatrix} = \begin{bmatrix} 201 & 0 & 100 \\ -400 & 1 & -200 \\ -400 & 0 & -199 \end{bmatrix}.
\]

**Exercise 5.52: Minimal polynomial example**

The matrix \( J \) has characteristic polynomial \( \pi_J(\lambda) = \det(J - \lambda I) = (2 - \lambda)^6(3 - \lambda)^2 \) and minimal polynomial \( \mu(\lambda) = (2 - \lambda)^3(3 - \lambda)^2 \).

**Exercise 5.53: Similar matrix polynomials**

For a given polynomial \( p(x) = \sum_k a_k x^k \), substituting \( B = S^{-1}AS \) for \( x \) gives

\[
p(B) = \sum_k a_k (S^{-1}AS)^k = \sum_k a_k S^{-1}A^kS = S^{-1}\sum_k a_k A^kS = S^{-1}p(A)S.
\]

**Exercise 5.54: Minimal polynomial of a diagonalizable matrix**

A matrix \( A \) is diagonalizable precisely when its Jordan form only has Jordan blocks of size one. Writing \( \lambda_1, \ldots, \lambda_k \) for the (distinct) eigenvalues of \( A \), it follows that \( A \) has minimal polynomial \( \mu_A(\lambda) = (\lambda_1 - \lambda) \cdots (\lambda_k - \lambda) \). In particular, for any positive integer \( n \), the identity matrix \( I \in \mathbb{R}^{n \times n} \) has minimal polynomial \( \mu_I(\lambda) = 1 - \lambda \).

**Exercise 5.59: Biorthogonal expansion**

The matrix \( A \) has characteristic polynomial \( \det(A - \lambda I) = (\lambda - 4)(\lambda - 1) \) and right eigenpairs \( (\lambda_1, x_1) = (4, [1, 1]^T) \) and \( (\lambda_2, x_2) = (1, [1, -2]^T) \). Since the right eigenvectors \( x_1, x_2 \) are linearly independent, there exists vectors \( y_1, y_2 \) satisfying \( \langle y_i, x_j \rangle = \delta_{ij} \). The set \( \{x_1, x_2\} \) forms a basis of \( \mathbb{C}^2 \), and the set \( \{y_1, y_2\} \) is called the dual basis.

How do we find such vectors \( y_1, y_2 \)? Any vector \([x_1, x_2]^T\) is orthogonal to the vector \([\alpha x_2, -\alpha x_1]^T\) for any \( \alpha \). Choosing \( \alpha \) appropriately, one finds \( y_1 = \frac{1}{3}[1, -1]^T, y_2 = \frac{1}{3}[2, 1]^T \). By Theorem 5.58, \( y_1 \) and \( y_2 \) are left eigenvectors of \( A \). For any vector \( v = [v_1, v_2]^T \in \mathbb{C}^2 \), Equation (5.20) then gives us the biorthogonal expansions

\[
v = \langle y_1, v \rangle x_1 + \langle y_2, v \rangle x_2 = \frac{1}{3}(v_1 - v_2)x_1 + \frac{1}{3}(2v_1 + v_2)x_2
\]

\[
= \langle x_1, v \rangle y_1 + \langle x_2, v \rangle y_2 = (v_1 + v_2)y_1 + (v_1 - 2v_2)y_2.
\]
Exercise 5.60: Generalized Rayleigh quotient

Suppose \((\lambda, x)\) is a right eigenpair for \(A\), so that \(Ax = \lambda x\). Then the generalized Rayleigh quotient for \(A\) is

\[
R(y, x) := \frac{y^*Ax}{y^*x} = \frac{y^*\lambda x}{y^*x} = \lambda,
\]

which is well defined whenever \(y^*x \neq 0\). On the other hand, if \((\lambda, y)\) is a left eigenpair for \(A\), then \(y^*A = \lambda y^*\) and it follows that

\[
R(y, x) := \frac{y^*Ax}{y^*x} = \frac{\lambda y^*x}{y^*x} = \lambda.
\]
CHAPTER 6

The Singular Value Decomposition

Exercise 6.14: SVD examples

(a) For \( A = [3, 4]^T \) we find a \( 1 \times 1 \) matrix \( A^T A = 25 \), which has the eigenvalue \( \lambda_1 = 25 \). This provides us with the singular value \( \sigma_1 = +\sqrt{\lambda_1} = 5 \) for \( A \). Hence the matrix \( A \) has rank 1 and a SVD of the form

\[
A = [U_1 \quad U_2] \begin{bmatrix} 5 \\ 0 \end{bmatrix} [V_1], \quad \text{with } U_1, U_2 \in \mathbb{R}^{2 \times 1}, \ V = V_1 \in \mathbb{R}.
\]

The eigenvector of \( A^T A \) that corresponds to the eigenvalue \( \lambda_1 = 25 \) is given by \( v_1 = 1 \), providing us with \( V = [1] \). Using Theorem 6.7.3, one finds \( u_1 = \frac{1}{5}[3, 4]^T \). Extending \( u_1 \) to an orthonormal basis for \( \mathbb{R}^2 \) gives \( u_2 = \frac{1}{5}[-4, 3]^T \). A SVD of \( A \) is therefore

\[
A = \frac{1}{5} \begin{bmatrix} 3 & -4 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 5 \\ 0 \end{bmatrix} [1].
\]

(b) One has

\[
A = \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 2 & 2 \end{bmatrix}, \quad A^T = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 2 \end{bmatrix}, \quad A^T A = \begin{bmatrix} 9 & 9 \\ 9 & 9 \end{bmatrix}.
\]

The eigenvalues of \( A^T A \) are the zeros of \( \text{det}(A^T A - \lambda I) = (9 - \lambda)^2 - 81 \), yielding \( \lambda_1 = 18 \) and \( \lambda_2 = 0 \), and therefore \( \sigma_1 = \sqrt{18} \) and \( \sigma_2 = 0 \). Note that since there is only one nonzero singular value, the rank of \( A \) is one. Following the dimensions of \( A \), one finds

\[
\Sigma = \begin{bmatrix} \sqrt{18} & 0 \\ 0 & 0 \end{bmatrix}.
\]

The normalized eigenvectors \( v_1, v_2 \) of \( A^T A \) corresponding to the eigenvalues \( \lambda_1, \lambda_2 \) are the columns of the matrix

\[
V = [v_1 \ v_2] = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}.
\]

Using Theorem 6.7.3 one finds \( u_1 \), which can be extended to an orthonormal basis \( \{u_1, u_2, u_3\} \) using Gram-Schmidt Orthogonalization (see Theorem 0.29). The vectors \( u_1, u_2, u_3 \) constitute a matrix

\[
U = [u_1 \ u_2 \ u_3] = \frac{1}{3} \begin{bmatrix} 1 & -2 & -2 \\ 2 & 2 & -1 \\ 2 & -1 & 2 \end{bmatrix}.
\]
A SVD of \( A \) is therefore given by
\[
A = \frac{1}{3} \begin{bmatrix} 1 & -2 & -2 \\ 2 & 2 & -1 \\ 2 & -1 & 2 \end{bmatrix} \begin{bmatrix} \sqrt{18} & 0 \\ 0 & 0 \\ 0 & \sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.
\]

Exercise 6.15: More SVD examples

(a) We have \( A = e_1 \) and \( A^T A = e_1^T e_1 = [1] \). This gives the eigenpair \( (\lambda_1, v_1) = (1, 1) \) of \( A^T A \). Hence \( \sigma_1 = 1 \) and \( \Sigma = e_1 = A \). As \( \Sigma = A \) and \( V = I \), we must have \( U = I_m \) yielding a singular value decomposition
\[
A = I_m e_1 I_1.
\]

(b) For \( A = e_n^T \), the matrix
\[
A^T A = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}
\]
has eigenpairs \((0, e_j)\) for \( j = 1, \ldots, n-1 \) and \((1, e_n)\). Then \( \Sigma = e_1^T \in \mathbb{R}^{1,n} \) and \( V = [e_n, e_{n-1}, \ldots, e_1] \in \mathbb{R}^{n,n} \). Using Theorem 6.7.3 we get \( u_1 = 1 \), yielding \( U = [1] \). A SVD for \( A \) is therefore given by
\[
A = e_n^T = [1] e_1^T [e_n, e_{n-1}, \ldots, e_1].
\]

(c) In this exercise
\[
A = \begin{bmatrix} -1 & 0 \\ 0 & 3 \end{bmatrix}, \quad A^T = A, \quad A^T A = \begin{bmatrix} 1 & 0 \\ 0 & 9 \end{bmatrix}.
\]
The eigenpairs of \( A^T A \) are given by \( (\lambda_1, v_1) = (9, e_2) \) and \( (\lambda_2, v_2) = (1, e_1) \), from which we find
\[
\Sigma = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}, \quad V = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.
\]
Using Theorem 6.7.3, one finds \( u_1 = e_2 \) and \( u_2 = -e_1 \), which constitute the matrix
\[
U = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.
\]
A SVD of \( A \) is therefore given by
\[
A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.
\]

Exercise 6.17: Counting dimensions of fundamental subspaces

Let \( A \) have singular value decomposition \( U \Sigma V^* \).

1. By items 1. and 3. of Theorem 6.16, \( \text{span}(A) \) and \( \text{span}(A^*) \) are vector spaces of the same dimension \( r \), implying that \( \text{rank}(A) = \text{rank}(A^*) \).

2. This statement is known as the rank-nullity theorem, and it follows immediately from combining items 1. and 4. in Theorem 6.16.

3. As \( \text{rank}(A^*) = \text{rank}(A) \) by 1., this follows by replacing \( A \) by \( A^* \) in 2.
Exercise 6.18: Rank and nullity relations

Let \( A = U \Sigma V^* \) be a singular value decomposition of a matrix \( A \in \mathbb{C}^{m \times n} \).

1. By Theorem 6.5.5, \( \text{rank}(A) \) is the number of positive eigenvalues of

\[
AA^* = U \Sigma V^* V \Sigma^* U^* = U D U^*,
\]

where \( D := \Sigma \Sigma^* \) is a diagonal matrix with real nonnegative elements. Since \( U D U^* \) is an orthogonal diagonalization of \( AA^* \), the number of positive eigenvalues of \( AA^* \) is the number of nonzero diagonal elements in \( D \). Moreover, \( \text{rank}(AA^*) \) is the number of positive eigenvalues of

\[
AA^*(AA^*)^* = AA^*AA^* = U \Sigma \Sigma^* \Sigma \Sigma^* V^* = UD^2 U^*,
\]

which is the number of nonzero diagonal elements in \( D^2 \), so that \( \text{rank}(A) = \text{rank}(AA^*) \). From a similar argument for \( \text{rank}(A^* A) \), we conclude that

\[
\text{rank}(A) = \text{rank}(AA^*) = \text{rank}(A^* A).
\]

2. Let \( r := \text{rank}(A) = \text{rank}(A^*) = \text{rank}(AA^*) = \text{rank}(A^* A) \). Applying Theorem 6.5, parts 3 and 4, to the singular value decompositions

\[
A = U \Sigma V^*, \quad A^* = V \Sigma U^*, \quad AA^* = U \Sigma \Sigma^* U^*, \quad A^* A = V \Sigma^* \Sigma V^*,
\]

one finds that \( \{v_{r+1}, \ldots, v_n\} \) is a basis for both \( \ker(A) \) and \( \ker(A^* A) \), while \( \{u_{r+1}, \ldots, u_m\} \) is a basis for both \( \ker(A^*) \) and \( \ker(AA^*) \). In particular it follows that

\[
\dim \ker(A) = \dim \ker(A^* A), \quad \dim \ker(A^*) = \dim \ker(AA^*),
\]

which is what needed to be shown.

Exercise 6.19: Orthonormal bases example

Given is the matrix

\[
A = \frac{1}{15} \begin{bmatrix}
14 & 4 & 16 \\
2 & 22 & 13
\end{bmatrix}.
\]

From Example 6.11 we know that \( B = A^T \) and hence \( A = U \Sigma V^T \) and \( B = V \Sigma^T U^T \), with

\[
V = \frac{1}{3} \begin{bmatrix}
1 & 2 \\
2 & -2 \\
2 & 1
\end{bmatrix}, \quad \Sigma = \begin{bmatrix}
2 & 0 \\
0 & 1 \\
0 & 0
\end{bmatrix}, \quad U = \frac{1}{5} \begin{bmatrix}
3 & 4 \\
4 & -3
\end{bmatrix}.
\]

From Theorem 6.16 we know that \( V_1 \) forms an orthonormal basis for \( \text{span}(A^T) = \text{span}(B) \), \( V_2 \) an orthonormal basis for \( \ker(A) \) and \( U_2 \) an orthonormal basis for \( \ker(A^T) = \ker(B) \). Hence

\[
\text{span}(B) = \alpha v_1 + \beta v_2, \quad \ker(A) = \gamma v_3 \quad \text{and} \quad \ker(B) = 0.
\]
Exercise 6.20: Some spanning sets

The matrices $A \in \mathbb{C}^{m \times n}$ and $A^*A$ have the same rank $r$ since they have the same number of singular values, so that the vector spaces $\text{span}(A^*A)$ and $\text{span}(A^*)$ have the same dimension. It is immediate from the definition that $\text{span}(A^*A) \subset \text{span}(A^*)$, and therefore $\text{span}(A^*A) = \text{span}(A^*)$.

Let $A = U_1 \Sigma_1 V_1^*$ be a singular value factorization of $A$. Taking the Hermitian transpose $A^* = V_1 \Sigma_1^* U_1^*$ one finds $\text{span}(A^*) \subset \text{span}(V_1)$. Moreover, since $V_1 \in \mathbb{C}^{n \times r}$ has orthonormal columns, it has the same rank as $A^*$, and we conclude $\text{span}(A^*) = \text{span}(V_1)$.

Exercise 6.21: Singular values and eigenpair of composite matrix

Given is a singular value decomposition $A = U \Sigma V^*$. Let $r = \text{rank}(A)$, so that $\sigma_1 \geq \cdots \geq \sigma_r > 0$ and $\sigma_{r+1} = \cdots = \sigma_n = 0$. Let $U = [U_1, U_2]$ and $V = [V_1, V_2]$ be partitioned accordingly and $\Sigma_1 = \text{diag}(\sigma_1, \ldots, \sigma_r)$ as in Equation (6.7), so that $A = U_1 \Sigma_1 V_1^*$ forms a singular value factorization of $A$.

By Theorem 6.16,

$$C_{pi} = \begin{bmatrix} 0 & A^* & 0 \\ \end{bmatrix} \begin{bmatrix} u_i \\ v_i \\ \end{bmatrix} = \begin{bmatrix} A \sigma_i p_i \\ A^* u_i \\ 0 \cdot p_i \\ \end{bmatrix} = \begin{cases} \sigma_i p_i & \text{for } i = 1, \ldots, r \\ 0 \cdot p_i & \text{for } i = r + 1, \ldots, n \end{cases}$$

$$C_{qi} = \begin{bmatrix} 0 & A^* & 0 \\ \end{bmatrix} \begin{bmatrix} u_i \\ -v_i \\ \end{bmatrix} = \begin{bmatrix} -A \sigma_i v_i \\ A^* u_i \\ 0 \cdot q_i \\ \end{bmatrix} = \begin{cases} -\sigma_i q_i & \text{for } i = 1, \ldots, r \\ 0 \cdot q_i & \text{for } i = r + 1, \ldots, n \end{cases}$$

$$C_{rj} = \begin{bmatrix} 0 & A^* & 0 \\ \end{bmatrix} \begin{bmatrix} u_j \\ 0 \\ \end{bmatrix} = \begin{bmatrix} 0 \\ A^* u_j \\ 0 \cdot r_j \\ \end{bmatrix} = 0 \cdot r_j, \text{ for } j = n + 1, \ldots, m.$$ 

This gives a total of $n + n + (m - n) = m + n$ eigen pairs.

Exercise 6.27: Rank example

We are given the singular value decomposition

$$A = U \Sigma V^T = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ \end{bmatrix} \begin{bmatrix} 6 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 6 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & -1 & -1 & -1 \\ -1 & 2 & -1 & -1 \\ -1 & -1 & 2 & -1 \\ -1 & -1 & -1 & 2 \end{bmatrix}.$$ 

Write $U = [u_1, u_2, u_3, u_4]$ and $V = [v_1, v_2, v_3]$. Clearly $r = \text{rank}(A) = 2$.

(a) A direct application of Theorem 6.16 with $r = 2$ gives

- $\{u_1, u_2\}$ is an orthonormal basis for $\text{span}(A)$,
- $\{u_3, u_4\}$ is an orthonormal basis for $\text{ker}(A^T)$,
- $\{v_1, v_2\}$ is an orthonormal basis for $\text{span}(A^T)$,
- $\{v_3, v_4\}$ is an orthonormal basis for $\text{ker}(A)$.

Since $U$ is orthogonal, $\{u_1, u_2, u_3, u_4\}$ is an orthonormal basis for $\mathbb{R}^4$. In particular $u_3, u_4$ are orthogonal to $u_1, u_2$, so that they span the orthogonal complement $\text{span}(A)^\perp$ to $\text{span}(A) = \text{span}\{u_1, u_2\}$.

(b) Applying Theorem 6.26 with $r = 1$ yields

$$\|A - B\|_F \geq \sqrt{\sigma_2^2 + \sigma_3^2} = \sqrt{6^2 + 0^2} = 6.$$
(c) Following Section 6.4.2, with \( D' := \text{diag}(\sigma_1, 0, \ldots, 0) \in \mathbb{R}^{n,n} \), take
\[
A_1 = A' := U \begin{bmatrix} D' & 0 \\ 0 & \end{bmatrix} V^T = \begin{bmatrix} 2 & 2 & 1 \\ 2 & 2 & 1 \\ 2 & 2 & 1 \end{bmatrix}.
\]

**Exercise 6.28: Another rank example**

(a) The matrix \( B = (b_{ij})_{ij} \in \mathbb{R}^{n,n} \) is defined by
\[
b_{ij} = \begin{cases} 
1 & \text{if } i = j; \\
-1 & \text{if } i < j; \\
-2^{2-n} & \text{if } (i,j) = (n,1); \\
0 & \text{otherwise.}
\end{cases}
\]
while the column vector \( x = (x_j)_j \in \mathbb{R}^n \) is given by
\[
x_j = \begin{cases} 
1 & \text{if } j = n; \\
2^{n-1-j} & \text{otherwise.}
\end{cases}
\]
For the final entry in the matrix product \( Bx \) one finds that
\[
(Bx)_n = \sum_{j=1}^{n} b_{nj}x_j = b_{n1}x_1 + b_{nn}x_n = -2^{2-n} \cdot 2^{n-2} + 1 \cdot 1 = 0.
\]
For any of the remaining indices \( i \neq n \), the \( i \)-th entry of the matrix product \( Bx \) can be expressed as
\[
(Bx)_i = \sum_{j=1}^{n} b_{ij}x_j = b_{im} + \sum_{j=1}^{n-1} 2^{n-1-j}b_{ij}
\]
\[
= -1 + 2^{n-1-i}b_{ii} + \sum_{j=i+1}^{n-1} 2^{n-1-j}b_{ij}
\]
\[
= -1 + 2^{n-1-i} - \sum_{j=i+1}^{n-1} 2^{n-1-j}
\]
\[
= -1 + 2^{n-1-i} - 2^{n-2-i} \sum_{j'=0}^{n-2-i} \left( \frac{1}{2} \right)^{j'}
\]
\[
= -1 + 2^{n-1-i} - 2^{n-2-i} \frac{1 - \left( \frac{1}{2} \right)^{n-1-i}}{1 - \frac{1}{2}}
\]
\[
= -1 + 2^{n-1-i} - 2^{n-1-i} \left( 1 - 2^{-(n-1-i)} \right)
\]
\[
= 0.
\]
As \( B \) has a nonzero kernel, it must be singular. The matrix \( A \), on the other hand, is nonsingular, as its determinant is \((-1)^n \neq 0\). The matrices \( A \) and \( B \) differ only in their \((n,1)\)-th entry, so one has \( \|A - B\|_F = \sqrt{|a_{n1} - b_{n1}|^2} = 2^{2-n} \). In other words, the tiniest perturbation can make a matrix with large determinant singular.
(b) Let $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$ be the singular values of $A$. Applying Theorem 6.26 for $r = \text{rank}(B) < n$, we obtain

$$\sigma_n \leq \sigma_n \sqrt{\left(\frac{\sigma_{r+1}}{\sigma_n}\right)^2 + \cdots + \left(\frac{\sigma_{n-1}}{\sigma_n}\right)^2} + 1 = \sqrt{\sigma_{r+1}^2 + \cdots + \sigma_n^2}$$

$$= \min_{\substack{C \in \mathbb{R}^{n,n} \\text{rank}(C) = r}} \|A - C\|_F \leq \|A - B\|_F = 2^{2-n}.$$ 

We conclude that the smallest singular value $\sigma_n$ can be at most $2^{2-n}$. 

39
CHAPTER 7

Matrix Norms

Exercise 7.4: Consistency of sum norm?

Observe that the sum norm is a matrix norm. This follows since it is equal to the $l_1$-norm of the vector $v = \text{vec}(A)$ obtained by stacking the columns of a matrix $A$ on top of each other.

Let $A = (a_{ij})_{ij}$ and $B = (b_{ij})_{ij}$ be matrices for which the product $AB$ is defined. Then

$$
\|AB\|_S = \sum_{i,j} \left| \sum_k a_{ik} b_{kj} \right| \leq \sum_{i,j,k} |a_{ik}| \cdot |b_{kj}| \leq \sum_{i,j,k,l} |a_{ik}| \cdot |b_{lj}| = \|A\|_S \|B\|_S,
$$

where the first inequality follows from the triangle inequality and multiplicative property of the absolute value $|\cdot|$. Since $A$ and $B$ were arbitrary, this proves that the sum norm is consistent.

Exercise 7.5: Consistency of max norm?

Observe that the max norm is a matrix norm. This follows since it is equal to the $l_\infty$-norm of the vector $v = \text{vec}(A)$ obtained by stacking the columns of a matrix $A$ on top of each other.

To show that the max norm is not consistent we use a counter example. Let $A = B = (1)_{i,j=1}^2$. Then

$$
\begin{vmatrix} 1 & 1 \\ 1 & 1 \end{vmatrix}^M = \begin{vmatrix} 1 & 1 \\ 1 & 1 \end{vmatrix}^M = 2 > 1 = \begin{vmatrix} 1 & 1 \\ 1 & 1 \end{vmatrix}^M \begin{vmatrix} 1 & 1 \\ 1 & 1 \end{vmatrix}^M,
$$

contradicting $\|AB\|_M \leq \|A\|_M \|B\|_M$.

Exercise 7.6: Consistency of modified max norm?

Exercise 7.5 shows that the max norm is not consistent. In this Exercise we show that the max norm can be modified so as to define a consistent matrix norm.

(a) Let $A \in \mathbb{C}^{m,n}$ and define $\|A\| := \sqrt{mn}\|A\|_M$ as in the Exercise. To show that $\|\cdot\|$ defines a consistent matrix norm we have to show that it fulfills the three matrix norm properties and that it is submultiplicative. Let $A, B \in \mathbb{C}^{m,n}$ be any matrices and $\alpha$ any scalar.

**Positivity:** Clearly $\|A\| = \sqrt{mn}\|A\|_M \geq 0$. Moreover,$$
\|A\| = 0 \iff a_{i,j} = 0 \ \forall i,j \iff A = 0.$$

**Homogeneity:** $\|\alpha A\| = \sqrt{mn}\|\alpha A\|_M = |\alpha|\sqrt{mn}\|A\|_M = |\alpha|\|A\|$. 

40
Subadditivity: One has
\[ \|A + B\| = \sqrt{nm}\|A + B\|_M \leq \sqrt{nm}\left(\|A\|_M + \|B\|_M\right) = \|A\| + \|B\|. \]

Submultiplicativity: One has
\[ \|AB\| = \sqrt{mn} \max_{1 \leq i \leq m, 1 \leq j \leq n} \left| \sum_{k=1}^{q} a_{i,k} b_{k,j} \right| \leq \sqrt{mn} \max_{1 \leq i \leq m, 1 \leq j \leq n} \left( \max_{1 \leq k \leq q} |b_{k,j}| \sum_{k=1}^{q} |a_{i,k}| \right) \leq q\sqrt{mn} \left( \max_{1 \leq i \leq m, 1 \leq k \leq q} |a_{i,k}| \right) \left( \max_{1 \leq k \leq q, 1 \leq j \leq n} |b_{k,j}| \right) = \|A\|_1 \|B\|_1. \]

(b) For any \( A \in \mathbb{C}^{m,n} \), let
\[ \|A\|^{(1)} := m\|A\|_M \quad \text{and} \quad \|A\|^{(2)} := n\|A\|_M. \]

Comparing with the solution of part (a) we see, that the points of positivity, homo-
ogeneity and subadditivity are fulfilled here as well, making \( \|A\|^{(1)} \) and \( \|A\|^{(2)} \) valid
matrix norms. Furthermore, for any \( A \in \mathbb{C}^{m,q}, B \in \mathbb{C}^{q,n} \),
\[ \|AB\|^{(1)} = m \max_{1 \leq i \leq m, 1 \leq j \leq n} \left| \sum_{k=1}^{q} a_{i,k} b_{k,j} \right| \leq m \left( \max_{1 \leq i \leq m, 1 \leq k \leq q} |a_{i,k}| \right) q \left( \max_{1 \leq k \leq q, 1 \leq j \leq n} |b_{k,j}| \right) = \|A\|^{(1)} \|B\|^{(1)}, \]
\[ \|AB\|^{(2)} = n \max_{1 \leq i \leq m, 1 \leq j \leq n} \left| \sum_{k=1}^{q} a_{i,k} b_{k,j} \right| \leq q \left( \max_{1 \leq i \leq m, 1 \leq k \leq q} |a_{i,k}| \right) n \left( \max_{1 \leq k \leq q, 1 \leq j \leq n} |b_{k,j}| \right) = \|A\|^{(2)} \|B\|^{(2)}, \]
which proves the submultiplicativity of both norms.

Exercise 7.8: The sum norm is subordinate to?

For any matrix \( A = (a_{ij})_{ij} \in \mathbb{C}^{m,n} \) and column vector \( x = (x_j)_{j} \in \mathbb{C}^{n} \), one has
\[ \|Ax\|_1 = \sum_{i=1}^{m} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \leq \sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}| |x_j| \leq \sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}| \sum_{k=1}^{n} |x_k| = \|A\|_S \|x\|_1, \]
which shows that the matrix norm \( \| \cdot \|_S \) is subordinate to the vector norm \( \| \cdot \|_1 \).
Exercise 7.9: The max norm is subordinate to?

Let $A = (a_{ij})_{ij} \in \mathbb{C}^{m,n}$ be a matrix and $x = (x_j)_j \in \mathbb{C}^n$ a column vector.

(a) One has
\[
\|Ax\|_\infty = \max_{i=1,\ldots,m} \left| \sum_{j=1}^n a_{ij} x_j \right| \leq \max_{i=1,\ldots,m} \left| \sum_{j=1}^n a_{ij} \right| \cdot \left| \sum_{j=1}^n x_j \right| = \|A\|_M \cdot \|x\|_1.
\]

(b) Assume that the maximum in the definition of $\|A\|_M$ is attained in column $l$, implying that $\|A\|_M = |a_{k,l}|$ for some $k$. Let $e_l$ be the $l$th standard basis vector. Then
\[
\|e_l\|_1 = 1 \quad \text{and} \quad \|Ae_l\|_\infty = \max_{i=1,\ldots,m} |a_{i,l}| = |a_{k,l}| \cdot 1 = \|A\|_M \cdot \|e_l\|_1,
\]
which is what needed to be shown.

(c) By (a), $\|A\|_M \geq \|Ax\|_\infty / \|x\|_1$ for all nonzero vectors $x$, implying that $\|A\|_M = \|A\|_M \cdot \|x\|_1 = \|A\|_M \cdot \|e_l\|_1$.

By (b), equality is attained for any standard basis vector $e_l$ for which there exists a $k$ such that $\|A\|_M = |a_{k,l}|$. We conclude that
\[
\|A\|_M = \max_{x \neq 0} \frac{\|Ax\|_\infty}{\|x\|_1},
\]
which means that $\|\cdot\|_M$ is the $(\infty,1)$-operator norm (see Definition 7.10).

Exercise 7.16: Spectral norm

Let $A = U\Sigma V^*$ be a singular value decomposition of $A$, and write $\sigma_1 := \|A\|_2$ for the biggest singular value of $A$. Since the orthogonal matrices $U$ and $V$ leave the Euclidean norm invariant,
\[
\frac{\|y^*Ax\|_2}{\|x\|_2} = \frac{\max_{\|x\|_2=1} \|y^*U\Sigma V^*x\|_2}{\|x\|_2} = \max_{\|x\|_2=1} |y^*\Sigma x| \leq \sigma_1 \frac{\|y\|_2}{\|x\|_2} \frac{\|x\|_2}{\|y\|_2} = \sigma_1.
\]
Moreover, this maximum is achieved for $x = y = e_1$, and we conclude
\[
\|A\|_2 = \sigma_1 = \max_{\|x\|_2=1} |y^*Ax|.
\]

Exercise 7.17: Spectral norm of the inverse

Let $\sigma_1 \geq \cdots \geq \sigma_n$ be the singular values of $A$. Since $A$ is nonsingular, $\sigma_n$ must be nonzero. Using Equations (7.10) and (6.16), we find
\[
\|A^{-1}\|_2 = \frac{1}{\sigma_n} = \frac{1}{\min_{0 \neq x \in \mathbb{C}^n} \frac{\|Ax\|_2}{\|x\|_2}} = \max_{0 \neq x \in \mathbb{C}^n} \frac{\|x\|_2}{\|Ax\|_2},
\]
which is what needed to be shown.

42
Exercise 7.18: $p$-norm example

We have

\[ A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \quad A^{-1} = \frac{1}{3} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}. \]

Using Theorem 7.12, one finds $\|A\|_1 = \|A\|_\infty = 3$ and $\|A^{-1}\|_1 = \|A^{-1}\|_\infty = 1$. The singular values $\sigma_1 \geq \sigma_2$ of $A$ are the square roots of the zeros of $0 = \det(A^T A - \lambda I) = (5 - \lambda)^2 - 16 = \lambda^2 - 10\lambda + 9 = (\lambda - 9)(\lambda - 1)$.

Using Theorem 7.14, we find $\|A\|_2 = \sigma_1 = 3$ and $\|A^{-1}\|_2 = \sigma_2^{-1} = 1$. Alternatively, since $A$ is symmetric positive definite, we know from (7.11) that $\|A\|_2 = \lambda_1$ and $\|A^{-1}\|_2 = 1/\lambda_2$, where $\lambda_1 = 3$ is the biggest eigenvalue of $A$ and $\lambda_2 = 1$ is the smallest.

Exercise 7.21: Unitary invariance of the spectral norm

Suppose $V$ is a rectangular matrix satisfying $V^* V = I$. Then

\[ \|VA\|_2^2 = \max_{\|x\|_2 = 1} \|VAx\|_2^2 = \max_{\|x\|_2 = 1} x^* A^* V^* V Ax = \max_{\|x\|_2 = 1} x^* A^* Ax = \max_{\|x\|_2 = 1} \|Ax\|_2^2 = \|A\|_2^2. \]

The result follows by taking square roots.

Exercise 7.22: $\|A u\|_2$ rectangular $A$

Let $u = [u_1, u_2]^T$ be any column vector satisfying $1 = u^T u = \|u\|_2^2$. Then $A u$, considered as a matrix, has operator 2-norm

\[ \max_{\|x\|_1 = 1} \|A u x\|_2 = \max \{ \|A u\|_2, \| - A u\|_2 \} = \|A u\|_2 \]

equal to its Euclidean norm, considered as a vector. In order for $\|A u\|_2 < \|A\|_2$ to hold, there needs to exist another unit vector $v$ for which one has the inequality $\|A u\|_2 < \|A v\|_2$ of Euclidean norms. In other words, we need to pick a matrix $A$ that scales more in the direction $v$ than in the direction $u$. For instance, if

\[ A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \quad u = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad v = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \]

then

\[ \|A\|_2 = \max_{\|x\|_2 = 1} \|Ax\|_2 \geq \|Av\|_2 = 2 > 1 = \|Au\|_2. \]

Exercise 7.23: $p$-norm of diagonal matrix

The eigenpairs of the matrix $A = \text{diag}(\lambda_1, \ldots, \lambda_n)$ are $(\lambda_1, e_1), \ldots, (\lambda_n, e_n)$. For $\rho(A) = \max\{\|\lambda_1\|, \ldots, \|\lambda_n\|\}$, one has

\[ \|A\|_p = \max_{(x_1, \ldots, x_n) \neq 0} \frac{(\|\lambda_1\| x_1^p + \cdots + \|\lambda_n\| x_n^p)^{1/p}}{(\|x_1^p + \cdots + x_n^p\|^{1/p})} \leq \max_{(x_1, \ldots, x_n) \neq 0} \frac{(\rho(A)^p |x_1^p + \cdots + \rho(A)^p |x_n^p|^{1/p})^{1/p}}{(\|x_1^p + \cdots + x_n^p\|^{1/p})} = \rho(A). \]
On the other hand, for \( e_j \) such that \( \rho(A) = |\lambda_j| \), one finds
\[
\|A\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} \geq \frac{\|Ae_j\|_p}{\|e_j\|_p} = \rho(A).
\]
Together, the above two statements imply that \( \|A\|_p = \rho(A) \) for any diagonal matrix \( A \) and any \( p \) satisfying \( 1 \leq p \leq \infty \).

**Exercise 7.24:** Spectral norm of a column vector

We write \( A \in \mathbb{C}^{m \times 1} \) for the matrix corresponding to the column vector \( a \in \mathbb{C}^m \). Write \( \|A\|_p \) for the operator \( p \)-norm of \( A \) and \( \|a\|_p \) for the vector \( p \)-norm of \( a \). In particular \( \|A\|_2 \) is the spectral norm of \( A \) and \( \|a\|_2 \) is the Euclidean norm of \( a \). Then
\[
\|A\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{|x|} = \max_{x \neq 0} \frac{|x||a|_p}{|x|} = \|a\|_p,
\]
proving (b). Note that (a) follows as the special case \( p = 2 \).

**Exercise 7.25:** Norm of absolute value matrix

(a) One finds
\[
|A| = \begin{bmatrix} |1 + i| & | -2i| \\ |1| & |1 - i| \end{bmatrix} = \begin{bmatrix} \sqrt{2} & 2 \\ 1 & \sqrt{2} \end{bmatrix}.
\]

(b) Let \( b_{i,j} \) denote the entries of \( |A| \). Observe that \( b_{i,j} = |a_{i,j}| = |b_{i,j}| \). Together with Theorem 7.12, these relations yield
\[
\|A\|_F = \left( \sum_{i=1}^m \sum_{j=1}^n |a_{i,j}|^2 \right)^{\frac{1}{2}} = \left( \sum_{i=1}^m \sum_{j=1}^n |b_{i,j}|^2 \right)^{\frac{1}{2}} = \|A\|_F,
\]
\[
\|A\|_1 = \max_{1 \leq j \leq n} \left( \sum_{i=1}^m |a_{i,j}| \right) = \max_{1 \leq j \leq n} \left( \sum_{i=1}^m |b_{i,j}| \right) = \|A\|_1,
\]
\[
\|A\|_\infty = \max_{1 \leq i \leq m} \left( \sum_{j=1}^n |a_{i,j}| \right) = \max_{1 \leq i \leq m} \left( \sum_{j=1}^n |b_{i,j}| \right) = \|A\|_\infty,
\]
which is what needed to be shown.

(c) To show this relation between the 2-norms of \( A \) and \( |A| \), we first examine the connection between the \( l_2 \)-norms of \( Ax \) and \( |A| \cdot |x| \), where \( x = (x_1, \ldots, x_n) \) and \( |x| = (|x_1|, \ldots, |x_n|) \). We find
\[
\|Ax\|_2 = \left( \sum_{i=1}^m \sum_{j=1}^n |a_{i,j}x_j|^2 \right)^{\frac{1}{2}} \leq \left( \sum_{i=1}^m \left( \sum_{j=1}^n |a_{i,j}| |x_j| \right)^2 \right)^{\frac{1}{2}} = \|A\| \cdot |x|_2.
\]

Now let \( x^* \) with \( \|x^*\|_2 = 1 \) be a vector for which \( \|A\|_2 = \|Ax^*\|_2 \). That is, let \( x^* \) be a unit vector for which the maximum in the definition of 2-norm is attained. Observe that \( |x^*| \) is then a unit vector as well, \( \|x^*\|_2 = 1 \). Then, by the above estimate of \( l_2 \)-norms and definition of the 2-norm,
\[
\|A\|_2 = \|Ax^*\|_2 \leq \|A\| \cdot |x^*|_2 \leq \|A\|_2.
\]
(d) By Theorem 7.12, we can solve this exercise by finding a matrix \( A \) for which \( A \) and \( |A| \) have different largest singular values. As \( A \) is real and symmetric, there exist \( a, b, c \in \mathbb{R} \) such that 

\[
A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \quad |A| = \begin{bmatrix} |a| & |b| \\ |b| & |c| \end{bmatrix},
\]

\[
A^T A = \begin{bmatrix} a^2 + b^2 & ab + bc \\ ab + bc & b^2 + c^2 \end{bmatrix}, \quad |A|^T |A| = \begin{bmatrix} a^2 + b^2 & |ab| + |bc| \\ |ab| + |bc| & b^2 + c^2 \end{bmatrix}.
\]

To simplify these equations we first try the case \( a + c = 0 \). This gives 

\[
A^T A = \begin{bmatrix} a^2 + b^2 & 0 \\ 0 & a^2 + b^2 \end{bmatrix}, \quad |A|^T |A| = \begin{bmatrix} a^2 + b^2 & 2|ab| \\ 2|ab| & a^2 + b^2 \end{bmatrix}.
\]

To get different norms we have to choose \( a, b \) in such a way that the maximal eigenvalues of \( A^T A \) and \( |A|^T |A| \) are different. Clearly \( A^T A \) has a unique eigenvalue \( \lambda := a^2 + b^2 \) and putting the characteristic polynomial \( \pi(\mu) = (a^2 + b^2 - \mu)^2 - 4|ab|^2 \) of \( |A|^T |A| \) to zero yields eigenvalues \( \mu_{\pm} := a^2 + b^2 \pm 2|ab| \). Hence \( |A|^T |A| \) has maximal eigenvalue \( \mu_+ = a^2 + b^2 + 2|ab| = \lambda + 2|ab| \). The spectral norms of \( A \) and \( |A| \) therefore differ whenever both \( a \) and \( b \) are nonzero. For example, when \( a = b = -c = 1 \) we find

\[
A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad \|A\|_2 = \sqrt{2}, \quad \|A\|_2 = 2.
\]

Exercise 7.32: Sharpness of perturbation bounds

Suppose \( Ax = b \) and \( Ay = b + e \). Let \( y_A \) and \( y_{A^{-1}} \) be unit vectors for which the maxima in the definition of the operator norms of \( A \) and \( A^{-1} \) are attained. That is, \( \|y_A\| = 1 = \|y_{A^{-1}}\| \), \( \|A\| = \|Ay_A\| \), and \( \|A^{-1}\| = \|A^{-1}y_{A^{-1}}\| \). If \( b = Ay_A \) and \( e = y_{A^{-1}} \), then

\[
\frac{\|y - x\|}{\|x\|} = \frac{\|A^{-1}e\|}{\|A^{-1}b\|} = \frac{\|A^{-1}y_{A^{-1}}\|}{\|y_{A^{-1}}\|} = \|A\|\|A^{-1}\| \frac{\|y_{A^{-1}}\|}{\|Ay_{A^{-1}}\|} = K \frac{\|e\|}{\|b\|},
\]

showing that the upper bound is sharp. If \( b = y_{A^{-1}} \) and \( e = Ay_A \), then

\[
\frac{\|y - x\|}{\|x\|} = \frac{\|A^{-1}e\|}{\|A^{-1}b\|} = \frac{\|y_A\|}{\|y_{A^{-1}}\|} = \frac{1}{\|A\|\|A^{-1}\|} \frac{\|Ay_A\|}{\|y_{A^{-1}}\|} = \frac{1}{K} \frac{\|e\|}{\|b\|},
\]

showing that the lower bound is sharp.

Exercise 7.33: Condition number of 2nd derivative matrix

Recall that \( T = \text{tridiag}(-1, 2, -1) \) and, by Exercise 1.10, \( T^{-1} \) is given by

\[
(T^{-1})_{ij} = (T^{-1})_{ji} = (1 - ih)j > 0, \quad 1 \leq j \leq i \leq m, \quad h = \frac{1}{m + 1}.
\]

From Theorems 7.12 and 7.14, we have the following explicit expressions for the 1-, 2- and \( \infty \)-norms

\[
\|A\|_1 = \max_{1 \leq i \leq n} \sum_{j=1}^{m} |a_{i,j}|, \quad \|A\|_2 = \sigma_1, \quad \|A^{-1}\|_2 = \frac{1}{\sigma_m}, \quad \|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{i,j}|
\]

for any matrix \( A \in \mathbb{C}^{m \times n} \), where \( \sigma_1 \) is the largest singular value of \( A \), \( \sigma_m \) the smallest singular value of \( A \), and we assumed \( A \) to be nonsingular in the third equation.
(a) For the matrix $T$ this gives $\|T\|_1 = \|T\|_\infty = m + 1$ for $m = 1, 2$ and $\|T\|_1 = \|T\|_\infty = 4$ for $m \geq 3$. For the inverse we get $\|T^{-1}\|_1 = \|T^{-1}\|_\infty = \frac{1}{2} = \frac{1}{8}h^{-2}$ for $m = 1$ and

$$\|T^{-1}\|_1 = \left\| \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix} \right\|_1 = 1 = \left\| \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix} \right\|_\infty = \|T^{-1}\|_\infty$$

for $m = 2$. For $m > 2$, one obtains

$$\sum_{i=1}^{m} \left| (T^{-1})_{ij} \right| = \sum_{i=1}^{j-1} (1 - jh)i + \sum_{i=j}^{m} (1 - ih)j$$

$$= \sum_{i=1}^{j-1} (1 - jh)i + \sum_{i=1}^{m} (1 - ih)j - \sum_{i=1}^{j-1} (1 - ih)j$$

$$= (1 - jh)(j - 1) + \frac{jm}{2} - (2 - jh)(j - 1)$$

$$= \frac{j}{2}(m + 1 - j)$$

$$= \frac{1}{2h}j - \frac{1}{2}j^2,$$

which is a quadratic function in $j$ that attains its maximum at $j = \frac{1}{2h} = \frac{m+1}{2}$. For odd $m > 1$, this function takes its maximum at integral $j$, yielding $\|T^{-1}\|_1 = \frac{1}{8}h^{-2}$. For even $m > 2$, on the other hand, the maximum over all integral $j$ is attained at $j = \frac{m}{2} = \frac{1-h}{h}$ or $j = \frac{m+2}{2} = \frac{1+h}{2h}$, which both give $\|T^{-1}\|_1 = \frac{1}{8}(h^{-2} - 1)$.

Similarly, we have for the infinity norm of $T^{-1}$

$$\sum_{j=1}^{m} \left| (T^{-1})_{ij} \right| = \sum_{j=1}^{i-1} (1 - ih)j + \sum_{j=i}^{m} (1 - jh)i = \frac{1}{2h}i - \frac{1}{2}i^2,$$

and hence $\|T^{-1}\|_\infty = \|T^{-1}\|_1$. This is what one would expect, as $T$ (and therefore $T^{-1}$) is symmetric. We conclude that the 1- and $\infty$-condition numbers of $T$ are

$$\text{cond}_1(T) = \text{cond}_\infty(T) = \begin{cases} 
2 & m = 1; \\ 
6 & m = 2; \\ 
h^{-2} & m \text{ odd, } m > 1; \\ 
h^{-2} - 1 & m \text{ even, } m > 2.
\end{cases}$$

(b) Since the matrix $T$ is symmetric, $T^T T = T^2$ and the eigenvalues of $T^T T$ are the squares of the eigenvalues $\lambda_1, \ldots, \lambda_n$ of $T$. As all eigenvalues of $T$ are positive, each singular value of $T$ is equal to an eigenvalue. Using that $\lambda_i = 2 - 2\cos(i\pi h)$, we find

$$\sigma_1 = |\lambda_n| = 2 - 2\cos(m\pi h) = 2 + 2\cos(\pi h),$$

$$\sigma_m = |\lambda_1| = 2 - 2\cos(\pi h).$$

It follows that

$$\text{cond}_2(T) = \frac{\sigma_1}{\sigma_m} = \frac{1 + \cos(\pi h)}{1 - \cos(\pi h)} = \cot^2 \left( \frac{\pi h}{2} \right).$$
(c) From \( \tan x > x \) we obtain \( \cot^2 x = \frac{1}{\tan^2 x} < \frac{1}{x^2} \). Using this and \( \cot^2 x > x^{-2} \) we find

\[
\frac{4}{\pi^2 h^2} - \frac{2}{3} < \text{cond}_2(T) < \frac{4}{\pi^2 h^2}.
\]

Exercise 7.44: When is a complex norm an inner product norm?

As in the Exercise, we let arbitrary vectors in \( \mathbb{C}^n \) be arbitrary vectors in \( \mathbb{C}^n \) and \( a \in \mathbb{C} \) be an arbitrary scalar.

1. Positive-definiteness. One has \( s(x, x) = \|x\|^2 \geq 0 \) and

\[
s(x, ix) = \frac{\|x + ix\|^2 - \|x - ix\|^2}{4} = \frac{(1 + i)x\langle \|x\|^2 - \|(1 - i)x\|^2}{4} = 0,
\]

so that \( \langle x, x \rangle = \|x\|^2 \geq 0 \), with equality holding precisely when \( x = 0 \).

2. Conjugate symmetry. Since \( s(x, y) \) is real, \( s(x, y) = s(y, x) \), \( s(ax, ay) = |a|^2 s(x, y) \), and \( s(x, -y) = -s(x, y) \),

\[
\langle y, x \rangle = s(y, x) - is(y, ix) = s(x, y) - is(ix, y) = s(x, y) - is(x, -iy) = \langle x, y \rangle.
\]

3. Linearity in the first argument. Assuming the parallelogram identity,

\[
2s(x, z) + 2s(y, z) = \frac{1}{2} \left[ \|x + z\|^2 - \frac{1}{2} \|x - z\|^2 + \frac{1}{2} \|y + z\|^2 - \frac{1}{2} \|y - z\|^2 \right]
\]

\[
= \frac{1}{2} \left[ \left\| z + \frac{x + y}{2} + \frac{x - y}{2} \right\|^2 - \frac{1}{2} \left\| z - \frac{x + y}{2} - \frac{x - y}{2} \right\|^2 \right] + \frac{1}{2} \left[ \left\| z + \frac{x + y}{2} - \frac{x - y}{2} \right\|^2 - \frac{1}{2} \left\| z - \frac{x + y}{2} + \frac{x - y}{2} \right\|^2 \right]
\]

\[
= \left\| z + \frac{x + y}{2} \right\|^2 - \left\| z - \frac{x + y}{2} \right\|^2
\]

\[
= 4s \left( \frac{x + y}{2}, z \right),
\]

implying that \( s(x + y, z) = s(x, z) + s(y, z) \). It follows that

\[
\langle x + y, z \rangle = s(x + y, z) + is(x + y, iz)
\]

\[
= s(x, z) + s(y, z) + is(x, iz) + is(y, iz)
\]

\[
= s(x, z) + is(x, iz) + s(y, z) + is(y, iz)
\]

\[
= \langle x, z \rangle + \langle y, z \rangle.
\]

That \( \langle ax, y \rangle = a \langle x, y \rangle \) follows, mutatis mutandis, from the proof of Theorem 7.42.
Exercise 7.45: p-norm for $p = 1$ and $p = \infty$

We need to verify the three properties that define a norm. Consider arbitrary vectors $\mathbf{x} = [x_1, \ldots, x_n]^T$ and $\mathbf{y} = [y_1, \ldots, y_n]$ in $\mathbb{R}^n$ and a scalar $a \in \mathbb{R}$. First we verify that $\| \cdot \|$ is a norm.

1. Positivity. Clearly $\|\mathbf{x}\| = |x_1| + \cdots + |x_n| \geq 0$, with equality holding precisely when $|x_1| = \cdots = |x_n| = 0$, which happens if and only if $\mathbf{x}$ is the zero vector.

2. Homogeneity. One has
   \[ \|a\mathbf{x}\| = |ax_1| + \cdots + |ax_n| = |a|(|x_1| + \cdots + |x_n|) = |a|\|\mathbf{x}\|. \]

3. Subadditivity. Using the triangle inequality for the absolute value,
   \[ \|\mathbf{x} + \mathbf{y}\| = |x_1 + y_1| + \cdots + |x_n + y_n| \leq |x_1| + |y_1| + \cdots + |x_n| + |y_n| = \|\mathbf{x}\| + \|\mathbf{y}\|. \]

Next we verify that $\| \cdot \|_{\infty}$ is a norm.

1. Positivity. Clearly $\|\mathbf{x}\|_{\infty} = \max\{|x_1|, \ldots, |x_n|\} \geq 0$, with equality holding precisely when $|x_1| = \cdots = |x_n| = 0$, which happens if and only if $\mathbf{x}$ is the zero vector.

2. Homogeneity. One has
   \[ \|a\mathbf{x}\|_{\infty} = \max\{|a||x_1|, \ldots, |a||x_n|\} = |a|\max\{|x_1|, \ldots, |x_n|\} = |a|\|\mathbf{x}\|_{\infty}. \]

3. Subadditivity. Using the triangle inequality for the absolute value,
   \[ \|\mathbf{x} + \mathbf{y}\|_{\infty} = \max\{|x_1 + y_1|, \ldots, |x_n + y_n|\} \leq \max\{|x_1| + |y_1|, \ldots, |x_n| + |y_n|\} \]
   \[ \leq \max\{|x_1|, \ldots, |x_n|\} + \max\{|y_1|, \ldots, |y_n|\} = \|\mathbf{x}\|_{\infty} + \|\mathbf{y}\|_{\infty} \]

Exercise 7.46: The p-norm unit sphere

In the plane, unit spheres for the 1-norm, 2-norm, and $\infty$-norm are

Exercise 7.47: Sharpness of p-norm inequality

Let $1 \leq p \leq \infty$. The vector $\mathbf{x}_1 = [1, 0, \ldots, 0]^T \in \mathbb{R}^n$ satisfies
\[ \|\mathbf{x}_1\|_p = (|1|^p + |0|^p + \cdots + |0|^p)^{1/p} = 1 = \max\{|1|, |0|, \ldots, |0|\} = \|\mathbf{x}_1\|_{\infty}, \]
and the vector $\mathbf{x}_u = [1, 1, \ldots, 1]^T \in \mathbb{R}^n$ satisfies
\[ \|\mathbf{x}_u\|_p = (|1|^p + \cdots + |1|^p)^{1/p} = n^{1/p} = n^{1/p} \max\{|1|, \ldots, |1|\} = n^{1/p}\|\mathbf{x}_u\|_{\infty}. \]
Exercise 7.48: $p$-norm inequalities for arbitrary $p$

Let $p$ and $q$ be integers satisfying $1 \leq q \leq p$, and let $x = [x_1, \ldots, x_n]^T \in \mathbb{C}^n$. Since $p/q \geq 1$, the function $f(z) = z^{p/q}$ is convex on $[0, \infty)$. For any $z_1, \ldots, z_n \in [0, \infty)$ and $\lambda_1, \ldots, \lambda_n \geq 0$ satisfying $\lambda_1 + \cdots + \lambda_n = 1$, Jensen’s inequality gives

$$\left( \sum_{i=1}^n \lambda_i z_i \right)^{p/q} = f \left( \sum_{i=1}^n \lambda_i z_i \right) \leq \sum_{i=1}^n \lambda_i f(z_i) = \sum_{i=1}^n \lambda_i z_i^{p/q}.$$ 

In particular for $z_i = |x_i|^q$ and $\lambda_1 = \cdots = \lambda_n = 1/n$, 

$$n^{-p/q} \left( \sum_{i=1}^n |x_i|^q \right)^{p/q} = \left( \sum_{i=1}^n \frac{1}{n} |x_i|^q \right)^{p/q} \leq \sum_{i=1}^n \frac{1}{n} (|x_i|^q)^{p/q} = n^{-1} \sum_{i=1}^n |x_i|^p.$$ 

Since the function $x \mapsto -\frac{x}{p}$ is monotone, we obtain

$$n^{-1/q} \|x\|_q = n^{-1/q} \left( \sum_{i=1}^n |x_i|^q \right)^{1/q} \leq n^{-1/p} \left( \sum_{i=1}^n |x_i|^p \right)^{1/p} = n^{-1/p} \|x\|_p,$$ 

from which the right inequality in the exercise follows.

The left inequality clearly holds for $x = 0$, so assume $x \neq 0$. Without loss of generality we can then assume $\|x\|_\infty = 1$, since $\|ax\|_p \leq \|ax\|_q$ if and only if $\|x\|_p \leq \|x\|_q$ for any nonzero scalar $a$. Then, for any $i = 1, \ldots, n$, one has $|x_i| \leq 1$, implying that $|x_i|^p \leq |x_i|^q$. Moreover, since $|x_i| = 1$ for some $i$, one has $|x_1|^q + \cdots + |x_n|^q \geq 1$, so that

$$\|x\|_p = \left( \sum_{i=1}^n |x_i|^p \right)^{1/p} \leq \left( \sum_{i=1}^n |x_i|^q \right)^{1/p} \leq \left( \sum_{i=1}^n |x_i|^q \right)^{1/q} = \|x\|_q.$$ 

Finally we consider the case $p = \infty$. The statement is obvious for $q = p$, so assume that $q$ is an integer. Then

$$\|x\|_q = \left( \sum_{i=1}^n |x_i|^q \right)^{1/q} \leq \left( \sum_{i=1}^n \|x\|_\infty^q \right)^{1/q} = n^{1/q} \|x\|_\infty,$$ 

proving the right inequality. Using that the map $x \mapsto x^{1/q}$ is monotone, the left inequality follows from

$$\|x\|_q^q = (\max_i |x_i|)^q \leq \sum_{i=1}^n |x_i|^q = \|x\|_q^q.$$
CHAPTER 8

The Classical Iterative Methods

Exercise 8.2: Richardson and Jacobi
If \( a_{11} = \cdots = a_{nn} = d \neq 0 \) and \( \alpha = 1/d \), Richardson’s method (8.1) yields, for \( i = 1, \ldots, n \),
\[
x_{k+1}(i) = x_k(i) + \frac{1}{d} \left( b_i - \sum_{j=1}^{n} a_{ij} x_k(j) \right)
\]
\[
= \frac{1}{d} \left( d x_k(i) - \sum_{j=1}^{n} a_{ij} x_k(j) + b_i \right)
\]
\[
= \frac{1}{a_{ii}} \left( a_{ii} x_k(i) - \sum_{j=1}^{n} a_{ij} x_k(j) + b_i \right)
\]
\[
= \frac{1}{a_{ii}} \left( - \sum_{j=1}^{i-1} a_{ij} x_k(j) - \sum_{j=i+1}^{n} a_{ij} x_k(j) + b_i \right),
\]
which is identical to Jacobi’s method (8.3).

Exercise 8.13: Convergence of the R-method when eigenvalues have positive real part (TODO)

Exercise 8.16: Example: GS converges, J diverges
The eigenvalues of \( A \) are the zeros of \( \det(A - \lambda I) = (-\lambda + 2a + 1)(\lambda + a - 1)^2 \). We find eigenvalues \( \lambda_1 := 2a + 1 \) and \( \lambda_2 := 1 - a \), the latter having algebraic multiplicity two. Whenever \(-1/2 < a < 1\) these eigenvalues are positive, implying that \( A \) is positive definite for such \( a \).

Let’s compute the spectral radius of \( G_J = I - D^{-1} A \), where \( D \) is the diagonal part of \( A \). The eigenvalues of \( G_J \) are the zeros of the characteristic polynomial
\[
\det(G_J - \lambda I) = \begin{vmatrix} -\lambda & -a & -a \\ -a & -\lambda & -a \\ -a & -a & -\lambda \end{vmatrix} = (-\lambda - 2a)(a - \lambda)^2,
\]
and we find spectral radius \( \rho(G_J) = \max\{|a|, |2a|\} \). It follows that \( \rho(G_J) > 1 \) whenever \( 1/2 < a < 1 \), in which case Theorem 8.27 implies that the Jacobi method does not converge (even though \( A \) is symmetric positive definite).
Exercise 8.17: Divergence example for J and GS

We compute the matrices $G_J$ and $G_1$ from $A$ and show that the spectral radii $\rho(G_J), \rho(G_1) \geq 1$. Once this is shown, Theorem 8.10 implies that the Jacobi method and Gauss-Seidel’s method diverge.

Write $A = D - A_L - A_R$ as in the book. From Equation (8.13), we find

$G_J = I - M_J^{-1}A = I - D^{-1}A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & \frac{3}{4} \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} -\frac{3}{4} & -2 \\ 0 & 0 \end{bmatrix},$

$G_1 = I - M_1^{-1}A = I - (D - A_L)^{-1}A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ -\frac{3}{4} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 0 & -2 \\ 0 & \frac{1}{2} \end{bmatrix}.$

From this, we find $\rho(G_J) = \sqrt{3}/2$ and $\rho(G_1) = 3/2$, both of which are bigger than 1.

Exercise 8.18: Strictly diagonally dominance; The J method

If $A = (a_{ij})_{ij}$ is strictly diagonally dominant, then it is nonsingular and $a_{11}, \ldots, a_{nn} \neq 0$. For the Jacobi method, one finds

$G = I - \text{diag}(a_{11}, \ldots, a_{nn})^{-1}A = \begin{bmatrix} 0 & -a_{12} & \ldots & -a_{1n} \\ -a_{21} & 0 & \ldots & -a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -a_{n1} & -a_{n2} & \ldots & 0 \end{bmatrix}.$

By Theorem 7.12, the $\infty$-norm can be expressed as the maximum, over all rows, of the sum of absolute values of the entries in a row. Using that $A$ is strictly diagonally dominant, one finds

$$\|G\|_{\infty} = \max_i \sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|} = \max_i \sum_{1 \leq j \leq n} \frac{|a_{ij}|}{|a_{ii}|} < 1.$$ 

As by Lemma 7.11 the $\infty$-norm is consistent, Corollary 8.9 implies that the Jacobi method converges for any strictly diagonally dominant matrix $A$.

Exercise 8.19: Strictly diagonally dominance; The GS method

Let $A = -A_L + D - A_R$ be decomposed as a sum of a lower triangular, a diagonal, and an upper triangular part. By Equation (8.14), the approximate solutions $x^{(k)}$ are related by

$$Dx^{(k+1)} = A_Lx^{(k+1)} + A_Rx^{(k)} + b$$

in the Gauss Seidel method. Let $x$ be the exact solution of $Ax = b$. It follows that the errors $\varepsilon^{(k)} := x^{(k)} - x$ are related by

$$D\varepsilon^{(k+1)} = A_L\varepsilon^{(k+1)} + A_R\varepsilon^{(k)}.$$

Let $r$ and $r_i$ be as in the exercise. Let $k \geq 0$ be arbitrary. We show by induction on $i$ that

$$|\varepsilon_j^{(k+1)}| \leq r\|\varepsilon^{(k)}\|_{\infty}, \quad \text{for } j = 1, \ldots, i.$$

51
For $i = 1$, the relation between the errors translates to

$$|e_{1}^{(k+1)}| = |a_{11}^{-1} - a_{12}e_{2}^{(k)} - \cdots - a_{1n}e_{n}^{(k)}| \leq r_{1}\|e^{(k)}\|_{\infty} \leq r\|e^{(k)}\|_{\infty}.$$ 

Fix $i \geq 2$ and assume that Equation (\ast) holds for all smaller $i$. The relation between the residuals then bounds $|e_{i}^{(k+1)}|$ as

$$|a_{ii}^{-1} - a_{ij}e_{j}^{(k+1)} - \cdots - a_{i(j-1)}e_{j-1}^{(k)} - a_{ij+1}e_{j+1}^{(k)} - \cdots - a_{in}e_{n}^{(k)}| \leq r_{j}\max\{r\|e^{(k)}\|_{\infty}, \|e^{(k)}\|_{\infty}\} = r_{j}\|e^{(k)}\|_{\infty} \leq r\|e^{(k)}\|_{\infty}.$$ 

Equation (\ast) then follows by induction.

If $A$ is strictly diagonally dominant, then $r < 1$ and

$$\lim_{k \to \infty} \|e^{(k)}\|_{\infty} \leq \|e^{(0)}\|_{\infty} \lim_{k \to \infty} r^{k} = 0.$$ 

We conclude that the Gauss Seidel method converges for strictly diagonally dominant matrices.

Exercise 8.23: Convergence example for fix point iteration

We show by induction that $x_{1}^{(k)} = x_{2}^{(k)} = 1 - a^{k}$ for every $k \geq 0$. Clearly the formula holds for $k = 0$. Assume the formula holds for some fixed $k$. Then

$$x^{(k+1)} = Gx^{(k)} + c = \begin{bmatrix} 0 & a \\ a & 0 \end{bmatrix} \begin{bmatrix} 1 - a^{k} \\ 1 - a^{k} \end{bmatrix} + \begin{bmatrix} 1 - a \\ 1 - a \end{bmatrix} = \begin{bmatrix} 1 - a^{k+1} \\ 1 - a^{k+1} \end{bmatrix},$$

It follows that the formula holds for any $k \geq 0$. When $|a| < 1$ we can evaluate the limit

$$\lim_{k \to \infty} x_{i}^{(k)} = \lim_{k \to \infty} 1 - a^{k} = 1 - \lim_{k \to \infty} a^{k} = 1, \quad \text{for } i = 1, 2.$$ 

When $|a| > 1$, however, $|x_{1}^{(k)}| = |x_{2}^{(k)}| = |1 - a^{k}|$ becomes arbitrary large with $k$ and $\lim_{k \to \infty} x_{i}^{(k)}$ diverges.

The eigenvalues of $G$ are the zeros of the characteristic polynomial $\lambda^{2} - a^{2} = (\lambda - a)(\lambda + a)$, and we find that $G$ has spectral radius $\rho(G) = 1 - \eta$, where $\eta := 1 - |a|$. Equation (8.32) yields an estimate $k = \log(10)s/(1 - |a|)$ for the smallest number of iterations $k$ so that $\rho(G)^{k} \leq 10^{-s}$. In particular, taking $a = 0.9$ and $s = 16$, one expects at least $k = 160\log(10) \approx 368$ iterations before $\rho(G)^{k} \leq 10^{-16}$. On the other hand, $0.9^{k} = |a|^{k} = 10^{-s} = 10^{-w}$ when $k \approx 350$, so in this case the estimate is fairly accurate.

Exercise 8.24: Estimate in Lemma 8.22 can be exact

As the eigenvalues of the matrix $G_{J}$ are the zeros of $\lambda^{2} - 1/4 = (\lambda - 1/2)(\lambda + 1/2) = 0$, one finds the spectral radius $\rho(G_{J}) = 1/2$. In this example, the Jacobi iteration process is described by

$$x^{(k+1)} = G_{J}x^{(k)} + c, \quad G_{J} = \begin{bmatrix} 0 & 1/2 \\ 1/2 & 0 \end{bmatrix}, \quad c = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$ 

The initial guess

$$x^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$
satisfies the formula $x_1^{(k)} = x_2^{(k)} = 1 - 2^{-k}$ for $k = 0$. Moreover, if this formula holds for some $k \geq 0$, one finds

$$x^{(k+1)} = G_J x^{(k)} + c = \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} 1 - 2^{1-k} \\ 1 - 2^{1-k} \end{bmatrix} + \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1 - 2^{-k} \\ 1 - 2^{-k} \end{bmatrix},$$

which means that it must then hold for $k + 1$ as well. By induction we can conclude that the formula holds for all $k \geq 0$.

At iteration $k$, each entry of the approximation $x^{(k)}$ differs by $2^{-k}$ from the fixed point, implying that $\|c^{(k)}\|_\infty = 2^{-k}$. Therefore, for given $s$, the error $\|c^{(k)}\|_\infty \leq 10^{-s}$ for the first time at $k = \lceil s \log(10)/\log(2) \rceil$. The bound from Lemma 8.22, on the other hand, yields $\hat{k} = 2s \log(10)$ in this case.

**Exercise 8.25: Slow spectral radius convergence**

In this exercise we show that the convergence of

$$\lim_{k \to \infty} \|A^k\|^{1/k}$$

can be quite slow. This makes it an impractical method for computing the spectral radius of $A$.

(a) The **Matlab** code

```matlab
1 n = 5
2 a = 10
3 l = 0.9
4
5 for k = n-1:200
6 L(k) = nchoosek(k,n-1)*a^(n-1)*l^(k-n+1);
7 end
8
9 stairs(L)
```

yields the following stairstep graph of $f$:

![Graph](image)

The command `max(L)` returns a maximum of $\approx 2.0589 \cdot 10^7$ of $f$ on the interval $n - 1 \leq k \leq 200$. Moreover, the code
\[ k = n-1; \]
\[ \textbf{while \ nchoosek}(k, n-1) \ast a^{-}(n-1) \ast l^{-}(k-n+1) \geq 10^{-8} \]
\[ k = k + 1; \]
\[ \textbf{end} \]

finds that \( f(k) \) dives for the first time below \( 10^{-8} \) at \( k = 470 \). We conclude that the matrix \( A^k \) is close to zero only for a very high power \( k \).

(b) Let \( E = E_1 := (A - \lambda I)/a \) be the \( n \times n \) matrix in the exercise, and write

\[ E_k := \begin{bmatrix} 0 & I_{n-k} \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{n,n}. \]

Clearly \( E^k = E_k \) for \( k = 1 \). Suppose that \( E^k = E_k \) for some \( k \) satisfying \( 1 \leq k \leq n-1 \). Using the rules of block multiplication,

\[ E^{k+1} = E^k E^1 = \begin{bmatrix} 0_{n-k,k} & I_{n-k} \\ 0_{k,k} & 0_{n-k,k} \end{bmatrix} \begin{bmatrix} 0_{k,1} & I_k \\ 0_{n-k,k+1} & 0_{n-k,k-1} \end{bmatrix} \]
\[ = \begin{bmatrix} 0_{n-k,k+1} & I_{n-k-1} \\ 0_{k,k+1} & 0_{n-k,k-1} \end{bmatrix} \]
\[ = E_{k+1}. \]

Alternatively, since

\[ (E)_{ij} = \begin{cases} 1 & \text{if } j = i + 1, \\ 0 & \text{otherwise}, \end{cases} \quad (E^k)_{ij} = \begin{cases} 1 & \text{if } j = i + k, \\ 0 & \text{otherwise}, \end{cases} \]

one has

\[ (E^{k+1})_{ij} = (E^k E)_{ij} = \sum_{\ell} (E^k)_{i\ell} (E)_{\ell j} = (E^k)_{i,i+k} (E)_{i+k,j} = 1 \cdot (E)_{i+k,j} \]

\[ = \begin{cases} 1 & \text{if } j = i + k + 1, \\ 0 & \text{otherwise}, \end{cases} \]

By induction we conclude that \( E^k = E_k \) for any \( k \) satisfying \( 1 \leq k \leq n \), with the convention that \( E^n = E_n = 0_{n,n} \). We summarize that the matrix \( E \) is \textit{nilpotent} of degree \( n \).

(c) Since the matrices \( E \) and \( I \) commute, the binomial theorem and (b) yield

\[ A^k = (aE + \lambda I)^k = \sum_{j=0}^{\min\{k,n-1\}} \binom{k}{j} \lambda^{k-j} a^j E^j. \]

Since \( (E^j)_{1,n} = 0 \) for \( 1 \leq j \leq n-2 \) and \( (E^{n-1})_{1,n} = 1 \), it follows that

\[ (A^k)_{1,n} = \sum_{j=0}^{\min\{k,n-1\}} \binom{k}{j} \lambda^{k-j} a^j (E^j)_{1,n} = \binom{k}{n-1} \lambda^{k-n+1} a^{n-1} = f(k), \]

which is what needed to be shown.
Exercise 8.31: A special norm (TODO)

Exercise 8.33: When is $A + E$ nonsingular?
Suppose $\rho(A^{-1}E) = \rho(A^{-1}(-E)) < 1$. By Theorem 8.32.2, $I + A^{-1}E$ is nonsingular and therefore so is the product $A(I + A^{-1}E) = A + E$.

Conversely, suppose $A + E$ is nonsingular. Then the inverse $C$ of $I - A^{-1}(-E) = A^{-1}(A + E)$ exists, implying that the series $\sum_{k=0}^{\infty} (A^{-1}(-E))^k$ converges (namely to $C$). By Theorem 8.32.1,

$$\rho(A^{-1}E) = \rho(A^{-1}(-E)) < 1.$$
CHAPTER 9

The Conjugate Gradient Method

Exercise 9.1: Paraboloid

Given is a quadratic function $Q(y) = \frac{1}{2} y^T A y - b^T y$, a decomposition $A = U D U^T$ with $U^T U = I$ and $D = \text{diag}(\lambda_1, \ldots, \lambda_n)$, new variables $v = [v_1, \ldots, v_n]^T := U^T y$, and a vector $c = [c_1, \ldots, c_n]^T := U^T b$. Then

$$Q(y) = \frac{1}{2} y^T U D U^T y - b^T y = \frac{1}{2} v^T D v - c^T v = \frac{1}{2} \sum_{j=1}^n \lambda_j v_j^2 - \sum_{j=1}^n c_j v_j,$$

which is what needed to be shown.

Exercise 9.4: Steepest descent iteration

In the method of Steepest Descent we choose, at the $k$th iteration, the search direction $p_k = r_k = b - Ax_k$ and optimal step length

$$\alpha_k := \frac{r_k^T r_k}{r_k^T A r_k}.$$

Given is a quadratic function

$$Q(x, y) = \frac{1}{2} \begin{bmatrix} x & y \end{bmatrix} A \begin{bmatrix} x \\ y \end{bmatrix} - b^T \begin{bmatrix} x \\ y \end{bmatrix}, \quad A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

and an initial guess $x_0 = [-1, -1/2]^T$ of its minimum. The corresponding residual is

$$r_0 = b - Ax_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} -1 \\ -1/2 \end{bmatrix} = \begin{bmatrix} 3/2 \\ 0 \end{bmatrix}.$$

Performing the steps in Equation (9.6) twice yields

$$t_0 = A r_0 = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 3/2 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 \\ -3/2 \end{bmatrix}, \quad \alpha_0 = \frac{r_0^T r_0}{r_0^T t_0} = \frac{9/4}{9/2} = \frac{1}{2},$$

$$x_1 = \begin{bmatrix} -1 \\ -1/2 \end{bmatrix} + \frac{3/2}{2} \begin{bmatrix} -1/4 \\ -1/2 \end{bmatrix} = \begin{bmatrix} -1/4 \\ -1/2 \end{bmatrix}, \quad r_1 = \begin{bmatrix} 3/2 \\ 0 \end{bmatrix} - \frac{1}{2} \begin{bmatrix} 3 \\ -3/2 \end{bmatrix} = \begin{bmatrix} 0 \\ 3/4 \end{bmatrix},$$

$$t_1 = A r_1 = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 3/2 \\ 0 \end{bmatrix} = \begin{bmatrix} -3/4 \\ 3/2 \end{bmatrix}, \quad \alpha_1 = \frac{r_1^T r_1}{r_1^T t_1} = \frac{9/16}{9/8} = \frac{1}{2},$$

$$x_2 = \begin{bmatrix} -1/4 \\ -1/2 \end{bmatrix} + \frac{3/4}{2} \begin{bmatrix} -1/4 \\ -1/8 \end{bmatrix} = \begin{bmatrix} -1/4 \\ -1/8 \end{bmatrix}, \quad r_2 = \begin{bmatrix} 0 \\ 3/4 \end{bmatrix} - \frac{1}{2} \begin{bmatrix} -3/4 \\ 3/2 \end{bmatrix} = \begin{bmatrix} 3/8 \\ 0 \end{bmatrix}.$$

Moreover, assume that for some $k \geq 1$ one has

$$t_{2k-2} = 3 \cdot 4^{1-k} \begin{bmatrix} 1 \\ -1/2 \end{bmatrix}, \quad x_{2k-1} = -4^{-k} \begin{bmatrix} 1/2 \\ 1 \end{bmatrix}, \quad r_{2k-1} = 3 \cdot 4^{-k} \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

56
By Exercise 9.7, 

\[ t_{2k-1} = 3 \cdot 4^{-k} \left[ \begin{array}{c} -1 \\ 2 \end{array} \right], \quad x_{2k} = -4^{-k} \left[ \begin{array}{c} 1 \\ 1/2 \end{array} \right], \quad r_{2k} = 3 \cdot 4^{-k} \left[ \begin{array}{c} 1/2 \\ 0 \end{array} \right]. \]

Then

\[ t_{2k} = 3 \cdot 4^{-k} \left[ \begin{array}{c} 2 \\ -1 \\ 2 \end{array} \right] \left[ \begin{array}{c} 1/2 \\ 0 \end{array} \right] = 3 \cdot 4^{1-(k+1)} \left[ \begin{array}{c} 1 \\ -1/2 \end{array} \right], \]

\[ \alpha_{2k} = \frac{r_{2k}^T r_{2k}}{r_{2k}^T t_{2k}} = \frac{9 \cdot 4^{-2k} \cdot (1/2)^2}{9 \cdot 4^{-2k} \cdot 1/2} = 1/2, \]

\[ x_{2k+1} = -4^{-k} \left[ \begin{array}{c} 1/2 \\ 0 \end{array} \right] + \frac{1}{2} \cdot 3 \cdot 4^{-k} \left[ \begin{array}{c} 1/2 \\ 0 \end{array} \right] = -4^{-(k+1)} \left[ \begin{array}{c} 1 \\ 2 \end{array} \right], \]

\[ r_{2k+1} = 3 \cdot 4^{-k} \left[ \begin{array}{c} 1/2 \\ 0 \end{array} \right] - \frac{1}{2} \cdot 3 \cdot 4^{1-(k+1)} \left[ \begin{array}{c} 1 \\ -1/2 \end{array} \right] = 3 \cdot 4^{-(k+1)} \left[ \begin{array}{c} 0 \\ 1 \end{array} \right], \]

\[ t_{2k+1} = 3 \cdot 4^{-(k+1)} \left[ \begin{array}{c} 2 \\ -1 \\ 2 \end{array} \right] \left[ \begin{array}{c} 0 \\ 1 \end{array} \right] = 3 \cdot 4^{-(k+1)} \left[ \begin{array}{c} -1 \\ 2 \end{array} \right], \]

\[ \alpha_{2k+1} = \frac{r_{2k+1}^T r_{2k+1}}{r_{2k+1}^T t_{2k+1}} = \frac{9 \cdot 4^{-2(k+1)}}{9 \cdot 4^{-2(k+1)} \cdot 2} = 1/2, \]

\[ x_{2k+2} = -4^{-(k+1)} \left[ \begin{array}{c} 1/2 \\ 0 \end{array} \right] + \frac{1}{2} \cdot 3 \cdot 4^{-(k+1)} \left[ \begin{array}{c} 0 \\ 1 \end{array} \right] = -4^{-(k+1)} \left[ \begin{array}{c} 1 \\ 2 \end{array} \right], \]

\[ r_{2k+2} = 3 \cdot 4^{-(k+1)} \left[ \begin{array}{c} 0 \\ 1 \end{array} \right] - \frac{1}{2} \cdot 3 \cdot 4^{-(k+1)} \left[ \begin{array}{c} -1 \\ 2 \end{array} \right] = 3 \cdot 4^{-(k+1)} \left[ \begin{array}{c} 0 \\ 1/2 \end{array} \right]. \]

Using the method of induction, we conclude that (1), (2), and \( \alpha_k = 1/2 \) hold for any \( k \geq 1 \).

**Exercise 9.7: Conjugate gradient iteration, II**

Using \( x^{(0)} = 0 \), one finds

\[ x^{(1)} = x^{(0)} + \frac{(b - Ax^{(0)}, b - Ax^{(0)})}{(b - Ax^{(0)}, Ab - A^2x^{(0)})} (b - Ax^{(0)}) = \frac{(b, b)}{(b, Ab)} b. \]

**Exercise 9.8: Conjugate gradient iteration, III**

By Exercise 9.7,

\[ x^{(1)} = \frac{(b, b)}{(b, Ab)} b = \frac{9}{18} \left[ \begin{array}{c} 0 \\ 3 \end{array} \right] = \left[ \begin{array}{c} 0 \\ 3/2 \end{array} \right]. \]

We find, in order,

\[ p^{(0)} = r^{(0)} = \left[ \begin{array}{c} 0 \\ 3 \end{array} \right], \quad \alpha_0 = \frac{1}{2}, \quad r^{(1)} = \left[ \begin{array}{c} 3 \\ 0 \end{array} \right], \]

\[ \beta_0 = \frac{1}{4}, \quad p^{(1)} = \left[ \begin{array}{c} 3 \\ 2 \end{array} \right], \quad \alpha_1 = \frac{2}{3}, \quad x^{(2)} = \left[ \begin{array}{c} 1 \\ 2 \end{array} \right]. \]

Since the residual vectors \( r^{(0)}, r^{(1)}, r^{(2)} \) must be orthogonal, it follows that \( r^{(2)} = 0 \) and \( x^{(2)} \) must be an exact solution. This can be verified directly by hand.
Exercise 9.9: The cg step length is optimal

For any fixed search direction \( p_k \), the step length \( \alpha_k \) is optimal if \( Q(x_{k+1}) \) is as small as possible, that is

\[
Q(x_{k+1}) = Q(x_k + \alpha_k p_k) = \min_{\alpha \in \mathbb{R}} f(\alpha),
\]

where, by (9.3),

\[
f(\alpha) := Q(x_k + \alpha p_k) = Q(x_k) - \alpha p_k^T r_k + \frac{1}{2} \alpha^2 p_k^T A p_k
\]
is a quadratic polynomial in \( \alpha \). Since \( A \) is assumed to be positive definite, necessarily \( p_k^T A p_k > 0 \). Therefore \( f \) has a minimum, which it attains at

\[
\alpha = \frac{p_k^T r_k}{p_k^T A p_k}.
\]

Applying (9.15) repeatedly, one finds that the search direction \( p_k \) for the conjugate gradient method satisfies

\[
p_k = r_k + \frac{r_k^T r_k}{r_{k-1}^T r_{k-1}} p_{k-1} = r_k + \frac{r_k^T r_k}{r_{k-1}^T r_{k-1}} \left( r_{k-1} + \frac{r_{k-1}^T r_{k-1}}{r_{k-2}^T r_{k-2}} p_{k-2} \right) = \cdots
\]

As \( p_0 = r_0 \), the difference \( p_k - r_k \) is a linear combination of the vectors \( r_{k-1}, \ldots, r_0 \), each of which is orthogonal to \( r_k \). It follows that \( p_k^T r_k = r_k^T r_k \) and that the step length \( \alpha \) is optimal for

\[
\alpha = \frac{r_k^T r_k}{p_k^T A p_k} = \alpha_k.
\]

Exercise 9.10: Starting value in cg

As in the exercise, we consider the conjugate gradient method for \( Ay = r_0 \), with \( r_0 = b - Ax_0 \). Starting with

\[
y_0 = 0, \quad s_0 = r_0 - Ay_0 = r_0, \quad q_0 = s_0 = r_0,
\]
one computes, for any \( k \geq 0 \),

\[
\gamma_k := \frac{s_k^T s_k}{q_k^T A q_k}, \quad y_{k+1} = y_k + \gamma_k q_k, \quad s_{k+1} = s_k - \gamma_k A q_k,
\]

\[
\delta_k := \frac{s_{k+1}^T s_{k+1}}{s_k^T s_k}, \quad q_{k+1} = s_{k+1} + \delta_k q_k.
\]

How are the iterates \( y_k \) and \( x_k \) related? As remarked above, \( s_0 = r_0 \) and \( q_0 = r_0 = p_0 \). Suppose \( s_k = r_k \) and \( q_k = p_k \) for some \( k \geq 0 \). Then

\[
s_{k+1} = s_k - \gamma_k A q_k = r_k - \frac{r_k^T r_k}{p_k^T A p_k} A p_k = r_k - \alpha_k A p_k = r_{k+1},
\]

\[
q_{k+1} = s_{k+1} + \delta_k q_k = r_{k+1} + \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k} p_k = p_{k+1}.
\]

It follows by induction that \( s_k = r_k \) and \( q_k = p_k \) for all \( k \geq 0 \). In addition,

\[
y_{k+1} - y_k = \gamma_k q_k = \frac{r_k^T r_k}{p_k^T A p_k} p_k = x_{k+1} - x_k, \quad \text{for any } k \geq 0,
\]

58
so that \( y_k = x_k - x_0 \).

**Exercise 9.15: The A-inner product**

We verify the axioms of Definition 0.20.

- **Positivity:** Since \( A \) is positive definite, \( x \neq 0 \iff \langle x, x \rangle = x^T Ax > 0 \). On the other hand, \( x = 0 \iff \langle x, x \rangle = x^T Ax = 0^T A 0 = 0 \). It follows that \( \langle x, x \rangle \geq 0 \) for all \( x \), with equality if and only if \( x = 0 \).

- **Symmetry:** One has \( \langle x, y \rangle = x^T Ay = (x^T Ay)^T = y^T A^T x = y^T Ax = \langle y, x \rangle \) for all vectors \( x \) and \( y \).

- **Linearity:** One has \( \langle ax + by, z \rangle = (ax + by)^T Az = ax^T Az + by^T Az = a\langle x, z \rangle + b\langle y, z \rangle \) for all real numbers \( a, b \) and vectors \( x, y, z \).

**Exercise 9.17: Program code for testing steepest descent**

Replacing the steps in (9.16) by those in (9.6), Algorithm 9.13 changes into the following algorithm for testing the method of Steepest Descent.

```matlab
function [V, K] = sdtest(m, a, d, tol, itmax)
    R = ones(m)/(m+1)^2; rho = sum(sum(R.*R)); rho0 = rho;
    V = zeros(m,m);
    T1 = sparse(toeplitz([d, a, zeros(1,m-2)]));
    for k=1:itmax
        if sqrt(rho/rho0) <= tol
            K = k; return
        end
        T = T1*R + R*T1;
        a = rho/sum(sum(R.*T)); V = V + a*R; R = R - a*T;
        rhos = rho; rho = sum(sum(R.*R));
    end
    K = itmax + 1;
end
```

To check that this program is correct, we compare its output with that of `cgtest`.

```matlab
[V1, K] = sdtest(50, -1, 2, 10^(-8), 1000000);
[V2, K] = cgtest(50, -1, 2, 10^(-8), 1000000);
surf(V2 - V1);
```

Running these commands yields Figure 1, which shows that the difference between both tests is of the order of \( 10^{-9} \), well within the specified tolerance.

As in Tables 9.12 and 9.14, we let the tolerance be \( \text{tol} = 10^{-8} \) and run `sdtest` for the \( m \times m \) grid for various \( m \), to find the number of iterations \( K_{sd} \) required before \( \|r_{K_{sd}}\|_2 \leq \text{tol} \cdot \|r_0\|_2 \). Choosing \( a = 1/9 \) and \( d = 5/18 \) yields the averaging matrix, and we find the following table.

<table>
<thead>
<tr>
<th>( n )</th>
<th>2500</th>
<th>10000</th>
<th>40000</th>
<th>1000000</th>
<th>4000000</th>
</tr>
</thead>
<tbody>
<tr>
<td>( K_{sd} )</td>
<td>37</td>
<td>35</td>
<td>32</td>
<td>26</td>
<td>24</td>
</tr>
</tbody>
</table>

Choosing \( a = -1 \) and \( d = 2 \) yields the Poisson matrix, and we find the following table.

59
Figure 1. For a $50 \times 50$ Poisson matrix and a tolerance of $10^{-8}$, the figure shows the difference of the outputs of cgtest and sdtest.

<table>
<thead>
<tr>
<th>$n$</th>
<th>100</th>
<th>400</th>
<th>1600</th>
<th>2500</th>
<th>10000</th>
<th>40000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_{sd}/n$</td>
<td>4.1900</td>
<td>4.0325</td>
<td>3.9112</td>
<td>3.8832</td>
<td>3.8235</td>
<td>3.7863</td>
</tr>
<tr>
<td>$K_{sd}$</td>
<td>419</td>
<td>1613</td>
<td>6258</td>
<td>9708</td>
<td>38235</td>
<td>151451</td>
</tr>
<tr>
<td>$K_J$</td>
<td>385</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$K_{GS}$</td>
<td>194</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$K_{SOR}$</td>
<td>35</td>
<td></td>
<td></td>
<td>164</td>
<td>324</td>
<td>645</td>
</tr>
<tr>
<td>$K_{cg}$</td>
<td>16</td>
<td>37</td>
<td>75</td>
<td>94</td>
<td>188</td>
<td>370</td>
</tr>
</tbody>
</table>

Here the number of iterations $K_J$, $K_{GS}$, and $K_{SOR}$ of the Jacobi, Gauss-Seidel and SOR methods are taken from Table 8.1, and $K_{cg}$ is the number of iterations in the Conjugate Gradient method.

Since $K_{sd}/n$ seems to tend towards a constant, it seems that the method of Steepest Descent requires $O(n)$ iterations for solving the Poisson problem for some given accuracy, as opposed to the $O(\sqrt{n})$ iterations required by the Conjugate Gradient method. The number of iterations in the method of Steepest Descent is comparable to the number of iterations in the Jacobi method, while the number of iterations in the Conjugate Gradient method is of the same order as in the SOR method.

The spectral condition number of the $m \times m$ Poisson matrix is $\kappa = (1+\cos(\pi h))/(1-\cos(\pi h))$. Theorem 9.16 therefore states that

\[
\|x - x_k\|_A \leq \left( \frac{\kappa - 1}{\kappa + 1} \right)^k = \cos^k \left( \frac{\pi}{m + 1} \right).
\]
function [x,K]=cg_leastSquares (A,b,x,tol,itmax)
    r=b-A'*A*x; p=r;
    for k=0:itmax
        if sqrt(rho/rho0)<= tol
            K=k;
            return
        end
        t=A*p; a=rho/(t'*t);
        x=x+a*p; r=r-a*A'*t;
        rhos=rho; rho=r'*r;
        p=r+(rho/rhos)*p;
    end
    K=itmax+1;
end

How can we relate this to the tolerance in the algorithm, which is specified in terms of the Euclidean norm? Since 

\[
\|x\|_2^2 = \frac{x^T A x}{x^T x}
\]

is the Rayleigh quotient of \(x\), Lemma 5.26 implies the bound

\[
\lambda_{\text{min}} \|x\|_2^2 \leq \|x\|^2_A \leq \lambda_{\text{max}} \|x\|_2^2.
\]

with \(\lambda_{\text{min}} = 4(1-\cos(\pi h))\) the smallest and \(\lambda_{\text{max}} = 4(1+\cos(\pi h))\) the largest eigenvalue of \(A\). Combining these bounds with Equation (\star) yields

\[
\frac{\|x - x_k\|_2}{\|x - x_0\|_2} \leq \sqrt{\kappa} \left(\frac{k - 1}{k + 1}\right)^k = \sqrt{\frac{1 + \cos\left(\frac{\pi}{m+1}\right)}{1 - \cos\left(\frac{\pi}{m+1}\right)}} \cos^k \left(\frac{\pi}{m + 1}\right).
\]

Replacing \(k\) by the number of iterations \(K_{sd}\) for the various values of \(m\) shows that this estimate holds for the tolerance of \(10^{-8}\).

Exercise 9.18: Using cg to solve normal equations

We need to perform Algorithm 9.11 with \(A^T A\) replacing \(A\) and \(A^T b\) replacing \(b\). For the system \(A^T A x = A^T b\), Equations (9.13), (9.14), and (9.15) become

\[
x^{(k+1)} = x^{(k)} + \alpha_k p^{(k)}, \quad \alpha_k = \frac{r^{(k)} T p^{(k)}}{p^{(k) T A^T A p^{(k)}}} = \frac{r^{(k)} T r^{(k)}}{(A p^{(k)})^T A p^{(k)}},
\]

\[
r^{(k+1)} = r^{(k)} - \alpha_k A^T A p^{(k)},
\]

\[
p^{(k+1)} = r^{(k+1)} + \beta_k p^{(k)}, \quad \beta_k = \frac{r^{(k+1) T r^{(k+1)}}}{r^{(k) T r^{(k)}}},
\]

with \(p^{(0)} = r^{(0)} = b - A^T A x^{(0)}\). Hence we only need to change the computation of \(r^{(0)}, \alpha_k, \text{ and } r^{(k+1)}\) in Algorithm 9.11, which yields the implementation in Listing 9.2.
Exercise 9.20: Maximum of a convex function

This is a special case of the maximum principle in convex analysis, which states that a convex function, defined on a compact convex set \( \Omega \), attains its maximum on the boundary of \( \Omega \).

Let \( f : [a, b] \to \mathbb{R} \) be a convex function. Consider an arbitrary point \( x = (1 - \lambda)a + \lambda b \in [a, b] \), with \( 0 \leq \lambda \leq 1 \). Since \( f \) is convex,

\[
f(x) = f((1 - \lambda)a + \lambda b) \leq (1 - \lambda)f(a) + \lambda f(b) = f(a) + \lambda(f(b) - f(a)).
\]

Since \( 0 \leq \lambda \leq 1 \), the right hand side is bounded by \( f(a) \) if \( f(b) - f(a) \) is negative, and by \( f(b) \) if \( f(b) - f(a) \) is positive. It follows that \( f(x) \leq \max\{f(a), f(b)\} \) and that \( f \) attains its maximum on the boundary of its domain of definition.

Exercise 9.25: Krylov space and cg iterations

(a) The Krylov spaces \( \mathbb{W}_k \) are defined as

\[
\mathbb{W}_k := \text{span}\{r^{(0)}, Ar^{(0)}, \ldots, A^{k-1}r^{(0)}\}.
\]

Taking \( A, b, x = 0 \), and \( r^{(0)} = b - Ax = b \) as in the Exercise, these vectors can be expressed as

\[
[r^{(0)}, Ar^{(0)}, A^2r^{(0)}] = [b, Ab, A^2b] = \begin{bmatrix} 4 & 0 & 20 \\ 0 & -4 & -16 \\ 0 & 0 & 4 \end{bmatrix}.
\]

(b) As \( x^{(0)} = 0 \) we have \( p^{(0)} = r^{(0)} = b \). We have for \( k = 0, 1, 2, \ldots \) Equations (9.13), (9.14), and (9.15),

\[
\begin{align*}
x^{(k+1)} &= x^{(k)} + \alpha_k p^{(k)}, \\
r^{(k+1)} &= r^{(k)} - \alpha_k A p^{(k)}, \\
p^{(k+1)} &= r^{(k+1)} + \beta_k p^{(k)},
\end{align*}
\]

which determine the approximations \( x^{(k)} \). For \( k = 0, 1, 2 \) these give

\[
\begin{align*}
\alpha_0 &= \frac{1}{2}, & x^{(1)} &= \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}, & r^{(1)} &= \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}, & \beta_0 = \frac{1}{4}, & p^{(1)} &= \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}, \\
\alpha_1 &= \frac{2}{3}, & x^{(2)} &= \begin{bmatrix} \frac{8}{3} \\ 4 \\ 0 \end{bmatrix}, & r^{(2)} &= \begin{bmatrix} 1 \\ 0 \\ 4 \end{bmatrix}, & \beta_1 = \frac{4}{9}, & p^{(2)} &= \begin{bmatrix} 1 \\ 4 \\ 8 \end{bmatrix}, \\
\alpha_2 &= \frac{3}{4}, & x^{(3)} &= \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}, & r^{(3)} &= \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, & \beta_2 = 0, & p^{(3)} &= \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.
\end{align*}
\]

(c) By definition we have \( \mathbb{W}_0 = \{0\} \). From the solution of part (a) we know that \( \mathbb{W}_k = \text{span}(b_0, Ab_0, \ldots, A^{k-1}b_0) \), where the vectors \( b, Ab \) and \( A^2b \) are linearly independent. Hence we have \( \text{dim} \mathbb{W}_k = k \) for \( k = 0, 1, 2, 3 \).
From (b) we know that the residual \( r^{(3)} = b - Ax^{(3)} = 0 \). Hence \( x^{(3)} \) is the exact solution to \( Ax = b \).

We observe that \( r^{(0)} = 4e_1, r^{(1)} = 2e_2 \) and \( r^{(2)} = (4/3)e_3 \) and hence the \( r^{(k)} \) for \( k = 0, 1, 2 \) are linear independent and orthogonal to each other. Thus we are only left to show that \( W_k \) is the span of \( r^{(0)}, \ldots, r^{(k-1)} \). We observe that \( b = r^{(0)}, Ab = 2r^{(0)} - 2r^{(1)} \) and \( A^2b = 5r^{(0)} - 8r^{(1)} + 3r^{(2)} \). Hence span\( (b, Ab, \ldots, Ab^{k-1}) = \text{span}(r^{(0)}, \ldots, r^{(k-1)}) \) for \( k = 1, 2, 3 \). We conclude that, for \( k = 1, 2, 3 \), the vectors \( r^{(0)}, \ldots, r^{(k-1)} \) form an orthogonal basis for \( W_k \).

One can verify directly that \( p^{(0)}, p^{(1)}, \text{ and } p^{(2)} \) are \( A \)-orthogonal. Moreover, observing that \( b = p^{(0)}, Ab = (5/2)p^{(0)} - 2p^{(1)} \), and \( A^2b = 7p^{(0)} - (28/3)p^{(1)} + 3p^{(2)} \), it follows that
\[
\text{span}(b, Ab, \ldots, Ab^{k-1}) = \text{span}(p^{(0)}, \ldots, p^{(k-1)}), \quad \text{for } k = 1, 2, 3.
\]
We conclude that, for \( k = 1, 2, 3 \), the vectors \( p^{(0)}, \ldots, p^{(k-1)} \) form an \( A \)-orthogonal basis for \( W_k \).

By computing the Euclidean norms of \( r^{(0)}, r^{(1)}, r^{(2)}, r^{(3)} \), we get
\[
\| r^{(0)} \|_2 = 4, \quad \| r^{(1)} \|_2 = 2, \quad \| r^{(2)} \|_2 = 4/3, \quad \| r^{(3)} \|_2 = 0.
\]
It follows that the sequence \( (\| r^{(k)} \|_2) \) is monotonically decreasing. Similarly, one finds
\[
\left( \| x^{(k)} - x \|_2 \right)^3_{k=0} = \left( \sqrt{10}, \sqrt{6}, \sqrt{14/9}, 0 \right),
\]
which is clearly monotonically decreasing.

**Exercise 9.28:** Another explicit formula for the Chebyshev polynomial

It is well known, and easily verified, that \( \cosh(x+y) = \cosh(x) \cosh(y) + \sinh(x) \sinh(y) \). Write \( P_n(t) = \cosh \left( n \cdot \arccosh(t) \right) \) for any integer \( n \geq 0 \). Writing \( \phi = \arccosh(t) \), and using that \( \cosh \) is even and \( \sinh \) is odd, one finds
\[
P_{n+1}(t) + P_{n-1}(t) \\
= \cosh \left( (n+1)\phi \right) + \cosh \left( (n-1)\phi \right) \\
= \cosh(n\phi) \cosh(\phi) + \sinh(n\phi) \sinh(\phi) + \cosh(n\phi) \cosh(\phi) - \sinh(n\phi) \sinh(\phi) \\
= 2 \cosh(\phi) \cosh(n\phi) \\
= 2tP_n(t).
\]
It follows that \( P_n(t) \) satisfies the same recurrence relation as \( T_n(t) \). Since in addition \( P_0(t) = 1 = T_0(t) \), necessarily \( P_n(t) = T_n(t) \) for any \( n \geq 0 \).
CHAPTER 10

Orthonormal and Unitary Transformations

Exercise 10.2: Reflector

Suppose \( x, y \in \mathbb{R}^n \) are column vectors with equal length \( l := \|x\|_2 = \|y\|_2 \), and write \( v := x - y \).

(a) Since \( v^Tv = x^Tx - y^Tx - x^Ty + y^Ty = 2l + 2y^Tx = 2v^Tx \),
we find
\[
\left( I - 2\frac{vv^T}{v^Tv} \right)x = x - \frac{v(2v^Tx)}{v^Tv} = x - \frac{vv^Tv}{v^Tv} = x - v = y.
\]

(b) Since \( \|x\|_2^2 = \|y\|_2^2 \),
\[
\langle x - y, x + y \rangle = \langle x, x \rangle - \langle y, x \rangle + \langle x, y \rangle - \langle y, y \rangle = \|x\|_2^2 - \|y\|_2^2 = 0,
\]
which means that \( x - y \) and \( x + y \) are orthogonal. \( Px \) is the orthogonal projection of \( x \) into \( \text{span}(x + y) \), because it satisfies
\[
\langle x - Px, \alpha(x + y) \rangle = \left\langle \frac{v^Txv}{v^Tv}, \alpha(x + y) \right\rangle = \left\langle \frac{1}{2}v^Tv, \alpha(x + y) \right\rangle = \left\langle \frac{1}{2}(x - y), \alpha(x + y) \right\rangle = 0,
\]
for an arbitrary element \( \alpha(x + y) \) in \( \text{span}(x + y) \).

Exercise 10.5: What does Algorithm housegen do when \( x = e_1 \)?

If \( x = e_1 \), then the algorithm yields \( \alpha = -\|e_1\|_2 = -1 \),
\[
u = \frac{x/\alpha - e_1}{\sqrt{1 - x_1/\alpha}} = \frac{e_1/(-1) - e_1}{\sqrt{1 - 1/(-1)}} = -\sqrt{2}e_1,
\]
and
\[
H = I - uu^T = \begin{bmatrix}
-1 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 1
\end{bmatrix}.
\]
Exercise 10.6: Examples of Householder transformations

(a) Let \( x \) and \( y \) be as in the exercise. As \( \|x\|_2 = \|y\|_2 \), we can apply Exercise 10.2 to obtain a vector \( v \) and a matrix \( Q \),

\[
v = x - y = \begin{bmatrix} -2 \\ 4 \end{bmatrix}, \quad Q = I - 2 \frac{vv^T}{v^Tv} = \frac{1}{5} \begin{bmatrix} 3 & 4 \\ 4 & -3 \end{bmatrix},
\]

such that \( Qx = y \). As explained in the text above Exercise 10.2, this matrix \( Q \) is a Householder transformation with \( u := \sqrt{2v}/\|v\|_2 \).

(b) Let \( x \) and \( y \) be as in the exercise. As \( \|x\|_2 = \|y\|_2 \), we can apply Exercise 10.2 to obtain a vector \( v \) and a Householder transformation \( Q \),

\[
v = x - y = \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}, \quad Q = I - 2 \frac{vv^T}{v^Tv} = \frac{1}{3} \begin{bmatrix} -1 & 2 & -2 \\ 2 & 2 & 1 \\ -2 & 1 & 2 \end{bmatrix},
\]

such that \( Qx = y \).

Exercise 10.7: \( 2 \times 2 \) Householder transformation

Let \( Q = I - uu^T \in \mathbb{R}^{2,2} \) be any Householder transformation. Then \( u = [u_1 \ u_2]^T \in \mathbb{R}^2 \) is a vector satisfying \( u_1^2 + u_2^2 = \|u\|_2^2 = 2 \), implying that the components of \( u \) are related via \( u_1^2 - 1 = 1 - u_2^2 \). Moreover, as \( 0 \leq u_1^2, u_2^2 \leq \|u\|^2 = 2 \), one has \( -1 \leq u_1^2 - 1 = 1 - u_2^2 \leq 1 \), and there exists an angle \( \phi' \in [0, 2\pi) \) such that \( \cos(\phi') = u_1^2 - 1 = 1 - u_2^2 \). For such an angle \( \phi' \), one has

\[
-u_1u_2 = \pm \sqrt{1 + \cos \phi' \sqrt{1 - \cos \phi'}} = \pm \sqrt{1 - \cos^2 \phi'} = \sin(\pm \phi').
\]

We thus find an angle \( \phi := \pm \phi' \) for which

\[
Q = \begin{bmatrix} 1 - u_1^2 & -u_1u_2 \\ -u_1u_2 & 1 - u_2^2 \end{bmatrix} = \begin{bmatrix} -\cos(\phi') & \sin(\pm \phi') \\ \sin(\pm \phi') & \cos(\phi') \end{bmatrix} = \begin{bmatrix} -\cos(\phi) & \sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{bmatrix}.
\]

Furthermore, we find

\[
Q \begin{bmatrix} \cos \phi \\ \sin \phi \end{bmatrix} = \begin{bmatrix} -\cos \phi & \sin \phi \\ \sin \phi & \cos \phi \end{bmatrix} \begin{bmatrix} \cos \phi \\ \sin \phi \end{bmatrix} = \begin{bmatrix} \sin^2 \phi - \cos^2 \phi \\ 2 \sin \phi \cos \phi \end{bmatrix} = \begin{bmatrix} -\cos(2\phi) \\ \sin(2\phi) \end{bmatrix}.
\]

When applied to the vector \([\cos \phi, \sin \phi]^T\) or \([\cos \phi, \sin \phi]^T\), therefore, \( Q \) doubles the angle and reflects the result in the \( y \)-axis.

Exercise 10.16: QR decomposition

That \( Q \) is orthonormal, and therefore unitary, can be shown directly by verifying that \( Q^TQ = I \). A direct computation shows that \( QR = A \). Moreover,

\[
R = \begin{bmatrix} 2 & 2 \\ 0 & 2 \\ 0 & 0 \end{bmatrix} = R_1 \begin{bmatrix} 0_{2,2} \end{bmatrix},
\]

where \( R_1 \) is upper triangular. It follows that \( A = QR \) is a QR decomposition.
A QR factorization is obtained by removing the parts of $Q$ and $R$ that don’t contribute anything to the product $QR$. Thus we find a QR factorization

$$A = Q_1R_1,$$

$Q_1 := \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \end{bmatrix}, \\
R_1 := \begin{bmatrix} 2 & 2 \\ 0 & 2 \end{bmatrix}.$$

**Exercise 10.17: Householder triangulation**

(a) Let

$$A = [a_1, a_2, a_3] = \begin{bmatrix} 1 & 0 & 1 \\ -2 & -1 & 0 \\ 2 & 2 & 1 \end{bmatrix}$$

be as in the Exercise. We wish to find Householder transformations $Q_1, Q_2$ that produce zeros in the columns $a_1, a_2, a_3$ of $A$. Applying Algorithm 10.4 to the first column of $A$, we find

$$u_1 = \frac{1}{\sqrt{3}} \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}, \quad Q_1 := (I - u_1u_1^T)A = \begin{bmatrix} -3 & -2 & -1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Next we need to map the bottom element $(Q_1A)_{3,2}$ of the second column to zero, without changing the first row of $Q_1A$. For this, we apply Algorithm 10.4 to the vector $[0, 1]^T$ to find

$$u_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad Q_2' := I - u_2u_2^T = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix},$$

which is a Householder transformation of size $2 \times 2$. Since

$$Q_2Q_1A := \begin{bmatrix} 1 & 0 \\ 0 & H_2 \end{bmatrix} Q_1A = \begin{bmatrix} -3 & -2 & -1 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix},$$

it follows that the Householder transformations $Q_1$ and $Q_2$ bring $A$ into upper triangular form.

(b) Clearly the matrix $Q_3 := -I$ is orthogonal and $R := Q_3Q_2Q_1A$ is upper triangular with positive diagonal elements. It follows that

$$A = QR,$$ 

$$Q := Q_1^TQ_2^TQ_3^T = Q_1Q_2Q_3,$$

is a QR factorization of $A$ of the required form.

**Exercise 10.20: QR using Gram-Schmidt, II**

Let

$$A = [a_1, a_2, a_3] = \begin{bmatrix} 1 & 3 & 1 \\ 1 & 3 & 7 \\ 1 & -1 & -4 \\ 1 & -1 & 2 \end{bmatrix}.$$
Applying Gram-Schmidt orthogonalization, we find

\[ \mathbf{v}_1 = \mathbf{a}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \]

\[ \rho_{12} = \frac{\mathbf{a}_2^T \mathbf{v}_1}{\mathbf{v}_1^T \mathbf{v}_1} = 1, \quad \mathbf{v}_2 = \mathbf{a}_2 - \rho_{12} \mathbf{v}_1 = \begin{bmatrix} 2 \\ 2 \\ -2 \\ -2 \end{bmatrix}, \]

\[ \rho_{13} = \frac{\mathbf{a}_3^T \mathbf{v}_1}{\mathbf{v}_1^T \mathbf{v}_1} = \frac{3}{2}, \quad \rho_{23} = \frac{\mathbf{a}_3^T \mathbf{v}_2}{\mathbf{v}_2^T \mathbf{v}_2} = \frac{5}{4}, \quad \mathbf{v}_3 = \mathbf{a}_3 - \rho_{13} \mathbf{v}_1 - \rho_{23} \mathbf{v}_2 = \begin{bmatrix} -3 \\ -3 \\ -3 \\ -3 \end{bmatrix}. \]

Hence we have

\[ \mathbf{V} = \begin{bmatrix} 1 & 2 & -3 \\ 1 & 2 & 3 \\ 1 & -2 & -3 \\ 1 & -2 & 3 \end{bmatrix}, \quad \hat{\mathbf{R}} = \frac{1}{4} \begin{bmatrix} 4 & 4 & 6 \\ 0 & 4 & 5 \\ 0 & 0 & 4 \end{bmatrix}, \quad \text{and} \quad \mathbf{D} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 6 \end{bmatrix}. \]

Using \( \mathbf{Q}_1 = \mathbf{V} \mathbf{D}^{-1} \) and \( \mathbf{R}_1 = \mathbf{D} \hat{\mathbf{R}} \), we obtain

\[ \mathbf{A} = \mathbf{Q}_1 \mathbf{R}_1 = \frac{1}{2} \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{bmatrix}. \]

**Exercise 10.22: Plane rotation**

Suppose

\[ \mathbf{x} = \begin{bmatrix} r \cos \alpha \\ r \sin \alpha \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix}. \]

Using the angle difference identities for the sine and cosine functions,

\[ \cos(\theta - \alpha) = \cos \theta \cos \alpha + \sin \theta \sin \alpha, \]
\[ \sin(\theta - \alpha) = \sin \theta \cos \alpha - \cos \theta \sin \alpha, \]

we find

\[ \mathbf{P} \mathbf{x} = r \begin{bmatrix} \cos \theta \cos \alpha + \sin \theta \sin \alpha \\ -\sin \theta \cos \alpha + \cos \theta \sin \alpha \end{bmatrix} = \begin{bmatrix} r \cos(\theta - \alpha) \\ -r \sin(\theta - \alpha) \end{bmatrix}. \]
Exercise 10.23: Solving upper Hessenberg system using rotations

To determine the number of arithmetic operations of Algorithm 10.24, we first consider the arithmetic operations in each step. Initially the algorithm stores the length of the matrix and adds the right hand side as the \((n + 1)\)-th column to the matrix. Such copying and storing operations do not count as arithmetic operations.

The second big step is the loop. Let us consider the arithmetic operations at the \(k\)-th iteration of this loop. First we have to compute the norm of a two dimensional vector, which comprises 4 arithmetic operations: two multiplications, one addition and one square root operation. Assuming \(r > 0\) we compute \(c\) and \(s\) each in one division, adding 2 arithmetic operations to our count. Computing the product of the Givens rotation and \(A\) includes 2 multiplications and one addition for each entry of the result. As we have \(2(n + 1 - k)\) entries, this amounts to \(6(n + 1 - k)\) arithmetic operations. The last operation in the loop is just the storage of two entries of \(A\), which again does not count as an arithmetic operation.

The final step of the whole algorithm is a backward substitution, known to require \(O(n^2)\) arithmetic operations. We conclude that the Algorithm uses

\[
O(n^2) + \sum_{k=1}^{n-1} (4 + 2 + 6(n + 1 - k)) = O(n^2) + 6 \sum_{k=1}^{n-1} (n + 2 - k)
\]

\[
= O(n^2) + 3n^2 + 9n - 12 = O(n^2)
\]

arithmetic operations.
CHAPTER 11

Least Squares

Exercise 11.7: Straight line fit (linear regression)

In each case, we are given an over-determined system $Ax = b$ with corresponding normal equations $A^*Ax = A^*b$.

(a) In this case $A = [1, 1, \ldots, 1]^T$, $x = [x_1]$, and $b = [y_1, y_2, \ldots, y_m]^T$, implying that $A^*A = [m]$ and $A^*b = [y_1 + y_2 + \cdots + y_m]$. The normal equation

$$mx_1 = y_1 + y_2 + \cdots + y_m$$

has the unique solution

$$x_1 = \frac{y_1 + y_2 + \cdots + y_m}{m},$$

which is the average of the values $y_1, y_2, \ldots, y_m$.

(b) In this case

$$A = \begin{bmatrix} 1 & t_1 & \cdots & t_m \\ 1 & t_2 & \cdots & t_m \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_m & \cdots & t_m \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad b = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix},$$

so that

$$A^*A = \begin{bmatrix} m & t_1 + \cdots + t_m \\ t_1 + \cdots + t_m & t_1^2 + \cdots + t_m^2 \end{bmatrix}, \quad A^*b = \begin{bmatrix} y_1 + \cdots + y_m \\ t_1y_1 + \cdots + t_my_m \end{bmatrix}.$$ 

The solution $x = [x_1, x_2]^T$ to the normal equations describes the line $y(t) = x_2t + x_1$ closest to the points $(t_1, y_1), \ldots, (t_m, y_m)$, in the sense that the total error

$$\|Ax - b\|^2_2 = \sum_{i=1}^{m} (x_1 + t_ix_2 - y_i)^2 = \sum_{i=1}^{m} (y(t_i) - y_i)^2$$

is minimal.

Exercise 11.8: Straight line fit using shifted power form

We are given an over-determined system $Ax = b$, with

$$A = \begin{bmatrix} 1 & t_1 - \hat{t} \\ 1 & t_2 - \hat{t} \\ \vdots & \vdots \\ 1 & t_m - \hat{t} \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad b = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}, \quad \hat{t} = \frac{t_1 + t_2 + \cdots + t_m}{m},$$

(a) The corresponding system of normal equations is $A^*Ax = A^*b$, where

$$A^*A = \begin{bmatrix} m & t_1 + \cdots + t_m - m\hat{t} \\ t_1 + \cdots + t_m - m\hat{t} & (t_1 - \hat{t})^2 + \cdots + (t_m - \hat{t})^2 \end{bmatrix}.$$
\[ A^*b = \left[ y_1 + \cdots + y_m \atop (t_1 - \hat{t}) y_1 + \cdots + (t_m - \hat{t}) y_m \right]. \]

The solution \( x = [x_1, x_2]^T \) to the normal equations describes the line \( y(t) = x_2(t - \hat{t}) + x_1 \) closest to the points \((t_1, y_1), \ldots, (t_m, y_m)\), in the sense that the total error

\[ \|Ax - b\|_2^2 = \sum_{i=1}^{m} (x_1 + (t_i - \hat{t})x_2 - y_i)^2 = \sum_{i=1}^{m} (y(t_i) - y_i)^2 \]

is minimal.

(b) Solving the normal equations, we find \( x_1 = 2.375 \) and \( x_2 = 0.87 \). Plotting the data \((t_i, y_i)\) and the fitted linear function \( y(t) = x_2(t - \hat{t}) + x_1 \) gives

![Graph showing linear fit](image)

Exercise 11.9: Fitting a circle to points

We are given the (in general overdetermined) system

\[(t_i - c_1)^2 + (y_i - c_2)^2 = r^2, \quad i = 1, \ldots, m.\]

(a) Let \( c_1 = x_1/2, c_2 = x_2/2, \) and \( r^2 = x_3 + c_1^2 + c_2^2 \) as in the Exercise. Then, for \( i = 1, \ldots, m, \)

\[ 0 = (t_i - c_1)^2 + (y_i - c_2)^2 - r^2 = \left( t_i - \frac{x_1}{2} \right)^2 + \left( y_i - \frac{x_2}{2} \right)^2 - x_3 - \left( \frac{x_1}{2} \right)^2 - \left( \frac{x_2}{2} \right)^2 = t_i^2 + y_i^2 - t_i x_1 - y_i x_2 - x_3, \]

from which Equation (11.6) follows immediately. Once \( x_1, x_2, \) and \( x_3 \) are determined, we can compute

\[ c_1 = \frac{x_1}{2}, \quad c_2 = \frac{x_2}{2}, \quad r = \sqrt{\frac{1}{4} x_1^2 + \frac{1}{4} x_2^2 + x_3}. \]

(b) The linear least square problem is to minimize \( \|Ax - b\|_2^2 \), with

\[ A = \begin{bmatrix} t_1 & y_1 & 1 \\ \vdots & \vdots & \vdots \\ t_m & y_m & 1 \end{bmatrix}, \quad b = \begin{bmatrix} t_1^2 + y_1^2 \\ \vdots \\ t_m^2 + y_m^2 \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}. \]

(c) Whether or not \( A \) has independent columns depends on the data \( t_i, y_i \). For instance, if \( t_i = y_i = 1 \) for all \( i \), then the columns of \( A \) are clearly dependent. In
general, $A$ has independent columns whenever we can find three points $(t_i, y_i)$ not on a straight line.

(d) For these points the matrix $A$ becomes

$$A = \begin{bmatrix} 1 & 4 & 1 \\ 3 & 2 & 1 \\ 1 & 0 & 1 \end{bmatrix},$$

which clearly is invertible. We find

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 & 4 & 1 \\ 3 & 2 & 1 \\ 1 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 17 \\ 13 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \\ -1 \end{bmatrix}.$$

It follows that $c_1 = 1$, $c_2 = 2$, and $r = 2$. The points $(t, y) = (1, 4), (3, 2), (1, 0)$ therefore all lie on the circle

$$(t - 1)^2 + (y - 2)^2 = 4,$$

as shown in the following picture.

Exercise 11.15: The generalized inverse

Let $A \in \mathbb{C}^{m,n}$ be a matrix of rank $r$ with singular value decomposition $A = U\Sigma V^*$. Define

$$B = A^\dagger := V\Sigma^j U^*,$$

$$\Sigma^j := \begin{bmatrix} \Sigma_i^{-1} & 0_{r,m-r} \\ 0_{n-r,r} & 0_{n-r,m-r} \end{bmatrix} \in \mathbb{R}^{n,m}.$$

Note that $\Sigma^j \Sigma = \Sigma$ and $\Sigma^j \Sigma^j = \Sigma^j$. Let us use this and the unitarity of $U$ and $V$ to show that $B$ satisfies the first two properties from the Exercise.

(1) $ABA = U\Sigma V^j V^* U^j U \Sigma V^* = U \Sigma \Sigma^j \Sigma V^* = U \Sigma V^* = A$

(2) $BAB = V \Sigma^j U^* U \Sigma V^j V^* U^* = V \Sigma^j \Sigma \Sigma^j U^* = V \Sigma U^* = B$

Moreover, since in addition the matrices $\Sigma \Sigma^j$ and $\Sigma^j \Sigma$ are Hermitian,

(3) $(BA)^* = A^* B^* = V \Sigma^j U^* U \Sigma^j V^* = V \Sigma^j \Sigma V^*$

$$= V (\Sigma^j \Sigma)^* V^* = V \Sigma^j \Sigma V^* = V \Sigma^j U^* U \Sigma V^* = BA$$

(4) $(AB)^* = B^* A^* = U \Sigma V^j V^* U^j U^* = U \Sigma^j \Sigma U^*$

$$= U (\Sigma^j \Sigma)^* U^* = U \Sigma^j \Sigma U^* = U \Sigma V^j V^* U^* = AB$$
Exercise 11.16: Uniqueness of generalized inverse

Denote the Properties to the left by (1_B), (2_B), (3_B), (4_B) and the Properties to the right by (1_C), (2_C), (3_C), (4_C). Then one uses, in order, (2_B), (4_B), (1_C), (4_C), (4_B), (1_B) or (2_C), (1_C) or (2_C), (3_C), (3_B), (3_B), (3_C), (2_C).

Exercise 11.17: Verify that a matrix is a generalized inverse

Let

\[ A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}, \quad B = \frac{1}{4} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \]

be as in the Exercise. One finds

\[ AB = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix} \cdot \frac{1}{4} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \]

\[ BA = \frac{1}{4} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \]

so that \((AB)^* = AB\) and \((BA)^* = BA\). Moreover,

\[ ABA = A(BA) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix} \cdot \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix} = A, \]

\[ BAB = (BA)B = \frac{1}{2} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \cdot \frac{1}{4} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} = B. \]

By Exercises 11.15 and 11.16, we conclude that \(B\) must be the pseudoinverse of \(A\).

Exercise 11.18: Linearly independent columns and generalized inverse

If \(A \in \mathbb{C}^{m,n}\) has independent columns then both \(A\) and \(A^*\) have rank \(n \leq m\). Then, by Theorem 6.18, \(A^*A\) must have rank \(n\) as well. Since \(A^*A\) is an \(n \times n\)-matrix of maximal rank, it is nonsingular and we can define \(B := (A^*A)^{-1}A^*\). We verify that \(B\) satisfies the four axioms of Exercise 11.15.

1. \(ABA = A(A^*A)^{-1}A^*A = A\)
2. \(BAB = (A^*A)^{-1}A^*A(A^*A)^{-1}A^* = (A^*A)^{-1}A^* = B\)
3. \((BA)^* = ((A^*A)^{-1}A^*)^* = I_n^* = I_n = (A^*A)^{-1}A^*A = BA\)
4. \((AB)^* = (A(A^*A)^{-1}A^*)^* = A((A^*A)^{-1})^*A^*
   = A(A^*A)^{-1}A^* = AB\)

It follows that \(B = A^\dagger\). The second claim follows similarly.

Alternatively, one can use the fact that the unique solution of the least squares problem is \(A^\dagger b\) and compare this with the solution of the normal equation.

Exercise 11.19: The generalized inverse of a vector

This is a special case of Exercise 11.18. In particular, if \(u\) is a nonzero vector, then \(u^*u = \langle u, u \rangle = ||u||^2\) is a nonzero number and \((u^*u)^{-1}u^*\) is defined. One can again check the axioms of Exercise 11.15 to show that this vector must be the pseudoinverse of \(u^*\).
Exercise 11.20: The generalized inverse of an outer product

Let \( A = uv^* \) be as in the Exercise. Since \( u \) and \( v \) are nonzero, the matrix

\[
B := \frac{A^*}{||u||_2^2 ||v||_2^2} = \frac{vu^*}{||u||_2^2 ||v||_2^2}
\]

is well defined. We verify the four axioms of Exercise 11.15 to show that \( B \) must be the pseudoinverse of \( A \).

\( (1) \) \( ABA = \frac{uv^*vu^*uv^*}{||u||_2^2 ||v||_2^2} = \frac{u||v||_2^2 ||u||_2^2 v^*}{||u||_2^2 ||v||_2^2} = uv^* = A; \)

\( (2) \) \( BAB = \frac{vu^*vu^*uv^*}{||u||_2^2 ||v||_2^2} = \frac{v||u||_2^2 ||v||_2^2 u^*}{||u||_2^2 ||v||_2^2} = vu^* = B; \)

\( (3) \) \( (BA)^* = \left( \frac{vu^*uv^*}{||u||_2^2 ||v||_2^2} \right)^* = \frac{vu^*uv^*}{||u||_2^2 ||v||_2^2} = BA; \)

\( (4) \) \( (AB)^* = \left( \frac{uv^*vu^*}{||u||_2^2 ||v||_2^2} \right)^* = \frac{uv^*vu^*}{||u||_2^2 ||v||_2^2} = AB. \)

This proves that \( B \) is the pseudoinverse of \( A \).

Exercise 11.21: The generalized inverse of a diagonal matrix

Let \( A := \text{diag}(\lambda_1, \ldots, \lambda_n) \) and \( B := \text{diag}(\lambda_1^*, \ldots, \lambda_n^*) \) as in the exercise. Note that, by definition, \( \lambda_j^* \) indeed represents the pseudoinverse of the number \( \lambda_j \) for any \( j \). It therefore satisfies the axioms of Exercise 11.15, something we shall use below. We now verify the axioms for \( B \) to show that \( B \) must be the pseudoinverse of \( A \).

\( (1) \) \( ABA = \text{diag}(\lambda_1^* \lambda_1, \ldots, \lambda_n^* \lambda_n) = \text{diag}(\lambda_1, \ldots, \lambda_n) = A; \)

\( (2) \) \( BAB = \text{diag}(\lambda_1^* \lambda_1, \ldots, \lambda_n^* \lambda_n) = \text{diag}(\lambda_1^*, \ldots, \lambda_n^*) = B; \)

\( (3) \) \( (BA)^* = \text{diag}(\lambda_1^* \lambda_1, \ldots, \lambda_n^* \lambda_n)^* = \text{diag}(\lambda_1^*, \ldots, \lambda_n^*) = BA; \)

\( (4) \) \( (AB)^* = \text{diag}(\lambda_1^*, \ldots, \lambda_n^*) = \text{diag}(\lambda_1^*, \ldots, \lambda_n^*) = AB. \)

This proves that \( B \) is the pseudoinverse of \( A \).

Exercise 11.22: Properties of the generalized inverse

Let \( A = U \Sigma V^* \) be a singular value decomposition of \( A \) and \( A = U_1 \Sigma_1 V_1^* \) the corresponding singular value factorization. By definition of the pseudo inverse, \( A^\dagger := V_1 \Sigma_1^{-1} U_1^* \).

\( a \) One has \( (A^\dagger)^* = (V_1 \Sigma_1^{-1} U_1^*)^* = U_1 \Sigma_1^{-*} V_1^* \). On the other hand, the matrix \( A^* \) has singular value factorization \( A^* = V_1 \Sigma_1^{-*} U_1^* \), so that its pseudo inverse is \( (A^*)^\dagger := U_1 \Sigma_1^{-*} V_1^* \) as well. We conclude that \( (A^\dagger)^* = (A^*)^\dagger \).

\( b \) Since \( A^\dagger := V_1 \Sigma_1^{-1} U_1^* \) is a singular value factorization, it has pseudo inverse \( (A^\dagger)^\dagger = (U_1^*)^* (\Sigma_1^{-1})^{-1} V_1^* = U_1 \Sigma_1^{-1} V_1^* = A \).

\( c \) Let \( \alpha \neq 0 \). Since the matrix \( \alpha A \) has singular value factorization \( U_1 (\alpha \Sigma_1) V_1^* \), it has pseudo inverse

\[
(\alpha A)^\dagger = V_1 (\alpha \Sigma_1)^{-1} U_1^* = \alpha^{-1} V_1 \Sigma_1^{-1} U_1^* = \alpha A^\dagger.
\]
Exercise 11.23: The generalized inverse of a product

(a) From the condition that $A$ has linearly independent columns we can deduce that $n \leq m$. Similarly it follows that $n \leq k$, hence $n \leq \min\{m, k\}$ and both matrices have maximal rank. As a consequence,

$$A = U_A \Sigma_A V_A^* = \begin{bmatrix} U_{A,1} & U_{A,2} \end{bmatrix} \begin{bmatrix} \Sigma_{A,1} \ 0 \end{bmatrix} V_A^*,$$

$$B = U_B \begin{bmatrix} \Sigma_{B,1} & 0 \end{bmatrix} \begin{bmatrix} V_{B,1} & V_{B,2} \end{bmatrix}^*.$$

This gives

$$A^\dagger A = V_A \begin{bmatrix} \Sigma_{A,1}^{-1} & 0 \end{bmatrix} \begin{bmatrix} U_{A,1} & U_{A,2} \end{bmatrix}^* \begin{bmatrix} \Sigma_{A,1} \ 0 \end{bmatrix} V_A$$

$$= V_A \Sigma_{A,1}^{-1} \Sigma_{A,1} V_A^* = I$$

and

$$B B^\dagger = U_B \begin{bmatrix} \Sigma_{B,1} & 0 \end{bmatrix} \begin{bmatrix} V_{B,1} & V_{B,2} \end{bmatrix}^* \begin{bmatrix} \Sigma_{B,1}^{-1} & 0 \end{bmatrix} U_B^*$$

$$= U_B \Sigma_{B,1} \Sigma_{B,1}^{-1} U_B^* = I.$$

Moreover we get

$$(AA^\dagger)^* = \begin{bmatrix} U_{A,1} & U_{A,2} \end{bmatrix} \begin{bmatrix} \Sigma_{A,1}^{-1} & 0 \end{bmatrix} V_A^* V_A \begin{bmatrix} \Sigma_{A,1} & 0 \end{bmatrix} \begin{bmatrix} U_{A,1} & U_{A,2} \end{bmatrix}^*$$

$$= \begin{bmatrix} U_{A,1} & U_{A,2} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_{A,1} & U_{A,2} \end{bmatrix}^*$$

$$= \begin{bmatrix} U_{A,1} & U_{A,2} \end{bmatrix} \begin{bmatrix} \Sigma_{A,1}^{-1} & 0 \end{bmatrix} V_A^* V_A \begin{bmatrix} \Sigma_{A,1}^{-1} & 0 \end{bmatrix} \begin{bmatrix} U_{A,1} & U_{A,2} \end{bmatrix}^*$$

$$= AA^\dagger$$

and

$$(B^\dagger B)^* = \begin{bmatrix} V_{B,1} & V_{B,2} \end{bmatrix} \begin{bmatrix} \Sigma_{B,1} \ 0 \end{bmatrix} U_B^* U_B \begin{bmatrix} \Sigma_{B,1}^{-1} & 0 \end{bmatrix} \begin{bmatrix} V_{B,1} & V_{B,2} \end{bmatrix}^*$$

$$= \begin{bmatrix} V_{B,1} & V_{B,2} \end{bmatrix} \begin{bmatrix} \Sigma_{B,1}^{-1} \ 0 \end{bmatrix} U_B^* U_B \begin{bmatrix} \Sigma_{B,1} & 0 \end{bmatrix} \begin{bmatrix} V_{B,1} & V_{B,2} \end{bmatrix}^*$$

$$= B^\dagger B.$$

We now let $E := AB$ and $F := B^\dagger A^\dagger$. Hence we want to show that $E^\dagger = F$. We do that by showing that $F$ satisfies the properties given in Exercise 11.15.

$$EFE = ABB^\dagger A^\dagger AB = AB = E$$

$$FEF = B^\dagger A^\dagger ABB^\dagger A^\dagger = B^\dagger A^\dagger = F$$

$$(FE)^* = (B^\dagger A^\dagger AB)^* = (B^\dagger B)^* = B^\dagger B = B^\dagger A^\dagger AB = FE$$

$$(EF)^* = (ABB^\dagger A^\dagger)^* = (AA^\dagger)^* = AA^\dagger = ABB^\dagger A^\dagger = EF$$

(b) We have

$$A = \begin{bmatrix} a & b \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} c \\ d \end{bmatrix}.$$
This gives

\[
A^T A = \begin{bmatrix} a^2 & ab \\ ab & b^2 \end{bmatrix} \quad \text{and} \quad B^T B = c^2 + d^2.
\]

Hence we want to choose \( a \) and \( b \) such that \( A^T A \) is diagonal and \( c \) and \( d \) such that it is the square of a nice number. Thus we set \( b = 0, c = 3 \) and \( d = 4 \), yielding

\[
A = \begin{bmatrix} a & 0 \end{bmatrix}, \quad A^T A = \begin{bmatrix} a^2 & 0 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \quad B^T B = 25 = 5^2.
\]

We hence can derive the following singular value decompositions and pseudoinverse.

\[
A = \begin{bmatrix} 1 \\ a \\ 0 \end{bmatrix} \frac{1}{a} \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad A^\dagger = \frac{1}{a} \begin{bmatrix} 1 \\ 0 \end{bmatrix},
\]

\[
B = \frac{1}{5} \begin{bmatrix} 3 \\ 4 \\ -4 \\ 3 \end{bmatrix} \begin{bmatrix} 5 \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad B^\dagger = \frac{1}{25} \begin{bmatrix} 3 & 4 \end{bmatrix}.
\]

We thus get

\[
(AB)^\dagger = (3a)^\dagger = \frac{1}{3a} \quad \text{and} \quad B^\dagger A^\dagger = \frac{3}{25a},
\]

and have

\[
(AB)^\dagger = \frac{1}{3a} \neq \frac{9}{25} : \frac{1}{3a} = \frac{3}{25a} = B^\dagger A^\dagger
\]

for all nonzero \( a \in \mathbb{R} \).

**Exercise 11.24: The generalized inverse of the conjugate transpose**

Let \( A \) have singular value factorization \( A = U_1 \Sigma_1 V_1^* \), so that \( A^* = V_1 \Sigma_1^* U_1^* \) and

\[
A^\dagger = V_1 \Sigma_1^{-1} U_1^*.
\]

Then \( A^* = A^\dagger \) if and only if \( \Sigma_1 = \Sigma_1^{-1} \), which happens precisely when all nonzero singular values of \( A \) are one.

**Exercise 11.25: Linearly independent columns**

By Exercise 11.18, if \( A \) has rank \( n \), then \( A^\dagger = (A^* A)^{-1} A^* \). Then \( A(A^* A)^{-1} A^* b = A A^\dagger b \), which is the orthogonal projection of \( b \) into \( \text{span}(A) \) by Theorem 11.10.

**Exercise 11.26: Analysis of the general linear system**

In this exercise, we can write

\[
\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \Sigma_1 = \text{diag}(\sigma_1, \ldots, \sigma_r), \quad \sigma_1 > \cdots > \sigma_r > 0.
\]

(a) As \( U \) is unitary, we have \( U^* U = I \). We find the following sequence of equivalences.

\[
Ax = b \iff U \Sigma V^* x = b \iff U^* U \Sigma (V^* x) = U^* b \iff \Sigma y = c,
\]

which is what needed to be shown.
(b) By (a), the linear system $Ax = b$ has a solution if and only if the system

$$
\begin{bmatrix}
\Sigma_1 & 0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
y_1 \\
y_r
\end{bmatrix}
= 
\begin{bmatrix}
c_1 \\
c_r \\
c_{r+1} \\
\vdots \\
c_n
\end{bmatrix}
= c
$$

has a solution $y$. Since $\sigma_1, \ldots, \sigma_r \neq 0$, this system has a solution if and only if $c_{r+1} = \cdots = c_n = 0$. We conclude that $Ax = b$ has a solution if and only if $c_{r+1} = \cdots = c_n = 0$.

(c) By (a), the linear system $Ax = b$ has a solution if and only if the system $\Sigma y = c$ has a solution. Hence we have the following three cases.

$r = n$:
Here $y_i = c_i/\sigma_i$ for $i = 1, \ldots, n$ provides the only solution to the system $\Sigma y = b$, and therefore $x = Vy$ is the only solution to $Ax = b$. It follows that the system has exactly one solution.

$r < n$, $c_i = 0$ for $i = r + 1, \ldots, n$:
Here each solution $y$ must satisfy $y_i = c_i/\sigma_i$ for $i = 1, \ldots, r$. The remaining $y_{r+1}, \ldots, y_n$, however, can be chosen arbitrarily. Hence we have infinitely many solutions to $\Sigma y = b$ as well as for $Ax = b$.

$r < n$, $c_i \neq 0$ for some $i$ with $r + 1 \leq i \leq n$:
In this case it is impossible to find a $y$ that satisfies $\Sigma y = b$, and therefore the system $Ax = b$ has no solution at all.

**Exercise 11.27: Fredholm’s Alternative**

Suppose that the system $Ax = b$ has a solution, i.e., $b \in \text{span}(A)$. Suppose in addition that $A^*y = 0$ has a solution, i.e., $y \in \ker(A^*)$. Since $(\text{span}(A))^\perp = \ker(A^*)$, one has $(y, b) = y^*b = 0$. Thus if the system $Ax = b$ has a solution, then we can not find a solution to $A^*y = 0$, $y^*b \neq 0$. Conversely if $y \in \ker(A^*)$ and $y^*b \neq 0$, then $b \notin (\ker(A^*))^\perp = \text{span}(A)$, implying that the system $Ax = b$ does not have a solution.

**Exercise 11.33: Condition number**

Let

$$
A = \begin{bmatrix}
1 & 2 \\
1 & 1 \\
1 & 1
\end{bmatrix}, \quad b = \begin{bmatrix}
b_1 \\
b_2 \\
b_3
\end{bmatrix}
$$

be as in the Exercise.

(a) By Exercise 11.18, the pseudoinverse of $A$ is

$$
A^\dagger = (A^T A)^{-1} A^T = \begin{bmatrix}
-1 & 1 & 1 \\
1 & -\frac{1}{2} & -\frac{1}{2}
\end{bmatrix}.
$$
Theorem 11.10 tells us that the orthogonal projection of \( b \) into \( \text{span}(A) \) is
\[
b_1 := AA^\dagger b = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 2b_1 \\ b_2 + b_3 \\ b_2 + b_3 \end{bmatrix},
\]
so that the orthogonal projection of \( b \) into \( \ker(A^T) \) is
\[
b_2 := (I - AA^\dagger) b = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{2} & -\frac{1}{2} \\ 0 & -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 0 \\ b_2 - b_3 \\ b_3 - b_2 \end{bmatrix},
\]
where we used that \( b = b_1 + b_2 \).

(b) By Theorem 7.12, the 2-norms \( \|A\|_2 \) and \( \|A^\dagger\|_2 \) can be found by computing the largest singular values of the matrices \( A \) and \( A^\dagger \). The largest singular value \( \sigma_1 \) of \( A \) is the square root of the largest eigenvalue \( \lambda_1 \) of \( A^T A \), which satisfies
\[
0 = \det(A^T A - \lambda_1 I) = \det \begin{bmatrix} 3 - \lambda_1 & 4 \\ 6 - \lambda_1 \end{bmatrix} = \lambda_1^2 - 9\lambda_1 + 2.
\]
It follows that \( \sigma_1 = \frac{1}{2} \sqrt{2} \sqrt{9 + \sqrt{73}} \). Similarly, the largest singular value \( \sigma_2 \) of \( A^\dagger \) is the square root of the largest eigenvalue \( \lambda_2 \) of \( A^{\dagger T} A^\dagger \), which satisfies
\[
0 = \det(A^{\dagger T} A^\dagger - \lambda_2 I) = \det \begin{bmatrix} 8 & -6 & -6 \\ -6 & 5 & 5 \\ -6 & 5 & 5 \end{bmatrix} - \lambda_2 I)
\]
\[
= -\frac{1}{2} \lambda_2 (2\lambda_2^2 - 9\lambda_2 + 1).
\]
Alternatively, we could have used that the largest singular value of \( A^\dagger \) is the inverse of the smallest singular value of \( A \) (this follows from the singular value factorization). It follows that \( \sigma_2 = \frac{1}{2} \sqrt{9 + \sqrt{73}} = \sqrt{2} \sqrt{9 - \sqrt{73}} \). We conclude
\[
K(A) = \|A\|_2 \cdot \|A^\dagger\|_2 = \sqrt{\frac{9 + \sqrt{73}}{9 - \sqrt{73}}} = \frac{1}{2\sqrt{2}} \left(9 + \sqrt{73}\right) \approx 6.203.
\]

**Exercise 11.34:** Equality in perturbation bound (TODO)

**Exercise 11.36:** Problem using normal equations

(a) Let \( A, b, \) and \( \varepsilon \) be as in the exercise. The normal equations \( A^T A x = A^T b \) are then
\[
\begin{bmatrix} 3 & 3 + \varepsilon \\ 3 + \varepsilon & (\varepsilon + 1)^2 + 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 7 \\ 7 + 2\varepsilon \end{bmatrix}.
\]
If \( \varepsilon \neq 0 \), inverting the matrix \( A^T A \) yields the unique solution
\[
\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{2\varepsilon^2} \begin{bmatrix} (\varepsilon + 1)^2 + 2 & -3 - \varepsilon \\ -3 - \varepsilon & 3 \end{bmatrix} \begin{bmatrix} 7 \\ 7 + 2\varepsilon \end{bmatrix} = \begin{bmatrix} \frac{5}{2} + \frac{1}{2\varepsilon} \\ -\frac{3}{2\varepsilon} \end{bmatrix}.
\]
If \( \varepsilon = 0 \), on the other hand, then any vector \( x = [x_1, x_2]^t \) with \( x_1 + x_2 = 7/3 \) is a solution.
(b) For $\varepsilon = 0$, we get the same solution as in (a). For $\varepsilon \neq 0$, however, the solution to the system
\[
\begin{bmatrix}
3 & 3 + \varepsilon \\
3 + \varepsilon & 3 + 2\varepsilon
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix} =
\begin{bmatrix}
7 \\
7 + 2\varepsilon
\end{bmatrix}
\]
is
\[
\begin{bmatrix}
x'_1 \\
x'_2
\end{bmatrix} = -\frac{1}{\varepsilon^2}
\begin{bmatrix}
3 + 2\varepsilon & -3 - \varepsilon \\
-3 - \varepsilon & 3
\end{bmatrix}
\begin{bmatrix}
7 \\
7 + 2\varepsilon
\end{bmatrix} = \begin{bmatrix}
2 - \frac{1}{\varepsilon} \\
\frac{1}{\varepsilon}
\end{bmatrix}.
\]
We can compare this to the solution of (a) by comparing the residuals,
\[
\left\| A \begin{bmatrix}
\frac{5}{2} + \frac{1}{2\varepsilon} \\
-\frac{1}{2\varepsilon}
\end{bmatrix} - b \right\|_2 = \left\| \begin{bmatrix}
\frac{1}{2} \\
-\frac{1}{2}
\end{bmatrix} \right\|_2 = \frac{1}{\sqrt{2}}
\]
\[
\leq \sqrt{2} = \left\| \begin{bmatrix}
0 \\
-1
\end{bmatrix} \right\|_2 = \left\| A \begin{bmatrix}
2 - \frac{1}{\varepsilon} \\
\frac{1}{\varepsilon}
\end{bmatrix} - b \right\|_2,
\]
which shows that the solution from (a) is more accurate.
CHAPTER 12

Numerical Eigenvalue Problems

Exercise 12.5: Continuity of eigenvalues

In this exercise \(\|\cdot\|\) denotes the Euclidean norm. For a given matrix \(A = (a_{ij})_{ij} \in \mathbb{R}^{n \times n}\), write
\[
A(t) := D + t(A - D), \quad D := \text{diag}(a_{11}, \ldots, a_{nn}), \quad t \in \mathbb{R},
\]
for the affine combinations of \(A\) and its diagonal part \(D\). Let \(t_1, t_2 \in [0, 1]\), with \(t_1 < t_2\), so that \(A(t_1), A(t_2)\) are convex combinations of \(A\) and \(D\). For any eigenvalue \(\mu\) of \(A(t_2)\), we are asked to show that \(A(t_1)\) has an eigenvalue \(\lambda\) such that
\[
|\lambda - \mu| \leq C(t_2 - t_1)^{1/n}, \quad C \leq 2\left(\|D\| + \|A - D\|\right).
\]
In particular, every eigenvalue of \(A(t)\) is a continuous function of \(t\).

Applying Theorem 12.8 with \(A(t_1)\) and \(E = A(t_2) - A(t_1)\), one finds that \(A(t_1)\) has an eigenvalue \(\lambda\) such that
\[
|\lambda - \mu| \leq \left(\|A(t_1)\| + \|A(t_2)\|\right)^{1-1/n}\|A(t_2) - A(t_1)\|^{1/n}.
\]
Applying the triangle inequality to the definition of \(A(t_1)\) and \(A(t_2)\), and using that the function \(x \mapsto x^{1-1/n}\) is monotone increasing,
\[
|\lambda - \mu| \leq \left(2\|D\| + (t_1 + t_2)\|A - D\|\right)^{1-1/n}\|A - D\|^{1/n}(t_2 - t_1)^{1/n}.
\]
Finally, using that \(t_1 + t_2 \leq 2\), that the function \(x \mapsto x^{1/n}\) is monotone increasing, and that \(\|A - D\| \leq 2\|D\| + 2\|A - D\|\), one obtains (\(*\)).

Exercise 12.6: Nonsingularity using Gerschgorin

We compute the Gerschgorin disks
\[
R_1 = R_4 = C_1 = C_4 = \{z \in \mathbb{C}: |z - 4| \leq 1\},
\]
\[
R_2 = R_3 = C_2 = C_3 = \{z \in \mathbb{C}: |z - 4| \leq 2\}.
\]
Then, by Gerschgorin’s Circle Theorem, each eigenvalue of \(A\) lies in
\[
(R_1 \cup \cdots \cup R_4) \cap (C_1 \cup \cdots \cup C_4) = \{z \in \mathbb{C}: |z - 4| \leq 2\}.
\]
In particular \(A\) has only nonzero eigenvalues, implying that \(A\) must be nonsingular.

Exercise 12.7: Gerschgorin, strictly diagonally dominant matrix

Suppose \(A\) is a strictly diagonally dominant matrix. For such a matrix, one finds Gerschgorin disks
\[
R_i = \left\{z \in \mathbb{C}: |z - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|\right\}.
\]
Since $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$ for all $i$, the origin is not an element of any of the $R_i$, and therefore neither of the union $\bigcup R_i$, nor of the intersection $(\bigcup R_i) \cap (\bigcup C_i)$ (which is smaller). Then, by Gershgorin’s Circle Theorem, $A$ only has nonzero eigenvalues, implying that $\det(A) = \det(A - 0 \cdot I) \neq 0$ and $A$ is nonsingular.

**Exercise 12.12: $\infty$-norm of a diagonal matrix**

Let $A = \text{diag}(\lambda_1, \ldots, \lambda_n)$ be a diagonal matrix. The spectral radius $\rho(A)$ is the absolute value of the biggest eigenvalue, say $\lambda_i$, of $A$. One has

$$\|A\|_\infty = \max_{\|x\|_\infty = 1} \|Ax\|_\infty = \max_{\|x\|_\infty = 1} \max \{|\lambda_1 x_1|, \ldots, |\lambda_n x_n|\} \leq \rho(A),$$

as $\lambda_1, \ldots, \lambda_n \leq \lambda_i = \rho(A)$ and since the components of any vector $x$ satisfy $x_1, \ldots, x_n \leq \|x\|_\infty$. Moreover, this bound is attained for the standard basis vector $x = e_i$, since $\|Ae_i\|_\infty = \lambda_i = \rho(A)$.

**Exercise 12.15: Number of arithmetic operations**

An arithmetic operation is a floating point operation, so we need not bother with any integer operations, like the computation of $k + 1$ in the indices. As we are only interested in the overall complexity, we count only terms that can contribute to this.

For the first line involving $C$, the multiplication $v' * C$ involves $(n - k)^2$ floating point multiplications and about $(n - k)^2$ floating point sums. Next, computing the outer product $v*(v' * C)$ involves $(n - k)^2$ floating point multiplications, and subtracting $C - v*(v' * C)$ needs $(n - k)^2$ subtractions. This line therefore involves $4(n - k)^2$ arithmetic operations. Similarly we find $4n(n - k)$ arithmetic operations for the line after that. These $4(n - k)^2 + 4n(n - k)$ arithmetic operations need to be carried out for $k = 1, \ldots, n - 2$, meaning that the algorithm requires of the order

$$N := \sum_{k=1}^{n-2} \left(4(n - k)^2 + 4n(n - k)\right)$$

arithmetic operations. This sum can be computed by either using the formulae for $\sum_{k=1}^{n-2} k$ and $\sum_{k=1}^{n-2} k^2$, or using that the highest order term can be found by evaluating an associated integral. One finds that the algorithm requires of the order

$$N \sim \int_0^n \left(4(n - k)^2 + 4n(n - k)\right) dk = \frac{10}{3} n^3$$

arithmetic operations.

**Exercise 12.17: Number of arithmetic operations**

The multiplication $v' * C$ involves $(n - k)^2$ floating point multiplications and about $(n - k)^2$ floating point sums. Next, computing the outer product $v*(v' * C)$ involves $(n - k)^2$ floating point multiplications, and subtracting $C - v*(v' * C)$ needs $(n - k)^2$ subtractions. In total we find $(n - k)^2$ arithmetic operations, which have to be carried out for $k = 1, \ldots, n - 2$, meaning that the algorithm requires of the order

$$N := \sum_{k=1}^{n-2} 4(n - k)^2$$

arithmetic operations.
arithmetic operations. This sum can be computed by either using the formulae for
\[ \sum_{k=1}^{n-2} k \] and \[ \sum_{k=1}^{n-2} k^2, \] or using that the highest order term can be found by evaluating an associated integral. One finds that the algorithm requires of the order
\[ N \sim \int_0^n 4(n-k)^2 dk = \frac{4}{3} n^3 \]
arithmetic operations.

Exercise 12.18: Tridiagonalize a symmetric matrix
From \( w = Ev, \beta = \frac{1}{2}v^Tw \) and \( z = w - \beta v \) we get \( z = w - v\beta = Ev - \frac{1}{2}vv^TEv \) and \( z^T = v^TE - \frac{1}{2}v^TEvv^T \). Using this yields
\[ G = (I - vv^T)E(I - vv^T) = E - vv^TE - Evv^T + vv^TEvv^T \]
\[ = E - v(v^TE - \frac{1}{2}v^TEvv^T) - (Ev - \frac{1}{2}vv^TEv)v^T \]
\[ = E - vz^T - zv^T. \]

Exercise 12.22: Counting eigenvalues
Let
\[ A = \begin{bmatrix} 4 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 4 \end{bmatrix}, \quad \alpha = 4.5. \]
Applying the recursive procedure described in Corollary 12.21, we find the diagonal elements \( d_1(\alpha), d_2(\alpha), d_3(\alpha), d_4(\alpha) \) of the matrix \( D \) in the factorization \( A - \alpha I = LDL^t \),
\[ d_1(\alpha) = 4 - 9/2 = -1/2, \]
\[ d_2(\alpha) = 4 - 9/2 - 1^2/(-1/2) = +3/2, \]
\[ d_3(\alpha) = 4 - 9/2 - 1^2/(+3/2) = -7/6, \]
\[ d_4(\alpha) = 4 - 9/2 - 1^2/(-7/6) = +5/14. \]
As precisely two of these are negative, Corollary 12.21 implies that there are precisely two eigenvalues of \( A \) strictly smaller than \( \alpha = 4.5 \). As
\[ \det(A - 4.5I) = \det(LDL^t) = d_1(\alpha)d_2(\alpha)d_3(\alpha)d_4(\alpha) \neq 0, \]
the matrix \( A \) does not have an eigenvalue equal to 4.5. We conclude that the remaining two eigenvalues must be bigger than 4.5.
Exercise 12.23: Overflow in LDL^T factorization

Since \( A_n \) is tridiagonal and strictly diagonally dominant, it has a unique LU factorization by Exercise 1.8. From Equations (1.4), one can determine the corresponding LDL^T factorization. For \( n = 1, 2, \ldots \), let \( d_{n,k} \), with \( k = 1, \ldots, n \), be the diagonal elements of the diagonal matrix \( D_n \) in a symmetric factorization of \( A_n \).

(a) We proceed by induction. Let \( n \geq 1 \) be any positive integer. For the first diagonal element, corresponding to \( k = 1 \), Equations (1.4) immediately yield \( 5 + \sqrt{24} < d_{n,1} = 10 \leq 10 \). Next, assume that \( 5 + \sqrt{24} \leq d_{n,k} \leq 10 \) for some \( 1 \leq k < n \). We show that this implies that \( 5 + \sqrt{24} < d_{n,k+1} \leq 10 \). First observe that \((5 + \sqrt{24})^2 = 25 + 10\sqrt{24} + 24 = 49 + 10\sqrt{24} \). From Equations (1.4) we know that \( d_{n,k+1} = 10 - 1/d_{n,k} \), which yields \( d_{n,k+1} < 10 \) since \( d_{n,k} > 0 \). Moreover, \( 5 + \sqrt{24} < d_{n,k} \) implies \( d_{n,k+1} = 10 - 1/d_{n,k} > 10 - \frac{1}{5 + \sqrt{24}} = \frac{50 + 10\sqrt{24} - 1}{5 + \sqrt{24}} = 5 + \sqrt{24} \).

Hence \( 5 + \sqrt{24} < d_{n,k+1} \leq 10 \), and we conclude that \( 5 + \sqrt{24} < d_{n,k} \leq 10 \) for any \( n \geq 1 \) and \( 1 \leq k \leq n \).

(b) We have \( A = LDL^T \) with \( L \) triangular and with ones on the diagonal. As a consequence,

\[
\det(A) = \det(L) \det(D) \det(L) = \det(D) = \prod_{i=1}^{n} d_i > (5 + \sqrt{24})^n.
\]

In Matlab an overflow is indicated by Matlab returning Inf as result. At my computer this happens at \( n = 310 \).

Exercise 12.24: Simultaneous diagonalization

Let \( A, B, U, D, \hat{A}, \) and \( D^{-1/2} \) be as in the Exercise.

(a) Since \( D^{-1/2} \), like any diagonal matrix, and \( A \) are symmetric, one has

\[
\hat{A}^T = D^{-1/2} U A^T U^T D^{-1/2} = D^{-1/2} U A U^T D^{-1/2} = \hat{A}
\]

(b) Since \( \hat{A} \) is symmetric, it admits an orthogonal diagonalization \( \hat{A} = \hat{U} \hat{D} \hat{U} \). Let \( E := U^T D^{-1/2} \hat{U} \). Then \( E \), as the product of three nonsingular matrices, is nonsingular. Its inverse is given explicitly by \( F := UD^{1/2} \), since

\[
FE = \hat{U} D^{1/2} U U^T D^{-1/2} \hat{U}^T = \hat{U} D^{1/2} D^{-1/2} \hat{U}^T = \hat{U} \hat{U}^T = I
\]

and similar \( EF = I \). Hence \( E^{-1} = F \) and \( E \) is nonsingular. Moreover, from \( \hat{A} = \hat{U} \hat{D} \hat{U} \) follows that \( \hat{U} \hat{A} \hat{U}^T = \hat{D} \), which gives

\[
E^T AE = \hat{U} D^{-1/2} U A U^T D^{-1/2} \hat{U}^T = \hat{U} \hat{A} \hat{U}^T = \hat{D}.
\]

Similarly \( B = U^T D U \) implies \( U B U^T = D \), which yields

\[
E^T BE = \hat{U} D^{-1/2} U B U^T D^{-1/2} \hat{U}^T = \hat{U} D^{-1/2} D \hat{D} \hat{D}^{-1/2} \hat{U}^T = \hat{U}.
\]

We conclude that for a symmetric matrix \( A \) and symmetric positive definite matrix \( B \), the congruence transformation \( X \mapsto E^T X E \) simultaneously diagonalizes the matrices \( A \) and \( B \), and even maps \( B \) to the identity matrix.
Exercise 12.25: Program code for one eigenvalue

(a) Let \( A = \text{tridiag}(c,d,c) \) and \( x \) be as in the Exercise. The following \textit{Matlab} program counts the number of eigenvalues \( k \) of \( A \) strictly less than \( x \).

```matlab
1 function k=count(c,d,x)
2 n = length(d);
3 k = 0; u = d(1)-x;
4 if u < 0
5 k = k+1;
6 end
7 for i = 2:n
8  umin = \|c(i-1)\| * eps;
9  if abs(u) < umin
10     if u < 0
11        u = -umin;
12     else
13        u = umin;
14    end
15 end
16 u = d(i)-x-c(i-1)^2/u;
17 if u < 0
18 k = k+1;
19 end
20 end
```

(b) Let \( A = \text{tridiag}(c,d,c) \) and \( m \) be as in the Exercise. The following \textit{Matlab} program computes a small interval \([a,b]\) around the \( m \)th eigenvalue \( \lambda_m \) of \( A \) and returns the point \( \lambda \) in the middle of this interval.

```matlab
1 function lambda = findeigv(c,d,m)
2 n = length(d);
3 a = d(1)-abs(c(1)); b = d(1)+abs(c(1));
4 for i = 2:n-1
5    a = \min(a, d(i)-abs(c(i-1))-abs(c(i)));
6    b = \max(b, d(i)+abs(c(i-1))+abs(c(i)));
7 end
8 a = \min(a, d(n)-abs(c(n-1)));
9 b = \max(b, d(n)+abs(c(n-1)));
10 h = b-a;
11 while abs(b-a) > eps*h
12    c0 = (a+b)/2;
13    k = count(c,d,c0);
14    if k < m
15        a = c0;
16    else
17        b = c0;
18    end
19 end
20 lambda = (a+b)/2;
```

(c) The following table shows a comparison between the values and errors obtained by the different methods.
<table>
<thead>
<tr>
<th>method</th>
<th>value</th>
<th>error</th>
</tr>
</thead>
<tbody>
<tr>
<td>exact</td>
<td>0.02413912051848666</td>
<td>0</td>
</tr>
<tr>
<td>findeigv</td>
<td>0.02413912051848621</td>
<td>4.44 · 10^{-16}</td>
</tr>
<tr>
<td>Matlab eig</td>
<td>0.02413912051848647</td>
<td>1.84 · 10^{-16}</td>
</tr>
</tbody>
</table>

Exercise 12.26: Determinant of upper Hessenberg matrix (TODO)
Exercise 13.4: Orthogonal vectors

In the Exercise it is implicitly assumed that $u^*u \neq 0$ and therefore $u \neq 0$. If $u$ and $Au - \lambda u$ are orthogonal, then

$$0 = \langle u, Au - \lambda u \rangle = u^* (Au - \lambda u) = u^*Au - \lambda u^*u.$$ 

Dividing by $u^*u$ yields

$$\lambda = \frac{u^*Au}{u^*u}.$$ 

Exercise 13.14: QR convergence detail (TODO)