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 (2009–2012) 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 2012, Georg Muntingh

Contents

CHAPTER 0

Preliminaries

Exercise 0.25: Linear combinations of convergent sequences

Since $\mathbf{x}_k \to \mathbf{x}$, $\mathbf{y}_k \to \mathbf{y}$, $a_k \to a$, and $b_k \to b$, Theorem 0.22 implies

$$\lim_{k\to\infty} |a_k - a| = \lim_{k\to\infty} \|\mathbf{x}_k - \mathbf{x}\| = \lim_{k\to\infty} |b_k - b| = \lim_{k\to\infty} \|\mathbf{y}_k - \mathbf{y}\| = 0.$$

It follows that

$$(\star) \qquad \lim_{k\to\infty} |a_k - a|\|\mathbf{x}_k\| + |a|\|\mathbf{x}_k - \mathbf{x}\| + |b_k - b|\|\mathbf{y}_k\| + |b|\|\mathbf{y}_k - \mathbf{y}\| = 0,$$

since all limits

$$\lim_{k\to\infty} |a_k - a|,\ \lim_{k\to\infty} |\mathbf{x}_k|,\ \lim_{k\to\infty} \|\mathbf{x}_k - \mathbf{x}\|,\ \lim_{k\to\infty} |b_k - b|,\ \lim_{k\to\infty} \|\mathbf{y}_k\|,\ \lim_{k\to\infty} \|\mathbf{y}_k - \mathbf{y}\|$$

exist.

By the triangle inequality and multiplicativity of the norm,

$$\begin{aligned}\|a_k\mathbf{x}_k + b_k\mathbf{y}_k - a\mathbf{x} - b\mathbf{y}\| &\leq \|(a_k - a + a)\mathbf{x}_k - a\mathbf{x}\| + \|(b_k - b + b)\mathbf{y}_k - b\mathbf{y}\| \\ &\leq |a_k - a|\|\mathbf{x}_k\| + |a|\|\mathbf{x}_k - \mathbf{x}\| + |b_k - b|\|\mathbf{y}_k\| + |b|\|\mathbf{y}_k - \mathbf{y}\|,\end{aligned}$$

which, together with $(\star)$ implies

$$\lim_{k\to\infty} \|a_k\mathbf{x}_k + b_k\mathbf{y}_k - a\mathbf{x} - b\mathbf{y}\| = 0.$$

Applying Theorem 0.22 once again, we conclude that

$$\lim_{k\to\infty} a_k\mathbf{x}_k + b_k\mathbf{y}_k = a\mathbf{x} + b\mathbf{y}.$$

Exercise 0.26: Coefficient norm

Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a basis of a vector space $\mathcal{V}$. As in Section 0.3.1, we define the function $\|\cdot\|_c : \mathcal{V} \longrightarrow \mathbb{R}$ by

$$\|\mathbf{x}\|_c := \max_{1\leq j\leq n} |c_j|, \qquad \text{for any } \mathbf{x} = \sum_{j=1}^n c_j\mathbf{v}_j \in \mathcal{V}.$$

To show that $\|\cdot\|_c$ is a norm on $\mathcal{V}$, we need to verify for all vectors

$$\mathbf{x} = \sum_{j=1}^n c_j\mathbf{v}_j, \qquad \mathbf{y} = \sum_{j=1}^n d_j\mathbf{v}_j$$

in $\mathcal{V}$ and any scalar a the axioms of Definition 0.17:

Positivity. Clearly $\|\mathbf{x}\|_c := \max_{1\leq j\leq n} |c_j| \geq 0$, with equality if $\mathbf{x} = \mathbf{0}$. Conversely if $\|\mathbf{x}\|_c = 0$, then $|c_j| = 0$ for every j, implying that $\mathbf{x} = \mathbf{0}$.

Homogeneity. One has

$$\|a\mathbf{x}\|_c = \max_{1\leq j\leq n} |ac_j| = \max_{1\leq j\leq n} |a| \cdot |c_j| = |a| \cdot \max_{1\leq j\leq n} |c_j| = |a| \cdot \|\mathbf{x}\|_c.$$

Subadditivity. One has

$$\|\mathbf{x}+\mathbf{y}\|_c = \max_{1\leq j\leq n} |c_j+d_j| \leq \max_{1\leq j\leq n} (|c_j|+|d_j|) \leq \max_{1\leq j\leq n} |c_j|+\max_{1\leq j\leq n} |d_j| = \|\mathbf{x}\|_c+\|\mathbf{y}\|_c.$$

Exercise 0.32: The $\mathbf{A}^T\mathbf{A}$ inner product

Assume that $\mathbf{A} \in \mathbb{R}^{m\times n}$ has linearly independent columns. We show that $\langle\cdot,\cdot\rangle_\mathbf{A} : (x,y) \longmapsto \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{y}$ satisfies the axioms of an inner product on a real vector space $\mathcal{V}$, as described in Definition 0.27. Let $\mathbf{x},\mathbf{y},\mathbf{z} \in \mathcal{V}$ and $a,b \in \mathbb{R}$, and let $\langle\cdot,\cdot\rangle$ be the standard inner product on $\mathcal{V}$.

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

Symmetry. One has $\langle\mathbf{x},\mathbf{y}\rangle_\mathbf{A} = \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{y} = (\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{y})^T = \mathbf{y}^T\mathbf{A}^T\mathbf{A}\mathbf{x} = \langle\mathbf{y},\mathbf{x}\rangle_\mathbf{A}$.

Linearity. One has $\langle a\mathbf{x}+b\mathbf{y},\mathbf{z}\rangle_\mathbf{A} = (a\mathbf{x}+b\mathbf{y})^T\mathbf{A}^T\mathbf{A}\mathbf{z} = a\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{z} + b\mathbf{y}^T\mathbf{A}^T\mathbf{A}\mathbf{z} = a\langle\mathbf{x},\mathbf{z}\rangle_\mathbf{A} + b\langle\mathbf{y},\mathbf{z}\rangle_\mathbf{A}$.

Exercise 0.33: Complex inner product as sums of norms

By the linearity in the first component and antilinearity in the second component of the complex inner product,

$$\|\mathbf{x}+\mathbf{y}\|^2 = \langle\mathbf{x}+\mathbf{y},\mathbf{x}+\mathbf{y}\rangle = \langle\mathbf{x},\mathbf{x}\rangle + \langle\mathbf{x},\mathbf{y}\rangle + \langle\mathbf{y},\mathbf{x}\rangle + \langle\mathbf{y},\mathbf{y}\rangle,$$

$$\|\mathbf{x}-\mathbf{y}\|^2 = \langle\mathbf{x}-\mathbf{y},\mathbf{x}-\mathbf{y}\rangle = \langle\mathbf{x},\mathbf{x}\rangle - \langle\mathbf{x},\mathbf{y}\rangle - \langle\mathbf{y},\mathbf{x}\rangle + \langle\mathbf{y},\mathbf{y}\rangle,$$

$$\|\mathbf{x}+i\mathbf{y}\|^2 = \langle\mathbf{x}+i\mathbf{y},\mathbf{x}+i\mathbf{y}\rangle = \langle\mathbf{x},\mathbf{x}\rangle - i\langle\mathbf{x},\mathbf{y}\rangle + i\langle\mathbf{y},\mathbf{x}\rangle - i^2\langle\mathbf{y},\mathbf{y}\rangle,$$

$$\|\mathbf{x}-i\mathbf{y}\|^2 = \langle\mathbf{x}-i\mathbf{y},\mathbf{x}-i\mathbf{y}\rangle = \langle\mathbf{x},\mathbf{x}\rangle + i\langle\mathbf{x},\mathbf{y}\rangle - i\langle\mathbf{y},\mathbf{x}\rangle - i^2\langle\mathbf{y},\mathbf{y}\rangle.$$

It follows that

$$\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2 + i\|\mathbf{x}+i\mathbf{y}\|^2 - i\|\mathbf{x}-i\mathbf{y}\|^2 = 4\langle\mathbf{x},\mathbf{y}\rangle.$$

Exercise 0.34: Angle between vectors in complex case

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

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

Note that taking $\mathbf{x}$ and $\mathbf{y}$ perpendicular yields zero, taking $\mathbf{x}$ and $\mathbf{y}$ 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

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

Exercise 0.53: The inverse of a general 2×2 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 0.54: The inverse of a 2×2 matrix

By Exercise 0.53, 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 0.55: Sherman-Morrison formula

A direct computation yields

$$\begin{aligned}
&(\mathbf{A}+\mathbf{B}\mathbf{C}^T)\big(\mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{B}(\mathbf{I}+\mathbf{C}^T\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}^T\mathbf{A}^{-1}\big)\\
=&\mathbf{I}-\mathbf{B}(\mathbf{I}+\mathbf{C}^T\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}^T\mathbf{A}^{-1}+\mathbf{B}\mathbf{C}^T\mathbf{A}^{-1}-\mathbf{B}\mathbf{C}^T\mathbf{A}^{-1}\mathbf{B}(\mathbf{I}+\mathbf{C}^T\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}^T\mathbf{A}^{-1}\\
=&\mathbf{I}+\mathbf{B}\mathbf{C}^T\mathbf{A}^{-1}-\mathbf{B}(\mathbf{I}+\mathbf{C}^T\mathbf{A}^{-1}\mathbf{B})(\mathbf{I}+\mathbf{C}^T\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}^T\mathbf{A}^{-1}\\
=&\mathbf{I}+\mathbf{B}\mathbf{C}^T\mathbf{A}^{-1}-\mathbf{B}\mathbf{C}^T\mathbf{A}^{-1}\\
=&\mathbf{I},
\end{aligned}$$

showing that the two matrices are inverse to each other.

Exercise 0.56: Cramer's rule; special case

Cramer's rule yields

$$x_1 = \begin{vmatrix} 3 & 2 \\ 6 & 1 \end{vmatrix} \Big/ \begin{vmatrix} 1 & 2 \\ 2 & 1 \end{vmatrix} = 3, \qquad x_2 = \begin{vmatrix} 1 & 3 \\ 2 & 6 \end{vmatrix} \Big/ \begin{vmatrix} 1 & 2 \\ 2 & 1 \end{vmatrix} = 0.$$

Exercise 0.57: Adjoint matrix; special case

We are given the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & -6 & 3 \\ 3 & -2 & -6 \\ 6 & 3 & 2 \end{bmatrix}.$$

Computing the cofactors of $\mathbf{A}$ gives

$$\operatorname{adj}_{\mathbf{A}}^{T} = \begin{bmatrix} (-1)^{1+1}\begin{vmatrix}-2 & -6\\ 3 & 2\end{vmatrix} & (-1)^{1+2}\begin{vmatrix}3 & -6\\ 6 & 2\end{vmatrix} & (-1)^{1+3}\begin{vmatrix}3 & -2\\ 6 & 3\end{vmatrix} \\ \\ (-1)^{2+1}\begin{vmatrix}-6 & 3\\ 3 & 2\end{vmatrix} & (-1)^{2+2}\begin{vmatrix}2 & 3\\ 6 & 2\end{vmatrix} & (-1)^{2+3}\begin{vmatrix}2 & -6\\ 6 & 3\end{vmatrix} \\ \\ (-1)^{3+1}\begin{vmatrix}-6 & 3\\ -2 & -6\end{vmatrix} & (-1)^{3+2}\begin{vmatrix}2 & 3\\ 3 & -6\end{vmatrix} & (-1)^{3+3}\begin{vmatrix}2 & -6\\ 3 & -2\end{vmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} 14 & 21 & 42\\ -42 & -14 & 21\\ 21 & -42 & 14 \end{bmatrix}^{T}.$$

One checks directly that $\operatorname{adj}_{\mathbf{A}}\mathbf{A} = \det(\mathbf{A})\mathbf{I}$, with $\det(\mathbf{A}) = 343$.

Exercise 0.59: 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

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

of the plane.

Exercise 0.60: 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 3, the area $A(T)$ can be expressed as

$$\begin{aligned} A(T) &= A(ABP_3) + A(P_3BCP_2) - A(ACP_2)\\ &= \frac{1}{2}x_3y_3 + (x_2 - x_3)y_2 + \frac{1}{2}(x_2 - x_3)(y_3 - y_2) - \frac{1}{2}x_2y_2\\ &= \frac{1}{2}\begin{vmatrix} 1 & 1 & 1\\ 0 & x_2 & x_3\\ 0 & y_2 & y_3 \end{vmatrix}, \end{aligned}$$

which is what needed to be shown.

Exercise 0.61: 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

$$(\star) \qquad D_N = \prod_{1 \le j < i \le N} (x_i - x_j)$$

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

Let us assume $(\star)$ 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_n^k times column k from column $k+1$ for $k = n-1, n-2, \ldots, 1$, we find

$$D_n = \begin{vmatrix} 1 & x_1 - x_n & x_1^2 - x_1x_n & \cdots & x_1^{n-1} - x_1^{n-2}x_n \\ 1 & x_2 - x_n & x_2^2 - x_2x_n & \cdots & x_2^{n-1} - x_2^{n-2}x_n \\ 1 & x_3 - x_n & x_3^2 - x_3x_n & \cdots & x_3^{n-1} - x_3^{n-2}x_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x_n & x_n^2 - x_nx_n & \cdots & x_n^{n-1} - x_n^{n-2}x_n \end{vmatrix}.$$

Next, by cofactor expansion along the last row and by the multilinearity in the rows,

$$\begin{aligned} D_n &= (-1)^{n-1} \cdot 1 \cdot \begin{vmatrix} x_1 - x_n & x_1^2 - x_1x_n & \cdots & x_1^{n-1} - x_1^{n-2}x_n \\ x_2 - x_n & x_2^2 - x_2x_n & \cdots & x_2^{n-1} - x_2^{n-2}x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_{n-1} - x_n & x_{n-1}^2 - x_{n-1}x_n & \cdots & x_{n-1}^{n-1} - x_{n-1}^{n-2}x_n \end{vmatrix} \\ &= (-1)^{n-1}(x_1 - x_n)(x_2 - x_n)\cdots(x_{n-1} - x_n)D_{n-1} \\ &= (x_n - x_1)(x_n - x_2)\cdots(x_n - x_{n-1}) \prod_{1 \le j < i \le n-1} (x_i - x_j) \\ &= \prod_{1 \le j < i \le n} (x_i - x_j). \end{aligned}$$

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

Exercise 0.62: Cauchy determinant

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

$$\mathbf{A} = (a_{i,j})_{i,j} = \left(\frac{1}{\alpha_i + \beta_j} \right)_{i,j} = \begin{bmatrix} \frac{1}{\alpha_1+\beta_1} & \frac{1}{\alpha_1+\beta_2} & \cdots & \frac{1}{\alpha_1+\beta_n} \\ \frac{1}{\alpha_2+\beta_1} & \frac{1}{\alpha_2+\beta_2} & \cdots & \frac{1}{\alpha_2+\beta_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\alpha_n+\beta_1} & \frac{1}{\alpha_n+\beta_2} & \cdots & \frac{1}{\alpha_n+\beta_n} \end{bmatrix}.$$

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

$$\mathbf{C}=(c_{i,j})_{i,j},\qquad c_{i,j}=\prod_{\substack{k=1\\k\neq j}}^{n}(\alpha_i+\beta_k).$$

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 $\mathbf{C}$ is a polynomial of degree $n-1$ in the variables α_i,β_j, the determinant of $\mathbf{C}$ must be a homogeneous polynomial of degree $n(n-1)$ in α_i,β_j.

By the multilinearity of the determinant, $\det\mathbf{C}=\prod_{i,j=1}^{n}(\alpha_i+\beta_j)\det\mathbf{A}$. Since $\mathbf{A}$ vanishes whenever $\alpha_i=\alpha_j$ or $\beta_i=\beta_j$ for $i\neq j$, the homogeneous polynomial $\det\mathbf{C}$ 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

$$(\star)\qquad \det\mathbf{C}=k\prod_{1\leq i<j\leq n}(\alpha_i-\alpha_j)\prod_{1\leq i<j\leq n}(\beta_i-\beta_j)$$

for some constant k. To determine k, we can evaluate $\det\mathbf{C}$ 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 $\mathbf{C}$ becomes a diagonal matrix with determinant

$$\det\mathbf{C}=\prod_{i=1}^{n}\prod_{\substack{k=1\\k\neq i}}^{n}(\alpha_i+\beta_k)=\prod_{i=1}^{n}\prod_{\substack{k=1\\k\neq i}}^{n}(\alpha_i-\alpha_k)=\prod_{1\leq i<k\leq n}(\alpha_i-\alpha_k)\prod_{1\leq i<k\leq n}(\alpha_k-\alpha_i).$$

Comparing with $(\star)$ shows that $k=1$. We conclude that

$$(\star\star)\qquad \det\mathbf{A}=\frac{\displaystyle\prod_{1\leq i<j\leq n}(\alpha_i-\alpha_j)\prod_{1\leq i<j\leq n}(\beta_i-\beta_j)}{\displaystyle\prod_{i,j=1}^{n}(\alpha_i+\beta_j)}.$$

(b) Deleting row l and column k from $\mathbf{A}$, results in the matrix $\mathbf{A}_{l,k}$ 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 $\mathbf{A}^{-1} = (b_{k,l})$ and by $(\star\star)$,

$$\begin{aligned}
b_{k,l} &:= (-1)^{k+l} \frac{\det \mathbf{A}_{l,k}}{\det \mathbf{A}} \\
&= (-1)^{k+l} \frac{\displaystyle\prod_{i,j=1}^{n} (\alpha_i + \beta_j) \prod_{\substack{1\le i<j\le n \\ i,j\neq l}} (\alpha_i - \alpha_j) \prod_{\substack{1\le i<j\le n \\ i,j\neq k}} (\beta_i - \beta_j)}{\displaystyle\prod_{\substack{i,j=1 \\ i\neq l \\ j\neq k}}^{n} (\alpha_i + \beta_j) \prod_{1\le i<j\le n} (\alpha_i - \alpha_j) \prod_{1\le i<j\le n} (\beta_i - \beta_j)} \\
&= (\alpha_l + \beta_k) \frac{\displaystyle\prod_{\substack{s=1 \\ s\neq l}}^{n} (\alpha_s + \beta_k)}{\displaystyle\prod_{\substack{s=1 \\ s\neq l}}^{n} (\alpha_s - \alpha_l)} \frac{\displaystyle\prod_{\substack{s=1 \\ s\neq k}}^{n} (\beta_s + \alpha_l)}{\displaystyle\prod_{\substack{s=1 \\ s\neq k}}^{n} (\beta_s - \beta_k)} \\
&= (\alpha_l + \beta_k) \prod_{\substack{s=1 \\ s\neq l}}^{n} \frac{\alpha_s + \beta_k}{\alpha_s - \alpha_l} \prod_{\substack{s=1 \\ s\neq k}}^{n} \frac{\beta_s + \alpha_l}{\beta_s - \beta_k},
\end{aligned}$$

which is what needed to be shown.

Exercise 0.63: Inverse of the Hilbert matrix

If we write

$$\alpha = [\alpha_1, \ldots, \alpha_n] = [1, 2, \ldots, n], \qquad \beta = [\beta_1, \ldots, \beta_n] = [0, 1, \ldots, n-1],$$

then the Hilbert matrix matrix is of the form $\mathbf{H}_n = (h_{i,j}) = \big(1/(\alpha_i + \beta_j)\big)$. By Exercise 0.62.(b), its inverse $\mathbf{T}_n = (t^n_{i,j}) := \mathbf{H}_n^{-1}$ is given by

$$t^n_{i,j} = (i + j - 1) \prod_{\substack{s=1 \\ s\neq j}}^{n} \frac{s + i - 1}{s - j} \prod_{\substack{s=1 \\ s\neq i}}^{n} \frac{s + j - 1}{s - i}, \qquad 1 \le i, j \le n.$$

We wish to show that

$$(\star) \qquad t^n_{i,j} = \frac{f(i) f(j)}{i + j - 1}, \qquad 1 \le i, j \le n,$$

where $f : \mathbb{N} \longrightarrow \mathbb{Q}$ is the sequence defined by

$$f(1) = -n, \qquad f(i+1) = \left(\frac{i^2 - n^2}{i^2}\right) f(i), \qquad \text{for } i = 1, 2, \ldots.$$

Clearly $(\star)$ holds when $i = j = 1$. Suppose that $(\star)$ holds for some (i, j). Then

$$\begin{aligned}
t^n_{i+1,j} &= (i+j) \prod_{\substack{s=1 \\ s \neq j}}^{n} \frac{s+1+i-1}{s-j} \prod_{\substack{s=1 \\ s \neq i+1}}^{n} \frac{s+j-1}{s-1-i} \\
&= (i+j) \frac{1}{(i+j)^2} \frac{\displaystyle\prod_{s=2}^{n+1} (s+i-1)}{\displaystyle\prod_{\substack{s=1 \\ s \neq j}}^{n} (s-j)} \frac{\displaystyle\prod_{s=1}^{n} (s+j-1)}{\displaystyle\prod_{\substack{s=0 \\ s \neq i}}^{n-1} (s-i)} \\
&= \frac{(i+j-1)^2 (n+i)(n-i)}{(i+j)i(-i)} \frac{\displaystyle\prod_{\substack{s=1 \\ s \neq j}}^{n} (s+i-1)}{\displaystyle\prod_{\substack{s=1 \\ s \neq j}}^{n} (s-j)} \frac{\displaystyle\prod_{\substack{s=1 \\ s \neq i}}^{n} (s+j-1)}{\displaystyle\prod_{\substack{s=1 \\ s \neq i}}^{n} (s-i)} \\
&= \frac{1}{i+j} \frac{i^2-n^2}{i^2} (i+j-1) \prod_{\substack{s=1 \\ s \neq j}}^{n} \frac{s+i-1}{s-j} (i+j-1) \prod_{\substack{s=1 \\ s \neq i}}^{n} \frac{s+j-1}{s-i} \\
&= \frac{1}{i+j} \frac{i^2-n^2}{i^2} f(i) f(j) \\
&= \frac{f(i+1) f(j)}{(i+1)+j-1},
\end{aligned}$$

so that $(\star)$ holds for $(i+1, j)$. Carrying out a similar calculation for $(i, j+1)$, or using the symmetry of $\mathbf{T}_n$, we conclude by induction that $(\star)$ holds for any i, j.

CHAPTER 1

Gaussian Elimination

Exercise 1.1: Matrix element as a quadratic form

Write $\mathbf{A} = (a_{ij})_{ij}$ and $\mathbf{e}_i = (\delta_{ik})_k$, 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,

$$\mathbf{e}_i^T \mathbf{A} \mathbf{e}_j = \mathbf{e}_i^T (\mathbf{A}\mathbf{e}_j) = \mathbf{e}_i^T \left(\sum_k a_{lk}\delta_{jk} \right)_l = \mathbf{e}_i^T \left(a_{lj}\right)_l = \sum_l \delta_{il} a_{lj} = a_{ij}.$$

Exercise 1.2: Outer product expansion of a matrix

Let δ_{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 $\mathbf{e}_i \mathbf{e}_j^T$ satisfies

$$\left(\mathbf{e}_i \mathbf{e}_j^T\right)_{kl} = \sum_o (\mathbf{e}_i)_{ko} (\mathbf{e}_j^T)_{ol} = (\mathbf{e}_i)_{k1} (\mathbf{e}_j^T)_{1l} = \delta_{ik}\delta_{jl}.$$

It follows that

$$\left(\sum_i \sum_j a_{ij} \mathbf{e}_i \mathbf{e}_j^T \right)_{kl} = \sum_i \sum_j a_{ij} \left(\mathbf{e}_i \mathbf{e}_j^T\right)_{kl} = \sum_i \sum_j a_{ij} \delta_{ik} \delta_{jl} = a_{kl}$$

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

Exercise 1.3: The product $\mathbf{A}^T\mathbf{A}$

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

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

$$(\mathbf{B})_{ij} = \left(\mathbf{A}^T\mathbf{A}\right)_{ij} = \sum_{k=1}^n a_{ki} a_{kj} = \mathbf{a}_{.i}^T \mathbf{a}_{.j} = \langle \mathbf{a}_{.i}, \mathbf{a}_{.j} \rangle,$$

which is what needed to be shown.

Exercise 1.4: Outer product expansion

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

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

For the outer product expansion of the columns of $\mathbf{A}$ and $\mathbf{B}$, on the other hand, we find $\left(\mathbf{a}_{:k}\mathbf{b}_{:k}^T\right)_{ij} = a_{ik}b_{jk}$. It follows that

$$\left(\mathbf{AB}^T\right)_{ij} = \sum_{k=1}^{n} a_{ik}b_{jk} = \sum_{k=1}^{n} \left(\mathbf{a}_{:k}\mathbf{b}_{:k}^T\right)_{ij}.$$

Exercise 1.5: System with many right hand sides; compact form

Let $\mathbf{A}, \mathbf{B}$, and $\mathbf{X}$ be as in the Exercise.

($\Longrightarrow$): Suppose $\mathbf{AX} = \mathbf{B}$. Multiplying this equation from the right by $\mathbf{e}_j$ yields $\mathbf{Ax}_{.j} = \mathbf{b}_{.j}$ for $j = 1, \ldots, p$.

($\Longleftarrow$): Suppose $\mathbf{Ax}_{.j} = \mathbf{b}_{.j}$ for $j = 1, \ldots, p$. Let $\mathbf{I} = \mathbf{I}_p$ denote the identity matrix. Then

$$\mathbf{AX} = \mathbf{AXI} = \mathbf{AX}[\mathbf{e}_1, \ldots, \mathbf{e}_p] = [\mathbf{AXe}_1, \ldots, \mathbf{AXe}_p]$$
$$= [\mathbf{Ax}_{.1}, \ldots, \mathbf{Ax}_{.p}] = [\mathbf{b}_{.1}, \ldots, \mathbf{b}_{.p}] = \mathbf{B}.$$

Exercise 1.6: Block multiplication example

The product $\mathbf{AB}$ of two matrices $\mathbf{A}$ and $\mathbf{B}$ is defined precisely when the number of columns of $\mathbf{A}$ is equal to the number of rows of $\mathbf{B}$. For both sides in the equation $\mathbf{AB} = \mathbf{A}_1\mathbf{B}_1$ to make sense, both pairs $(\mathbf{A}, \mathbf{B})$ and $(\mathbf{A}_1, \mathbf{B}_1)$ need to be compatible in this way. Conversely, if the number of columns of $\mathbf{A}$ equals the number of rows of $\mathbf{B}$ and the number of columns of $\mathbf{A}_1$ equals the number of rows of $\mathbf{B}_1$, then there exists integers m, p, n, and s with $1 \leq s \leq p$ such that

$$\mathbf{A} \in \mathbb{C}^{m,p},\ \mathbf{B} \in \mathbb{C}^{p,n},\ \mathbf{A}_1 \in \mathbb{C}^{m,s},\ \mathbf{A}_2 \in \mathbb{C}^{m,p-s},\ \mathbf{B}_1 \in \mathbb{C}^{s,n}.$$

Then

$$(\mathbf{AB})_{ij} = \sum_{k=1}^{p} a_{ik}b_{kj} = \sum_{k=1}^{s} a_{ik}b_{kj} + \sum_{k=s+1}^{p} a_{ik} \cdot 0 = (\mathbf{A}_1\mathbf{B}_1)_{ij}.$$

Exercise 1.7: Another block multiplication example (TODO)

Exercise 1.14: Column oriented backsolve (TODO)

Exercise 1.17: Computing the inverse of a triangular matrix

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

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

$$[\mathbf{A}\mathbf{b}_1, \ldots, \mathbf{A}\mathbf{b}_n] = \mathbf{A}[\mathbf{b}_1, \ldots, \mathbf{b}_n] = \mathbf{A}\mathbf{B} = \mathbf{I} = [\mathbf{e}_1, \ldots, \mathbf{e}_n].$$

The kth column in this matrix equation can be partioned into blocks, as

$$\left[\begin{array}{ccc|ccc}
a_{11} & \cdots & a_{1,k} & a_{1,k+1} & \cdots & a_{1,n} \\
 & \ddots & \vdots & \vdots & \ddots & \vdots \\
 & & a_{k,k} & a_{k,k+1} & \cdots & a_{k,n} \\ \hline
 & & & a_{k+1,k+1} & \cdots & a_{k+1,n} \\
 & & & & \ddots & \vdots \\
 & & & & & a_{n,n}
\end{array}\right]
\left[\begin{array}{c}
b_{1k} \\ \vdots \\ b_{kk} \\ \hline 0 \\ \vdots \\ 0
\end{array}\right]
=
\left[\begin{array}{c}
0 \\ \vdots \\ 0 \\ 1 \\ \hline 0 \\ \vdots \\ 0
\end{array}\right].$$

Evaluating the upper block matrix multiplication then yields (1.5). By Lemma 1.9, the matrix $\mathbf{B}$ is upper triangular, implying that the other entries $b_{k+1,k}, \ldots, b_{n,k}$ in $\mathbf{b}_k$ are zero. Solving the above system thus yields the kth column of $\mathbf{B}$.

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 $\mathbf{A}$. It follows that the column $\mathbf{b}_k$ computed at step k can be stored in row (or column) k of $\mathbf{A}$ without altering the remaining computations.

Exercise 1.21: Gaussian elimination example (TODO)

Exercise 1.22: 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

$$\begin{aligned}1+2+\cdots+m &= P_m'(1)\\ &= \lim_{x\to 1}\frac{1-(m+1)x^m+mx^{m+1}}{(x-1)^2}\\ &= \lim_{x\to 1}\frac{-m(m+1)x^{m-1}+m(m+1)x^m}{2(x-1)}\\ &= \frac{1}{2}m(m+1),\end{aligned}$$

establishing (1.10). 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 (1.12). 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

$$\begin{aligned}&\lim_{x\to 1}\frac{-2+(m^2+m)x^{m-1}+2(1-m^2)x^m+(m^2-m)x^{m+1}}{(x-1)^3}\\ &=\lim_{x\to 1}\frac{(m-1)(m^2+m)x^{m-2}+2m(1-m^2)x^{m-1}+(m+1)(m^2-m)x^m}{3(x-1)^2}\\ &=\lim_{x\to 1}\frac{(m-2)(m-1)(m^2+m)x^{m-3}+2(m-1)m(1-m^2)x^{m-2}+m(m+1)(m^2-m)x^{m-1}}{6(x-1)}\\ &=\frac{1}{3}(m-1)m(m+1),\end{aligned}$$

establishing (1.13). Finally,

$$\begin{aligned}1^2+2^2+\cdots+m^2 &= \sum_{k=1}^{m}k^2=\sum_{k=1}^{m}\big((k-1)k+k\big)=\sum_{k=1}^{m}(k-1)k+\sum_{k=1}^{m}k\\ &= \frac{1}{3}(m-1)m(m+1)+\frac{1}{2}m(m+1)=\frac{1}{3}(m+1)(m+\frac{1}{2})m,\end{aligned}$$

which establishes (1.11).

Exercise 1.23: Operations (TODO)

Exercise 1.24: Multiplying triangular matrices

Computing the (i,j)-th entry of the matrix $\mathbf{AB}$ amounts to computing the inner product of the ith row $\mathbf{a}_{i:}^T$ of $\mathbf{A}$ and the jth column $\mathbf{b}_{:j}$ of $\mathbf{B}$. Because of the triangular nature of $\mathbf{A}$ and $\mathbf{B}$, only the first i entries of $\mathbf{a}_{i:}^T$ can be nonzero and only the first j entries of $\mathbf{b}_{:j}$ can be nonzero. The computation $\mathbf{a}_{i:}^T\mathbf{b}_{:j}$ 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

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

$$
\begin{aligned}
&= \sum_{i=1}^{n} \big(-i + i(i+1) + (n-i)(2i-1) \big) = \sum_{i=1}^{n} \big(-i^2 + 2ni - n + i \big) \\
&= -n^2 + (2n+1)\sum_{i=1}^{n} i - \sum_{i=1}^{n} i^2 \\
&= -n^2 + \frac{1}{2}n(n+1)(2n+1) - \frac{1}{6}n(n+1)(2n+1) \\
&= -n^2 + \frac{1}{3}n(n+1)(2n+1) = \frac{2}{3}n^3 + \frac{1}{3}n = \frac{1}{3}n(2n^2+1)
\end{aligned}
$$

arithmetic operations. A similar calculation gives the same result for the product $\mathbf{BA}$.

Exercise 1.25: Matrix formulation of Gaussian elimination

(a) By the last line of (1.7), the (i,j)-th entry a_{ij}^{k+1} of the matrix $\mathbf{A}_{k+1}$ is computed by

$$a_{ij}^{k+1} = a_{ij}^{k} - l_{ik}^{k} a_{kj}^{k}, \qquad \text{for } i \in \{k+1,\ldots,n\} \text{ and } j \in \{k,\ldots,n\}.$$

At other entries (i,j), the matrix $\mathbf{A}_{k+1}$ agrees with $\mathbf{A}_k$, so

$$a_{ij}^{k+1} = a_{ij}^{k}, \qquad \text{for } i \notin \{k+1,\ldots,n\} \quad \text{or} \quad j \notin \{k,\ldots,n\}.$$

System (1.14) is the matrix form of this linear system.

(b) Combining the relations $\mathbf{A}_{k+1} = \mathbf{M}_k \mathbf{A}_k$ for $k = 1,\ldots,n-1$, one finds

$$\mathbf{A}_n = \mathbf{M}_{n-1}\mathbf{A}_{n-1} = \mathbf{M}_{n-1}\left(\mathbf{M}_{n-2}\mathbf{A}_{n-2}\right) = \cdots = \mathbf{M}_{n-1}\cdots\mathbf{M}_1\mathbf{A}_1.$$

One can formally prove this statement using the method of induction. Since the matrices $\mathbf{M}_1,\ldots,\mathbf{M}_{n-1}$ are unit lower triangular and since the product of any unit lower triangular matrices is again unit lower triangular, the product $\mathbf{M} := \mathbf{M}_{n-1}\cdots\mathbf{M}_1$ is unit lower triangular. It follows that $\mathbf{A}_n = \mathbf{M}\mathbf{A}_1$, which proves the exercise.

(c) The matrices $\mathbf{M}_k$ are invertible, as they each have determinant 1. In fact, since

$$\left(\mathbf{I} + \begin{bmatrix}\mathbf{0}\\ \mathbf{l}_k\end{bmatrix}\mathbf{e}_k^T\right)\left(\mathbf{I} - \begin{bmatrix}\mathbf{0}\\ \mathbf{l}_k\end{bmatrix}\mathbf{e}_k^T\right) = \mathbf{I}^2 - \begin{bmatrix}\mathbf{0}\\ \mathbf{l}_k\end{bmatrix}\underbrace{\mathbf{e}_k^T\begin{bmatrix}\mathbf{0}\\ \mathbf{l}_k\end{bmatrix}}_{0}\mathbf{e}_k^T = \mathbf{I},$$

one finds

$$\mathbf{M}_k^{-1} = \left(\mathbf{I} + \begin{bmatrix}\mathbf{0}\\ \mathbf{l}_k\end{bmatrix}\mathbf{e}_k^T\right).$$

We conclude that

$$\mathbf{M}^{-1} = \mathbf{M}_1^{-1}\cdots\mathbf{M}_{n-1}^{-1} = \left(\mathbf{I} + \begin{bmatrix}\mathbf{0}\\ \mathbf{l}_1\end{bmatrix}\mathbf{e}_1^T\right)\cdots\left(\mathbf{I} + \begin{bmatrix}\mathbf{0}\\ \mathbf{l}_{n-1}\end{bmatrix}\mathbf{e}_{n-1}^T\right).$$

Exercise 1.31: Using PLU of A to solve $\mathbf{A}^T\mathbf{x} = \mathbf{b}$

If $\mathbf{A} = \mathbf{PLR}$, then $\mathbf{A}^T = \mathbf{R}^T\mathbf{L}^T\mathbf{P}^T$. The matrix $\mathbf{L}^T$ is upper triangular and the matrix $\mathbf{R}^T$ is lower triangular, implying that $\mathbf{R}^T\mathbf{L}^T$ is an LU factorization of $\mathbf{A}^T\mathbf{P}$. Since $\mathbf{A}$ is nonsingular, the matrix $\mathbf{R}^T$ must be nonsingular, and we can apply Algorithms 1.12 and 1.13 to economically solve the systems $\mathbf{R}^T\mathbf{z} = \mathbf{b}$, $\mathbf{L}^T\mathbf{y} = \mathbf{z}$, and $\mathbf{P}^T\mathbf{x} = \mathbf{y}$, to find a solution $\mathbf{x}$ to the system $\mathbf{R}^T\mathbf{L}^T\mathbf{P}^T\mathbf{x} = \mathbf{A}^T\mathbf{x} = \mathbf{b}$.

Exercise 1.32: Using PLU to compute the determinant

If $\mathbf{A} = \mathbf{PLU}$, then

$$\det(\mathbf{A}) = \det(\mathbf{PLU}) = \det(\mathbf{P}) \det(\mathbf{L}) \det(\mathbf{U})$$

and the determinant of $\mathbf{A}$ can be computed from the determinants of $\mathbf{P}$, $\mathbf{L}$, and $\mathbf{U}$. Since the latter two matrices are triangular, their determinants are simply the products of their diagonal entries. The matrix $\mathbf{P}$, 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 1.33: Using PLU to compute the inverse (TODO)

CHAPTER 2

Examples of Linear Systems

Exercise 2.6: LU factorization of 2nd derivative matrix

Let $\mathbf{L} = (l_{ij})_{ij}$, $\mathbf{U} = (r_{ij})_{ij}$ and $\mathbf{T}$ be as in the exercise. Clearly $\mathbf{L}$ is unit lower triangular and $\mathbf{U}$ is upper triangular. We compute the product $\mathbf{LU}$ 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,

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

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

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

$$(\mathbf{LU})_{i-1,i} = 1 \cdot -1 = -1, \qquad \text{for } i = 2, \dots, m;$$

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

It follows that $\mathbf{T} = \mathbf{LU}$ is an LU factorization.

Another way to show that $\mathbf{T} = \mathbf{LU}$ is by induction. For $m = 1$, one has $\mathbf{L}_1\mathbf{U}_1 = 1 \cdot 2 = \mathbf{T}_1$. Now let $m > 1$ be arbitrary and assume that $\mathbf{L}_m\mathbf{U}_m = \mathbf{T}_m$. With

$$\mathbf{a} := [0, \dots, 0, -\frac{m}{m+1}]^T, \qquad \mathbf{b} := [0, \dots, 0, -1]^T,$$

block multiplication yields

$$\mathbf{L}_{m+1}\mathbf{U}_{m+1} = \begin{bmatrix} \mathbf{L}_m & \mathbf{0} \\ \mathbf{a}^T & 1 \end{bmatrix} \begin{bmatrix} \mathbf{U}_m & \mathbf{b} \\ \mathbf{0} & \frac{m+2}{m+1} \end{bmatrix} =$$

$$\begin{bmatrix} \mathbf{T}_m & \mathbf{L}_m\mathbf{b} \\ \mathbf{a}^T\mathbf{U}_m & \mathbf{a}^T\mathbf{b} + \frac{m+2}{m+1} \end{bmatrix} = \begin{bmatrix} \mathbf{T}_m & \mathbf{b} \\ \mathbf{b}^T & 2 \end{bmatrix} = \mathbf{T}_{m+1}.$$

By induction, we can then conclude that $\mathbf{T}_m = \mathbf{L}_m\mathbf{U}_m$ for all $m \geq 1$.

Exercise 2.7: Inverse of 2nd derivative matrix

Let $\mathbf{S} = (s_{ij})_{ij}$ be defined by

$$s_{ij} = s_{ji} = \left(1 - \frac{i}{m+1}\right) j, \qquad \text{for } 1 \leq j \leq i \leq m.$$

In order to show that $\mathbf{S} = \mathbf{T}^{-1}$, we multiply $\mathbf{S}$ by $\mathbf{T}$ 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 $1 \le j < i \le m$, we find

$$\begin{aligned}
(\mathbf{ST})_{i,j} &= \sum_{k=1}^{m} s_{i,k}\mathbf{T}_{k,j} = -s_{i,j-1} + 2s_{i,j} - s_{i,j+1} \\
&= \left(1 - \frac{i}{m+1}\right)(-j + 1 + 2j - j - 1) = 0, \\
(\mathbf{ST})_{j,i} &= \sum_{k=1}^{m} s_{j,k}\mathbf{T}_{k,i} = -s_{j,i-1} + 2s_{j,i} - s_{j,i+1} \\
&= -\left(1 - \frac{i-1}{m+1}\right) j + 2\left(1 - \frac{i}{m+1}\right) j - \left(1 - \frac{i+1}{m+1}\right) j \\
&= -j + 2j - j + j \cdot \frac{i - 1 - 2i + i + 1}{m+1} = 0, \\
(\mathbf{ST})_{i,i} &= \sum_{k=1}^{m} s_{i,k}\mathbf{T}_{k,i} = -s_{i,i-1} + 2s_{i,i} - s_{i,i+1} \\
&= -\left(1 - \frac{i}{m+1}\right)(i-1) + 2\left(1 - \frac{i}{m+1}\right) i - \left(1 - \frac{i+1}{m+1}\right) i = 1
\end{aligned}$$

which means that $\mathbf{ST} = \mathbf{I}$. Moreover, since $\mathbf{S}, \mathbf{T}$, and $\mathbf{I}$ are symmetric, transposing this equation yields $\mathbf{TS} = \mathbf{I}$. We conclude that $\mathbf{S} = \mathbf{T}^{-1}$.

Exercise 2.8: Central difference approximation of 2nd derivative

If all h_i equal to the same number h, then

$$\lambda_i = \mu_i = \frac{2h}{h+h} = 1, \qquad \delta_i = \frac{y_{i+1} - y_i}{h}, \qquad \beta_i = 3(\delta_{i-1} + \delta_i) = 3\frac{y_{i+1} - y_{i-1}}{h},$$

which is what needed to be shown.

Exercise 2.9: Two point boundary value problem (TODO)

Exercise 2.10: Two point boundary value problem; computation (TODO)

Exercise 2.18: Spline evaluation (TODO)

Exercise 2.20: Bounding the moments

Let us write $\mathbf{Ax} = \mathbf{b}$ for the system (2.23), with $\mathbf{A} = (a_{ij})_{ij}$, $\mathbf{x} = [\mu_2, \ldots, \mu_n]$ and $\mathbf{b} = [b_1, \ldots, b_{n-1}]$. The matrix $\mathbf{A}$ is strictly diagonally dominant, since

$$\sigma_i := |a_{ii}| - \sum_{j \ne i} |a_{ij}| \ge 2$$

for any i. Theorem 2.4 therefore yields the bound

$$\max_{2 \le j \le n} |\mu_j| \le \max_{1 \le j \le n-1} \frac{|b_j|}{\sigma_j} \le \frac{1}{2} \max_{1 \le j \le n-1} |b_j|.$$

Exercise 2.21: Moment equations for 1st derivative boundary conditions

Since the matrix in (2.28) is strictly diagonally dominant, the system has a unique solution $[\mu_1, \mu_2, \ldots, \mu_n, \mu_{n+1}]$. Let $g(x)$ be as in (2.15), with piecewise polynomials

$$(\star) \qquad p_i(x) = c_{i1} + c_{i2}(x - x_i) + c_{i3}(x - x_i)^2 + c_{i4}(x - x_i)^3, \qquad i = 1, \ldots, n,$$

of degree at most three with coefficients

$$(\star\star) \qquad c_{i1} = y_i,\ c_{i2} = \frac{y_{i+1} - y_i}{h} - \frac{h}{3}\mu_i - \frac{h}{6}\mu_{i+1},\ c_{i3} = \frac{\mu_i}{2},\ c_{i4} = \frac{\mu_{i+1} - \mu_i}{6h},$$

for $i = 1, \ldots, n$.

To show that $g(x)$ is a C^2 *cubic spline*, we have to verify the smoothness conditions $p_{i-1}(x_i) = p_i(x_i)$, $p'_{i-1}(x_i) = p'_i(x_i)$, and $p''_{i-1}(x_i) = p''_i(x_i)$ for $i = 2, \ldots, n-1$. First let us show continuity. Evaluating p_i in $(\star)$ at $x = x_i$ and using that $c_{i1} = y_i$, it follows that $p_i(x_i) = y_i$ for $i = 1, \ldots, n$. On the other hand, evaluating the adjacent polynomials p_{i-1} at the x_i yields

$$\begin{aligned}
p_{i-1}(x_i) &= c_{i-1,1} + c_{i-1,2}h + c_{i-1,3}h^2 + c_{i-1,4}h^3 \\
&= y_{i-1} + y_i - y_{i-1} - \frac{h^2}{3}\mu_{i-1} - \frac{h^2}{6}\mu_i + \frac{h^2}{2}\mu_{i-1} + \frac{h^2}{6}(\mu_i - \mu_{i-1}) \\
&= y_i
\end{aligned}$$

for $i = 2, 3, \ldots, n+1$, implying that g is continuous and $g(x_{n+1}) = y_{n+1}$.

Secondly, let us show that g is C^1. Taking the derivative of $(\star)$ yields

$$(\star\star\star) \quad p'_i(x) = c_{i2} + 2c_{i3}(x - x_i) + 3c_{i4}(x - x_i)^2, \qquad i = 1, \ldots, n.$$

For $i = 2, 3, \ldots, n$,

$$\begin{aligned}
p'_{i-1}(x_i) - p'_i(x_i) &= c_{i-1,2} + 2c_{i-1,3}h + 3c_{i-1,4}h^2 - c_{i,2} \\
&= \frac{y_i - y_{i-1}}{h} - \frac{h}{3}\mu_{i-1} - \frac{h}{6}\mu_i + \mu_{i-1}h + \frac{h}{2}(\mu_i - \mu_{i-1}) \\
&\quad - \frac{y_{i+1} - y_i}{h} + \frac{h}{3}\mu_i + \frac{h}{6}\mu_{i+1} \\
&= \frac{h}{6}(\mu_{i-1} + 4\mu_i + \mu_{i+1}) - \frac{y_{i+1} - 2y_i + y_{i-1}}{h},
\end{aligned}$$

which is equal to zero by rows $2, 3, \ldots, n$ in (2.28). It follows that the function g has continuous first derivatives.

Thirdly, let us show that $g \in C^2$. Differentiating $(\star\star\star)$ gives

$$p''_i(x) = 2c_{i,3} + 6c_{i,4}(x - x_i), \qquad i = 1, \ldots, n,$$

from which it follows that $g''(x_i) = p''_i(x_i) = 2c_{i,3} = \mu_i$. One finds

$$p''_{i-1}(x_i) - p''_i(x_i) = 2c_{i-1,3} + 6c_{i-1,4}h - \mu_i = \mu_{i-1} + 6\frac{\mu_i - \mu_{i-1}}{6h}h - \mu_i = 0,$$

for $i = 2, 3, \ldots, n$, which implies that g has continuous second derivatives.

Now that we have established that g is a piecewise cubic polynomial of C^2 smoothness interpolating the data (x_i, y_i) for $i = 1, 2, \ldots, n+1$, let us show that it satisfies the first derivative boundary conditions (2.16). Evaluating $g'(x)$ at $x = a = x_1$ and

$x = b = x_{n+1}$ gives

$$\begin{aligned} g'(a) &= p_1'(x_1) = c_{1,2} = \frac{y_2 - y_1}{h} - \frac{h}{6}(2\mu_1 + \mu_2) = s_1, \\ g'(b) &= p_n'(x_{n+1}) = c_{n,2} + 2c_{n,3}h + 3c_{n,4}h^2 \\ &= \frac{y_{n+1} - y_n}{h} + \frac{h}{6}(\mu_n + 2\mu_{n+1}) = s_{n+1}. \end{aligned}$$

where we used the first and last row in the matrix equation (2.28).

Let us now show uniqueness of the spline g. Suppose that there are two cubic splines g_1, g_2 satisfying (2.13) and (2.16). Then the difference $g_{12} := g_1 - g_2$ is a cubic spline satisfying (2.13) and (2.16) with $y_1 = \cdots = y_n = 0$ and $s_1 = s_{n+1} = 0$. For this cubic spline, the right hand side in (2.28) is the zero vector. Since the matrix in (2.28) is nonsingular, it follows that vector $[\mu_1, \mu_2, \ldots, \mu_n, \mu_{n+1}]^T$ must be the zero vector. Equations ($\star$) and ($\star\star$) imply that g_{12} must be the zero spline, and we conclude that the splines g_1 and g_2 must be equal.

Exercise 2.22: Proof of minimal 2nd derivative property (TODO)

CHAPTER 3

LU Factorizations

Exercise 3.9: 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

$$\mathbf{L} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}, \qquad \mathbf{U} = \begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix}$$

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 3.10: 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

$$\mathbf{L} = \begin{bmatrix}1 & 0\\ 1 & 1\end{bmatrix}, \qquad \mathbf{U} = \begin{bmatrix}1 & 1\\ 0 & 0\end{bmatrix}$$

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 3.11: LU and determinant

Suppose $\mathbf{A}$ has an LU factorization $\mathbf{A} = \mathbf{LU}$. Then, by Lemma 3.5, $\mathbf{A}_{[k]} = \mathbf{L}_{[k]}\mathbf{U}_{[k]}$ 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 $\det(\mathbf{L}_{[k]}) = 1$, $\det(\mathbf{U}_{[k]}) = u_{11} \cdots u_{kk}$ and

$$\det(\mathbf{A}_{[k]}) = \det(\mathbf{L}_{[k]}\mathbf{U}_{[k]}) = \det(\mathbf{L}_{[k]}) \det(\mathbf{U}_{[k]}) = u_{11} \cdots u_{kk}$$

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

Exercise 3.12: Diagonal elements in U

From Exercise 3.11, we know that $\det(\mathbf{A}_{[k]}) = u_{11} \cdots u_{kk}$ for $k = 1, \ldots, n$. Since $\mathbf{A}$ is nonsingular, its determinant $\det(\mathbf{A}) = u_{11} \cdots u_{nn}$ is nonzero. This implies that $\det(\mathbf{A}_{[k]}) = u_{11} \cdots u_{kk} \neq 0$ for $k = 1, \ldots, n$, yielding $a_{11} = u_{11}$ for $k = 1$ and a well-defined quotient

$$\frac{\det(\mathbf{A}_{[k]})}{\det(\mathbf{A}_{[k-1]})} = \frac{u_{1,1} \cdots u_{k-1,k-1} u_{k,k}}{u_{1,1} \cdots u_{k-1,k-1}} = u_{k,k},$$

for $k = 2, \ldots, n$.

Exercise 3.16: Making a block LU into an LU (TODO)

CHAPTER 4

The Kronecker Product

Exercise 4.2: 2×2 Poisson matrix

For $m = 2$, the Poisson matrix $\mathbf{A}$ 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, $\mathbf{A}$ is strictly diagonally dominant.

Exercise 4.5: Properties of Kronecker products (TODO)

Exercise 4.14: 2nd derivative matrix is positive definite

Applying Lemma 4.11 to the case that $a = -1$ and $d = 2$, one finds that the eigenvalues λ_j of the matrix $\mathrm{tridiag}(-1, 2, -1) \in \mathbb{R}^{m,m}$ 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 $j = 1, \dots, m$. Moreover, as $|\cos(x)| < 1$ for any $x \in (0, \pi)$, it follows that $\lambda_j > 0$ for $j = 1, \dots, m$. Since, in addition, $\mathrm{tridiag}(-1, 2, -1)$ is symmetric, Lemma 3.29 implies that the matrix $\mathrm{tridiag}(-1, 2, -1)$ is symmetric positive definite.

Exercise 4.15: 1D test matrix is positive definite?

The statement of this exercise is a generalization of the statement of Exercise 4.14. Consider a matrix $M = \mathrm{tridiag}(a, d, a) \in \mathbb{R}^{m,m}$ for which $d > 0$ and $d \geq 2|a|$. By Lemma 4.11, the eigenvalues λ_j, with $j = 1, \dots, m$, 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 $\mathrm{sgn}(a)$ for the sign of a. Then

$$\lambda_j \geq 2|a| \left[1 + \frac{a}{|a|} \cos\left(\frac{j\pi}{m+1}\right)\right] = 2|a| \left[1 + \mathrm{sgn}(a) \cos\left(\frac{j\pi}{m+1}\right)\right] > 0,$$

again because $|\cos(x)| < 1$ for any $x \in (0, \pi)$. Since, in addition, M is symmetric, Lemma 3.29 implies that M is symmetric positive definite.

Exercise 4.16: Eigenvalues 2×2 for 2D test matrix

One has

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

which means that $(\lambda, \mathbf{x})$ is an eigenpair of $\mathbf{A}$. For $j = k = 1$ and $m = 2$, Theorem 4.13.1 implies that

$$\mathbf{x}_{1,1} = \mathbf{s}_1 \otimes \mathbf{s}_1 = \begin{bmatrix} \sqrt{3}/2 \\ \sqrt{3}/2 \end{bmatrix} \otimes \begin{bmatrix} \sqrt{3}/2 \\ \sqrt{3}/2 \end{bmatrix} = \begin{bmatrix} 3/4 \\ 3/4 \\ 3/4 \\ 3/4 \end{bmatrix} \propto \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \mathbf{x}.$$

Equation (4.22), 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 4.17: Nine point scheme for Poisson problem

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

$$\begin{bmatrix} v_{00} & v_{01} & v_{02} & v_{03} \\ v_{10} & & & v_{13} \\ v_{20} & & & v_{23} \\ v_{30} & v_{31} & v_{32} & v_{33} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & & & 0 \\ 0 & & & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

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

$$\begin{aligned}
&20v_{11} - 4v_{01} - 4v_{10} - 4v_{21} - 4v_{12} - v_{00} - v_{20} - v_{02} - v_{22} \\
&\qquad\qquad = \frac{1}{18}(8f_{11} + f_{01} + f_{10} + f_{21} + f_{12}), \\
&20v_{21} - 4v_{11} - 4v_{20} - 4v_{31} - 4v_{22} - v_{10} - v_{30} - v_{12} - v_{32} \\
&\qquad\qquad = \frac{1}{18}(8f_{21} + f_{11} + f_{20} + f_{31} + f_{22}), \\
&20v_{12} - 4v_{02} - 4v_{11} - 4v_{22} - 4v_{13} - v_{01} - v_{21} - v_{03} - v_{23} \\
&\qquad\qquad = \frac{1}{18}(8f_{12} + f_{02} + f_{11} + f_{22} + f_{13}), \\
&20v_{22} - 4v_{12} - 4v_{21} - 4v_{32} - 4v_{23} - v_{11} - v_{31} - v_{13} - v_{33} \\
&\qquad\qquad = \frac{1}{18}(8f_{22} + f_{12} + f_{21} + f_{32} + f_{23}),
\end{aligned}$$

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

$$\begin{aligned}
20v_{11} - 4v_{21} - 4v_{12} - v_{22} &= \frac{1}{18}(8f_{11} + f_{01} + f_{10} + f_{21} + f_{12}), \\
20v_{21} - 4v_{11} - 4v_{22} - v_{12} &= \frac{1}{18}(8f_{21} + f_{11} + f_{20} + f_{31} + f_{22}), \\
20v_{12} - 4v_{11} - 4v_{22} - v_{21} &= \frac{1}{18}(8f_{12} + f_{02} + f_{11} + f_{22} + f_{13}),
\end{aligned}$$

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

(b) For $f(x,y) = 2\pi^2 \sin(\pi x)\sin(\pi y)$, one finds

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

Substituting these values in our linear system, we obtain

$$\begin{bmatrix} 20 & -4 & -4 & -1 \\ -4 & 20 & -1 & -4 \\ -4 & -1 & 20 & -4 \\ -1 & -4 & -4 & 20 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{21} \\ v_{12} \\ v_{22} \end{bmatrix} = \frac{8+1+1}{18}\frac{3\pi^2}{2} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 5\pi^2/6 \\ 5\pi^2/6 \\ 5\pi^2/6 \\ 5\pi^2/6 \end{bmatrix}.$$

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

Exercise 4.18: Matrix equation for nine point scheme

(a) Let

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

be of equal dimensions. Implicitly assuming the boundary condition

$$(\star) \qquad v_{0,k} = v_{m+1,k} = v_{j,0} = v_{j,m+1} = 0, \qquad \text{for } j,k = 0,\ldots,m+1,$$

the (j,k)-th entry of $\mathbf{TV} + \mathbf{VT}$ can be written as

$$4v_{j,k} - v_{j-1,k} - v_{j+1,k} - v_{j,k-1} - v_{j,k+1}.$$

(Compare Equations (4.3) – (4.5).) Similarly, writing out two matrix products, the (j,k)-th entry of $\mathbf{TVT} = \mathbf{T}(\mathbf{VT})$ is found to be

$$\begin{array}{lll} -1(-1v_{j-1,k-1} & +2v_{j-1,k} & -1v_{j-1,k+1}) \\ +2(-1v_{j,k-1} & +2v_{j,k} & -1v_{j,k+1}) \\ -1(-1v_{j+1,k-1} & +2v_{j+1,k} & -1v_{j+1,k+1}) \end{array} = \begin{array}{lll} +v_{j-1,k-1} & -2v_{j-1,k} & +v_{j-1,k+1} \\ -2v_{j,k-1} & +4v_{j,k} & -2v_{j,k+1} \\ +v_{j+1,k-1} & -2v_{j+1,k} & +v_{j+1,k+1} \end{array}.$$

Together, these observations yield that the System (4.24) is equivalent to $(\star)$ and

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

(b) It is a direct consequence of Lemma 4.10 that this equation can be rewritten to one of the form $\mathbf{Ax} = \mathbf{b}$, where

$$\mathbf{A} = \mathbf{T}\otimes\mathbf{I} + \mathbf{I}\otimes\mathbf{T} - \frac{1}{6}\mathbf{T}\otimes\mathbf{T}, \quad \mathbf{x} = \mathrm{vec}(\mathbf{V}), \quad \mathbf{b} = h^2\mathrm{vec}(\mu\mathbf{F}).$$

Exercise 4.19: Biharmonic equation

(a) Writing $v = -\nabla^2 u$, the second line in Equation (4.26) is equivalent to

$$u(s,t) = v(s,t) = 0, \qquad \text{for } (s,t) \in \partial\Omega,$$

while the first line is equivalent to

$$f(s,t) = \nabla^4 u(s,t) = \nabla^2\big(\nabla^2 u(s,t)\big) = -\nabla^2 v(s,t), \qquad \text{for } (s,t) \in \Omega.$$

(b) By Lemma 4.10,

$$(\mathbf{A} \otimes \mathbf{B})\mathrm{vec}(\mathbf{V}) = \mathrm{vec}(\mathbf{F}) \iff \mathbf{A}\mathbf{V}\mathbf{B}^T = \mathbf{F},$$

$$(\mathbf{A} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{B})\mathrm{vec}(\mathbf{V}) = \mathrm{vec}(\mathbf{F}) \iff \mathbf{A}\mathbf{V} + \mathbf{V}\mathbf{B}^T = \mathbf{F},$$

whenever $\mathbf{A} \in \mathbb{R}^{r,r}, \mathbf{B} \in \mathbb{R}^{s,s}, \mathbf{F}, \mathbf{V} \in \mathbb{R}^{r,s}$ (the identity matrices are assumed to be of the appropriate dimensions). Using $\mathbf{T} = \mathbf{T}^T$, these equations imply that

$$\mathbf{T}\mathbf{V} + \mathbf{V}\mathbf{T} = h^2\mathbf{F} \iff (\mathbf{T} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{T})\mathrm{vec}(\mathbf{V}) = h^2\mathrm{vec}(\mathbf{F}),$$

$$\mathbf{T}\mathbf{U} + \mathbf{U}\mathbf{T} = h^2\mathbf{V} \iff (\mathbf{T} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{T})\mathrm{vec}(\mathbf{U}) = h^2\mathrm{vec}(\mathbf{V}).$$

Substituting the equation for $\mathrm{vec}(\mathbf{V})$ into the equation for $\mathrm{vec}(\mathbf{F})$, one obtains the equation

$$\mathbf{A}\mathrm{vec}(\mathbf{U}) = h^4\mathrm{vec}(\mathbf{F}), \qquad \text{where } \mathbf{A} := (\mathbf{T} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{T})^2,$$

which is a linear system of m^2 equations.

(c) The equations $h^2\mathbf{V} = (\mathbf{T}\mathbf{U} + \mathbf{U}\mathbf{T})$ and $\mathbf{T}\mathbf{V} + \mathbf{V}\mathbf{T} = h^2\mathbf{F}$ together yield the normal form

$$\mathbf{T}(\mathbf{T}\mathbf{U} + \mathbf{U}\mathbf{T}) + (\mathbf{T}\mathbf{U} + \mathbf{U}\mathbf{T})\mathbf{T} = \mathbf{T}^2\mathbf{U} + 2\mathbf{T}\mathbf{U}\mathbf{T} + \mathbf{U}\mathbf{T}^2 = h^4\mathbf{F}.$$

The vector form is given in (b). Using the distributive property of matrix multiplication and the mixed product rule of Lemma 4.6, the matrix $\mathbf{A} = (\mathbf{T} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{T})^2$ can be rewritten as

$$\begin{aligned}\mathbf{A} &= (\mathbf{T} \otimes \mathbf{I})(\mathbf{T} \otimes \mathbf{I}) + (\mathbf{T} \otimes \mathbf{I})(\mathbf{I} \otimes \mathbf{T}) + (\mathbf{I} \otimes \mathbf{T})(\mathbf{T} \otimes \mathbf{I}) + (\mathbf{I} \otimes \mathbf{T})(\mathbf{I} \otimes \mathbf{T})\\ &= \mathbf{T}^2 \otimes \mathbf{I} + 2\mathbf{T} \otimes \mathbf{T} + \mathbf{I} \otimes \mathbf{T}^2.\end{aligned}$$

Writing $\mathbf{x} := \mathrm{vec}(U)$ and $\mathbf{b} := h^4\mathrm{vec}(\mathbf{F})$, the linear system of (b) can be written as $\mathbf{A}\mathbf{x} = \mathbf{b}$.

(d) Since $\mathbf{T}$ and $\mathbf{I}$ are symmetric positive definite, Lemma 4.9.3 implies that $\mathbf{M} := (\mathbf{T} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{T})$ is symmetric positive definite as well. The square of any symmetric positive definite matrix is symmetric positive definite as well, implying that $\mathbf{A} = \mathbf{M}^2$ is symmetric positive definite. Let us now show this more directly by calculating the eigenvalues of $\mathbf{A}$.

By Lemma 4.11, we know the eigenpairs $(\lambda_i, \mathbf{s}_i)$, where $i = 1, \ldots, m$ of the matrix $\mathbf{T}$. By Lemma 4.8, it follows that the eigenpairs of $\mathbf{M}$ are $(\lambda_i + \lambda_j, \mathbf{s}_i \otimes \mathbf{s}_j)$ for $i, j = 1, \ldots, m$. If $\mathbf{B}$ is any matrix with eigenpairs $(\mu_i, \mathbf{v}_i)$, where $i = 1, \ldots, m$, then $\mathbf{B}^2$ has eigenpairs $(\mu_i^2, \mathbf{v}_i)$, as

$$\mathbf{B}^2\mathbf{v}_i = \mathbf{B}(\mathbf{B}\mathbf{v}_i) = \mathbf{B}(\mu_i\mathbf{v}_i) = \mu_i(\mathbf{B}\mathbf{v}_i) = \mu_i^2\mathbf{v}_i, \qquad \text{for } i = 1, \ldots, m.$$

It follows that $\mathbf{A} = \mathbf{M}^2$ has eigenpairs $\big((\lambda_i + \lambda_j)^2, \mathbf{s}_i \otimes \mathbf{s}_j\big)$, for $i, j = 1, \ldots, m$. (Note that we can verify this directly by multiplying $\mathbf{A}$ by $\mathbf{s}_i \otimes \mathbf{s}_j$ and using the mixed product rule.) Since the λ_i are positive, the eigenvalues of $\mathbf{A}$ are positive. We conclude that $\mathbf{A}$ is symmetric positive definite.

Writing $\mathbf{A} = \mathbf{T}^2 \otimes \mathbf{I} + 2\mathbf{T} \otimes \mathbf{T} + \mathbf{I} \otimes \mathbf{T}^2$ and computing the block structure of each of these terms, one finds that $\mathbf{A}$ 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 5

Fast Direct Solution of a Large Linear System

Exercise 5.5: Fourier matrix

By Equation (5.6), the Fourier matrix $\mathbf{F}_N$ has entries

$$(\mathbf{F}_N)_{j,k} = \omega_N^{(j-1)(k-1)}, \qquad \omega_N := e^{-\frac{2\pi}{N}i} = \cos\left(\frac{2\pi}{N}\right) - i\sin\left(\frac{2\pi}{N}\right).$$

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

$$\mathbf{F}_4 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{bmatrix}.$$

Computing the transpose and Hermitian transpose gives

$$\mathbf{F}_4^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{bmatrix} = \mathbf{F}_4, \qquad \mathbf{F}_4^H = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{bmatrix} \neq \mathbf{F}_4,$$

which is what needed to be shown.

Exercise 5.6: Sine transform as Fourier transform

According to Lemma 5.2, the Discrete Sine Transform can be computed from the Discrete Fourier Transform by $(\mathbf{S}_m\mathbf{x})_k = \frac{i}{2}(\mathbf{F}_{2m+2}\mathbf{z})_{k+1}$, where

$$\mathbf{z} = [0, x_1, \ldots, x_m, 0, -x_m, \ldots, -x_1]^T.$$

For $m = 1$ this means that

$$\mathbf{z} = [0, x_1, 0, -x_1]^T \quad \text{and} \quad \mathbf{S}_1 x_1 = \frac{i}{2}(\mathbf{F}_4\mathbf{z})_2.$$

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

$$\mathbf{S}_1 x_1 = \sin(\pi h)x_1 = \sin\left(\frac{\pi}{2}\right)x_1 = x_1,$$

while computing the Fourier transform gives

$$\mathbf{F}_4\mathbf{z} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{bmatrix} \begin{bmatrix} 0 \\ x_1 \\ 0 \\ -x_1 \end{bmatrix} = \begin{bmatrix} 0 \\ -2ix_1 \\ 0 \\ 2ix_1 \end{bmatrix}.$$

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

$$\mathbf{S}_1 x_1 = x_1 = \frac{i}{2}(\mathbf{F}_4\mathbf{z})_2,$$

which is what needed to be shown.

Exercise 5.7: Explicit solution of the discrete Poisson equation

For any integer $m \geq 1$, let $h = 1/(m+1)$. For $j = 1, \ldots, m$, let $\lambda_j = 4\sin^2\big(j\pi h/2\big)$, $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$, and $\mathbf{S} = (s_{jk})_{jk} = \big(\sin(jk\pi h)\big)_{jk}$. By Section 5.2, the solution to the discrete Poisson equation is $\mathbf{V} = \mathbf{SXS}$, where $\mathbf{X}$ is found by solving $\mathbf{DX}+\mathbf{XD} = 4h^4\mathbf{SFS}$. Since $\mathbf{D}$ is diagonal, one has

$$x_{pr} = 4h^4 \frac{(\mathbf{SFS})_{pr}}{\lambda_p + \lambda_r} = 4h^4 \sum_{k=1}^{m}\sum_{l=1}^{m} \frac{s_{pk}f_{kl}s_{lr}}{\lambda_p + \lambda_r}$$

so that

$$\begin{aligned} v_{ij} &= \sum_{p=1}^{m}\sum_{r=1}^{m} s_{ip}x_{pr}s_{rj} = 4h^4 \sum_{p=1}^{m}\sum_{r=1}^{m}\sum_{k=1}^{m}\sum_{l=1}^{m} \frac{s_{ip}s_{pk}s_{lr}s_{rj}}{\lambda_p + \lambda_r} f_{kl} \\ &= h^4 \sum_{p=1}^{m}\sum_{r=1}^{m}\sum_{k=1}^{m}\sum_{l=1}^{m} \frac{\sin\left(\frac{ip\pi}{m+1}\right)\sin\left(\frac{pk\pi}{m+1}\right)\sin\left(\frac{lr\pi}{m+1}\right)\sin\left(\frac{rj\pi}{m+1}\right)}{\sin^2\left(\frac{p\pi}{2(m+1)}\right) + \sin^2\left(\frac{r\pi}{2(m+1)}\right)} f_{kl}, \end{aligned}$$

which is what needed to be shown.

Exercise 5.8: Improved version of Algorithm 5.1

Given is that

$(\star)$ $\qquad \mathbf{TV} + \mathbf{VT} = h^2\mathbf{F}.$

Let $\mathbf{T} = \mathbf{SDS}^{-1}$ be the orthogonal diagonalization of $\mathbf{T}$ from Equation (5.4), and write $\mathbf{X} = \mathbf{VS}$ and $\mathbf{C} = h^2\mathbf{FS}$.

(a) Multiplying Equation $(\star)$ from the right by $\mathbf{S}$, one obtains

$$\mathbf{TX} + \mathbf{XD} = \mathbf{TVS} + \mathbf{VSD} = \mathbf{TVS} + \mathbf{VTS} = h^2\mathbf{FS} = \mathbf{C}.$$

(b) Writing $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_m]$, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]$ and applying the rules of block multiplication, we find

$$\begin{aligned} [\mathbf{c}_1, \ldots, \mathbf{c}_m] &= \mathbf{C} \\ &= \mathbf{TX} + \mathbf{XD} \\ &= \mathbf{T}[\mathbf{x}_1, \ldots, \mathbf{x}_m] + \mathbf{X}[\lambda_1\mathbf{e}_1, \ldots, \lambda_m\mathbf{e}_m] \\ &= [\mathbf{Tx}_1 + \lambda_1\mathbf{Xe}_1, \ldots, \mathbf{Tx}_m + \lambda_m\mathbf{Xe}_m] \\ &= [\mathbf{Tx}_1 + \lambda_1\mathbf{x}_1, \ldots, \mathbf{Tx}_m + \lambda_m\mathbf{x}_m] \\ &= [(\mathbf{T} + \lambda_1\mathbf{I})\mathbf{x}_1, \ldots, (\mathbf{T} + \lambda_m\mathbf{I})\mathbf{x}_m], \end{aligned}$$

which is equivalent to System (5.10). To find $\mathbf{X}$, we therefore need to solve the m tridiagonal linear systems of (5.10). Since the eigenvalues $\lambda_1, \ldots, \lambda_m$ are positive, each matrix $\mathbf{T} + \lambda_j\mathbf{I}$ is diagonally dominant. By Theorem 2.5, every such matrix is nonsingular and has a unique LU factorization. Algorithms 2.1 and 2.2 then solve the corresponding system $(\mathbf{T} + \lambda_j\mathbf{I})\mathbf{x}_j = \mathbf{c}_j$ in $O(\delta m)$ operations for some constant δ. Doing this for all m columns $\mathbf{x}_1, \ldots, \mathbf{x}_m$, one finds the matrix $\mathbf{X}$ in $O(\delta m^2)$ operations.

(c) To find $\mathbf{V}$, we first find $\mathbf{C} = h^2\mathbf{FS}$ by performing $O(2m^3)$ operations. Next we find $\mathbf{X}$ as in step b) by performing $O(\delta m^2)$ operations. Finally we compute $\mathbf{V} = 2h\mathbf{XS}$ by performing $O(2m^3)$ operations. In total, this amounts to $O(4m^3)$ operations.

(d) As explained in Section 5.3, multiplying by the matrix $\mathbf{S}$ 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 5.9: Fast solution of 9 point scheme

Analogously to Section 5.2, we use the relations between the matrices $\mathbf{T}, \mathbf{S}, \mathbf{X}, \mathbf{D}$ to rewrite Equation (4.25).

$$\begin{aligned}
& \mathbf{TV} + \mathbf{VT} - \frac{1}{6}\mathbf{TVT} = h^2\mu\mathbf{F} \\
\iff \quad & \mathbf{TSXS} + \mathbf{SXST} - \frac{1}{6}\mathbf{TSXST} = h^2\mu\mathbf{F} \\
\iff \quad & \mathbf{STSXS}^2 + \mathbf{S}^2\mathbf{XSTS} - \frac{1}{6}\mathbf{STSXSTS} = h^2\mu\mathbf{SFS} \\
\iff \quad & \mathbf{S}^2\mathbf{DXS}^2 + \mathbf{S}^2\mathbf{XS}^2\mathbf{D} - \frac{1}{6}\mathbf{S}^2\mathbf{DXS}^2\mathbf{D} = h^2\mu\mathbf{SFS} \\
\iff \quad & \mathbf{DX} + \mathbf{XD} - \frac{1}{6}\mathbf{DXD} = 4h^4\mu\mathbf{SFS} = 4h^4\mathbf{G}
\end{aligned}$$

Writing $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$, the (j,k)-th entry of $\mathbf{DX} + \mathbf{XD} - \frac{1}{6}\mathbf{DXD}$ 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

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

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

$$\begin{aligned}
\sigma_j + \sigma_k - \frac{2}{3}\sigma_j\sigma_k &> \sigma_j + \sigma_k - \sigma_j\sigma_k \\
&= 2 - \cos^2\alpha - \cos^2\beta - (1 - \cos^2\alpha)(1 - \cos^2\beta) \\
&= 1 - \cos^2\alpha\cos^2\beta \\
&\geq 1 - \cos^2\beta \\
&\geq 0.
\end{aligned}$$

Let $\mathbf{A} = \mathbf{T} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{T} - \frac{1}{6}\mathbf{T} \otimes \mathbf{T}$ be as in Exercise 4.18.(b) and $\mathbf{s}_i$ as in Section 5.2. Applying the mixed-product rule, one obtains

$$\mathbf{A}(\mathbf{s}_i \otimes \mathbf{s}_j) = (\mathbf{T} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{T})(\mathbf{s}_i \otimes \mathbf{s}_j) - \frac{1}{6}(\mathbf{T} \otimes \mathbf{T})(\mathbf{s}_i \otimes \mathbf{s}_j) =$$

$$(\lambda_i + \lambda_j)(\mathbf{s}_i \otimes \mathbf{s}_j) - \frac{1}{6}\lambda_i\lambda_j(\mathbf{s}_i \otimes \mathbf{s}_j) = (\lambda_i + \lambda_j - \frac{1}{6}\lambda_i\lambda_j)(\mathbf{s}_i \otimes \mathbf{s}_j).$$

The matrix $\mathbf{A}$ therefore has eigen vectors $\mathbf{s}_i \otimes \mathbf{s}_j$, 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 $\mathbf{A}$ is positive definite. It follows that the System (4.24) always has a (unique) solution.

Exercise 5.10: Algorithm for fast solution of 9 point scheme

The following describes an algorithm for solving System (4.24).

Algorithm 1 A method for solving the discrete Poisson problem (4.24)

```
Require: An integer $m$ denoting the grid size, a matrix $\mu\mathbf{F} \in \mathbb{R}^{m,m}$ of function values.
Ensure: The solution $\mathbf{V}$ to the discrete Poisson problem (4.24).
1: $h \leftarrow \frac{1}{m+1}$
2: $\mathbf{S} \leftarrow \big(\sin(jk\pi h)\big)_{j,k=1}^{m}$
3: $\sigma \leftarrow \big(\sin^2\big(\frac{j\pi h}{2}\big)\big)_{j=1}^{m}$
4: $\mathbf{G} \leftarrow \mathbf{S}\mu\mathbf{F}\mathbf{S}$
5: $\mathbf{X} \leftarrow \Big(\frac{h^4 g_{i,j}}{\sigma_i+\sigma_j-\frac{2}{3}\sigma_i\sigma_j}\Big)_{j,k=1}^{m}$
6: $\mathbf{V} \leftarrow \mathbf{S}\mathbf{X}\mathbf{S}$
```

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

step	1	2	3	4	5	6
complexity	$\mathcal{O}(1)$	$\mathcal{O}(m^2)$	$\mathcal{O}(m)$	$\mathcal{O}(m^3)$	$\mathcal{O}(m^2)$	$\mathcal{O}(m^3)$

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

Exercise 5.11: Fast solution of biharmonic equation

From Exercise 4.19 we know that $\mathbf{T} \in \mathbb{R}^{m\times m}$ is the second derivative matrix. According to Lemma 4.11, the eigenpairs $(\lambda_j, \mathbf{s}_j)$, with $j = 1,\dots,m$, of $\mathbf{T}$ are given by

$$\begin{aligned}\mathbf{s}_j &= [\sin(j\pi h), \sin(2j\pi h), \dots, \sin(mj\pi h)]^T,\\ \lambda_j &= 2 - 2\cos(j\pi h) = 4\sin^2(j\pi h/2),\end{aligned}$$

and satisfy $\mathbf{s}_j^T\mathbf{s}_k = \delta_{j,k}/(2h)$ for all j,k, where $h := 1/(m+1)$. Using, in order, that $\mathbf{U} = \mathbf{SXS}$, $\mathbf{TS} = \mathbf{SD}$, and $\mathbf{S}^2 = \mathbf{I}/(2h)$, one finds that

$$\begin{aligned} & h^4\mathbf{F} = \mathbf{T}^2\mathbf{U} + 2\mathbf{TUT} + \mathbf{UT}^2\\ \iff\quad & h^4\mathbf{F} = \mathbf{T}^2\mathbf{SXS} + 2\mathbf{TSXST} + \mathbf{SXST}^2\\ \iff\quad & h^4\mathbf{SFS} = \mathbf{ST}^2\mathbf{SXS}^2 + 2\mathbf{STSXSTS} + \mathbf{S}^2\mathbf{XST}^2\mathbf{S}\\ \iff\quad & h^4\mathbf{SFS} = \mathbf{S}^2\mathbf{D}^2\mathbf{XS}^2 + 2\mathbf{S}^2\mathbf{DXS}^2\mathbf{D} + \mathbf{S}^2\mathbf{XS}^2\mathbf{D}^2\\ \iff\quad & h^4\mathbf{SFS} = \mathbf{ID}^2\mathbf{XI}/(4h^2) + 2\mathbf{IDXID}/(4h^2) + \mathbf{IXID}^2/(4h^2)\\ \iff\quad & 4h^6\mathbf{G} = \mathbf{D}^2\mathbf{X} + 2\mathbf{DXD} + \mathbf{XD}^2,\end{aligned}$$

where $\mathbf{G} := \mathbf{SFS}$. The (j,k)-th entry of the latter matrix equation is

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

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

$$x_{jk} = \frac{4h^6 g_{jk}}{(\lambda_j+\lambda_k)^2} = \frac{4h^6 g_{jk}}{\big(4\sin^2(j\pi h/2) + 4\sin^2(k\pi h/2)\big)^2} = \frac{h^6 g_{jk}}{4(\sigma_j+\sigma_k)^2}.$$

Exercise 5.12: Algorithm for fast solution of biharmonic equation

In order to derive an algorithm that computes $\mathbf{U}$ in Problem 4.19, we can adjust Algorithm 5.1 by replacing the computation of the matrix $\mathbf{X}$ by the formula from Exercise 5.11. This adjustment does not change the complexity of Algorithm 5.1, which therefore remains $\mathcal{O}(\delta n^{3/2})$. The new algorithm can be implemented in Matlab as in Listing 5.1.

Listing 5.1. A simple fast solution to the biharmonic equation

```
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 5.13: Check algorithm for fast solution of biharmonic equation

The Matlab function from Listing 5.2 directly solves the standard form $\mathbf{Ax} = \mathbf{b}$ of Equation (4.28), making sure to return a matrix of the same dimension as the implementation from Listing 5.1.

Listing 5.2. A direct solution to the biharmonic equation

```
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
```

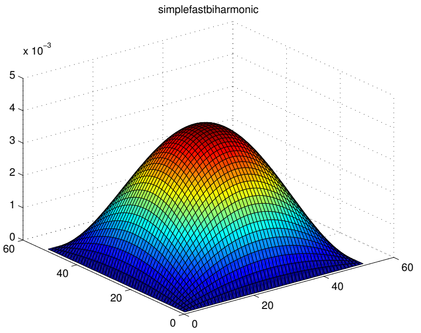
After specifying $m = 4$ by issuing the command `F = ones(4,4)`, the commands `simplefastbiharmonic(F)` and `standardbiharmonic(F)` both return the matrix

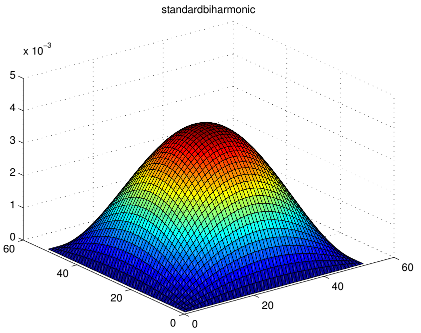
$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.0015 & 0.0024 & 0.0024 & 0.0015 & 0 \\ 0 & 0.0024 & 0.0037 & 0.0037 & 0.0024 & 0 \\ 0 & 0.0024 & 0.0037 & 0.0037 & 0.0024 & 0 \\ 0 & 0.0015 & 0.0024 & 0.0024 & 0.0015 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

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 5.3.

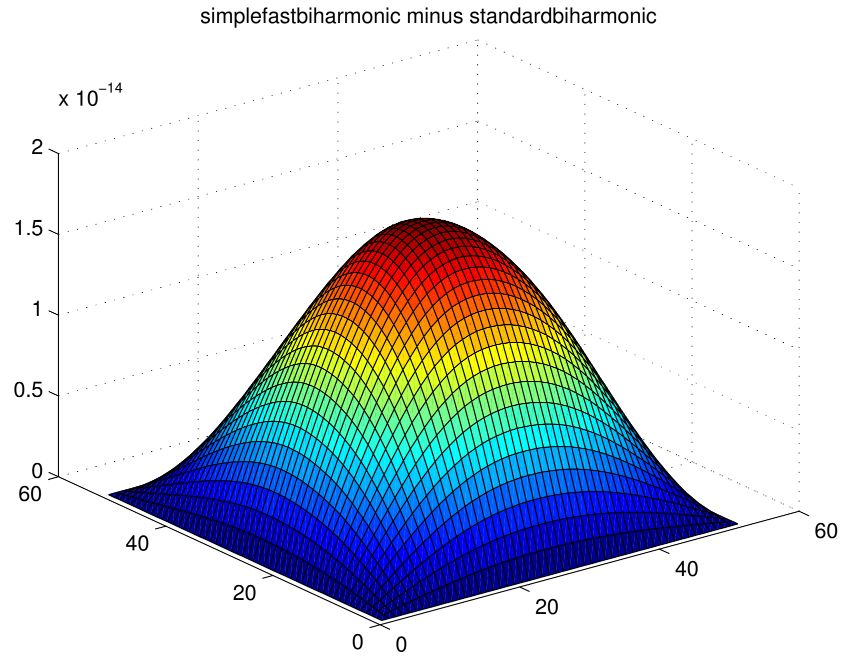
Listing 5.3. Solving the biharmonic equation and plotting the result

```
F = ones(50, 50);
U = simplefastbiharmonic(F);
V = standardbiharmonic(F);
surf(U);
surf(V);
```





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 `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 5.14: Fast solution of biharmonic equation using 9 point rule (TODO)

CHAPTER 6

Matrix Reduction by Similarity Transformations

Exercise 6.5: Unitary matrix

Suppose $\|\mathbf{U}\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ for all $\mathbf{x} \in \mathbb{C}^n$. Applying Equation (25) twice, one finds

$$\begin{aligned}
&4\,\langle \mathbf{U}\mathbf{x}, \mathbf{U}\mathbf{y}\rangle \\
&= \|\mathbf{U}(\mathbf{x}+\mathbf{y})\|_2^2 - \|\mathbf{U}(\mathbf{x}-\mathbf{y})\|_2^2 + i\|\mathbf{U}(\mathbf{x}-i\mathbf{y})\|_2^2 - i\|\mathbf{U}(\mathbf{x}+i\mathbf{y})\|_2^2 \\
&= \|\mathbf{x}+\mathbf{y}\|_2^2 - \|\mathbf{x}-\mathbf{y}\|_2^2 + i\|\mathbf{x}-i\mathbf{y}\|_2^2 - i\|\mathbf{x}+i\mathbf{y}\|_2^2 \\
&= 4\,\langle \mathbf{x}, \mathbf{y}\rangle
\end{aligned}$$

for any $\mathbf{x}, \mathbf{y}$ in $\mathbb{C}^n$. It follows that $\langle \mathbf{U}\mathbf{x}, \mathbf{U}\mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{y}\rangle$ for any $\mathbf{x}, \mathbf{y}$ in $\mathbb{C}^n$.

Exercise 6.23: Find eigenpair example

As $\mathbf{A}$ 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 $\mathbf{A}\mathbf{x}_1 = \lambda_1\mathbf{x}_1$ and $\mathbf{A}\mathbf{x}_2 = \lambda_2\mathbf{x}_2$, one finds (valid choices of) eigenpairs, for instance

$$(\lambda_1, \mathbf{x}_1) = (1, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}), \qquad (\lambda_2, \mathbf{x}_2) = (2, \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}).$$

Exercise 6.24: Idempotent matrix

Suppose that $(\lambda, \mathbf{x})$ is an eigenpair of a matrix $\mathbf{A}$ satisfying $\mathbf{A}^2 = \mathbf{A}$. Then

$$\lambda\mathbf{x} = \mathbf{A}\mathbf{x} = \mathbf{A}^2\mathbf{x} = \lambda\mathbf{A}\mathbf{x} = \lambda^2\mathbf{x}.$$

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 6.25: Nilpotent matrix

Suppose that $(\lambda, \mathbf{x})$ is an eigenpair of a matrix $\mathbf{A}$ satisfying $\mathbf{A}^k = \mathbf{0}$ for some natural number k. Then

$$\mathbf{0} = \mathbf{A}^k\mathbf{x} = \lambda\mathbf{A}^{k-1}\mathbf{x} = \lambda^2\mathbf{A}^{k-2}\mathbf{x} = \cdots = \lambda^k\mathbf{x}.$$

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 6.26: Eigenvalues of a unitary matrix

Let $\mathbf{x}$ be an eigenvector corresponding to λ. Then $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ and, as a consequence, $\mathbf{x}^*\mathbf{A}^* = \mathbf{x}^*\overline{\lambda}$. To use that $\mathbf{A}^*\mathbf{A} = \mathbf{I}$, it is tempting to multiply the left hand sides of these equations, yielding

$$|\lambda|^2\|\mathbf{x}\|^2 = \mathbf{x}^*\overline{\lambda}\lambda\mathbf{x} = \mathbf{x}^*\mathbf{A}^*\mathbf{A}\mathbf{x} = \mathbf{x}^*\mathbf{I}\mathbf{x} = \|\mathbf{x}\|^2.$$

Since $\mathbf{x}$ is an eigenvector, it must be nonzero. Nonzero vectors have nonzero norms, and we can therefore divide the above equation by $\|\mathbf{x}\|^2$, 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 6.27: Nonsingular approximation of a singular matrix

Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of the matrix $\mathbf{A}$. As the matrix $\mathbf{A}$ is singular, its determinant $\det(\mathbf{A}) = \lambda_1 \cdots \lambda_n$ is zero, implying that one of its eigenvalues is zero. If all the eigenvalues of $\mathbf{A}$ 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, $\det(\mathbf{A} - \lambda\mathbf{I})$ is zero for $\lambda = \lambda_1, \ldots, \lambda_n$, and nonzero otherwise. In particular $\det(\mathbf{A} - \varepsilon\mathbf{I})$ is nonzero for any $\varepsilon \in (0, \varepsilon_0)$, and $\mathbf{A} - \varepsilon\mathbf{I}$ will be nonsingular in this interval. This is what we needed to prove.

Exercise 6.28: Companion matrix

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

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

By the rules of determinant evaluation, we can substract 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

$$\pi_\mathbf{A}(\lambda) = \det\begin{bmatrix} -q_{n-1} - \lambda & -q_{n-2} & \cdots & -q_1 & -f(\lambda) \\ 1 & -\lambda & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} = (-1)^n f(\lambda),$$

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 6.32: Schur decomposition example

The matrix $\mathbf{U}$ is unitary, as $\mathbf{U}^*\mathbf{U} = \mathbf{U}^T\mathbf{U} = \mathbf{I}$. One directly verifies that

$$\mathbf{R} := \mathbf{U}^T\mathbf{A}\mathbf{U} = \begin{bmatrix} -1 & -1 \\ 0 & 4 \end{bmatrix}.$$

Since this matrix is upper triangular, $\mathbf{A} = \mathbf{U}\mathbf{R}\mathbf{U}^T$ is a Schur decomposition of $\mathbf{A}$.

Exercise 6.35: Skew-Hermitian matrix

By definition, a matrix $\mathbf{C}$ is *skew-Hermitian* if $\mathbf{C}^* = -\mathbf{C}$.

"$\Longrightarrow$": Suppose that $\mathbf{C} = \mathbf{A} + i\mathbf{B}$, with $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m,m}$, is skew-Hermitian. Then

$$-\mathbf{A} - i\mathbf{B} = -\mathbf{C} = \mathbf{C}^* = (\mathbf{A} + i\mathbf{B})^* = \mathbf{A}^T - i\mathbf{B}^T,$$

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

"$\Longleftarrow$": Suppose that we are given matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m,m}$ such that $\mathbf{A}$ is skew-Hermitian and $\mathbf{B}$ is real symmetric. Let $\mathbf{C} = \mathbf{A} + i\mathbf{B}$. Then

$$\mathbf{C}^* = (\mathbf{A} + i\mathbf{B})^* = \mathbf{A}^T - i\mathbf{B}^T = -\mathbf{A} - i\mathbf{B} = -(\mathbf{A} + i\mathbf{B}) = -\mathbf{C},$$

meaning that $\mathbf{C}$ is skew-Hermitian.

Exercise 6.36: Eigenvalues of a skew-Hermitian matrix

Let $\mathbf{A}$ be a skew-Hermitian matrix and consider a Schur triangularization $\mathbf{A} = \mathbf{U}\mathbf{R}\mathbf{U}^*$ of $\mathbf{A}$. Then

$$\mathbf{R} = \mathbf{U}^*\mathbf{A}\mathbf{U} = \mathbf{U}^*(-\mathbf{A}^*)\mathbf{U} = -\mathbf{U}^*\mathbf{A}^*\mathbf{U} = -(\mathbf{U}^*\mathbf{A}\mathbf{U})^* = -\mathbf{R}^*.$$

Since $\mathbf{R}$ differs from $\mathbf{A}$ by a similary transform, their eigenvalues coincide (use the multiplicative property of the determinant to show that

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \det(\mathbf{U}^*)\det(\mathbf{U}\mathbf{R}\mathbf{U}^* - \lambda\mathbf{I}))\det(\mathbf{U}) = \det(\mathbf{R} - \lambda\mathbf{I}).)$$

As $\mathbf{R}$ is a triangular matrix, its eigenvalues λ_i appear on its diagonal. From the equation $\mathbf{R} = -\mathbf{R}^*$ it then follows that $\lambda_i = -\overline{\lambda}_i$, implying that each λ_i is purely imaginary.

Exercise 6.47: Eigenvalue perturbation for Hermitian matrices

As proved in Theorem 3.26, a positive semidefinite matrix has no negative eigenvalues, implying that $\beta_n \geq 0$. It immediately follows from $\alpha_i + \beta_n \leq \gamma_i$ that in this case $\gamma_i \geq \alpha_i$.

Exercise 6.49: Hoffman-Wielandt

The matrix $\mathbf{A}$ has eigenvalues 0 and 4, and the matrix $\mathbf{B}$ 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

$$16 = \sum_{j=1}^{n} |\mu_{i_j} - \lambda_j|^2 \leq \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij} - b_{ij}|^2 = 12,$$

which clearly cannot be valid. The Hoffman-Wielandt Theorem cannot be applied to these matrices, because $\mathbf{B}$ is not normal,

$$\mathbf{B}^H\mathbf{B} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \neq \begin{bmatrix} 2 & -2 \\ -2 & 2 \end{bmatrix} = \mathbf{B}\mathbf{B}^H.$$

Exercise 6.54: Biorthogonal expansion (TODO)

Exercise 6.55: Generalized Rayleigh quotient

Suppose $(\lambda, \mathbf{x})$ is a right eigenpair for $\mathbf{A}$, so that $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$. Then the generalized Rayleight quotient for $\mathbf{A}$ is

$$R(\mathbf{y}, \mathbf{x}) := \frac{\mathbf{y}^*\mathbf{A}\mathbf{x}}{\mathbf{y}^*\mathbf{x}} = \frac{\mathbf{y}^*\lambda\mathbf{x}}{\mathbf{y}^*\mathbf{x}} = \lambda,$$

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

$$R(\mathbf{y}, \mathbf{x}) := \frac{\mathbf{y}^*\mathbf{A}\mathbf{x}}{\mathbf{y}^*\mathbf{x}} = \frac{\lambda\mathbf{y}^*\mathbf{x}}{\mathbf{y}^*\mathbf{x}} = \lambda.$$

Exercise 6.58: Jordan example (TODO)

Exercise 6.59: Big Jordan example (TODO)

Exercise 6.61: Jordan block example (TODO)

Exercise 6.62: Powers of a Jordan block (TODO)

Exercise 6.64: Minimal polynomial example (TODO)

Exercise 6.65: Similar matrix polynomials (TODO)

Exercise 6.66: Minimal polynomial of a diagonalizable matrix (TODO)

CHAPTER 7

The Singular Value Decomposition

Exercise 7.11: SVD examples

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

$$\mathbf{A} = \begin{bmatrix}\mathbf{U}_1 & \mathbf{U}_2\end{bmatrix}\begin{bmatrix}5\\0\end{bmatrix}\begin{bmatrix}\mathbf{V}_1\end{bmatrix}, \qquad \text{with } \mathbf{U}_1, \mathbf{U}_2 \in \mathbb{R}^{2,1},\ \mathbf{V} = \mathbf{V}_1 \in \mathbb{R}.$$

The eigenvector of $\mathbf{A}^T\mathbf{A}$ that corresponds to the eigenvalue $\lambda_1 = 25$ is given by $\mathbf{v}_1 = 1$, providing us with $\mathbf{V} = \begin{bmatrix}1\end{bmatrix}$. Using Equation (7.6), one finds $\mathbf{u}_1 = \frac{1}{5}[3,4]^T$. Extending $\mathbf{u}_1$ to an orthonormal basis for $\mathbb{R}^2$ gives $\mathbf{u}_2 = \frac{1}{5}[-4,3]^T$. A SVD of $\mathbf{A}$ is therefore

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

(b) One has

$$\mathbf{A} = \begin{bmatrix}1 & 1\\2 & 2\\2 & 2\end{bmatrix}, \qquad \mathbf{A}^T = \begin{bmatrix}1 & 2 & 2\\1 & 2 & 2\end{bmatrix}, \qquad \mathbf{A}^T\mathbf{A} = \begin{bmatrix}9 & 9\\9 & 9\end{bmatrix}.$$

The eigenvalues of $\mathbf{A}^T\mathbf{A}$ are the zeros of $\det(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}) = (9-\lambda)^2 - 81$, 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 $\mathbf{A}$ is one. Following the dimensions of $\mathbf{A}$, one finds

$$\mathbf{\Sigma} = \begin{bmatrix}\sqrt{18} & 0\\0 & 0\\0 & 0\end{bmatrix}.$$

The normalized eigenvectors $\mathbf{v}_1, \mathbf{v}_2$ of $\mathbf{A}^T\mathbf{A}$ corresponding to the eigenvalues λ_1, λ_2 are the columns of the matrix

$$\mathbf{V} = [\mathbf{v}_1\ \mathbf{v}_2] = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & -1\\1 & 1\end{bmatrix}.$$

Using Equation (7.6) one finds $\mathbf{u}_1$, which can be extended to an orthonormal basis $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$ using Gram-Schmidt Orthogonalization (see Theorem 0.38). The vectors $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ constitute a matrix

$$\mathbf{U} = [\mathbf{u}_1\ \mathbf{u}_2\ \mathbf{u}_3] = \frac{1}{3}\begin{bmatrix}1 & -2 & -2\\2 & 2 & -1\\2 & -1 & 2\end{bmatrix}.$$

A SVD of $\mathbf{A}$ is therefore given by

$$\mathbf{A} = \frac{1}{3}\begin{bmatrix} 1 & -2 & -2 \\ 2 & 2 & -1 \\ 2 & -1 & 2 \end{bmatrix} \begin{bmatrix} \sqrt{18} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$

Exercise 7.12: More SVD examples

(a) We have $\mathbf{A} = \mathbf{e}_1$ and $\mathbf{A}^T\mathbf{A} = \mathbf{e}_1^T\mathbf{e}_1 = \begin{bmatrix} 1 \end{bmatrix}$. This gives the eigenpair $(\lambda_1, \mathbf{v}_1) = (1, 1)$ of $\mathbf{A}^T\mathbf{A}$. Hence $\sigma_1 = 1$ and $\mathbf{\Sigma} = \mathbf{e}_1 = \mathbf{A}$. As $\mathbf{\Sigma} = \mathbf{A}$ and $\mathbf{V} = \mathbf{I}_1$ we must have $\mathbf{U} = \mathbf{I}_m$ yielding a singular value decomposition

$$\mathbf{A} = \mathbf{I}_m\mathbf{e}_1\mathbf{I}_1.$$

(b) For $\mathbf{A} = \mathbf{e}_n^T$, the matrix

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix}.$$

has eigenpairs $(0, \mathbf{e}_j)$ for $j = 1, \ldots, n-1$ and $(1, \mathbf{e}_n)$. Then $\mathbf{\Sigma} = \mathbf{e}_1^T \in \mathbb{R}^{1,n}$ and $\mathbf{V} = \begin{bmatrix} \mathbf{e}_n, \mathbf{e}_{n-1}, \ldots, \mathbf{e}_1 \end{bmatrix} \in \mathbb{R}^{n,n}$. Using Equation (7.6) we get $\mathbf{u}_1 = 1$, yielding $\mathbf{U} = \begin{bmatrix} 1 \end{bmatrix}$. A SVD for $\mathbf{A}$ is therefore given by

$$\mathbf{A} = \mathbf{e}_n^T = \begin{bmatrix} 1 \end{bmatrix} \mathbf{e}_1^T \begin{bmatrix} \mathbf{e}_n, \mathbf{e}_{n-1}, \ldots, \mathbf{e}_1 \end{bmatrix}.$$

(c) In this exercise

$$\mathbf{A} = \begin{bmatrix} -1 & 0 \\ 0 & 3 \end{bmatrix}, \qquad \mathbf{A}^T = \mathbf{A}, \qquad \mathbf{A}^T\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 9 \end{bmatrix}.$$

The eigenpairs of $\mathbf{A}^T\mathbf{A}$ are given by $(\lambda_1, \mathbf{v}_1) = (9, \mathbf{e}_2)$ and $(\lambda_2, \mathbf{v}_2) = (1, \mathbf{e}_1)$, from which we find

$$\mathbf{\Sigma} = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{V} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Using Equation (7.6), one finds $\mathbf{u}_1 = \mathbf{e}_2$ and $\mathbf{u}_2 = -\mathbf{e}_1$, which constitute the matrix

$$\mathbf{U} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

A SVD of $\mathbf{A}$ is therefore given by

$$\mathbf{A} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Exercise 7.13: Singular values of a normal matrix

Suppose $\mathbf{A}$ is normal. By Theorem 6.37, there exist a diagonal matrix $\mathbf{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ and a unitary matrix $\mathbf{U}$ such that $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^*$. The eigenvalues of $\mathbf{A}^*\mathbf{A} = \mathbf{U}\mathbf{D}^*\mathbf{D}\mathbf{U}^*$ are $\overline{\lambda}_i\lambda_i = |\lambda_i|^2$, with $i = 1, \ldots, n$. By definition, the singular values of $\mathbf{A}$ are the nonnegative square roots of the eigenvalues of $\mathbf{A}^*\mathbf{A}$, that is, $|\lambda_1|, \ldots, |\lambda_n|$. We conclude that the singular values of $\mathbf{A}$ are the absolute values of its eigenvalues.

Exercise 7.18: Orthonormal bases example

Given is the matrix

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

From Example 7.9 we know that $\mathbf{B} = \mathbf{A}^T$ and hence $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ and $\mathbf{B} = \mathbf{V}\boldsymbol{\Sigma}^T\mathbf{U}^T$, with

$$\mathbf{V} = \frac{1}{3}\left[\begin{array}{cc|c} 1 & 2 & 2 \\ 2 & -2 & 1 \\ 2 & 1 & -2 \end{array}\right], \qquad \boldsymbol{\Sigma} = \left[\begin{array}{cc|c} 2 & 0 & 0 \\ 0 & 1 & 0 \end{array}\right], \qquad \mathbf{U} = \frac{1}{5}\begin{bmatrix} 3 & 4 \\ 4 & -3 \end{bmatrix}.$$

From Theorem 7.15 we know that $\mathbf{V}_1$ forms an orthonormal basis for $\operatorname{span}(\mathbf{A}^T) = \operatorname{span}(\mathbf{B})$, $\mathbf{V}_2$ an orthonormal basis for $\ker(\mathbf{A})$ and $\mathbf{U}_2$ an orthonormal basis for $\ker(\mathbf{A}^T) = \ker(\mathbf{B})$. Hence

$$\operatorname{span}(\mathbf{B}) = \alpha\mathbf{v}_1 + \beta\mathbf{v}_2, \qquad \ker(\mathbf{A}) = \gamma\mathbf{v}_3 \qquad \text{and} \qquad \ker(\mathbf{B}) = \mathbf{0}.$$

From Theorem 0.43 we know that

$$\mathbb{C}^3 = \operatorname{span}(\mathbf{B}) \oplus \ker(\mathbf{B}^*) = \operatorname{span}(\mathbf{B}) \oplus \ker(\mathbf{A})$$

is an orthogonal decomposition of $\mathbb{C}^3$.

Exercise 7.19: Some spanning sets

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

Let $\mathbf{A} = \mathbf{U}_1\boldsymbol{\Sigma}_1\mathbf{V}_1^*$ be a singular value factorization of $\mathbf{A}$. Taking the Hermitian transpose $\mathbf{A}^* = \mathbf{V}_1\boldsymbol{\Sigma}_1^*\mathbf{U}_1^*$ one finds $\operatorname{span}(\mathbf{A}^*) \subset \operatorname{span}(\mathbf{V}_1)$. Moreover, since $\mathbf{V}_1 \in \mathbb{C}^{n\times r}$ has orthonormal columns, it has the same rank as $\mathbf{A}^*$, and we conclude $\operatorname{span}(\mathbf{A}^*) = \operatorname{span}(\mathbf{V}_1)$.

Exercise 7.20: Singular values and eigenpair of composite matrix

Given is a singular value decomposition $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^*$. Let $r = \operatorname{rank}(\mathbf{A})$, so that $\sigma_1 \geq \cdots \geq \sigma_r > 0$ and $\sigma_{r+1} = \cdots = \sigma_n = 0$. Let $\mathbf{U} = [\mathbf{U}_1, \mathbf{U}_2]$ and $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2]$ be partitioned accordingly and $\boldsymbol{\Sigma}_1 = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ as in Equation (7.7), so that $\mathbf{A} = \mathbf{U}_1\boldsymbol{\Sigma}_1\mathbf{V}_1^*$ forms a singular value factorization of $\mathbf{A}$.

By Theorem 7.15,

$$\mathbf{C}\mathbf{p}_i = \begin{bmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^* & \mathbf{0} \end{bmatrix}\begin{bmatrix} \mathbf{u}_i \\ \mathbf{v}_i \end{bmatrix} = \begin{bmatrix} \mathbf{A}\mathbf{v}_i \\ \mathbf{A}^*\mathbf{u}_i \end{bmatrix} = \begin{cases} \sigma_i\mathbf{p}_i & \text{for } i = 1, \ldots, r \\ 0 \cdot \mathbf{p}_i & \text{for } i = r+1, \ldots, n \end{cases}$$

$$\mathbf{C}\mathbf{q}_i = \begin{bmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^* & \mathbf{0} \end{bmatrix}\begin{bmatrix} \mathbf{u}_i \\ -\mathbf{v}_i \end{bmatrix} = \begin{bmatrix} -\mathbf{A}\mathbf{v}_i \\ \mathbf{A}^*\mathbf{u}_i \end{bmatrix} = \begin{cases} -\sigma_i\mathbf{q}_i & \text{for } i = 1, \ldots, r \\ 0 \cdot \mathbf{q}_i & \text{for } i = r+1, \ldots, n \end{cases}$$

$$\mathbf{C}\mathbf{r}_j = \begin{bmatrix} \mathbf{0} & \mathbf{A} \\ \mathbf{A}^* & \mathbf{0} \end{bmatrix}\begin{bmatrix} \mathbf{u}_j \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{A}^*\mathbf{u}_j \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} = 0 \cdot \mathbf{r}_j, \text{ for } j = n+1, \ldots, m.$$

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

Exercise 7.25: Rank example

We are given the singular value decomposition

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 6 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ \frac{2}{3} & -\frac{1}{3} & -\frac{2}{3} \\ \frac{1}{3} & -\frac{2}{3} & \frac{2}{3} \end{bmatrix}.$$

Write $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_m]$ and $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_n]$. Clearly $r = \operatorname{rank}(\mathbf{A}) = 2$.

(a) A direct application of Theorem 7.15 with $r = 2$ gives

$\{\mathbf{u}_1, \mathbf{u}_2\}$ is an orthonormal basis for $\operatorname{span}(\mathbf{A})$,
$\{\mathbf{u}_3, \mathbf{u}_4\}$ is an orthonormal basis for $\ker(\mathbf{A}^T)$,
$\{\mathbf{v}_1, \mathbf{v}_2\}$ is an orthonormal basis for $\operatorname{span}(\mathbf{A}^T)$,
$\{\mathbf{v}_3, \mathbf{v}_4\}$ is an orthonormal basis for $\ker(\mathbf{A})$.

By Theorem 0.43, $\{\mathbf{u}_3, \mathbf{u}_4\}$ is also an orthonormal basis for $\operatorname{span}(\mathbf{A})^\perp$.

(b) Applying Theorem 7.24 with $r = 1$ yields

$$\|\mathbf{A} - \mathbf{B}\|_F \geq \sqrt{\sigma_2^2 + \sigma_3^2} = \sqrt{6^2 + 0^2} = 6.$$

(c) Following Section 7.4.2, with $\mathbf{D}' := \operatorname{diag}(\sigma_1, 0, \dots, 0) \in \mathbb{R}^{n,n}$, take

$$\mathbf{A}_1 = \mathbf{A}' := \mathbf{U} \begin{bmatrix} \mathbf{D}' \\ \mathbf{0} \end{bmatrix} \mathbf{V}^T = \begin{bmatrix} 2 & 2 & 1 \\ 2 & 2 & 1 \\ 2 & 2 & 1 \\ 2 & 2 & 1 \end{bmatrix}.$$

Exercise 7.26: Another rank example

(a) The matrix $\mathbf{B} = (b_{ij})_{ij} \in \mathbb{R}^{n,n}$ 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 $\mathbf{x} = (x_j)_j \in \mathbb{R}^n$ 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 $\mathbf{Bx}$ one finds that

$$(\mathbf{Bx})_n = \sum_{j=1}^n b_{nj} x_j = b_{n1} x_1 + b_{nn} x_n = -2^{2-n} \cdot 2^{n-2} + 1 \cdot 1 = 0.$$

For any of the remaining indices $i \neq n$, the i-th entry of the matrix product $\mathbf{Bx}$ can be expressed as

$$\begin{aligned} (\mathbf{Bx})_i = \sum_{j=1}^n b_{ij} x_j &= b_{in} + \sum_{j=1}^{n-1} 2^{n-1-j} b_{ij} \\ &= -1 + 2^{n-1-i} b_{ii} + \sum_{j=i+1}^{n-1} 2^{n-1-j} b_{ij} \end{aligned}$$

$$
\begin{aligned}
&= -1 + 2^{n-1-i} - \sum_{j=i+1}^{n-1} 2^{n-1-j} \\
&= -1 + 2^{n-1-i} - 2^{n-2-i} \sum_{j'=0}^{n-2-i} \left(\frac{1}{2}\right)^{j'} \\
&= -1 + 2^{n-1-i} - 2^{n-2-i} \frac{1 - \left(\frac{1}{2}\right)^{n-1-i}}{1 - \frac{1}{2}} \\
&= -1 + 2^{n-1-i} - 2^{n-1-i}\left(1 - 2^{-(n-1-i)}\right) \\
&= 0.
\end{aligned}
$$

As $\mathbf{B}$ has a nonzero kernel, it must be singular. The matrix $\mathbf{A}$, on the other hand, is nonsingular, as its determinant is $(-1)^n \neq 0$. The matrices $\mathbf{A}$ and $\mathbf{B}$ differ only in their $(n, 1)$-th entry, so one has $\|\mathbf{A} - \mathbf{B}\|_F = \sqrt{|a_{n1} - b_{n1}|^2} = 2^{2-n}$. 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 $\mathbf{A}$. Applying Theorem 7.24 for $r = \operatorname{rank}(\mathbf{B}) < n$, we obtain

$$
\begin{aligned}
\sigma_n \leq \sigma_n \sqrt{\left(\frac{\sigma_{r+1}}{\sigma_n}\right)^2 + \cdots + \left(\frac{\sigma_{n-1}}{\sigma_n}\right)^2 + 1} &= \sqrt{\sigma_{r+1}^2 + \cdots + \sigma_n^2} \\
= \min_{\substack{\mathbf{C} \in \mathbb{R}^{n,n} \\ \operatorname{rank}(\mathbf{C}) = r}} \|\mathbf{A} - \mathbf{C}\|_F \leq \|\mathbf{A} - \mathbf{B}\|_F &= 2^{2-n}.
\end{aligned}
$$

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

CHAPTER 8

Matrix Norms

Exercise 8.4: 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 $\mathbf{v} = \mathrm{vec}(\mathbf{A})$ obtained by stacking the columns of a matrix $\mathbf{A}$ on top of each other.

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

$$\begin{aligned}\|\mathbf{AB}\|_S = \sum_{i,j}\left|\sum_k a_{ik}b_{kj}\right| &\le \sum_{i,j,k}|a_{ik}|\cdot|b_{kj}| \\ &\le \sum_{i,j,k,l}|a_{ik}|\cdot|b_{lj}| = \sum_{i,k}|a_{ik}|\sum_{l,j}|b_{lj}| = \|\mathbf{A}\|_S\|\mathbf{B}\|_S,\end{aligned}$$

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

Exercise 8.5: Consistency of max norm?

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

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

$$\left\|\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}\right\|_M = \left\|\begin{bmatrix}2 & 2\\ 2 & 2\end{bmatrix}\right\|_M = 2 > 1 = \left\|\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}\right\|_M\left\|\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}\right\|_M,$$

contradicting $\|\mathbf{AB}\|_M \le \|\mathbf{A}\|_M\|\mathbf{B}\|_M$.

Exercise 8.6: Consistency of modified max norm?

Exercise 8.5 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 $\mathbf{A} \in \mathbb{C}^{m,n}$ and define $\|\mathbf{A}\| := \sqrt{mn}\|\mathbf{A}\|_M$ 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 $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{m,n}$ be any matrices and α any scalar.

Positivity: Clearly $\|\mathbf{A}\| = \sqrt{mn}\|\mathbf{A}\|_M \ge 0$. Moreover,

$$\|\mathbf{A}\| = 0 \iff a_{i,j} = 0\ \forall i,j \iff \mathbf{A} = 0.$$

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

Subadditivity: One has

$$\|\mathbf{A}+\mathbf{B}\| = \sqrt{nm}\|\mathbf{A}+\mathbf{B}\|_M \leq \sqrt{nm}\Big(\|\mathbf{A}\|_M + \|\mathbf{B}\|_M\Big) = \|\mathbf{A}\| + \|\mathbf{B}\|.$$

Submultiplicativity: One has

$$\begin{aligned}
\|\mathbf{AB}\| &= \sqrt{mn} \max_{\substack{1\leq i\leq m\\ 1\leq j\leq n}} \left|\sum_{k=1}^{q} a_{i,k}b_{k,j}\right| \\
&\leq \sqrt{mn} \max_{\substack{1\leq i\leq m\\ 1\leq j\leq n}} \sum_{k=1}^{q} |a_{i,k}||b_{k,j}| \\
&\leq \sqrt{mn} \max_{1\leq i\leq m} \left(\max_{\substack{1\leq k\leq q\\ 1\leq j\leq n}} |b_{k,j}| \sum_{k=1}^{q} |a_{i,k}| \right) \\
&\leq q\sqrt{mn} \left(\max_{\substack{1\leq i\leq m\\ 1\leq k\leq q}} |a_{i,k}| \right) \left(\max_{\substack{1\leq k\leq q\\ 1\leq j\leq n}} |b_{k,j}| \right) \\
&= \|\mathbf{A}\|\|\mathbf{B}\|.
\end{aligned}$$

(b) For any $\mathbf{A} \in \mathbb{C}^{m,n}$, let

$$\|\mathbf{A}\|^{(1)} := m\|\mathbf{A}\|_M \qquad \text{and} \qquad \|\mathbf{A}\|^{(2)} := n\|\mathbf{A}\|_M.$$

Comparing with the solution of part (a) we see, that the points of positivity, homogeneity and subadditivity are fulfilled here as well, making $\|\mathbf{A}\|^{(1)}$ and $\|\mathbf{A}\|^{(2)}$ valid matrix norms. Furthermore, for any $\mathbf{A} \in \mathbb{C}^{m,q}, \mathbf{B} \in \mathbb{C}^{q,n}$,

$$\begin{aligned}
\|\mathbf{AB}\|^{(1)} &= m \max_{\substack{1\leq i\leq m\\ 1\leq j\leq n}} \left|\sum_{k=1}^{q} a_{i,k}b_{k,j}\right| \leq m \left(\max_{\substack{1\leq i\leq m\\ 1\leq k\leq q}} |a_{i,k}| \right) q \left(\max_{\substack{1\leq k\leq q\\ 1\leq j\leq n}} |b_{k,j}| \right) \\
&= \|\mathbf{A}\|^{(1)}\|\mathbf{B}\|^{(1)},
\end{aligned}$$

$$\begin{aligned}
\|\mathbf{AB}\|^{(2)} &= n \max_{\substack{1\leq i\leq m\\ 1\leq j\leq n}} |\sum_{k=1}^{q} a_{i,k}b_{k,j}| \leq q \left(\max_{\substack{1\leq i\leq m\\ 1\leq k\leq q}} |a_{i,k}| \right) n \left(\max_{\substack{1\leq k\leq q\\ 1\leq j\leq n}} |b_{k,j}| \right) \\
&= \|\mathbf{A}\|^{(2)}\|\mathbf{B}\|^{(2)},
\end{aligned}$$

which proves the submultiplicativity of both norms.

Exercise 8.8: The sum norm is subordinate to?

For any matrix $\mathbf{A} = (a_{ij})_{ij} \in \mathbb{C}^{m,n}$ and column vector $\mathbf{x} = (x_j)_j \in \mathbb{C}^n$, one has

$$\|\mathbf{Ax}\|_1 = \sum_{i=1}^{m} \left|\sum_{j=1}^{n} a_{ij}x_j\right| \leq \sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|\cdot|x_j| \leq \sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}| \sum_{k=1}^{n} |x_k| = \|\mathbf{A}\|_S\|\mathbf{x}\|_1,$$

which shows that the matrix norm $\|\cdot\|_S$ is subordinate to the vector norm $\|\cdot\|_1$.

Exercise 8.9: The max norm is subordinate to?

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

(a) One has

$$\begin{aligned}\|\mathbf{Ax}\|_\infty &= \max_{i=1,\ldots,m} \left| \sum_{j=1}^n a_{ij}x_j \right| \leq \max_{i=1,\ldots,m} \sum_{j=1}^n |a_{ij}| \cdot |x_j| \leq \max_{\substack{i=1,\ldots,m\\ j=1,\ldots,n}} |a_{ij}| \sum_{j=1}^n |x_j| \\ &= \|\mathbf{A}\|_M \|\mathbf{x}\|_1.\end{aligned}$$

(b) Assume that the maximum in the definition of $\|\mathbf{A}\|_M$ is attained in column l, implying that $\|\mathbf{A}\|_M = |a_{k,l}|$ for some k. Let $\mathbf{e}_l$ be the lth standard basis vector. Then $\|\mathbf{e}_l\|_1 = 1$ and

$$\|\mathbf{Ae}_l\|_\infty = \max_{i=1,\ldots,m} |a_{i,l}| = |a_{k,l}| = |a_{k,l}| \cdot 1 = \|\mathbf{A}\|_M \cdot \|\mathbf{e}_l\|_1,$$

which is what needed to be shown.

(c) By (a), $\|\mathbf{A}\|_M \geq \|\mathbf{Ax}\|_\infty / \|\mathbf{x}\|_1$ for all nonzero vectors $\mathbf{x}$, implying that

$$\|\mathbf{A}\|_M \geq \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{Ax}\|_\infty}{\|\mathbf{x}\|_1}.$$

By (b), equality is attained for any standard basis vector $\mathbf{e}_l$ for which there exists a k such that $\|\mathbf{A}\|_M = |a_{k,l}|$. We conclude that

$$\|\mathbf{A}\|_M = \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{Ax}\|_\infty}{\|\mathbf{x}\|_1},$$

which means that $\|\cdot\|_M$ is the $(\infty, 1)$-operator norm (see Definition 8.10).

Exercise 8.15: Spectral norm of the inverse

Let $\sigma_1 \geq \cdots \geq \sigma_n$ be the singular values of $\mathbf{A}$. Since $\mathbf{A}$ is nonsingular, σ_n must be nonzero. Using Equations (8.9) and (7.14), we find

$$\|\mathbf{A}^{-1}\|_2 = \frac{1}{\sigma_n} = \frac{1}{\displaystyle\min_{\mathbf{0}\neq\mathbf{x}\in\mathbb{C}^n} \frac{\|\mathbf{Ax}\|_2}{\|\mathbf{x}\|_2}} = \max_{\mathbf{0}\neq\mathbf{x}\in\mathbb{C}^n} \frac{\|\mathbf{x}\|_2}{\|\mathbf{Ax}\|_2},$$

which is what needed to be shown.

Exercise 8.16: p-norm example

We have

$$\mathbf{A} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \qquad \mathbf{A}^{-1} = \frac{1}{3}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

Using Theorem 8.12, one finds $\|\mathbf{A}\|_1 = \|\mathbf{A}\|_\infty = 3$ and $\|\mathbf{A}^{-1}\|_1 = \|\mathbf{A}^{-1}\|_\infty = 1$. The singular values $\sigma_1 \geq \sigma_2$ of $\mathbf{A}$ are the square roots of the zeros of

$$0 = \det(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I}) = (5-\lambda)^2 - 16 = \lambda^2 - 10\lambda + 9 = (\lambda - 9)(\lambda - 1).$$

Using Theorem 8.14, we find $\|\mathbf{A}\|_2 = \sigma_1 = 3$ and $\|\mathbf{A}^{-1}\|_2 = \sigma_2^{-1} = 1$.

Exercise 8.20: Univariance of spectral norm (TODO)

Exercise 8.21: $\|\mathbf{AU}\|_2$ rectangular $\mathbf{A}$ (TODO)

Exercise 8.22: p-norm of diagonal matrix

The eigenpairs of the matrix $\mathbf{A} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ are $(\lambda_1, \mathbf{e}_1), \ldots, (\lambda_n, \mathbf{e}_n)$. For $\rho(\mathbf{A}) = \max\{|\lambda_1|, \ldots, |\lambda_n|\}$, one has

$$\begin{aligned}\|\mathbf{A}\|_p &= \max_{(x_1,\ldots,x_n)\neq \mathbf{0}} \frac{(|\lambda_1 x_1|^p + \cdots + |\lambda_n x_n|^p)^{1/p}}{(|x_1|^p + \cdots + |x_n|^p)^{1/p}} \\ &\leq \max_{(x_1,\ldots,x_n)\neq \mathbf{0}} \frac{\left(\rho(\mathbf{A})^p |x_1|^p + \cdots + \rho(\mathbf{A})^p |x_n|^p\right)^{1/p}}{(|x_1|^p + \cdots + |x_n|^p)^{1/p}} = \rho(\mathbf{A}).\end{aligned}$$

On the other hand, for $\mathbf{e}_j$ such that $\rho(\mathbf{A}) = |\lambda_j|$, one finds

$$\|\mathbf{A}\|_p = \max_{\mathbf{x}\neq 0} \frac{\|\mathbf{Ax}\|_p}{\|\mathbf{x}\|_p} \geq \frac{\|\mathbf{Ae}_j\|_p}{\|\mathbf{e}_j\|_p} = \rho(\mathbf{A}).$$

Together, the above two statements imply that $\|\mathbf{A}\|_p = \rho(\mathbf{A})$ for any diagonal matrix $\mathbf{A}$ and any p satisfying $1 \leq p \leq \infty$.

Exercise 8.23: Spectral norm of a column vector (TODO)

Exercise 8.24: Norm of absolute value matrix

(a) One finds

$$|\mathbf{A}| = \begin{bmatrix} |1+i| & |-2| \\ |1| & |1-i| \end{bmatrix} = \begin{bmatrix} \sqrt{2} & 2 \\ 1 & \sqrt{2} \end{bmatrix}.$$

(b) Let $b_{i,j}$ denote the entries of $|\mathbf{A}|$. Observe that $b_{i,j} = |a_{i,j}| = |b_{i,j}|$. Together with Theorem 8.12, these relations yield

$$\begin{aligned}\|\mathbf{A}\|_F &= \left(\sum_{i=1}^m \sum_{j=1}^n |a_{i,j}|^2\right)^{\frac{1}{2}} = \left(\sum_{i=1}^m \sum_{j=1}^n |b_{i,j}|^2\right)^{\frac{1}{2}} = \|\,|\mathbf{A}|\,\|_F, \\ \|\mathbf{A}\|_1 &= \max_{1\leq j\leq n}\left(\sum_{i=1}^m |a_{i,j}|\right) = \max_{1\leq j\leq n}\left(\sum_{i=1}^m |b_{i,j}|\right) = \|\,|\mathbf{A}|\,\|_1, \\ \|\mathbf{A}\|_\infty &= \max_{1\leq i\leq m}\left(\sum_{j=1}^n |a_{i,j}|\right) = \max_{1\leq i\leq m}\left(\sum_{j=1}^n |b_{i,j}|\right) = \|\,|\mathbf{A}|\,\|_\infty,\end{aligned}$$

which is what needed to be shown.

(c) To show this relation between the 2-norms of $\mathbf{A}$ and $|\mathbf{A}|$, we first examine the connection between the l_2-norms of $\mathbf{Ax}$ and $|\mathbf{A}| \cdot |\mathbf{x}|$, where $\mathbf{x} = (x_1, \ldots, x_n)$ and $|\mathbf{x}| = (|x_1|, \ldots, |x_n|)$. We find

$$\|\mathbf{Ax}\|_2 = \left(\sum_{i=1}^m \left|\sum_{j=1}^n a_{i,j} x_j\right|^2\right)^{\frac{1}{2}} \leq \left(\sum_{i=1}^m \left(\sum_{j=1}^n |a_{i,j}||x_j|\right)^2\right)^{\frac{1}{2}} = \|\,|\mathbf{A}| \cdot |\mathbf{x}|\,\|_2.$$

Now let $\mathbf{x}^*$ with $\|\mathbf{x}^*\|_2 = 1$ be a vector for which $\|\mathbf{A}\|_2 = \|\mathbf{Ax}^*\|_2$. That is, let $\mathbf{x}^*$ be a unit vector for which the maximum in the definition of 2-norm is attained. Observe that $|\mathbf{x}^*|$ is then a unit vector as well, $\| \, |\mathbf{x}^*| \, \|_2 = 1$. Then, by the above estimate of l_2-norms and definition of the 2-norm,

$$\|\mathbf{A}\|_2 = \|\mathbf{A}\mathbf{x}^*\|_2 \leq \| \, |\mathbf{A}| \cdot |\mathbf{x}^*| \, \|_2 \leq \| \, |\mathbf{A}| \, \|_2.$$

(d) By Theorem 8.12, we can solve this exercise by finding a matrix $\mathbf{A}$ for which $\mathbf{A}$ and $|\mathbf{A}|$ have different largest singular values. As $\mathbf{A}$ is real and symmetric, there exist $a, b, c \in \mathbb{R}$ such that

$$\mathbf{A} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \qquad |\mathbf{A}| = \begin{bmatrix} |a| & |b| \\ |b| & |c| \end{bmatrix},$$
$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} a^2 + b^2 & ab + bc \\ ab + bc & b^2 + c^2 \end{bmatrix}, \qquad |\mathbf{A}|^T|\mathbf{A}| = \begin{bmatrix} a^2 + b^2 & |ab| + |bc| \\ |ab| + |bc| & b^2 + c^2 \end{bmatrix}.$$

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

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} a^2 + b^2 & 0 \\ 0 & a^2 + b^2 \end{bmatrix}, \qquad |\mathbf{A}|^T|\mathbf{A}| = \begin{bmatrix} a^2 + b^2 & 2|ab| \\ 2|ab| & a^2 + b^2 \end{bmatrix}.$$

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

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \qquad \|\mathbf{A}\|_2 = \sqrt{2}, \qquad \| \, |\mathbf{A}| \, \|_2 = 2.$$

Exercise 8.25: Spectral norm (TODO)

Exercise 8.26: Absolute norms (TODO)

Exercise 8.27: Is the spectral norm an absolute norm?

By the solution for Exercise 8.24(d), we know that

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

is a matrix for which $\|\mathbf{A}\|_2 = \sqrt{2} \neq 2 = \| \, |\mathbf{A}| \, \|_2$.

Exercise 8.34: Sharpness of perturbation bounds (TODO)

Exercise 8.35: Condition number of 2nd derivative matrix

Recall that $\mathbf{T} = \text{tridiag}(-1, 2, -1)$ and, by Exercise 2.7, $\mathbf{T}^{-1}$ is given by

$$\left(\mathbf{T}^{-1}\right)_{ij} = \left(\mathbf{T}^{-1}\right)_{ji} = (1-ih)j > 0, \quad 1 \le j \le i \le m, \quad h = \frac{1}{m+1}.$$

From Theorems 8.12 and 8.14, we have the following explicit expressions for the 1-, 2- and ∞-norms

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

for any matrix $\mathbf{A} \in \mathbb{C}^{m,n}$, where σ_1 is the largest singular value of $\mathbf{A}$, σ_m the smallest singular value of $\mathbf{A}$, and we assumed $\mathbf{A}$ to be nonsingular in the third equation.

(a) For the matrix $\mathbf{T}$ this gives $\|\mathbf{T}\|_1 = \|\mathbf{T}\|_\infty = m+1$ for $m = 1, 2$ and $\|\mathbf{T}\|_1 = \|\mathbf{T}\|_\infty = 4$ for $m \ge 3$. For the inverse we get $\|\mathbf{T}^{-1}\|_1 = \|\mathbf{T}^{-1}\|_\infty = \frac{1}{2} = \frac{1}{8}h^{-2}$ for $m = 1$ and

$$\|\mathbf{T}^{-1}\|_1 = \left\|\frac{1}{3}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\right\|_1 = 1 = \left\|\frac{1}{3}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\right\|_\infty = \|\mathbf{T}^{-1}\|_\infty$$

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

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

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 $\|\mathbf{T}^{-1}\|_1 = \frac{1}{8}h^{-2}$. For even $m > 2$, on the other hand, the maximum over all integral j is attained at $j = \frac{m}{2} = \frac{1-h}{2h}$ or $j = \frac{m+2}{2} = \frac{1+h}{2h}$, which both give $\|\mathbf{T}^{-1}\|_1 = \frac{1}{8}(h^{-2} - 1)$.

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

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

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

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

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

$$\sigma_1 = |\lambda_m| = 2 - 2\cos(m\pi h) = 2 + 2\cos(\pi h),$$
$$\sigma_m = |\lambda_1| = 2 - 2\cos(\pi h).$$

It follows that

$$\mathrm{cond}_2(\mathbf{T}) = \frac{\sigma_1}{\sigma_m} = \frac{1 + \cos(\pi h)}{1 - \cos(\pi h)} = \cot^2\left(\frac{\pi h}{2}\right).$$

(c) From $\tan x > x$ we obtain $\cot^2 x = \frac{1}{\tan^2 x} < \frac{1}{x^2}$. Using this and $\cot^2 x > x^{-2} - \frac{2}{3}$ we find

$$\frac{4}{\pi^2 h^2} - \frac{2}{3} < \mathrm{cond}_2(\mathbf{T}) < \frac{4}{\pi^2 h^2}.$$

Exercise 8.43: p-norm for $p = 1$ and $p = \infty$ (TODO)

Exercise 8.44: The p-norm unit sphere (TODO)

Exercise 8.45: Sharpness of p-norm inequality (TODO)

Exercise 8.46: p-norm inequalities for arbitrary p (TODO)

CHAPTER 9

The Classical Iterative Methods

Exercise 9.9: Slow spectral radius convergence

In this Exercise we show that the convergence of

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

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

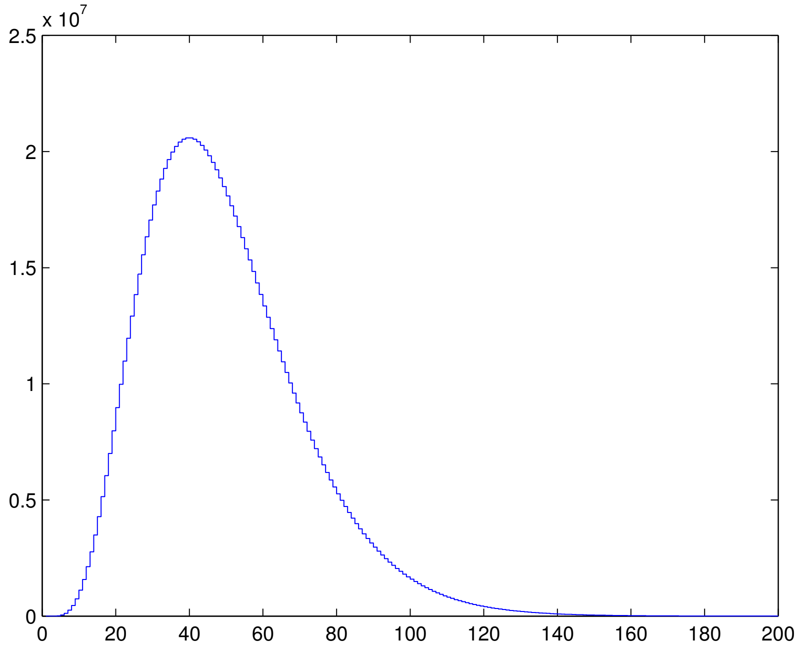
(a) The Matlab code

```
n = 5
a = 10
l = 0.9

for k = n-1:200
    L(k) = nchoosek(k,n-1)*a^(n-1)*l^(k-n+1);
end

stairs(L)
```

yields the following stairstep graph of f:



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;

while nchoosek(k,n-1)*a^(n-1)*l^(k-n+1) >= 10^(-8)
    k = k + 1;
end

k
```

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

(b) Denote by $\mathbf{E}_n$ the $n \times n$ matrix $\mathbf{E}$ as in the exercise, leaving out the subscript whenever its size is understood. Clearly $\mathbf{E}^k$ is of the required form for $k = 1$. For any $1 \leq k \leq n-1$, the induction hypothesis yields

$$\begin{aligned}
\mathbf{E}^{1+k} &= \mathbf{E}^1\mathbf{E}^k \\
&= \left[\begin{array}{c|cc} \mathbf{0}_{n-k-1,1} & \mathbf{I}_{n-k-1} & \mathbf{0}_{n-k-1,k} \\ \hline \mathbf{0}_{k+1,1} & \mathbf{0}_{k+1,n-k-1} & \begin{smallmatrix}\mathbf{I}_k \\ \mathbf{0}_{1,k}\end{smallmatrix} \end{array}\right] \left[\begin{array}{cc|c} \mathbf{0}_{1,k} & 1 & \mathbf{0}_{1,n-k-1} \\ \hline \mathbf{0}_{n-1,k} & \mathbf{0}_{n-1,1} & \begin{smallmatrix}\mathbf{I}_{n-k-1} \\ \mathbf{0}_{k,n-k-1}\end{smallmatrix} \end{array}\right] \\
&= \left[\begin{array}{c|c} \mathbf{0}_{n-k-1,k+1} & \mathbf{I}_{n-k-1} \\ \hline \mathbf{0}_{k+1,k+1} & \mathbf{0}_{k+1,n-k-1} \end{array}\right].
\end{aligned}$$

Similarly, one verifies that $\mathbf{E}^n = \mathbf{E}^1\mathbf{E}^{n-1} = \mathbf{0}_{n,n}$. We summarize that the matrix $\mathbf{E}$ is *nilpotent* of degree n.

(c) By the binomial theorem and (b), one has

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

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

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

which is what needed to be shown.

Exercise 9.11: A special norm (TODO)

Exercise 9.12: When is $\mathbf{A} + \mathbf{E}$ nonsingular?

Suppose $\rho(\mathbf{A}^{-1}\mathbf{E}) = \rho\big(\mathbf{A}^{-1}(-\mathbf{E})\big) < 1$. By Theorem 9.10.2, $\mathbf{I} + \mathbf{A}^{-1}\mathbf{E}$ is nonsingular and therefore so is the product $\mathbf{A}(\mathbf{I} + \mathbf{A}^{-1}\mathbf{E}) = \mathbf{A} + \mathbf{E}$.

Conversely, suppose $\mathbf{A} + \mathbf{E}$ is nonsingular. Then the inverse $\mathbf{C}$ of $\mathbf{I} - \mathbf{A}^{-1}(-\mathbf{E}) = \mathbf{A}^{-1}(\mathbf{A} + \mathbf{E})$ exists, implying that the series $\sum_{k=0}^{\infty} \big(\mathbf{A}^{-1}(-\mathbf{E})\big)^k$ converges (namely to $\mathbf{C}$). By Theorem 9.10.1,

$$\rho(\mathbf{A}^{-1}\mathbf{E}) = \rho\big(\mathbf{A}^{-1}(-\mathbf{E})\big) < 1.$$

Exercise 9.17: Divergence example for J and GS

We compute the matrices $\mathbf{G}_J$ and $\mathbf{G}_1$ from $\mathbf{A}$ and show that that the spectral radii $\rho(\mathbf{G}_J), \rho(\mathbf{G}_1) \geq 1$. Once this is shown, Theorem 9.15 implies that the Jacobi method and Gauss-Seidel's method diverge.

Write $\mathbf{A} = \mathbf{D} - \mathbf{A}_L - \mathbf{A}_R$ as in the book. From Equation (9.14), we find

$$\mathbf{G}_J = \mathbf{I} - \mathbf{M}_J^{-1}\mathbf{A} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 0 & -2 \\ -\frac{3}{4} & 0 \end{bmatrix},$$

$$\mathbf{G}_1 = \mathbf{I} - \mathbf{M}_1^{-1}\mathbf{A} = \mathbf{I} - (\mathbf{D} - \mathbf{A}_L)^{-1}\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ -\frac{3}{4} & \frac{1}{4} \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$
$$= \begin{bmatrix} 0 & -2 \\ 0 & \frac{3}{2} \end{bmatrix}.$$

From this, we find $\rho(\mathbf{G}_J) = \sqrt{3/2}$ and $\rho(\mathbf{G}_1) = 3/2$, both of which are bigger than 1.

Exercise 9.18: J and GS on spline matrix (TODO)

Exercise 9.19: Strictly diagonally dominance; The J method

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

$$\mathbf{G} = \mathbf{I} - \operatorname{diag}(a_{11}, \ldots, a_{nn})^{-1}\mathbf{A} = \begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & -\frac{a_{13}}{a_{11}} & \cdots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & -\frac{a_{23}}{a_{22}} & \cdots & -\frac{a_{2n}}{a_{22}} \\ -\frac{a_{31}}{a_{33}} & -\frac{a_{32}}{a_{33}} & 0 & \cdots & -\frac{a_{3n}}{a_{33}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & -\frac{a_{n3}}{a_{nn}} & \cdots & 0 \end{bmatrix}.$$

By Theorem 8.12, the ∞-norm can be expressed as the maximum, over all rows, of the sum of absolute values of the entries in a row. Using that $\mathbf{A}$ is strictly diagonally dominant, one finds

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

As by Lemma 8.11 the ∞-norm is consistent, Corollary 9.16 implies that the Jacobi method converges for any strictly diagonally dominant matrix $\mathbf{A}$.

Exercise 9.20: Strictly diagonally dominance; The GS method

Let $\mathbf{A} = -\mathbf{A}_L + \mathbf{D} - \mathbf{A}_R$ be decomposed as a sum of a lower triangular, a diagonal, and an upper triangular part. By Equation (9.15), the approximate solutions $\mathbf{x}^{(k)}$ are related by

$$\mathbf{D}\mathbf{x}^{(k+1)} = \mathbf{A}_L\mathbf{x}^{(k+1)} + \mathbf{A}_R\mathbf{x}^{(k)} + \mathbf{b}$$

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

$$\mathbf{D}\varepsilon^{(k+1)} = \mathbf{A}_L\varepsilon^{(k+1)} + \mathbf{A}_R\varepsilon^{(k)}.$$

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

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

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

$$|\varepsilon_1^{(k+1)}| = |a_{11}|^{-1} \left| -a_{12}\varepsilon_2^{(k)} - \cdots - a_{1n}\varepsilon_n^{(k)} \right| \leq r_1\|\varepsilon^{(k)}\|_\infty \leq r\|\varepsilon^{(k)}\|_\infty.$$

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

$$|a_{jj}|^{-1} \left| -a_{j1}\varepsilon_1^{(k+1)} - \cdots - a_{j,j-1}\varepsilon_{j-1}^{(k+1)} - a_{j,j+1}\varepsilon_{j+1}^{(k)} - \cdots - a_{jn}\varepsilon_n^{(k)} \right|$$

$$\leq r_j \max\{r\|\varepsilon^{(k)}\|_\infty, \|\varepsilon^{(k)}\|_\infty\} = r_j\|\varepsilon^{(k)}\|_\infty \leq r\|\varepsilon^{(k)}\|_\infty.$$

Equation $(\star)$ then follows by induction.

If $\mathbf{A}$ is strictly diagonally dominant, then $r < 1$ and

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

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

Exercise 9.22: Estimate in Lemma 9.21 can be exact

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

$$\mathbf{x}^{(k+1)} = \mathbf{G}_J\mathbf{x}^{(k)} + \mathbf{c}, \qquad \mathbf{G}_J = \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix}, \qquad \mathbf{c} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}.$$

The initial guess

$$\mathbf{x}^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

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

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

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 $\mathbf{x}^{(k)}$ differs by 2^{-k} from the fixed point, implying that $\|\varepsilon^{(k)}\|_\infty = 2^{-k}$. Therefore, for given s, the error $\|\varepsilon^{(k)}\|_\infty \leq 10^{-s}$ for the first time at $k = \lceil s\log(10)/\log(2)\rceil$. The bound from Lemma 9.21, on the other hand, yields $\tilde{k} = 2s\log(10)$ in this case.

Exercise 9.25: The GS method converges, but not the J method

The eigenvalues of $\mathbf{A}$ are the zeros of $\det(\mathbf{A}-\lambda\mathbf{I}) = (-\lambda+2a+1)(\lambda+a-1)^2$. 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 $\mathbf{A}$ is positive definite for such a.

Let's compute that spectral radius of $\mathbf{G}_J = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A}$, where $\mathbf{D}$ is the diagonal part of $\mathbf{A}$. The eigenvalues of $\mathbf{G}_J$ are the zeros of the characteristic polynomial

$$\det(\mathbf{G}_J - \lambda\mathbf{I}) = \begin{vmatrix} -\lambda & -a & -a \\ -a & -\lambda & -a \\ -a & -a & -\lambda \end{vmatrix} = (-\lambda-2a)(a-\lambda)^2,$$

and we find spectral radius $\rho(\mathbf{G}_J) = \max\{|a|, |2a|\}$. It follows that $\rho(\mathbf{G}_J) > 1$ whenever $1/2 < a < 1$, in which case Theorem 9.5 implies that the Jacobi method does not converge (even though $\mathbf{A}$ is symmetric positive definite).

Exercise 9.28: 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

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

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

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

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 $\mathbf{G}$ are the zeros of the characteristic polynomial $\lambda^2 - a^2 = (\lambda-a)(\lambda+a)$, and we find that $\mathbf{G}$ has spectral radius $\rho(\mathbf{G}) = 1-\eta$, where $\eta := 1-|a|$. Equation (9.22) yields an estimate $\tilde{k} = \log(10)s/(1-|a|)$ for the smallest number of iterations k so that $\rho(\mathbf{G})^k \leq 10^{-s}$. In particular, taking $a = 0.9$ and $s = 16$, one expects at least $\tilde{k} = 160\log(10) \approx 368$ iterations before $\rho(\mathbf{G})^k \leq 10^{-16}$. On the other hand, $0.9^k = |a|^k = 10^{-s} = 10^{-16}$ when $k \approx 350$, so in this case the estimate is fairly accurate.

CHAPTER 10

The Conjugate Gradient Method

Exercise 10.1: Paraboloid

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

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

which is what needed to be shown.

Exercise 10.4: Steepest descent iteration

In the method of Steepest Descent we choose, at the kth iteration, the search direction $\mathbf{p}_k = \mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k$ and optimal step length

$$\alpha_k := \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{r}_k^T\mathbf{A}\mathbf{r}_k}.$$

Given is a quadratic function

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

and an initial guess $\mathbf{x}_0 = [-1, -1/2]^T$ of its minimum. The corresponding residual is

$$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} -1 \\ -1/2 \end{bmatrix} = \begin{bmatrix} 3/2 \\ 0 \end{bmatrix}.$$

Performing the steps in Equation (10.6) twice yields

$$\mathbf{t}_0 = \mathbf{A}\mathbf{r}_0 = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} 3/2 \\ 0 \end{bmatrix} = \begin{bmatrix} 3 \\ -3/2 \end{bmatrix}, \quad \alpha_0 = \frac{\mathbf{r}_0^T\mathbf{r}_0}{\mathbf{r}_0^T\mathbf{t}_0} = \frac{9/4}{9/2} = \frac{1}{2},$$

$$\mathbf{x}_1 = \begin{bmatrix} -1 \\ -1/2 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 3/2 \\ 0 \end{bmatrix} = \begin{bmatrix} -1/4 \\ -1/2 \end{bmatrix}, \quad \mathbf{r}_1 = \begin{bmatrix} 3/2 \\ 0 \end{bmatrix} - \frac{1}{2}\begin{bmatrix} 3 \\ -3/2 \end{bmatrix} = \begin{bmatrix} 0 \\ 3/4 \end{bmatrix}$$

$$\mathbf{t}_1 = \mathbf{A}\mathbf{r}_1 = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} 0 \\ 3/4 \end{bmatrix} = \begin{bmatrix} -3/4 \\ 3/2 \end{bmatrix}, \quad \alpha_1 = \frac{\mathbf{r}_1^T\mathbf{r}_1}{\mathbf{r}_1^T\mathbf{t}_1} = \frac{9/16}{9/8} = \frac{1}{2},$$

$$\mathbf{x}_2 = \begin{bmatrix} -1/4 \\ -1/2 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} 0 \\ 3/4 \end{bmatrix} = \begin{bmatrix} -1/4 \\ -1/8 \end{bmatrix}, \quad \mathbf{r}_2 = \begin{bmatrix} 0 \\ 3/4 \end{bmatrix} - \frac{1}{2}\begin{bmatrix} -3/4 \\ 3/2 \end{bmatrix} = \begin{bmatrix} 3/8 \\ 0 \end{bmatrix}.$$

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

$$(\star) \qquad \mathbf{t}_{2k-2} = 3 \cdot 4^{1-k}\begin{bmatrix} 1 \\ -1/2 \end{bmatrix}, \quad \mathbf{x}_{2k-1} = -4^{-k}\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \mathbf{r}_{2k-1} = 3 \cdot 4^{-k}\begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

$$(\star\star) \qquad \mathbf{t}_{2k-1} = 3\cdot 4^{-k}\begin{bmatrix}-1\\2\end{bmatrix}, \quad \mathbf{x}_{2k} = -4^{-k}\begin{bmatrix}1\\1/2\end{bmatrix}, \quad \mathbf{r}_{2k} = 3\cdot 4^{-k}\begin{bmatrix}1/2\\0\end{bmatrix}.$$

Then

$$\mathbf{t}_{2k} = 3\cdot 4^{-k}\begin{bmatrix}2 & -1\\-1 & 2\end{bmatrix}\begin{bmatrix}1/2\\0\end{bmatrix} = 3\cdot 4^{1-(k+1)}\begin{bmatrix}1\\-1/2\end{bmatrix},$$

$$\alpha_{2k} = \frac{\mathbf{r}_{2k}^T\mathbf{r}_{2k}}{\mathbf{r}_{2k}^T\mathbf{t}_{2k}} = \frac{9\cdot 4^{-2k}\cdot(\frac{1}{2})^2}{9\cdot 4^{-2k}\cdot\frac{1}{2}} = \frac{1}{2},$$

$$\mathbf{x}_{2k+1} = -4^{-k}\begin{bmatrix}1\\1/2\end{bmatrix} + \frac{1}{2}\cdot 3\cdot 4^{-k}\begin{bmatrix}1/2\\0\end{bmatrix} = -4^{-(k+1)}\begin{bmatrix}1\\2\end{bmatrix},$$

$$\mathbf{r}_{2k+1} = 3\cdot 4^{-k}\begin{bmatrix}1/2\\0\end{bmatrix} - \frac{1}{2}\cdot 3\cdot 4^{1-(k+1)}\begin{bmatrix}1\\-1/2\end{bmatrix} = 3\cdot 4^{-(k+1)}\begin{bmatrix}0\\1\end{bmatrix},$$

$$\mathbf{t}_{2k+1} = 3\cdot 4^{-(k+1)}\begin{bmatrix}2 & -1\\-1 & 2\end{bmatrix}\begin{bmatrix}0\\1\end{bmatrix} = 3\cdot 4^{-(k+1)}\begin{bmatrix}-1\\2\end{bmatrix},$$

$$\alpha_{2k+1} = \frac{\mathbf{r}_{2k+1}^T\mathbf{r}_{2k+1}}{\mathbf{r}_{2k+1}^T\mathbf{t}_{2k+1}} = \frac{9\cdot 4^{-2(k+1)}}{9\cdot 4^{-2(k+1)}\cdot 2} = \frac{1}{2},$$

$$\mathbf{x}_{2k+2} = -4^{-(k+1)}\begin{bmatrix}1\\2\end{bmatrix} + \frac{1}{2}\cdot 3\cdot 4^{-(k+1)}\begin{bmatrix}0\\1\end{bmatrix} = -4^{-(k+1)}\begin{bmatrix}1\\1/2\end{bmatrix},$$

$$\mathbf{r}_{2k+2} = 3\cdot 4^{-(k+1)}\begin{bmatrix}0\\1\end{bmatrix} - \frac{1}{2}\cdot 3\cdot 4^{-(k+1)}\begin{bmatrix}-1\\2\end{bmatrix} = 3\cdot 4^{-(k+1)}\begin{bmatrix}1/2\\0\end{bmatrix},$$

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

Exercise 10.6: 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 Ω, attains its maximum on the boundary of Ω.

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,

$$f(x) = f\big((1-\lambda)a + \lambda b\big) \leq (1-\lambda)f(a) + \lambda f(b) = f(a) + \lambda\big(f(b) - f(a)\big).$$

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 $f(x) \leq \max\big(f(a), f(b)\big)$ and that f attains its maximum on the boundary of its domain of definition.

Exercise 10.7: The A-inner product

We verify the axioms of Definition 0.27.

Positivity: Since $\mathbf{A}$ is positive definite, $\mathbf{x} \neq \mathbf{0} \implies \langle\mathbf{x},\mathbf{x}\rangle = \mathbf{x}^T\mathbf{A}\mathbf{x} > 0$. On the other hand, $\mathbf{x} = \mathbf{0} \implies \langle\mathbf{x},\mathbf{x}\rangle = \mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{0}^T\mathbf{A}\mathbf{0} = 0$. It follows that $\langle\mathbf{x},\mathbf{x}\rangle \geq 0$ for all $\mathbf{x}$, with equality if and only if $\mathbf{x} = \mathbf{0}$.

Symmetry: One has $\langle\mathbf{x},\mathbf{y}\rangle = \mathbf{x}^T\mathbf{A}\mathbf{y} = (\mathbf{x}^T\mathbf{A}\mathbf{y})^T = \mathbf{y}^T\mathbf{A}^T\mathbf{x} = \mathbf{y}^T\mathbf{A}\mathbf{x} = \langle\mathbf{y},\mathbf{x}\rangle$ for all vectors $\mathbf{x}$ and $\mathbf{y}$.

Linearity: One has $\langle a\mathbf{x}+b\mathbf{y},\mathbf{z}\rangle = (a\mathbf{x}+b\mathbf{y})^T\mathbf{A}\mathbf{z} = a\mathbf{x}^T\mathbf{A}\mathbf{z} + b\mathbf{y}^T\mathbf{A}\mathbf{z} = a\langle\mathbf{x},\mathbf{z}\rangle + b\langle\mathbf{y},\mathbf{z}\rangle$ for all real numbers a, b and vectors $\mathbf{x},\mathbf{y},\mathbf{z}$.

Exercise 10.9: A test for the error bound (TODO)

Exercise 10.10: Orthogonality in steepest descent

It follows from (10.6) that

$$\mathbf{r}_k^T\mathbf{r}_{k+1} = \mathbf{r}_k^T\left(\mathbf{r}_k - \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{r}_k^T\mathbf{A}\mathbf{r}_k}\mathbf{A}\mathbf{r}_k\right) = \mathbf{r}_k^T\mathbf{r}_k - \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{r}_k^T\mathbf{A}\mathbf{r}_k}\mathbf{r}_k^T\mathbf{A}\mathbf{r}_k = 0.$$

There is, however, no reason to believe that the other residuals are orthogonal.

Exercise 10.12: Conjugate gradient iteration, II

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

$$\mathbf{x}^{(1)} = \mathbf{x}^{(0)} + \frac{(\mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}, \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)})}{(\mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}, \mathbf{A}\mathbf{b} - \mathbf{A}^2\mathbf{x}^{(0)})}(\mathbf{b} - \mathbf{A}\mathbf{x}^{(0)}) = \frac{(\mathbf{b},\mathbf{b})}{(\mathbf{b},\mathbf{A}\mathbf{b})}\mathbf{b}.$$

Exercise 10.13: Conjugate gradient iteration, III

By Exercise 10.12,

$$\mathbf{x}^{(1)} = \frac{(\mathbf{b},\mathbf{b})}{(\mathbf{b},\mathbf{A}\mathbf{b})}\mathbf{b} = \frac{9}{18}\begin{bmatrix}0\\3\end{bmatrix} = \begin{bmatrix}0\\3/2\end{bmatrix}.$$

We find, in order,

$$\mathbf{p}^{(0)} = \mathbf{r}^{(0)} = \begin{bmatrix}0\\3\end{bmatrix}, \qquad \alpha_0 = \frac{1}{2},\ \mathbf{r}^{(1)} = \begin{bmatrix}\frac{3}{2}\\0\end{bmatrix},$$

$$\beta_0 = \frac{1}{4}, \qquad \mathbf{p}^{(1)} = \begin{bmatrix}\frac{3}{2}\\\frac{3}{4}\end{bmatrix}, \qquad \alpha_1 = \frac{2}{3},\ \mathbf{x}^{(2)} = \begin{bmatrix}1\\2\end{bmatrix}.$$

Since the residual vectors $\mathbf{r}^{(0)}, \mathbf{r}^{(1)}, \mathbf{r}^{(2)}$ must be orthogonal, it follows that $\mathbf{r}^{(2)} = \mathbf{0}$ and $\mathbf{x}^{(2)}$ must be an exact solution. This can be verified directly by hand.

Exercise 10.16: The cg step length is optimal

For any fixed search direction $\mathbf{p}_k$, the step length α_k is optimal if $Q(\mathbf{x}_{k+1})$ is as small as possible, that is

$$Q(\mathbf{x}_{k+1}) = Q(\mathbf{x}_k + \alpha_k\mathbf{p}_k) = \min_{\alpha\in\mathbb{R}} f(\alpha),$$

where, by (10.3),

$$f(\alpha) := Q(\mathbf{x}_k + \alpha\mathbf{p}_k) = Q(\mathbf{x}_k) - \alpha\mathbf{p}_k^T\mathbf{r}_k + \frac{1}{2}\alpha^2\mathbf{p}_k^T\mathbf{A}\mathbf{p}_k$$

is a quadratic polynomial in α. Since $\mathbf{A}$ is assumed to be positive definite, necessarily $\mathbf{p}_k^T\mathbf{A}\mathbf{p}_k > 0$. Therefore f has a minimum, which it attains at

$$\alpha = \frac{\mathbf{p}_k^T\mathbf{r}_k}{\mathbf{p}_k^T\mathbf{A}\mathbf{p}_k}.$$

Applying (10.22) repeatedly, one finds that the search direction $\mathbf{p}_k$ for the conjugate gradient method satisfies

$$\mathbf{p}_k = \mathbf{r}_k + \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{r}_{k-1}^T\mathbf{r}_{k-1}}\mathbf{p}_{k-1} = \mathbf{r}_k + \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{r}_{k-1}^T\mathbf{r}_{k-1}}\left(\mathbf{r}_{k-1} + \frac{\mathbf{r}_{k-1}^T\mathbf{r}_{k-1}}{\mathbf{r}_{k-2}^T\mathbf{r}_{k-2}}\mathbf{p}_{k-2}\right) = \cdots$$

As $\mathbf{p}_0 = \mathbf{r}_0$, the difference $\mathbf{p}_k - \mathbf{r}_k$ is a linear combination of the vectors $\mathbf{r}_{k-1}, \ldots, \mathbf{r}_0$, each of which is orthogonal to $\mathbf{r}_k$. It follows that $\mathbf{p}_k^T\mathbf{r}_k = \mathbf{r}_k^T\mathbf{r}_k$ and that the step length α is optimal for

$$\alpha = \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{p}_k^T\mathbf{A}\mathbf{p}_k} = \alpha_k.$$

Exercise 10.17: Starting value in cg

As in the exercise, we consider the conjugate gradient method for $\mathbf{A}\mathbf{y} = \mathbf{r}_0$, with $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$. Starting with

$$\mathbf{y}_0 = \mathbf{0}, \qquad \mathbf{s}_0 = \mathbf{r}_0 - \mathbf{A}\mathbf{y}_0 = \mathbf{r}_0, \qquad \mathbf{q}_0 = \mathbf{s}_0 = \mathbf{r}_0,$$

one computes, for any $k \geq 0$,

$$\gamma_k := \frac{\mathbf{s}_k^T\mathbf{s}_k}{\mathbf{q}_k^T\mathbf{A}\mathbf{q}_k}, \qquad \mathbf{y}_{k+1} = \mathbf{y}_k + \gamma_k\mathbf{q}_k, \qquad \mathbf{s}_{k+1} = \mathbf{s}_k - \gamma_k\mathbf{A}\mathbf{q}_k,$$

$$\delta_k := \frac{\mathbf{s}_{k+1}^T\mathbf{s}_{k+1}}{\mathbf{s}_k^T\mathbf{s}_k}, \qquad \mathbf{q}_{k+1} = \mathbf{s}_{k+1} + \delta_k\mathbf{q}_k.$$

How are the iterates $\mathbf{y}_k$ and $\mathbf{x}_k$ related? As remarked above, $\mathbf{s}_0 = \mathbf{r}_0$ and $\mathbf{q}_0 = \mathbf{r}_0 = \mathbf{p}_0$. Suppose $\mathbf{s}_k = \mathbf{r}_k$ and $\mathbf{q}_k = \mathbf{p}_k$ for some $k \geq 0$. Then

$$\mathbf{s}_{k+1} = \mathbf{s}_k - \gamma_k\mathbf{A}\mathbf{q}_k = \mathbf{r}_k - \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{p}_k^T\mathbf{A}\mathbf{p}_k}\mathbf{A}\mathbf{p}_k = \mathbf{r}_k - \alpha_k\mathbf{A}\mathbf{p}_k = \mathbf{r}_{k+1},$$

$$\mathbf{q}_{k+1} = \mathbf{s}_{k+1} + \delta_k\mathbf{q}_k = \mathbf{r}_{k+1} + \frac{\mathbf{r}_{k+1}^T\mathbf{r}_{k+1}}{\mathbf{r}_k^T\mathbf{r}_k}\mathbf{p}_k = \mathbf{p}_{k+1}.$$

It follows by induction that $\mathbf{s}_k = \mathbf{r}_k$ and $\mathbf{q}_k = \mathbf{p}_k$ for all $k \geq 0$. In addition,

$$\mathbf{y}_{k+1} - \mathbf{y}_k = \gamma_k\mathbf{q}_k = \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{p}_k^T\mathbf{A}\mathbf{p}_k}\mathbf{p}_k = \mathbf{x}_{k+1} - \mathbf{x}_k, \qquad \text{for any } k \geq 0,$$

so that $\mathbf{y}_k = \mathbf{x}_k - \mathbf{x}_0$.

Exercise 10.23: Krylov space and cg iterations

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

$$\mathbb{W}_k := \operatorname{span}\left\{\mathbf{r}^{(0)}, \mathbf{A}\mathbf{r}^{(0)}, \ldots, \mathbf{A}^{k-1}\mathbf{r}^{(0)}\right\}.$$

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

$$\left[\mathbf{r}^{(0)}, \mathbf{A}\mathbf{r}^{(0)}, \mathbf{A}^2\mathbf{r}^{(0)}\right] = \left[\mathbf{b}, \mathbf{A}\mathbf{b}, \mathbf{A}^2\mathbf{b}\right] = \left[\ \begin{bmatrix}4\\0\\0\end{bmatrix}, \begin{bmatrix}8\\-4\\0\end{bmatrix}, \begin{bmatrix}20\\-16\\4\end{bmatrix}\ \right].$$

(b) As $\mathbf{x}^{(0)} = \mathbf{0}$ we have $\mathbf{p}^{(0)} = \mathbf{r}^{(0)} = \mathbf{b}$. We have for $k = 0, 1, 2, \dots$ Equations (10.20), (10.21), and (10.22),

$$\begin{aligned} \mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} + \alpha_k \mathbf{p}^{(k)}, & \alpha_k &= \frac{\mathbf{r}^{(k)\,T}\mathbf{r}^{(k)}}{\mathbf{p}^{(k)\,T}\mathbf{A}\mathbf{p}^{(k)}}, \\ \mathbf{r}^{(k+1)} &= \mathbf{r}^{(k)} - \alpha_k \mathbf{A}\mathbf{p}^{(k)}, & & \\ \mathbf{p}^{(k+1)} &= \mathbf{r}^{(k+1)} + \beta_k \mathbf{p}^{(k)}, & \beta_k &= \frac{\mathbf{r}^{(k+1)\,T}\mathbf{r}^{(k+1)}}{\mathbf{r}^{(k)\,T}\mathbf{r}^{(k)}}, \end{aligned}$$

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

$$\begin{aligned} &\alpha_0 = \frac{1}{2}, \quad \mathbf{x}^{(1)} = \begin{bmatrix} 2 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{r}^{(1)} = \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}, \quad \beta_0 = \frac{1}{4}, \quad \mathbf{p}^{(1)} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}, \\ &\alpha_1 = \frac{2}{3}, \quad \mathbf{x}^{(2)} = \frac{1}{3}\begin{bmatrix} 8 \\ 4 \\ 0 \end{bmatrix}, \quad \mathbf{r}^{(2)} = \frac{1}{3}\begin{bmatrix} 0 \\ 0 \\ 4 \end{bmatrix}, \quad \beta_1 = \frac{4}{9}, \quad \mathbf{p}^{(2)} = \frac{1}{9}\begin{bmatrix} 4 \\ 8 \\ 12 \end{bmatrix}, \\ &\alpha_2 = \frac{3}{4}, \quad \mathbf{x}^{(3)} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}, \quad \mathbf{r}^{(3)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \beta_2 = 0, \quad \mathbf{p}^{(3)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \end{aligned}$$

(c) By definition we have $\mathbb{W}_0 = \{\mathbf{0}\}$. From the solution of part (a) we know that $\mathbb{W}_k = \operatorname{span}(\mathbf{b}_0, \mathbf{A}\mathbf{b}_0, \dots, \mathbf{A}^{k-1}\mathbf{b}_0)$, where the vectors $\mathbf{b}$, $\mathbf{A}\mathbf{b}$ and $\mathbf{A}^2\mathbf{b}$ are linearly independent. Hence we have $\dim \mathbb{W}_k = k$ for $k = 0, 1, 2, 3$.

From (b) we know that the residual $\mathbf{r}^{(3)} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(3)} = \mathbf{0}$. Hence $\mathbf{x}^{(3)}$ is the exact solution to $\mathbf{A}\mathbf{x} = \mathbf{b}$.

We observe that $\mathbf{r}^{(0)} = 4\mathbf{e}_1$, $\mathbf{r}^{(1)} = 2\mathbf{e}_2$ and $\mathbf{r}^{(2)} = (4/3)\mathbf{e}_3$ and hence the $\mathbf{r}^{(k)}$ for $k = 0, 1, 2$ are linear independent and orthogonal to each other. Thus we are only left to show that $\mathbb{W}_k$ is the span of $\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(k-1)}$. We observe that $\mathbf{b} = \mathbf{r}^{(0)}$, $\mathbf{A}\mathbf{b} = 2\mathbf{r}^{(0)} - 2\mathbf{r}^{(1)}$ and $\mathbf{A}^2\mathbf{b} = 5\mathbf{r}^{(0)} - 8\mathbf{r}^{(1)} + 3\mathbf{r}^{(2)}$. Hence $\operatorname{span}(\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}\mathbf{b}^{k-1}) = \operatorname{span}(\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(k-1)})$ for $k = 1, 2, 3$. We conclude that, for $k = 1, 2, 3$, the vectors $\mathbf{r}^{(0)}, \dots, \mathbf{r}^{(k-1)}$ form an orthogonal basis for $\mathbb{W}_k$.

One can verify directly that $\mathbf{p}^{(0)}, \mathbf{p}^{(1)}$, and $\mathbf{p}^{(2)}$ are $\mathbf{A}$-orthogonal. Moreover, observing that $\mathbf{b} = \mathbf{p}^{(0)}$, $\mathbf{A}\mathbf{b} = (5/2)\mathbf{p}^{(0)} - 2\mathbf{p}^{(1)}$, and $\mathbf{A}^2\mathbf{b} = 7\mathbf{p}^{(0)} - (28/3)\mathbf{p}^{(1)} + 3\mathbf{p}^{(2)}$, it follows that

$$\operatorname{span}(\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}\mathbf{b}^{k-1}) = \operatorname{span}(\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(k-1)}), \quad \text{for } k = 1, 2, 3.$$

We conclude that, for $k = 1, 2, 3$, the vectors $\mathbf{p}^{(0)}, \dots, \mathbf{p}^{(k-1)}$ form an $\mathbf{A}$-orthogonal basis for $\mathbb{W}_k$.

By computing the Euclidean norms of $\mathbf{r}^{(0)}, \mathbf{r}^{(1)}, \mathbf{r}^{(2)}, \mathbf{r}^{(3)}$, we get

$$\|\mathbf{r}^{(0)}\|_2 = 4, \qquad \|\mathbf{r}^{(1)}\|_2 = 2, \qquad \|\mathbf{r}^{(2)}\|_2 = 4/3, \qquad \|\mathbf{r}^{(3)}\|_2 = 0.$$

It follows that the sequence $(\|\mathbf{r}^{(k)}\|)_k$ is monotonically decreasing.

Similarly, one finds

$$\left(\|\mathbf{x}^{(k)} - \mathbf{x}\|_2\right)_{k=0}^{3} = \left(\sqrt{10}, \sqrt{6}, \sqrt{14/9}, 0\right),$$

which is clearly monotonically decreasing.

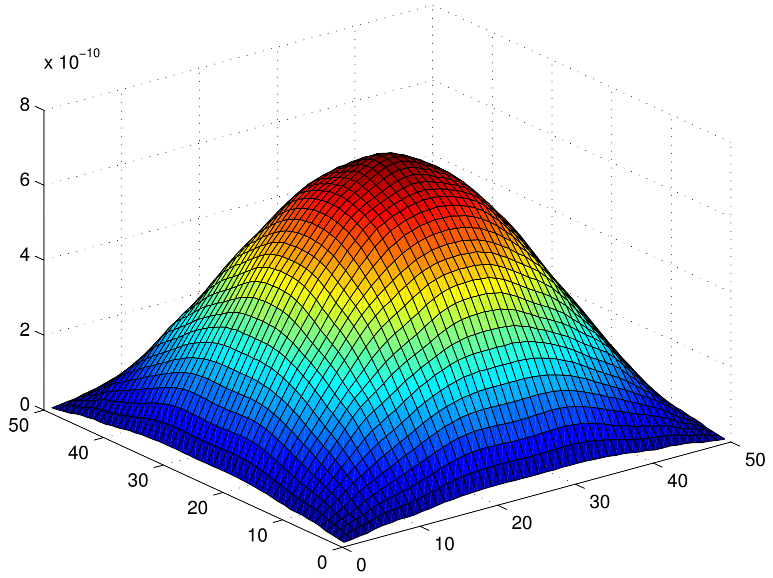


FIGURE 1. For a 50×50 Poisson matrix and a tolerance of 10^{-8}, the figure shows the difference of the outputs of cgtest and sdtest.

Exercise 10.24: Program code for testing steepest descent

Replacing the steps in (10.23) by those in (10.6), Algorithm 10.19 changes into the following algorithm for testing the method of Steepest Descent.

Listing 10.1. Testing the method of Steepest Descent

```
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;
```

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

```
[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 10.20 and 10.21, we let the tolerance be 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 $||\mathbf{r}_{K_{\text{sd}}}||_2 \leq \text{tol} \cdot ||\mathbf{r}_0||_2$. Choosing $a = 1/9$ and $d = 5/18$ yields the averaging matrix, and we find the following table.

n	2 500	10 000	40 000	1 000 000	4 000 000
K_{sd}	37	35	32	26	24

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

n	100	400	1 600	2 500	10 000	40 000
K_{sd}/n	4.1900	4.0325	3.9112	3.8832	3.8235	3.7863
K_{sd}	419	1 613	6 258	9 708	38 235	151 451
K_{J}	385			8 386		
K_{GS}	194			4 194		
K_{SOR}	35			164	324	645
K_{cg}	16	37	75	94	188	370

Here the number of iterations K_{J}, K_{GS}, and K_{SOR} of the Jacobi, Gauss-Seidel and SOR methods are taken from Table 9.1, 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 $\mathcal{O}(n)$ iterations for solving the Poisson problem for some given accuracy, as opposed to the $\mathcal{O}(\sqrt{n})$ 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.

As remarked below Theorem 10.22, the spectral condition number of the $m \times m$ Poisson matrix is $\kappa = \big(1+\cos(\pi h)\big)/\big(1-\cos(\pi h)\big)$. Theorem 10.8 therefore states that

$$(\star) \qquad \frac{||\mathbf{x}-\mathbf{x}_k||_{\mathbf{A}}}{||\mathbf{x}-\mathbf{x}_0||_{\mathbf{A}}} \leq \left(\frac{\kappa-1}{\kappa+1}\right)^k = \cos^k\left(\frac{\pi}{m+1}\right).$$

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

$$\frac{||\mathbf{x}||_{\mathbf{A}}^2}{||\mathbf{x}||_2^2} = \frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$$

is the Rayleigh quotient of $\mathbf{x}$, Lemma 6.42 implies the bound

$$\lambda_{\text{min}}||\mathbf{x}||_2^2 \leq ||\mathbf{x}||_{\mathbf{A}}^2 \leq \lambda_{\text{max}}||\mathbf{x}||_2^2,$$

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

$$\frac{||\mathbf{x}-\mathbf{x}_k||_2}{||\mathbf{x}-\mathbf{x}_0||_2} \leq \sqrt{\kappa}\left(\frac{\kappa-1}{\kappa+1}\right)^k = \sqrt{\frac{1+\cos\left(\frac{\pi}{m+1}\right)}{1-\cos\left(\frac{\pi}{m+1}\right)}}\cos^k\left(\frac{\pi}{m+1}\right).$$

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}.

Listing 10.2. Conjugate gradient method for least squares

```
function [x,K]=cg_leastSquares (A,b,x,tol,itmax)
r=b-A'*A*x; p=r;
rho=r'*r; rho0=rho;
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;
```

Exercise 10.25: Compare Richardson and steepest descent

Both the R method and the method of Steepest Descent are iteratively methods for approximating the solution to the equation $\mathbf{Ax} = \mathbf{b}$. In the k-th iteration, both methods find a better approximation $\mathbf{x}_k$ by moving along the same search direction, namely the residual $\mathbf{r}_{k-1} := \mathbf{b} - \mathbf{A}\mathbf{x}_{k-1}$.

The methods differ in how far they move along this direction. In the R method, the step length is the same for all iterations, and chosen to minimize the spectral radius of the matrix $\mathbf{G}(\alpha)$, thereby guaranteeing the fastest overall convergence. In the Steepest Descent method, on the other hand, the step length is not in general the same for each iteration. In each iteration, it is chosen so as to minimize the corresponding quadratic function $Q(y) = \frac{1}{2}\mathbf{y}^T\mathbf{A}\mathbf{y} - \mathbf{b}^T\mathbf{y}$ along this search direction. It is a *greedy algorithm*, in the sense that it makes the maximal improvement in each iteration.

Exercise 10.26: Using cg to solve normal equations

We need to perform Algorithm 10.18 with $\mathbf{A}^T\mathbf{A}$ replacing $\mathbf{A}$ and $\mathbf{A}^T\mathbf{b}$ replacing $\mathbf{b}$. For the system $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$, Equations (10.20), (10.21), and (10.22) become

$$\begin{aligned}
\mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} + \alpha_k \mathbf{p}^{(k)}, & \alpha_k &= \frac{\mathbf{r}^{(k)\,T}\mathbf{r}^{(k)}}{\mathbf{p}^{(k)\,T}\mathbf{A}^T\mathbf{A}\mathbf{p}^{(k)}} = \frac{\mathbf{r}^{(k)\,T}\mathbf{r}^{(k)}}{(\mathbf{A}\mathbf{p}^{(k)})^T\mathbf{A}\mathbf{p}^{(k)}}, \\
\mathbf{r}^{(k+1)} &= \mathbf{r}^{(k)} - \alpha_k \mathbf{A}^T\mathbf{A}\mathbf{p}^{(k)}, & & \\
\mathbf{p}^{(k+1)} &= \mathbf{r}^{(k+1)} + \beta_k \mathbf{p}^{(k)}, & \beta_k &= \frac{\mathbf{r}^{(k+1)\,T}\mathbf{r}^{(k+1)}}{\mathbf{r}^{(k)\,T}\mathbf{r}^{(k)}},
\end{aligned}$$

with $\mathbf{p}^{(0)} = \mathbf{r}^{(0)} = \mathbf{b} - \mathbf{A}^T\mathbf{A}\mathbf{x}^{(0)}$. Hence we only need to change the computation of $\mathbf{r}^{(0)}$, α_k, and $\mathbf{r}^{(k+1)}$ in Algorithm 10.18, which the implementation of Listing 10.2.

Exercise 10.31: An explicit formula for the Chebyshev polynomial (TODO)

CHAPTER 11

Orthonormal and Unitary Transformations

Exercise 11.2: Reflector

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

(a) Since

$$\mathbf{v}^T\mathbf{v} = \mathbf{x}^T\mathbf{x} - \mathbf{y}^T\mathbf{x} - \mathbf{x}^T\mathbf{y} + \mathbf{y}^t\mathbf{y} = 2l + 2\mathbf{y}^T\mathbf{x} = 2\mathbf{v}^T\mathbf{x},$$

we find

$$\left(\mathbf{I} - 2\frac{\mathbf{v}\mathbf{v}^T}{\mathbf{v}^T\mathbf{v}}\right)\mathbf{x} = \mathbf{x} - \frac{\mathbf{v}(2\mathbf{v}^T\mathbf{x})}{\mathbf{v}^T\mathbf{v}} = \mathbf{x} - \frac{\mathbf{v}\mathbf{v}^T\mathbf{v}}{\mathbf{v}^T\mathbf{v}} = \mathbf{x} - \mathbf{v} = \mathbf{y}.$$

(b) Since $\|\mathbf{x}\|_2^2 = \|\mathbf{y}\|_2^2$,

$$\langle \mathbf{x} - \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle - \langle \mathbf{y}, \mathbf{x} \rangle + \langle \mathbf{x}, \mathbf{y} \rangle - \langle \mathbf{y}, \mathbf{y} \rangle = \|\mathbf{x}\|_2^2 - \|\mathbf{y}\|_2^2 = 0,$$

which means that $\mathbf{x} - \mathbf{y}$ and $\mathbf{x} + \mathbf{y}$ are orthogonal. By Theorem 0.41, $\mathbf{Px}$ is the orthogonal projection of $\mathbf{x}$ into span$(\mathbf{x} + \mathbf{y})$, because it satisfies

$$\begin{aligned}\langle \mathbf{x} - \mathbf{Px}, \alpha(\mathbf{x} + \mathbf{y}) \rangle &= \left\langle \frac{\mathbf{v}^T\mathbf{x}\mathbf{v}}{\mathbf{v}^T\mathbf{v}}, \alpha(\mathbf{x} + \mathbf{y}) \right\rangle = \left\langle \frac{1}{2}\frac{\mathbf{v}^T\mathbf{v}}{\mathbf{v}^T\mathbf{v}}\mathbf{v}, \alpha(\mathbf{x} + \mathbf{y}) \right\rangle \\ &= \left\langle \frac{1}{2}(\mathbf{x} - \mathbf{y}), \alpha(\mathbf{x} + \mathbf{y}) \right\rangle = 0,\end{aligned}$$

for an arbitrary element $\alpha(\mathbf{x} + \mathbf{y})$ in span$(\mathbf{x} + \mathbf{y})$.

Exercise 11.5: What does Algorithm housegen do when $\mathbf{x} = \mathbf{e}_1$?

If $\mathbf{x} = \mathbf{e}_1$, then the algorithm yields $\alpha = -\|\mathbf{e}_1\|_2 = -1$,

$$\mathbf{u} = \frac{\mathbf{x}/\alpha - \mathbf{e}_1}{\sqrt{1 - x_1/\alpha}} = \frac{\mathbf{e}_1/(-1) - \mathbf{e}_1}{\sqrt{1 - 1/(-1)}} = -\sqrt{2}\mathbf{e}_1,$$

and

$$\mathbf{H} = \mathbf{I} - \mathbf{u}\mathbf{u}^T = \begin{bmatrix} -1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Exercise 11.6: Examples of Householder transformations

(a) Let $\mathbf{x}$ and $\mathbf{y}$ be as in the exercise. As $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2$, we can apply Exercise 11.2 to obtain a vector $\mathbf{v}$ and a matrix $\mathbf{Q}$,

$$\mathbf{v} = \mathbf{x} - \mathbf{y} = \begin{bmatrix} -2 \\ 4 \end{bmatrix}, \qquad \mathbf{Q} = \mathbf{I} - 2\frac{\mathbf{v}\mathbf{v}^t}{\mathbf{v}^t\mathbf{v}} = \frac{1}{5}\begin{bmatrix} 3 & 4 \\ 4 & -3 \end{bmatrix},$$

such that $\mathbf{Q}\mathbf{x} = \mathbf{y}$. As explained in the text above Exercise 11.2, this matrix $\mathbf{Q}$ is a Householder transformation with $\mathbf{u} := \sqrt{2}\mathbf{v}/\|\mathbf{v}\|_2$.

(b) Let $\mathbf{x}$ and $\mathbf{y}$ be as in the exercise. As $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2$, we can apply Exercise 11.2 to obtain a vector $\mathbf{v}$ and a Householder transformation $\mathbf{Q}$,

$$\mathbf{v} = \mathbf{x} - \mathbf{y} = \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}, \qquad \mathbf{Q} = \mathbf{I} - 2\frac{\mathbf{v}\mathbf{v}^t}{\mathbf{v}^t\mathbf{v}} = \frac{1}{3}\begin{bmatrix} -1 & 2 & -2 \\ 2 & 2 & 1 \\ -2 & 1 & 2 \end{bmatrix},$$

such that $\mathbf{Q}\mathbf{x} = \mathbf{y}$.

Exercise 11.7: 2×2 Householder transformation

Let $\mathbf{Q} = \mathbf{I} - \mathbf{u}\mathbf{u}^T \in \mathbb{R}^{2,2}$ be any Householder transformation. Then $\mathbf{u} = [u_1 \; u_2]^T \in \mathbb{R}^2$ is a vector satisfying $u_1^2 + u_2^2 = \|\mathbf{u}\|_2^2 = 2$, implying that the components of $\mathbf{u}$ are related via $u_1^2 - 1 = 1 - u_2^2$. Moreover, as $0 \le u_1^2, u_2^2 \le \|\mathbf{u}\|^2 = 2$, one has $-1 \le u_1^2 - 1 = 1 - u_2^2 \le 1$, 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 ϕ', one has

$$-u_1 u_2 = \pm\sqrt{1 + \cos\phi'}\sqrt{1 - \cos\phi'} = \pm\sqrt{1 - \cos^2\phi'} = \sin(\pm\phi').$$

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

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

Furthermore, we find

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

When applied to the vector $[\cos\phi, \sin\phi]^T$, therefore, $\mathbf{Q}$ doubles the angle and reflects the result in the y-axis.

Exercise 11.15: QR decomposition

That $\mathbf{Q}$ is orthonormal, and therefore unitary, can be shown directly by verifying that $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$. A direct computation shows that $\mathbf{Q}\mathbf{R} = \mathbf{A}$. Moreover,

$$\mathbf{R} = \begin{bmatrix} 2 & 2 \\ 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} =: \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{0}_{2,2} \end{bmatrix},$$

where $\mathbf{R}_1$ is upper triangular. It follows that $\mathbf{A} = \mathbf{Q}\mathbf{R}$ is a QR decomposition.

A QR factorization is obtained by removing the parts of $\mathbf{Q}$ and $\mathbf{R}$ that don't contribute anything to the product $\mathbf{QR}$. Thus we find a QR factorization

$$\mathbf{A} = \mathbf{Q}_1\mathbf{R}_1, \qquad \mathbf{Q}_1 := \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \end{bmatrix}, \qquad \mathbf{R}_1 := \begin{bmatrix} 2 & 2 \\ 0 & 2 \end{bmatrix}.$$

Exercise 11.16: Householder triangulation

(a) Let

$$\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3] = \begin{bmatrix} 1 & 0 & 1 \\ -2 & -1 & 0 \\ 2 & 2 & 1 \end{bmatrix}$$

be as in the Exercise. We wish to find Householder transformations $\mathbf{Q}_1, \mathbf{Q}_2$ that produce zeros in the columns $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ of $\mathbf{A}$. Applying Algorithm 11.4 to the first column of $\mathbf{A}$, we find

$$\mathbf{u}_1 = \frac{1}{\sqrt{3}}\begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}, \qquad \mathbf{Q}_1\mathbf{A} := (\mathbf{I} - \mathbf{u}_1\mathbf{u}_1^T)\mathbf{A} = \begin{bmatrix} -3 & -2 & -1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Next we need to map the bottom element $(\mathbf{Q}_1\mathbf{A})_{3,2}$ of the second column to zero, without changing the first row of $\mathbf{Q}_1\mathbf{A}$. For this, we apply Algorithm 11.4 to the vector $[0, 1]^T$ to find

$$\mathbf{u}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \text{and} \qquad \mathbf{Q}_2' := \mathbf{I} - \mathbf{u}_2\mathbf{u}_2^T = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix},$$

which is a Householder transformation of size 2 × 2. Since

$$\mathbf{Q}_2\mathbf{Q}_1\mathbf{A} := \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_2 \end{bmatrix}\mathbf{Q}_1\mathbf{A} = \begin{bmatrix} -3 & -2 & -1 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix},$$

it follows that the Householder transformations $\mathbf{Q}_1$ and $\mathbf{Q}_2$ bring $\mathbf{A}$ into upper triangular form.

(b) Clearly the matrix $\mathbf{Q}_3 := -\mathbf{I}$ is orthogonal and $\mathbf{R} := \mathbf{Q}_3\mathbf{Q}_2\mathbf{Q}_1\mathbf{A}$ is upper triangular with positive diagonal elements. It follows that

$$\mathbf{A} = \mathbf{QR}, \qquad \mathbf{Q} := \mathbf{Q}_1^T\mathbf{Q}_2^T\mathbf{Q}_3^T = \mathbf{Q}_1\mathbf{Q}_2\mathbf{Q}_3,$$

is a QR factorization of $\mathbf{A}$ of the required form.

Exercise 11.19: QR using Gram-Schmidt, II

Let

$$\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3] = \begin{bmatrix} 1 & 3 & 1 \\ 1 & 3 & 7 \\ 1 & -1 & -4 \\ 1 & -1 & 2 \end{bmatrix}.$$

Applying Gram-Schmidt orthogonalization, we find

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

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

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

Hence we have

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

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 11.21: Plane rotation

Suppose

$$\mathbf{x} = \begin{bmatrix} r\cos\alpha \\ r\sin\alpha \end{bmatrix}, \qquad \mathbf{P} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}.$$

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 11.22: Solving upper Hessenberg system using rotations

To determine the number of arithmetic operations of Algorithm 11.23, 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 $\mathbf{A}$ 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 $\mathbf{A}$, 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

$$
\begin{aligned}
&O(n^2) + \sum_{k=1}^{n-1} \bigl(4 + 2 + 6(n+1-k)\bigr) = O(n^2) + 6\sum_{k=1}^{n-1}(n+2-k) \\
&= O(n^2) + 3n^2 + 9n - 12 = O(n^2)
\end{aligned}
$$

arithmetic operations.

CHAPTER 12

Least Squares

Exercise 12.10: Straight line fit (linear regression)

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

(a) In this case $\mathbf{A} = [1, 1, \dots, 1]^T$, $\mathbf{x} = [x_1]$, and $\mathbf{b} = [y_1, y_2 \dots, y_m]^T$, implying that $\mathbf{A}^*\mathbf{A} = [m]$ and $\mathbf{A}^*\mathbf{b} = [y_1 + y_2 + \cdots + y_m]$. 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 $y_1, y_2, \dots, y_m$.

(b) In this case

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

so that

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

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

$$\sum_{i=1}^{m} (y(t) - y_i)^2$$

is minimal.

Exercise 12.11: Straight line fit using shifted power form (TODO)

Exercise 12.12: Fitting a circle to points

We are given the (in general overdetermined) system

$$(t_i - c_1)^2 + (y_i - c_2)^2 = r^2, \qquad i = 1, \dots, m.$$

(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 $i = 1, \dots, m$,

$$\begin{aligned} 0 &= (t_i - c_1)^2 + (y_i - c_2)^2 - r^2 \\ &= \left(t_i - \frac{x_1}{2}\right)^2 + \left(y_i - \frac{x_2}{2}\right)^2 - x_3 - \left(\frac{x_1}{2}\right)^2 - \left(\frac{x_2}{2}\right)^2 \\ &= t_i^2 + y_i^2 - t_i x_1 - y_i x_2 - x_3, \end{aligned}$$

from which Equation (12.8) follows immediately. Once x_1, x_2, and x_3 are determined, we can compute

$$c_1 = \frac{x_1}{2}, \qquad c_2 = \frac{x_2}{2}, \qquad r = \sqrt{\frac{1}{4}x_1^2 + \frac{1}{4}x_2^2 + x_3}.$$

(b) The linear least square problem is to minimize $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$, with

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

(c) Whether or not $\mathbf{A}$ 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 $\mathbf{A}$ are clearly dependent. In general, $\mathbf{A}$ has independent columns whenever we can find three points (t_i, y_i) not on a straight line.

(d) For these points the matrix $\mathbf{A}$ becomes

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

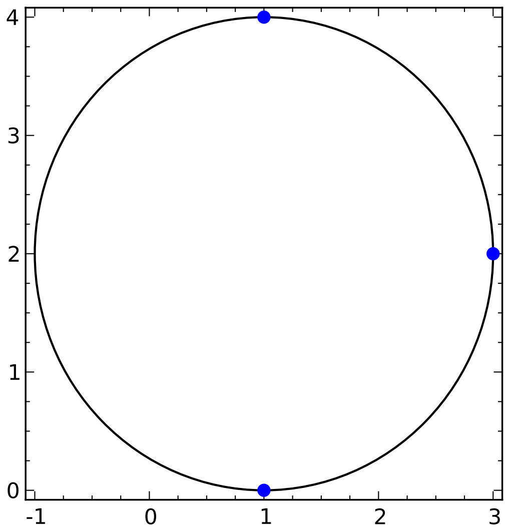
which clearly is invertible. We find

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

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 12.16: The generalized inverse

Let $\mathbf{A} \in \mathbb{C}^{m,n}$ be a matrix of rank r with singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ and corresponding singular value factorization $\mathbf{A} = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^*$. Define

$$\mathbf{B} = \mathbf{A}^\dagger := \mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{U}^*, \qquad \mathbf{\Sigma}^\dagger := \begin{bmatrix} \mathbf{\Sigma}_1^{-1} & \mathbf{0}_{r,m-r} \\ \mathbf{0}_{n-r,r} & \mathbf{0}_{n-r,m-r} \end{bmatrix} \in \mathbb{R}^{n,m}.$$

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

(1) $\mathbf{ABA} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*\mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{U}^*\mathbf{U}\mathbf{\Sigma}\mathbf{V}^* = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^\dagger\mathbf{\Sigma}\mathbf{V}^* = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^* = \mathbf{A}$

(2) $\mathbf{BAB} = \mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{U}^*\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*\mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{U}^* = \mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{\Sigma}\mathbf{\Sigma}^\dagger\mathbf{U}^* = \mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{U}^* = \mathbf{B}$

Moreover, since in addition the matrices $\mathbf{\Sigma}\mathbf{\Sigma}^\dagger$ and $\mathbf{\Sigma}^\dagger\mathbf{\Sigma}$ are Hermitian,

(3) $(\mathbf{BA})^* = \mathbf{A}^*\mathbf{B}^* = \mathbf{V}\mathbf{\Sigma}^*\mathbf{U}^*\mathbf{U}\mathbf{\Sigma}^{\dagger*}\mathbf{V}^* = \mathbf{V}\mathbf{\Sigma}^*\mathbf{\Sigma}^{\dagger*}\mathbf{V}^*$
$= \mathbf{V}(\mathbf{\Sigma}^\dagger\mathbf{\Sigma})^*\mathbf{V}^* = \mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{\Sigma}\mathbf{V}^* = \mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{U}^*\mathbf{U}\mathbf{\Sigma}\mathbf{V}^* = \mathbf{BA}$

(4) $(\mathbf{AB})^* = \mathbf{B}^*\mathbf{A}^* = \mathbf{U}\mathbf{\Sigma}^{\dagger*}\mathbf{V}^*\mathbf{V}\mathbf{\Sigma}^*\mathbf{U}^* = \mathbf{U}\mathbf{\Sigma}^{\dagger*}\mathbf{\Sigma}^*\mathbf{U}^*$
$= \mathbf{U}(\mathbf{\Sigma}\mathbf{\Sigma}^\dagger)^*\mathbf{U}^* = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^\dagger\mathbf{U}^* = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*\mathbf{V}\mathbf{\Sigma}^\dagger\mathbf{U}^* = \mathbf{AB}$

Exercise 12.17: 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), (1_B), (3_C), (2_C).

Exercise 12.18: Verify that a matrix is a generalized inverse

Let

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}, \qquad \mathbf{B} = \frac{1}{4}\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

be as in the Exercise. One finds

$$\mathbf{AB} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix} \frac{1}{4}\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$\mathbf{BA} = \frac{1}{4}\begin{bmatrix}1 & 1 & 0\\1 & 1 & 0\end{bmatrix}\begin{bmatrix}1 & 1\\1 & 1\\0 & 0\end{bmatrix} = \frac{1}{2}\begin{bmatrix}1 & 1\\1 & 1\end{bmatrix},$$

so that $(\mathbf{AB})^* = \mathbf{AB}$ and $(\mathbf{BA})^* = \mathbf{BA}$. Moreover,

$$\mathbf{ABA} = \mathbf{A}(\mathbf{BA}) = \begin{bmatrix}1 & 1\\1 & 1\\0 & 0\end{bmatrix}\frac{1}{2}\begin{bmatrix}1 & 1\\1 & 1\end{bmatrix} = \begin{bmatrix}1 & 1\\1 & 1\\0 & 0\end{bmatrix} = \mathbf{A},$$

$$\mathbf{BAB} = (\mathbf{BA})\mathbf{B} = \frac{1}{2}\begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}\frac{1}{4}\begin{bmatrix}1 & 1 & 0\\1 & 1 & 0\end{bmatrix} = \frac{1}{4}\begin{bmatrix}1 & 1 & 0\\1 & 1 & 0\end{bmatrix} = \mathbf{B}.$$

By Exercises 12.16 and 12.17, we conclude that $\mathbf{B}$ must be the pseudoinverse of $\mathbf{A}$.

Exercise 12.19: Linearly independent columns and generalized inverse

If $\mathbf{A} \in \mathbb{C}^{m,n}$ has independent columns then both $\mathbf{A}$ and $\mathbf{A}^*$ have rank $n \le m$. Then, by Theorem 7.17, $\mathbf{A}^*\mathbf{A}$ must have rank n as well. Since $\mathbf{A}^*\mathbf{A}$ is an $n \times n$-matrix of maximal rank, it is nonsingular and we can define $\mathbf{B} := (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$. We verify that $\mathbf{B}$ satisfies the four axioms of Exercise 12.16.

(1) $\mathbf{ABA} = \mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{A} = \mathbf{A}$

(2) $\mathbf{BAB} = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^* = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^* = \mathbf{B}$

(3) $(\mathbf{BA})^* = \big((\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{A}\big)^* = \mathbf{I}_n^* = \mathbf{I}_n = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{A} = \mathbf{BA}$

(4) $(\mathbf{AB})^* = \big(\mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\big)^* = \mathbf{A}\big((\mathbf{A}^*\mathbf{A})^{-1}\big)^*\mathbf{A}^*$
$= \mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^* = \mathbf{AB}$

It follows that $\mathbf{B} = \mathbf{A}^\dagger$. The second claim follows similarly.

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

Exercise 12.20: The generalized inverse of a vector

This is a special case of Exercise 12.19. In particular, if $\mathbf{u}$ is a nonzero vector, then $\mathbf{u}^*\mathbf{u} = \langle\mathbf{u},\mathbf{u}\rangle = \|\mathbf{u}\|^2$ is a nonzero number and $(\mathbf{u}^*\mathbf{u})^{-1}\mathbf{u}^*$ is defined. One can again check the axioms of Exercise 12.16 to show that this vector must be the pseudoinverse of $\mathbf{u}^*$.

Exercise 12.21: The generalized inverse of an outer product

Let $\mathbf{A} = \mathbf{uv}^*$ be as in the Exercise. Since $\mathbf{u}$ and $\mathbf{v}$ are nonzero, the matrix

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

is well defined. We verify the four axioms of Exercise 12.16 to show that $\mathbf{B}$ must be the pseudoinverse of $\mathbf{A}$.

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

(2) $\mathbf{BAB} = \dfrac{\mathbf{vu}^*\mathbf{uv}^*\mathbf{vu}^*}{\|\mathbf{u}\|_2^4\|\mathbf{v}\|_2^4} = \dfrac{\mathbf{v}\|\mathbf{u}\|_2^2\|\mathbf{v}\|_2^2\mathbf{u}^*}{\|\mathbf{u}\|_2^4\|\mathbf{v}\|_2^4} = \dfrac{\mathbf{vu}^*}{\|\mathbf{u}\|_2^2\|\mathbf{v}\|_2^2} = \mathbf{B};$

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

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

This proves that $\mathbf{B}$ is the pseudoinverse of $\mathbf{A}$.

Exercise 12.22: The generalized inverse of a diagonal matrix

Let $\mathbf{A} := \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ and $\mathbf{B} := \operatorname{diag}(\lambda_1^\dagger, \ldots, \lambda_n^\dagger)$ as in the exercise. Note that, by definition, $\lambda_j^\dagger$ indeed represents the pseudoinverse of the number λ_j for any j. It therefore satisfies the axioms of Exercise 12.16, something we shall use below. We now verify the axioms for $\mathbf{B}$ to show that $\mathbf{B}$ must be the pseudoinverse of $\mathbf{A}$.

(1) $\mathbf{ABA} = \operatorname{diag}(\lambda_1\lambda_1^\dagger\lambda_1, \ldots, \lambda_n\lambda_n^\dagger\lambda_n) = \operatorname{diag}(\lambda_1, \ldots, \lambda_n) = \mathbf{A}$;

(2) $\mathbf{BAB} = \operatorname{diag}(\lambda_1^\dagger\lambda_1\lambda_1^\dagger, \ldots, \lambda_n^\dagger\lambda_n\lambda_n^\dagger) = \operatorname{diag}(\lambda_1^\dagger, \ldots, \lambda_n^\dagger) = \mathbf{B}$;

(3) $(\mathbf{BA})^* = (\operatorname{diag}(\lambda_1^\dagger\lambda_1, \ldots, \lambda_n^\dagger\lambda_n))^* = \operatorname{diag}(\lambda_1^\dagger\lambda_1, \ldots, \lambda_n^\dagger\lambda_n) = \mathbf{BA}$;

(4) $(\mathbf{AB})^* = (\operatorname{diag}(\lambda_1\lambda_1^\dagger, \ldots, \lambda_n\lambda_n^\dagger))^* = \operatorname{diag}(\lambda_1\lambda_1^\dagger, \ldots, \lambda_n\lambda_n^\dagger) = \mathbf{AB}$.

This proves that $\mathbf{B}$ is the pseudoinverse of $\mathbf{A}$.

Exercise 12.23: Properties of the generalized inverse (TODO)

Exercise 12.24: The generalized inverse of a product

(a) From the condition that $\mathbf{A}$ has linearly independent columns we can deduce that $n \le m$. Similarly it follows that $n \le k$, hence $n \le \min\{m, k\}$ and both matrices have maximal rank. As a consequence,

$$\mathbf{A} = \mathbf{U_A}\boldsymbol{\Sigma}_\mathbf{A}\mathbf{V}_\mathbf{A}^* = \begin{bmatrix}\mathbf{U}_{\mathbf{A},1} & \mathbf{U}_{\mathbf{A},2}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{A},1}\\ \mathbf{0}\end{bmatrix}\mathbf{V}_\mathbf{A}^*,$$

$$\mathbf{B} = \mathbf{U_B}\begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{B},1} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}_{\mathbf{B},1} & \mathbf{V}_{\mathbf{B},2}\end{bmatrix}^*.$$

This gives

$$\begin{aligned}\mathbf{A}^\dagger\mathbf{A} &= \mathbf{V_A}\begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{A},1}^{-1} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{U}_{\mathbf{A},1} & \mathbf{U}_{\mathbf{A},2}\end{bmatrix}^*\begin{bmatrix}\mathbf{U}_{\mathbf{A},1} & \mathbf{U}_{\mathbf{A},2}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{A},1}\\ \mathbf{0}\end{bmatrix}\mathbf{V}_\mathbf{A}^*\\ &= \mathbf{V_A}\boldsymbol{\Sigma}_{\mathbf{A},1}^{-1}\boldsymbol{\Sigma}_{\mathbf{A},1}\mathbf{V}_\mathbf{A}^* = \mathbf{I}\end{aligned}$$

and

$$\begin{aligned}\mathbf{B}\mathbf{B}^\dagger &= \mathbf{U_B}\begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{B},1} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}_{\mathbf{B},1} & \mathbf{V}_{\mathbf{B},2}\end{bmatrix}^*\begin{bmatrix}\mathbf{V}_{\mathbf{B},1} & \mathbf{V}_{\mathbf{B},2}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{B},1}^{-1}\\ \mathbf{0}\end{bmatrix}\mathbf{U}_\mathbf{B}^*\\ &= \mathbf{U_B}\boldsymbol{\Sigma}_{\mathbf{B},1}\boldsymbol{\Sigma}_{\mathbf{B},1}^{-1}\mathbf{U}_\mathbf{B}^* = \mathbf{I}.\end{aligned}$$

Moreover we get

$$\begin{aligned}(\mathbf{A}\mathbf{A}^\dagger)^* &= \begin{bmatrix}\mathbf{U}_{\mathbf{A},1} & \mathbf{U}_{\mathbf{A},2}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{A},1}^{-1}\\ \mathbf{0}\end{bmatrix}\mathbf{V}_\mathbf{A}^*\mathbf{V_A}\begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{A},1} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{U}_{\mathbf{A},1} & \mathbf{U}_{\mathbf{A},2}\end{bmatrix}^*\\ &= \begin{bmatrix}\mathbf{U}_{\mathbf{A},1} & \mathbf{U}_{\mathbf{A},2}\end{bmatrix}\begin{bmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{U}_{\mathbf{A},1} & \mathbf{U}_{\mathbf{A},2}\end{bmatrix}^*\\ &= \begin{bmatrix}\mathbf{U}_{\mathbf{A},1} & \mathbf{U}_{\mathbf{A},2}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{A},1}\\ \mathbf{0}\end{bmatrix}\mathbf{V}_\mathbf{A}^*\mathbf{V_A}\begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{A},1}^{-1} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{U}_{\mathbf{A},1} & \mathbf{U}_{\mathbf{A},2}\end{bmatrix}^*\\ &= \mathbf{A}\mathbf{A}^\dagger\end{aligned}$$

and

$$\begin{aligned}(\mathbf{B}^\dagger\mathbf{B})^* &= \begin{bmatrix}\mathbf{V}_{\mathbf{B},1} & \mathbf{V}_{\mathbf{B},2}\end{bmatrix}\begin{bmatrix}\mathbf{\Sigma}_{\mathbf{B},1}\\ \mathbf{0}\end{bmatrix}\mathbf{U}_\mathbf{B}^*\mathbf{U}_\mathbf{B}\begin{bmatrix}\mathbf{\Sigma}_{\mathbf{B},1}^{-1} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}_{\mathbf{B},1} & \mathbf{V}_{\mathbf{B},2}\end{bmatrix}^*\\ &= \begin{bmatrix}\mathbf{V}_{\mathbf{B},1} & \mathbf{V}_{\mathbf{B},2}\end{bmatrix}\begin{bmatrix}\mathbf{\Sigma}_{\mathbf{B},1}^{-1}\\ \mathbf{0}\end{bmatrix}\mathbf{U}_\mathbf{B}^*\mathbf{U}_\mathbf{B}\begin{bmatrix}\mathbf{\Sigma}_{\mathbf{B},1} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}_{\mathbf{B},1} & \mathbf{V}_{\mathbf{B},2}\end{bmatrix}^*\\ &= \mathbf{B}^\dagger\mathbf{B}.\end{aligned}$$

We now let $\mathbf{E} := \mathbf{AB}$ and $\mathbf{F} := \mathbf{B}^\dagger\mathbf{A}^\dagger$. Hence we want to show that $\mathbf{E}^\dagger = \mathbf{F}$. We do that by showing that $\mathbf{F}$ satisfies the properties given in Exercise 12.16.

$$\begin{aligned}\mathbf{EFE} &= \mathbf{ABB}^\dagger\mathbf{A}^\dagger\mathbf{AB} = \mathbf{AB} = \mathbf{E}\\ \mathbf{FEF} &= \mathbf{B}^\dagger\mathbf{A}^\dagger\mathbf{ABB}^\dagger\mathbf{A}^\dagger = \mathbf{B}^\dagger\mathbf{A}^\dagger = \mathbf{F}\\ (\mathbf{FE})^* &= (\mathbf{B}^\dagger\mathbf{A}^\dagger\mathbf{AB})^* = (\mathbf{B}^\dagger\mathbf{B})^* = \mathbf{B}^\dagger\mathbf{B} = \mathbf{B}^\dagger\mathbf{A}^\dagger\mathbf{AB} = \mathbf{FE}\\ (\mathbf{EF})^* &= (\mathbf{ABB}^\dagger\mathbf{A}^\dagger)^* = (\mathbf{AA}^\dagger)^* = \mathbf{AA}^\dagger = \mathbf{ABB}^\dagger\mathbf{A}^\dagger = \mathbf{EF}\end{aligned}$$

(b) We have

$$\mathbf{A} = \begin{bmatrix}a & b\end{bmatrix} \qquad \text{and} \qquad \mathbf{B} = \begin{bmatrix}c\\ d\end{bmatrix}.$$

This gives

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix}a^2 & ab\\ ab & b^2\end{bmatrix} \qquad \text{and} \qquad \mathbf{B}^T\mathbf{B} = c^2 + d^2.$$

Hence we want to choose a and b such that $\mathbf{A}^T\mathbf{A}$ 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

$$\mathbf{A} = \begin{bmatrix}a & 0\end{bmatrix}, \qquad \mathbf{A}^T\mathbf{A} = \begin{bmatrix}a^2 & 0\\ 0 & 0\end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix}3\\ 4\end{bmatrix}, \qquad \mathbf{B}^T\mathbf{B} = 25 = 5^2.$$

We hence can derive the following singular value decompositions and pseudoinverse.

$$\mathbf{A} = \begin{bmatrix}1\end{bmatrix}\begin{bmatrix}a & 0\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}, \qquad \mathbf{A}^\dagger = \frac{1}{a}\begin{bmatrix}1\\ 0\end{bmatrix},$$

$$\mathbf{B} = \frac{1}{5}\begin{bmatrix}3 & -4\\ 4 & 3\end{bmatrix}\begin{bmatrix}5\\ 0\end{bmatrix}\begin{bmatrix}1\end{bmatrix}, \qquad \mathbf{B}^\dagger = \frac{1}{25}\begin{bmatrix}3 & 4\end{bmatrix}.$$

We thus get

$$(\mathbf{AB})^\dagger = (3a)^\dagger = \frac{1}{3a} \qquad \text{and} \qquad \mathbf{B}^\dagger\mathbf{A}^\dagger = \frac{3}{25a},$$

and have

$$(\mathbf{AB})^\dagger = \frac{1}{3a} \neq \frac{9}{25}\cdot\frac{1}{3a} = \frac{3}{25a} = \mathbf{B}^\dagger\mathbf{A}^\dagger$$

for all nonzero $a \in \mathbb{R}$.

Exercise 12.25: The generalized inverse of the conjugate transpose (TODO)

Exercise 12.26: Linearly independent columns

By Exercise 12.19, if $\mathbf{A}$ has rank n, then $\mathbf{A}^\dagger = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$. Then $\mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{b} = \mathbf{A}\mathbf{A}^\dagger\mathbf{b}$, which is the orthogonal projection of $\mathbf{b}$ into $\operatorname{span}(\mathbf{A})$ by Theorem 12.13.

Exercise 12.27: Analysis of the general linear system

In this exercise, we can write

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \qquad \mathbf{\Sigma}_1 = \operatorname{diag}(\sigma_1, \dots, \sigma_r), \qquad \sigma_1 > \cdots > \sigma_r > 0.$$

(a) As $\mathbf{U}$ is unitary, we have $\mathbf{U}^*\mathbf{U} = \mathbf{I}$. We find the following sequence of equivalences.

$$\mathbf{A}\mathbf{x} = \mathbf{b} \iff \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*\mathbf{x} = \mathbf{b} \iff \mathbf{U}^*\mathbf{U}\mathbf{\Sigma}(\mathbf{V}^*\mathbf{x}) = \mathbf{U}^*\mathbf{b} \iff \mathbf{\Sigma}\mathbf{y} = \mathbf{c},$$

which is what needed to be shown.

(b) By (a), the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ has a solution if and only if the system

$$\begin{bmatrix} \mathbf{\Sigma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{y} = \begin{bmatrix} \sigma_1 y_1 \\ \vdots \\ \sigma_r y_r \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} c_1 \\ \vdots \\ c_r \\ c_{r+1} \\ \vdots \\ c_n \end{bmatrix} = \mathbf{c}$$

has a solution $\mathbf{y}$. Since $\sigma_1, \dots, \sigma_r \neq 0$, this system has a solution if and only if $c_{r+1} = \cdots = c_n = 0$. We conclude that $\mathbf{A}\mathbf{x} = \mathbf{b}$ has a solution if and only if $c_{r+1} = \cdots = c_n = 0$.

(c) The linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ has a solution if and only if the system $\mathbf{\Sigma}\mathbf{y} = \mathbf{c}$ has a solution. Hence we have the following three cases.

$r = n$:
Here $y_i = c_i/\sigma_i$ for $i = 1, \dots, n$ provides the only solution to the system $\mathbf{\Sigma}\mathbf{y} = \mathbf{b}$, and therefore $\mathbf{x} = \mathbf{V}\mathbf{y}$ is the only solution to $\mathbf{A}\mathbf{x} = \mathbf{b}$. It follows that the system has exactly one solution.

$r < n$, $c_i = 0$ for $i = r + 1, \dots, n$:
Here each solution $\mathbf{y}$ must satisfy $y_i = c_i/\sigma_i$ for $i = 1, \dots, r$. The remaining $y_{r+1}, \dots, y_n$, however, can be chosen arbitrarily. Hence we have infinitely many solutions to $\mathbf{\Sigma}\mathbf{y} = \mathbf{b}$ as well as for $\mathbf{A}\mathbf{x} = \mathbf{b}$.

$r < n$, $c_i \neq 0$ for some i with $r + 1 \leq i \leq n$:
In this case it is impossible to find a $\mathbf{y}$ that satisfies $\mathbf{\Sigma}\mathbf{y} = \mathbf{b}$, and therefore the system $\mathbf{A}\mathbf{x} = \mathbf{b}$ has no solution at all.

Exercise 12.28: Fredholm's Alternative

Assume $\mathbf{b} \in \operatorname{span}(\mathbf{A})$. This means that the system $\mathbf{A}\mathbf{x} = \mathbf{b}$ has a solution. Let $\mathbf{y} \in \ker(\mathbf{A}^*)$ so that $\mathbf{A}^*\mathbf{y} = \mathbf{0}$. We know that $(\operatorname{span}(\mathbf{A}))^\perp = \ker(\mathbf{A}^*)$, hence $\langle \mathbf{y}, \mathbf{b} \rangle = \mathbf{y}^*\mathbf{b} = 0$. Thus if the system $\mathbf{A}\mathbf{x} = \mathbf{b}$ has a solution, then we can not find a $\mathbf{y} \in \ker(\mathbf{A}^*)$ such that $\mathbf{y}^*\mathbf{b} = 0$. Conversely if $\mathbf{y} \in \ker(\mathbf{A}^*)$ and $\mathbf{y}^*\mathbf{b} = 0$ than $\mathbf{b}$ can not be in the $\operatorname{span}(\mathbf{A})$ and thus the system $\mathbf{A}\mathbf{x} = \mathbf{b}$ can not have a solution.

Exercise 12.36: Condition number

Let

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

be as in the Exercise.

(a) By Exercise 12.19, the pseudoinverse of $\mathbf{A}$ is

$$\mathbf{A}^\dagger = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T = \begin{bmatrix} -1 & 1 & 1 \\ 1 & -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}.$$

Theorem 12.13 tells us that the orthogonal projection of $\mathbf{b}$ into $\operatorname{span}(\mathbf{A})$ is

$$\mathbf{b}_1 := \mathbf{A}\mathbf{A}^\dagger\mathbf{b} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 2b_1 \\ b_2 + b_3 \\ b_2 + b_3 \end{bmatrix},$$

while the orthogonal projection of $\mathbf{b}$ into $\ker(\mathbf{A}^T)$ is

$$\mathbf{b}_2 := (\mathbf{I} - \mathbf{A}\mathbf{A}^\dagger)\mathbf{b} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{2} & -\frac{1}{2} \\ 0 & -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 0 \\ b_2 - b_3 \\ b_3 - b_2 \end{bmatrix}.$$

(b) By Theorem 8.12, the 2-norms $\|\mathbf{A}\|_2$ and $\|\mathbf{A}^\dagger\|_2$ can be found by computing the largest singular values of the matrices $\mathbf{A}$ and $\mathbf{A}^\dagger$. The largest singular value σ_1 of $\mathbf{A}$ is the square root of the largest eigenvalue λ_1 of $\mathbf{A}^T\mathbf{A}$, which satisfies

$$0 = \det(\mathbf{A}^T\mathbf{A} - \lambda_1\mathbf{I}) = \det \begin{bmatrix} 3 - \lambda_1 & 4 \\ 4 & 6 - \lambda_1 \end{bmatrix} = \lambda_1^2 - 9\lambda_1 + 2.$$

It follows that $\sigma_1 = \frac{1}{2}\sqrt{2}\sqrt{9 + \sqrt{73}}$. Similarly, the largest singular value σ_2 of $\mathbf{A}^\dagger$ is the square root of the largest eigenvalue λ_2 of $\mathbf{A}^{\dagger T}\mathbf{A}^\dagger$, which satisfies

$$0 = \det(\mathbf{A}^{\dagger T}\mathbf{A}^\dagger - \lambda_2\mathbf{I}) = \det\left(\frac{1}{4}\begin{bmatrix} 8 & -6 & -6 \\ -6 & 5 & 5 \\ -6 & 5 & 5 \end{bmatrix} - \lambda_2\mathbf{I}\right)$$
$$= -\frac{1}{2}\lambda_2\big(2\lambda_2^2 - 9\lambda_2 + 1\big).$$

Alternatively, we could have used that the largest singular value of $\mathbf{A}^\dagger$ is the inverse of the smallest singular value of $\mathbf{A}$ (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

$$K(\mathbf{A}) = \|\mathbf{A}\|_2 \cdot \|\mathbf{A}^\dagger\|_2 = \sqrt{\frac{9 + \sqrt{73}}{9 - \sqrt{73}}} = \frac{1}{2\sqrt{2}}\left(9 + \sqrt{73}\right) \approx 6.203.$$

Exercise 12.37: Equality in perturbation bound (TODO)

Exercise 12.39: Problem using normal equations

(a) Let $\mathbf{A}$, $\mathbf{b}$, and ε be as in the exercise. The normal equations $\mathbf{A}^t\mathbf{A}\mathbf{x} = \mathbf{A}^t\mathbf{b}$ 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 $\mathbf{A}^t\mathbf{A}$ yields the unique solution

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \frac{1}{2\varepsilon^2} \begin{bmatrix} (\varepsilon+1)^2+2 & -3-\varepsilon \\ -3-\varepsilon & 3 \end{bmatrix} \begin{bmatrix} 7 \\ 7+2\varepsilon \end{bmatrix} = \begin{bmatrix} \frac{5}{2}+\frac{1}{2\varepsilon} \\ -\frac{1}{2\varepsilon} \end{bmatrix}.$$

If $\varepsilon = 0$, on the other hand, then any vector $\mathbf{x} = [x_1, x_2]^t$ 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,

$$\begin{aligned} \left\| \mathbf{A} \begin{bmatrix} \frac{5}{2}+\frac{1}{2\varepsilon} \\ -\frac{1}{2\varepsilon} \end{bmatrix} - \mathbf{b} \right\|_2 = \left\| \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ 0 \end{bmatrix} \right\|_2 &= \frac{1}{\sqrt{2}} \\ &\leq \sqrt{2} = \left\| \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} \right\|_2 = \left\| \mathbf{A} \begin{bmatrix} 2-\frac{1}{\varepsilon} \\ \frac{1}{\varepsilon} \end{bmatrix} - \mathbf{b} \right\|_2, \end{aligned}$$

which shows that the solution from (a) is more accurate.

CHAPTER 13

Numerical Eigenvalue Problems

Exercise 13.7: Continuity of eigenvalues (TODO)

Exercise 13.9: Nonsingularity using Gerschgorin

We compute the Gerschgorin disks

$$R_1 = R_4 = C_1 = C_4 = \{z \in \mathbb{C} : |z-4| \le 1\},$$
$$R_2 = R_3 = C_2 = C_3 = \{z \in \mathbb{C} : |z-4| \le 2\}.$$

Then, by Gerschgorin's Circle Theorem, each eigenvalue of $\mathbf{A}$ lies in

$$(R_1 \cup \cdots \cup R_4) \cap (C_1 \cup \cdots \cup C_4) = \{z \in \mathbb{C} : |z-4| \le 2\}.$$

In particular $\mathbf{A}$ has only nonzero eigenvalues, implying that $\mathbf{A}$ must be nonsingular.

Exercise 13.10: Gerschgorin, strictly diagonally dominant matrix

Suppose $\mathbf{A}$ is a strictly diagonally dominant matrix. For such a matrix, one finds Gerschgorin disks

$$R_i = \left\{ z \in \mathbb{C} : |z - a_{ii}| \le \sum_{j \ne i} |a_{ij}| \right\}$$

Since $|a_{ii}| > \sum_{j\ne i} |a_{ij}|$ 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 Gerschgorin's Circle Theorem, $\mathbf{A}$ only has nonzero eigenvalues and must be nonsingular.

Exercise 13.12: 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 $(n-k)^2$ substractions. 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

$$N \sim \int_0^n \big(4(n-k)^2 + 4n(n-k)\big)\, dk = \frac{10}{3} n^3$$

arithmetic operations.

Exercise 13.14: 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 $(n-k)^2$ substractions. In total we find $4(n-k)^2$ 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. 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 13.15: Tridiagonalize a symmetric matrix

From $\mathbf{w} = \mathbf{E}\mathbf{v}$, $\beta = \frac{1}{2}\mathbf{v}^T\mathbf{w}$ and $\mathbf{z} = \mathbf{w} - \beta\mathbf{v}$ we get $\mathbf{z} = \mathbf{w} - \mathbf{v}\beta = \mathbf{E}\mathbf{v} - \frac{1}{2}\mathbf{v}\mathbf{v}^T\mathbf{E}\mathbf{v}$ and $\mathbf{z}^T = \mathbf{v}^T\mathbf{E} - \frac{1}{2}\mathbf{v}^T\mathbf{E}\mathbf{v}\mathbf{v}^T$. Using this yields

$$\begin{aligned}
\mathbf{G} &= (\mathbf{I} - \mathbf{v}\mathbf{v}^T)\mathbf{E}(\mathbf{I} - \mathbf{v}\mathbf{v}^T) = \mathbf{E} - \mathbf{v}\mathbf{v}^T\mathbf{E} - \mathbf{E}\mathbf{v}\mathbf{v}^T + \mathbf{v}\mathbf{v}^T\mathbf{E}\mathbf{v}\mathbf{v}^T \\
&= \mathbf{E} - \mathbf{v}(\mathbf{v}^T\mathbf{E} - \frac{1}{2}\mathbf{v}^T\mathbf{E}\mathbf{v}\mathbf{v}^T) - (\mathbf{E}\mathbf{v} - \frac{1}{2}\mathbf{v}\mathbf{v}^T\mathbf{E}\mathbf{v})\mathbf{v}^T \\
&= \mathbf{E} - \mathbf{v}\mathbf{z}^T - \mathbf{z}\mathbf{v}^T.
\end{aligned}$$

Exercise 13.19: Counting eigenvalues

Let

$$\mathbf{A} = \begin{bmatrix} 4 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 4 \end{bmatrix}, \qquad \alpha = 4.5.$$

Applying the recursive procedure described in Corollary 13.18, we find the diagonal elements $d_1(\alpha), d_2(\alpha), d_3(\alpha), d_4(\alpha)$ of the matrix $\mathbf{D}$ in the factorization $\mathbf{A} - \alpha\mathbf{I} = \mathbf{L}\mathbf{D}\mathbf{L}^t$,

$$\begin{aligned}
d_1(\alpha) &= 4 - 9/2 = -1/2, \\
d_2(\alpha) &= 4 - 9/2 - 1^2/(-1/2) = +3/2, \\
d_3(\alpha) &= 4 - 9/2 - 1^2/(+3/2) = -7/6,
\end{aligned}$$

$$d_4(\alpha) = 4 - 9/2 - 1^2/(-7/6) = +5/14.$$

As precisely two of these are negative, Corollary 13.18 implies that there are precisely two eigenvalues of $\mathbf{A}$ strictly smaller than $\alpha = 4.5$. As

$$\det(\mathbf{A} - 4.5\mathbf{I}) = \det(\mathbf{LDL}^t) = d_1(\alpha)d_2(\alpha)d_3(\alpha)d_4(\alpha) \neq 0,$$

the matrix $\mathbf{A}$ does not have an eigenvalue equal to 4.5. We conclude that the remaining two eigenvalues must be bigger than 4.5.

Exercise 13.20: Overflow in 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 $\mathbf{D}_n$ in a symmetric factorization of $\mathbf{A}_n$.

(a) We proceed by induction. Let $n \geq 1$ be any positive integer. For the first diagonal element, corresponding to $k = 1$, Equations (2.4) immediately yield $5 + \sqrt{24} < d_{n,1} = 10 \leq 10$. Next, assume that $5 + \sqrt{24} < d_{n,k} \leq 10$ for some $1 \leq k < n$. We shall show that this implies that $5 + \sqrt{24} < d_{n,k+1} \leq 10$. First observe that $\big(5 + \sqrt{24}\big)^2 = 25 + 10\sqrt{24} + 24 = 49 + 10\sqrt{24}$. From Equations (2.4) 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

$$d_{n,k+1} = 10 - \frac{1}{d_{n,k}} > 10 - \frac{1}{5 + \sqrt{24}} = \frac{50 + 10\sqrt{24} - 1}{5 + \sqrt{24}} = 5 + \sqrt{24}.$$

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 $\mathbf{A} = \mathbf{LDL}^T$ with $\mathbf{L}$ triangular and with ones on the diagonal. As a consequence,

$$\det(\mathbf{A}) = \det(\mathbf{L})\det(\mathbf{D})\det(\mathbf{L}) = \det(\mathbf{D}) = \prod_{i=1}^{n} d_i > \big(5 + \sqrt{24}\big)^n.$$

In Matlab an overflow is indicated by Matlab returning `Inf` as result. At my computer this happens at $n_0 = 310$.

Exercise 13.21: Simultaneous diagonalization

Let $\mathbf{A}, \mathbf{B}, \mathbf{U}, \mathbf{D}, \hat{\mathbf{A}}$, and $\mathbf{D}^{-1/2}$ be as in the Exercise.

(a) Since $\mathbf{D}^{-\frac{1}{2}}$, like any diagonal matrix, and $\mathbf{A}$ are symmetric, one has

$$\hat{\mathbf{A}}^T = \mathbf{D}^{-\frac{1}{2}}{}^T \mathbf{U}\mathbf{A}^T\mathbf{U}^T\mathbf{D}^{-\frac{1}{2}}{}^T = \mathbf{D}^{-\frac{1}{2}}\mathbf{U}\mathbf{A}\mathbf{U}^T\mathbf{D}^{-\frac{1}{2}} = \hat{\mathbf{A}}$$

(b) Since $\hat{\mathbf{A}}$ is symmetric, it admits an orthogonal diagonalization $\hat{\mathbf{A}} = \hat{\mathbf{U}}^T\hat{\mathbf{D}}\hat{\mathbf{U}}$. Let $\mathbf{E} := \mathbf{U}^T\mathbf{D}^{-\frac{1}{2}}\hat{\mathbf{U}}^T$, $\mathbf{F} := \hat{\mathbf{U}}\mathbf{D}^{\frac{1}{2}}\mathbf{U}$, and observe that $\mathbf{D}^{\frac{1}{2}}\mathbf{D}^{-\frac{1}{2}} = \mathbf{I}$. Then

$$\mathbf{FE} = \hat{\mathbf{U}}\mathbf{D}^{\frac{1}{2}}\mathbf{U}\mathbf{U}^T\mathbf{D}^{-\frac{1}{2}}\hat{\mathbf{U}}^T = \hat{\mathbf{U}}\mathbf{D}^{\frac{1}{2}}\mathbf{D}^{-\frac{1}{2}}\hat{\mathbf{U}}^T = \hat{\mathbf{U}}\hat{\mathbf{U}}^T = \mathbf{I}$$

and similar $\mathbf{EF} = \mathbf{I}$. Hence $\mathbf{E}^{-1} = \mathbf{F}$ and $\mathbf{E}$ is nonsingular. Moreover, from $\hat{\mathbf{A}} = \hat{\mathbf{U}}^T\hat{\mathbf{D}}\hat{\mathbf{U}}$ follows that $\hat{\mathbf{U}}\hat{\mathbf{A}}\hat{\mathbf{U}}^T = \hat{\mathbf{D}}$, which gives

$$\mathbf{E}^T\mathbf{AE} = \hat{\mathbf{U}}\mathbf{D}^{-\frac{1}{2}}\mathbf{U}\mathbf{A}\mathbf{U}^T\mathbf{D}^{-\frac{1}{2}}\hat{\mathbf{U}}^T = \hat{\mathbf{U}}\hat{\mathbf{A}}\hat{\mathbf{U}}^T = \hat{\mathbf{D}}.$$

Similarly $\mathbf{B} = \mathbf{U}^T\mathbf{DU}$ implies $\mathbf{UBU}^T = \mathbf{D}$, which yields

$$\mathbf{E}^T\mathbf{BE} = \hat{\mathbf{U}}\mathbf{D}^{-\frac{1}{2}}\mathbf{U}\mathbf{B}\mathbf{U}^T\mathbf{D}^{-\frac{1}{2}}\hat{\mathbf{U}}^T = \hat{\mathbf{U}}\mathbf{D}^{-\frac{1}{2}}\mathbf{D}^{\frac{1}{2}}\mathbf{D}^{\frac{1}{2}}\mathbf{D}^{-\frac{1}{2}}\hat{\mathbf{U}}^T = \mathbf{I}.$$

We conclude that for a symmetric matrix $\mathbf{A}$ and symmetric positive definite matrix $\mathbf{B}$, the congruence transformation $\mathbf{X} \longmapsto \mathbf{E}^T\mathbf{X}\mathbf{E}$ simultaneously diagonalizes the matrices $\mathbf{A}$ and $\mathbf{B}$, and even maps $\mathbf{B}$ to the identity matrix.

Exercise 13.22: Program code for one eigenvalue

(a) Let $\mathbf{A} = \text{tridiag}(\mathbf{c}, \mathbf{d}, \mathbf{c})$ and x be as in the Exercise. The following Matlab program counts the number of eigenvalues k of $\mathbf{A}$ strictly less than x.

```
function k=count(c,d,x)
n = length(d);
k = 0; u = d(1)-x;
if u < 0
    k = k+1;
end
for i = 2:n
    umin = abs(c(i-1))*eps;
    if abs(u) < umin
        if u < 0
      u = -umin;
  else
            u = umin;
        end
    end
    u = d(i)-x-c(i-1)^2/u;
    if u < 0
        k = k+1;
    end
end
```

(b) Let $\mathbf{A} = \text{tridiag}(\mathbf{c}, \mathbf{d}, \mathbf{c})$ and m be as in the Exercise. The following Matlab program computes a small interval $[a, b]$ around the mth eigenvalue λ_m of $\mathbf{A}$ and returns the point λ in the middle of this interval.

```
function lambda = findeigv(c,d,m)
n = length(d);
a = d(1)-abs(c(1)); b = d(1)+abs(c(1));
for i = 2:n-1
    a = min(a, d(i)-abs(c(i-1))-abs(c(i)));
    b = max(b, d(i)+abs(c(i-1))+abs(c(i)));
end
a = min(a, d(n)-abs(c(n-1)));
b = max(b, d(n)+abs(c(n-1)));
h = b-a;
while abs(b-a) > eps*h
    c0 = (a+b)/2;
    k = count(c,d,c0);
    if k < m
        a = c0;
    else
        b = c0;
    end
end
lambda = (a+b)/2;
```

(c) The following table shows a comparison between the values and errors obtained by the different methods.

method	value	error
exact	0.02413912051848666	0
`findeigv`	0.02413912051848621	$4.44 \cdot 10^{-16}$
Matlab `eig`	0.02413912051848647	$1.84 \cdot 10^{-16}$

Exercise 13.23: Determinant of upper Hessenberg matrix (TODO)

Exercise 13.25: ∞-norm of a diagonal matrix

Let $\mathbf{A} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ be a diagonal matrix. The spectral radius $\rho(\mathbf{A})$ coincides with the biggest eigenvalue, say λ_i, of $\mathbf{A}$. One has

$$\|\mathbf{A}\|_\infty = \max_{\|\mathbf{x}\|_\infty = 1} \|\mathbf{A}\mathbf{x}\|_\infty = \max_{\|\mathbf{x}\|_\infty = 1} \max\{|\lambda_1 x_1|, \ldots, |\lambda_n x_n|\} \leq \rho(\mathbf{A}),$$

as $\lambda_1, \ldots, \lambda_n \leq \lambda_i = \rho(\mathbf{A})$ and since the components of any vector $\mathbf{x}$ satisfy $x_1, \ldots, x_n \leq \|\mathbf{x}\|_\infty$. Moreover, this bound is attained for the standard basis vector $\mathbf{x} = \mathbf{e}_i$, since $\|\mathbf{A}\mathbf{e}_i\|_\infty = \lambda_i = \rho(\mathbf{A})$.

CHAPTER 14

The QR Algorithm

Exercise 14.3: Orthogonal vectors

In the Exercise it is implicitly assumed that $\mathbf{u}^*\mathbf{u} \neq 0$ and therefore $\mathbf{u} \neq 0$. If $\mathbf{u}$ and $\mathbf{A}\mathbf{u} - \lambda\mathbf{u}$ are orthogonal, then

$$0 = \langle \mathbf{u}, \mathbf{A}\mathbf{u} - \lambda\mathbf{u} \rangle = \mathbf{u}^*(\mathbf{A}\mathbf{u} - \lambda\mathbf{u}) = \mathbf{u}^*\mathbf{A}\mathbf{u} - \lambda\mathbf{u}^*\mathbf{u}.$$

Dividing by $\mathbf{u}^*\mathbf{u}$ yields

$$\lambda = \frac{\mathbf{u}^*\mathbf{A}\mathbf{u}}{\mathbf{u}^*\mathbf{u}}.$$

Exercise 14.13: QR convergence detail (TODO)