101MAT4/101MT4B: Solved Problems

Compilation date: March 3, 2020

Jan Chleboun

Contents

Introduction

This collection has been created to help the students attending the course `101MAT4` or `101MT4B`. All the included problems originate from the course lectures and problem solving practices. Although the solutions are presented in a rather concise form, I believe that they are accessible even for students with limited experience in the subject area.

On May 23, 2019, misprints were corrected on pages 9, 30, 35, and 37. On March 3, 2020, Problem 2.1 was extended.

1 Gaussian Elimination

Let us solve the following system of linear algebraic equations

$$\begin{aligned} 10x_1 - 5x_2 + 5x_3 + 10x_4 &= 5,\\ x_1 + x_2 - x_3 - x_4 &= 0,\\ x_1 + 2x_2 - x_3 + x_4 &= 5,\\ -x_1 + x_2 + x_3 - x_4 &= 4. \end{aligned}$$

First, we use the augmented matrix form:

$$\left(\begin{array}{rrrr|r} 10 & -5 & 5 & 10 & 5\\ 1 & 1 & -1 & -1 & 0\\ 1 & 2 & -1 & 1 & 5\\ -1 & 1 & 1 & -1 & 4 \end{array}\right) \overset{\spadesuit}{\sim} \left(\begin{array}{rrrr|r} 2 & -1 & 1 & 2 & 1\\ 1 & 1 & -1 & -1 & 0\\ 1 & 2 & -1 & 1 & 5\\ -1 & 1 & 1 & -1 & 4 \end{array}\right) \overset{\clubsuit}{\sim} \left(\begin{array}{rrrr|r} 1 & 1 & -1 & -1 & 0\\ 1 & 2 & -1 & 1 & 5\\ 2 & -1 & 1 & 2 & 1\\ -1 & 1 & 1 & -1 & 4 \end{array}\right)$$

$$\overset{\heartsuit}{\sim} \left(\begin{array}{rrrr|r} 1 & 1 & -1 & -1 & 0\\ 0 & 1 & 0 & 2 & 5\\ 0 & -3 & 3 & 4 & 1\\ 0 & 2 & 0 & -2 & 4 \end{array}\right) \overset{\square}{\sim} \left(\begin{array}{rrrr|r} 1 & 1 & -1 & -1 & 0\\ 0 & 1 & 0 & 2 & 5\\ 0 & 0 & 3 & 10 & 16\\ 0 & 0 & 0 & -6 & -6 \end{array}\right)$$

The first row was divided by 5 ($\spadesuit$), then the order of the 1st, 2nd, and 3rd row changed ($\clubsuit$). In the next step $\heartsuit$, the first row is subtracted from the second row and the first row multiplied by two is subtracted from the third row, and the first row is added to the fourth row. Finally ($\square$), the second row multiplied by three was added to the third row and the second row multiplied by two was subtracted from the fourth row.

The reduced row echelon form (after dividing the last row by -6) represents the equivalent system:

$$\begin{aligned} x_1 + x_2 - x_3 - x_4 &= 0,\\ x_2 + 2x_4 &= 5,\\ 3x_3 + 10x_4 &= 16,\\ x_4 &= 1. \end{aligned}$$

Rank of the matrix as well as of the augmented matrix is four, which is also the number of unknowns. The system has a unique solution.[1]

[1] Recall that if the rank of the matrix is less than the rank of the augmented matrix, then no solution exists. If both ranks are equal but less than the number of unknowns, then the system has an infinite number of solutions.

The last equation gives $x_4 = 1$, which is immediately substituted into the third equation, i.e., into $3x_3 + 10 = 16$, to get $x_3 = 2$. Using again $x_4 = 1$, we obtain $x_2 = 3$ from the second equation. Finally, $x_1 = 0$ from the first equation.

Let us check the correctness of the result

$$\begin{pmatrix} 10 & -5 & 5 & 10 \\ 1 & 1 & -1 & -1 \\ 1 & 2 & -1 & 1 \\ -1 & 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} 0 \\ 3 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 5 \\ 0 \\ 5 \\ 4 \end{pmatrix}.$$

Let us pay attention to a system $Ax = b$ with a singular matrix A. Take, for example, the following augmented matrix:

$$\left(\begin{array}{cccc|c} -1 & 2 & 1 & -1 & -1 \\ 0 & -1 & -1 & 3 & 3 \\ 1 & 3 & 2 & -4 & -4 \\ 5 & 2 & 0 & 4 & 4 \end{array}\right) \overset{\spadesuit}{\sim} \left(\begin{array}{cccc|c} -1 & 2 & 1 & -1 & -1 \\ 0 & -1 & -1 & 3 & 3 \\ 0 & 5 & 3 & -5 & -5 \\ 0 & 12 & 5 & -1 & -1 \end{array}\right) \overset{\clubsuit}{\sim} \left(\begin{array}{cccc|c} -1 & 2 & 1 & -1 & -1 \\ 0 & -1 & -1 & 3 & 3 \\ 0 & 0 & -2 & 10 & 10 \\ 0 & 0 & -7 & 35 & 35 \end{array}\right)$$

$$\overset{\heartsuit}{\sim} \left(\begin{array}{cccc|c} -1 & 2 & 1 & -1 & -1 \\ 0 & -1 & -1 & 3 & 3 \\ 0 & 0 & 1 & -5 & -5 \end{array}\right).$$

The first row was added to the third row and the first row multiplied by five was added to the last row ($\spadesuit$). The second row multiplied by five was added to the third row and the second row multiplied by twelve was added to the fourth row ($\clubsuit$). The third row was divided by -2 and the fourth row was divided by -7, these two rows are now identical and one of them can be deleted ($\heartsuit$).

We have four unknowns but only three equations:

$$\begin{aligned} -x_1 + 2x_2 + x_3 - x_4 &= -1, \\ - x_2 - x_3 + 3x_4 &= 3,. \\ x_3 - 5x_4 &= -5 \end{aligned}$$

One unknown will play the role of a parameter p, the others will be expressed via that p. Let us choose $x_4 \equiv p$. Then $x_3 = 5p - 5$, $x_2 = 2 - 2p$, and $x_1 = 0$. The solution x in the vector form

$$x = \begin{pmatrix} 0 \\ 2 - 2p \\ -5 + 5p \\ p \end{pmatrix} = u + pv, \text{ where } u = \begin{pmatrix} 0 \\ 2 \\ -5 \\ 0 \end{pmatrix} \text{ and } v = \begin{pmatrix} 0 \\ -2 \\ 5 \\ 1 \end{pmatrix},$$

where $p \in \mathbb{R}$ is arbitrary.

How the correctness can be checked? Notice that for $p \in \mathbb{R}$

$$b = Ax = A(u + pv) = Au + A(pv) = Au + pAv. \tag{1}$$

From (1), we obtain $b - Au = pAv$ that is valid for all $p \in \mathbb{R}$. This can be true only if Av is a zero vector[2] $o = (0, 0, 0, 0)^{\mathrm{T}}$.

To check the correctness of $x = u + pv$, we can proceed in two steps: $Av = o$ and $Au = b$.

[2] ... because $b - Au$ does not depend on $p \in \mathbb{R}$.

It may happen that the solution to a linear system $Ax = b$ depends on, for instance, two parameters, that is, $x = u+pv+qw$, where $p, q \in \mathbb{R}$. Then, $Au = b$ and $Av = o = Aw$; see the next example.

Let us consider another system with a singular matrix:

$$\left(\begin{array}{cccc|c} 0 & -8 & 8 & 2 & -22 \\ 2 & -4 & 6 & 3 & -13 \\ 6 & -4 & 10 & 7 & -17 \\ 4 & -4 & 8 & 5 & -15 \end{array}\right) \overset{\spadesuit}{\sim} \left(\begin{array}{cccc|c} 2 & -4 & 6 & 3 & -13 \\ 0 & -4 & 4 & 1 & -11 \\ 6 & -4 & 10 & 7 & -17 \\ 4 & -4 & 8 & 5 & -15 \end{array}\right) \overset{\clubsuit}{\sim} \left(\begin{array}{cccc|c} 2 & -4 & 6 & 3 & -13 \\ 0 & -4 & 4 & 1 & -11 \\ 0 & 8 & -8 & -2 & 22 \\ 0 & 4 & -4 & -1 & 11 \end{array}\right)$$

$$\overset{\heartsuit}{\sim} \left(\begin{array}{cccc|c} 2 & -4 & 6 & 3 & -13 \\ 0 & -4 & 4 & 1 & -11 \end{array}\right) \sim \left(\begin{array}{cccc|c} 1 & 0 & 1 & 1 & -1 \\ 0 & 4 & -4 & -1 & 11 \end{array}\right).$$

The second row was divided by two and interchanged with the first row ($\spadesuit$). The first row was multiplied by three and subtracted from the third row; the first row was multiplied by two and subtracted from the fourth row ($\clubsuit$). The third and fourth rows are both equivalent to the second row and can be deleted ($\heartsuit$). Finally, the second row is multiplied by -1, added to the first row and the resulting row is divided by two.

There are four unknowns but only two linearly independent equations in the system

$$\begin{array}{rcl} x_1 \quad + \quad x_3 + x_4 & = & -1, \\ 4x_2 - 4x_3 - x_4 & = & 11 \end{array}.$$

As a consequence, we will get infinitely many solutions that will depend on two parameters.

Let us choose $x_4 = p$ and $x_3 = q$. Then $x_2 = 11/4 + p/4 + q$ and $x_1 = -1 - p - q$. That is,

$$x = \begin{pmatrix} -1-p-q \\ \frac{11}{4} + \frac{p}{4} + q \\ q \\ p \end{pmatrix} = u + vp + wq, \text{ where } u = \begin{pmatrix} -1 \\ \frac{11}{4} \\ 0 \\ 0 \end{pmatrix}, \; v = \begin{pmatrix} -1 \\ 1/4 \\ 0 \\ 1 \end{pmatrix}, \; w = \begin{pmatrix} -1 \\ 1 \\ 1 \\ 0 \end{pmatrix},$$

and p, q are arbitrary real numbers.

The solution x is correct if $Av = o$, $Aw = o$, and $Ax = b$, where $b = (-22, -13, -17, -15)^{\mathrm{T}}$.

2 Eigenvalues and Eigenvectors

Problem 2.1: Let $A = \begin{pmatrix} 5 & -8 \\ -2 & 5 \end{pmatrix}$ be given. Find the eigenvalues and eigenvectors of the matrices B and B^{-1}, where $B = A^3$, i.e., $B = AAA$. Further, let μ_1 and μ_2 be the eigenvalues of B such that $\mu_1 < \mu_2$, and let $\widehat{v}_1$ and $\widehat{v}_2$ be the respective eigenvectors associated with μ_1 and μ_2 and such that their first component is equal to 4. Calculate $w = B^{-1}(3\widehat{v}_1 - 1458\widehat{v}_2)$.

Solution: If λ and u is an eigenpair of A, then $Au = \lambda u$ implies $Bu = AAAu = AA\lambda u = A\lambda^2 u = \lambda^3 u$ and $B^{-1}u = \lambda^{-3}u$.

The eigenvalues of A are the roots of the characteristic polynomial

$$\det(A - \lambda I) = \det \begin{pmatrix} 5-\lambda & -8 \\ -2 & 5-\lambda \end{pmatrix} = (5-\lambda)^2 - 16 = \lambda^2 - 10\lambda + 9,$$

that is, numbers 1 and 9. As a consequence, the values 1 and $9^3 = 3^6 = 729$ are the eigenvalues of B and the values 1 and $9^{-3} = 3^{-6} = 1/729$ are the eigenvalues of B^{-1}.

The matrices A, B, and B^{-1} have the same eigenvectors $v_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix} p$ and $v_2 = \begin{pmatrix} 2 \\ -1 \end{pmatrix} q$, where $p, q \in \mathbb{C} \setminus \{0\}$.

We observe that be choosing $p = 2 = q$, we get $\widehat{v}_1 = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$ and $\widehat{v}_2 = \begin{pmatrix} 4 \\ -2 \end{pmatrix}$. The vectors $\widehat{v}_1, \widehat{v}_2$ are the eigenvectors of A, B, and B^{-1}. Since $B\widehat{v}_1 = \widehat{v}_1$ and $B\widehat{v}_2 = 729\widehat{v}_2$, we obtain

$$w = B^{-1}(3\widehat{v}_1 - 1458\widehat{v}_2) = 3B^{-1}\widehat{v}_1 - 1458B^{-1}\widehat{v}_2 = 3\frac{1}{1}\widehat{v}_1 - 1458\frac{1}{729}\widehat{v}_2 = 3\widehat{v}_1 - 2\widehat{v}_2$$
$$= 3\begin{pmatrix} 4 \\ 2 \end{pmatrix} - 2\begin{pmatrix} 4 \\ -2 \end{pmatrix} = \begin{pmatrix} 4 \\ 10 \end{pmatrix}.$$

Problem 2.2: Let the matrices

$$A = \begin{pmatrix} -1 & 1 \\ 3 & 1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 4 & a \\ 1 & b \end{pmatrix}$$

be given. Find $a, b \in \mathbb{R}$ such that both matrices have the same eigenvalues.

Solution: The characteristic polynomial of A, i.e., $(-1-\lambda)(1-\lambda) - 3$, has two roots, namely, 2 and -2.

The characteristic equation of B is given by $(4-\lambda)(b-\lambda) - a = 0$, and the substitution $\lambda = 2$ and $\lambda = -2$ leads to the system

$$2b - a = 4$$
$$6b - a = -12$$

whose solution is $a = -12$, $b = -4$.

Problem 2.3: Let a matrix A and its eigenvector u be given, where

$$A = \begin{pmatrix} -2 & 2 & -7 \\ 7 & 3 & 7 \\ -2 & -2 & 3 \end{pmatrix}, \quad u = \begin{pmatrix} 23 \\ 0 \\ -23 \end{pmatrix} p, \; p \in \mathbb{C} \setminus \{0\}.$$

Find both all the eigenvalues and all the remaining eigenvectors of the matrix A.

Solution: Instead of using u, we can make calculations simpler by taking the eigenvector $\hat{u} = u/23 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$ that is associated with the same eigenvalue as u. Since $A\hat{u} = \begin{pmatrix} 5 \\ 0 \\ -5 \end{pmatrix}$, the associated eigenvalue $\lambda_1 = 5$.

We can proceed in two ways.

1) Let us take the equations $\lambda_1 + \lambda_2 + \lambda_3 = \operatorname{tr} A$ and $\lambda_1\lambda_2\lambda_3 = \det A$, and apply $\operatorname{tr} A = 4$, $\det A = -60$, and $\lambda_1 = 5$. We arrive at two equations

$$\lambda_2 + \lambda_3 = -1,$$
$$\lambda_2\lambda_3 = -12$$

whose solution is $\lambda_2 = 3$ and $\lambda_3 = -4$.

Let us determine the eigenvectors associated with $\lambda_2 = 3$. We have to solve the homogeneous system

$$\begin{pmatrix} -5 & 2 & -7 \\ 7 & 0 & 7 \\ -2 & -2 & 0 \end{pmatrix} \sim \begin{pmatrix} -5 & 2 & -7 \\ 1 & 0 & 1 \\ -1 & -1 & 0 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 1 \\ -5 & 2 & -7 \\ -1 & -1 & 0 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & -2 \\ 0 & -1 & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix},$$

that is,[3]

$$\begin{aligned} x_1 \quad + x_3 &= 0, \\ x_2 - x_3 &= 0. \end{aligned}$$

The solution is $x_3 = p$, $x_2 = p$, and $x_1 = -p$. The associated eigenvector:

$$v = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} p, \quad \text{where } p \in \mathbb{C} \setminus \{0\}.$$

For $\lambda_3 = -4$, we solve the homogeneous system

$$\begin{pmatrix} 2 & 2 & -7 \\ 7 & 7 & 7 \\ -2 & -2 & 7 \end{pmatrix} \sim \begin{pmatrix} 1 & 1 & 1 \\ 2 & 2 & -7 \\ -2 & -2 & 7 \end{pmatrix} \sim \begin{pmatrix} 1 & 1 & 1 \\ 2 & 2 & -7 \end{pmatrix} \sim \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & -9 \end{pmatrix},$$

that is,

$$\begin{aligned} x_1 + x_2 + x_3 &= 0, \\ -9x_3 &= 0. \end{aligned}$$

The solution is $x_3 = 0$, $x_2 = q$, and $x_1 = -q$. The associated eigenvector:

$$w = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} q, \quad \text{where } q \in \mathbb{C} \setminus \{0\}.$$

2) Since we know the eigenvalue -5, we can divide the characteristic polynomial of A

$$\begin{aligned} &(-2-\lambda)(3-\lambda)(3-\lambda) - 28 + 98 - 14(3-\lambda) - 14(3-\lambda) + 14(-2-\lambda) \\ &\quad = (-2-\lambda)(9 - 6\lambda + \lambda^2) + 70 - 84 + 28\lambda - 28 - 14\lambda \\ &\quad = -\lambda^3 + 4\lambda^2 + 17\lambda - 60 \end{aligned}$$

by $\lambda - 5$ to get

$$-\lambda^2 - \lambda + 12 = -(\lambda - 3)(\lambda + 4),$$

a polynomial whose roots $\lambda_2 = 3$ and $\lambda_3 = -4$ are the two unknown eigenvalues.

The eigenvectors are calculated in the same way as in the first approach.

Remark: The latter method is computationally more demanding than the former one because the cubic characteristic polynomial has to be calculated and than divided by the linear term.

[3]The matrix of the system must be singular! A nonsingular matrix (in the reduced row echelon form) would indicate an error in either the eigenvalue or matrix transformation.

Problem 2.4: Find the eigenpairs of

$$A = \begin{pmatrix} -2 & -1 & -3 \\ 4 & 3 & 3 \\ 1 & 1 & 2 \end{pmatrix}.$$

(Hint: The eigenvalues are small integers.)

Solution: Thanks to the hint, the roots of the cubic characteristic polynomial

$$\begin{aligned}(-2-\lambda)(3-\lambda)(2-\lambda) - 3 - 12 + 3(3-\lambda) + 4(2-\lambda) + 3(2+\lambda) \\ = (\lambda^2 - 4)(3-\lambda) + 8 - 4\lambda = -\lambda^3 + 3\lambda^2 - 4\end{aligned}$$

can be found by the trial and error method. The eigenvalues are $2, 2, -1$. By solving the respective homogenous systems, we obtain the eigenvectors

$$v = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} q, \quad w = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} q, \quad \text{where } p, q \in \mathbb{C} \setminus \{0\}.$$

Problem 2.5: Find the eigenpairs of $A = \begin{pmatrix} 1 & 0 & 5 \\ 0 & 2 & 0 \\ -2 & 0 & 3 \end{pmatrix}$. If A^{-1} exists, determine the eigenpairs of A^{-1}.

Solution: To get the characteristic polynomial, we calculate $\det(A - \lambda I) = (2-\lambda)\big((1-\lambda)(3-\lambda)+10\big) = (2-\lambda)(\lambda^2 - 4\lambda + 13)$. The roots are $\lambda_1 = 2$, $\lambda_2 = 2+3\mathrm{i}$, and $\lambda_3 = 2-3\mathrm{i}$.

(*Remark:* We can partly check the correctness of the eigenvalues. Their sum is equal to $\operatorname{tr} A = 6$, and their product is equal to $2 \times (2^2 + 3^2) = 26 = \det A$.)

All the eigenvalues are nonzero, so A is nonsingular. The eigenvalues of A^{-1} are the reciprocals of the eigenvalues of A. That is,

$$\mu_1 = 1/2, \ \mu_2 = \frac{1}{\lambda_2} = \frac{1}{2+3\mathrm{i}} \frac{2-3\mathrm{i}}{2-3\mathrm{i}} = \frac{2-3\mathrm{i}}{4+9} = \frac{2}{13} - \frac{3}{13}\mathrm{i}.$$

Since the eigenvalues are complex conjugate, we get $\mu_3 = \frac{2}{13} + \frac{3}{13}\mathrm{i}$ (we obtain the same result if we calculate $\mu_3 = 1/\lambda_3$).

Let us calculate the eigenvector associated with $\lambda_1 = 2$. We have to solve the system $(A - \lambda_1 I)v_1 = 0$, that is,

$$\begin{pmatrix} -1 & 0 & 5 \\ 0 & 0 & 0 \\ -2 & 0 & 1 \end{pmatrix} \sim \begin{pmatrix} -1 & 0 & 5 \\ -2 & 0 & 1 \end{pmatrix}.$$

we get $v_1 = c(0, 1, 0)^{\mathrm{T}}$, where $c \in \mathbb{C} \setminus \{0\}$.

The calculation of the eigenvector associated with $\lambda_2 = 2 + 3\mathrm{i}$ is more complex. We have to solve the system $(A - \lambda_2 I)v_2 = 0$, that is,

$$\begin{pmatrix} -1-3\mathrm{i} & 0 & 5 \\ 0 & -3\mathrm{i} & 0 \\ -2 & 0 & 1-3\mathrm{i} \end{pmatrix} \sim \begin{pmatrix} -1-3\mathrm{i} & 0 & 5 \\ 0 & -3\mathrm{i} & 0 \\ 0 & 0 & 0 \end{pmatrix} \sim \begin{pmatrix} -1-3\mathrm{i} & 0 & 5 \\ 0 & 1 & 0 \end{pmatrix}.$$

We multiplied the first row by $2/(-1-3\mathrm{i})$ and added the product to the third row. Since

$$\frac{5\cdot 2}{-1-3\mathrm{i}}+1-3\mathrm{i}=-\frac{10}{1+3\mathrm{i}}\frac{1-3\mathrm{i}}{1-3\mathrm{i}}+1-3\mathrm{i}=-\frac{10-30\mathrm{i}}{1+9}+1-3\mathrm{i}=0,$$

the third row vanishes (the matrix $A-\lambda_2 I$ is singular), which is also a proof that the eigenvalue is correct . Finally, we divided the second row by $-3\mathrm{i}$.

The equivalent system

$$\begin{aligned}-(1+3\mathrm{i})x_1 \qquad &+5x_3=0,\\ x_2 \qquad &=0\end{aligned}$$

leads to the conclusion $x_2=0$. We choose $x_3=q$, then

$$(-1-3\mathrm{i})x_1+5q=0\Longrightarrow x_3=\frac{5q}{1+3\mathrm{i}}\frac{1-3\mathrm{i}}{1-3\mathrm{i}}=\frac{5-15\mathrm{i}}{10}q=(1-3\mathrm{i})r,$$

where $q=2r$. The eigenvector $v_2=r(1-3\mathrm{i},0,2)^{\mathrm{T}}$, where $r\in\mathbb{C}\setminus\{0\}$.[4]

The theory says that the eigenvectors are complex conjugate. This implies $v_3=s(1+3\mathrm{i},0,2)^{\mathrm{T}}$, where $s\in\mathbb{C}\setminus\{0\}$. Also, we can calculate v_3 by solving the homogeneous system $(A-\lambda_3 I)v_3=0$.

Let us check that $Av_3=(2-3\mathrm{i})v_3$. We have (without loss of generality, we can consider $s=1$)

$$Av_3=\begin{pmatrix}1+3\mathrm{i}+10\\0\\-2-6\mathrm{i}+6\end{pmatrix}=\begin{pmatrix}11+3\mathrm{i}\\0\\4-6\mathrm{i}\end{pmatrix},\quad (2-3\mathrm{i})\begin{pmatrix}1+3\mathrm{i}\\0\\2\end{pmatrix}=\begin{pmatrix}2-3\mathrm{i}+6\mathrm{i}+9\\0\\4-6\mathrm{i}\end{pmatrix},$$

and the equality holds.

The eigenvectors of A^{-1} are again v_1, v_2, and v_3. The respective associated eigenvalues are μ_1, μ_2, and μ_3.

Remark: The matrix A has only real elements, but two of its eigenpairs are complex.

Problem 2.6: Let a matrix A dependent on $a,b\in\mathbb{R}$ and vectors u and v be given as

$$A=\begin{pmatrix}a&b&1\\5&5&1\\-5&-4&0\end{pmatrix},\quad u=\begin{pmatrix}-4\\5\\0\end{pmatrix},\quad v=\begin{pmatrix}-1\\0\\5\end{pmatrix}.$$

Find values $a,b\in\mathbb{R}$ such that u and v become the eigenvectors of A; determine the associated eigenvalues. Next, identify the last eigenpair.

Solution: Since

$$\lambda_1\begin{pmatrix}-4\\5\\0\end{pmatrix}=Au=\begin{pmatrix}-4a+5b\\5\\0\end{pmatrix}\quad\text{and}\quad\lambda_2\begin{pmatrix}-1\\0\\5\end{pmatrix}=Av=\begin{pmatrix}-a+5\\0\\5\end{pmatrix},$$

we can compare the second (third) components and we obtain $\lambda_1=1$ and $\lambda_2=1$. Then the parameters a and b have to solve the equations

$$-4a+5b=-4,\quad -a+5=-1.$$

[4]Let us emphasize that, for instance, $v_2=s(10,0,2+6\mathrm{i})^{\mathrm{T}}$, where $s\in\mathbb{C}\setminus\{0\}$, defines the same set of vectors. Indeed, since r is and arbitrary nonzero complex number, it can be expressed as $r=s(1+3\mathrm{i})$.

That is, $a = 6$ and $b = 4$.

Next, $\lambda_1 + \lambda_2 + \lambda_3 = a + 5 \Rightarrow 6 + 5 = 1 + 1 + \lambda_3 \Rightarrow \lambda_3 = 9$.

By solving the homogeneous system

$$\begin{pmatrix} -3 & 4 & 1 \\ 5 & -4 & 1 \\ -5 & -4 & -9 \end{pmatrix} \sim \begin{pmatrix} -3 & 4 & 1 \\ 5 & -4 & 1 \\ 0 & -8 & -8 \end{pmatrix} \sim \begin{pmatrix} -3 & 4 & 1 \\ 0 & 8 & 8 \\ 0 & -8 & -8 \end{pmatrix} \sim \begin{pmatrix} -3 & 4 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix},$$

we obtain $(-1, -1, 1)^{\mathrm{T}} q$, $q \in \mathbb{C} \setminus \{0\}$, the eigenvector associated with $\lambda_3 = 9$.

Problem 2.7: a) Find $a, b, c, d \in \mathbb{R}$ such that (λ_1, v_1), (λ_2, v_2), and (λ_3, v_3) are the eigenpairs of the matrix A, where

$$A = \begin{pmatrix} 0 & -2 & d \\ a & b & c \\ -2 & 1 & 0 \end{pmatrix}, \quad v_1 = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}, \quad v_2 = \begin{pmatrix} -2 \\ 1 \\ 1 \end{pmatrix}, \quad v_3 = \begin{pmatrix} -2 \\ 4 \\ 1 \end{pmatrix}.$$

b) Check the correctness of the identified parameters a, b, c, d.
c) Let a matrix B be defined as $B = 1600A^{-1}A^{-1} - 2A - 50I$, where A^{-1} is the inverse of A and I is the 3×3 identity matrix. Calculate all the eigenpairs of B.
d) Calculate the vector $u = Bv$, where $v = v_1/29 + v_2/2 + v_3$.
e) Find a vector w solving $Aw = f$, where $f = 8v_1 + 10v_2 + 24v_3$.

Solution: **a)** From $Av_i = \lambda_i v_i$, $i = 1, 2, 3$, we infer (by comparing the third components of Av_i and $\lambda_i v_i$) that $\lambda_1 = -4$, $\lambda_2 = 5$, and $\lambda_3 = 8$. By using these results and by comparing the other components of Av_i and $\lambda_i v_i$, we arrive at

$$\begin{gathered} a - 2b + c = 8, \\ -2a + b + c = 5, \\ -2a + 4b + c = 32, \\ 4 + d = -4, \ -2 + d = -10, \ -8 + d = -16. \end{gathered}$$

We get $d = -8$. To identify the other parameters, we solve the following nonhomogeneous system

$$\left(\begin{array}{ccc|c} 1 & -2 & 1 & 8 \\ -2 & 1 & 1 & 5 \\ -2 & 4 & 1 & 32 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & -2 & 1 & 8 \\ 0 & -3 & 3 & 21 \\ 0 & 3 & 0 & 27 \end{array}\right).$$

We easily infer $b = 9$, $c = 16$, $a = 10$ from the last augmented matrix.

b) It is sufficient to check the relationships between the matrix A and its eigenpairs, that is,

$$\begin{pmatrix} 0 & -2 & -8 \\ 10 & 9 & 16 \\ -2 & 1 & 0 \end{pmatrix} v_i = \lambda_i v_i, \qquad \text{where } i = 1, 2, 3.$$

c) It is known from $Av_i = \lambda_i v_i$, that $AAv_i = \lambda_i^2 v_i$, $A^{-1}v_i = \dfrac{1}{\lambda_i} v_i$, and $A^{-1}A^{-1}v_i = \dfrac{1}{\lambda_i^2} v_i$, where $i = 1, 2, 3$. Moreover, the identity matrix I has a triple eigenvalue 1 and each vector is its eigenvector. We observe that

$$Bv_i = \frac{1600}{\lambda_i^2} v_i - 2\lambda_i v_i - 50 v_i = \left(\frac{1600}{\lambda_i^2} - 2\lambda_i - 50 \right) v_i = \mu_i v_i, \quad i = 1, 2, 3, \qquad (2)$$

where μ_i stands for the eigenvalue of B. As a consequence of (2), if λ_i is an eigenvalue of A, then $\mu_i = 1600\lambda_i^{-2} - 2\lambda_i - 50$:

$$\lambda_1 = -4 \Rightarrow \mu_1 = 100 + 8 - 50 = 58,$$
$$\lambda_2 = 5 \Rightarrow \mu_2 = 64 - 10 - 50 = 4,$$
$$\lambda_3 = 8 \Rightarrow \mu_3 = 25 - 16 - 50 = -41.$$

This means that $(58, v_1)$, $(4, v_2)$, and $(-41, v_3)$ are the eigenpairs of B.

d) Since we know the eigenpairs of B, the vector Bv is easily calculated:

$$Bv = B(v_1/29 + v_2/2 + v_3) = Bv_1/29 + Bv_2/2 + Bv_3$$
$$= 2v_1 + 2v_2 - 41v_3 = \begin{pmatrix} 2 - 4 + 82 \\ -4 + 2 - 164 \\ 2 + 2 - 41 \end{pmatrix} = \begin{pmatrix} 80 \\ -166 \\ -37 \end{pmatrix}.$$

e) To find w, let us utilize the eigenvalues of A:

$$f = 8v_1 + 10v_2 + 24v_3 = -2Av_1 + 2Av_2 + 3Av_3 = A(-2v_1 + 2v_2 + 3v_3)$$
$$\Rightarrow w = -2v_1 + 2v_2 + 3v_3 = \begin{pmatrix} -2 - 4 - 6 \\ 4 + 2 + 12 \\ -2 + 2 + 3 \end{pmatrix} = \begin{pmatrix} -12 \\ 18 \\ 3 \end{pmatrix}.$$

Problem 2.8: The following matrices are given:

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad A = \begin{pmatrix} -1 & -1 \\ -1 & -1 \end{pmatrix}, \quad B = A^5 = AAAAA.$$

Calculate
a) the eigenvalues of B;
b) $\|B\|_2$;
c) the eigenvalues of $17I - B$.

Solution: a) By solving the characteristic equation, we can infer $\lambda_1 = 0$ and $\lambda_2 = -2$, the eigenvalues of A. However, a more elegant way is to realize that $\lambda_1 = 0$ because A is singular. Since $\lambda_1 + \lambda_2 = -1 - 1$, we obtain $\lambda_2 = -2.\lambda_2 = -2$. Then $\mu_1 = 0$ and $\mu_2 = (-2)^5 = -32$ are the eigenvalues of B.

b) Since A is symmetric, B is symmetric, too (why?), and $\|B\|_2 = 32$.

c) Let $C \equiv 17I - B$. If θ, v is an eigenpair of C, then $\theta v = Cv = (17I - B)v = 17v - Bv$ implies $Bv = (17 - \theta)v$. In other words, v is also an eigenvector of B that is associated with the eigenvalue $17 - \theta$. If, on the other hand, μ and w is an eigenpair of B, then $Cw = (17I - B)w = 17w - Bw = (17 - \mu)w$, so that w is the eigenvector of C associated with the eigenvalue $\theta = 17 - \mu$ of C.

We can summarize: 1) the matrices C and B have an identical set of eigenvectors; 2) the relationship between the respective eigenvalues is $\theta = 17 - \mu$. Since $\mu_1 = 0$ and $\mu_2 = -32$, the eigenvalues of the matrix $17I - B$ are $\theta_1 = 17$ and $\theta_2 = 49$.

3 Gershgorin Theorem

Problem 3.1: Use the Gershgorin theorem to find an upper and lower estimate of the condition number of the matrix C, where $C = \begin{pmatrix} 545 & -2 & 7 & 31 \\ -2 & -10 & 1 & -2 \\ 7 & 1 & -310 & 27 \\ 31 & -2 & 27 & -540 \end{pmatrix}$.

To calculate the condition number $\kappa(C)$, use the spectral norm $\|\cdot\|_2$. (Remark: If a nonsingular matrix is real and symmetric, its inverse is also real and symmetric.)

Solution: $\kappa(C) = \|C\|_2\|C^{-1}\|_2$.

The matrix C is real and symmetric which means that $\|C\|_2 = |\tilde{\lambda}|$, where $\tilde{\lambda}$ is the eigenvalue of C that is the most distant from the origin (all the eigenvalues of C are real).

Gershgorin disc reduce to intervals because all the eigenvalues are real. In particular,

$$\begin{aligned} K_1 &= \{z \in \mathbb{R} : \ |545 - z| \leq 40\} = [505, 585], \\ K_2 &= \{z \in \mathbb{R} : \ |-10 - z| \leq 5\} = [-15, -5], \\ K_3 &= \{z \in \mathbb{R} : \ |-310 - z| \leq 35\} = [-345, -275], \\ K_4 &= \{z \in \mathbb{R} : \ |-540 - z| \leq 60\} = [-600, -480]. \end{aligned}$$

The intervals of the absolute values of the eigenvalues:

$$\widetilde{K}_2 = [5, 15], \quad \widetilde{K}_3 = [275, 345], \quad \widetilde{K}_4 = [480, 600].$$

Since the intervals $K_1, \ldots, K_4$ are mutually disjoint, each contains exactly one eigenvalue. We observe that $K_1 \cap \widetilde{K}_4 \neq \emptyset$, even $K_1 \subset \widetilde{K}_4$. It is sure that K_1 contains one eigenvalue and K_4 also contains one eigenvalue. As a consequence, $|\tilde{\lambda}| \in [505, 600]$ (the most distant eigenvalue cannot be closer to the origin than 505, because one eigenvalue is in K_1. We conclude that $\|C\|_2 \in [505, 600]$.

The eigenvalues of C^{-1} are the reciprocals of the eigenvalues of C. The eigenvalues of C^{-1} must lie in the intervals $I_1 = [1/585, 1/505]$, $I_2 = [-1/5, -1/15]$, $I_3 = [-1/275, -1/345]$, and $I_4 = [-1/480, -1/600]$. The absolute values of the eigenvalues of C^{-1} lies in

$$\widetilde{I}_2 = [1/15, 1/5], \quad \widetilde{I}_3 = [1/345, 1/274], \quad \widetilde{I}_4 = [1/600, 1, 480].$$

The most distant interval is $\widetilde{I}_2$ that, moreover, does not intersect any of the intervals I_1, $\widetilde{I}_3$, and $\widetilde{I}_4$. This observation implies $\|C^{-1}\|_2 \in [1/15, 1/5]$.

We conclude that $\kappa(C) \in [505/15, 600/5] = [101/3, 120]$.

Problem 3.2: By applying the Gershgorin theorem, find both the maximum value $a \in \mathbb{R}$ and the minimum value $b \in \mathbb{R}$ such that $a \leq \kappa(C) \leq b$, where $\kappa(C)$ is the condition number of the matrix $C = \begin{pmatrix} -12 & 2 \\ 2 & 10 \end{pmatrix}$. Use the spectral norm $\|\cdot\|_2$ to calculate $\kappa(C)$.

Next, calculate the exact values $\|C\|_2$, $\|C^{-1}\|_2$, and $\kappa(C)$. Is the exact condition number less than 4/3? Check whether the exact condition number meets the Gershgorin-theorem-based estimate.

Solution: $\kappa(C) = \|C\|_2\|C^{-1}\|_2$.

The matrix C is real and symmetric, so $\|C\|_2 = |\tilde{\lambda}|$, where $\tilde{\lambda}$ it the eigenvalue of C with the maximum absolute value (all the eigenvalues of C are real).

According to the Gershgorin theorem, all the eigenvalues lie in the union of the following discs:

$$K_1 = \{z \in \mathbb{R} : \ | -12 - z| \leq 2\} = [-14, -10],$$
$$K_2 = \{z \in \mathbb{R} : \ |10 - z| \leq 2\} = [8, 12].$$

By introducing $\widetilde{K}_1 = [10, 14]$, we see that $\widetilde{K}_1 \cap K_2 \neq \emptyset$. The distance of the most distant eigenvalue from the origin must be at least 10 (this observation is similar to that in Problem 3.1). It holds that $\|C\|_2 \in [10,\ 14]$.

The eigenvalues of C^{-1} are the reciprocals of the eigenvalues of C. The eigenvalues of C^{-1} lie in the intervals $[-1/10, -1/14]$ and $[1/12, 1/8]$. The norm $\|C^{-1}\|_2$ is estimated on the basis of $[1/14, 1/10]$ and $[1/12, 1/8]$. These intervals have a nonempty intersection but neither is the subset of the other. We get $\|C^{-1}\|_2 \in [1/12, 1/8]$.

We conclude that $\kappa(C) \in [10/12,\ 14/8] = [5/6,\ 7/4]$.

Let us determine $\kappa(C)$, the exact condition number. The roots of the characteristic polynomial

$$\det(C - \lambda I) = (-12 - \lambda)(10 - \lambda) - 4 = \lambda^2 + 2\lambda - 124$$

are

$$\frac{-2 \pm \sqrt{4 + 4 \cdot 124}}{2} = \frac{-2 \pm \sqrt{500}}{2} = \frac{-2 \pm 10\sqrt{5}}{2} = -1 \pm 5\sqrt{5}.$$

We obtain $\|C\|_2 = 1 + 5\sqrt{5}$.

The eigenvalues of C^{-1}

$$\frac{1}{-1 + 5\sqrt{5}} = \frac{1}{-1 + 5\sqrt{5}} \frac{5\sqrt{5} + 1}{5\sqrt{5} + 1} = \frac{5\sqrt{5} + 1}{125 - 1} = \frac{5\sqrt{5} + 1}{124},$$
$$\frac{1}{-1 - 5\sqrt{5}} = \frac{1}{-1 - 5\sqrt{5}} \frac{-5\sqrt{5} + 1}{-5\sqrt{5} + 1} = \frac{-5\sqrt{5} + 1}{125 - 1} = \frac{-5\sqrt{5} + 1}{124}.$$

implies $\|C^{-1}\|_2 = \frac{5\sqrt{5}+1}{124}$, so that

$$\kappa(C) = \|C\|_2 \|C^{-1}\|_2 = \frac{5\sqrt{5} + 1}{124}(1 + 5\sqrt{5}) = \frac{125 + 10\sqrt{5} + 1}{124} = \frac{63 + 5\sqrt{5}}{62}$$
$$\leq \frac{63 + 5 \cdot 3}{62} = \frac{78}{62} = \frac{39}{31} \leq \frac{40}{30} = \frac{4}{3}.$$

We conclude that $\kappa(C) = \dfrac{63 + 5\sqrt{5}}{62}$ is less than $4/3 = 16/12 < 21/12 = 7/4$. At the same time, we see that $\kappa(C) \geq 1$. By combining these two inequalities, we obtain $\kappa(C) \in [5/6,\ 7/4]$.

4 Vector and Matrix Norms

Problem 4.1: The matrices A, B, C and vectors v, w are given as

$$A = \begin{pmatrix} 2 & -1 & 2 & a \\ -1 & 0 & 4 & -4 \\ 5 & -1 & 2 & 4 \\ -3 & 2 & -4 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} -1 & 3 \\ 3 & -1 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \quad v = \begin{pmatrix} 2 \\ 3 \\ -1 \\ -2 \end{pmatrix}, \quad w = Av.$$

Find a negative real number a such that $\|w\|_2 = 7\sqrt{2}$. Also calculate $\|w\|_1$, $\|w\|_\infty$, $\|A\|_1$, and $\|A\|_\infty$. Finally, determine $\|B\|_2$ and $\|C\|_2$.

Solution: $w = (-1-2a, 2, -3, 6)^{\mathrm{T}}$ and $\|w\|_2 = \sqrt{49 + (1+2a)^2}$ lead to $a = -4$.
Then $w = (7, 2, -3, 6)^{\mathrm{T}}$ and $\|w\|_1 = 7 + 2 + 3 + 6 = 18$.
$\|w\|_\infty = 7$.
$\|A\|_1 = \max\{11, 4, 12, 13\} = 13$.
$\|A\|_\infty = \max\{9, 9, 12, 10\} = 12$.
The matrix B is symmetric. Its eigenvalues are the roots of the characteristic polynomial $\lambda^2 + 2\lambda - 8$, namely, $\lambda_1 = 2$ and $\lambda_2 = -4$. As a consequence, $\|B\|_2 = 4$.
The matrix C is not symmetric. $\|C\|_2 = \sqrt{\varrho(C^{\mathrm{T}}C)}$, where the matrix $C^{\mathrm{T}}C = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ has the double eigenvalue 2, that is, $\|C\|_2 = \sqrt{2}$.

Problem 4.2: Let the matrix A and the vector z be:

$$A = \begin{pmatrix} 1 & a & 2 \\ -2 & 0 & b \\ c & 1 & 2 \end{pmatrix}, \quad z = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}.$$

Let the vectors v and w be defined by means of A and z as follows: $v = Az$ and $w = A^{\mathrm{T}}z$. Find $a, b, c \in \mathbb{R}$ such that $v = w$ and $\|v\|_2 = 2\sqrt{2}$.

Solution:

$$v = Az = \begin{pmatrix} 1 & a & 2 \\ -2 & 0 & b \\ c & 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3-a \\ -2+b \\ c+1 \end{pmatrix}, \tag{3}$$

$$w = A^{\mathrm{T}}z = \begin{pmatrix} 1 & -2 & c \\ a & 0 & 1 \\ 2 & b & 2 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3+c \\ a+1 \\ 4-b \end{pmatrix}.$$

The equality $v = w$ leads to three equations

$$\begin{aligned} a \quad + c &= 0, \\ a - b \quad &= -3, \\ b + c &= 3 \end{aligned}$$

that are easily solvable. We obtain $a = -p$, $b = 3 - p$ and $c = p$, where p is a real parameter.

The right-hand side of (3) gives $\|v\|_2^2$ expressed via p

$$\begin{aligned} \|v\|_2^2 &= \|(3+p, 1-p, p+1)\|_2^2 = (3+p)^2 + (1-p)^2 + (p+1)^2 \\ &= 9 + 6p + p^2 + 1 - 2p + p^2 + p^2 + 2p + 1 = 3p^2 + 6p + 11, \end{aligned}$$

which together with $\|v\|_2^2 = 8$ results in the equation

$$3p^2 + 6p + 11 = 8 \text{ rearranged as } p^2 + 2p + 1 = 0$$

and solved by $p_{1,2} = -1$.

The parameters are $a = 1$, $b = 4$, and $c = -1$.

By substituting a, b, c into A, we can calculate $v = Az$, $w = A^{\mathrm{T}}z$,$\|v\|_2$ and check the validity of $v = w$ and $\|v\|_2 = 2\sqrt{2}$.

Problem 4.3: The vector $v = (a, 1)^{\mathrm{T}}$ depends on a parameter a. Find all $a \in \mathbb{R}$ such that $\sqrt{2}\|v\|_2 = \|v\|_1$.

Solution: The squared equality reads $2\|v\|_2^2 = \|v\|_1^2$, that is, $2(a^2 + 1) = a^2 + 2|a| + 1$. After rearranging, we obtain $a^2 - 2|a| + 1 = 0$ or $(|a| - 1)^2 = 0$. It is directly seen that $a = 1$ of $a = -1$. The correctness of the results can be easily checked.

5 Inner Product

Let us recall that the inner (scalar) product of two elements of a vector space V is a map that maps ordered pairs of elements of V to real numbers. The inner product of $x, y \in V$ is denoted by (x, y) or $\langle x, y\rangle$ or $\langle\langle x, y\rangle\rangle$, for example.

Definition: If for any triple $x, y, z \in V$ and $\forall\alpha \in \mathbb{R}$ hold

$$(x, y) = (y, x), \tag{4}$$
$$(x + z, y) = (x, y) + (z, y), \tag{5}$$
$$(\alpha x, y) = \alpha(x, y), \tag{6}$$
$$(x, x) \geq 0, \tag{7}$$
$$(x, x) = 0 \Rightarrow x = 0^*, \quad 0^* \text{ is the zero element of } V, \tag{8}$$

then the map $(\cdot, \cdot) : V \times V \to \mathbb{R}$ is called the inner product on the vector space V.

Simple consequences can be easily inferred. Take, for instance, (4) and (5). Then

$$(x, y + z) = (y + z, x) = (y, x) + (z, x) = (x, y) + (x, z).$$

Similarly, (4) a (6) implies

$$(x, \alpha y) = (\alpha y, x) = \alpha(y, x) = \alpha(x, y).$$

Problem 5.1: Let $\mathbb{R}^3$ be the vector space of ordered triples of real numbers. Does the expression $\langle u, v\rangle = u_1v_1 + 2u_2v_2 + 4u_3v_3$, where $u = (u_1, u_2, u_3) \in \mathbb{R}^3$ and $v = (v_1, v_2, v_3) \in \mathbb{R}^3$, define an inner product on $\mathbb{R}^3$?

Solution: . Let us check whether (4)-(8) hold for $\langle u, v\rangle$. Besides u and v, we will also use $z = (z_1, z_2, z_3)$ and $\alpha \in \mathbb{R}$. Let us emphasize that u, v, z are arbitrary vectors from $\mathbb{R}^3$ and that α represents an arbitrary real number.

$$\begin{aligned}
\langle u, v\rangle &= u_1v_1 + 2u_2v_2 + 4u_3v_3 = v_1u_1 + 2v_2u_2 + 4v_3u_3 = \langle v, u\rangle \Rightarrow (4),\\
\langle u + z, v\rangle &= (u_1 + z_1)v_1 + 2(u_2 + z_2)v_2 + 4(u_3 + z_3)v_3\\
&= u_1v_1 + 2u_2v_2 + 4u_3v_3 + z_1v_1 + 2z_2v_2 + 4z_3v_3 = \langle u, v\rangle + \langle z, v\rangle \Rightarrow (5),\\
\langle \alpha u, v\rangle &= \alpha u_1v_1 + 2\alpha u_2v_2 + 4\alpha u_3v_3 = \alpha\langle u, v\rangle \Rightarrow (6),\\
\langle u, u\rangle &= u_1u_1 + 2u_2u_2 + 4u_3u_3 \geq 0 \Rightarrow (7).
\end{aligned}$$

If $\langle u,u\rangle = u_1^2 + 2u_2^2 + 4u_3^2 = 0$, then $u_1 = u_2 = u_3 = 0$, that is, $u = 0^*$, where 0^* stands for the zero vector $(0,0,0)$. This proves (8).

The map is an inner product on $\mathbb{R}^3$.

Problem 5.2: Let $V = \mathbb{R}^3$ a $W = \{(u_1,u_2,u_3) \in \mathbb{R}^3 |\ u_2 = 0\}$. Does the expression $\langle u,v\rangle = 2u_1v_1 - 3u_2v_2 + 5u_3v_3$, where $u,v \in \mathbb{R}^3$, define an inner product
1) on V;
2) on W?

Solution: 1) Let us choose $u = (0,1,0) \in \mathbb{R}^3$. Then $\langle u,u\rangle = -3 < 0$ and (7) is not valid. The map is not an inner product on the space $\mathbb{R}^3$.

2) If $u,v \in W$, then $\langle u,v\rangle = 2u_1v_1 + 5u_3v_3$. By performing the same steps as in Problem 5.1, we check the validity of (4)-(8) for any vectors $u,v,z \in W$. The map is an inner product on W.

Problem 5.3: Let P be the space of all polynomials defined on the interval $[0,1]$ and let P be provided with the inner product $(u,v) = \int_0^1 u(x)v(x)\,\mathrm{d}x$ and the norm $\|u\| = \sqrt{(u,u)}$, where $u,v \in P$. Find all real numbers a,b,c such that the two following conditions hold simultaneously:
1) $p(x) = ax^2 + bx + c$ is orthogonal to all linear polynomials;
2) $\|p\| = \sqrt{5}$.

Solution: Let us realize that 1) is equivalent to the orthogonality of p to 1 and to x. We obtain two equations for three unknowns a,b,c:

$$(p,1) = \int_0^1 (ax^2 + bx + c)\,\mathrm{d}x = \left[a\frac{x^3}{3} + b\frac{x^2}{2} + cx\right]_0^1 = \frac{a}{3} + \frac{b}{2} + c = 0$$

and

$$(p,x) = \int_0^1 (ax^3 + bx^2 + cx)\,\mathrm{d}x = \left[a\frac{x^4}{4} + b\frac{x^3}{3} + c\frac{x^2}{2}\right]_0^1 = \frac{a}{4} + \frac{b}{3} + \frac{c}{2} = 0.$$

We express a and b by means of c, i.e, $a = 6c$ and $b = -6c$. Then $p = 6cx^2 - 6cx + c$. According to 2),

$$\begin{aligned}\|p\|^2 &= c^2\int_0^1 (6x^2 - 6x + 1)^2\,\mathrm{d}x = c^2\int_0^1 (36x^4 - 72x^3 + 48x^2 - 12x + 1)\,\mathrm{d}x \\ &= c^2(36/5 - 72/4 + 48/3 - 12/2 + 1) = c^2(36/5 - 18 + 16 - 6 + 1) \\ &= c^2(36/5 - 7) = c^2/5\end{aligned}$$

is equal to 5. We conclude $c = \pm 5$ and obtain two solutins: $p = 30x^2 - 30x + 5$ and $p = -30x^2 + 30x - 5$.

Problem 5.4: Let the space $\mathbb{R}^3$ and the common inner product (v,w), $v,w \in \mathbb{R}^3$, be given. Let us assume that $u_1,u_2,u_3,u_4 \in \mathbb{R}^3$ are mutually orthogonal vectors and that $(u_k,u_k) = k^2$ for $k = 1,2,3,4$. Let the vectors u and z be defined as

$$u = -3u_1 + 2u_2 + \frac{1}{3}u_3 - u_4 \qquad \text{a} \qquad z = -u_1 + \frac{1}{4}u_2 + \frac{1}{3}u_3 + u_4.$$

Calculate (u, z).

Solution: The orthogonality implies

$$(u,z) = (-3u_1, -u_1) + \left(2u_2, \frac{1}{4}u_2\right) + \left(\frac{1}{3}u_3, \frac{1}{3}u_3\right) + (-u_4, u_4) = 3\cdot 1 + \frac{1}{2}\cdot 2^2 + \frac{1}{9}\cdot 3^2 - 4^2 = -10.$$

6 Iterative Methods

Problem 6.1: Let a system of linear algebraic equations $Cx = y$ be given, where

$$C = \begin{pmatrix} 4 & 1 & 2 \\ 1 & 2 & 0 \\ 1 & 1 & 3 \end{pmatrix} \quad \text{and} \quad y = \begin{pmatrix} 8 \\ 2 \\ 6 \end{pmatrix}.$$

a) By applying the Jacobi iterative method with the starting vector $x^{(0)} = (1,\, 1,\, 1)^{\mathrm{T}}$, calculate $x^{(1)}$ and $x^{(2)}$, the first and the second iterations.

b) Does the sequence of the Jacobi approximations $x^{(k)}$, $k = 1, 2, 3, \ldots$, converge to y, the exact solution of $Cx = y$?

c) Do not solve the system but estimate the difference $\|\widehat{x} - x^{(2)}\|_\infty$, where $\widehat{x}$ is the exact solution. Can we guarantee that $\|\widehat{x} - x^{(2)}\|_\infty < 0.4$?

d) Can we guarantee that $\|\widehat{x} - x^{(25)}\|_\infty \leq 0.01$?

(This may help: $\log_{10} 2 \approx 0{,}301$, $\log_{10} 3 \approx 0{,}477$, $\log_{10} 5 \approx 0{,}699$, $\log_{10} 7 \approx 0{,}845$.)

e) By using an estimate, infer the minimum k such that $\|\widehat{x} - x^{(k)}\|_\infty \leq 0.001$

Solution: a) Let us infer the iteration matrix A:

$$A = D^{-1}\hat{C} = \begin{pmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 0 & -1 & -2 \\ -1 & 0 & 0 \\ -1 & -1 & 0 \end{pmatrix} = - \begin{pmatrix} 0 & 1/4 & 1/2 \\ 1/2 & 0 & 0 \\ 1/3 & 1/3 & 0 \end{pmatrix}.$$

Next,

$$b = D^{-1}y = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}.$$

We can start the iterative process $x^{(k+1)} = Ax^{(k)} + b$, where $k = 0, 1, \ldots,$

$$x^{(1)} = - \begin{pmatrix} 0 & 1/4 & 1/2 \\ 1/2 & 0 & 0 \\ 1/3 & 1/3 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix} = - \begin{pmatrix} 3/4 \\ 1/2 \\ 2/3 \end{pmatrix} + \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 5/4 \\ 1/2 \\ 4/3 \end{pmatrix},$$

$$x^{(2)} = - \begin{pmatrix} 0 & 1/4 & 1/2 \\ 1/2 & 0 & 0 \\ 1/3 & 1/3 & 0 \end{pmatrix} \begin{pmatrix} 5/4 \\ 1/2 \\ 4/3 \end{pmatrix} + \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix} = - \begin{pmatrix} 1/8 + 4/6 \\ 5/8 \\ 5/12 + 1/6 \end{pmatrix} + \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}$$

$$= - \begin{pmatrix} 19/24 \\ 5/8 \\ 7/12 \end{pmatrix} + \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 29/24 \\ 3/8 \\ 17/12 \end{pmatrix}.$$

b) The convergence follows from $\|A\|_\infty = \frac{3}{4} < 1$ or from the fact that the matrix C has a dominant main diagonal.

c) The estimate

$$\|\widehat{x} - x^{(k)}\| \leq \|A\|^k \|x_0\| + \frac{\|A\|^k}{1 - \|A\|} \|b\|$$

leads to

$$\|\widehat{x} - x^{(2)}\|_\infty \leq \|A\|_\infty^2 \|x_0\|_\infty + \frac{\|A\|_\infty^2}{1 - \|A\|_\infty} \|b\|_\infty = \left(\frac{3}{4}\right)^2 \cdot 1 + \frac{\left(\frac{3}{4}\right)^2}{\frac{1}{4}} \cdot 2 = \frac{81}{16}.$$

Another estimate

$$\|\widehat{x} - x^{(k)}\| \leq \frac{\|A\|}{1 - \|A\|} \|x^{(k)} - x^{(k-1)}\|$$

gives

$$\|\widehat{x} - x^{(2)}\|_\infty \leq \frac{\|A\|_\infty}{1 - \|A\|_\infty} \|x^{(2)} - x^{(1)}\|_\infty = 3 \left\| \left(\frac{-1}{24}, \frac{-1}{8}, \frac{1}{12}\right)^{\mathrm{T}} \right\|_\infty = \frac{3}{8}.$$

Yes, the latter estimate guarantees that $\|\widehat{x} - x^{(2)}\|_\infty \leq 3/8 < 0.4$.[5]

d) Analogously to c) (the first estimate), we obtain

$$\|\widehat{x} - x^{(25)}\|_\infty \leq \|A\|_\infty^{25} \|x_0\|_\infty + \frac{\|A\|_\infty^{25}}{1 - \|A\|_\infty} \|b\|_\infty = \left(\frac{3}{4}\right)^{25} \left(1 + \frac{2}{\frac{1}{4}}\right) = 9 \left(\frac{3}{4}\right)^{25}.$$

We ask whether $2 \log_{10} 3 + 25 \log_{10} \left(\frac{3}{4}\right)$ is less than -2 (i.e., $\log_{10} 0.01$). A short calculation

$$2 \cdot 0.477 + 25(0.477 - 0.602) = 0.954 - 25 \cdot 0.125 = -2.171 < -2$$

shows that $\|\widehat{x} - x^{(25)}\|_\infty \leq 0.01$.

e) We proceed as in d) and begin with

$$\|\widehat{x} - x^{(k)}\|_\infty \leq \|A\|_\infty^k \|x_0\|_\infty + \frac{\|A\|_\infty^k}{1 - \|A\|_\infty} \|b\|_\infty = \left(\frac{3}{4}\right)^k \left(1 + \frac{2}{\frac{1}{4}}\right) = 9 \left(\frac{3}{4}\right)^k.$$

We search for a minimum natural number k such that

$$\begin{aligned}
2 \log_{10} 3 + k \log_{10} \left(\frac{3}{4}\right) &\leq \log_{10} 0.001 \\
k \log_{10} \left(\frac{3}{4}\right) &\leq -3 - 2 \log_{10} 3 \\
k \left(\log_{10} 3 - 2 \log_{10} 2\right) &\leq -3 - 2 \log_{10} 3 \\
k &\geq \frac{-3 - 2 \log_{10} 3}{\log_{10} 3 - 2 \log_{10} 2} \\
k &\geq \frac{-3 - 0.954}{0.477 - 0.602} \\
k &\geq 31.6.
\end{aligned}$$

[5] By observing that $\frac{3}{8} ? \frac{4}{10} \Leftrightarrow 30 ? 32$, we infer that ? stands for $<$.

According to the last inequality, $\|\widehat{x} - x^{(32)}\|_\infty < 0.001$.

Problem 6.2: a) Let $C = \begin{pmatrix} 4 & 1 & 2 \\ 1 & 2 & 0 \\ 1 & 0 & 3 \end{pmatrix}$ a $y = \begin{pmatrix} 36 \\ 18 \\ 15 \end{pmatrix}$. Solve the system $Cx = y$ approximately by performing two iterations of the Jacobi algorithm with the starting vector $x^{(0)} = (12,\, 4,\, 2)^{\mathrm{T}}$. That is, $x^{(1)}$ and $x^{(2)}$ will be calculated.

b) Does the sequence $x^{(k)}$ generated by the Jacobi algorithm, where $k \to \infty$, converge to the exact solution $\widehat{x}$?

c) The difference $\widehat{x} - x^{(k)}$ can be estimated in different norms, namely, $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$. Which norm should be used to obtain the fastest decrease of the estimate with respect to $k \to \infty$?

d) Perform one iteration of the Gauss-Seidel algorithm with the starting vector $x_{\mathrm{GS}}^{(0)} = (12,\, 4,\, 2)^{\mathrm{T}}$ to obtain $x_{\mathrm{GS}}^{(1)}$. Avoid the calculation of the iteration matrix.

e) Use the Gaussian elimination to solve $Cx = y$ exactly and check the correctness of the result. Calculate $\|x - x^{(1)}\|_1$ and $\|x - x_{\mathrm{GS}}^{(1)}\|_1$. Which iterative method gave a more accurate result?

Solution: a) Jacobi iteration matrix A:

$$A = D^{-1}\hat{C} = \begin{pmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix} \begin{pmatrix} 0 & -1 & -2 \\ -1 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -1/4 & -1/2 \\ -1/2 & 0 & 0 \\ -1/3 & 0 & 0 \end{pmatrix}.$$

Next,

$$b = D^{-1}y = \begin{pmatrix} 9 \\ 9 \\ 5 \end{pmatrix}.$$

The Jacobi iteration process $x^{(k+1)} = Ax^{(k)} + b$, where $k = 0, 1, \ldots$, than gives

$$x^{(1)} = \begin{pmatrix} 0 & -1/4 & -1/2 \\ -1/2 & 0 & 0 \\ -1/3 & 0 & 0 \end{pmatrix} \begin{pmatrix} 12 \\ 4 \\ 2 \end{pmatrix} + \begin{pmatrix} 9 \\ 9 \\ 5 \end{pmatrix} = \begin{pmatrix} -2+9 \\ -6+9 \\ -4+5 \end{pmatrix} = \begin{pmatrix} 7 \\ 3 \\ 1 \end{pmatrix},$$

$$x^{(2)} = \begin{pmatrix} 0 & -1/4 & -1/2 \\ -1/2 & 0 & 0 \\ -1/3 & 0 & 0 \end{pmatrix} \begin{pmatrix} 7 \\ 3 \\ 1 \end{pmatrix} + \begin{pmatrix} 9 \\ 9 \\ 5 \end{pmatrix} = \begin{pmatrix} \frac{-5+36}{4} \\ \frac{-7+18}{2} \\ \frac{-7+15}{3} \end{pmatrix} = \begin{pmatrix} 31/4 \\ 11/2 \\ 8/3 \end{pmatrix}.$$

b) The convergence is a consequence of $\|A\|_\infty = 3/4 < 1$ or $\|A\|_1 = 5/6 < 1$ or the fact that A is diagonally dominant.

c) The speed of the decrease is determined by $\|A\|_*^k$, where * stand for a particular matrix norm index.[6] The norms $\|A\|_\infty$ and $\|A\|_1$ are already known (see b)); we have to calculate $\|A\|_2 = \sqrt{\varrho(A^{\mathrm{T}}A)}$.

$$A^{\mathrm{T}}A = \begin{pmatrix} 0 & -1/2 & -1/3 \\ -1/4 & 0 & 0 \\ -1/2 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & -1/4 & -1/2 \\ -1/2 & 0 & 0 \\ -1/3 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \frac{13}{36} & 0 & 0 \\ 0 & \frac{1}{16} & \frac{1}{8} \\ 0 & \frac{1}{8} & \frac{1}{4} \end{pmatrix}.$$

[6] See the first estimate in the solution of Problem 6.1 c).

Characteristic polynomial

$$\left(\frac{13}{36}-\lambda\right)\left(\frac{1}{16}-\lambda\right)\left(\frac{1}{4}-\lambda\right)-\left(\frac{13}{36}-\lambda\right)\frac{1}{64}$$
$$=\left(\frac{13}{36}-\lambda\right)\left(\frac{1}{64}-\frac{\lambda}{16}-\frac{\lambda}{4}+\lambda^2-\frac{1}{64}\right)=\left(\frac{13}{36}-\lambda\right)\lambda\left(\lambda-\frac{5}{16}\right)$$

has three roots, namely, 0, 5/16, and 13/36. It holds $5/16 < 5/15 = 1/3 = 12/36 < 13/36$ and

$$\|A\|_2=\sqrt{\frac{13}{36}}=\frac{\sqrt{13}}{6}<\frac{4}{6}=\frac{8}{12}<\frac{9}{12}=\frac{3}{4}=\|A\|_\infty<\|A\|_1.$$

We conclude that $\|\cdot\|_2$ should be used in the estimate.

d) Gauss-Seidel iteration[7]

$$x_{\mathrm{GS},1}^{(1)}=-\frac{1}{4}(4+2\cdot 2)+\frac{36}{4}=7,$$
$$x_{\mathrm{GS},2}^{(1)}=-\frac{1}{2}(7+0)+\frac{18}{2}=\frac{11}{2},$$
$$x_{\mathrm{GS},3}^{(1)}=-\frac{1}{3}(7+0)+\frac{15}{3}=\frac{8}{3}.$$

e) Gaussian elimination

$$\left(\begin{array}{ccc|c}4&1&2&36\\1&2&0&18\\1&0&3&15\end{array}\right)\sim\left(\begin{array}{ccc|c}1&0&3&15\\1&2&0&18\\4&1&2&36\end{array}\right)\sim\left(\begin{array}{ccc|c}1&0&3&15\\0&2&-3&3\\0&1&-10&-24\end{array}\right)\sim\left(\begin{array}{ccc|c}1&0&3&15\\0&2&-3&3\\0&0&17&51\end{array}\right).$$

Then $\widehat{x}=(6,6,3)^{\mathrm{T}}$. To check the result, we simply multiply $\widehat{x}$ by C.

We see from $\|\widehat{x}-x^{(1)}\|_1=\|(-1,3,2)\|_1=6$ and $\|\widehat{x}-x_{\mathrm{GS}}^{(1)}\|_1=\|(-1,1/2,1/3)\|_1=\dfrac{6+3+2}{6}=\dfrac{11}{6}$ that the Gauss-Seidel method delivered a more accurate result.

7 Linear 2nd Order Ordinary Differential Equations

This chapter is based on Chapter 2 of *O. Zindulka: Matematika 3, Česká technika – nakladatelství ČVUT, Praha, 2007.*

We will deal with a nonhomogeneous equation

$$y''(x)+p(x)y'(x)+q(x)y(x)=f(x), \tag{9}$$

[7] Although the matrix form of the iteration, i.e.,

$$x^{(k+1)}=(D-L)^{-1}Ux^{(k)}+(D-L)^{-1}y,\quad k=0,1,\ldots,$$

is elegant, it is not used in calculations. The following formula is used instead:

$$x_i^{(k+1)}=-\frac{1}{c_{ii}}\left(\sum_{j=1}^{i-1}c_{ij}x_j^{(k+1)}+\sum_{j=i+1}^{n}c_{ij}x_j^{(k)}\right)+\frac{y_i}{c_{ii}},\quad i=1,2,\ldots,n.$$

where p, q, and f are known functions, whereas y is to be identified from (9) and boundary or initial conditions (see Problem 7.4.1, for instance). For brevity, y will often be used instead of $y(x)$. Next,

$$y'' + p(x)y' + q(x)y = 0, \tag{10}$$

the homogeneous parallel to (9) will also be considered.

Superposition principle
If y_1 and y_2 solve the respective equations

$$y_1'' + p(x)y_1' + q(x)y_1(x) = f_1(x), \tag{11}$$
$$y_2'' + p(x)y_2' + q(x)y_2(x) = f_2(x), \tag{12}$$

then, for any $r, s \in \mathbb{R}$,

$$y = ry_1 + sy_2 \quad \text{is a solution to } y'' + p(x)y' + q(x)y(x) = rf_1(x) + sf_2(x). \tag{13}$$

Indeed, the statement (13) is easily proved by inserting $y = ry_1 + sy_2$ into $y'' + p(x)y' + q(x)y(x)$ and by applying (11)–(12).

As a consequence, we observe that

- If y_P and y_H is a solution of (9) and (10), respectively, then $y = y_\mathrm{P} + y_\mathrm{H}$ is a solution of (9).
- If both y_1 and y_2 are solutions of (9), then $y = y_1 - y_2$ is a solution of (10).

The observation gives a strategy for finding all the solutions to (9) and solving an initial or boundary value problem:
(A) Find two[8] linearly independent solutions of (10) and denote them y_1 and y_2, for instance. Then any solution y_H of (10) is a linear combination of y_1 and y_2, that is,

$$y_\mathrm{H} = c_1y_1 + c_2y_2,$$

(B) Find an arbitrary solution y_P of (9).
(C) Define a *general solution* of (9) as

$$y = y_\mathrm{P} + y_\mathrm{H} = y_\mathrm{P} + c_1y_1 + c_2y_2.$$

(D) If initial or boundary conditions are prescribed, apply them to the general solution y and identify the parameters c_1 and c_2.

Let us recall two methods for finding a particular solution y_P.

7.1 Variation of Constants Method

The method is based on the fact that if y_1 and y_2 are two linearly independent solutions of (10) and f is the right-hand side of (9), then functions v_1 and v_2 exist such that

$$v_1'y_1 + v_2'y_2 = 0, \tag{14}$$
$$v_1'y_1' + v_2'y_2' = f, \tag{15}$$

[8]It can be proved that the null space of the equation (10) has dimension equal to two.

and $y_P = v_1 y_1 + v_2 y_2$ is a solution of (9).

On the basis of the equations (14)–(15), it is possible to infer a direct formula that defines the particular solution:

$$y_P = -y_1 \int \frac{y_2 f}{W}\,\mathrm{d}x + y_2 \int \frac{y_1 f}{W}\,\mathrm{d}x, \tag{16}$$

where

$$W = y_1 y_2' - y_1' y_2 \tag{17}$$

is the Wronski determinant.

7.2 Parameter Identification Method (aka Special Right-Hand Side)

This method is applicable to the equation (9) where the parameters p and q are constant and the right-hand side has a special form, that is,

$$y'' + ay' + by = \mathrm{e}^{\alpha x}\left(P_1(x)\cos\beta x + P_2(x)\sin\beta x\right), \tag{18}$$

where P_1 and P_2 are polynomials.

Let us recall that a *characteristic equation*

$$\lambda^2 + a\lambda + b = 0 \tag{19}$$

is associated with (18).

The method is based on the following statement:
For the equation (18), polynomials Q_1 and Q_2 exist such that

$$y_P(x) = x^n \mathrm{e}^{\alpha x}\left(Q_1(x)\cos\beta x + Q_2(x)\sin\beta x\right) \tag{20}$$

is a particular solution of the equation (18). The degree of Q_1 and Q_2 is less than or equal to the maximum degree of the polynomials P_1 and P_2. Moreover, let $\mu = \alpha + \mathrm{i}\beta$ be a complex number defined by α and β (see the right-hand side of (18)), then
$n = 0$ if μ is not a characteristic number (root) of (19);
$n = 1$ if μ is a characteristic number (root) of (19) with multiplicity one;
$n = 2$ if μ is a characteristic number (root) of (19) with multiplicity two.

7.3 Linearly Independent Solutions y_1 and y_2

In case of the equation with constant parameters, see the left-hand side of (18), the two linearly independent solutions y_1 and y_2 (see (A) on page 20) are determined by the roots of the characteristic equation (19):

- if $\lambda_1, \lambda_2 \in \mathbb{R}$ are two <u>different</u> roots of (19), then $y_1(x) = \mathrm{e}^{\lambda_1 x}$ and $y_2(x) = \mathrm{e}^{\lambda_2 x}$
- if $\lambda \in \mathbb{R}$ is a <u>double</u> root of (19), then $y_1(x) = \mathrm{e}^{\lambda x}$ and $y_2(x) = x\mathrm{e}^{\lambda x}$
- if $\lambda_1 = \alpha + \mathrm{i}\beta$ and $\lambda_2 = \alpha - \mathrm{i}\beta$, where $\beta \neq 0$, are two different complex roots of (19), then $y_1(x) = \mathrm{e}^{\alpha x}\cos\beta x$ and $y_2(x) = \mathrm{e}^{\alpha x}\sin\beta x$

7.4 Solved Problems

Problem 7.4.1: a) Find a general solution of $y'' + y = x^2$.
b) Find a solution such that $y(0) = 2$ and $y'(0) = 3$.

Solution: a) The roots of the characteristic equation $\lambda^2 + 1 = 0$ are $\lambda_1 = \mathrm{i}$ and $\lambda_2 = -\mathrm{i}$. As a consequence, $y_1(x) = \cos x$ and $y_2(x) = \sin x$ are two linearly independent solutions of the homogenous equation $y'' + y = 0$, see Section 7.3.

To infer y_P, let us apply the method presented in Section 7.1. The Wronski determinant, see (17), $W = y_1 y_2' - y_1' y_2 = \cos^2 x + \sin^2 x = 1$.

By virtue of (16) and integration by parts,

$$\begin{aligned} y_\mathrm{P} &= -\cos x \int x^2 \sin x \,\mathrm{d}x + \sin x \int x^2 \cos x \,\mathrm{d}x \\ &= -\cos x(-x^2 \cos x + 2\cos x + 2x \sin x) + \sin x(x^2 \sin x - 2\sin x + 2x \cos x) \\ &= x^2 - 2. \end{aligned} \tag{21}$$

The general solution, see (C) on page 20,

$$y = y_\mathrm{P} + y_\mathrm{H} = y_\mathrm{P} + c_1 y_1 + c_2 y_2 = x^2 - 2 + c_1 \cos x + c_2 \sin x, \tag{22}$$

where $c_1, c_2 \in \mathbb{R}$.

For comparison, let us show how the special right-hand side approach (see Section 7.2) is used to obtain y_P.

The right-hand side of the ordinary differential equation (that is, x^2) coincides with the right-hand side of (18) if $\alpha = 0$, $\beta = 0$, $P_1(x) = x^2$, and $P_2(x) = 0$. We observe that $\mu = \alpha + \mathrm{i}\beta = 0$.

Since $\lambda_1 = \mathrm{i} \neq \mu \neq -\mathrm{i} = \lambda_2$, $n = 0$ according to Section 7.2. The maximum degree of P_1 and P_2 is equal to 2, therefore we have to assume that Q_1 is also a quadratic polynomial. We do not need a polynomial Q_2 because $Q_2(x) \sin 0x = 0$ independently of Q_2. Referring to (20), we can say that

$$y_\mathrm{P} = x^0 \mathrm{e}^0 (Q_1(x) \cdot 1 + 0) = ax^2 + bx + c$$

for some real numbers a, b, and c that have to be identified. To this end, we infer $y_\mathrm{P}'' = 2a$ and together with y_P substitute into $y'' + y = x^2$. We obtain

$$2a + ax^2 + bx + c = x^2. \tag{23}$$

By comparing the left- and right-hand side of (23), we arrive at $a = 1$, $b = 0$, and $2a + c = 0$, that is, $c = -2$. As a consequence,

$$y_\mathrm{P} = x^2 - 2,$$

which is (21).

b) We search for c_1 and c_2 such that (22) fulfills the initial conditions $y(0) = 2$ and $y'(0) = 3$. First,

$$y(0) = -2 + c_1 = 2 \;\Rightarrow\; c_1 = 4.$$

Second,

$$y'(0) = (2x - c_1 \sin x + c_2 \cos(x))|_{x=0} = c_2 = 3.$$

In summary, $y = x^2 - 2 + 4\cos x + 3\sin x$.

Problem 7.4.2: Find a general solution of $y'' - 4y' - 5y = \mathrm{e}^{2x}(2x + 3)$.

Solution: The characteristic polynominal $\lambda^2 - 4\lambda - 5 = 0$ has two roots, namely $\lambda_1 = -1$ and $\lambda_1 = 5$. As a consequence, $y_1(x) = \mathrm{e}^{-x}$ and $y_2(x) = \mathrm{e}^{5x}$ are two linearly independent solutions of the homogenous equation $y'' - 4y' - 5y = 0$, see Section 7.3. We define

$$y_{\mathrm{H}}(x) = c_1 y_1(x) + c_2 y_2(x) = c_1 \mathrm{e}^{-x} + c_2 \mathrm{e}^{5x},$$

where $c_1, c_2 \in \mathbb{R}$.

Since the special right-hand side approach is, if applicable, less laborious than the variation of constants, we will use the former approach.

The right-hand side of the ordinary differential equation (that is, $\mathrm{e}^{2x}(2x+3)$) coincides with the right-hand side of (18) if $\alpha = 2$, $\beta = 0$, $P_1(x) = 2x + 3$, and $P_2(x) = 0$. We observe that $\mu = \alpha + \mathrm{i}\beta = 2$.

Since $\lambda_1 = -1 \neq \mu \neq 5 = \lambda_2$, $n = 0$ according to Section 7.2. The maximum degree of P_1 and P_2 is equal to 1, therefore we have to assume that Q_1 is also a quadratic polynomial. We do not need a polynomial Q_2 because $Q_2(x)\sin 0x = 0$ independently of Q_2. Referring to (20), we can say that

$$y_{\mathrm{P}} = \mathrm{e}^{2x} Q_1(x) = \mathrm{e}^{2x}(ax + b)$$

for some real numbers a, and b that have to be identified. To this end, we infer

$$y'_{\mathrm{P}} = \mathrm{e}^{2x}(2ax + 2b + a) \text{ and } y''_{\mathrm{P}} = 4\mathrm{e}^{2x}(ax + b + a)$$

and substitute into $y'' - 4y' - 5y = \mathrm{e}^{2x}(2x + 3)$. We obtain

$$4\mathrm{e}^{2x}(ax + b + a) - 4\mathrm{e}^{2x}(2ax + 2b + a) - 5\mathrm{e}^{2x}(ax + b) = \mathrm{e}^{2x}(2x + 3).$$

This equation is easily simplified to

$$-9ax - 9b = 2x + 3.$$

By comparing the left- and right-hand side, we arrive at $a = -2/9$, $b = -1/3$. As a consequence,

$$y_{\mathrm{P}} = -\frac{1}{9}\mathrm{e}^{2x}(2x + 3),$$

$$y = y_{\mathrm{P}} + y_{\mathrm{H}} = y_{\mathrm{P}} + c_1 y_1 + c_2 y_2 = -\frac{2x + 3}{9}\mathrm{e}^{2x} + c_1 \mathrm{e}^{-x} + c_2 \mathrm{e}^{5x}.$$

Problem 7.4.3: Find a general solution of $y'' - 4y' + 4y = \mathrm{e}^{2x}(2x + 3)$.

Solution: The characteristic polynominal $\lambda^2 - 4\lambda + 4 = 0$ has one double root, namely $\lambda = 2$. As a consequence, $y_1(x) = \mathrm{e}^{2x}$ and $y_2(x) = x\mathrm{e}^{2x}$ are two linearly independent solutions of the homogenous equation $y'' - 4y' + 4y = 0$, see Section 7.3. We define

$$y_{\mathrm{H}}(x) = c_1 y_1(x) + c_2 y_2(x) = \mathrm{e}^{2x}(c_1 + c_2 x),$$

where $c_1, c_2 \in \mathbb{R}$.

Since the special right-hand side approach is, if applicable, less laborious than the variation of constants, we will use the former approach.

The right-hand side of the ordinary differential equation (that is, $\mathrm{e}^{2x}(2x+3)$) coincides with the right-hand side of (18) if $\alpha = 2$, $\beta = 0$, $P_1(x) = 2x+3$, and $P_2(x) = 0$. We observe that $\mu = \alpha + \mathrm{i}\beta = 2$ coincides with the double root of the characteristic polynomial. This means that $n = 2$ according to Section 7.2. The maximum degree of P_1 and P_2 is equal to 1, therefore we have to assume that Q_1 is also a quadratic polynomial. We do not need a polynomial Q_2 because $Q_2(x)\sin 0x = 0$ independently of Q_2. Referring to (20), we can say that

$$y_{\mathrm{P}} = x^2\mathrm{e}^{2x}Q_1(x) = x^2\mathrm{e}^{2x}(ax+b)$$

for some real numbers a and b that have to be identified. To this end, we infer

$$\begin{aligned} y'_{\mathrm{P}} &= \mathrm{e}^{2x}\left(2ax^3 + (2b+3a)x^2 + 2bx\right), \\ y''_{\mathrm{P}} &= \mathrm{e}^{2x}\left(4ax^3 + (12a+4b)x^2 + (6a+8b)x + 2b\right) \end{aligned}$$

and apply the expressions to the original equation, that is, to $y''_{\mathrm{P}} - 4y'_{\mathrm{P}} + 4y_{\mathrm{P}} = \mathrm{e}^{2x}(2x+3)$. We obtain

$$\begin{aligned} &\mathrm{e}^{2x}\left(4ax^3 + (12a+4b)x^2 + (6a+8b)x + 2b\right) \\ &\qquad - 4\mathrm{e}^{2x}\left(2ax^3 + (2b+3a)x^2 + 2bx\right) + 4x^2\mathrm{e}^{2x}(ax+b) \\ &= \mathrm{e}^{2x}(2x+3). \end{aligned}$$

After some effort, this equation is simplified to

$$6ax + 2b = 2x + 3.$$

By comparing the left- and right-hand side, we arrive at $a = 1/3$, $b = 3/2$. As a consequence,

$$\begin{aligned} y_{\mathrm{P}} &= \mathrm{e}^{2x}\left(\frac{3}{2}x^2 + \frac{1}{3}x^3\right), \\ y = y_{\mathrm{P}} + y_{\mathrm{H}} &= y_{\mathrm{P}} + c_1y_1 + c_2y_2 = \mathrm{e}^{2x}\left(c_1 + c_2x + \frac{3}{2}x^2 + \frac{1}{3}x^3\right), \quad c_1, c_2 \in \mathbb{R}, \end{aligned}$$

where the last expression stands for the general solution of the differential equation.

Problem 7.4.4: Find a general solution of $y'' - 4y' + 4y = 6\mathrm{e}^{2x}(2x+3) - 50\sin x$.

Solution: We observe that the right-hand side expression $6\mathrm{e}^{2x}(2x+3) - 50\sin x$ cannot be obtained from the right-hand side of (18) by "tuning" the parameters α, β, P_1, and P_2. The special right-hand side approach is still applicable, however.

Indeed, let us recall the superposition principle (11)–(13) and define

$$y'' - 4y' + 4y = \mathrm{e}^{2x}(2x+3), \tag{24}$$

$$y'' - 4y' + 4y = \sin x. \tag{25}$$

If y_{P_1} solves (24) and y_{P_2} solves (25), then

$$y_{\mathrm{P}} = 6y_{\mathrm{P}_1} - 50y_{\mathrm{P}_2} \tag{26}$$

is a particular solution to our problem.

We already know that

$$y_{\mathrm{P}_1} = \mathrm{e}^{2x}\left(\frac{3}{2}x^2 + \frac{1}{3}x^3\right),$$

see Problem 7.4.3. We also have

$$y_{\mathrm{H}}(x) = \mathrm{e}^{2x}(c_1 + c_2 x).$$

It remains to solve (25).

The right-hand side of the ordinary differential equation (25), that is, $\sin x$, coincides with the right-hand side of (18) if $\alpha = 0$, $\beta = 1$, $P_1(x) = 0$, and $P_2(x) = 1$. We observe that $\mu = \alpha + \mathrm{i}\beta = \mathrm{i}$ is not the root ($= 2$) of the characteristic polynomial. This means that $n = 0$ according to Section 7.2. The maximum degree of P_1 and P_2 is equal to 0, therefore we have to assume that[9] Q_1 and Q_2 are constants. Referring to (20), we can say that

$$y_{\mathrm{P}_2} = a\cos x + b\sin x$$

for some real numbers a and b that have to be identified. To this end we infer

$$y'_{\mathrm{P}_2} = -a\sin x + b\cos x \quad \text{and} \quad y''_{\mathrm{P}_2} = -a\cos x - b\sin x$$

and substitute into $y'' - 4y' + 4y = \sin x$. We obtain

$$(-a\cos x - b\sin x) - 4(-a\sin x + b\cos x) + 4(a\cos x + b\sin x) = \sin x.$$

This equation is simplified to

$$(3a - 4b)\cos x + (4a + 3b)\sin x = \sin x.$$

By comparing the left- and right-hand side, we arrive at the system

$$\begin{aligned} 3a - 4b &= 0, \\ 4a + 3b &= 1 \end{aligned}$$

whose solution is $a = 4/25$ and $b = 3/25$. As a consequence (see (26) for y_{P}),

$$\begin{aligned} y_{\mathrm{P}_2} &= \frac{4}{25}\cos x + \frac{3}{25}\sin x, \\ y &= y_{\mathrm{P}} + y_{\mathrm{H}} = 6y_{\mathrm{P}_1} - 50y_{\mathrm{P}_2} + y_{\mathrm{H}} \\ &= \mathrm{e}^{2x}\left(9x^2 + 2x^3\right) - 8\cos x - 6\sin x + \mathrm{e}^{2x}\left(c_1 + c_2 x\right) \\ &= \mathrm{e}^{2x}\left(c_1 + c_2 x + 9x^2 + 2x^3\right) - 8\cos x - 6\sin x. \end{aligned}$$

Other solved examples are available at

`http://mat.fsv.cvut.cz/BAKALARI/ma3si/files/dif_rov.pdf`

Although this collection is in Czech, the language of mathematics is international. That is, at least the definition of problems and the solution expressions are understandable for non-Czech speakers. The relevant methods have been explained in Section 7.

[9] The polynomial Q_1 can be **non-zero** even if $P_1 = 0$!!!

8 Solvability of 1D Boundary Value Problems

Let us recall the eigenpairs related to a few basic boundary value problems for the differential equation $u'' + \lambda u = 0$.

1) Boundary conditions $u(a) = 0 = u(b)$.
The set of eigenvalues and the associated eigenfunctions:

$$\lambda_k = \left(\frac{k\pi}{b-a}\right)^2, \quad u_k(x) = \sin\frac{k\pi(x-a)}{b-a}, \quad \text{where } k = 1, 2, \dots \tag{27}$$

2) Boundary conditions $u(a) = 0 = u'(b)$.
The set of eigenvalues and the associated eigenfunctions:

$$\lambda_k = \left(\frac{(k-1/2)\pi}{b-a}\right)^2, \quad u_k(x) = \sin\frac{(k-1/2)\pi(x-a)}{b-a}, \quad \text{where } k = 1, 2, \dots \tag{28}$$

3) Boundary conditions $u'(a) = 0 = u(b)$.
The set of eigenvalues and the associated eigenfunctions:

$$\lambda_k = \left(\frac{(k-1/2)\pi}{b-a}\right)^2, \quad u_k(x) = \cos\frac{(k-1/2)\pi(x-a)}{b-a}, \quad \text{where } k = 1, 2, \dots \tag{29}$$

Problem 8.1: How many solutions are there to the boundary value problem (BVP)

$$u'' - \pi u = \ln(\pi + x), \quad u(0) = 0,\ u(\pi) = 0?$$

Solution: Since $-\pi$ is not the eigenvalue of the BVP (see (27) or realize that the BVP has only positive eigenvalues), the BVP has exactly one solution.

Problem 8.2: Find out how the solvability of the BVP

$$u'' + \lambda u = \sin(-x) + 5\sin 3x - 8\sin 7x + \frac{1}{2}\sin 22x,$$
$$u(0) = u(\pi) = 0$$

depends on $\lambda \in \mathbb{R}$:

Solution: The eigenvalues of the above BVP form the set $V \equiv \{k^2 : \ k = 1, 2, 3, \dots\}$; each eigenfunction associated with $\lambda = k^2$ is a multiple of $u_k = \sin kx$. We observe that the right-hand side of the equation comprises[10] eigenfunctions, that is, $f \equiv -u_1 + 5u_3 - 8u_7 + \frac{1}{2}u_{22}$, see (27).

On the basis of the Fredholm theorem, we can make the following conclusion:

1. If $\lambda \in \mathbb{R} \setminus V$, then the BVP has a unique solution.
2. If $k = 1, 3, 7, 22$, then $(u_k, f) \neq 0$. As a consequence, the BVP does not have any solution for $\lambda = 1, 9, 49, 22^2$ (it holds that $22^2 = 484$).
3. If $\lambda \in V \setminus \{1, 9, 49, 22^2\}$, then the BVP has infinitely many solutions.

[10] Let us note that $\sin(-x) = -\sin(x)$.

Problem 8.3: Let the following BVP be given

$$u'' + u = f,$$
$$u(-\pi/2) = 0 = u(\pi/2).$$

Discuss the solvability of the BVP if a) $f \equiv x^2$; b) $f \equiv x^3$.

Solution: The eigenvalues are (see (27)) given by $k^2 \frac{\pi^2}{(b-a)^2}$, where $k = 1, 2, 3, \dots$ In the BVP, $\lambda = 1$ is the first eigenvalue ($k = 1$) and $u_1(x) = \sin \frac{k\pi(x-a)}{b-a}\Big|_{k=1} = \sin\left(x + \frac{\pi}{2}\right) = \cos x$ is the associated eigenfunction.

We observe that u_1 is *even* (i.e., $u_1(-x) = u_1(x)$) on the interval $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ and, moreover, positive in the interval $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. This observation will help us to avoid the detailed calculation of the inner products (i.e., definite integrals) below.

a) It holds that $(u_1, f) \neq 0$ because $u_1(x)x^2$ is positive in $\left(-\frac{\pi}{2}, 0\right) \cup \left(0, \frac{\pi}{2}\right)$. The BVP has no solution.

b) It holds that $(u_1, f) = 0$ because $u_1(x)x^3$ is an odd function (i.e., $u_1(-x)(-x)^3 = -u_1(x)x^3$) in the interval $(-\pi/2, \pi/2)$. The BVP has infinitely many solutions.

Problem 8.4: Find out how the solvability of the BVP

$$u'' + \lambda u = 2\sin x + \sin(-3x) + 3\sin(-5x) - \frac{1}{3}\sin 7x + \frac{1}{2}\sin 3x,$$
$$u(0) = u'(\pi/2) = 0$$

depends on the parameter $\lambda \in \mathbb{R}$.

Solution: The eigenvalues form the set (see (28)) $V \equiv \{(2k-1)^2 : \ k = 1, 2, \dots\}$ and $u_k = \sin\big((2k-1)x\big)$ is the eigenfunction associated with the eigenvalue $\lambda = (2k-1)^2$.

We easily discover that the right-hand side of the equation is a linear combination of eigenfunctions. In detail, $f \equiv 2u_1 - u_2 - 3u_3 - \frac{1}{3}u_4 + \frac{1}{2}u_2$.

1. If $\lambda \notin V$, then the BVP has a unique solution.
2. If $k = 1, 2, 3, 4$, then $(u_k, f) \neq 0$. As a consequence, the BVP has no solution if $\lambda = 1, 9, 25, 49$.
3. If $\lambda \in V \setminus \{1, 9, 25, 49\}$, then the BVP has infinitely many solutions.

Problem 8.5: Find out how the solvability of the BVP

$$u'' + \lambda u = \cos(-2x) + 3\cos 6x - 2\cos 14x,$$
$$u'(0) = u(\pi/4) = 0$$

depends on the parameter $\lambda \in \mathbb{R}$.

Solution: The eigenvalues form the set (see (29)) $V \equiv \{4(2k-1)^2 : \ k = 1, 2, \dots\}$ and $u_k = \cos\big(2(2k-1)x\big)$ is the eigenfunction associated with the eigenvalue $\lambda = 4(2k-1)^2$. The right-hand side of the equation is a linear combination of eigenfunctions (let us recall $\cos(-\alpha) = \cos(\alpha)$) $f \equiv u_1 + 3u_2 - 2u_4$.

1. If $\lambda \in \mathbb{R} \setminus V$, than the BVP has a unique solution.
2. If $k = 1, 2, 4$, then $(u_k, f) \neq 0$. As a consequence, the BVP has no solution for $\lambda = 4, 36, 196$.
3. If $\lambda \in V \setminus \{4, 36, 196\}$, then the associated eigenfunctions are orthogonal to f. As a consequence, the BVP has infinitely many solutions.

9 Positive Definiteness of Operators

If $u \in C^1([a,b])$ and $u(a) = 0$ or $u(b) = 0$, then the Friedrichs inequality

$$\int_a^b u'^2(x)\,\mathrm{d}x \geq \frac{2}{(b-a)^2}\int_a^b u^2(x)\,\mathrm{d}x. \tag{30}$$

can be beneficial in the proof of the positive definiteness of an operator.

Problem 9.1: The following boundary value problem is given

$$-(x^2+x-2)u'' - (2x+1)u' + 7\sin(8x^2)\,u = \ln(1+x),$$
$$u(0) = 0, \quad u'(1/2) = 0.$$

Find an operator A that is both symmetric on its domain of definition $\mathcal{D}(A)$ and such that $\forall u \in \mathcal{D}(A)$ $(Au, u) \geq c\|u\|^2_{L^2(0,1/2)}$; specify the positive and u-independent constant c.

Solution: First, we rewrite the equation in to the divergence form, i.e.,

$$-((x^2+x-2)u')' + 7\sin(8x^2)\,u = \ln(1+x).$$

Since the function $(x^2 + x - 2)$ is negative on $[0, 1/2]$, we multiply the equation by -1 to get (as we will see) a positive definiteness of the associated operator defined as $Au \stackrel{\text{def}}{=} -((2-x^2-x)u')' - 7\sin(8x^2)\,u$ where
$u \in \mathcal{D}(A) = \{\eta \in C^2([0,1/2]) :\ \eta(0) = 0,\ \eta'(1/2) = 0\}$.

Let us show that A is symmetric: $\forall u, v \in \mathcal{D}(A)$

$$\begin{aligned}
(Au, v) &= \int_0^{1/2} \Big(-\big((2-x^2-x)u'(x)\big)'\Big)\,v(x)\,\mathrm{d}x \\
&= -\big[(2-x^2-x)u'(x)v(x)\big]_{x=0}^{x=1/2} + \int_0^{1/2}(2-x^2-x)u'v'\,\mathrm{d}x \\
&\overset{v(0)=0=u'(1/2)}{=} \int_0^{1/2}(2-x^2-x)v'u'\,\mathrm{d}x, \\
(u, Av) &= \int_0^{1/2} u(x)\Big(-\big((2-x^2-x)v'(x)\big)'\Big)\,\mathrm{d}x \\
&= -\big[(2-x^2-x)v'(x)u(x)\big]_{x=0}^{x=1/2} + \int_0^{1/2}(2-x^2-x)u'v'\,\mathrm{d}x \\
&\overset{u(0)=0=v'(1/2)}{=} \int_0^{1/2}(2-x^2-x)v'u'\,\mathrm{d}x,
\end{aligned}$$

that is, $(Au, v) = (u, Av)$. We applied the integration by parts and the boundary conditions included in the definition of $\mathcal{D}(A)$.

To show that the operator A is positive definite, we apply the Friedrichs inequality (30)

$$\begin{aligned}(Au, u) &= \int_0^{1/2} (2 - x^2 - x)u'^2 \,\mathrm{d}x + \int_0^{1/2} (-7)\sin(8x^2)\, u^2 \,\mathrm{d}x \\ &\geq \int_0^{1/2} \left(\min_{t\in[0,1/2]} (2 - t^2 - t) \right) u'^2 \,\mathrm{d}x + \int_0^{1/2} \left(\min_{t\in[0,1/2]} \left(-7\sin(8t^2)\right) \right) u^2 \,\mathrm{d}x \\ &\geq \frac{5}{4} \int_0^{1/2} u'^2 \,\mathrm{d}x - 7\int_0^{1/2} u^2 \,\mathrm{d}x \overset{\text{Fr. ineq.}}{\geq} \frac{5}{4} 8 \int_0^{1/2} u^2 \,\mathrm{d}x - 7\int_0^{1/2} u^2 \,\mathrm{d}x \\ &= 3\|u\|^2_{L^2(0,1/2)}.\end{aligned}$$

Thus $c = 3$ or $c \in (0, 3]$.

The operator equation: Find $u \in \mathcal{D}(A)$ such that $Au = -\ln(1 + x)$.

10 Positive Definiteness of Operators and the Ritz Method

Problem 10.1: Define a symmetric positive definite operator A (and also $\mathcal{D}(A)$, its domain of definition)[11] that represents the following boundary value problem

$$\begin{gathered}(x^2 + 4)u'' + 2xu' = -x^2, \\ u(-1) = 0, \quad u(1) = 0.\end{gathered}$$

Write the problem as an operator equation.

Use $a \in \mathbb{R}$ and the Ritz method to identify $w = a(x + 1)(x - 1)$, an approximate solution to the above problem. To identify $a \in \mathbb{R}$, use the Ritz method.

Solution: In the divergence form of the equation

$$-\left(-(x^2 + 4)u'\right)' = -x^2,$$

the function $-(x^2 + 4)$ is negative on $[-1, 1]$. To obtain a positive definite operator, we have to multiply the entire equation by -1. We arrive at $Au \overset{\text{def}}{=} -\left((x^2 + 4)u'\right)'$ and $\mathcal{D}(A) = \{\eta \in C^2([-1, 1]) : \ \eta(-1) = 0 = \eta(1)\}$.

Operator equation: Find $u \in \mathcal{D}(A)$ such that

$$Au = x^2.$$

[11] Show that the operator is symmetric on $\mathcal{D}(A)$ and that $\forall v \in \mathcal{D}(A)$ $(Av, v) \geq c\|v\|^2_{L^2(-1,1)}$, where $c > 0$ is a v-indpendent constant.

The operator A is symmetric because $\forall u, v \in \mathcal{D}(A)$

$$
\begin{aligned}
(Au, v) &= \int_{-1}^{1} \Big(- \big((x^2+4)u'(x)\big)' \Big) v(x)\,\mathrm{d}x \\
&= - \big[(x^2+4)u'(x)v(x)\big]_{x=-1}^{x=1} + \int_{-1}^{1} (x^2+4)u'v'\,\mathrm{d}x \overset{v(-1)=0=v(1)}{=} \int_{-1}^{1} (x^2+4)v'u'\,\mathrm{d}x, \\
(u, Av) &= \int_{-1}^{1} u(x) \Big(- \big((x^2+4)v'(x)\big)' \Big)\,\mathrm{d}x \\
&= - \big[(x^2+4)v'(x)u(x)\big]_{x=-1}^{x=1} + \int_{-1}^{1} (x^2+4)u'v'\,\mathrm{d}x \overset{u(-1)=0=u(1)}{=} \int_{-1}^{1} (x^2+4)v'u'\,\mathrm{d}x,
\end{aligned}
$$

that is, $(Au, v) = (u, Av)$. We applied the integration by parts and the boundary conditions included in $\mathcal{D}(A)$.

The operator A is positive definite because $\forall v \in \mathcal{D}(A)$

$$
\begin{aligned}
(Av, v) &= \int_{-1}^{1} (x^2+4)v'^2\,\mathrm{d}x \geq \int_{-1}^{1} \left(\min_{t\in[-1,1]} (t^2+4) \right) v'^2\,\mathrm{d}x \\
&\geq 4 \int_{-1}^{1} v'^2\,\mathrm{d}x \overset{(30)}{\geq} 4\frac{2}{2^2} \|v\|^2_{L^2(-1,1)},
\end{aligned}
$$

that is, $c = 2$ or $c \in (0, 2]$.

In the one-dimensional subspace generated by $a\omega$, where $a \in \mathbb{R}$ and $\omega = (x+1)(x-1)$, the minimum of the energy functional is attained determined $a = \dfrac{(f, \omega)}{(A\omega, \omega)}$, where $f = x^2$ and $\omega = x^2 - 1$.

Let us calculate

$$
(f, \omega) = \int_{-1}^{1} x^2(x^2-1)\,\mathrm{d}x = \int_{-1}^{1} (x^4 - x^2)\,\mathrm{d}x = \left[\frac{x^5}{5} - \frac{x^3}{3} \right]_{-1}^{1} = \frac{2}{5} - \frac{2}{3} = -\frac{4}{15}.
$$

$$
\begin{aligned}
(A\omega, \omega) &= \int_{-1}^{1} (x^2+4)(\omega')^2\,\mathrm{d}x = \int_{-1}^{1} (x^2+4)4x^2\,\mathrm{d}x = \int_{-1}^{1} (4x^4 + 16x^2)\,\mathrm{d}x \\
&= \left[\frac{4}{5}x^5 + \frac{16}{3}x^3 \right]_{-1}^{1} = \frac{8}{5} + \frac{32}{3} = \frac{184}{15}.
\end{aligned}
$$

As a consequence, $a = -1/46$ and $w(x) = (1 - x^2)/46$, $x \in [-1, 1]$.

Problem 10.2: Define an operator A (and its domain of definition $\mathcal{D}(A)$) associated with the BVP

$$
\begin{aligned}
-(1/2 + x)u'' - u' &= x, \\
u(0) = 0, \quad u(1) &= 0.
\end{aligned}
$$

Show that the operator is symmetric and positive definite. By applying the Ritz method, find an approximate solution $w = ax(x-1)$, where $a \in \mathbb{R}$ is to be determined.

Solution: Let us write the equation in the divergence form, i.e.,

$$
-\left((1/2 + x)u'\right)' = x.
$$

We observe that $1/2 + x$ is a positive function on $[0, 1]$; there is no need to multiply the equation by -1. We are ready to define the operator $Au \stackrel{\text{def}}{=} -\left((1/2 + x)u'\right)'$ and its domain of definition $\mathcal{D}(A) = \{\eta \in C^2([0, 1]) : \ \eta(0) = 0 = \eta(1)\}$.

A is symmetric and positive definite:[12]

$$\begin{aligned}
(Au, v) &= -\int_0^1 \left((1/2 + x)u'\right)' v \, \mathrm{d}x = \int_0^1 (1/2 + x)u'v' \, \mathrm{d}x \quad \forall u, v \in \mathcal{D}(A), \\
(u, Av) &= \int_0^1 (1/2 + x)u'v' \, \mathrm{d}x \quad \forall u, v \in \mathcal{D}(A), \\
(Av, v) &= \int_0^1 (1/2 + x)v'^2 \, \mathrm{d}x \geq \int_0^1 \left(\min_{t\in[0,1]} (1/2 + t)\right) v'^2 \, \mathrm{d}x \\
&= \frac{1}{2}\int_0^1 v'^2 \, \mathrm{d}x \geq \frac{1}{2}\frac{2}{1}\int_0^1 v^2 \, \mathrm{d}x \quad \forall v \in \mathcal{D}(A).
\end{aligned}$$

The BVP a the operator equation: find $u \in \mathcal{D}(A)$ such that $Au = x$.

The minimization of the functional of energy leads to $a = \dfrac{(f, \omega)}{(A\omega, \omega)}$, where $f = x$ and $\omega = x(x - 1)$. In particular,

$$(f, \omega) = \int_0^1 (x^3 - x^2) \, \mathrm{d}x = \left[\frac{x^4}{4} - \frac{x^3}{3}\right]_0^1 = -\frac{1}{12}.$$

$$\begin{aligned}
(A\omega, \omega) &= \int_0^1 (1/2 + x) \left((x(x - 1))'\right)^2 \, \mathrm{d}x \\
&= \int_0^1 (1/2 + x)(2x - 1)^2 \, \mathrm{d}x = \int_0^1 (1/2 + x)(4x^2 - 4x + 1) \, \mathrm{d}x \\
&= \int_0^1 (4x^3 - 2x^2 - x + 1/2) \, \mathrm{d}x = \left[x^4 - \frac{2x^3}{3} - \frac{x^2}{2} + \frac{x}{2}\right]_0^1 = 1/3.
\end{aligned}$$

To conclude, we have $a = -1/4$ and the approximate solution $w = -x(x - 1)/4 = (x - x^2)/4$.

Problem 10.3: The same boundary value problem as in Example 10.2. The approximate solution is sought as a linear combination of two basis functions v_1 and v_2, that is,

$$w = c_1 v_1 + c_2 v_2, \text{ where } v_1 = x(x - 1), \ v_2 = x^2(x - 1) \text{ and } c_1, c_2 \in \mathbb{R}.$$

Infer the system of linear equations for the unknowns c_1 and c_2.

Solution: See Example 10.2 for the definition of the operator as well as for the proof of its symmetry and positive definiteness. The road to the approximate solution is partly different, however.

We have to minimize

$$F(u) = (Au, u) - 2(f, u) = (u, u)_A - 2(f, u)$$

[12] The proof is based on the integration ba parts and on the boundary conditions included in $\mathcal{D}(A)$. We refer to Example 10.1 for the details of the technique.

(the functional of the energy) on the subspace generated by the two given basis functions. That is,

$$\begin{aligned}&\min_{c_1,c_2\in\mathbb{R}} F(c_1v_1+c_2v_2)\\&=\min_{c_1,c_2\in\mathbb{R}}\left(c_1^2(v_1,v_1)_A+2c_1c_2(v_1,v_2)_A+c_2^2(v_2,v_2)_A-2c_1(f,v_1)-2c_2(f,v_2)\right).\end{aligned}$$

The point of the minimum is characterized by

$$\frac{\partial F}{\partial c_1}=0,\quad \frac{\partial F}{\partial c_2}=0$$

that leads to the system of equation in a general form

$$\begin{aligned}(v_1,v_1)_Ac_1+(v_1,v_2)_Ac_2&=(f,v_1),\\(v_1,v_2)_Ac_1+(v_2,v_2)_Ac_2&=(f,v_2).\end{aligned}$$

The values

$$(f,v_1)=-\frac{1}{12},\quad (v_1,v_1)_A=\frac{1}{3}$$

have already been calculated in Example 10.2. It remains to deal with

$$\begin{aligned}(f,v_2)&=\int_0^1 x^3(x-1)\,\mathrm{d}x=-\frac{1}{20},\\(v_1,v_2)_A&=\int_0^1\left(\frac{1}{2}+x\right)\left(x^2-x\right)'\left(x^3-x^2\right)'\,\mathrm{d}x=\frac{1}{5},\\(v_2,v_2)_A&=\int_0^1\left(\frac{1}{2}+x\right)\left(x^3-x^2\right)'\left(x^3-x^2\right)'\,\mathrm{d}x=\frac{1}{6}.\end{aligned}$$

The linear system

$$\begin{aligned}\frac{1}{3}c_1+\frac{1}{5}c_2&=-\frac{1}{12},\\\frac{1}{5}c_1+\frac{1}{6}c_2&=-\frac{1}{20}.\end{aligned}$$

finishes the example.

Problem 10.4: Use the Ritz method to find an approximate solution of

$$\begin{aligned}-(x+2)u''-u'+x^2u&=1,\\u(-1)=1,\quad u(1)&=-5.\end{aligned}$$

Search for the approximate solution in the form $w=(x+1)(1-x)$, where $x\in[-1,1]$.[13]

[13] In the Ritz method, the integration can by simplified by taking into account that the integral of an odd function is equal to zero and that the integral of an even function on the interval $[-1,1]$ is equal to the value of the integral over $[0,1]$ multiplied by two.

Solution: The solution u is searched for as the sum of two functions, i.e., $u = \hat{u} + \phi$, where $\phi = -3x - 2$ complies with the boundary conditions and $\hat{u}$ has homogeneous Dirichlet boundary conditions. Let us substitute $u = \hat{u} + \phi$ into the equation

$$\begin{aligned} -(x+2)(\hat{u}+\phi)'' - (\hat{u}+\phi)' + x^2(\hat{u}+\phi) &= 1, \\ -(x+2)\hat{u}'' - \hat{u}' + x^2\hat{u} &= 1 + (x+2)\phi'' + \phi' - x^2\phi \\ -(x+2)\hat{u}'' - \hat{u}' + x^2\hat{u} &= -2 + 2x^2 + 3x^3, \\ -((x+2)\hat{u}')' + x^2\hat{u} &= 3x^3 + 2x^2 - 2, \\ \hat{u}(-1) = 0, \quad \hat{u}(1) &= 0. \end{aligned}$$

The operator $A\hat{u} \stackrel{\text{def}}{=} -((x+2)\hat{u}')' + x^2\hat{u}$ is symmetric and positive definite on $\mathcal{D}(A) = \{v \in C^2([-1,1]) : \ v(-1) = 0 = v(1)\}$.

The approximate solution of $A\hat{u} = f$, where $\hat{u} \in \mathcal{D}(A)$, is determined by $a = (f,w)/(Aw,w)$.

$$\begin{aligned} (f,w) &= \int_{-1}^{1} (3x^3 + 2x^2 - 2)(1-x^2)\,\mathrm{d}x = \int_{-1}^{1} (-3x^5 - 2x^4 + 3x^3 + 4x^2 - 2)\,\mathrm{d}x \\ &= 4\int_0^1 (-x^4 + 2x^2 - 1)\,\mathrm{d}x = 4\left[-\frac{x^5}{5} + \frac{2x^3}{3}\right]_0^1 - 4 \\ &= 4\left(-\frac{1}{5} + \frac{2}{3} - 1\right) = -\frac{32}{15}. \end{aligned}$$

$$\begin{aligned} (Aw,w) &= \int_{-1}^{1} \left(4(x+2)x^2 + x^2(1-x^2)^2\right)\,\mathrm{d}x = \int_{-1}^{1} \left(x^6 - 2x^4 + 4x^3 + 9x^2\right)\,\mathrm{d}x \\ &= \int_{-1}^{1} \left(x^6 - 2x^4 + 9x^2\right)\,\mathrm{d}x = 2\int_0^1 \left(x^6 - 2x^4 + 9x^2\right)\,\mathrm{d}x \\ &= 2\left[\frac{x^7}{7} - \frac{2x^5}{5} + \frac{9x^3}{3}\right]_0^1 = 2\left(\frac{1}{7} - \frac{2}{5} + 3\right) = \frac{192}{35}. \end{aligned}$$

We obtain $a = \dfrac{-\frac{32}{15}}{\frac{192}{35}} = -\dfrac{7}{18}$, so that

$$u_{\text{Ritz}} = -\frac{7}{18}w + \phi = -\frac{7}{18}(1-x^2) - 3x - 2 = \frac{7}{18}x^2 - 3x - \frac{43}{18}.$$

11 Finite Element Method

Problem 11.1:[14] Consider the boundary value problem

$$xu'' + u' = -1, \tag{31}$$

$$u(1) = u(4) = 0. \tag{32}$$

[14]This example originates from the example kindly provided by Dr. Jaroslav Novotný, a colleague of mine.

a) Properly define an associated differential operator A and its domain of definition.
b) Show that A is a linear operator.
c) Show that A is a symmetric operator.
d) Show that A is a positive definite operator.
e) By applying the finite element (FE) method, approximately solve the boundary value problem on a two-dimensional space of continuous piecewise linear functions defined on a uniform mesh. That is, assemble and solve the relevant system of equations.
f) Evaluate the approximate solution at $x = \frac{3}{2}$ and graph the approximate solution on the interval $[1, 4]$.

Solution: **a)** By rewriting (31) into the divergence form

$$-(-xu')' = -1,$$

and realizing that $-x$ is a negative function on the interval $[1, 4]$[15], we conclude that (31) should be multiplied by -1, that is,

$$-(xu')' = 1. \tag{33}$$

The operator $Au \stackrel{\text{def}}{=} -(xu')'$ and its domain of definition
$\mathcal{D}(A) = \{u \in C^2([1, 4]) : \ u(1) = u(4) = 0\}$.[16]
Operator equation: Find $u \in \mathcal{D}(A)$ such that $Au = 1$.

b) We observe that for all $v \in \mathcal{D}(A)$ and for all $\alpha \in \mathbb{R}$ it holds that

$$A(\alpha v)(x) = -(x\alpha v'(x))' = \alpha(-(xv'(x))') = \alpha A(v)(x).$$

Next, for all $v, w \in \mathcal{D}(A)$ we have

$$A(v + w)(x) = -(x(v(x) + w(x))')' = -(xv'(x))' - (xw'(x))' = A(v)(x) + A(w)(x).$$

We conclude that the operator A is linear on $\mathcal{D}(A)$.

c) Symmetry: $\forall v, w \in \mathcal{D}(A)$

$$\begin{aligned}(Av, w) = \int_1^4 \big(xv'(x)\big)' w(x)\,\mathrm{d}x &= -\left[xv'(x)w(x)\right]_{x=1}^{x=4} + \int_1^4 xv'(x)w'(x)\,\mathrm{d}x \\ &\stackrel{(32)}{=} \int_1^4 xv'(x)w'(x)\,\mathrm{d}x \stackrel{\text{def}}{=} (v, w)_A;\end{aligned}$$

$$\begin{aligned}(v, Aw) = \int_1^4 \big(xw'(x)\big)' v(x)\,\mathrm{d}x &= -\left[xw'(x)v(x)\right]_{x=1}^{x=4} + \int_1^4 xw'(x)v'(x)\,\mathrm{d}x \\ &\stackrel{(32)}{=} \int_1^4 xv'(x)w'(x)\,\mathrm{d}x.\end{aligned}$$

We conclude[17] that $\forall v, w \in \mathcal{D}(A) \quad (Av, w) = (v, Aw)$; the operator A is symmetric $\mathcal{D}(A)$.

[15] The ends of the interval are determined by (32).
[16] Or, more exactly, $\mathcal{D}(A) = \{u \in C^2((1, 4)) \cap C([1, 4]) : \ u(1) = u(4) = 0\}$.
[17] It is not difficult to see that any boundary value problem in the divergence form and with the boundary conditions given by $u(a) = 0 = u(b)$, or $u'(a) = 0 = u(b)$, or $u(a) = 0 = u'(b)$, or $u'(a) = 0 = u'(b)$ is symmetric.

d) Positive definitenes: $\forall w \in \mathcal{D}(A)$

$$\begin{aligned}(Aw, w) &= \int_1^4 x\big(w'(x)\big)^2 \,\mathrm{d}x \geq \int_1^4 \left(\min_{t\in[1,4]} t\right) \big(w'(x)\big)^2 \,\mathrm{d}x \\ &= \int_1^4 \big(w'(x)\big)^2 \,\mathrm{d}x \overset{(30)}{\geq} \frac{2}{(4-1)^2} \int_1^4 w^2(x)\,\mathrm{d}x = \frac{2}{9}\int_1^4 w^2(x)\,\mathrm{d}x;\end{aligned}$$

the operator A is positive definite on $\mathcal{D}(A)$.

e) The two-dimensional space of functions is formed by linear combinations of two basis functions. These are continuous and piecewise linear, i.e., "hat" functions associated with two inner nodes of the FE mesh. The mesh is uniform (the nodes are equaly spaced), which means that the mesh is formed by the nodes $x_0 = 1$, $x_1 = 2$, $x_2 = 3$, and $x_3 = 4$. Let the basis functions be denoted by u_1 and u_2. It holds[18] $u_i(x_j) = \delta_{ij}$, where $i = 1, 2$ and $j = 0, 1, 2, 3$.

An approximate solution u_{FEM} is sought as a linear combination of u_1 and u_2. That is, $u_{\mathrm{FEM}}(x) = \alpha_1 u_1(x) + \alpha_2 u_2(x)$, where α_1, α_2 are real numbers that will be calculated.

Remarks: 1) The basis function u_1, u_2 comply with the boundary conditions (32).
2) The approximate solution u_{FEM} has to minimize the energy functional $F(u) \overset{\mathrm{def}}{=} (Au, u) - 2(f, u)$, where $f = 1$ (see (31)) over all linear combinations of u_1 and u_2. In other words, the function g defined as

$$g(\alpha_1, \alpha_2) \overset{\mathrm{def}}{=} F(\alpha_1 u_1 + \alpha_2 u_2)$$

attains its minimum at such $(\widehat{\alpha}_1, \widehat{\alpha}_2)$ that define u_{FEM} minimizing F.

At the minimum point, the partial derivatives of g vanish, which is the feature that leads to the following system of linear algebraic equations for unknown $\alpha_1, \alpha_2 \in \mathbb{R}$:

$$\begin{aligned}(u_1, u_1)_A \alpha_1 + (u_1, u_2)_A \alpha_2 &= (f, u_1), \\ (u_1, u_2)_A \alpha_1 + (u_2, u_2)_A \alpha_2 &= (f, u_2),\end{aligned}$$

where the symmetry $(u_1, u_2)_A = (u_2, u_1)_A$ has been applied. To set up this system of linear algebraic equations, its coefficients $(u_i, u_j)_A$ and right-hand side values (f, u_i), where $i, j = 1, 2$, have to be calculated. □

[18] Kronecker delta $\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$.

Let us calculate[19]

$$\begin{aligned}
(u_1,u_1)_A &= \int_1^4 xu_1'(x)u_1'(x)\,\mathrm{d}x = \int_1^2 xu_1'(x)u_1'(x)\,\mathrm{d}x + \int_2^3 xu_2'(x)u_2'(x)\,\mathrm{d}x \\
&= \int_1^2 x1^2\,\mathrm{d}x + \int_2^3 x(-1)^2\,\mathrm{d}x = \left[\frac{x^2}{2}\right]_1^2 + \left[\frac{x^2}{2}\right]_2^3 = \frac{3}{2}+\frac{5}{2} = 4, \\
(u_1,u_2)_A &= \int_1^4 xu_1'(x)u_2'(x)\,\mathrm{d}x = \int_2^3 xu_1'(x)u_2'(x)\,\mathrm{d}x \\
&= \int_2^3 x(-1)1\,\mathrm{d}x = \left[\frac{-x^2}{2}\right]_2^3 = -\frac{5}{2}, \\
(u_2,u_1)_A &= (u_1,u_2)_A = -\frac{5}{2}, \\
(u_2,u_2)_A &= \int_1^4 xu_2'(x)u_2'(x)\,\mathrm{d}x = \int_2^3 xu_2'(x)u_2'(x)\,\mathrm{d}x + \int_3^4 xu_2'(x)u_2'(x)\,\mathrm{d}x \\
&= \int_2^3 x1^2\,\mathrm{d}x + \int_3^4 x(-1)^2\,\mathrm{d}x = \left[\frac{x^2}{2}\right]_2^3 + \left[\frac{x^2}{2}\right]_3^4 = \frac{5}{2}+\frac{7}{2} = 6.
\end{aligned}$$

We could infer the expressions defining u_1, u_2:

$$\begin{aligned}
u_1(x) &= \begin{cases} x-1 & \text{if } x\in[1,2], \\ -x+3 & \text{if } x\in(2,3], \\ 0 & \text{elsewhere,} \end{cases} \\
u_2(x) &= \begin{cases} x-2 & \text{if } x\in[2,3], \\ -x+4 & \text{if } x\in(3,4], \\ 0 & \text{elsewhere,} \end{cases}
\end{aligned}$$

but, thanks to the fact that f is constant, this is not necessary. The area under the graph of u_i, $i=1,2$, is equal to 1, and we obtain

$$\begin{aligned}
(f,u_1) &= \int_1^4 1u_1(x)\,\mathrm{d}x = 1, \\
(f,u_2) &= \int_1^4 1u_2(x)\,\mathrm{d}x = 1.
\end{aligned}$$

We are at the point of setting up the FE system of equations:

$$\begin{aligned}
4\alpha_1 - \frac{5}{2}\alpha_2 &= 1, \\
-\frac{5}{2}\alpha_1 + 6\alpha_2 &= 1.
\end{aligned}$$

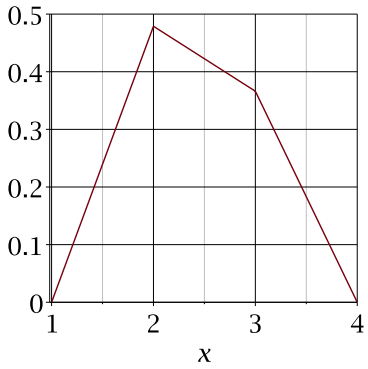
After some labor, we infer $\alpha_1 = \frac{34}{71}$ and $\alpha_2 = \frac{26}{71}$, that is,

$$u_{\mathrm{FEM}}(x) = \frac{34}{71}u_1(x) + \frac{26}{71}u_2(x).$$

[19] The derivative of u_1 and u_2 is equal to 0, 1, or -1 as can be seen from a sketch of the graphs of u_1 and u_2.

f) By substituting[20] $3/2$ for x, we arrive at $u_{\mathrm{FEM}}\left(\frac{3}{2}\right) = \frac{34}{71}\cdot\frac{1}{2} = \frac{17}{71}$.

Since the approximate solution u_{FEM} is a linear combination of continuos piecewise linear basis functions, u_{FEM} is also a continuous piecewise linear function. Moreover, $u_{\mathrm{FEM}}(2) = \alpha_1$ and $u_{\mathrm{FEM}}(3) = \alpha_2$ by virtue of $u_1(2) = 1 = u(3)$ and $u_1(3) = 0 = u(2)$.



Problem 11.2: A BVP is defined on the interval $[0, 2]$:

$$(x-5)u'' + u' = 6 \quad \text{with the boundary condition} \quad u(0) = u(2) = 0.$$

Apply the finite element method and infer $Mc = b$, the relevant system of linear algebraic equations. Use the mesh defined by the uniformly distributed nodes $x_0 = 0, x_1, x_2, x_3, x_4 = 2$ as well as the associated continuous piecewise linear basis functions v_i such that $v_i(x_j) = \delta_{ij}$.[21] Is the matrix M diagonally dominant?

Solution: The operator in the divergence form (no multiplication by -1!): $Au = -\left((5-x)u'\right)'$ with the domain of definition $\mathcal{D}(A) = \{\eta \in C^2([0,2]) : \ \eta(0) = \eta(2) = 0\}$. The operator is symmetric (due to the boundary conditions and and the divergence form) and positive definite ($5 - x > 0$ in $[0, 2]$, and the Friedrichs inequality can be applied).

The operator equation: Find $u \in \mathcal{D}(A)$ such that $Au = f$, where $f = 6$.

The system $Mc = b$ is determined by $M = (m_{ij})$, $m_{ij} = (v_i, v_j)_A$, and $b = (b_i)$, $b_i = (f, v_i)$, where $i, j = 1, 2, 3$.[22]

Since only differentiated basis functions appear in the energy inner product, it is necessary to know only the first derivatives. These are easy to calculate without having a formula defining the basis functions.[23] The derivative is equal to 0 outside the support

[20] Again, the formulae for u_1 and u_2 are not necessary. Indeed, as $3/2$ is the center of the interval $[1, 2]$, we see that $u_1(3/2) = 1/2$ and $u_2(3/2) = 0$.

[21] Kronecker delta $\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$.

[22] The inner product $(v_i, v_j)_A$ is inferred from the inner product (Av_i, v_j) by the integration by parts.

[23] For those who are interested, I list the expressions that define the basis functions on their respective supports:

$$v_1(x) = \begin{cases} 2x, & x \in [0,\ 1/2], \\ 2 - 2x, & x \in (1/2,\ 1], \end{cases} \quad v_2(x) = \begin{cases} -1 + 2x, & x \in [1/2,\ 1], \\ 3 - 2x, & x \in (1,\ 3/2], \end{cases} \quad v_3(x) = \begin{cases} -2 + 2x, & x \in [1,\ 3/2], \\ 4 - 2x, & x \in (3/2,\ 2]. \end{cases}$$

I emphasize that, in this particular problem, we do not need these formulae to evaluate the inner products.

of a basis function v_i, or 2 (where v_i increses), or -2 (where v_i decreases).

$$
\begin{aligned}
(v_1, v_1)_A &= \int_0^2 (5-x) v_1'(x) v_1'(x)\,\mathrm{d}x = \int_0^{1/2} (5-x) 2^2\,\mathrm{d}x + \int_{1/2}^1 (5-x) 2^2\,\mathrm{d}x \\
&= 10 - 4\left[\frac{x^2}{2}\right]_0^{1/2} + 10 - 4\left[\frac{x^2}{2}\right]_{1/2}^1 = 20 - \frac{1}{2} + -4\left[\frac{1}{2} - \frac{1}{8}\right] = 18, \\
(v_1, v_2)_A &= \int_0^2 (5-x) v_1'(x) v_2'(x)\,\mathrm{d}x = \int_{1/2}^1 (5-x) 2(-2)\,\mathrm{d}x \\
&= -10 + 4\left[\frac{x^2}{2}\right]_{1/2}^1 = -10 + 4\left(\frac{1}{2} - \frac{1}{8}\right) = -10 + \frac{3}{2} = -\frac{17}{2}, \\
(v_1, v_3)_A &= 0, \\
(v_2, v_1)_A &= (v_1, v_2)_A = -\frac{17}{2}, \\
(v_2, v_2)_A &= \int_0^2 (5-x) v_2'(x) v_2'(x)\,\mathrm{d}x = \int_{1/2}^1 (5-x) 2^2\,\mathrm{d}x + \int_1^{3/2} (5-x)(-2)^2\,\mathrm{d}x \\
&= 4\int_{1/2}^{3/2} (5-x)\,\mathrm{d}x = 20 - 4\left[\frac{x^2}{2}\right]_{1/2}^{3/2} = 20 - 4\left[\frac{9}{8} - \frac{1}{8}\right] = 16, \\
(v_2, v_3)_A &= \int_0^2 (5-x) v_2'(x) v_3'(x)\,\mathrm{d}x = \int_1^{3/2} (5-x) 2(-2)\,\mathrm{d}x \\
&= -10 + 4\left[\frac{x^2}{2}\right]_1^{3/2} = -10 + 4\left[\frac{9}{8} - \frac{1}{2}\right] = -10 + 4\frac{5}{8} = -\frac{15}{2}, \\
(v_3, v_1)_A &= (v_1, v_3)_A = 0, \\
(v_3, v_2)_A &= (v_2, v_3)_A = -\frac{15}{2}, \\
(v_3, v_3)_A &= \int_0^2 (5-x) v_3'(x) v_3'(x)\,\mathrm{d}x = \int_1^{3/2} (5-x) 2^2\,\mathrm{d}x + \int_{3/2}^2 (5-x)(-2)^2\,\mathrm{d}x \\
&= 4\int_1^2 (5-x)\,\mathrm{d}x = 20 - 4\left[\frac{x^2}{2}\right]_1^2 = 20 - 4\left[2 - \frac{1}{2}\right] = 14.
\end{aligned}
$$

The fact that f is a constant substantially simplifies the evaluation of the other inner products. The integration reduces to the calculation of the area determined by the triangular graph of a basis function and the horizontal axis. The area is equal to $1/2$ for each v_i.

$$
\begin{aligned}
(f, v_1) &= \int_0^2 6 v_1(x)\,\mathrm{d}x = 3, \\
(f, v_2) &= 3, \\
(f, v_3) &= 3.
\end{aligned}
$$

The augmented matrix of the system $Mc = b$:

$$
\left(\begin{array}{rrr|l}
18 & -\frac{17}{2} & 0 & 3 \\
-\frac{17}{2} & 16 & -\frac{15}{2} & 3 \\
0 & -\frac{15}{2} & 14 & 3
\end{array}\right).
$$

Although $16 = |-17/2|+|-15/2|$, the matrix M is diagonally dominant. It is sufficient to choose $h_1 = h_3 = 1$ and $h_2 = 3/2$, for instance, in the inequality that characterizes the diagonally dominant matrices.

Problem 11.3: Apply the finite element method to the following BVP

$$\begin{aligned} -(x^3+1)u'' - 3x^2u' &= 8 + 9x^2, \\ u(1) = 2, \quad u(3) &= -4, \end{aligned}$$

and infer the matrix M from the system $Mc = b$ originating from the method. Use continuous piecewise linear functions ψ_i defined at the mesh nodes $x_0 = 1$, $x_1 = 4/3$, $x_2 = 2$, $x_3 = 5/2$, and $x_4 = 3$ and such that $\psi_i(x_j) = \delta_{ij}$, where δ_{ij} is the Kronecker delta and $i = 1, 2, 3$, $j = 0, 1, \ldots, 4$.

Solution: One of possible approaches to the problem is based on the transformation to a BVP with homogeneous boundary conditions. That is, the solution u is searched for in the form of the sum of two functions. In detail, $u = u_0 + \phi$, where $u_0(1) = 0 = u_0(3)$ and $\phi(1) = 2$, $\phi(3) = -4$. Let us choose $\phi(x) = -3x + 5$ an substitute $u = u_0 + \phi$ into the equation:

$$\begin{aligned} -(x^3+1)(u_0+\phi)'' - 3x^2(u_0+\phi)' &= 8 + 9x^2, \\ -(x^3+1)u_0'' - 3x^2u_0' + 3x^2\phi' &= 8 + 9x^2, \\ -(x^3+1)u_0'' - 3x^2u_0' &= 8 + 9x^2 - 9x^2, \\ -\big((x^3+1)u_0'\big)' &= 8. \end{aligned}$$

The operator $Au_0 \stackrel{\text{def}}{=} -\left((x^3+1)u_0'\right)'$ is symmetric and positive definite on its domain $\mathcal{D}(A) = \{\eta \in C^2([0,2]) : \ \eta(1) = \eta(3) = 0\}$. We apply the FE method on the BVP: Find $u_0 \in \mathcal{D}(A)$ such that

$$Au_0 = 8.$$

To calculate the elements of M, we need the derivatives of ψ_i. If $x \in [x_0, x_1]$, then the derivative of ψ_1 is equal to $(1-0)/(x_1-x_0) = 3$. If $x \in [x_1, x_2]$, then the derivative of ψ_1 is equal to $(0-1)/(x_2-x_1) = -3/2$ and the derivative of ψ_2 is equal to $(1-0)/(x_2-x_1) = 3/2$.

The derivatives of ψ_2 and ψ_3 are equal to 2 or -2 on the other subintervals.

$$\begin{aligned}
m_{11} = (\psi_1, \psi_1)_A &= \int_1^{4/3} (x^3+1)9\,\mathrm{d}x + \int_{4/3}^{2} (x^3+1)9/4\,\mathrm{d}x \\
&= \frac{9}{4}\left[x^4+4x\right]_1^{4/3} + \frac{9}{16}\left[x^4+4x\right]_{4/3}^{2} \\
&= \frac{9}{4}\left(\frac{256}{81}+\frac{16}{3}-1-4\right) + \frac{9}{16}\left(16+8-\frac{256}{81}-\frac{16}{3}\right) \\
&= \frac{64}{9}+12-\frac{9}{4}-9+9+\frac{9}{2}-\frac{16}{9}-3 = \frac{16}{3}+9+\frac{9}{4} = \frac{199}{12}, \\
m_{12} = (\psi_1, \psi_2)_A &= \int_{4/3}^{2} (x^3+1)(-9/4)\,\mathrm{d}x = -\frac{9}{16}\left[x^4+4x\right]_{4/3}^{2} \\
&= -\frac{9}{16}\left(16+8-\frac{256}{81}-\frac{16}{3}\right) = -9-\frac{9}{2}+\frac{16}{9}+3 = -\frac{157}{18}, \\
m_{22} = (\psi_2, \psi_2)_A &= \int_{4/3}^{2} (x^3+1)9/4\,\mathrm{d}x + \int_{2}^{5/2} (x^3+1)4\,\mathrm{d}x \\
&= \frac{9}{16}\left[x^4+4x\right]_{4/3}^{2} + \left[x^4+4x\right]_{2}^{5/2} \\
&= \frac{157}{18}+\frac{625}{16}+10-16-8 = \frac{4865}{144}, \\
m_{23} = (\psi_2, \psi_3)_A &= \int_{2}^{5/2} (x^3+1)(-4)\,\mathrm{d}x = -\left[x^4+4x\right]_{2}^{5/2} = \\
-\frac{625}{16}-10+16+8 &= -\frac{401}{16}, \\
m_{33} = (\psi_3, \psi_3)_A &= \int_{2}^{3} (x^3+1)4\,\mathrm{d}x\left[x^4+4x\right]_{2}^{3} = 81+12-16-8 = 69.
\end{aligned}$$

Matrix

$$M = \begin{pmatrix} 199/12 & -157/18 & 0 \\ -157/18 & 4865/144 & -401/16 \\ 0 & -401/16 & 69 \end{pmatrix}.$$

To infer the right-hand side vector b, we only need to calculate the area determined by the graph of ψ_i, $i = 1, 2, 3$. We obtain $b = (4, 14/3, 4)^{\mathrm{T}}$.

Let us assume that we have c, the solution to $Mc = b$. As a consequence, the approximate solution of the BVP with the homogeneous boundary conditions is

$$u_{0,\mathrm{FEM}} = c_1\psi_1 + c_2\psi_2 + c_3\psi_3$$

and the approximate solution to the original BVP reads

$$u_{\mathrm{FEM}} = u_{0,\mathrm{FEM}} - 3x + 5.$$

12 Finite-Difference Method in 1D: Boundary Value Problems

Problem 12.1: Use the finite-difference method with the step size $h = 1/2$ and find the approximate solution of the BVP

$$-\left(\frac{1}{4}+x^2\right)u'' - 2xu' = -\frac{60}{2+x},$$
$$u(-1) = 0, \quad u(1) = 0,$$

at points $x_1 = -1/2$, $x_2 = 0$, and $x_3 = 1/2$. (To approximate the first derivative, use the central difference, i.e., $u'(x_i) \approx (U_{i+1} - U_{i-1})/(2h)$.)

Solution: Let U_i denote the approximate solution at $x_i = -1 + ih$, where $i = 0, 1, 2, 3, 4$ and $h = 1/2$. Let us recall that $U_0 = 0$ and $U_4 = 0$. To slightly simplify the equation,[24] we multiply it by -1:

$$(1/4 + x^2)u'' + 2xu' = 60/(2+x).$$

To infer a finite-difference equation at $x_1 = -1/2$, we substitute for the derivatives and the known nodal values:

$$\frac{2}{4}\frac{U_0 - 2U_1 + U_2}{\frac{1}{4}} + 2\frac{-1}{2}(U_2 - U_0) = 60\frac{2}{3},$$
$$-4U_1 + U_2 = 40.$$

The finite-difference equation at $x_2 = 0$:

$$\frac{1}{4}\frac{U_1 - 2U_2 + U_3}{\frac{1}{4}} + 0 = 60\frac{1}{2},$$
$$U_1 - 2U_2 + U_3 = 30.$$

The finite-difference equation at $x_3 = 1/2$:

$$\frac{2}{4}\frac{U_2 - 2U_3 + U_4}{\frac{1}{4}} + 2\frac{1}{2}(U_4 - U_2) = 60\frac{2}{5},$$
$$U_2 - 4U_3 = 24.$$

Matrix form:

$$\begin{pmatrix} -4 & 1 & 0 & | & 40 \\ 1 & -2 & 1 & | & 30 \\ 0 & 1 & -4 & | & 24 \end{pmatrix} \sim \begin{pmatrix} 1 & -2 & 1 & | & 30 \\ -4 & 1 & 0 & | & 40 \\ 0 & 1 & -4 & | & 24 \end{pmatrix} \sim \begin{pmatrix} 1 & -2 & 1 & | & 30 \\ 0 & -7 & 4 & | & 160 \\ 0 & 1 & -4 & | & 24 \end{pmatrix}$$
$$\sim \begin{pmatrix} 1 & -2 & 1 & | & 30 \\ 0 & 1 & -4 & | & 24 \\ 0 & -7 & 4 & | & 160 \end{pmatrix} \sim \begin{pmatrix} 1 & -2 & 1 & | & 30 \\ 0 & 1 & -4 & | & 24 \\ 0 & 0 & -24 & | & 328 \end{pmatrix}.$$

[24] The main goal is to set up the algebraic equations. Their solution does not depend on the multiplicative parameter applied to *both* sides of the equation.

Solution, i.e., nodal values at x_3, x_2, and x_1, respectively,

$$U_3 = -\frac{328}{24} = -\frac{41}{3},$$
$$U_2 = 24 + 4U_3 = \frac{72}{3} - \frac{164}{3} = -\frac{92}{3},$$
$$U_1 = 30 - U_3 + 2U_2 = \frac{90}{3} + \frac{41}{3} - \frac{184}{3} = -\frac{53}{3}.$$

Problem 12.2: Use the finite-difference method with the step size $h = 1/2$ and find the approximate solution of the BVP[25]

$$-\left(\frac{1}{4} + x^2\right) u'' - 2xu' = -\frac{60}{2+x},$$
$$u(-1) = 0, \quad u'(1) = 2,$$

at points $x_1 = -1/2$, $x_2 = 0$, and $x_3 = 1/2$. (To approximate the first derivative, use the central difference, i.e., $u'(x_i) \approx (U_{i+1} - U_{i-1})/(2h)$.)

Solution: Let U_i denote the approximate solution at $x_i = -1 + ih$, where $i = 0, 1, 2, 3, 4, 5$ and $h = 1/2$; we have $U_0 = 0$. The node x_5 lies outside the interval $[-1, 1]$ and is associated with an unknown value U_5 of the solution extended beyond the interval $[-1, 1]$. To slightly simplify the equation,[26] we multiply it by -1:

$$(1/4 + x^2)u'' + 2xu' = 60/(2 + x).$$

The finite-difference equation at $x_1 = -1/2$:

$$\frac{2}{4}\frac{U_0 - 2U_1 + U_2}{\frac{1}{4}} + 2\frac{-1}{2}(U_2 - U_0) = 60\frac{2}{3},$$
$$-4U_1 + U_2 = 40.$$

The finite-difference equation at $x_2 = 0$:

$$\frac{1}{4}\frac{U_1 - 2U_2 + U_3}{\frac{1}{4}} + 0 = 60\frac{1}{2},$$
$$U_1 - 2U_2 + U_3 = 30.$$

The finite-difference equation at $x_3 = 1/2$:

$$\frac{2}{4}\frac{U_2 - 2U_3 + U_4}{\frac{1}{4}} + 2\frac{1}{2}(U_4 - U_2) = 60\frac{2}{5},$$
$$U_2 - 4U_3 + 3U_4 = 24.$$

[25] The equation is taken from Problem 12.1, but the boundary condition is partly different.

[26] The main goal is to set up the algebraic equations. Their solution does not depend on the multiplicative parameter applied to *both* sides of the equation.

We set up *two*[27] equation at $x_4 = 1$:
a) The equation associated with the boundary condition at $x_4 = 1$ (i.e., $u'(x_4) \approx \dfrac{U_5 - U_3}{2\frac{1}{2}}$):

$$U_5 - U_3 = 2. \tag{34}$$

b) The finite-difference equation associated with the differential equation at $x_4 = 1$

$$\frac{5}{4}\frac{U_3 - 2U_4 + U_5}{\frac{1}{4}} + 2(U_5 - U_3) = 60\frac{1}{3},$$
$$3U_3 - 10U_4 + 7U_5 = 20. \tag{35}$$

In (35), the unknown U_5 is replaced through (34)

$$3U_3 - 10U_4 + 7(2 + U_3) = 20,$$
$$10U_3 - 10U_4 = 6,$$
$$5U_3 - 5U_4 = 3.$$

The resulting system in the form of the augmented matrix

$$\left(\begin{array}{cccc|c} -4 & 1 & 0 & 0 & 40 \\ 1 & -2 & 1 & 0 & 30 \\ 0 & 1 & -4 & 3 & 24 \\ 0 & 0 & 5 & -5 & 3 \end{array}\right) \sim \left(\begin{array}{cccc|c} 1 & -2 & 1 & 0 & 30 \\ -4 & 1 & 0 & 0 & 40 \\ 0 & 1 & -4 & 3 & 24 \\ 0 & 0 & 5 & -5 & 3 \end{array}\right) \sim \left(\begin{array}{cccc|c} 1 & -2 & 1 & 0 & 30 \\ 0 & -7 & 4 & 0 & 160 \\ 0 & 1 & -4 & 3 & 24 \\ 0 & 0 & 5 & -5 & 3 \end{array}\right)$$
$$\sim \left(\begin{array}{cccc|c} 1 & -2 & 1 & 0 & 30 \\ 0 & 1 & -4 & 3 & 24 \\ 0 & -7 & 4 & 0 & 160 \\ 0 & 0 & 5 & -5 & 3 \end{array}\right) \sim \left(\begin{array}{cccc|c} 1 & -2 & 1 & 0 & 30 \\ 0 & 1 & -4 & 3 & 24 \\ 0 & 0 & -24 & 21 & 328 \\ 0 & 0 & 5 & -5 & 3 \end{array}\right) \sim \left(\begin{array}{cccc|c} 1 & -2 & 1 & 0 & 30 \\ 0 & 1 & -4 & 3 & 24 \\ 0 & 0 & -1 & 7/8 & 41/3 \\ 0 & 0 & 1 & -1 & 3/5 \end{array}\right)$$
$$\sim \left(\begin{array}{cccc|c} 1 & -2 & 1 & 0 & 30 \\ 0 & 1 & -4 & 3 & 24 \\ 0 & 0 & -1 & 7/8 & 41/3 \\ 0 & 0 & 0 & -1/8 & 214/15 \end{array}\right).$$

Nodal values

$$U_4 = -\frac{1712}{15},$$
$$U_3 = -\left(\frac{41}{3} - \frac{7}{8}U_4\right) = -\left(\frac{41}{3} + \frac{7 \cdot 214}{15}\right) = -\frac{1703}{15},$$
$$U_2 = 24 + 4U_3 - 3U_4 = -\frac{1316}{15},$$
$$U_1 = 30 - U_3 + 2U_2 = -\frac{479}{15}.$$

[27]The boundary condition can also be approximated by $u'(x_4) \approx \dfrac{U_4 - U_3}{\frac{1}{2}}$. In this approach, the use of the auxiliary node x_5 is avoided, but the error of the approximate solution is proportional to h. The use of the auxiliary node and the central difference makes the error proportional to h^2, i.e., one order better.

Problem 12.3: Use the finite-difference method with the step size $h = 1/2$ infer the linear equations for the approximate solution of the BVP[28]

$$-\left(\frac{1}{4}+x^2\right)u'' - 2xu' = -\frac{60}{2+x},$$
$$u'(-1) + 3u(-1) = -2, \quad u(1) = 5$$

at points $x_1 = -1/2$, $x_2 = 0$, and $x_3 = 1/2$. Do not solve the linear algebraic equations. (To approximate the first derivative, use the central difference, i.e., $u'(x_i) \approx (U_{i+1} - U_{i-1})/(2h)$.)

Solution: Let U_i denote the approximate solution at $x_i = -1+ih$, where $i = -1, 0, 1, 2, 3, 4$ and $h = 1/2$; we have $U_0 = 0$. The node $x_{-1} = -3/2$ lies outside the interval $[-1, 1]$ and is associated with an unknown value U_{-1} of the solution extended beyond the interval $[-1, 1]$. It is $U_4 = 5$ at x_4 by virtue of the boundary condition. To slightly simplify the equation,[29] we multiply it by -1:

$$(1/4 + x^2)u'' + 2xu' = 60/(2 + x).$$

We set up *two*[30] equation at $x_0 = -1$:

a) The equation associated with the boundary condition at x_0 (i.e., $u'(x_0) \approx \dfrac{U_1 - U_{-1}}{2\frac{1}{2}}$:

$$U_1 - U_{-1} + 3U_0 = -2,$$
$$U_{-1} = 2 + 3U_0 + U_1. \tag{36}$$

b) The finite-difference equation associated with the differential equation at $x_0 = -1$

$$\frac{5}{4}\frac{U_1 - 2U_0 + U_{-1}}{\frac{1}{4}} - 2(U_1 - U_{-1}) = 60,$$
$$3U_1 - 10U_0 + 7U_{-1} = 60. \tag{37}$$

In (37), the unknown U_{-1} is replaced through (36)

$$3U_1 - 10U_0 + 7(2 + 3U_0 + U_1) = 60,$$
$$11U_0 + 10U_1 = 46.$$

The finite-difference equation at $x_1 = -1/2$:

$$\frac{2}{4}\frac{U_0 - 2U_1 + U_2}{\frac{1}{4}} + 2\frac{-1}{2}(U_2 - U_0) = 60\frac{2}{3},$$
$$3U_0 - 4U_1 + U_2 = 40.$$

[28] The equation is taken from Problems 12.1 and 12.2, but the boundary condition is partly different.

[29] The main goal is to set up the algebraic equations. Their solution does not depend on the multiplicative parameter applied to *both* sides of the equation.

[30] The boundary condition can also be approximated by $u'(x_0) \approx \dfrac{U_1 - U_0}{\frac{1}{2}}$. In this approach, the use of the auxiliary node x_{-1} is avoided, but the error of the approximate solution is proportional to h. The use of the auxiliary node and the central difference makes the error proportional to h^2, i.e., one order better.

The finite-difference equation at $x_2 = 0$:

$$\frac{1}{4}\frac{U_1 - 2U_2 + U_3}{\frac{1}{4}} + 0 = 60\frac{1}{2},$$
$$U_1 - 2U_2 + U_3 = 30.$$

The finite-difference equation at $x_3 = 1/2$:

$$\frac{2}{4}\frac{U_2 - 2U_3 + U_4}{\frac{1}{4}} + 2\frac{1}{2}(U_4 - U_2) = 60\frac{2}{5},$$
$$U_2 - 4U_3 = 9.$$

The linear system for the unknowns $(U_0, U_1, U_2, U_3)^{\mathrm{T}}$ in the matrix form

$$\left(\begin{array}{cccc|c} 11 & 10 & 0 & 0 & 46 \\ 3 & -4 & 1 & 0 & 40 \\ 0 & 1 & -2 & 1 & 30 \\ 0 & 0 & 1 & -4 & 9 \end{array}\right).$$

Problem 12.4: Solve the problem

$$\left(\frac{1}{2} + x\right)u''(x) - 4xu(x) = 4x,$$
$$u(0) = 1, \quad u(2) = -1$$

approximately by the finite-difference method with the step size $1/2$ and calculate the approximate solution at $x = 3/2$.

Solution: Let U_i denote the approximate solution at $x_i = ih$, where $i = 0, 1, 2, 3, 4$ and $h = 1/2$; we have $U_0 = 1$ and $U_4 = -1$.

The finite-difference equation at $x_1 = \dfrac{1}{2}$:

$$1\frac{1 - 2U_1 + U_2}{\frac{1}{4}} - 2U_1 = 2,$$
$$-5U_1 + 2U_2 = -1.$$

The finite-difference equation at $x_2 = 1$:

$$\frac{3}{2}\frac{U_1 - 2U_2 + U_3}{\frac{1}{4}} - 4U_2 = 4,$$
$$3U_1 - 8U_2 + 3U_3 = 2.$$

The finite-difference equation at $x_3 = 3/2$:

$$2\frac{U_2 - 2U_3 - 1}{\frac{1}{4}} - 6U_3 = 6,$$
$$4U_2 - 11U_3 = 7.$$

The linear system for the unknowns $(U_1, U_2, U_3)^{\mathrm{T}}$ in the matrix form

$$\left(\begin{array}{ccc|c} -5 & 2 & 0 & -1 \\ 3 & -8 & 3 & 2 \\ 0 & 4 & -11 & 7 \end{array}\right) \sim \left(\begin{array}{ccc|c} -5 & 2 & 0 & -1 \\ 0 & -34 & 15 & 7 \\ 0 & 4 & -11 & 7 \end{array}\right) \sim \left(\begin{array}{ccc|c} -5 & 2 & 0 & -1 \\ 0 & -34 & 15 & 7 \\ 0 & 0 & -157 & 133 \end{array}\right).$$

We immediately get $U_3 = -\dfrac{133}{157}$.

13 Finite-Difference Method in 1D: Eigenvalues

Problem 13.1: Use the finite-difference method and find the two smallest approximate eigenvalues of the problem

$$u'' + \lambda 4^x u = 0,$$
$$u(0) = 0, \quad u(3/2) = 0.$$

Subdivide the interval into three subintervals of the same length.

Solution: The inner mesh nodes are $x_1 = 1/2$ and $x_2 = 1$; the step size $h = 1/2$. The exact values $u(x_1)$ and $u(x_2)$ will be approximated by U_1 and U_2. The eigenvalue will be approximated by μ.

The finite-difference equations (with the boundary conditions already applied)

$$\frac{-2U_1 + U_2}{\frac{1}{4}} + \mu 4^{\frac{1}{2}} U_1 = 0,$$
$$\frac{U_1 - 2U_2}{\frac{1}{4}} + \mu 4^1 U_2 = 0$$

can be arranged as follows

$$(\mu - 4)U_1 + 2U_2 = 0, \tag{38}$$
$$U_1 + (\mu - 2)U_2 = 0. \tag{39}$$

This system has a *nonzero* solution if and only if the determinant of the system matrix is equal to zero. That is,

$$(\mu - 4)(\mu - 2) - 2 = 0.$$

The equation has two solutions, namely, $\mu_1 = 3 - \sqrt{3} \approx 1{,}27$ and $\mu_2 = 3 + \sqrt{3} \approx 4{,}73$.

An equivalent way to the above eigenvalues originates in (38)-(39) written as

$$4U_1 - 2U_2 = \mu U_1,$$
$$-U_1 + 2U_2 = \mu U_2.$$

This is the eigenvalue problem for the matrix

$$A = \begin{pmatrix} 4 & -2 \\ -1 & 2 \end{pmatrix}.$$

The eigenvalues are $\mu_1 = 3 - \sqrt{3}$ and $\mu_2 = 3 + \sqrt{3}$.

Problem 13.2: Use the finite-difference method and find the two smallest approximate eigenvalues of the problem

$$u'' + \lambda \frac{2+x}{1-x} u = 0,$$
$$u(-1) = 0, \quad u(1/2) = 0.$$

Subdivide the interval into three subintervals of the same length.

Solution: The inner mesh nodes are $x_1 = -1/2$ and $x_2 = 0$; the step size $h = 1/2$. The exact values $u(x_1)$ and $u(x_2)$ will be approximated by U_1 and U_2. The eigenvalue will be approximated by μ.

The finite-difference equations (with the boundary conditions already applied)

$$\frac{-2U_1 + U_2}{\frac{1}{4}} + \mu \frac{\frac{3}{2}}{\frac{3}{2}} U_1 = 0,$$
$$-8U_1 + 4U_2 + \mu U_1 = 0; \tag{40}$$
$$\frac{U_1 - 2U_2}{\frac{1}{4}} + \mu 2U_2 = 0,$$
$$4U_1 - 8U_2 + \mu 2U_2 = 0 \tag{41}$$

can be arranged as follows

$$(\mu - 8)U_1 + 4U_2 = 0,$$
$$4U_1 + (2\mu - 8)U_2 = 0.$$

This system has a *nonzero* solution if and only if the determinant of the system matrix is equal to zero. That is,

$$0 = (\mu - 8)(2\mu - 8) - 16 = 2\left[(\mu - 8)(\mu - 4) - 8\right] = 2\left[\mu^2 - 12\mu + 24\right].$$

The equation has two solutions, namely,

$$\mu_{1,2} = \frac{12 \pm \sqrt{12^2 - 4 \cdot 24}}{2} = \frac{12 \pm \sqrt{4 \cdot 12}}{2} = 6 \pm \sqrt{12} = 6 \pm 2\sqrt{3}.$$

These are the approximate eigenvalues.[31]

Problem 13.3: a) Consider the following eigenproblem:

$$u'' + \lambda(4 - 4x^2)u = 0, \quad u(-1) = 0, \quad u(1) = 0.$$

Use the finite difference method and find (approximately) the first three eigenvalues and order them increasingly. Divide the interval $[-1, 1]$ into four subintervals of the same length.

b) Take the smallest and the second smallest eigenvalue and find the respective associated approximate eigenfunctions such that they attain the value 1 at $x = -1/2$ (normalization condition), sketch their graph.

[31] Again, it is possible to formulate a matrix eigenvalue problem as in Problem 13.1.

Solution: **a)** The inner points of the mesh are $x_1 = -1/2$, $x_2 = 0$, and $x_2 = 1/2$, the mesh size $h = 1/2$. The exact unknown values $u(x_i)$ will be approximated by unknown values U_i. The exact unknown eigenvalue will be approximated by a value μ.

Let us infer the finite-difference equations at $x_1 = -1/2$, $x_2 = 0$, and $x_2 = 1/2$. The boundary conditions will be used to simplify the equations associated with $x_1 = -1/2$ and $x_2 = 1/2$.

$$\frac{-2U_1 + U_2}{\frac{1}{4}} + \mu 3U_1 = 0,$$

$$-8U_1 + 4U_2 + 3\mu U_1 = 0; \tag{42}$$

$$\frac{U_1 - 2U_2 + U_3}{\frac{1}{4}} + \mu 4U_2 = 0,$$

$$4U_1 - 8U_2 + 4U_3 + 4\mu U_2 = 0, \tag{43}$$

$$\frac{U_2 - 2U_3}{\frac{1}{4}} + \mu 3U_3 = 0,$$

$$4U_2 - 8U_3 + 3\mu U_3 = 0, \tag{44}$$

From (42)–(44), we get a linear system

$$\begin{array}{l} (3\mu - 8)U_1 + 4U_2 = 0, \\ U_1 + (\mu - 2)U_2 + U_3 = 0, \\ 4U_2 + (3\mu - 8)U_3 = 0, \end{array} \qquad \text{or, in the matrix form,} \quad \begin{pmatrix} 3\,\mu - 8 & 4 & 0 \\ 1 & \mu - 2 & 1 \\ 0 & 4 & 3\,\mu - 8 \end{pmatrix},$$

that has a *nonzero* solution if and only if

$$\begin{aligned} 0 &= (3\mu - 8)^2(\mu - 2) - 8(3\mu - 8) = (3\mu - 8)\left((3\mu - 8)(\mu - 2) - 8\right) \\ &= (3\mu - 8)\left(3\mu^2 - 14\mu + 8\right). \end{aligned}$$

One root is immediately obvious, that is, $\mu_1 = 8/3$. The other approximate eigenvalues are the roots of the quadratic equation

$$3\mu^2 - 14\mu + 8 = 0.$$

Let us calculate: $D = 14^2 - 4 \cdot 3 \cdot 8 = 4(49 - 24) = 100$ implies

$$\mu_{2,3} = \frac{14 \pm \sqrt{D}}{6} = \frac{14 \pm 10}{6} \;\Rightarrow\; \mu_2 = 2/3,\ \mu_3 = 4.$$

Let us rearrange the eigenvalues to get an increasing sequence: $\mu_1 = 2/3$, $\mu_2 = 8/3$, and $\mu_3 = 4$. [32]

[32] The same triple can be obtained by solving the matrix eigenvalue problem $Aw = \mu w$ that corresponds to (42)–(44), where $w = (U_1, U_2, U_3)^{\mathrm{T}}$. In detail,

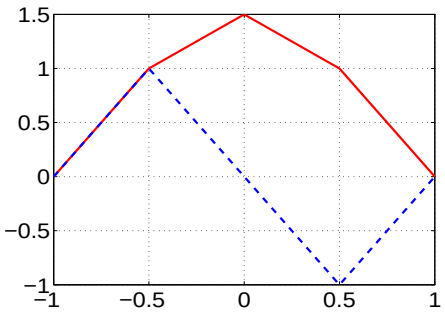
$$A = \begin{pmatrix} 8/3 & -4/3 & 0 \\ -1 & 2 & -1 \\ 0 & -4/3 & 8/3 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ U_3 \end{pmatrix} = \mu \begin{pmatrix} U_1 \\ U_2 \\ U_3 \end{pmatrix}.$$

b) The eigenvector u_1 associated with $\mu_1 = 2/3$ solves the homogeneous system

$$\begin{pmatrix} -6 & 4 & 0 \\ 1 & -4/3 & 1 \\ 0 & 4 & -6 \end{pmatrix} \sim \begin{pmatrix} -3 & 2 & 0 \\ 3 & -4 & 3 \\ 0 & 2 & -3 \end{pmatrix} \sim \begin{pmatrix} -3 & 2 & 0 \\ 0 & -2 & 3 \\ 0 & 2 & -3 \end{pmatrix} \sim \begin{pmatrix} -3 & 2 & 0 \\ 0 & -2 & 3 \end{pmatrix},$$

that is $u_1 = (1, 3/2, 1)^{\mathrm{T}}$.

The eigenvector u_2 associated with $\mu_2 = 8/3$ solves the homogeneous system $\begin{pmatrix} 0 & 4 & 0 \\ 1 & 2/3 & 1 \\ 0 & 4 & 0 \end{pmatrix}$, that is $u_2 = (1, 0, -1)^{\mathrm{T}}$. The vectors u_1 and u_2 represent the approximate eigenfunction value at $x = -1/2$, $x = 0$, and $x = 1/2$. We observe that the normalization condition is fulfilled, that is, the vectors u_1 and u_2 can be used without any modification. The eigenfunctions are equal to zero at $x = 0$ and $x = 2$ by virtue of the boundary conditions. See the graph.



14 Finite Difference Method in 2D: Poisson Equation

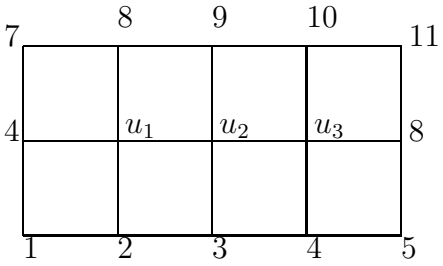
Problem 14.1: Let $\Omega = (0, 2) \times (0, 1)$ and let the boundary of Ω be denoted by Γ. Consider the following boundary value problem

$$\begin{aligned} -\Delta u &= -48x \text{ in } \Omega, \\ u &= 2x + 6y + 1 \text{ on } \Gamma \end{aligned}$$

and apply the finite difference method to find its approximate solution. Use the discretization step $h = 1/2$ in both x- and y-direction.

Solution:

The Dirichlet boundary condition gives the nodal values on the boundary Γ. The unknown nodal values are u_1, u_2, and u_3; see the sketch on the right.



The general difference equation at (x_i, y_j) (five-point stencil for $-\Delta$)[33]:

$$-\frac{U_{i+1}^j - 2U_i^j + U_{i-1}^j}{h_x^2} - \frac{U_i^{j+1} - 2U_i^j + U_i^{j-1}}{h_y^2} = f_i^j.$$

If $h_x = h_y = h$, which is our case, then

$$-U_{i+1}^j - U_{i-1}^j - U_i^{j+1} - U_i^{j-1} + 4U_i^j = h^2 f_i^j. \tag{45}$$

[33] Note the sign!

It is[34] $u_1 \equiv U_1^1$, $u_2 \equiv U_2^1$, and $u_3 \equiv U_3^1$. We evaluate $h^2 f_i^j$ at the inner nodes and, using the boundary values and (45), infer the equations

$$\begin{aligned} 4u_1 - u_2 - 2 - 4 - 8 &= -6, \\ -u_1 + 4u_2 - u_3 - 3 - 9 &= -12, \\ -u_2 + 4u_3 - 4 - 8 - 10 &= -18. \end{aligned}$$

In the matrix form,

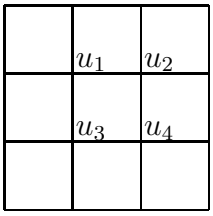
$$\left(\begin{array}{ccc|c} 4 & -1 & 0 & 8 \\ -1 & 4 & -1 & 0 \\ 0 & -1 & 4 & 4 \end{array}\right) \sim \left(\begin{array}{ccc|c} 4 & -1 & 0 & 8 \\ 0 & 15 & -4 & 8 \\ 0 & -1 & 4 & 4 \end{array}\right) \sim \left(\begin{array}{ccc|c} 4 & -1 & 0 & 8 \\ 0 & 15 & -4 & 8 \\ 0 & 0 & 56 & 68 \end{array}\right);$$

solution $(u_1, u_2, u_3)^{\mathrm{T}} = (31/14, 6/7, 17/14)^{\mathrm{T}}$.

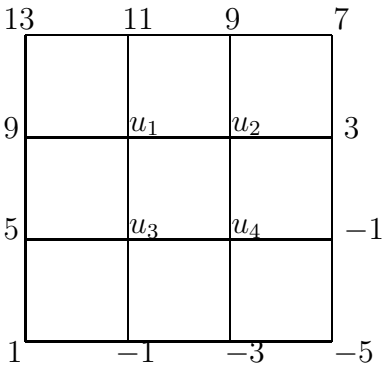
Problem 14.2: Let $\Omega = \{(x, y) \in \mathbb{R}^2 : \ x \in (0, 2), \ y \in (0, 2)\}$ and let the boundary of Ω be denoted by Γ. Consider the following boundary value problem and apply the finite difference method to find its approximate solution. Use the discretization step $h = 2/3$ in both x- and y-direction. The unknown values are denoted by $u_1, \ldots, u_4$ as shown below. Solve the resulting system of linear algebraic equations by Gaussian elimination

and check the correctness of the solution.

$$\begin{aligned} -\Delta u(x, y) &= 54(x - y) \ \text{ for } (x, y) \in \Omega, \\ u(x, y) &= 1 - 3x + 6y \ \text{ for } (x, y) \in \partial\Omega. \end{aligned}$$



Solution:



First, the values of u at the boundary nodes are calculated. Second, equations at the inner nodes are inferred. Finite difference equation at (x_i, y_j) is given by the five-point stencil:

$$-\frac{U_{i+1}^j - 2U_i^j + U_{i-1}^j}{h^2} - \frac{U_i^{j+1} - 2U_i^j + U_i^{j-1}}{q^2} = f_i^j.$$

If $h \equiv h_x = h_y$ (rectangular mesh), then

$$-U_{i+1}^j - U_{i-1}^j - U_i^{j+1} - U_i^{j-1} + 4U_i^j = h^2 f_i^j.$$

[34] In U_i^j, the subscript refers to the numbering in the horizontal (i.e., x-) direction, the superscript to the numbering in the vertical direction.

Since we have $u_1 \equiv U_1^2$, $u_2 \equiv U_2^2$, $u_3 \equiv U_1^1$, and $u_4 \equiv U_2^2$, we obtain

$$\begin{aligned}
-9 - 11 + 4u_1 - u_2 - u_3 &= \frac{4}{9}\frac{-2}{3}54 (= -16),\\
4u_1 - u_2 - u_3 &= 4;\\
-u_1 + 4u_2 - u_4 - 3 - 9 &= \frac{4}{9}0,\\
-u_1 + 4u_2 - u_4 &= 12;\\
-5 + 1 - u_1 + 4u_3 - u_4 &= \frac{4}{9}0,\\
-u_1 + 4u_3 - u_4 &= 4;\\
-u_2 - u_3 + 4u_4 + 1 + 3 &= \frac{4}{9}\frac{2}{3}54 (= 16),\\
-u_2 - u_3 + 4u_4 &= 12.
\end{aligned}$$

Let us solve the system (note that to make calculations simpler a shorter, the order of equations is changed and some equations are multiplied by -1)

$$\left(\begin{array}{cccc|c} 4 & -1 & -1 & 0 & 4\\ -1 & 4 & 0 & -1 & 12\\ -1 & 0 & 4 & -1 & 4\\ 0 & -1 & -1 & 4 & 12 \end{array}\right) \sim \left(\begin{array}{cccc|c} 1 & 0 & -4 & 1 & -4\\ 0 & 1 & 1 & -4 & -12\\ -1 & 4 & 0 & -1 & 12\\ 4 & -1 & -1 & 0 & 4 \end{array}\right) \sim \left(\begin{array}{cccc|c} 1 & 0 & -4 & 1 & -4\\ 0 & 1 & 1 & -4 & -12\\ 0 & 4 & -4 & 0 & 8\\ 0 & -1 & 15 & -4 & 20 \end{array}\right)$$

$$\sim \left(\begin{array}{cccc|c} 1 & 0 & -4 & 1 & -4\\ 0 & 1 & 1 & -4 & -12\\ 0 & 0 & -2 & 4 & 14\\ 0 & 0 & 16 & -8 & 8 \end{array}\right) \sim \left(\begin{array}{cccc|c} 1 & 0 & -4 & 1 & -4\\ 0 & 1 & 1 & -4 & -12\\ 0 & 0 & -1 & 2 & 7\\ 0 & 0 & 2 & -1 & 1 \end{array}\right) \sim \left(\begin{array}{cccc|c} 1 & 0 & -4 & 1 & -4\\ 0 & 1 & 1 & -4 & -12\\ 0 & 0 & -1 & 2 & 7\\ 0 & 0 & 0 & 3 & 15 \end{array}\right).$$

Solution $u = (u_1,\, u_2,\, u_3,\, u_4)^{\mathrm{T}} = (3,\, 5,\, 3,\, 5)^{\mathrm{T}}$.
Its correctness can easily be checked[35]

$$\begin{pmatrix} 4 & -1 & -1 & 0 & 4\\ -1 & 4 & 0 & -1 & 12\\ -1 & 0 & 4 & -1 & 4\\ 0 & -1 & -1 & 4 & 12 \end{pmatrix} \begin{pmatrix} 3\\ 5\\ 3\\ 5 \end{pmatrix} = \begin{pmatrix} 12 - 5 - 3\\ -3 + 20 - 5\\ -3 + 12 - 5\\ -5 - 3 + 20 \end{pmatrix} = \begin{pmatrix} 4\\ 12\\ 4\\ 12 \end{pmatrix},$$

which is the right-hand side of the system of equations.

15 Finite Difference Method in 2D: Heat Equation

Problem 15.1: Solve

$$\begin{aligned}
\frac{\partial u}{\partial t} &= 2\frac{\partial^2 u}{\partial x^2} \quad \text{in } \Omega = (0, 9) \times (0, 1),\\
u(x, 0) &= -x^2 + 8x + 9 \quad \text{for } 0 \le x \le 9,\\
u(0, t) &= 9 - 4t \quad \text{for } 0 \le t \le 1,\\
u(9, t) &= 8t \quad \text{for } 0 \le t \le 1
\end{aligned}$$

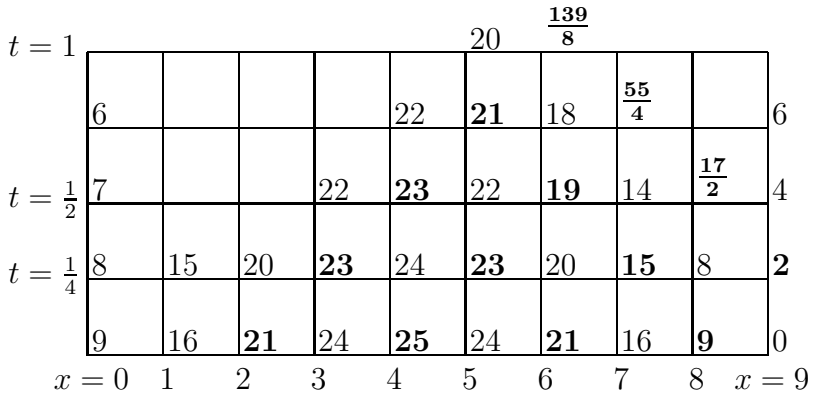
[35] To check whether u is a correct solution of $Au = f$, we calculate $\widehat{f} = Au$ and check whether $\widehat{f} = f$.

approximately by the finite difference method. Use the *explicit* four-point stencil and evaluate the approximate solution at points $B_1 = (5, 1)$ and $B_2 = (6, 1)$. Choose the discretization parameters h and τ (choose the values that are acceptable and result in simplest calculations).

Solution: The points B_1 and B_2 have to be mesh nodes. That is, $h = 1$. The stability condition $\tau \le h^2/(2a^2)$ gives $\tau = 1/4$ because $a^2 = 2$.

At $t = 0$, the nodal values are determined by the initial condition; see the figure.

Explicit scheme for $h = 1$ and $\tau = 1/4$: $U_i^{k+1} = \frac{1}{2}\left(U_{i+1}^k + U_{i-1}^k\right)$.



The boldface nodal values show the path leading to 139/8, the non-boldface nodal values indicate the path leading to the nodal value 20.

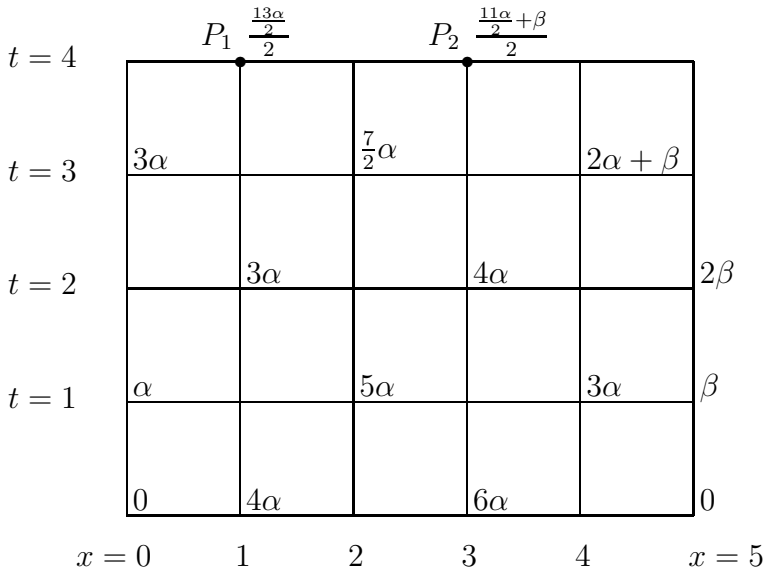
We conclude that the value at B_1 is equal to 20, and the value at B_2 is equal 139/8.

Problem 15.2: Let a heat conduction experiment be controlled by two parameters, namely α and β. These parameters control the initial temperature and the temperature at the end of a metal rod; see the following mathematical model:

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{1}{2}\frac{\partial^2 u}{\partial x^2} \quad \text{in } \Omega = \{(x,t) \in \mathbb{R}^2 :\ x \in (0,5),\ t \in (0,8)\} \\ u(x,0) &= \alpha(5x - x^2), \quad \text{where } 0 \le x \le 5, \\ u(0,t) &= \alpha t, \quad \text{where } 0 \le t \le 8, \\ u(5,t) &= \beta t, \quad \text{where } 0 \le t \le 8. \end{aligned}$$

Apply the *explicit* four-point scheme to calculate the approximate temperature at $P_1 = [1, 4]$ and $P_2 = [3, 4]$, and, moreover, find $\alpha, \beta \in \mathbb{R}$ such that the calculated temperature is equal to 26 units at P_1 and 15 units at P_2. Choose proper discretization parameters h and τ as large as possible.

Solution:



The points P_1, P_2 have to coincide with two mesh nodes. Since their x-coordinates are equal to 1 and 3, we conclude that the largest h we can consider is equal to one. The stability condition $\tau \leq h^2/(2a^2)$, where $a^2 = \frac{1}{2}$, then implies $\tau \leq 1$. Our choice: $h = 1$, $\tau = 1$.

As a consequence, the general explicit scheme reduces to $U_i^{k+1} = \frac{1}{2}\left(U_{i+1}^k + U_{i-1}^k\right)$. Moreover, we observe that to calculate the approximate solution at P_1 and P_2, we do not need to evaluate the solution at all the mesh nodes; see the figure above.

The scheme can only be applied when the initial and boundary nodal values are known; these are given by the expressions defining $u(x, 0)$, $u(0, t)$, and $u(5, t)$; see the figure. Since the values of the parameters α and β have not been identified yet, the nodal values are expressions in α and β.

We infer that $\dfrac{\frac{13\alpha}{2}}{2} = 26$. Consequently, $\alpha = 8$. After substituting $\alpha = 8$ into $\dfrac{\frac{11\alpha}{2} + \beta}{2} = 15$, we obtain $44 + \beta = 30$, that is, $\beta = -14$.

Problem 15.3: Solve the heat problem

$$\frac{\partial u}{\partial t} = \frac{1}{4}\frac{\partial^2 u}{\partial x^2} \quad \text{in } G = (0, 2) \times (0, 1),$$

$$u(x, 0) = 8 - 8(x - 1)^4 \quad \text{for } x \in [0, 2], \tag{46}$$

$$u(0, t) = 2t \quad \text{for } t \in [0, 1], \tag{47}$$

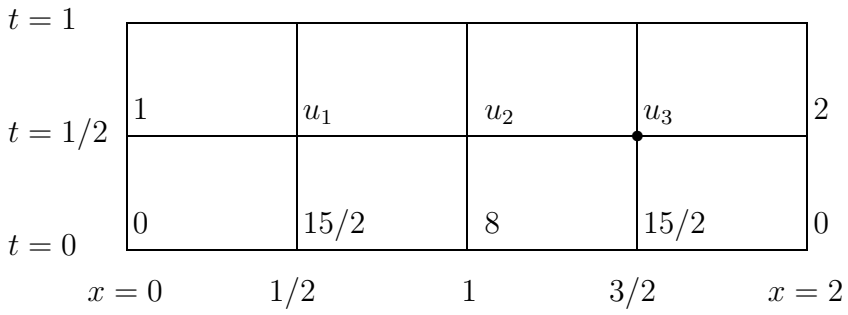
$$u(2, t) = 4t \quad \text{for } t \in [0, 1], \tag{48}$$

by the finite difference method with the *implicit* four-point scheme and calculate the approximate solution at the point given by $x = 3/2$ and $t = 1/2$. Choose the discretization parameters h and τ that make calculations simple.

Solution: The point $[3/2, 1/2]$ must be a mesh node. The larger h, the less equations will be formed, but $h = 3/2$ is not acceptable because the mesh should also cover the point $[2, 0]$, the bottom right vertex of G. The implicit scheme is unconditionally stable, thus the length of the time step τ is independent of h. The proper choice is $h = 1/2$ and $\tau = 1/2$.

The mesh has five nodes in the x-direction and three nodes in the t-direction. The nodal values are known for $t = 0$, see (46) and the figure. If $t = 1/2$, then $u(0, 1/2) \approx$

$U_0^1 = 2\tau = 1$ (see (47)) and $u(2, 1/2) \approx U_4^1 = 4\tau = 2$ (see (48)). Let us introduce $u_1 \equiv U_1^1$, $u_2 \equiv U_2^1$, and $u_3 \equiv U_3^1$; see the figure.



We will apply the implicit scheme (given in a general notation here)

$$\frac{U_i^k - U_i^{k-1}}{\tau} = a^2 \frac{U_{i-1}^k - 2U_i^k + U_{i+1}^k}{h^2}.$$

First equation ($x = 1/2$, $t = 1/2$, $a^2 = 1/4$)

$$\frac{u_1 - 15/2}{\frac{1}{2}} = \frac{1}{4} \frac{1 - 2u_1 + u_2}{\frac{1}{4}},$$
$$4u_1 - u_2 = 16.$$

Second equation ($x = 1$, $t = 1/2$)

$$\frac{u_2 - 8}{\frac{1}{2}} = \frac{1}{4} \frac{u_1 - 2u_2 + u_3}{\frac{1}{4}},$$
$$-u_1 + 4u_2 - u_3 = 16.$$

Third equation ($x = 3/2$, $t = 1/2$)

$$\frac{u_3 - 15/2}{\frac{1}{2}} = \frac{1}{4} \frac{u_2 - 2u_3 + 2}{\frac{1}{4}},$$
$$-u_2 + 4u_3 = 17.$$

Matrix form:

$$\left(\begin{array}{ccc|c} 4 & -1 & 0 & 16 \\ -1 & 4 & -1 & 16 \\ 0 & -1 & 4 & 17 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & -4 & 1 & -16 \\ -4 & 1 & 0 & -16 \\ 0 & -1 & 4 & 17 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & -4 & 1 & -16 \\ 0 & -15 & 4 & -80 \\ 0 & -1 & 4 & 17 \end{array}\right)$$
$$\sim \left(\begin{array}{ccc|c} 1 & -4 & 1 & -16 \\ 0 & 1 & -4 & -17 \\ 0 & -15 & 4 & -80 \end{array}\right) \sim \left(\begin{array}{ccc|c} 1 & -4 & 1 & -16 \\ 0 & 1 & -4 & -17 \\ 0 & 0 & -56 & -335 \end{array}\right).$$

We infer $u_3 = \dfrac{335}{56}$ from the row echelon form.