Numerical Linear Algebra
A Solution Manual

Georg Muntingh and Christian Schulz

Contents

CHAPTER 0

A Short Review of Linear Algebra

Exercise 0.25: 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 0.26: The inverse of a 2×2 matrix

By Exercise 0.25, 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 0.27: Sherman-Morrison formula

A direct computation yields

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

showing that the two matrices are inverse to each other.

Exercise 0.29: 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 0.30: 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

$$\begin{aligned}\mathrm{adj}_{\mathbf{A}}^{\mathrm{T}} &= \begin{bmatrix} (-1)^{1+1}\begin{vmatrix}-2 & -6\\ 3 & 2\end{vmatrix} & (-1)^{1+2}\begin{vmatrix}3 & -6\\ 6 & 2\end{vmatrix} & (-1)^{1+3}\begin{vmatrix}3 & -2\\ 6 & 3\end{vmatrix} \\ (-1)^{2+1}\begin{vmatrix}-6 & 3\\ 3 & 2\end{vmatrix} & (-1)^{2+2}\begin{vmatrix}2 & 3\\ 6 & 2\end{vmatrix} & (-1)^{2+3}\begin{vmatrix}2 & -6\\ 6 & 3\end{vmatrix} \\ (-1)^{3+1}\begin{vmatrix}-6 & 3\\ -2 & -6\end{vmatrix} & (-1)^{3+2}\begin{vmatrix}2 & 3\\ 3 & -6\end{vmatrix} & (-1)^{3+3}\begin{vmatrix}2 & -6\\ 3 & -2\end{vmatrix} \end{bmatrix} \\ &= \begin{bmatrix} 14 & 21 & 42 \\ -42 & -14 & 21 \\ 21 & -42 & 14 \end{bmatrix}^{\mathrm{T}}.\end{aligned}$$

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

Exercise 0.31: 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 0.32: 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 1, 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 0.33: 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

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

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

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

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

Exercise 0.34: Cauchy determinant

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

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

which is what needed to be shown.

Exercise 0.35: 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 0.34.(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

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

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

Diagonally dominant tridiagonal matrices; three examples

Exercise 1.12: The shifted power basis is a basis

We know that the set of polynomials of degree n is a vector space of dimension $n+1$: They are spanned by $\{x^k\}_{k=0}^n$, and these are linearly independent (if a linear combination of these is zero, then it has in particular $n+1$ zeros (since every x is a zero), and it follows from the fundamental theorem of algebra that the linear combination must be zero). Since the shifted power basis also has $n+1$ vectors which are polynomials, all we need to show is that they are linearly independent. Suppose then that

$$\sum_{j=0}^{n} a_j (x - x_i)^j = 0.$$

In particular we can then pick $n+1$ distinct values z_k for x so that this is zero. But then the polynomial $\sum_{j=0}^n a_j x^j$ has the $n+1$ different zeros $z_k - x_i$. Since the $\{x^k\}_{k=0}^n$ are linearly independent, it follows that all $a_j = 0$, so that the shifted power basis also is a basis.

Exercise 1.25: 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,

$$\begin{aligned}
&(\mathbf{LU})_{11} = 1 \cdot 2 = 2; \\
&(\mathbf{LU})_{ii} = -\frac{i-1}{i} \cdot -1 + 1 \cdot \frac{i+1}{i} = 2, && \text{for } i = 2, \ldots, m; \\
&(\mathbf{LU})_{i,i-1} = -\frac{i-1}{i} \cdot \frac{i}{i-1} = -1, && \text{for } i = 2, \ldots, m; \\
&(\mathbf{LU})_{i-1,i} = 1 \cdot -1 = -1, && \text{for } i = 2, \ldots, m; \\
&(\mathbf{LU})_{ij} = 0, && \text{for } |i-j| \geq 2.
\end{aligned}$$

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

One can also show this by induction using the `trifactor`-algorithm. Since $\mathbf{T}$ and $\mathbf{U}$ have the same super-diagonal, we must have $c_m = -1$ for all m. Assume now that $\mathbf{L}_m \mathbf{U}_m = \mathbf{T}_m$, and that $l_{m-1} = -(m-1)/m$ and $u_m = (m+1)/m$. From the `trifactor`-algorithm,

$$\begin{aligned}
l_m &= a_m / u_m = -1/((m+1)/m) = -m/(m+1) \\
u_{m+1} &= d_{m+1} - l_m c_m = 2 - m/(m+1) = (m+2)/(m+1).
\end{aligned}$$

This shows that the `trifactor`-algorithm produces the desired terms in $\mathbf{L}_{m+1}$ and $\mathbf{U}_{m+1}$ as well.

Another way to show this by induction is as follows. 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

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

block multiplication yields

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

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

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

Exercise 1.26: Inverse of 2nd derivative matrix

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

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

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 $1 \leq j < i \leq m$, we find

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

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 1.27: 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 1.28: Two point boundary value problem

(a) For $j = 1, \ldots, m$, we get when we gather terms that

$$h^2 f(x_j) = \left(-1 - \frac{h}{2} r(x_j)\right) v_{j-1} + (2 + h^2 q(x_j)) v_j + \left(-1 + \frac{h}{2} r(x_j)\right) v_{j+1}$$

From this we get the desired formula for a_j, c_j, and d_j, and the right hand sides b_j for $2 \leq j \leq m-1$.
For $j = 1$, since v_0 is known we have to move $\left(-1 - \frac{h}{2} r(x_0)\right) v_0 = a_1 g_0$ over to the right hand side, so that we obtain $b_0 = h^2 f(x_1) - a_1 g_0$.
For $j = m$, since v_{m+1} is known we have to move $\left(-1 + \frac{h}{2} r(x_m)\right) v_{m+1} = c_m g_1$ over to the right hand side, so that we obtain $b_m = h^2 f(x_m) - c_m g_1$. This leads to the tridiagonal system $\mathbf{A}\mathbf{v} = \mathbf{b}$ in the exercise.

(b) One has When $h|r(x)|/2 < 1$ for all $x \in [a, b]$, we see that $a, j, c_j \in (-2, 0)$. It follows that $|a_j| + |c_j| = 1 + \frac{h}{2} r(x_m) + 1 + \frac{h}{2} r(x_m) = 2$. Since $q(x_j) \geq 0$, $|d_j| = d_j \geq 2$, so that $\mathbf{A}$ is weakly diagonally dominant. Since $|c_j| = 1 + \frac{h}{2} r(x_j) < 2$, and $|d_j| > 2$ it follows in particular that $|d_1| > |c_1|$. Clearly also all $a_j > 0$ since $h|r(x)|/2 < 1$, and since also $|d_j| > 2$, in particular $d_n \neq 0$, so that all the conditions in the theorem are fulfilled.

(c) We can use the method `trisolve` to find the $v_1, \ldots, v_m$. Note that the indexing of the a_j should be shifted with one in this exercise, to be compatible with the notation used in `tridiag(`a_j`,` d_j`,` c_j`)` (a_j and d_j have the same index when they are in the same column of the matrix. In this exercise they have the same index when they are in the same row).

Exercise 1.29: Two point boundary value problem; computation

(a) and (c) The provided values for r, f, q give that $a_j = c_j = -1$, $d_j = 2 + h^2$. The initial conditions are $g_0 = 1$, $g_1 = 0$, so that $\mathbf{b} = (h^2 + 1, h^2, \ldots, h^2)$. The code can look as follows

```
for m = [9 19 39 79, 159]
    h = 1/(m+1);
    x = h*(1:m)';
    [l, u] = trifactor( -ones(1, m - 1), (2+h^2)*ones(1, m), -ones(1, m - 1));
    b = h^2*ones(m, 1); b(1) = b(1) + 1;
    v = trisolve(l, u, -ones(1, m - 1), b);
    err = max(abs( (1-sinh(x)/sinh(1)) - v))
    log(err)/log(h)
end
```

The code also solves (c); If the error is proportional to h^p, then $err = Ch^p$ for some C. But then $p = (\log(err) - \log C)/\log h \approx \log(err)/\log h$ for small h, which is the quantity computed inside the `for`-loop. It seems that this converges to 3, so that one would guess that the error is proportional to h^3.

(b)

```
m = 9
h = 1/(m+1);
x = h*(1:m)';
```

```
[l, u] = trifactor( -ones(1, m - 1), (2+h^2)*ones(1, m), -ones(1, m - 1));
b = h^2*ones(m, 1); b(1) = b(1) + 1;
v = trisolve(l, u, -ones(1, m - 1), b);
plot(x, (1-sinh(x)/sinh(1)), x, v)
legend('Exact solution', 'Estimated solution')
```

Exercise 1.30: Approximate force

Since $\sin x$ has Taylor series $x - x^3/3! + x^5/5! - \ldots$, We have that $\sin(\pi h/2) = \pi h/2 + O(h^3)$. If we square both sides we obtain $\sin^2(\pi h/2) = \pi^2 h^2/4 + O(h^4)$. From this we obtain that $4\sin^2(\pi h/2)R/(h^2L^2) = \pi^2 R/L^2 + O(h^2)$.

Exercise 1.38: 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,

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

Exercise 1.39: Outer product expansion of a matrix

Clearly $\mathbf{e}_i\mathbf{e}_j^{\mathrm{T}}$ is the matrix $E_{i,j}$ with 1 at entry (i,j), and zero elsewhere. Clearly also $A = \sum_{i,j} a_{i,j}E_{i,j} = \sum_{i,j} a_{i,j}\mathbf{e}_i\mathbf{e}_j^{\mathrm{T}}$.

Exercise 1.40: The product $\mathbf{A}^{\mathrm{T}}\mathbf{A}$

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

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

which is what was needed to be shown.

Exercise 1.41: Outer product expansion

Recall that the matrix product of $\mathbf{A} \in \mathbb{C}^{m,n}$ and $\mathbf{B}^{\mathrm{T}} = \mathbf{C} \in \mathbb{C}^{n,p}$ 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 $\left(\mathbf{a}_{:k}\mathbf{b}_{:k}^{\mathrm{T}}\right)_{ij} = a_{ik}b_{jk}$. It follows that

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

Exercise 1.42: 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 $\mathbf{A}\mathbf{x}_{\cdot j} = \mathbf{b}_{\cdot j}$ for $j = 1, \ldots, p$.

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

$$\begin{aligned}\mathbf{AX} &= \mathbf{AXI} = \mathbf{AX}[\mathbf{e}_1, \ldots, \mathbf{e}_p] = [\mathbf{AXe}_1, \ldots, \mathbf{AXe}_p] \\ &= [\mathbf{A}\mathbf{x}_{\cdot 1}, \ldots, \mathbf{A}\mathbf{x}_{\cdot p}] = [\mathbf{b}_{\cdot 1}, \ldots, \mathbf{b}_{\cdot p}] = \mathbf{B}.\end{aligned}$$

Exercise 1.43: 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 1.44: Another block multiplication example

Since the matrices have compatible dimensions, a direct computation gives

$$\mathbf{CAB} = \begin{bmatrix}1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & \mathbf{C}_1\end{bmatrix}\begin{bmatrix}\lambda & \mathbf{a}^{\mathrm{T}} \\ \mathbf{0} & \mathbf{A}_1\end{bmatrix}\begin{bmatrix}1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & \mathbf{B}_1\end{bmatrix} = \begin{bmatrix}\lambda & \mathbf{a}^{\mathrm{T}} \\ \mathbf{0} & \mathbf{C}_1\mathbf{A}_1\end{bmatrix}\begin{bmatrix}1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & \mathbf{B}_1\end{bmatrix} = \begin{bmatrix}\lambda & \mathbf{a}^{\mathrm{T}}\mathbf{B}_1 \\ \mathbf{0} & \mathbf{C}_1\mathbf{A}_1\mathbf{B}_1\end{bmatrix}.$$

CHAPTER 2

Gaussian eliminations and LU Factorizations

Exercise 2.8: Column oriented backsolve

If $\mathbf{A}$ is upper triangular, suppose that we after $n-k$ steps of the algorithm have reduced our system to one of the form

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,k} \\ 0 & a_{2,1} & \cdots & a_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{k,k} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix}$$

Clearly then $x_k = b_k/a_{k,k}$ (this explains the first statement inside the `for`-loop). Eliminating the x_k-variable we obtain the system

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,k-1} \\ 0 & a_{2,1} & \cdots & a_{2,k-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{k-1,k-1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{k-1} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{k-1} \end{bmatrix} - x_k \begin{bmatrix} a_{1,k} \\ a_{2,k} \\ \vdots \\ a_{k-1,k} \end{bmatrix}.$$

This means that the right hand side $\mathbf{b}$ should be updated by subtracting `A(1:(k-1),k)*x(k)`. If A is d-banded, $A_{1,k} = \cdots = A_{k-d-1,k} = 0$, so that this is the same as subtracting `A(lk:(k-1),k)*x(k)` with `lk` being the maximum of 1 and $k-d$. This explains the second part inside the `for`-loop. Finally we end up with a 1×1-matrix, so to find x_1 we only need to divide with $a_{1,1}$.

Exercise 2.11: 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 says that $\mathbf{A}\mathbf{b}_k = \mathbf{e}_k$. Let $\mathbf{b}_k = (b_{1k}, \ldots, b_{nk})^{\mathrm{T}}$. Since the last $n-k$ components of $\mathbf{e}_k$ are 0, back subsitution yields that $b_{k+1,k} = \ldots = b_{n,k} = 0$, so that $\mathbf{B}$ is upper triangular (as stated also by Lemma 1.35). Splitting $\mathbf{A}$ into blocks $\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ 0 & \mathbf{A}_{22} \end{bmatrix}$ where $\mathbf{A}_{11}$ has size $k \times k$ ($\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are then upper triangular),

we get

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ 0 & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} b_{1k} \\ \vdots \\ b_{kk} \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \left[\begin{array}{c} \mathbf{A}_{11} \begin{bmatrix} b_{1k} \\ \vdots \\ b_{kk} \end{bmatrix} \\ \hline 0 \end{array}\right] = \left[\begin{array}{c} \mathbf{e}_k \\ \hline 0 \end{array}\right],$$

so that we need to solve

$$\mathbf{A}_{11} \begin{bmatrix} b_{1k} \\ \vdots \\ b_{kk} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1,k} \\ 0 & \ddots & \vdots \\ 0 & 0 & a_{k,k} \end{bmatrix} \begin{bmatrix} b_{1k} \\ \vdots \\ b_{kk} \end{bmatrix} = \mathbf{e}_k. \tag{1}$$

This yields (2.10) for solving for the kth column of $\mathbf{B}$ (note that the Matlab notation $I(1:k,k)$ yields $\mathbf{e}_k$).

Let us consider the number of arithmetic operations needed to compute the inverse. In finding $\mathbf{b}_k$ we need to solve a $k \times k$ triangular system. Solving for x_1 we need to compute $k-1$ multiplications, $k-2$ additions, and one division. This gives a total number of $2k-2$ arithmetic operations. Solving for x_2 needs $2k-4$ operations, and so on, all the way down to x_{k-1} which needs 2 operations. Solving for $x_k = 1/a_{k,k}$ needs an additional division, so that we need to perform

$$1 + \sum_{r=1}^{k-1} 2r = 1 + (k-1)k$$

operations. Since we solve a triangular system for any $1 \le k \le n$, we end up with a total of

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

arithmetic operations. Here we used the formulas we deduced in Exercise 2.13.

Usually we are just interesting in the "leading term" for the number of operations (here $n^3/3$). This can be obtained more simply by approximating the sums with integrals as in the book: solving the $k \times k$ triangular system can be solved in $1 + \sum_{r=1}^{k-1} 2r \approx \int_{r=1}^{k-1} 2r \approx (k-1)^2 \approx k^2$ operations, and adding together the number of operations for all k we obtain $\sum_{k=1}^{n} k^2 \approx \int_{k=1}^{n} k^2 dk \approx n^3/3$ operations.

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. A Matlab implementation which stores the inverse (in-place) in $\mathbf{A}$ can thus look as follows:

```
n = 8;
A = rand(n);
A = triu(A);
U=A;
for k=n:-1:1
    U(k,k) = 1/U(k,k);
    for r=k-1:-1:1
```

```
        U(r, k) = -U(r,r+1:k)*U(r+1:k,k)/U(r,r);
    end
end
U*A
```

A Python implementation can look as follows:

```
from numpy import *

n = 8
A = matrix(random.random((n,n)))
A=triu(A)
U=A.copy()
for k in range(n-1,-1,-1):
    U[k,k] = 1/U[k,k]
    for r in range(k-1,-1,-1):
        U[r, k] = -U[r, (r+1):(k+1)]*U[(r+1):(k+1),k]/U[r,r]
print U*A
```

In the code, `r` and `k` are row- and column indices, respecively. Inside the `for`-loop we compute x_r for the system in Equation (1). The contribution from $x_{r+1}, \ldots, x_k$ can be written as a dot product, which here is computed as a matrix product (the minus sign comes from that we isolate x_r on the left hand side). Note that k goes from n and downwards. If we did this the other way we would overwrite matrix entries needed for later calculations.

Exercise 2.13: Finite sums of integers

There are many ways to prove these identities. The quickest is perhaps by induction. We choose instead an approach based on what is called a generating function. This approach does not assume knowledge of the sum-expressions we want to derive, and the approach also works in a wide range of other circumstances.

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

$$P_m(x) := 1 + x + \cdots + x^m = \frac{1 - x^{m+1}}{1 - x}.$$

Then

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

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

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 (2.12). In addition it follows that

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

which establishes (2.14). 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 (2.15). 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 (2.13).

Exercise 2.14: Multiplying triangular matrices

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

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

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

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

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

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

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

Exercise 2.23: 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 2.24: LU and determinant

Suppose $\mathbf{A}$ has an LU factorization $\mathbf{A} = \mathbf{LU}$. Then, by Lemma 2.16, $\mathbf{A}_{[k]} = \mathbf{L}_{[k]}\mathbf{U}_{[k]}$ 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 2.25: Diagonal elements in U

From Exercise 2.24, 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 2.31: Making a block LU into an LU

We can write a block LU factorization of $\mathbf{A}$ as

$$\mathbf{A} = \mathbf{LU} = \begin{bmatrix} I & 0 & \cdots & 0 \\ \mathbf{L}_{21} & I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{L}_{m1} & \mathbf{L}_{m2} & \cdots & I \end{bmatrix} \begin{bmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} & \cdots & \mathbf{U}_{1m} \\ 0 & \mathbf{U}_{22} & \cdots & \mathbf{U}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & \mathbf{U}_{mm} \end{bmatrix}$$

(i.e. the blocks are denoted $\mathbf{L}_{ij}$, $\mathbf{U}_{ij}$). We now assume that $\mathbf{U}_{ii}$ has an LU factorization $\tilde{\mathbf{L}}_{ii}\tilde{\mathbf{U}}_{ii}$ ($\tilde{\mathbf{L}}_{ii}$ unit lower triangular, $\tilde{\mathbf{U}}_{ii}$ upper triangular), and define $\hat{\mathbf{L}} = \mathbf{L}\mathrm{diag}(\tilde{\mathbf{L}}_{ii})$, $\hat{\mathbf{U}} = \mathrm{diag}(\tilde{\mathbf{L}}_{ii}^{-1})\mathbf{U}$. We get that

$$\begin{aligned} \hat{\mathbf{L}} = \mathbf{L}\mathrm{diag}(\tilde{\mathbf{L}}_{ii}) &= \begin{bmatrix} I & 0 & \cdots & 0 \\ \mathbf{L}_{21} & I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{L}_{m1} & \mathbf{L}_{m2} & \cdots & I \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{L}}_{11} & 0 & \cdots & 0 \\ 0 & \tilde{\mathbf{L}}_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tilde{\mathbf{L}}_{mm} \end{bmatrix} \\ &= \begin{bmatrix} \tilde{\mathbf{L}}_{11} & 0 & \cdots & 0 \\ \mathbf{L}_{21}\tilde{\mathbf{L}}_{11} & \tilde{\mathbf{L}}_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{L}_{m1}\tilde{\mathbf{L}}_{11} & \mathbf{L}_{m2}\tilde{\mathbf{L}}_{22} & \cdots & \tilde{\mathbf{L}}_{mm} \end{bmatrix} \end{aligned}$$

This shows that $\hat{\mathbf{L}}$ has the blocks $\tilde{\mathbf{L}}_{ii}$ on the diagonal, and since these are unit lower triangular, it follows that also $\hat{\mathbf{L}}$ is unit lower triangular. Also,

$$\begin{aligned} \hat{\mathbf{U}} = \mathrm{diag}(\tilde{\mathbf{L}}_{ii}^{-1})\mathbf{U} &= \begin{bmatrix} \tilde{\mathbf{L}}_{11}^{-1} & 0 & \cdots & 0 \\ 0 & \tilde{\mathbf{L}}_{22}^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tilde{\mathbf{L}}_{mm}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} & \cdots & \mathbf{U}_{1m} \\ 0 & \mathbf{U}_{22} & \cdots & \mathbf{U}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & \mathbf{U}_{mm} \end{bmatrix} \\ &= \begin{bmatrix} \tilde{\mathbf{L}}_{11}^{-1}\mathbf{U}_{11} & \tilde{\mathbf{L}}_{11}^{-1}\mathbf{U}_{12} & \cdots & \tilde{\mathbf{L}}_{11}^{-1}\mathbf{U}_{1m} \\ 0 & \tilde{\mathbf{L}}_{22}^{-1}\mathbf{U}_{22} & \cdots & \tilde{\mathbf{L}}_{22}^{-1}\mathbf{U}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & \tilde{\mathbf{L}}_{mm}^{-1}\mathbf{U}_{mm} \end{bmatrix} \\ &= \begin{bmatrix} \tilde{\mathbf{L}}_{11}^{-1}\tilde{\mathbf{L}}_{11}\tilde{\mathbf{U}}_{11} & \tilde{\mathbf{L}}_{11}^{-1}\mathbf{U}_{12} & \cdots & \tilde{\mathbf{L}}_{11}^{-1}\mathbf{U}_{1m} \\ 0 & \tilde{\mathbf{L}}_{22}^{-1}\tilde{\mathbf{L}}_{22}\tilde{\mathbf{U}}_{22} & \cdots & \tilde{\mathbf{L}}_{22}^{-1}\mathbf{U}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & \tilde{\mathbf{L}}_{mm}^{-1}\tilde{\mathbf{L}}_{mm}\tilde{\mathbf{U}}_{mm} \end{bmatrix} \\ &= \begin{bmatrix} \tilde{\mathbf{U}}_{11} & \tilde{\mathbf{L}}_{11}^{-1}\mathbf{U}_{12} & \cdots & \tilde{\mathbf{L}}_{11}^{-1}\mathbf{U}_{1m} \\ 0 & \tilde{\mathbf{U}}_{22} & \cdots & \tilde{\mathbf{L}}_{22}^{-1}\mathbf{U}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & \tilde{\mathbf{U}}_{mm} \end{bmatrix} \end{aligned}$$

where we inserted $\mathbf{U}_{ii} = \tilde{\mathbf{L}}_{ii}\tilde{\mathbf{U}}_{ii}$. This shows $\hat{\mathbf{U}}$ has the blocks $\tilde{\mathbf{U}}_{ii}$ on the diagonal, and since these are upper triangular, it follows that also $\hat{\mathbf{U}}$ is upper triangular.

Exercise 2.36: Using PLU of A to solve $\mathbf{A}^{\mathrm{T}}\mathbf{x} = \mathbf{b}$

If $\mathbf{A} = \mathbf{PLR}$, then $\mathbf{A}^{\mathrm{T}} = \mathbf{R}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}$. The matrix $\mathbf{L}^{\mathrm{T}}$ is upper triangular and the matrix $\mathbf{R}^{\mathrm{T}}$ is lower triangular, implying that $\mathbf{R}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}}$ is an LU factorization of $\mathbf{A}^{\mathrm{T}}\mathbf{P}$. Since $\mathbf{A}$ is nonsingular, the matrix $\mathbf{R}^{\mathrm{T}}$ must be nonsingular, and we can apply Algorithms 2.6 and 2.7 to economically solve the systems $\mathbf{R}^{\mathrm{T}}\mathbf{z} = \mathbf{b}$, $\mathbf{L}^{\mathrm{T}}\mathbf{y} = \mathbf{z}$, and $\mathbf{P}^{\mathrm{T}}\mathbf{x} = \mathbf{y}$, to find a solution $\mathbf{x}$ to the system $\mathbf{R}^{\mathrm{T}}\mathbf{L}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\mathbf{x} = \mathbf{A}^{\mathrm{T}}\mathbf{x} = \mathbf{b}$.

Exercise 2.37: 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 2.38: Using PLU to compute the inverse

Solving an $n \times n$-triangular system takes n^2 operations, as is clear from the `rforwardsolve` and `rbacksolve` algorithms. From Exercise 2.11 it is thus clear that inverting an upper /lower triangular matrix takes $\sum_{k=1}^{n} k^2 \approx n^3/3$ operations (see Exercise 2.13). Inverting both $\mathbf{L}$ and $\mathbf{U}$ thus takes $2n^3/3 \approx G_n$ operations. According to Exercise 2.14, it takes approximately G_n arithmetic operations to multiply an upper and a lower triangular matrix. It thus takes approximately $G_n + G_n = 2G_n$ operations to compute $\mathbf{U}^{-1}\mathbf{L}^{-1}$.

CHAPTER 3

LDL* Factorization and Positive definite Matrices

Exercise 3.20: Positive definite characterizations

We check the equivalent statements of Theorem 3.18 for the matrix

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

1. Obviously $\mathbf{A}$ is symmetric. In addition $\mathbf{A}$ is positive definite, because

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

   for any nonzero vector $[x, y]^{\mathrm{T}} \in \mathbb{R}^2$.
2. The eigenvalues of $\mathbf{A}$ are the roots of the characteristic equation

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

   Hence the eigenvalues are $\lambda = 1$ and $\lambda = 3$, which are both positive.
3. The leading principal submatrices of $\mathbf{A}$ are $[2]$ and $\mathbf{A}$ itself, which both have positive determinants.
4. If we assume as in a Cholesky factorization that B is lower triangular we have that

   $$\mathbf{B}\mathbf{B}^{\mathrm{T}} = \begin{bmatrix} b_{11} & 0 \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{21} \\ 0 & b_{22} \end{bmatrix} = \begin{bmatrix} b_{11}^2 & b_{11}b_{21} \\ b_{21}b_{11} & b_{21}^2 + b_{22}^2 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

   Since $b_1 1^2 = 2$ we can choose $b_{11} = \sqrt{2}$. $b_{11}b_{21} = 1$ then gives that $b_{21} = 1/\sqrt{2}$, and $b_{21}^2 + b_{22}^2 = 2$ finally gives $b_{22} = \sqrt{2 - 1/2} = \sqrt{3/2}$ (we chose the positive square root). This means that we can choose

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

   This could also have been obtained by writing down an LDL-factorization (as in the proof for its existence), and then multiplying in the square root of the diagonal matrix.

CHAPTER 4

Orthonormal and Unitary Transformations

Exercise 4.4: The $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ inner product

Assume that $\mathbf{A} \in \mathbb{R}^{m\times n}$ has linearly independent columns. We show that

$$\langle \cdot, \cdot \rangle_{\mathbf{A}} : (x, y) \longmapsto \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{y}$$

satisfies the axioms of an inner product on a real vector space $\mathcal{V}$, as described in Definition 4.1. 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 $\langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}} = \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x} = \langle \mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{x} \rangle \geq 0$, 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 $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} = \mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{y} = (\mathbf{x}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{y})^{\mathrm{T}} = \mathbf{y}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x} = \langle \mathbf{y}, \mathbf{x} \rangle_{\mathbf{A}}$.

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

Exercise 4.5: 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 4.18: What does Algorithm **housegen** do when $\mathbf{x} = \mathbf{e}_1$?

If $\mathbf{x} = \mathbf{e}_1$, then the algorithm yields $\rho = 1$, and $a = -\|\mathbf{e}_1\|_2 = -1$. We then get $\mathbf{z} = \mathbf{e}_1$, and

$$\mathbf{u} = \frac{\mathbf{z} + \mathbf{e}_1}{\sqrt{1 + z_1}} = \frac{2\mathbf{e}_1}{\sqrt{2}} = \sqrt{2}\mathbf{e}_1$$

and

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

Exercise 4.19: 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 what we did in Example 4.15 to obtain a vector $\mathbf{v}$ and a matrix $\mathbf{H}$,

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

such that $\mathbf{H}\mathbf{x} = \mathbf{y}$. As explained in the text above Example 4.15, this matrix $\mathbf{H}$ 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 what we did in Example 4.15 to obtain a vector $\mathbf{v}$ and a Householder transformation $\mathbf{H}$,

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

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

Exercise 4.20: 2×2 Householder transformation

Let $\mathbf{H} = \mathbf{I} - \mathbf{u}\mathbf{u}^{\mathrm{T}} \in \mathbb{R}^{2,2}$ be any Householder transformation. Then $\mathbf{u} = [u_1\ u_2]^{\mathrm{T}} \in \mathbb{R}^2$ 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

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

Furthermore, we find

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

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

Exercise 4.28: QR decomposition

That $\mathbf{Q}$ is orthonormal, and therefore unitary, can be shown directly by verifying that $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{I}$. A direct computation shows that $\mathbf{QR} = \mathbf{A}$. 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 $\mathbf{A} = \mathbf{QR}$ 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 4.29: 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 $\mathbf{H}_1, \mathbf{H}_2$ that produce zeros in the columns $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ of $\mathbf{A}$. Applying Algorithm 4.17 to the first column of $\mathbf{A}$, we find first that $a = -3$, $z = (1/3, -2/3, 2/3)^{\mathrm{T}}$, and then

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

Next we need to map the bottom element $(\mathbf{H}_1\mathbf{A})_{3,2}$ of the second column to zero, without changing the first row of $\mathbf{H}_1\mathbf{A}$. For this, we apply Algorithm 4.17 to the vector $(0, 1)^{\mathrm{T}}$ to find $a = -1$, $\mathbf{z} = (0, 1)^{\mathrm{T}}$, and then

$$\mathbf{u}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \text{and} \qquad \mathbf{H}_2 := \mathbf{I} - \mathbf{u}_2\mathbf{u}_2^{\mathrm{T}} = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix},$$

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

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

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

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

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

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

Exercise 4.32: 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

$$\mathbf{v}_1 = \mathbf{a}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \mathbf{q}_1 = \frac{1}{2}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix},$$

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

$$\frac{\mathbf{a}_3^{\mathrm{T}}\mathbf{v}_1}{\mathbf{v}_1^{\mathrm{T}}\mathbf{v}_1} = \frac{3}{2}, \qquad \frac{\mathbf{a}_3^{\mathrm{T}}\mathbf{v}_2}{\mathbf{v}_2^{\mathrm{T}}\mathbf{v}_2} = \frac{5}{4},$$

$$\mathbf{v}_3 = \mathbf{a}_3 - \frac{\mathbf{a}_3^{\mathrm{T}}\mathbf{v}_1}{\mathbf{v}_1^{\mathrm{T}}\mathbf{v}_1}\mathbf{v}_1 - \frac{\mathbf{a}_3^{\mathrm{T}}\mathbf{v}_2}{\mathbf{v}_2^{\mathrm{T}}\mathbf{v}_2}\mathbf{v}_2 = \begin{bmatrix} -3 \\ 3 \\ -3 \\ 3 \end{bmatrix}, \qquad \mathbf{q}_3 = \frac{1}{2}\begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix}.$$

Since $(\mathbf{R}_1)_{11} = \|\mathbf{v}_1\| = 2$, $(\mathbf{R}_1)_{22} = \|\mathbf{v}_2\| = 4$, $(\mathbf{R}_1)_{33} = \|\mathbf{v}_3\| = 6$, and since also $(\mathbf{R}_1)_{ij} = (\mathbf{a}_j)^T\mathbf{q}_i = ||\mathbf{v}_i||(\mathbf{a}_j^{\mathrm{T}}\mathbf{v}_i)/(\mathbf{v}_i^{\mathrm{T}}\mathbf{v}_i)$ for $i > j$ we get that

$$(\mathbf{R}_1)_{12} = 2 \times 1 = 2, \qquad (\mathbf{R}_1)_{13} = 2 \times \frac{3}{2} = 3, \qquad (\mathbf{R}_1)_{23} = 4 \times \frac{5}{4} = 5,$$

so that

$$Q_1 = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \mathbf{q}_3 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & 1 \end{bmatrix} \qquad R_1 = \begin{bmatrix} 2 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{bmatrix}$$

and

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

Exercise 4.34: 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

$$\mathbf{Px} = r\begin{bmatrix} \cos\theta\cos\alpha + \sin\theta\sin\alpha \\ -\sin\theta\cos\alpha + \cos\theta\sin\alpha \end{bmatrix} = \begin{bmatrix} r\cos(\theta - \alpha) \\ -r\sin(\theta - \alpha) \end{bmatrix}.$$

Exercise 4.35: Solving upper Hessenberg system using rotations

To determine the number of arithmetic operations of Algorithm 4.36, 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 $\mathcal{O}(n^2)$ arithmetic operations. We conclude that the Algorithm uses

$$\mathcal{O}(n^2) + \sum_{k=1}^{n-1}\big(4 + 2 + 6(n+1-k)\big) = \mathcal{O}(n^2) + 6\sum_{k=1}^{n-1}(n+2-k)$$
$$= \mathcal{O}(n^2) + 3n^2 + 9n - 12 = \mathcal{O}(4n^2)$$

arithmetic operations.

CHAPTER 5

Eigenpairs and Similarity Transformations

Exercise 5.9: 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 5.10: 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 5.11: 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 5.12: 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 5.13: 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 5.17: 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

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

It follows that the eigenvectors span a space of dimension 2, and this means that $\mathbf{A}$ is defective.

Exercise 5.22: Jordan example

This exercise shows that it matters in which order we solve for the columns of S. One would here need to find the second column first before solving for the other two. The matrices given are

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

we are asked to find $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3]$ satisfying

$$[\mathbf{A}\mathbf{s}_1, \mathbf{A}\mathbf{s}_2, \mathbf{A}\mathbf{s}_3] = \mathbf{A}\mathbf{S} = \mathbf{S}\mathbf{J} = [\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3]\mathbf{J} = \left[\mathbf{s}_1, \mathbf{s}_1 + \mathbf{s}_2, \mathbf{s}_3\right].$$

The equations for the first and third columns say that $\mathbf{s}_1$ and $\mathbf{s}_3$ are eigenvectors for $\lambda = 1$, so that they can be found by row reducing $\mathbf{A} - \mathbf{I}$:

$$\mathbf{A} - \mathbf{I} = \begin{bmatrix} 2 & 0 & 1 \\ -4 & 0 & -2 \\ -4 & 0 & -2 \end{bmatrix} \sim \begin{bmatrix} 2 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

$(1, 0, -2)^{\mathrm{T}}$ and $(0, 1, 0)^{\mathrm{T}}$ thus span the set of eigenvectors for $\lambda = 1$.

$\mathbf{s}_2$ can be found by solving $\mathbf{A}\mathbf{s}_2 = \mathbf{s}_1 + \mathbf{s}_2$, so that $(\mathbf{A} - \mathbf{I})\mathbf{s}_2 = \mathbf{s}_1$. This means that $(\mathbf{A} - \mathbf{I})^2\mathbf{s}_2 = (\mathbf{A} - \mathbf{I})\mathbf{s}_1 = \mathbf{0}$, so that $\mathbf{s}_2 \in \ker(\mathbf{A} - \mathbf{I})^2$. A simple computation shows that $(\mathbf{A} - \mathbf{I})^2 = 0$ so that any $\mathbf{s}_2$ will do, but we must also choose $\mathbf{s}_2$ so that $(\mathbf{A} - \mathbf{I})\mathbf{s}_2 = \mathbf{s}_1$ is an eigenvector of $\mathbf{A}$. Since $\mathbf{A} - \mathbf{I}$ has rank one, we may choose any $\mathbf{s}_2$ so that $(\mathbf{A} - \mathbf{I})\mathbf{s}_2$ is nonzero. In particular we can choose $\mathbf{s}_2 = \mathbf{e}_1$, and then $\mathbf{s}_1 = (\mathbf{A} - \mathbf{I})\mathbf{s}_2 = (2, -4, -4)^{\mathrm{T}}$.

We can also choose $\mathbf{s}_3 = (0, 1, 0)^{\mathrm{T}}$, since it is an eigenvector not spanned by the $\mathbf{s}_1$ and $\mathbf{s}_2$ which we just defined. All this means that we can set

$$\mathbf{S} = \begin{bmatrix} 2 & 1 & 0 \\ -4 & 0 & 1 \\ -4 & 0 & 0 \end{bmatrix}.$$

Exercise 5.24: Properties of the Jordan form

Let $\mathbf{J} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}$ be the Jordan form of the matrix $\mathbf{A}$ as in Theorem 5.19. Items 1. – 3. are easily shown by induction, making use of the rules of block multiplication in 2. and 3. For Item 4., write $\mathbf{E}_m := \mathbf{J}_m(\lambda) - \lambda\mathbf{I}_m$, with $\mathbf{J}_m(\lambda)$ the Jordan block of order m. By the binomial theorem,

$$\mathbf{J}_m(\lambda)^r = (\mathbf{E}_m + \lambda\mathbf{I}_m)^r = \sum_{k=0}^{r} \binom{r}{k} \mathbf{E}_m^k (\lambda\mathbf{I}_m)^{r-k} = \sum_{k=0}^{r} \binom{r}{k} \lambda^{r-k} \mathbf{E}_m^k.$$

Since $\mathbf{E}_m^k = \mathbf{0}$ for any $k \geq m$, we obtain

$$\mathbf{J}_m(\lambda)^r = \sum_{k=0}^{\min\{r, m-1\}} \binom{r}{k} \lambda^{r-k} \mathbf{E}_m^k.$$

Exercise 5.25: Powers of a Jordan block

Let $\mathbf{S}$ be as in Exercise 5.22. $\mathbf{J}$ is block-diagonal so that we can write

$$(\star) \qquad \mathbf{J}^n = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}^n = \begin{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}^n & 0 \\ 0 & 1^n \end{bmatrix} = \begin{bmatrix} 1 & n & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

where we used property 4. in exercise 5.24 on the upper left block. It follows that

$$\mathbf{A}^{100} = (\mathbf{S}\mathbf{J}\mathbf{S}^{-1})^{100} = \mathbf{S}\mathbf{J}^{100}\mathbf{S}^{-1} = \begin{bmatrix} 2 & 1 & 1 \\ -4 & 0 & 0 \\ -4 & 0 & -2 \end{bmatrix} \begin{bmatrix} 1 & 100 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 \\ -4 & 0 & 0 \\ -4 & 0 & -2 \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} 2 & 1 & 1 \\ -4 & 0 & 0 \\ -4 & 0 & -2 \end{bmatrix} \begin{bmatrix} 1 & 100 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & -\frac{1}{4} & 0 \\ 1 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & -\frac{1}{2} \end{bmatrix} = \begin{bmatrix} 201 & 0 & 100 \\ -400 & 1 & -200 \\ -400 & 0 & -199 \end{bmatrix}.$$

Exercise 5.27: Big Jordan example

The matrix $\mathbf{A}$ has Jordan form $\mathbf{A} = \mathbf{S}\mathbf{J}\mathbf{S}^{-1}$, with

$$\mathbf{J} = \left[\begin{array}{cc|ccc|cc|c} 3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{array}\right], \quad \mathbf{S} = \frac{1}{9}\begin{bmatrix} -14 & 9 & -5 & 6 & 0 & -8 & 9 & 9 \\ -28 & 18 & -10 & 12 & 0 & -7 & 0 & 0 \\ -42 & 27 & -15 & 18 & 0 & -6 & 0 & -9 \\ -56 & 36 & -20 & 24 & 0 & -5 & 0 & 0 \\ -70 & 45 & -16 & 12 & 9 & -4 & 0 & 0 \\ -84 & 54 & -12 & 9 & 0 & -3 & 0 & 0 \\ -98 & 63 & -8 & 6 & 0 & -2 & 0 & 0 \\ -49 & 0 & -4 & 3 & 0 & -1 & 0 & 0 \end{bmatrix}.$$

Exercise 5.30: Schur decomposition example

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

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

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

Exercise 5.34: 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

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

which implies that $\mathbf{A}^{\mathrm{T}} = -\mathbf{A}$ and $\mathbf{B} = \mathbf{B}^{\mathrm{T}}$ (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

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

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

Exercise 5.35: 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 5.49: Eigenvalue perturbation for Hermitian matrices

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

Exercise 5.51: 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 5.54: Biorthogonal expansion

The matrix $\mathbf{A}$ has characteristic polynomial $\det(\mathbf{A} - \lambda\mathbf{I}) = (\lambda - 4)(\lambda - 1)$ and right eigenpairs $(\lambda_1, \mathbf{x}_1) = (4, [1, 1]^{\mathrm{T}})$ and $(\lambda_2, \mathbf{x}_2) = (1, [1, -2]^{\mathrm{T}})$. Since the right eigenvectors $\mathbf{x}_1, \mathbf{x}_2$ are linearly independent, there exists vectors $\mathbf{y}_1, \mathbf{y}_2$ satisfying $\langle \mathbf{y}_i, \mathbf{x}_j \rangle = \delta_{ij}$. The set $\{\mathbf{x}_1\mathbf{x}_2\}$ forms a basis of $\mathbb{C}^2$, and the set $\{\mathbf{y}_1, \mathbf{y}_2\}$ is called the *dual basis*.

How do we find such vectors $\mathbf{y}_1, \mathbf{y}_2$? Any vector $[x_1, x_2]^{\mathrm{T}}$ is orthogonal to the vector $[\alpha x_2, -\alpha x_1]^{\mathrm{T}}$ for any α. Choosing α appropriately, one finds $\mathbf{y}_1 = \frac{1}{3}[1, -1]^{\mathrm{T}}, \mathbf{y}_2 = \frac{1}{3}[2, 1]^{\mathrm{T}}$. By Theorem 5.53, $\mathbf{y}_1$ and $\mathbf{y}_2$ are left eigenvectors of $\mathbf{A}$. For any vector $\mathbf{v} = [v_1, v_2]^{\mathrm{T}} \in \mathbb{C}^2$, Equation (5.21) then gives us the biorthogonal expansions

$$\begin{aligned} \mathbf{v} &= \langle \mathbf{y}_1, \mathbf{v} \rangle \mathbf{x}_1 + \langle \mathbf{y}_2, \mathbf{v} \rangle \mathbf{x}_2 = \frac{1}{3}(v_1 - v_2)\mathbf{x}_1 + \frac{1}{3}(2v_1 + v_2)\mathbf{x}_2 \\ &= \langle \mathbf{x}_1, \mathbf{v} \rangle \mathbf{y}_1 + \langle \mathbf{x}_2, \mathbf{v} \rangle \mathbf{y}_2 = (v_1 + v_2)\mathbf{y}_1 + (v_1 - 2v_2)\mathbf{y}_2. \end{aligned}$$

Exercise 5.57: 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

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

CHAPTER 6

The Singular Value Decomposition

Exercise 6.7: SVD examples

(a) For $\mathbf{A} = [3,4]^{\mathrm{T}}$ we find a 1×1 matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A} = 25$, 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 $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ 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 part 3 of Theorem 6.5, one finds $\mathbf{u}_1 = \frac{1}{5}[3,4]^{\mathrm{T}}$. Extending $\mathbf{u}_1$ to an orthonormal basis for $\mathbb{R}^2$ gives $\mathbf{u}_2 = \frac{1}{5}[-4,3]^{\mathrm{T}}$. 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

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

The eigenvalues of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ are the zeros of $\det(\mathbf{A}^{\mathrm{T}}\mathbf{A} - \lambda\mathbf{I}) = (9-\lambda)^2 - 81$, 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 $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ 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 part 3 of Theorem 6.5 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 4.9). 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 6.8: More SVD examples

(a) We have $\mathbf{A} = \mathbf{e}_1$ and $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{e}_1^{\mathrm{T}}\mathbf{e}_1 = \begin{bmatrix}1\end{bmatrix}$. This gives the eigenpair $(\lambda_1, \mathbf{v}_1) = (1, 1)$ of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$. 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 $\mathbf{A} = \mathbf{e}_n^{\mathrm{T}}$, the matrix

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

has eigenpairs $(0, \mathbf{e}_j)$ for $j = 1, \ldots, n-1$ and $(1, \mathbf{e}_n)$. Then $\mathbf{\Sigma} = \mathbf{e}_1^{\mathrm{T}} \in \mathbb{R}^{1,n}$ and $\mathbf{V} = \begin{bmatrix}\mathbf{e}_n, \mathbf{e}_{n-1}, \ldots, \mathbf{e}_1\end{bmatrix} \in \mathbb{R}^{n,n}$. Using part 3 of Theorem 6.5 we get $\mathbf{u}_1 = 1$, yielding $\mathbf{U} = \begin{bmatrix}1\end{bmatrix}$. A SVD for $\mathbf{A}$ is therefore given by

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

(c) In this exercise

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

The eigenpairs of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ 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 part 3 of Theorem 6.5 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 6.16: Counting dimensions of fundamental subspaces

Let $\mathbf{A}$ have singular value decomposition $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$.

1. By parts 1. and 3. of Theorem 6.15, $\mathrm{span}(\mathbf{A})$ and $\mathrm{span}(\mathbf{A}^*)$ are vector spaces of the same dimension r, implying that $\mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\mathbf{A}^*)$.

2. This statement is known as the *rank-nullity theorem*, and it follows immediately from combining parts 1. and 4. in Theorem 6.15.

3. As $\mathrm{rank}(\mathbf{A}^*) = \mathrm{rank}(\mathbf{A})$ by 1., this follows by replacing $\mathbf{A}$ by $\mathbf{A}^*$ in 2.

Exercise 6.17: Rank and nullity relations

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

1. By part 5 of Theorem 6.4, $\operatorname{rank}(\mathbf{A})$ is the number of positive eigenvalues of

$$\mathbf{A}\mathbf{A}^* = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^*\mathbf{V}\boldsymbol{\Sigma}^*\mathbf{U}^* = \mathbf{U}\mathbf{D}\mathbf{U}^*,$$

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

$$\mathbf{A}\mathbf{A}^*(\mathbf{A}\mathbf{A}^*)^* = \mathbf{A}\mathbf{A}^*\mathbf{A}\mathbf{A}^* = \mathbf{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^*\boldsymbol{\Sigma}\boldsymbol{\Sigma}^*\mathbf{V}^* = \mathbf{U}\mathbf{D}^2\mathbf{U}^*,$$

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

$$\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{A}\mathbf{A}^*) = \operatorname{rank}(\mathbf{A}^*\mathbf{A}).$$

2. Let $r := \operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{A}^*) = \operatorname{rank}(\mathbf{A}\mathbf{A}^*) = \operatorname{rank}(\mathbf{A}^*\mathbf{A})$. Applying Theorem 6.4, parts 3 and 4, to the singular value decompositions

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^*,\ \mathbf{A}^* = \mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^*,\ \mathbf{A}\mathbf{A}^* = \mathbf{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^*\mathbf{U}^*,\ \mathbf{A}^*\mathbf{A} = \mathbf{V}\boldsymbol{\Sigma}^*\boldsymbol{\Sigma}\mathbf{V}^*,$$

one finds that $\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$ is a basis for both $\ker(\mathbf{A})$ and $\ker(\mathbf{A}^*\mathbf{A})$, while $\{\mathbf{u}_{r+1}, \ldots \mathbf{u}_m\}$ is a basis for both $\ker(\mathbf{A}^*)$ and $\ker(\mathbf{A}\mathbf{A}^*)$. In particular it follows that

$$\dim\ker(\mathbf{A}) = \dim\ker(\mathbf{A}^*\mathbf{A}), \qquad \dim\ker(\mathbf{A}^*) = \dim\ker(\mathbf{A}\mathbf{A}^*),$$

which is what needed to be shown.

Exercise 6.18: Orthonormal bases example

Given is the matrix

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

From Example 6.6 we know that $\mathbf{B} = \mathbf{A}^{\mathrm{T}}$ and hence $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}$ and $\mathbf{B} = \mathbf{V}\boldsymbol{\Sigma}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}$, 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 6.15 we know that $\mathbf{V}_1$ forms an orthonormal basis for $\operatorname{span}(\mathbf{A}^{\mathrm{T}}) = \operatorname{span}(\mathbf{B})$, $\mathbf{V}_2$ an orthonormal basis for $\ker(\mathbf{A})$ and $\mathbf{U}_2$ an orthonormal basis for $\ker(\mathbf{A}^{\mathrm{T}}) = \ker(\mathbf{B})$. Hence

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

Exercise 6.19: 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 $\mathbf{A} = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^*$ be a singular value factorization of $\mathbf{A}$. Taking the Hermitian transpose $\mathbf{A}^* = \mathbf{V}_1\mathbf{\Sigma}_1^*\mathbf{U}_1^*$ 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 6.20: Singular values and eigenpair of composite matrix

Given is a singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$. 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 $\mathbf{\Sigma}_1 = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ as in Equation (6.7), so that $\mathbf{A} = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^*$ forms a singular value factorization of $\mathbf{A}$.

By Theorem 6.15,

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

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

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

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

Exercise 6.26: Rank example

We are given the singular value decomposition

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

Write $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4]$ and $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3]$. Clearly $r = \operatorname{rank}(\mathbf{A}) = 2$.

(a) A direct application of Theorem 6.15 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 $\ker(\mathbf{A}^{\mathrm{T}})$,
$\{\mathbf{v}_1, \mathbf{v}_2\}$ is an orthonormal basis for $\operatorname{span}(\mathbf{A}^{\mathrm{T}})$,
$\{\mathbf{v}_3\}$ is an orthonormal basis for $\ker(\mathbf{A})$.

Since $\mathbf{U}$ is orthogonal, $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4\}$ is an orthonormal basis for $\mathbb{R}^4$. In particular $\mathbf{u}_3, \mathbf{u}_4$ are orthogonal to $\mathbf{u}_1, \mathbf{u}_2$, so that they span the orthogonal complement $\operatorname{span}(\mathbf{A})^{\perp}$ to $\operatorname{span}(\mathbf{A}) = \operatorname{span}\{\mathbf{u}_1, \mathbf{u}_2\}$.

(b) Applying Theorem 6.25 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 6.3.2, with $\mathbf{D}' := \operatorname{diag}(\sigma_1, 0, \ldots, 0) \in \mathbb{R}^{n,n}$, take

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

Exercise 6.27: 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

$$\begin{aligned} (\mathbf{Bx})_i &= \sum_{j=1}^{n} b_{ij} x_j = b_{in} + \sum_{j=1}^{n-1} 2^{n-1-j} b_{ij} \\ &= -1 + 2^{n-1-i} b_{ii} + \sum_{j=i+1}^{n-1} 2^{n-1-j} b_{ij} \\ &= -1 + 2^{n-1-i} - \sum_{j=i+1}^{n-1} 2^{n-1-j} \\ &= -1 + 2^{n-1-i} - 2^{n-2-i} \sum_{j'=0}^{n-2-i} \left(\frac{1}{2}\right)^{j'} \\ &= -1 + 2^{n-1-i} - 2^{n-2-i} \frac{1 - \left(\frac{1}{2}\right)^{n-1-i}}{1 - \frac{1}{2}} \\ &= -1 + 2^{n-1-i} - 2^{n-1-i} \left(1 - 2^{-(n-1-i)}\right) \\ &= 0. \end{aligned}$$

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 6.25 for $r = \text{rank}(\mathbf{B}) < n$, we obtain

$$\sigma_n \leq \sqrt{\sigma_{r+1}^2 + \cdots + \sigma_n^2} = \min_{\substack{\mathbf{C} \in \mathbb{R}^{n,n} \\ \text{rank}(\mathbf{C}) = r}} \|\mathbf{A} - \mathbf{C}\|_F \leq \|\mathbf{A} - \mathbf{B}\|_F = 2^{2-n}.$$

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

CHAPTER 7

Norms and Perturbation theory for linear systems

Exercise 7.7: 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 $\mathbf{v} = \text{vec}(\mathbf{A})$ 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

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

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 7.8: 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 $\mathbf{v} = \text{vec}(\mathbf{A})$ 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 $\mathbf{A} = \mathbf{B} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. 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 $\|\mathbf{AB}\|_M \leq \|\mathbf{A}\|_M \|\mathbf{B}\|_M$.

Exercise 7.9: Consistency of modified max norm?

Exercise 7.8 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.

(1) Positivity. Clearly $\|\mathbf{A}\| = \sqrt{mn}\|\mathbf{A}\|_M \geq 0$. Moreover,

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

(2) Homogeneity. $\|\alpha\mathbf{A}\| = \sqrt{mn}\|\alpha\mathbf{A}\|_M = |\alpha|\sqrt{mn}\|\mathbf{A}\|_M = |\alpha|\|\mathbf{A}\|$.

(3) Subadditivity. One has

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

(4) 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 7.11: 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 7.12: 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

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

(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

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

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 7.13).

Exercise 7.19: Spectral norm

Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ be a singular value decomposition of $\mathbf{A}$, and write $\sigma_1 := \|\mathbf{A}\|_2$ for the biggest singular value of $\mathbf{A}$. Since the orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$ leave the Euclidean norm invariant,

$$\begin{aligned}&\max_{\|\mathbf{x}\|_2 = 1 = \|\mathbf{y}\|_2} |\mathbf{y}^* \mathbf{A} \mathbf{x}| = \max_{\|\mathbf{x}\|_2 = 1 = \|\mathbf{y}\|_2} |\mathbf{y}^* \mathbf{U} \mathbf{\Sigma} \mathbf{V}^* \mathbf{x}| = \max_{\|\mathbf{x}\|_2 = 1 = \|\mathbf{y}\|_2} |\mathbf{y}^* \mathbf{\Sigma} \mathbf{x}| \\ \leq &\max_{\|\mathbf{x}\|_2 = 1 = \|\mathbf{y}\|_2} \sigma_1 |\mathbf{y}^* \mathbf{x}| \leq \max_{\|\mathbf{x}\|_2 = 1 = \|\mathbf{y}\|_2} \sigma_1 \|\mathbf{y}\|_2 \|\mathbf{x}\|_2 = \sigma_1.\end{aligned}$$

Moreover, this maximum is achieved for $\mathbf{x} = \mathbf{y} = \mathbf{e}_1$, and we conclude

$$\|\mathbf{A}\|_2 = \sigma_1 = \max_{\|\mathbf{x}\|_2 = 1 = \|\mathbf{y}\|_2} |\mathbf{y}^* \mathbf{A} \mathbf{x}|.$$

Exercise 7.20: 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 (7.17) and (**??**), we find

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

which is what needed to be shown.

Exercise 7.21: 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 7.15, 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

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

Using Theorem 7.17, we find $\|\mathbf{A}\|_2 = \sigma_1 = 3$ and $\|\mathbf{A}^{-1}\|_2 = \sigma_2^{-1} = 1$. Alternatively, since $\mathbf{A}$ is symmetric positive definite, we know from (7.18) that $\|\mathbf{A}\|_2 = \lambda_1$ and $\|\mathbf{A}^{-1}\|_2 = 1/\lambda_2$, where $\lambda_1 = 3$ is the biggest eigenvalue of $\mathbf{A}$ and $\lambda_2 = 1$ is the smallest.

Exercise 7.24: Unitary invariance of the spectral norm

Suppose $\mathbf{V}$ is a rectangular matrix satisfying $\mathbf{V}^*\mathbf{V} = \mathbf{I}$. Then

$$\begin{aligned} \|\mathbf{VA}\|_2^2 &= \max_{\|\mathbf{x}\|_2=1} \|\mathbf{VAx}\|_2^2 = \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^*\mathbf{A}^*\mathbf{V}^*\mathbf{VAx} \\ &= \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^*\mathbf{A}^*\mathbf{Ax} = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2^2 = \|\mathbf{A}\|_2^2. \end{aligned}$$

The result follows by taking square roots.

Exercise 7.25: $\|\mathbf{AU}\|_2$ rectangular $\mathbf{A}$

Let $\mathbf{U} = [u_1, u_2]^{\mathrm{T}}$ be any 2×1 matrix satisfying $1 = \mathbf{U}^{\mathrm{T}}\mathbf{U}$. Then $\mathbf{AU}$ is a 2×1-matrix, and clearly the operator 2-norm of a 2×1-matrix equals its euclidean norm (when viewed as a vector):

$$\left\| \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \begin{bmatrix} x \end{bmatrix} \right\|_2 = \left\| \begin{bmatrix} a_1 x \\ a_2 x \end{bmatrix} \right\|_2 = |x| \left\| \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \right\|_2.$$

In order for $\|\mathbf{AU}\|_2 < \|\mathbf{A}\|_2$ to hold, we need to find a vector $\mathbf{v}$ with $\|\mathbf{v}\|_2 = 1$ so that $\|\mathbf{AU}\|_2 < \|\mathbf{Av}|_2$. In other words, we need to pick a matrix $\mathbf{A}$ that scales more in the direction $\mathbf{v}$ than in the direction $\mathbf{U}$. For instance, if

$$\mathbf{A} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{U} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad \mathbf{v} = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

then

$$\|\mathbf{A}\|_2 = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{Ax}\|_2 \geq \|\mathbf{Av}\|_2 = 2 > 1 = \|\mathbf{AU}\|_2.$$

Exercise 7.26: p-norm of diagonal matrix

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

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

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

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

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 7.27: Spectral norm of a column vector

We write $\mathbf{A} \in \mathbb{C}^{m,1}$ for the matrix corresponding to the column vector $\mathbf{a} \in \mathbb{C}^m$. Write $\|\mathbf{A}\|_p$ for the operator p-norm of $\mathbf{A}$ and $\|\mathbf{a}\|_p$ for the vector p-norm of $\mathbf{a}$. In particular $\|\mathbf{A}\|_2$ is the spectral norm of $\mathbf{A}$ and $\|\mathbf{a}\|_2$ is the Euclidean norm of $\mathbf{a}$. Then

$$\|\mathbf{A}\|_p = \max_{x\neq 0} \frac{\|\mathbf{A}x\|_p}{|x|} = \max_{x\neq 0} \frac{|x|\|\mathbf{a}\|_p}{|x|} = \|\mathbf{a}\|_p,$$

proving (b). Note that (a) follows as the special case $p = 2$.

Exercise 7.28: 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 7.15, these relations yield

$$\|\mathbf{A}\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{i,j}|^2\right)^{\frac{1}{2}} = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} |b_{i,j}|^2\right)^{\frac{1}{2}} = \| \, |\mathbf{A}| \, \|_F,$$

$$\|\mathbf{A}\|_1 = \max_{1\leq j\leq n} \left(\sum_{i=1}^{m} |a_{i,j}|\right) = \max_{1\leq j\leq n} \left(\sum_{i=1}^{m} |b_{i,j}|\right) = \| \, |\mathbf{A}| \, \|_1,$$

$$\|\mathbf{A}\|_\infty = \max_{1\leq i\leq m} \left(\sum_{j=1}^{n} |a_{i,j}|\right) = \max_{1\leq i\leq m} \left(\sum_{j=1}^{n} |b_{i,j}|\right) = \| \, |\mathbf{A}| \, \|_\infty,$$

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 $\mathbf{A}\mathbf{x}$ and $|\mathbf{A}| \cdot |\mathbf{x}|$, where $\mathbf{x} = (x_1, \ldots, x_n)$ and $|\mathbf{x}| = (|x_1|, \ldots, |x_n|)$. We find

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

Now let $\mathbf{x}^*$ with $\|\mathbf{x}^*\|_2 = 1$ be a vector for which $\|\mathbf{A}\|_2 = \|\mathbf{A}\mathbf{x}^*\|_2$. 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 7.15, 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

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

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

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

To get different norms we have to choose a, b in such a way that the maximal eigenvalues of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ and $|\mathbf{A}|^{\mathrm{T}}|\mathbf{A}|$ are different. Clearly $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ 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 $|\mathbf{A}|^{\mathrm{T}}|\mathbf{A}|$ to zero yields eigenvalues $\mu_\pm := a^2 + b^2 \pm 2|ab|$. Hence $|\mathbf{A}|^{\mathrm{T}}|\mathbf{A}|$ 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 7.35: Sharpness of perturbation bounds

Suppose $\mathbf{Ax} = \mathbf{b}$ and $\mathbf{Ay} = \mathbf{b} + \mathbf{e}$. Let $K = K(\mathbf{A}) = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ be the condition number of $\mathbf{A}$. Let $\mathbf{y}_\mathbf{A}$ and $\mathbf{y}_{\mathbf{A}^{-1}}$ be unit vectors for which the maxima in the definition of the operator norms of $\mathbf{A}$ and $\mathbf{A}^{-1}$ are attained. That is, $\|\mathbf{y}_\mathbf{A}\| = 1 = \|\mathbf{y}_{\mathbf{A}^{-1}}\|$, $\|\mathbf{A}\| = \|\mathbf{A}\mathbf{y}_\mathbf{A}\|$, and $\|\mathbf{A}^{-1}\| = \|\mathbf{A}^{-1}\mathbf{y}_{\mathbf{A}^{-1}}\|$. If $\mathbf{b} = \mathbf{A}\mathbf{y}_\mathbf{A}$ and $\mathbf{e} = \mathbf{y}_{\mathbf{A}^{-1}}$, then

$$\frac{\|\mathbf{y}-\mathbf{x}\|}{\|\mathbf{x}\|} = \frac{\|\mathbf{A}^{-1}\mathbf{e}\|}{\|\mathbf{A}^{-1}\mathbf{b}\|} = \frac{\|\mathbf{A}^{-1}\mathbf{y}_{\mathbf{A}^{-1}}\|}{\|\mathbf{y}_\mathbf{A}\|} = \|\mathbf{A}^{-1}\| = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|\frac{\|\mathbf{y}_{\mathbf{A}^{-1}}\|}{\|\mathbf{A}\mathbf{y}_\mathbf{A}\|} = K\frac{\|\mathbf{e}\|}{\|\mathbf{b}\|},$$

showing that the upper bound is sharp. If $\mathbf{b} = \mathbf{y}_{\mathbf{A}^{-1}}$ and $\mathbf{e} = \mathbf{A}\mathbf{y}_\mathbf{A}$, then

$$\frac{\|\mathbf{y}-\mathbf{x}\|}{\|\mathbf{x}\|} = \frac{\|\mathbf{A}^{-1}\mathbf{e}\|}{\|\mathbf{A}^{-1}\mathbf{b}\|} = \frac{\|\mathbf{y}_\mathbf{A}\|}{\|\mathbf{A}^{-1}\mathbf{y}_{\mathbf{A}^{-1}}\|} = \frac{1}{\|\mathbf{A}^{-1}\|} = \frac{1}{\|\mathbf{A}\|\|\mathbf{A}^{-1}\|}\frac{\|\mathbf{A}\mathbf{y}_\mathbf{A}\|}{\|\mathbf{y}_{\mathbf{A}^{-1}}\|} = \frac{1}{K}\frac{\|\mathbf{e}\|}{\|\mathbf{b}\|},$$

showing that the lower bound is sharp.

Exercise 7.36: Condition number of 2nd derivative matrix

Recall that $\mathbf{T} = \mathrm{tridiag}(-1, 2, -1)$ and, by Exercise 1.26, $\mathbf{T}^{-1}$ 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 7.15 and 7.17, 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 $m \geq 3$. 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

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

b) Since the matrix $\mathbf{T}$ is symmetric, $\mathbf{T}^{\mathrm{T}}\mathbf{T} = \mathbf{T}^2$ and the eigenvalues of $\mathbf{T}^{\mathrm{T}}\mathbf{T}$ 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

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

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

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

(d) For $p = 2$, substitute $h = 1/(m+1)$ in c) and use that $4/\pi^2 < 1/2$. For $p = 1, \infty$ we need to show due to a) that

$$\frac{4}{\pi^2} h^{-2} - 2/3 < \frac{1}{2} h^{-2} \leq \frac{1}{2} h^{-2}.$$

when m is odd, and that

$$\frac{4}{\pi^2} h^{-2} - 2/3 < \frac{1}{2}(h^{-2} - 1) \leq \frac{1}{2} h^{-2}.$$

when m is even. The right hand sides in these equations are obvious. The left equation for m odd is also obvious since $4/\pi^2 < 1/2$. The left equation for m even is also obvious since $-2/3 < -1/2$.

Exercise 7.47: When is a complex norm an inner product norm?

As in the Exercise, we let

$$\langle \mathbf{x}, \mathbf{y} \rangle = s(\mathbf{x}, \mathbf{y}) + i s(\mathbf{x}, i\mathbf{y}), \qquad s(\mathbf{x}, \mathbf{y}) = \frac{\|\mathbf{x} + \mathbf{y}\|^2 - \|\mathbf{x} - \mathbf{y}\|^2}{4}.$$

We need to verify the three properties that define an inner product. Let $\mathbf{x}, \mathbf{y}, \mathbf{z}$ be arbitrary vectors in $\mathbb{C}^m$ and $a \in \mathbb{C}$ be an arbitrary scalar.

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

$$\begin{aligned} s(\mathbf{x}, i\mathbf{x}) &= \frac{\|\mathbf{x} + i\mathbf{x}\|^2 - \|\mathbf{x} - i\mathbf{x}\|^2}{4} = \frac{\|(1+i)\mathbf{x}\|^2 - \|(1-i)\mathbf{x}\|^2}{4} \\ &= \frac{(|1+i| - |1-i|)\|\mathbf{x}\|^2}{4} = 0, \end{aligned}$$

so that $\langle \mathbf{x}, \mathbf{x} \rangle = \|\mathbf{x}\|^2 \geq 0$, with equality holding precisely when $\mathbf{x} = \mathbf{0}$.

(2) Conjugate symmetry. Since $s(\mathbf{x}, \mathbf{y})$ is real, $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$, $s(a\mathbf{x}, a\mathbf{y}) = |a|^2 s(\mathbf{x}, \mathbf{y})$, and $s(\mathbf{x}, -\mathbf{y}) = -s(\mathbf{x}, \mathbf{y})$,

$$\overline{\langle \mathbf{y}, \mathbf{x} \rangle} = s(\mathbf{y}, \mathbf{x}) - i s(\mathbf{y}, i\mathbf{x}) = s(\mathbf{x}, \mathbf{y}) - i s(i\mathbf{x}, \mathbf{y}) = s(\mathbf{x}, \mathbf{y}) - i s(\mathbf{x}, -i\mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle.$$

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

$$\begin{aligned}
2s(\mathbf{x},\mathbf{z})+2s(\mathbf{y},\mathbf{z}) &= \frac{1}{2}\|\mathbf{x}+\mathbf{z}\|^2-\frac{1}{2}\|\mathbf{z}-\mathbf{x}\|^2+\frac{1}{2}\|\mathbf{y}+\mathbf{z}\|^2-\frac{1}{2}\|\mathbf{z}-\mathbf{y}\|^2\\
&= \frac{1}{2}\left\|\mathbf{z}+\frac{\mathbf{x}+\mathbf{y}}{2}+\frac{\mathbf{x}-\mathbf{y}}{2}\right\|^2-\frac{1}{2}\left\|\mathbf{z}-\frac{\mathbf{x}+\mathbf{y}}{2}-\frac{\mathbf{x}-\mathbf{y}}{2}\right\|^2+\\
&\quad \frac{1}{2}\left\|\mathbf{z}+\frac{\mathbf{x}+\mathbf{y}}{2}-\frac{\mathbf{x}-\mathbf{y}}{2}\right\|^2-\frac{1}{2}\left\|\mathbf{z}-\frac{\mathbf{x}+\mathbf{y}}{2}+\frac{\mathbf{x}-\mathbf{y}}{2}\right\|^2\\
&= \left\|\mathbf{z}+\frac{\mathbf{x}+\mathbf{y}}{2}\right\|^2+\left\|\frac{\mathbf{x}-\mathbf{y}}{2}\right\|^2-\left\|\mathbf{z}-\frac{\mathbf{x}+\mathbf{y}}{2}\right\|^2-\left\|\frac{\mathbf{x}-\mathbf{y}}{2}\right\|^2\\
&= \left\|\mathbf{z}+\frac{\mathbf{x}+\mathbf{y}}{2}\right\|^2-\left\|\mathbf{z}-\frac{\mathbf{x}+\mathbf{y}}{2}\right\|^2\\
&= 4s\left(\frac{\mathbf{x}+\mathbf{y}}{2},\mathbf{z}\right),
\end{aligned}$$

implying that $s(\mathbf{x}+\mathbf{y},\mathbf{z})=s(\mathbf{x},\mathbf{z})+s(\mathbf{y},\mathbf{z})$. It follows that

$$\begin{aligned}
\langle \mathbf{x}+\mathbf{y},\mathbf{z}\rangle &= s(\mathbf{x}+\mathbf{y},\mathbf{z})+is(\mathbf{x}+\mathbf{y},i\mathbf{z})\\
&= s(\mathbf{x},\mathbf{z})+s(\mathbf{y},\mathbf{z})+is(\mathbf{x},i\mathbf{z})+is(\mathbf{y},i\mathbf{z})\\
&= s(\mathbf{x},\mathbf{z})+is(\mathbf{x},i\mathbf{z})+s(\mathbf{y},\mathbf{z})+is(\mathbf{y},i\mathbf{z})\\
&= \langle \mathbf{x},\mathbf{z}\rangle+\langle \mathbf{y},\mathbf{z}\rangle.
\end{aligned}$$

That $\langle a\mathbf{x},\mathbf{y}\rangle = a\langle \mathbf{x},\mathbf{y}\rangle$ follows, *mutatis mutandis*, from the proof of Theorem 7.45.

Exercise 7.48: p-norm for $p=1$ and $p=\infty$

We need to verify the three properties that define a norm. Consider arbitrary vectors $\mathbf{x}=[x_1,\dots,x_n]^{\mathrm{T}}$ and $\mathbf{y}=[y_1,\dots,y_n]$ in $\mathbb{R}^n$ and a scalar $a\in\mathbb{R}$. First we verify that $\|\cdot\|_1$ is a norm.

(1) Positivity. Clearly $\|\mathbf{x}\|_1=|x_1|+\cdots+|x_n|\geq 0$, with equality holding precisely when $|x_1|=\cdots=|x_n|=0$, which happens if and only if $\mathbf{x}$ is the zero vector.

(2) Homogeneity. One has

$$\|a\mathbf{x}\|_1=|ax_1|+\cdots+|ax_n|=|a|(|x_1|+\cdots+|x_n|)=|a|\|\mathbf{x}\|_1.$$

(3) Subadditivity. Using the triangle inequality for the absolute value,

$$\|\mathbf{x}+\mathbf{y}\|_1=|x_1+y_1|+\cdots+|x_n+y_n|\leq |x_1|+|y_1|+\cdots+\cdots|x_n|+|y_n|=\|\mathbf{x}\|_1+\|\mathbf{y}\|_1.$$

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

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

(2) Homogeneity. One has

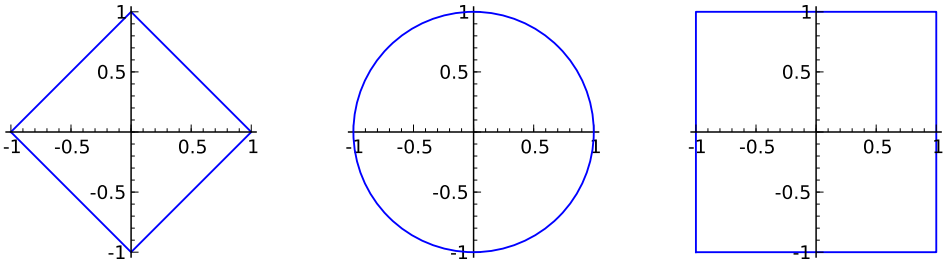
$$\|a\mathbf{x}\|_\infty=\max\{|a||x_1|,\dots,|a||x_n|\}=|a|\max\{|x_1|,\dots,|x_n|\}=|a|\|\mathbf{x}\|_\infty.$$

(3) Subadditivity. Using the triangle inequality for the absolute value,

$$\begin{aligned}
\|\mathbf{x}+\mathbf{y}\|_\infty &= \max\{|x_1+y_1|,\dots,|x_n+y_n|\}\leq \max\{|x_1|+|y_1|,\dots,|x_n|+|y_n|\}\\
&\leq \max\{|x_1|,\dots,|x_n|\}+\max\{|y_1|,\dots,|y_n|\}=\|\mathbf{x}\|_\infty+\|\mathbf{y}\|_\infty
\end{aligned}$$

Exercise 7.49: The p-norm unit sphere

In the plane, unit spheres for the 1-norm, 2-norm, and ∞-norm are



Exercise 7.50: Sharpness of p-norm inequality

Let $1 \leq p \leq \infty$. The vector $\mathbf{x}_l = [1, 0, \ldots, 0]^{\mathrm{T}} \in \mathbb{R}^n$ satisfies

$$\|\mathbf{x}_l\|_p = (|1|^p + |0|^p + \cdots + |0|^p)^{1/p} = 1 = \max\{|1|, |0|, \ldots, |0|\} = \|\mathbf{x}_l\|_\infty,$$

and the vector $\mathbf{x}_u = [1, 1, \ldots, 1]^{\mathrm{T}} \in \mathbb{R}^n$ satisfies

$$\|\mathbf{x}_u\|_p = (|1|^p + \cdots + |1|^p)^{1/p} = n^{1/p} = n^{1/p} \max\{|1|, \ldots, |1|\} = n^{1/p}\|\mathbf{x}_u\|_\infty.$$

Exercise 7.51: p-norm inequalities for arbitrary p

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

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

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

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

Since the function $x \longmapsto x^{1/p}$ is monotone, we obtain

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

from which the right inequality in the exercise follows.

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

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

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

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

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

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

CHAPTER 8

Least Squares

Exercise 8.10: Fitting a circle to points

We are given the (in general overdetermined) system

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

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

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

from which Equation (8.5) 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

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

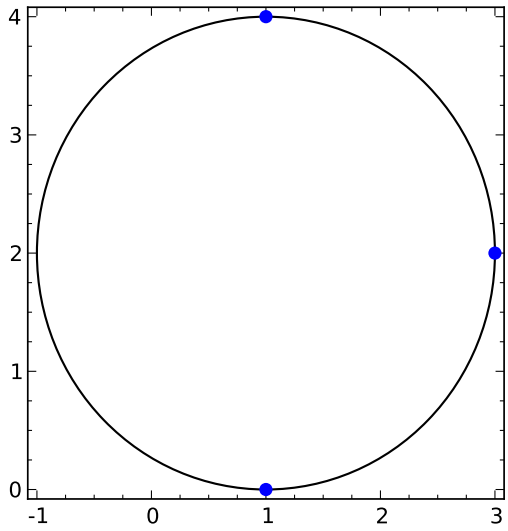
which clearly is invertible. We find

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

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 8.17: The generalized inverse

We let $\mathbf{A} = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^*$ and $\mathbf{B} = \mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^*$. Using that $\mathbf{U}_1^*\mathbf{U}_1 = \mathbf{V}_1^*\mathbf{V}_1 = \mathbf{I}$ and that $\mathbf{\Sigma}_1$ is diagonal we get

(1) $\mathbf{ABA} = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^*\mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^*\mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^* = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_1\mathbf{V}_1^* = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^* = \mathbf{A}$

(2) $\mathbf{BAB} = \mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^*\mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^*\mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^* = \mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^* = \mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^* = \mathbf{B}$

(3)

$$(\mathbf{BA})^* = \mathbf{A}^*\mathbf{B}^* = \mathbf{V}_1\mathbf{\Sigma}_1^*\mathbf{U}_1^*\mathbf{U}_1(\mathbf{\Sigma}_1^{-1})^*\mathbf{V}_1^* = \mathbf{V}_1\mathbf{\Sigma}_1^*(\mathbf{\Sigma}_1^{-1})^*\mathbf{V}_1^* = \mathbf{V}_1\mathbf{V}_1^*$$
$$\mathbf{BA} = \mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^*\mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^* = \mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_1\mathbf{V}_1^* = \mathbf{V}_1\mathbf{V}_1^*$$

(4)

$$(\mathbf{AB})^* = \mathbf{B}^*\mathbf{A}^* = \mathbf{U}_1(\mathbf{\Sigma}_1^{-1})^*\mathbf{V}_1^*\mathbf{V}_1\mathbf{\Sigma}_1^*\mathbf{U}_1^* = \mathbf{U}_1(\mathbf{\Sigma}_1^{-1})^*\mathbf{\Sigma}_1^*\mathbf{U}_1^* = \mathbf{U}_1\mathbf{U}_1^*$$
$$\mathbf{AB} = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^*\mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^* = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^* = \mathbf{U}_1\mathbf{U}_1^*$$

Exercise 8.18: 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), (2_B), (2_C), (3_C), (3_B), (1_B), (3_C), and (2_C).

Exercise 8.19: 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 8.17 and 8.18, we conclude that $\mathbf{B}$ must be the pseudoinverse of $\mathbf{A}$.

Exercise 8.20: 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 $n \leq m$. Then, by Exercise 6.17, $\mathbf{A}^*\mathbf{A}$ 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 8.17.

(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 8.21: The generalized inverse of a vector

This is a special case of Exercise 8.20. 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 8.17 to show that this vector must be the pseudoinverse of $\mathbf{u}^*$.

Exercise 8.22: The generalized inverse of an outer product

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

$$\mathbf{A} = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^* = \frac{\mathbf{u}}{\|\mathbf{u}\|_2}\left[\|\mathbf{u}\|_2\|\mathbf{v}\|_2\right]\frac{\mathbf{v}^*}{\|\mathbf{v}\|_2}$$

is a singular value factorization of $\mathbf{A}$. But then

$$\mathbf{A}^\dagger = \mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^* = \frac{\mathbf{v}}{\|\mathbf{v}\|_2}\left[\frac{1}{\|\mathbf{u}\|_2\|\mathbf{v}\|_2}\right]\frac{\mathbf{u}^*}{\|\mathbf{u}\|_2} = \frac{1}{\|\mathbf{u}\|_2^2\|\mathbf{v}\|_2^2}\mathbf{v}\mathbf{u}^* = \frac{\mathbf{A}^*}{\alpha}.$$

Exercise 8.23: 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 8.17, 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 8.24: Properties of the generalized inverse

Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$ be a singular value decomposition of $\mathbf{A}$ and $\mathbf{A} = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^*$ the corresponding singular value factorization. By definition of the pseudo inverse, $\mathbf{A}^\dagger := \mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^*$.

(a) One has $(\mathbf{A}^\dagger)^* = (\mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^*)^* = \mathbf{U}_1\mathbf{\Sigma}_1^{-*}\mathbf{V}_1^*$. On the other hand, the matrix $\mathbf{A}^*$ has singular value factorization $\mathbf{A}^* = \mathbf{V}_1\mathbf{\Sigma}_1^*\mathbf{U}_1^*$, so that its pseudo inverse is $(\mathbf{A}^*)^\dagger := \mathbf{U}_1\mathbf{\Sigma}_1^{-*}\mathbf{V}_1^*$ as well. We conclude that $(\mathbf{A}^\dagger)^* = (\mathbf{A}^*)^\dagger$.

(b) Since $\mathbf{A}^\dagger := \mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^*$ is a singular value factorization, it has pseudo inverse $(\mathbf{A}^\dagger)^\dagger = (\mathbf{U}_1^*)^*(\mathbf{\Sigma}_1^{-1})^{-1}\mathbf{V}_1^* = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^* = \mathbf{A}$.

(c) Let $\alpha \neq 0$. Since the matrix $\alpha\mathbf{A}$ has singular value factorization $\mathbf{U}_1(\alpha\mathbf{\Sigma}_1)\mathbf{V}_1^*$, it has pseudo inverse

$$(\alpha\mathbf{A})^\dagger = \mathbf{V}_1(\alpha\mathbf{\Sigma}_1)^{-1}\mathbf{U}_1^* = \alpha^{-1}\mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^* = \alpha^{-1}\mathbf{A}^\dagger.$$

Exercise 8.25: The generalized inverse of a product

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

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

where $\mathbf{\Sigma}_{\mathbf{A},1}$ and $\mathbf{\Sigma}_{\mathbf{B},1}$ are invertible, and $\mathbf{V}_\mathbf{A}$ and $\mathbf{U}_\mathbf{B}$ are unitary. This gives

$$\mathbf{A}^\dagger\mathbf{A} = \mathbf{V}_\mathbf{A}\mathbf{\Sigma}_{\mathbf{A},1}^{-1}\mathbf{U}_{\mathbf{A},1}^*\mathbf{U}_{\mathbf{A},1}\mathbf{\Sigma}_{\mathbf{A},1}\mathbf{V}_\mathbf{A}^* = \mathbf{V}_\mathbf{A}\mathbf{\Sigma}_{\mathbf{A},1}^{-1}\mathbf{\Sigma}_{\mathbf{A},1}\mathbf{V}_\mathbf{A}^* = \mathbf{V}_\mathbf{A}\mathbf{V}_\mathbf{A}^* = \mathbf{I}$$
$$\mathbf{B}\mathbf{B}^\dagger = \mathbf{U}_\mathbf{B}\mathbf{\Sigma}_{\mathbf{B},1}\mathbf{V}_{\mathbf{B},1}^*\mathbf{V}_{\mathbf{B},1}\mathbf{\Sigma}_{\mathbf{B},1}^{-1}\mathbf{U}_\mathbf{B}^* = \mathbf{U}_\mathbf{B}\mathbf{\Sigma}_{\mathbf{B},1}\mathbf{\Sigma}_{\mathbf{B},1}^{-1}\mathbf{U}_\mathbf{B}^* = \mathbf{U}_\mathbf{B}\mathbf{U}_\mathbf{B} = \mathbf{I}.$$

We know already from Exercise 8.17 that $(\mathbf{A}\mathbf{A}^\dagger)^* = \mathbf{A}\mathbf{A}^\dagger$ and $(\mathbf{B}^\dagger\mathbf{B})^* = \mathbf{B}^\dagger\mathbf{B}$. 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 8.17.

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

(b) Let $\mathbf{A} = \mathbf{u}^*$ and $\mathbf{B} = \mathbf{v}$, where $\mathbf{u}$ and $\mathbf{v}$ are column vectors. From exercises 8.21and 8.22 we have that $\mathbf{A}^\dagger = \mathbf{u}/\|\mathbf{u}\|_2^2$, and $\mathbf{B}^\dagger = \mathbf{v}^*/\|\mathbf{v}\|_2^2$. We have that

$$(\mathbf{AB})^\dagger = (\mathbf{u}^*\mathbf{v})^\dagger = 1/(\mathbf{u}^*\mathbf{v}) \qquad \mathbf{B}^\dagger\mathbf{A}^\dagger = \mathbf{v}^*\mathbf{u}/(\|\mathbf{v}\|_2^2\|\mathbf{u}\|_2^2).$$

If these are to be equal we must have that $(\mathbf{u}^*\mathbf{v})^2 = \|\mathbf{v}\|_2^2\|\mathbf{u}\|_2^2$. We must thus have equality in the triangle inequality, and this can happen only if $\mathbf{u}$ and $\mathbf{v}$ are parallel. It is thus enough to find $\mathbf{u}$ and $\mathbf{v}$ which are not parallel, in order to produce a counterexample.

Exercise 8.26: The generalized inverse of the conjugate transpose

Let $\mathbf{A}$ have singular value factorization $\mathbf{A} = \mathbf{U}_1\mathbf{\Sigma}_1\mathbf{V}_1^*$, so that $\mathbf{A}^* = \mathbf{V}_1\mathbf{\Sigma}_1^*\mathbf{U}_1^*$ and $\mathbf{A}^\dagger = \mathbf{V}_1\mathbf{\Sigma}_1^{-1}\mathbf{U}_1^*$. Then $\mathbf{A}^* = \mathbf{A}^\dagger$ if and only if $\mathbf{\Sigma}_1^* = \mathbf{\Sigma}_1^{-1}$, which happens precisely when all nonzero singular values of $\mathbf{A}$ are one.

Exercise 8.27: Linearly independent columns

By Exercise 8.20, if $\mathbf{A}$ has rank n, then $\mathbf{A}^\dagger = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$. Then $\mathbf{A}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{b} = \mathbf{AA}^\dagger\mathbf{b}$, which is the orthogonal projection of $\mathbf{b}$ into span($\mathbf{A}$) by Theorem 8.12.

Exercise 8.28: Analysis of the general linear system

In this exercise, we can write

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

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

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

which is what needed to be shown.

(b) By (a), the linear system $\mathbf{Ax} = \mathbf{b}$ 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 $\sigma_1, \ldots, \sigma_r \neq 0$, this system has a solution if and only if $c_{r+1} = \cdots = c_n = 0$. We conclude that $\mathbf{Ax} = \mathbf{b}$ has a solution if and only if $c_{r+1} = \cdots = c_n = 0$.

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

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

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

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

Exercise 8.29: Fredholm's Alternative

Suppose that the system $\mathbf{Ax} = \mathbf{b}$ has a solution, i.e., $\mathbf{b} \in \text{span}(\mathbf{A})$. Suppose in addition that $\mathbf{A}^*\mathbf{y} = \mathbf{0}$ has a solution, i.e., $\mathbf{y} \in \ker(\mathbf{A}^*)$. Since $(\text{span}(\mathbf{A}))^\perp = \ker(\mathbf{A}^*)$, one has $\langle \mathbf{y}, \mathbf{b} \rangle = \mathbf{y}^*\mathbf{b} = 0$. Thus if the system $\mathbf{Ax} = \mathbf{b}$ has a solution, then we can not find a solution to $\mathbf{A}^*\mathbf{y} = \mathbf{0}$, $\mathbf{y}^*\mathbf{b} \neq 0$. Conversely if $\mathbf{y} \in \ker(\mathbf{A}^*)$ and $\mathbf{y}^*\mathbf{b} \neq 0$, then $\mathbf{b} \notin (\ker(\mathbf{A}^*))^\perp = \text{span}(\mathbf{A})$, implying that the system $\mathbf{Ax} = \mathbf{b}$ does not have a solution.

Exercise 8.32: 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 8.20, the pseudoinverse of $\mathbf{A}$ is

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

Theorem 8.12 tells us that the orthogonal projection of $\mathbf{b}$ into $\text{span}(\mathbf{A})$ is

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

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

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

where we used that $\mathbf{b} = \mathbf{b}_1 + \mathbf{b}_2$.

(b) By Theorem 7.15, 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 $\mathbf{A}^{\mathrm{T}}\mathbf{A}$, which satisfies

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

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 8.35: Problem using normal equations

(a) Let $\mathbf{A}$, $\mathbf{b}$, and ε be as in the exercise. The normal equations $\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x} = \mathbf{A}^{\mathrm{T}}\mathbf{b}$ 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 $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ 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 $\mathbf{x} = [x_1, x_2]^{\mathrm{T}}$ with $x_1 + x_2 = 7/3$ is a solution.

(b) For $\varepsilon = 0$, we get the same solution as in (a). For $\varepsilon \neq 0$, however, the solution to the system

$$\begin{bmatrix} 3 & 3+\varepsilon \\ 3+\varepsilon & 3+2\varepsilon \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 7 \\ 7+2\varepsilon \end{bmatrix}$$

is

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

We can compare this to the solution of (a) by comparing the residuals,

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

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

CHAPTER 9

The Kronecker Product

Exercise 9.2: 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 9.5: Properties of Kronecker products

Let be given matrices $\mathbf{A}, \mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{p \times q}$, $\mathbf{B}, \mathbf{B}_1, \mathbf{B}_2 \in \mathbb{R}^{r \times s}$, and $\mathbf{C} \in \mathbb{R}^{t \times u}$. Then $(\lambda \mathbf{A}) \otimes (\mu \mathbf{B}) = \lambda\mu(\mathbf{A} \otimes \mathbf{B})$ by definition of the Kronecker product and since

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

The identity $(\mathbf{A}_1 + \mathbf{A}_2) \otimes \mathbf{B} = (\mathbf{A}_1 \otimes \mathbf{B}) + (\mathbf{A}_2 \otimes \mathbf{B})$ follows from

$$\begin{aligned} &\begin{bmatrix} (\mathbf{A}_1 + \mathbf{A}_2)b_{11} & (\mathbf{A}_1 + \mathbf{A}_2)b_{12} & \cdots & (\mathbf{A}_1 + \mathbf{A}_2)b_{1s} \\ (\mathbf{A}_1 + \mathbf{A}_2)b_{21} & (\mathbf{A}_1 + \mathbf{A}_2)b_{22} & \cdots & (\mathbf{A}_1 + \mathbf{A}_2)b_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ (\mathbf{A}_1 + \mathbf{A}_2)b_{r1} & (\mathbf{A}_1 + \mathbf{A}_2)b_{r2} & \cdots & (\mathbf{A}_1 + \mathbf{A}_2)b_{rs} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{A}_1b_{11} + \mathbf{A}_2b_{11} & \mathbf{A}_1b_{12} + \mathbf{A}_2b_{12} & \cdots & \mathbf{A}_1b_{1s} + \mathbf{A}_2b_{1s} \\ \mathbf{A}_1b_{21} + \mathbf{A}_2b_{21} & \mathbf{A}_1b_{22} + \mathbf{A}_2b_{22} & \cdots & \mathbf{A}_1b_{2s} + \mathbf{A}_2b_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_1b_{r1} + \mathbf{A}_2b_{r1} & \mathbf{A}_1b_{r2} + \mathbf{A}_2b_{r2} & \cdots & \mathbf{A}_1b_{rs} + \mathbf{A}_2b_{rs} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{A}_1b_{11} & \mathbf{A}_1b_{12} & \cdots & \mathbf{A}_1b_{1s} \\ \mathbf{A}_1b_{21} & \mathbf{A}_1b_{22} & \cdots & \mathbf{A}_1b_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_1b_{r1} & \mathbf{A}_1b_{r2} & \cdots & \mathbf{A}_1b_{rs} \end{bmatrix} + \begin{bmatrix} \mathbf{A}_2b_{11} & \mathbf{A}_2b_{12} & \cdots & \mathbf{A}_2b_{1s} \\ \mathbf{A}_2b_{21} & \mathbf{A}_2b_{22} & \cdots & \mathbf{A}_2b_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_2b_{r1} & \mathbf{A}_2b_{r2} & \cdots & \mathbf{A}_2b_{rs} \end{bmatrix}. \end{aligned}$$

A similar argument proves $\mathbf{A} \otimes (\mathbf{B}_1 + \mathbf{B}_2) = (\mathbf{A} \otimes \mathbf{B}_1) + (\mathbf{A} \otimes \mathbf{B}_2)$, and therefore the bilinearity of the Kronecker product. The associativity $(\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C})$

follows from

$$
\begin{aligned}
&= \begin{bmatrix} \mathbf{A}b_{11} & \cdots & \mathbf{A}b_{1s} \\ \vdots & & \vdots \\ \mathbf{A}b_{r1} & \cdots & \mathbf{A}b_{rs} \end{bmatrix} \otimes \mathbf{C} \\
&= \left[\begin{array}{ccc|c|ccc} \mathbf{A}b_{11}c_{11} & \cdots & \mathbf{A}b_{1s}c_{11} & & \mathbf{A}b_{11}c_{1u} & \cdots & \mathbf{A}b_{1s}c_{1u} \\ \vdots & & \vdots & \cdots & \vdots & & \vdots \\ \mathbf{A}b_{r1}c_{11} & \cdots & \mathbf{A}b_{rs}c_{11} & & \mathbf{A}b_{r1}c_{1u} & \cdots & \mathbf{A}b_{rs}c_{1u} \\ \hline & \vdots & & & & \vdots & \\ \hline \mathbf{A}b_{11}c_{t1} & \cdots & \mathbf{A}b_{1s}c_{t1} & & \mathbf{A}b_{11}c_{tu} & \cdots & \mathbf{A}b_{1s}c_{tu} \\ \vdots & & \vdots & \cdots & \vdots & & \vdots \\ \mathbf{A}b_{r1}c_{t1} & \cdots & \mathbf{A}b_{rs}c_{t1} & & \mathbf{A}b_{r1}c_{tu} & \cdots & \mathbf{A}b_{rs}c_{tu} \end{array}\right] \\
&= \mathbf{A} \otimes \left[\begin{array}{ccc|c|ccc} b_{11}c_{11} & \cdots & b_{1s}c_{11} & & b_{11}c_{1u} & \cdots & b_{1s}c_{1u} \\ \vdots & & \vdots & \cdots & \vdots & & \vdots \\ b_{r1}c_{11} & \cdots & b_{rs}c_{11} & & b_{r1}c_{1u} & \cdots & b_{rs}c_{1u} \\ \hline & \vdots & & & & \vdots & \\ \hline b_{11}c_{t1} & \cdots & b_{1s}c_{t1} & & b_{11}c_{tu} & \cdots & b_{1s}c_{tu} \\ \vdots & & \vdots & \cdots & \vdots & & \vdots \\ b_{r1}c_{t1} & \cdots & b_{rs}c_{t1} & & b_{r1}c_{tu} & \cdots & b_{rs}c_{tu} \end{array}\right] \\
&= \mathbf{A} \otimes \begin{bmatrix} \mathbf{B}c_{11} & \cdots & \mathbf{B}c_{1u} \\ \vdots & & \vdots \\ \mathbf{B}c_{t1} & \cdots & \mathbf{B}c_{tu} \end{bmatrix}.
\end{aligned}
$$

Exercise 9.9: 2nd derivative matrix is positive definite

Applying Lemma 1.31 to the case that $a = -1$ and $d = 2$, one finds that the eigenvalues λ_j of the matrix $\text{tridiag}(-1, 2, -1) \in \mathbb{R}^{m,m}$ 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 $j = 1, \ldots, m$. Moreover, as $|\cos(x)| < 1$ for any $x \in (0, \pi)$, it follows that $\lambda_j > 0$ for $j = 1, \ldots, m$. Since, in addition, $\text{tridiag}(-1, 2, -1)$ is symmetric, Lemma 3.16 implies that the matrix $\text{tridiag}(-1, 2, -1)$ is symmetric positive definite.

Exercise 9.10: 1D test matrix is positive definite?

The statement of this exercise is a generalization of the statement of Exercise 9.9. Consider a matrix $M = \text{tridiag}(a, d, a) \in \mathbb{R}^{m,m}$ for which $d > 0$ and $d \geq 2|a|$. By Lemma 1.31, the eigenvalues λ_j, with $j = 1, \ldots, m$, of the matrix M are

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

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

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

again because $|\cos(x)| < 1$ for any $x \in (0, \pi)$. Since, in addition, M is symmetric, Lemma 3.16 again implies that M is symmetric positive definite.

Exercise 9.11: Eigenvalues for 2D test matrix of order 4

One has

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

which means that $(\lambda, \mathbf{x})$ is an eigenpair of $\mathbf{A}$. For $j = k = 1$ and $m = 2$, Property 1. of Theorem 9.8 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 (9.15), 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 9.12: 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

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

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

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

$$20v_{21} - 4v_{11} - 4v_{22} - v_{12} = \frac{1}{18}(8f_{21} + f_{11} + f_{20} + f_{31} + f_{22}),$$
$$20v_{12} - 4v_{11} - 4v_{22} - v_{21} = \frac{1}{18}(8f_{12} + f_{02} + f_{11} + f_{22} + f_{13}),$$
$$20v_{22} - 4v_{12} - 4v_{21} - v_{11} = \frac{1}{18}(8f_{22} + f_{12} + f_{21} + f_{32} + f_{23}).$$

(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 9.13: 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 (9.4) – (9.5).) Similarly, writing out two matrix products, the (j, k)-th entry of $\mathbf{TVT} = \mathbf{T}(\mathbf{VT})$ is found to be

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

Together, these observations yield that the System (9.17) 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 properties 7 and 8 of Theorem 9.7 that this equation can be rewritten to one of the form $\mathbf{Ax} = \mathbf{b}$, where

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

Exercise 9.14: Biharmonic equation

(a) Writing $v = -\nabla u$, the second line in Equation (9.19) 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

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

(b) By property 8 of Theorem 9.7,

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

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 $\mathbf{T} = \mathbf{T}^{\mathrm{T}}$, this equation implies

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

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

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

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

which is a linear system of m^2 equations.

(c) The equations $h^2\mathbf{V} = (\mathbf{TU} + \mathbf{UT})$ and $\mathbf{TV} + \mathbf{VT} = h^2\mathbf{F}$ together yield the normal form

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

The vector form is given in (b). Using the distributive property of matrix multiplication and the mixed product rule of Lemma 9.6, 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 $\mathbf{x} := \text{vec}(\mathbf{U})$ and $\mathbf{b} := h^4\text{vec}(\mathbf{F})$, the linear system of (b) can be written as $\mathbf{Ax} = \mathbf{b}$.

(d) Since $\mathbf{T}$ and $\mathbf{I}$ are symmetric positive definite, property 6 of Theorem 9.7 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 1.31, we know the eigenpairs $(\lambda_i, \mathbf{s}_i)$, where $i = 1, \ldots, m$, of the matrix $\mathbf{T}$. By property 5 of Theorem 9.7, 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 10

Fast Direct Solution of a Large Linear System

Exercise 10.5: Fourier matrix

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

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

which is what needed to be shown.

Exercise 10.6: Sine transform as Fourier transform

According to Lemma 10.2, 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

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

For $m = 1$ this means that

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

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

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

while computing the Fourier transform gives

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

Multiplying the Fourier transform with $\frac{i}{2}$, one finds $\frac{i}{2}\mathbf{F}_4\mathbf{z} = \mathbf{z}$, so that $\frac{i}{2}(\mathbf{F}_4\mathbf{z})_2 = x_1 = (\mathbf{S}_1\mathbf{x})_1$, which is what we needed to show.

Exercise 10.7: Explicit solution of the discrete Poisson equation

For any integer $m \geq 1$, let $h = 1/(m+1)$. For $j = 1, \ldots, m$, let $\lambda_j = 4\sin^2\big(j\pi h/2\big)$, $\mathbf{D} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$, and $\mathbf{S} = (s_{jk})_{jk} = \big(\sin(jk\pi h)\big)_{jk}$. By Section 10.2, 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 $(\mathbf{DX}+\mathbf{XD})_{pr} = (\lambda_p + \lambda_r)x_{pr}$, so that

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

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 10.8: Improved version of Algorithm 10.1

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 (10.4), 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 (10.9). To find $\mathbf{X}$, we therefore need to solve the m tridiagonal linear systems of (10.9). Since the eigenvalues $\lambda_1, \ldots, \lambda_m$ are positive, each matrix $\mathbf{T}+\lambda_j\mathbf{I}$ is diagonally dominant. By Theorem 1.24, every such matrix is nonsingular and has a unique LU factorization. Algorithms 1.8 and 1.9 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 10.3, 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 10.9: Fast solution of 9 point scheme

Analogously to Section 10.2, we use the relations between the matrices $\mathbf{T}, \mathbf{S}, \mathbf{X}, \mathbf{D}$ to rewrite Equation (9.18).

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

Writing $\mathbf{D} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$, 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 9.13.(b) and $\mathbf{s}_i$ as in Section 10.2. 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 eigenvectors $\mathbf{s}_i \otimes \mathbf{s}_j$, 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 (9.17) always has a (unique) solution.

Exercise 10.10: Algorithm for fast solution of 9 point scheme

The following describes an algorithm for solving System (9.17).

Algorithm 1 A method for solving the discrete Poisson problem (9.17)

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 (9.17).

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: $\mathbf{V} \leftarrow \mathbf{SXS}$

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 10.11: Fast solution of biharmonic equation

From Exercise 9.14 we know that $\mathbf{T} \in \mathbb{R}^{m\times m}$ is the second derivative matrix. According to Lemma 1.31, the eigenpairs $(\lambda_j, \mathbf{s}_j)$, with $j = 1, \ldots, m$, of $\mathbf{T}$ are given by

$$\mathbf{s}_j = [\sin(j\pi h), \sin(2j\pi h), \ldots, \sin(mj\pi h)]^{\mathrm{T}},$$
$$\lambda_j = 2 - 2\cos(j\pi h) = 4\sin^2(j\pi h/2),$$

and satisfy $\mathbf{s}_j^{\mathrm{T}}\mathbf{s}_k = \delta_{j,k}/(2h)$ 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 10.12: Algorithm for fast solution of biharmonic equation

In order to derive an algorithm that computes $\mathbf{U}$ in Problem 9.14, we can adjust Algorithm 10.1 by replacing the computation of the matrix $\mathbf{X}$ by the formula from Exercise 10.11. This adjustment does not change the complexity of Algorithm 10.1, which therefore remains $\mathcal{O}(\delta n^{3/2})$. The new algorithm can be implemented in Matlab as in Listing 10.1.

```
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
```

Listing 10.1. A simple fast solution to the biharmonic equation

Exercise 10.13: Check algorithm for fast solution of biharmonic equation

The Matlab function from Listing 10.2 directly solves the standard form $\mathbf{Ax} = \mathbf{b}$ of Equation (9.21), making sure to return a matrix of the same dimension as the implementation from Listing 10.1.

```
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
```

Listing 10.2. A direct solution to the biharmonic equation

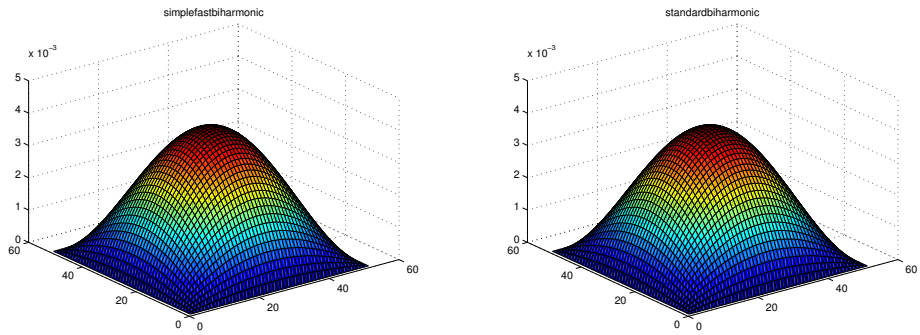
After specifying $m = 4$ by issuing the command `F = ones(4,4)`, the commands `simplefastbiharmonic(F)` and `standardbiharmonic(F)` both return the matrix

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

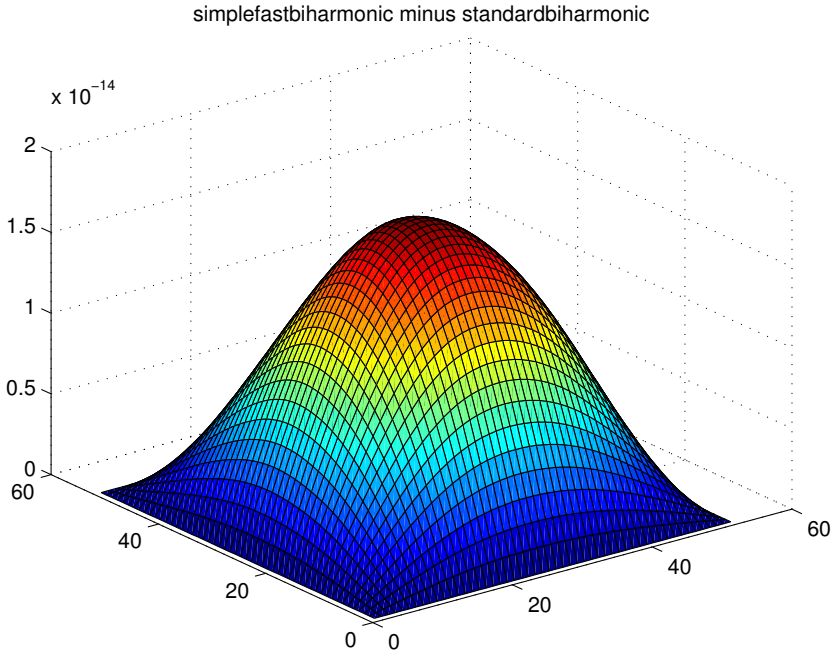
For large m, it is more insightful to *plot* the data returned by our Matlab functions. For $m = 50$, we solve and plot our system with the commands in Listing 10.3.

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

Listing 10.3. Solving the biharmonic equation and plotting the result



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.

CHAPTER 11

The Classical Iterative Methods

Exercise 11.12: Richardson and Jacobi

If $a_{11} = \cdots = a_{nn} = d \neq 0$ and $\alpha = 1/d$, Richardson's method (11.18) yields, for $i = 1, \ldots, n$,

$$\begin{aligned}
\mathbf{x}_{k+1}(i) &= \mathbf{x}_k(i) + \frac{1}{d}\left(b_i - \sum_{j=1}^{n} a_{ij}\mathbf{x}_k(j)\right) \\
&= \frac{1}{d}\left(d\mathbf{x}_k(i) - \sum_{j=1}^{n} a_{ij}\mathbf{x}_k(j) + b_i\right) \\
&= \frac{1}{a_{ii}}\left(a_{ii}\mathbf{x}_k(i) - \sum_{j=1}^{n} a_{ij}\mathbf{x}_k(j) + b_i\right) \\
&= \frac{1}{a_{ii}}\left(-\sum_{j=1}^{i-1} a_{ij}\mathbf{x}_k(j) - \sum_{j=i+1}^{n} a_{ij}\mathbf{x}_k(j) + b_i\right),
\end{aligned}$$

which is identical to Jacobi's method (11.2).

Exercise 11.13: Convergence of the R-method when eigenvalues have positive real part

We can write Richardson's method as $\mathbf{x}_{k+1} = \mathbf{G}\mathbf{x}_k + \mathbf{c}$, with $\mathbf{G} = \mathbf{I} - \alpha\mathbf{A}$, $\mathbf{c} = \alpha\mathbf{b}$. We know from Theorem 11.9 that the method converges if and only if $\rho(G) < 1$. The eigenvalues of $\mathbf{I} - \alpha\mathbf{A}$ are $1 - \alpha\lambda_j$, and we have that

$$|1 - \alpha\lambda_j|^2 = 1 + \alpha^2|\lambda_j|^2 - 2\alpha\Re(\lambda_j) = 1 + \alpha^2|\lambda_j|^2 - 2\alpha u_j.$$

This is less than 1 if and only if $\alpha^2|\lambda_j|^2 < 2\alpha u_j$. This can only hold if $\alpha > 0$, since $u_j > 0$. Dividing with α we get that $\alpha|\lambda_j|^2 < 2u_j$, so that $\alpha < 2u_j/|\lambda_j|^2$ (since $|\lambda_j| > 0$ since $u_j \neq 0$). We thus have that $\rho(G) < 1$ if and only if $\alpha < \min_j(2u_j/|\lambda_j|^2)$, and the result follows.

Exercise 11.16: Example: GS converges, J diverges

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 the 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 11.9 implies that the Jacobi method does not converge (even though $\mathbf{A}$ is symmetric positive definite).

Exercise 11.17: 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 11.9 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 (11.12), 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 11.18: 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 7.15, 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 7.14 the ∞-norm is consistent, Corollary 11.8 implies that the Jacobi method converges for any strictly diagonally dominant matrix $\mathbf{A}$.

Exercise 11.19: 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 (11.3), 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 $\epsilon_k := \mathbf{x}_k - \mathbf{x}$ are related by

$$\mathbf{D}\epsilon_{k+1} = \mathbf{A}_L\epsilon_{k+1} + \mathbf{A}_R\epsilon_k.$$

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

$$(\star) \qquad |\epsilon_{k+1}(j)| \leq r\|\epsilon_k\|_\infty, \qquad \text{for } j = 1, 2 \ldots, n.$$

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

$$|\epsilon_{k+1}(1)| = |a_{11}|^{-1}\,|-a_{12}\epsilon_k(2) - \cdots - a_{1n}\epsilon_k(n)| \leq r_1\|\epsilon_k\|_\infty \leq r\|\epsilon_k\|_\infty.$$

Assume that Equation $(\star)$ holds for $1, \ldots, j-1$. The relation between the residuals then bounds $|\epsilon_{k+1}(j)|$ as

$$|a_{jj}|^{-1}\,|-a_{j,1}\epsilon_{k+1}(1) - \cdots - a_{j,j-1}\epsilon_{k+1}(j-1) - a_{j,j+1}\epsilon_k(j+1) - \cdots - a_{j,n}\epsilon_k(n)|$$

$$\leq r_j \max\{r\|\epsilon_k\|_\infty, \|\epsilon_k\|_\infty\} = r_j\|\epsilon_k\|_\infty \leq r\|\epsilon_k\|_\infty.$$

Equation $(\star)$ then follows by induction, and it also follows that $\|\epsilon_{k+1}\|_\infty \leq r\|\epsilon_k\|_\infty$

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

$$\lim_{k\to\infty} \|\epsilon_k\|_\infty \leq \|\epsilon_0\|_\infty \lim_{k\to\infty} r^k = 0.$$

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

Exercise 11.23: Convergence example for fix point iteration

We show by induction that $\mathbf{x}_k(1) = \mathbf{x}_k(2) = 1 - a^k$ 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

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

When $|a| > 1$, however, $|\mathbf{x}_k(1)| = |\mathbf{x}_k(2)| = |1 - a^k|$ becomes arbitrary large with k and $\lim_{k\to\infty} \mathbf{x}_k(i)$ 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 (11.31) 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.

Exercise 11.24: Estimate in Lemma 11.22 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 $\mathbf{x}_k(1) = \mathbf{x}_k(2) = 1 - 2^{-k}$ for $k = 0$. Moreover, if this formula holds for some $k \geq 0$, one finds

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

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 $\|\epsilon_k\|_\infty = 2^{-k}$. Therefore, for given s, the error $\|\epsilon_k\|_\infty \leq 10^{-s}$ for the first time at $k \approx s \log(10)/\log(2)$. The bound $-s \log(10)/\log(\rho(\mathbf{G}))$ gives the same.

Exercise 11.25: 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 $\mathbf{A}$.

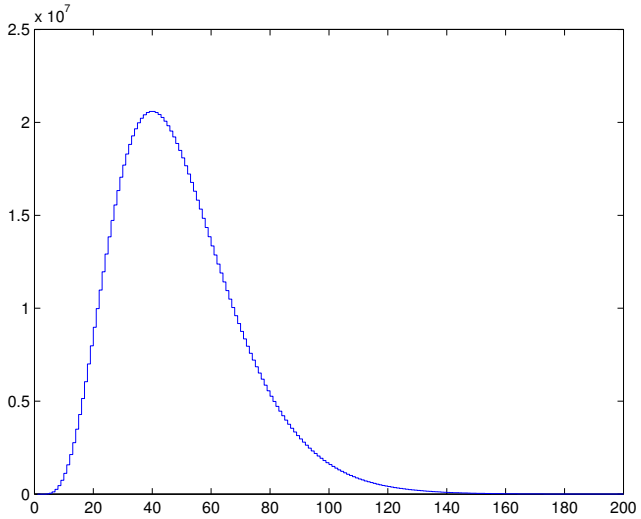
(a) The Matlab code

```
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) Let $\mathbf{E} = \mathbf{E}_1 := (\mathbf{A} - \lambda\mathbf{I})/a$ be the $n \times n$ matrix in the exercise, and write

$$\mathbf{E}_k := \begin{bmatrix} \mathbf{0} & \mathbf{I}_{n-k} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{n,n}.$$

Clearly $\mathbf{E}^k = \mathbf{E}_k$ for $k = 1$. Suppose that $\mathbf{E}^k = \mathbf{E}_k$ for some k satisfying $1 \leq k \leq n-1$. Using the rules of block multiplication,

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

Alternatively, since

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

one has

$$(\mathbf{E}^{k+1})_{ij} = (\mathbf{E}^k\mathbf{E})_{ij} = \sum_{\ell}(\mathbf{E}^k)_{i\ell}(\mathbf{E})_{\ell j} = (\mathbf{E}^k)_{i,i+k}(\mathbf{E})_{i+k,j} = 1\cdot(\mathbf{E})_{i+k,j}$$
$$= \begin{cases} 1 & \text{if } j = i+k+1, \\ 0 & \text{otherwise,} \end{cases}$$

By induction we conclude that $\mathbf{E}^k = \mathbf{E}_k$ for any k satisfying $1 \le k \le n$, with the convention that $\mathbf{E}^n = \mathbf{E}_n = \mathbf{0}_{n,n}$. We summarize that the matrix $\mathbf{E}$ is *nilpotent* of degree n.

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

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

Since $(\mathbf{E}^j)_{1,n} = 0$ for $1 \le j \le n-2$ 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 11.31: A special norm

We show that $\|\cdot\|_t$ inherits the three properties that define a norm from the operator norm $\|\cdot\|_1$. For arbitrary matrices $\mathbf{A}, \mathbf{B}$ and scalar a, we have

(1) Positivity. One has $\|\mathbf{B}\|_t = \|\mathbf{D}_t\mathbf{U}^*\mathbf{B}\mathbf{U}\mathbf{D}_t^{-1}\|_1 \ge 0$, with equality holding precisely when $\mathbf{D}_t\mathbf{U}^*\mathbf{B}\mathbf{U}\mathbf{D}_t^{-1}$ is the zero matrix, which happens if and only if $\mathbf{B}$ is the zero matrix.

(2) Homogeneity. For any scalar $a \in \mathbb{C}$,

$$\|a\mathbf{B}\|_t = \|a\mathbf{D}_t\mathbf{U}^*\mathbf{B}\mathbf{U}\mathbf{D}_t^{-1}\|_1 = |a|\cdot\|\mathbf{D}_t\mathbf{U}^*\mathbf{B}\mathbf{U}\mathbf{D}_t^{-1}\|_1 = |a|\cdot\|\mathbf{B}\|_t.$$

(3) Subadditivity. One has

$$\begin{aligned}\|\mathbf{A}+\mathbf{B}\|_t &= \|\mathbf{D}_t\mathbf{U}^*(\mathbf{A}+\mathbf{B})\mathbf{U}\mathbf{D}_t^{-1}\|_1 \\ &\le \|\mathbf{D}_t\mathbf{U}^*\mathbf{A}\mathbf{U}\mathbf{D}_t^{-1}\|_1 + \|\mathbf{D}_t\mathbf{U}^*\mathbf{B}\mathbf{U}\mathbf{D}_t^{-1}\|_1 \\ &= \|\mathbf{A}\|_t + \|\mathbf{B}\|_t.\end{aligned}$$

Since $\|\cdot\|_1$ is an operator norm, it is consistent. For any matrices $\mathbf{A}, \mathbf{B}$ for which the product $\mathbf{AB}$ is defined, therefore,

$$\begin{aligned}\|\mathbf{AB}\|_t &= \|\mathbf{D}_t\mathbf{U}^*\mathbf{A}\mathbf{B}\mathbf{U}\mathbf{D}_t^{-1}\|_1 \\ &= \|\mathbf{D}_t\mathbf{U}^*\mathbf{A}\mathbf{U}\mathbf{D}_t^{-1}\mathbf{D}_t\mathbf{U}^*\mathbf{B}\mathbf{U}\mathbf{D}_t^{-1}\|_1 \\ &\le \|\mathbf{D}_t\mathbf{U}^*\mathbf{A}\mathbf{U}\mathbf{D}_t^{-1}\|_1\|\mathbf{D}_t\mathbf{U}^*\mathbf{B}\mathbf{U}\mathbf{D}_t^{-1}\|_1 \\ &= \|\mathbf{A}\|_t\|\mathbf{B}\|_t,\end{aligned}$$

proving that $\|\cdot\|_t$ is consistent.

Exercise 11.33: When is $\mathbf{A}+\mathbf{E}$ nonsingular?

Suppose $\rho(\mathbf{A}^{-1}\mathbf{E}) = \rho\big(\mathbf{A}^{-1}(-\mathbf{E})\big) < 1$. By part 2 of Theorem 11.32, $\mathbf{I} + \mathbf{A}^{-1}\mathbf{E}$ is nonsingular and therefore so is the product $\mathbf{A}(\mathbf{I} + \mathbf{A}^{-1}\mathbf{E}) = \mathbf{A} + \mathbf{E}$.

CHAPTER 12

The Conjugate Gradient Method

Exercise 12.1: A-norm

Let $\mathbf{A} = \mathbf{L}\mathbf{L}^*$ be a Cholesky factorization of $\mathbf{A}$, i.e. $\mathbf{L}$ is lower triangular with positive diagonal elements. The $\mathbf{A}$-norm then takes the form $\|\mathbf{x}\|_\mathbf{A} = \sqrt{\mathbf{x}^T\mathbf{L}\mathbf{L} * \mathbf{x}} = \|\mathbf{L}^*\mathbf{x}\|$. Let us verify the three properties of a vector norm:

(1) **Positivity**: Clearly $\|\mathbf{x}\|_\mathbf{A} = \|\mathbf{L}^*\mathbf{x}\| \geq 0$. Since L^* in nonsingular, $\|\mathbf{x}\|_\mathbf{A} = \|\mathbf{L}^*\mathbf{x}\| = 0$ if and only if $\mathbf{L}^*\mathbf{x} = 0$ if and only if $\mathbf{x} = 0$.

(2) **Homogeneity**: $\|a\mathbf{x}\|_A = \|\mathbf{L}^*(a\mathbf{x})\| = \|a\mathbf{L}^*\mathbf{x}\| = |a|\|\mathbf{L}^*(\mathbf{x})\| = |a|\|\mathbf{x}\|_\mathbf{A}$.

(3) **Subadditivity**:

$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|_\mathbf{A} &= \|\mathbf{L}^*(\mathbf{x}+\mathbf{y})\| = \|\mathbf{L}^*\mathbf{x}+\mathbf{L}^*\mathbf{y}\| \\ &\leq \|\mathbf{L}^*\mathbf{x}\| + \|\mathbf{L}^*\mathbf{y}\| = \|\mathbf{x}\|_\mathbf{A} + \|\mathbf{y}\|_\mathbf{A}.\end{aligned}$$

Exercise 12.2: Paraboloid

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

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

which is what needed to be shown.

Exercise 12.5: 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

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

Given is a quadratic function

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

and an initial guess $\mathbf{x}_0 = [-1, -1/2]^\mathrm{T}$ 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 (12.7) twice yields

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Exercise 12.8: Conjugate gradient iteration, II

Using $\mathbf{x}_0 = \mathbf{0}$, one finds

$$\mathbf{x}_1 = \mathbf{x}_0 + \frac{(\mathbf{b} - \mathbf{A}\mathbf{x}_0)^T(\mathbf{b} - \mathbf{A}\mathbf{x}_0)}{(\mathbf{b} - \mathbf{A}\mathbf{x}_0)^T\mathbf{A}(\mathbf{b} - \mathbf{A}^2\mathbf{x}_0)}(\mathbf{b} - \mathbf{A}\mathbf{x}_0) = \frac{\mathbf{b}^T\mathbf{b}}{\mathbf{b}^T A\mathbf{b}}\mathbf{b}.$$

Exercise 12.9: Conjugate gradient iteration, III

By Exercise 12.8,

$$\mathbf{x}_1 = \frac{\mathbf{b}^T\mathbf{b}}{\mathbf{b}^T\mathbf{A}\mathbf{b}}\mathbf{b} = \frac{9}{18}\begin{bmatrix}0\\3\end{bmatrix} = \begin{bmatrix}0\\3/2\end{bmatrix}.$$

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 12.10: 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 (12.4),

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

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

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

Applying (12.16) repeatedly, one finds that the search direction $\mathbf{p}_k$ for the conjugate gradient method satisfies

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

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 $\mathbf{p}_k^{\mathrm{T}}\mathbf{r}_k = \mathbf{r}_k^{\mathrm{T}}\mathbf{r}_k$ and that the step length α is optimal for

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

Exercise 12.11: 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 $k \geq 0$,

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

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

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

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

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

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,

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

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

Exercise 12.17: Program code for testing steepest descent

Replacing the steps in (12.17) by those in (12.7), Algorithm 12.14 changes into the following algorithm for 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;
```

Listing 12.1. Testing the method of Steepest Descent

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 12.13 and 12.15, 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 $||\mathbf{r}_{K_{\mathrm{sd}}}||_2 \leq \mathrm{tol} \cdot ||\mathbf{r}_0||_2$. 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

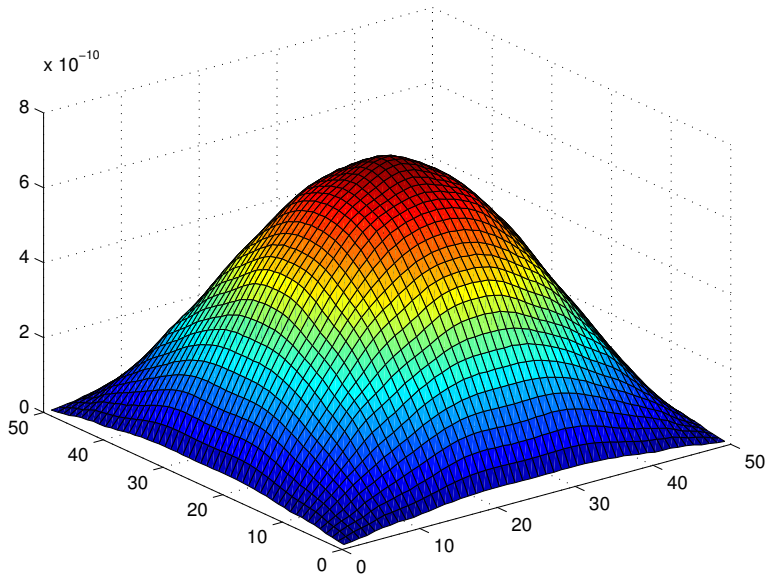


FIGURE 1. For a 50×50 Poisson matrix and a tolerance of 10^{-8}, the figure shows the difference of the outputs of cgtest and sdtest.

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 11.1, 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.

The spectral condition number of the $m \times m$ Poisson matrix is $\kappa = \big(1+\cos(\pi h)\big)/(1-\cos\big(\pi h)\big)$. Theorem 12.16 therefore states that

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

```
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;
```

Listing 12.2. Conjugate gradient method for least squares

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

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

is the Rayleigh quotient of $\mathbf{x}$, Lemma 5.44 implies the bound

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

with $\lambda_{\min} = 4\big(1-\cos(\pi h)\big)$ the smallest and $\lambda_{\max} = 4\big(1+\cos(\pi h)\big)$ 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}.

Exercise 12.18: Using cg to solve normal equations

We need to perform Algorithm 12.12 with $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ replacing $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}\mathbf{b}$ replacing $\mathbf{b}$. For the system $\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x} = \mathbf{A}^{\mathrm{T}}\mathbf{b}$, Equations (12.14), (12.15), and (12.16) become

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

with $\mathbf{p}_0 = \mathbf{r}_0 = \mathbf{b} - \mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{x}_0$. Hence we only need to change the computation of $\mathbf{r}_0$, α_k, and $\mathbf{r}_{k+1}$ in Algorithm 12.12, which yields the implementation in Listing 12.2.

Exercise 12.23: 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 $k = 0, 1, 2, \ldots$ Equations (12.14), (12.15), and (12.16),

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

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

$$\begin{aligned}
&\alpha_0 = \frac{1}{2}, && \mathbf{x}_1 = \begin{bmatrix}2\\0\\0\end{bmatrix}, && \mathbf{r}_1 = \begin{bmatrix}0\\2\\0\end{bmatrix}, && \beta_0 = \frac{1}{4}, && \mathbf{p}_1 = \begin{bmatrix}1\\2\\0\end{bmatrix},\\
&\alpha_1 = \frac{2}{3}, && \mathbf{x}_2 = \frac{1}{3}\begin{bmatrix}8\\4\\0\end{bmatrix}, && \mathbf{r}_2 = \frac{1}{3}\begin{bmatrix}0\\0\\4\end{bmatrix}, && \beta_1 = \frac{4}{9}, && \mathbf{p}_2 = \frac{1}{9}\begin{bmatrix}4\\8\\12\end{bmatrix},\\
&\alpha_2 = \frac{3}{4}, && \mathbf{x}_3 = \begin{bmatrix}3\\2\\1\end{bmatrix}, && \mathbf{r}_3 = \begin{bmatrix}0\\0\\0\end{bmatrix}, && \beta_2 = 0, && \mathbf{p}_3 = \begin{bmatrix}0\\0\\0\end{bmatrix}.
\end{aligned}$$

(c) By definition we have $\mathbb{W}_0 = \{\mathbf{0}\}$. From the solution of part (a) we know that $\mathbb{W}_k = \operatorname{span}(\mathbf{b}_0, \mathbf{A}\mathbf{b}_0, \ldots, \mathbf{A}^{k-1}\mathbf{b}_0)$, 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 $\mathbf{r}_0, \ldots, \mathbf{r}_{k-1}$. 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 $\operatorname{span}(\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}\mathbf{b}^{k-1}) = \operatorname{span}(\mathbf{r}_0, \ldots, \mathbf{r}_{k-1})$ for $k = 1, 2, 3$. We conclude that, for $k = 1, 2, 3$, the vectors $\mathbf{r}_0, \ldots, \mathbf{r}_{k-1}$ 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

$$\operatorname{span}(\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}\mathbf{b}^{k-1}) = \operatorname{span}(\mathbf{p}_0, \ldots, \mathbf{p}_{k-1}), \quad \text{for } k = 1, 2, 3.$$

We conclude that, for $k = 1, 2, 3$, the vectors $\mathbf{p}_0, \ldots, \mathbf{p}_{k-1}$ 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

$$\big\|\mathbf{r}_0\big\|_2 = 4, \qquad \big\|\mathbf{r}_1\big\|_2 = 2, \qquad \big\|\mathbf{r}_2\big\|_2 = 4/3, \qquad \big\|\mathbf{r}_3\big\|_2 = 0.$$

It follows that the sequence $(\|\mathbf{r}_k\|)_k$ is monotonically decreasing. Similarly, one finds

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

which is clearly monotonically decreasing.

Exercise 12.26: Another explicit formula for the Chebyshev polynomial

It is well known, and easily verified, that $\cosh(x+y) = \cosh(x)\cosh(y)+\sinh(x)\sinh(y)$. Write $P_n(t) = \cosh\big(n \cdot \operatorname{arccosh}(t)\big)$ for any integer $n \geq 0$. Writing $\phi = \operatorname{arccosh}(t)$, and using that cosh is even and sinh is odd, one finds

$$\begin{aligned}
&P_{n+1}(t) + P_{n-1}(t) \\
&= \cosh\big((n+1)\phi\big) + \cosh\big((n-1)\phi\big) \\
&= \cosh(n\phi)\cosh(\phi) + \sinh(n\phi)\sinh(\phi) + \cosh(n\phi)\cosh(\phi) - \sinh(n\phi)\sinh(\phi) \\
&= 2\cosh(\phi)\cosh(n\phi) \\
&= 2tP_n(t).
\end{aligned}$$

It follows that $P_n(t)$ satisfies the same recurrence relation as $T_n(t)$. Since in addition $P_0(t) = 1 = T_0(t)$, necessarily $P_n(t) = T_n(t)$ for any $n \geq 0$.

Exercise 12.28: 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,

$$\begin{aligned}
f(x) = f\big((1-\lambda)a + \lambda b\big) &\leq (1-\lambda)f(a) + \lambda f(b) \\
&\leq (1-\lambda)\max\{f(a), f(b)\} + \lambda \max\{f(a), f(b)\} = \max\{f(a), f(b)\}.
\end{aligned}$$

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

CHAPTER 13

Numerical Eigenvalue Problems

Exercise 13.5: Nonsingularity using Gerschgorin

We compute the Gerschgorin disks

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

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

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

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

Exercise 13.6: Gerschgorin, strictly diagonally dominant matrix

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

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

Since $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$ for all i, the origin is not an element of any of the R_i, and therefore neither of the union $\bigcup R_i$, nor of the intersection $(\bigcup R_i) \cap (\bigcup C_i)$ (which is smaller). Then, by Gerschgorin's Circle Theorem, $\mathbf{A}$ only has nonzero eigenvalues, implying that $\det(\mathbf{A}) = \det(\mathbf{A} - 0 \cdot \mathbf{I}) \neq 0$ and $\mathbf{A}$ is nonsingular.

Exercise 13.8: Continuity of eigenvalues

For a given matrix $\mathbf{A} = (a_{ij})_{ij} \in \mathbb{R}^{n \times n}$, write

$$\mathbf{A}(t) := \mathbf{D} + t(\mathbf{A} - \mathbf{D}), \qquad \mathbf{D} := \operatorname{diag}(a_{11}, \ldots, a_{nn}), \qquad t \in \mathbb{R},$$

for the affine combinations of $\mathbf{A}$ and its diagonal part $\mathbf{D}$. Let $t_1, t_2 \in [0, 1]$, with $t_1 < t_2$, so that $\mathbf{A}(t_1), \mathbf{A}(t_2)$ are *convex* combinations of $\mathbf{A}$ and $\mathbf{D}$. For any eigenvalue μ of $\mathbf{A}(t_2)$, we are asked to show that $\mathbf{A}(t_1)$ has an eigenvalue λ such that

$$(\star) \qquad |\lambda - \mu| \leq C(t_2 - t_1)^{1/n}, \qquad C \leq 2\big(\|\mathbf{D}\|_2 + \|\mathbf{A} - \mathbf{D}\|_2\big).$$

In particular, every eigenvalue of $\mathbf{A}(t)$ is a continuous function of t.

Applying Theorem 13.7 with $\mathbf{A}(t_1)$ and $\mathbf{E} = \mathbf{A}(t_2) - \mathbf{A}(t_1)$, one finds that $\mathbf{A}(t_1)$ has an eigenvalue λ such that

$$|\lambda - \mu| \leq \big(\|\mathbf{A}(t_1)\|_2 + \|\mathbf{A}(t_2)\|_2\big)^{1-1/n} \|\mathbf{A}(t_2) - \mathbf{A}(t_1)\|_2^{1/n}.$$

Applying the triangle inequality to the definition of $\mathbf{A}(t_1)$ and $\mathbf{A}(t_2)$, and using that the function $x \longmapsto x^{1-1/n}$ is monotone increasing,

$$|\lambda - \mu| \leq \Big(2\|\mathbf{D}\|_2 + (t_1 + t_2)\|\mathbf{A} - \mathbf{D}\|_2\Big)^{1-1/n} \|(\mathbf{A} - \mathbf{D})\|_2^{1/n} (t_2 - t_1)^{1/n}.$$

Finally, using that $t_1 + t_2 \leq 2$, that the function $x \longmapsto x^{1/n}$ is monotone increasing, and that $\|(\mathbf{A} - \mathbf{D})\|_2 \leq 2\|\mathbf{D}\|_2 + 2\|(\mathbf{A} - \mathbf{D})\|_2$, one obtains $(\star)$.

Exercise 13.12: ∞-norm of a diagonal matrix

Let $\mathbf{A} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ be a diagonal matrix. The spectral radius $\rho(\mathbf{A})$ is the absolute value of 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 $\|\mathbf{A}\mathbf{e}_i\|_\infty = \lambda_i = \rho(\mathbf{A})$.

Exercise 13.15: Number of arithmetic operations

An arithmetic operation is a floating point operation, so we need not bother with any integer operations, like the computation of $k + 1$ in the indices. As we are only interested in the overall complexity, we count only terms that can contribute to this.

For the first line involving C, the multiplication `v'*C` involves $(n-k)^2$ floating point multiplications and about $(n-k)^2$ floating point sums. Next, computing the outer product `v*(v'*C)` involves $(n-k)^2$ floating point multiplications, and subtracting `C - v*(v'*C)` needs $(n-k)^2$ 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

$$N := \sum_{k=1}^{n-2} \big(4(n-k)^2 + 4n(n-k)\big)$$

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 13.17: Number of arithmetic operations

The multiplication `v'*C` involves $(n-k)^2$ floating point multiplications and about $(n-k)^2$ floating point sums. Next, computing the outer product `v*(v'*C)` involves $(n-k)^2$ floating point multiplications, and subtracting `C - v*(v'*C)` needs $(n-k)^2$ 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 13.18: Tridiagonalize a symmetric matrix

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

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

Exercise 13.22: 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 13.21, 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 $\mathbf{A} - \alpha\mathbf{I} = \mathbf{L}\mathbf{D}\mathbf{L}^{\mathrm{T}}$,

$$d_1(\alpha) = 4 - 9/2 = -1/2,$$

$$d_2(\alpha) = 4 - 9/2 - 1^2/(-1/2) = +3/2,$$

$$d_3(\alpha) = 4 - 9/2 - 1^2/(+3/2) = -7/6,$$

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

As precisely two of these are negative, Corollary 13.21 implies that there are precisely two eigenvalues of $\mathbf{A}$ strictly smaller than $\alpha = 4.5$. As

$$\det(\mathbf{A} - 4.5\mathbf{I}) = \det(\mathbf{L}\mathbf{D}\mathbf{L}^{\mathrm{T}}) = d_1(\alpha)d_2(\alpha)d_3(\alpha)d_4(\alpha) \neq 0,$$

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 13.23: Overflow in LDL$^{\mathrm{T}}$ factorization

Since $\mathbf{A}_n$ is tridiagonal and strictly diagonally dominant, it has a unique LU factorization by Theorem 1.10. From Equations (1.16), one can determine the corresponding LDL* 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 (1.16) 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 show that this implies that $5 + \sqrt{24} < d_{n,k+1} \leq 10$. First observe that $\left(5 + \sqrt{24}\right)^2 = 25 + 10\sqrt{24} + 24 = 49 + 10\sqrt{24}$. From Equations (1.16) 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 $\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^{\mathrm{T}}$ with $\mathbf{L}$ triangular and with ones on the diagonal. As a consequence,

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

In Matlab an overflow is indicated by Matlab returning `Inf` as result. At my computer this happens at $n = 310$.

Exercise 13.24: 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

$$\hat{\mathbf{A}}^{\mathrm{T}} = {\mathbf{D}^{-\frac{1}{2}}}^{\mathrm{T}} \mathbf{U} \mathbf{A}^{\mathrm{T}} \mathbf{U}^{\mathrm{T}} {\mathbf{D}^{-\frac{1}{2}}}^{\mathrm{T}} = \mathbf{D}^{-\frac{1}{2}} \mathbf{U} \mathbf{A} \mathbf{U}^{\mathrm{T}} \mathbf{D}^{-\frac{1}{2}} = \hat{\mathbf{A}}$$

(b) Since $\hat{\mathbf{A}}$ is symmetric, it admits an orthogonal diagonalization $\hat{\mathbf{A}} = \hat{\mathbf{U}}^{\mathrm{T}} \hat{\mathbf{D}} \hat{\mathbf{U}}$. Let $\mathbf{E} := \mathbf{U}^{\mathrm{T}} \mathbf{D}^{-\frac{1}{2}} \hat{\mathbf{U}}^{\mathrm{T}}$. Then $\mathbf{E}$, as the product of three nonsingular matrices, is nonsingular. Its inverse is given explicitly by $\mathbf{F} := \hat{\mathbf{U}} \mathbf{D}^{\frac{1}{2}} \mathbf{U}$, since

$$\mathbf{F}\mathbf{E} = \hat{\mathbf{U}} \mathbf{D}^{\frac{1}{2}} \mathbf{U} \mathbf{U}^{\mathrm{T}} \mathbf{D}^{-\frac{1}{2}} \hat{\mathbf{U}}^{\mathrm{T}} = \hat{\mathbf{U}} \mathbf{D}^{\frac{1}{2}} \mathbf{D}^{-\frac{1}{2}} \hat{\mathbf{U}}^{\mathrm{T}} = \hat{\mathbf{U}} \hat{\mathbf{U}}^{\mathrm{T}} = \mathbf{I}$$

and similar $\mathbf{E}\mathbf{F} = \mathbf{I}$. Hence $\mathbf{E}^{-1} = \mathbf{F}$ and $\mathbf{E}$ is nonsingular. Moreover, from $\hat{\mathbf{A}} = \hat{\mathbf{U}}^{\mathrm{T}} \hat{\mathbf{D}} \hat{\mathbf{U}}$ follows that $\hat{\mathbf{U}} \hat{\mathbf{A}} \hat{\mathbf{U}}^{\mathrm{T}} = \hat{\mathbf{D}}$, which gives

$$\mathbf{E}^{\mathrm{T}} \mathbf{A} \mathbf{E} = \hat{\mathbf{U}} \mathbf{D}^{-\frac{1}{2}} \mathbf{U} \mathbf{A} \mathbf{U}^{\mathrm{T}} \mathbf{D}^{-\frac{1}{2}} \hat{\mathbf{U}}^{\mathrm{T}} = \hat{\mathbf{U}} \hat{\mathbf{A}} \hat{\mathbf{U}}^{\mathrm{T}} = \hat{\mathbf{D}}.$$

Similarly $\mathbf{B} = \mathbf{U}^{\mathrm{T}} \mathbf{D} \mathbf{U}$ implies $\mathbf{U} \mathbf{B} \mathbf{U}^{\mathrm{T}} = \mathbf{D}$, which yields

$$\mathbf{E}^{\mathrm{T}} \mathbf{B} \mathbf{E} = \hat{\mathbf{U}} \mathbf{D}^{-\frac{1}{2}} \mathbf{U} \mathbf{B} \mathbf{U}^{\mathrm{T}} \mathbf{D}^{-\frac{1}{2}} \hat{\mathbf{U}}^{\mathrm{T}} = \hat{\mathbf{U}} \mathbf{D}^{-\frac{1}{2}} \mathbf{D}^{\frac{1}{2}} \mathbf{D}^{\frac{1}{2}} \mathbf{D}^{-\frac{1}{2}} \hat{\mathbf{U}}^{\mathrm{T}} = \mathbf{I}.$$

We conclude that for a symmetric matrix $\mathbf{A}$ and symmetric positive definite matrix $\mathbf{B}$, the congruence transformation $\mathbf{X} \longmapsto \mathbf{E}^{\mathrm{T}} \mathbf{X} \mathbf{E}$ simultaneously diagonalizes the matrices $\mathbf{A}$ and $\mathbf{B}$, and even maps $\mathbf{B}$ to the identity matrix.

Exercise 13.25: 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 13.26: Determinant of upper Hessenberg matrix (TODO)

CHAPTER 14

The QR Algorithm

Exercise 14.4: 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

$$\lambda = \frac{\mathbf{u}^*\mathbf{A}\mathbf{u}}{\mathbf{u}^*\mathbf{u}}.$$