Numerical methods, midterm test I (2018/19 autumn, group A)
Solutions

PROBLEM 1. (6p) We are going to approximate the limit $\lim\limits_{x\to 1}\dfrac{x^{3/2}-x}{\sqrt{x}-1}$ by substituting $x = 0.99$ into the fraction in the present form. We use a calculator that uses a decimal number system with 2-digit-long mantissa (there is no restriction to the characteristic). Calculate the value of the fraction, explain the result, and give a better way of the calculation.

Solution: We round after each operation: $fl(0.99^{3/2}) = fl(0.985037) = 0.99$, thus the numerator is 0. In the denominator we have $fl(\sqrt{0.99}) = fl(0.994987) = 0.99$, thus the value of the denominator is -0.01. The value of the fraction is 0.

We can reformulate the fraction as follows

$$\frac{x^{3/2}-x}{\sqrt{x}-1} = \frac{x(\sqrt{x}-1)}{\sqrt{x}-1} = x,$$

which gives 0.99 on the given calculator. This value is much closer to the exact limit 1. The previous result was inaccurate due to the finite precision and cancellation.

PROBLEM 2. (6p) Let us consider the two linear systems

$$a)\quad \begin{bmatrix}6 & 1\\ 1 & 8\end{bmatrix}\overline{\mathbf{x}} = \begin{bmatrix}1\\ 2\end{bmatrix}, \qquad b)\quad \begin{bmatrix}6 & 7\\ 7 & 8\end{bmatrix}\overline{\mathbf{x}} = \begin{bmatrix}1\\ 2\end{bmatrix}.$$

Which of the two systems is the most sensitive to the change of the coefficients? For that system, give an upper bound for the change of the solution of the system in 1-norm in the case when we add real numbers that are not greater than 0.02 in absolute value to each element of the coefficient matrix and to the right-hand-side vector.

Solution: We must compute the condition numbers of the two coefficient matrices. The inverse of a matrix

$$\begin{bmatrix}a & b\\ b & c\end{bmatrix}$$

has the form

$$\frac{1}{ac-b^2}\begin{bmatrix}c & -b\\ -b & a\end{bmatrix}.$$

Thus the condition number of the first matrix (in maximum or 1-norm) is 81/47, and of the second one 225. The most sensitive equation is the second one.

The upper bound for the change of the solution can be given using the formula

$$\frac{\|\delta\overline{\mathbf{x}}\|_1}{\|\overline{\mathbf{x}}\|_1} \le \frac{\kappa_1(\mathbf{A})}{1-\kappa_1(\mathbf{A})\|\delta\mathbf{A}\|_1/\|\mathbf{A}\|_1}\left(\frac{\|\delta\mathbf{A}\|_1}{\|\mathbf{A}\|_1}+\frac{\|\delta\overline{\mathbf{b}}\|_1}{\|\overline{\mathbf{b}}\|_1}\right) = \frac{225}{1-225\cdot 0.04/15}\left(\frac{0.04}{15}+\frac{0.04}{3}\right) = 9.$$

Thus, the relative change cannot be greater than 900%.

PROBLEM 3. (7p) Show that if $\mathbf{A}$ is an M-matrix then the matrix $\mathbf{C} = (1/\omega)\mathbf{D} - \mathbf{R}$ is also an M-matrix for all $\omega \in (0, 1]$ (here $\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{R}$ is the usual splittig of $\mathbf{A}$ in the iterative methods). Show that if $\overline{\mathbf{g}}$ is a majorizing vector of $\mathbf{A}$ then it is valid the estimation

$$\|\mathbf{C}^{-1}\|_\infty \le \frac{\|\overline{\mathbf{g}}\|_\infty}{\min_i(\mathbf{A}\overline{\mathbf{g}})_i}.$$

Solution: Because $\mathbf{A}$ is an M matrix, $\mathbf{L}+\mathbf{R} \ge 0$ (offdiagonal is nonpositive), $\mathbf{D} > 0$ and there exists a majorizing vector $\overline{\mathbf{g}} > 0$ such that $\mathbf{A}\overline{\mathbf{g}} > 0$.

The offdiagonal of $\mathbf{C}$ is trivially nonpositive, because these elements are the offdiagonal elements of $\mathbf{A}$ as well.

We show that the majorizing vector $\overline{\mathbf{g}}$ of $\mathbf{A}$ majorizes also $\mathbf{C}$.

$$\mathbf{C}\overline{\mathbf{g}} = ((1/\omega)\mathbf{D} - \mathbf{R})\overline{\mathbf{g}} \geq (\mathbf{D} - \mathbf{R})\overline{\mathbf{g}} \geq (\mathbf{D} - \mathbf{R} - \mathbf{L})\overline{\mathbf{g}} = \mathbf{A}\overline{\mathbf{g}} > 0.$$

Because $\overline{\mathbf{g}}$ majorizes $\mathbf{C}$ and due to the above estimation we have

$$\|\mathbf{C}^{-1}\|_\infty \leq \frac{\|\overline{\mathbf{g}}\|_\infty}{\min_i(\mathbf{C}\overline{\mathbf{g}})_i} \leq \frac{\|\overline{\mathbf{g}}\|_\infty}{\min_i(\mathbf{A}\overline{\mathbf{g}})_i}.$$

Problem 4. (7p) The upper triangular matrix of the Cholesky decomposition of a matrix $\mathbf{A}$ has the form

$$\mathbf{F} = \begin{bmatrix} 1 & 1 & 1 & 2 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Compute the determinant of $\mathbf{A}$, give the LU decomposition of $\mathbf{A}$ and the solution of the system $\mathbf{A}\overline{\mathbf{x}} = [3, 4, 5, 9]^T$. Decide whether we can use the relaxed Gauss–Seidel method with relaxation parameter $\omega = 1.9$ to solve this system.

Solution: $\mathbf{A}$ has the Cholesky factorization in the form $\mathbf{A} = \mathbf{F}^T\mathbf{F}$. $det(\mathbf{A}) = det(\mathbf{F}^T)det(\mathbf{F}) = 1$, moreover now the Cholesky factorization is an LU factorization too (because the LU factorization is unique if $det(\mathbf{A}) \neq 0$). The system can be solved using two simple back-substitutions: first we solve $\mathbf{F}^T\overline{\mathbf{y}} = [3, 4, 5, 9]^T$, we get $\overline{\mathbf{y}} = [3, 1, 1, 1]^T$. Then we solve $\mathbf{F}\overline{\mathbf{x}} = \overline{\mathbf{y}}$. We get the solution of the system in the form $\overline{\mathbf{x}} = [1, 0, 0, 1]^T$. The relaxed Gauss–Seidel method works for symmetric positive definite systems with ω parameter from (0,2). Thus, the method is applicable to the present system.

Problem 5. (7p) Use the Jacobi method to solve the linear system
$$\begin{aligned} 6x_1 - x_2 &= 1 \\ -x_1 + 3x_2 - x_3 &= 2 \\ -x_2 + 2x_3 &= 1. \end{aligned}$$
Construct the iteration and estimate the number of iterations needed to approximate the exact solution of the system within the absolute error tolerance 10^{-6} in 1-norm. We start the iteration from the zero vector.

Solution: The Jacobi method has the iteration

$$\overline{\mathbf{x}}_{k+1} = \mathbf{B}\overline{\mathbf{x}}_k + \overline{\mathbf{f}} = \begin{bmatrix} 0 & 1/6 & 0 \\ 1/3 & 0 & 1/3 \\ 0 & 1/2 & 0 \end{bmatrix} \overline{\mathbf{x}}_k + \begin{bmatrix} 1/6 \\ 2/3 \\ 1/2 \end{bmatrix}.$$

Because $\|\mathbf{B}\|_1 = 2/3$ and $\overline{\mathbf{x}}_1 = \overline{\mathbf{f}}$, it is valid the estimation

$$\|\overline{\mathbf{x}}_k - \overline{\mathbf{x}}^\star\|_1 \leq \frac{(2/3)^k}{1 - 2/3} \cdot 4/3 \leq 10^{-6},$$

that shows that 38 iterations are enough to achieve the required error tolerance.

Problem 6. (7p) We are going to give the QR decomposition of the matrix $\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 0 & \sqrt{3} \\ 0 & 0 \end{bmatrix}$ using Householder reflections. The first Householder reflection, which belongs to the first column, is the matrix $\mathbf{H}_1 = \begin{bmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$. Give the QR decomposition of $\mathbf{A}$ and solve the over-determined system $\mathbf{A}\overline{\mathbf{x}} = [0,0,1,1]^T$ by the use of the QR decomposition.

Solution: The second Householder matrix has the form

$$\mathbf{H}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1/2 & \sqrt{3}/2 & 0 \\ 0 & \sqrt{3}/2 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

and

$$\mathbf{R} = \mathbf{H}_2\mathbf{H}_1\mathbf{A} = \begin{bmatrix} -1 & -1 \\ 0 & 2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

$$\mathbf{Q} = \mathbf{H}_1\mathbf{H}_2 = \begin{bmatrix} 0 & 1/2 & -\sqrt{3}/2 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & \sqrt{3}/2 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Because $\mathbf{Q}^T\overline{\mathbf{b}} = [0, \sqrt{3}/2, \star, \star]^T$, to compute the $\overline{\mathbf{x}}_{LS}$ solution, we have to solve the system

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

By back-substitution we obtain that $\overline{\mathbf{x}}_{LS} = [-\sqrt{3}/4, \sqrt{3}/4]^T$.

Numerical methods, midterm test I (2018/19 autumn, group B) Solutions

PROBLEM 1. (6p) We are going to approximate the limit $\lim_{x\to 1} \dfrac{x - x^{3/2}}{1-\sqrt{x}}$ by substituting $x = 0.999$ into the fraction in the present form. We use a calculator that uses a decimal number system with 3-digit-long mantissa (there is no restriction to the characteristic). Calculate the value of the fraction, explain the result, and give a better way of the calculation.

Solution: We round after each operation: $fl(0.999^{3/2}) = fl(0.99850037) = 0.999$, thus the numerator is 0. In the denominator we have $fl(\sqrt{0.999}) = fl(0.99949987) = 0.999$, thus the value of the denominator is 0.001. The value of the fraction is 0.

We can reformulate the fraction as follows

$$\frac{x - x^{3/2}}{1-\sqrt{x}} = \frac{x(1-\sqrt{x})}{1-\sqrt{x}} = x,$$

which gives 0.999 on the given calculator. This value is much closer to the exact limit 1. The previous result was inaccurate due to the finite precision and cancellation.

PROBLEM 2. (6p) Let us consider the two linear systems

$$a)\quad \begin{bmatrix} 7 & 6 \\ 6 & 5 \end{bmatrix} \overline{\mathbf{x}} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \qquad b)\quad \begin{bmatrix} 7 & 1 \\ 1 & 5 \end{bmatrix} \overline{\mathbf{x}} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}.$$

Which of the two systems is the most sensitive to the change of the coefficients? For that system, give an upper bound for the change of the solution of the system in 1-norm in the case when we add real numbers that are not greater than 0.01 in absolute value to each element of the coefficient matrix and to the right-hand-side vector.

Solution: We must compute the condition numbers of the two coefficient matrices. The inverse of a matrix

$$\begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

has the form

$$\frac{1}{ac-b^2}\begin{bmatrix} c & -b \\ -b & a \end{bmatrix}.$$

Thus condition number of the first matrix (in maximum or 1-norm) is 169, and of the second one 64/34. The most sensitive equation is the first one.

The upper bound for the change of the solution can be given using the formula

$$\frac{\|\delta\overline{\mathbf{x}}\|_1}{\|\overline{\mathbf{x}}\|_1} \le \frac{\kappa_1(\mathbf{A})}{1-\kappa_1(\mathbf{A})\|\delta\mathbf{A}\|_1/\|\mathbf{A}\|_1}\left(\frac{\|\delta\mathbf{A}\|_1}{\|\mathbf{A}\|_1} + \frac{\|\delta\overline{\mathbf{b}}\|_1}{\|\overline{\mathbf{b}}\|_1}\right) = \frac{169}{1-169\cdot 0.02/13}\left(\frac{0.02}{13} + \frac{0.02}{3}\right) = 1.87.$$

Thus, the relative change cannot be greater than 187%.

PROBLEM 3. (7p) Show that if $\mathbf{A}$ is an M-matrix then the matrix $\mathbf{B} = (1/\omega)\mathbf{D} - \mathbf{L}$ is also an M-matrix for all $\omega \in (0,1]$ (here $\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{R}$ is the usual splittig of $\mathbf{A}$ in the iterative methods). Show that if $\overline{\mathbf{g}}$ is a majorizing vector of $\mathbf{A}$ then it is valid the estimation

$$\|\mathbf{B}^{-1}\|_\infty \le \frac{\|\overline{\mathbf{g}}\|_\infty}{\min_i(\mathbf{A}\overline{\mathbf{g}})_i}.$$

Solution: Because $\mathbf{A}$ is an M matrix, $\mathbf{L} + \mathbf{R} \ge 0$ (offdiagonal is nonpositive), $\mathbf{D} > 0$ and there exists a majorizing vector $\overline{\mathbf{g}} > 0$ such that $\mathbf{A}\overline{\mathbf{g}} > 0$.

The offdiagonal of $\mathbf{B}$ is trivially nonpositive, because these elements are the offdiagonal elements of $\mathbf{A}$ as well.

We show that the majorizing vector $\overline{\mathbf{g}}$ of $\mathbf{A}$ majorizes also $\mathbf{B}$.

$$\mathbf{B}\overline{\mathbf{g}} = ((1/\omega)\mathbf{D} - \mathbf{L})\overline{\mathbf{g}} \geq (\mathbf{D} - \mathbf{L})\overline{\mathbf{g}} \geq (\mathbf{D} - \mathbf{L} - \mathbf{R})\overline{\mathbf{g}} = \mathbf{A}\overline{\mathbf{g}} > 0.$$

Because $\overline{\mathbf{g}}$ majorizes $\mathbf{B}$ and due to the above estimation we have

$$\|\mathbf{B}^{-1}\|_\infty \leq \frac{\|\overline{\mathbf{g}}\|_\infty}{\min_i(\mathbf{B}\overline{\mathbf{g}})_i} \leq \frac{\|\overline{\mathbf{g}}\|_\infty}{\min_i(\mathbf{A}\overline{\mathbf{g}})_i}.$$

PROBLEM 4. (7p) The upper triangular matrix of the Cholesky decomposition of a matrix $\mathbf{A}$ has the form

$$\mathbf{F} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 2 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Compute the determinant of $\mathbf{A}$, give the LU decomposition of $\mathbf{A}$ and the solution of the system $\mathbf{A}\overline{\mathbf{x}} = [2,3,3,4]^T$. Decide whether we can use the relaxed Gauss–Seidel method with relaxation parameter $\omega = 0.1$ to solve this system.

Solution: $\mathbf{A}$ has the Cholesky factorization in the form $\mathbf{A} = \mathbf{F}^T\mathbf{F}$. $det(\mathbf{A}) = det(\mathbf{F}^T)det(\mathbf{F}) = 1$, moreover now the Cholesky factorization is an LU factorization too (because the LU factorization is unique if $det(\mathbf{A}) \neq 0$). The system can be solved using two simple back-substitutions: first we solve $\mathbf{F}^T\overline{\mathbf{y}} = [2,3,3,4]^T$, we get $\overline{\mathbf{y}} = [2,1,0,0]^T$. Then we solve $\mathbf{F}\overline{\mathbf{x}} = \overline{\mathbf{y}}$. We get the solution of the system in the form $\overline{\mathbf{x}} = [1,1,0,0]^T$. The relaxed Gauss–Seidel method works for symmetric positive definite systems with ω parameter from (0,2). Thus the method is applicable to the present system.

PROBLEM 5. (7p) Use the Jacobi method to solve the linear system
$$\begin{aligned} 5x_1 - x_2 &= 1 \\ -x_1 + 4x_2 - x_3 &= 3 \\ -x_2 + 2x_3 &= 1. \end{aligned}$$
Construct the iteration and estimate the number of iterations needed to approximate the exact solution of the system within the absolute error tolerance 10^{-4} in 1-norm. We start the iteration from the zero vector.

Solution: The Jacobi method has the iteration

$$\overline{\mathbf{x}}_{k+1} = \mathbf{B}\overline{\mathbf{x}}_k + \overline{\mathbf{f}} = \begin{bmatrix} 0 & 1/5 & 0 \\ 1/4 & 0 & 1/4 \\ 0 & 1/2 & 0 \end{bmatrix} \overline{\mathbf{x}}_k + \begin{bmatrix} 1/5 \\ 3/4 \\ 1/2 \end{bmatrix}.$$

Because $\|\mathbf{B}\|_1 = 7/10$ and $\overline{\mathbf{x}}_1 = \overline{\mathbf{f}}$, it is valid the estimation

$$\|\overline{\mathbf{x}}_k - \overline{\mathbf{x}}^\star\|_1 \leq \frac{(7/10)^k}{1 - 7/10} \cdot 29/20 \leq 10^{-4},$$

that shows that 31 iterations are enough to achieve the required error tolerance.

PROBLEM 6. (7p) We are going to give the QR decomposition of the matrix $\mathbf{A} = \begin{bmatrix} 0 & \sqrt{3} \\ 0 & 1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}$ using Householder reflections. The first Householder reflection, which belongs to the first column, is the matrix $\mathbf{H}_1 = \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$. Give the QR decomposition of $\mathbf{A}$ and solve the over-determined system $\mathbf{A}\overline{\mathbf{x}} = [1, 0, 0, 1]^T$ by the use of the QR decomposition.

Solution: The second Householder matrix has the form

$$\mathbf{H}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1/2 & \sqrt{3}/2 & 0 \\ 0 & \sqrt{3}/2 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

and

$$\mathbf{R} = \mathbf{H}_2\mathbf{H}_1\mathbf{A} = \begin{bmatrix} -1 & -1 \\ 0 & -2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

$$\mathbf{Q} = \mathbf{H}_1\mathbf{H}_2 = \begin{bmatrix} 0 & -\sqrt{3}/2 & -1/2 & 0 \\ 0 & -1/2 & \sqrt{3}/2 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Because $\mathbf{Q}^T\overline{\mathbf{b}} = [0, -\sqrt{3}/2, \star, \star]^T$, to compute the $\overline{\mathbf{x}}_{LS}$ solution, we have to solve the system

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

By back-substitution we obtain that $\overline{\mathbf{x}}_{LS} = [-\sqrt{3}/4, \sqrt{3}/4]^T$.