Prof. Ch. Schwab
B. Fitzpatrick
J. Zech

Spring Term 2016

Numerische Mathematik I

ETH Zürich
D-MATH

Homework Problem Sheet 3

For some problems, parts of the solution are already given. Fill in the gaps and complete the proofs where you see a red band at the left margin.
Introduction. LU-decomposition, matrix-norms, condition.

Problem 3.1 Matrix Norms

(3.1a) For $\mathbf{A} \in \mathbb{R}^{n\times n}$ let $\rho(\mathbf{A}) := \max_{1\leq i\leq n} |\lambda_i|$ be the spectral radius of $\mathbf{A}$ (i.e. the maximum of the absolute value of all eigenvalues of $\mathbf{A}$). Is ρ a norm? Prove your claim.

Solution:

(3.1b) Define *submultiplicativity* of a matrix norm $\|\cdot\|_M$. Moreover, define *compatibility* of a matrix norm $\|\cdot\|_M$ with vector norms $\|\cdot\|_V$, $\|\cdot\|_W$.

HINT: See [NMI, Sect. 0.7].

Solution: A matrix norm $\|\cdot\|_M : \mathbb{C}^{m\times n} \to \mathbb{R}$ is submultiplicative if for all $\mathbf{A} \in \mathbb{C}^{a\times b}$ and $\mathbf{B} \in \mathbb{C}^{b\times c}$, we have

A matrix norm $\|\cdot\|_M$ on $\mathbb{C}^{m\times n}$ is compatible with vector norms $\|\cdot\|_V$ and $\|\cdot\|_W$ on $\mathbb{C}^n$ and $\mathbb{C}^m$ if

(3.1c) Show that the Frobenius norm $\|\cdot\|_F$ on $\mathbb{C}^{m\times n}$ is compatible with the vector norm $\|\cdot\|_2$.

Solution: For $\mathbf{A} \in \mathbb{C}^{m\times n}$, the Frobenius norm is given by

$$\|\mathbf{A}\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2\right)^{1/2}.$$

Consequently, we have

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

(3.1d) Let $\mathbf{I} \in \mathbb{R}^{n\times n}$ be the identity matrix and let $\mathbf{F} \in \mathbb{R}^{n\times n}$ be a matrix such that $\|\mathbf{F}\|_M < 1$ for a submultiplicative matrix norm $\|\cdot\|_M$. Show that $\mathbf{I}+\mathbf{F}$ is invertible and determine a formula for the inverse $(\mathbf{I}+\mathbf{F})^{-1}$.

Is the condition $\|\mathbf{F}\|_M < 1$ necessary for the existence of the inverse $(\mathbf{I}+\mathbf{F})^{-1}$? If so, state reasons; if not, provide a necessary and a sufficient condition.

Solution: Define $\mathbf{B} = -\mathbf{F}$. Then for all $m \in \mathbb{N}$, we have

$$\begin{aligned} & (\mathbf{I}-\mathbf{B})(\mathbf{I}+\mathbf{B}+\mathbf{B}^2+\ldots+\mathbf{B}^m) = \mathbf{I}-\mathbf{B}^{m+1} \\ \iff \quad & (\mathbf{I}+\mathbf{F})(\mathbf{I}-\mathbf{F}+\mathbf{F}^2-\ldots+(-\mathbf{F})^m) = \mathbf{I}-(-\mathbf{F})^{m+1}. \end{aligned}$$

(3.1e) Show that for $\mathbf{I}$, $\mathbf{F}$ and $\|\cdot\|_M$ as in subproblem (3.1d), we have

$$\left\|(\mathbf{I}+\mathbf{F})^{-1}\right\|_M \leqslant \|\mathbf{I}\|_M - 1 + \frac{1}{1-\|\mathbf{F}\|_M}$$

if the inverse exists.

Solution: If the inverse matrix $(\mathbf{I}+\mathbf{F})^{-1}$ exists, we can use the Neumann series, the triangle inequality and submultiplicativity to get:

(3.1f) Let $\lambda \in \mathbb{C}$ be an eigenvalue with eigenvector $\mathbf{x} \in \mathbb{C}^n$ of $\mathbf{A} \in \mathbb{C}^{n\times n}$ such that $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$. Using the decomposition $\mathbf{A} = \mathbf{D} + \mathbf{U}$, $\mathbf{D} = \operatorname{diag}(\mathbf{A})$, show that there exists an $i \in \{1,\ldots,n\}$ such that

$$|\lambda - a_{ii}| \leqslant \sum_{j=1, j\neq i}^{n} |a_{ij}|.$$

This is known as the Gershgorin circle theorem.

Solution: As a start, we decompose $\mathbf{A}$ into $\mathbf{D}+\mathbf{U}$ as indicated in the problem statement, which results in

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \quad \Longleftrightarrow \quad (\mathbf{D}+\mathbf{U})\mathbf{x} = \lambda\mathbf{x} \quad \Longleftrightarrow \quad (\lambda\mathbf{I}-\mathbf{D})\mathbf{x} = \mathbf{U}\mathbf{x}.$$

Now choose i such that $|x_i| = \max_{j\in\{1,\ldots,n\}} |x_j|$. Then, we have

(3.1g) Let $\mathbf{A} \in \mathbb{R}^{m\times n}$. First show that $\|\mathbf{Q}\mathbf{A}\|_F = \|\mathbf{A}\|_F$, if $\mathbf{Q} \in \mathbb{R}^{m\times m}$ is orthogonal. Then show that $\|\mathbf{A}\|_F^2 = \sum_{i=1}^{p} \sigma_i(\mathbf{A})^2$, where $p = \min\{m,n\}$ and $\sigma_i(\mathbf{A})$ is the i^{th} singular value of $\mathbf{A}$.

Solution:

(3.1h) Consider the matrix

$$\mathbf{A} := \frac{2}{7}\begin{pmatrix} 3 & -1 \\ 1 & 3 \end{pmatrix}.$$

Compute $\|\mathbf{A}\|_F$ and $\|\mathbf{A}\|_p$ for $p \in \{1, 2, \infty\}$. What can you say about the convergence of the series $\sum_{j\in\mathbb{N}} \mathbf{A}^j$?

Solution: Recall that $\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}(\mathbf{A}^\top \mathbf{A})}$ and

$$\|\mathbf{A}\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^{n} |a_{ij}|, \quad \|\mathbf{A}\|_\infty = \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}|, \quad \|\mathbf{A}\|_F = \sqrt{\sum_{i,j=1}^{n} |a_{ij}|^2}.$$

Problem 3.2 An Estimate of the Condition Number

In [NMI, Sect. 2.4.3] we conducted a careful analysis of the sensitivity of the solution map for linear systems of equations with respect to perturbations of the right hand side vector and the system matrix. Neumann series arguments and norm estimates for inverses were important tools. This problem will revisit them.

Let $\mathbf{A}, \Delta\mathbf{A} \in \mathbb{C}^{n\times n}$ be matrices such that $\mathbf{A}$ is regular. Prove the following properties:

(3.2a) If $\|\Delta\mathbf{A}\|_2 \|\mathbf{A}^{-1}\|_2 < 1$, then $\mathbf{A} + \Delta\mathbf{A}$ is regular.

Solution: The 2-norms for vectors and matrices are compatible, we have

$$\|\mathbf{x}\|_2 = \left\|\mathbf{A}^{-1}\mathbf{A}\mathbf{x}\right\|_2 \leqslant \left\|\mathbf{A}^{-1}\right\|_2 \left\|\mathbf{A}\mathbf{x}\right\|_2. \tag{3.2.1}$$

Moreover, by the triangle inequality,

$$\|\mathbf{A}\mathbf{x}\|_2 = \|-\Delta\mathbf{A}\mathbf{x} + \mathbf{A}\mathbf{x} + \Delta\mathbf{A}\mathbf{x}\|_2 \leqslant \|\Delta\mathbf{A}\mathbf{x}\|_2 + \|(\mathbf{A}+\Delta\mathbf{A})\mathbf{x}\|_2. \tag{3.2.2}$$

(3.2b) If $\mathbf{B} \in \mathbb{C}^{n\times n}$ is singular, then $1 \leqslant \|\mathbf{A}^{-1}\|_2 \|\mathbf{A}-\mathbf{B}\|_2$.

Solution:

(3.2c) For the condition number in the 2-norm, namely $\kappa_2(\mathbf{A}) := \|\mathbf{A}\|_2 \|\mathbf{A}^{-1}\|_2$, we have

$$\kappa_2(\mathbf{A})^{-1} \leqslant \inf\left\{ \frac{\|\mathbf{A}-\mathbf{B}\|_2}{\|\mathbf{A}\|_2} \,\middle|\, \mathbf{B} \in \mathbb{C}^{n\times n} \text{ singular} \right\}.$$

Solution: Using the result from the subproblem (3.2b), we get

Problem 3.3 Complexity of Pivoting Strategies

(3.3a) Determine the exact number of comparisons in the $\mathbf{LU}$ decomposition with column-wise pivoting ([NMI, Alg. 2.28]).

Solution: In the k^{th} execution of the loop, we have to find the absolute maximum of all elements $a_{ik}^{(k-1)}$ with $i \in \{k, \dots, n\}$, i. e. the maximum of $n-k+1$ elements. To do that, we need $n-k$ comparisons, so the total number of comparisons is

(3.3b) Determine the exact number of comparisons in the $\mathbf{LU}$ decomposition with complete pivoting (Full Pivoting Remark at the end of [NMI, Ch. 2.5]).

Solution:

(3.3c) Determine the costs of the Cholesky decomposition in [NMI, Alg. 2.37] assuming that floating point operations all have unit costs of 1.

Solution: We first calculate the costs c_1 for $\left(a_{ij} - \sum_{k=1}^{i-1} r_{ki} r_{kj}\right)/r_{ii}$ in the inner loop and get

The costs c_2 for $\left(a_{jj} - \sum_{k=1}^{j-1} r_{kj}^2\right)^{1/2}$ in the outer loop are given by

Inserting these into sums representing the inner and the outer loop, respectively, the result is

Problem 3.4 Gaussian Elimination with Pivoting for Structured Sparse Linear System

If the coefficient matrix of an $n \times n$ linear system of equations has special properties, Gaussian elimination can often be carried out with a computational effort that is much lower that the $\frac{2}{3}n^3 + \mathcal{O}(n^2)$ as counted in [NMI, Sect. 2.2]. In particular, we can usually benefit from *sparsity* of the matrix, that is, the property that only $\mathcal{O}(n)$ of its entries do not vanish. We saw this in the case of banded matrices in [NMI, Sect. 2.8] and in this problem we will come across banded matrices in disguise.

We consider a block partitioned linear system of equations

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad , \quad \mathbf{A} := \begin{pmatrix} \mathbf{D}_1 & \mathbf{C} \\ \mathbf{C} & \mathbf{D}_2 \end{pmatrix} \in \mathbb{R}^{2n\times 2n} , \tag{3.4.1}$$

where all the $n \times n$-matrices $\mathbf{D}_1$, $\mathbf{D}_2$, and $\mathbf{C}$ are diagonal. Hence, the matrix $\mathbf{A}$ can be described through three n-vectors $\mathbf{d}_1$, $\mathbf{c}$, and $\mathbf{d}_2$, which provide the diagonals of the matrix blocks. These vectors will be passed as arguments `d1`, `c`, and `d2` to the MATLAB codes below.

(3.4a) Write an *efficient* MATLAB function `function y = multA(d1,c,d2,x)` that returns $\mathbf{y} := \mathbf{A}\mathbf{x}$ in the column vector `y`. The argument `x` passes a column vector $\mathbf{x} \in \mathbb{R}^{2n}$.

(3.4b) Count the number of elementary floating point operations for a call to `multA` as you have implemented it in sub-problem (3.4a).

(3.4c) Assuming that the LU-decomposition of $\mathbf{A}$ exists, compute the matrices $\mathbf{L}$ and $\mathbf{U}$ of the factorization in Equation 3.4.1.

(3.4d) Write an *efficient* MATLAB function

```
function x = solveA(d1,c,d2,b)
```

that solves $\mathbf{A}\mathbf{x} = \mathbf{b}$ with Gaussian elimination *with partial pivoting* (see [NMI, Sect. 2.5]). The MATLAB \-operator must not be used.

HINT: Test your code with arguments `d1 = (1:n)'`, `d2 = -d1`, `c = ones(n,1)`, `b = [d1;d1]`. Compare with reference solution obtained in MATLAB by `x = A \ b`.

(3.4e) Analyze the asymptotic computational effort of your implementation of `solveA` in terms of the problem size parameter $n \to \infty$?

(3.4f) Runtime measurements in MATLAB can be carried out by means of the commands `tic` and `toc`. Please study their documentation (type "`help tic`" in MATLAB to view the documentation).

(3.4g) Determine the asymptotic complexity of the algorithm `solveA` in a numerical experiment using the MATLAB routines `tic` and `toc`. As test case use `d1 = (1:n)'`, `d2 = -d1`, `c = ones(n,1)`, `b = [d1;d1]` for $n = 2^2, \ldots, 2^{12}$. Create a suitable plot of the runtime versus n that allows to read off the asymptotic complexity.

(3.4h) In the previous subproblem you found the asymptotic complexity of the algorithm. Write down the asymptotic behaviour of the computational time as the dimension n tends to infinity using the Landau notation. Numerically estimate the constant occuring in the Landau notation (with time measured in seconds on your computer).

(3.4i) Find a permutation matrix $\mathbf{P} \in \mathbb{R}^{n \times n}$ such that $\mathbf{P}\mathbf{A}\mathbf{P}^\top$ becomes a banded matrix (see [NMI, Def. 2.40]) with bandwidth 3.

Problem 3.5 Cramer's Rule vs. Gaussian Elimination

Define

$$\mathbf{A}(\phi) := \begin{pmatrix} \frac{1}{\sqrt{2}} & \cos\phi \\ \frac{1}{\sqrt{2}} & \sin\phi \end{pmatrix} \quad , \quad \mathbf{b}(\phi) = \mathbf{A}(\phi)\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

The 2×2 linear system of equations $\mathbf{A}(\phi)\mathbf{x} = \mathbf{b}(\phi)$ with exact solution $\mathbf{x} = (x_1\ x_2)^\top = (1\ 1)^\top$ can be solved in two ways:

1. You may use Gaussian elimination employing the \-operator in MATLAB.
2. You may use Cramer's rule, that is $x_i = \det(\mathbf{A}_i)/\det(\mathbf{A})$ where $\mathbf{A}_i$ is obtained by replacing the i-th column of $\mathbf{A}$ by the vector $\mathbf{b}$. For the 2×2 system $\mathbf{A}\mathbf{x} = \mathbf{b}$, one has

$$x_1 = \frac{b_1 a_{22} - b_2 a_{12}}{a_{11}a_{22} - a_{21}a_{12}} \,, \qquad x_2 = \frac{a_{11}b_2 - a_{21}b_1}{a_{11}a_{22} - a_{21}a_{12}} \,. \tag{3.5.1}$$

(3.5a) Implement both solution strategies in MATLAB in order to obtain numerical solutions $\widetilde{\mathbf{x}}(\phi)$ (affected by roundoff) to $\mathbf{A}(\phi)\mathbf{x} = \mathbf{b}(\phi)$ for $\phi = \pi/4 - 10^{-9} : 10^{-11} : \pi/4 + 10^{-9}$. Plot in a suitable scale the condition number of $\mathbf{A}(\phi)$ in the 2-norm, the relative residual

$$\frac{\|\mathbf{b}(\phi) - \mathbf{A}(\phi)\widetilde{\mathbf{x}}(\phi)\|_2}{\|\mathbf{b}(\phi)\|_2}$$

as function of ϕ for the two approaches and the relative (forward) error

$$\frac{\|\widetilde{\mathbf{x}}(\phi) - \mathbf{x}\|_2}{\|\mathbf{x}\|_2}$$

vs. the function ϕ for both strategies.

(3.5b) Write down what you would say, if you had to step in front of the class and explain the observations made in (3.5a). Which solution strategy should be preferred?

Published on March 9, 2016.

To be submitted on March 22, 2016.

MATLAB: Submit all file in the online system. Include the files that generate the plots. Label all your plots. Include commands to run your functions. Comment on your results.

References

[NMI] Lecture Notes for the course "Numerische Mathematik I".

Last modified on March 9, 2016