Prof. Dr. Günter Bärwolf
Dr. Antoine Laurain
Dr. Christian Schröder
Tobias Ahnert
Ute Kandler
Dr. Nico Schlömer

Hanoi Fall-School 2013 on Numerical Analysis

Exercise sheet 1

Numerical Linear Algebra – "LU factorization"

You will find all the material for the first exercise in:

https://github.com/ukandler/hanoi2013_numerical_analyses

Exercise 1.1 Solve the system $\begin{bmatrix} 4 & 1 & 4 \\ 8 & 4 & 6 \\ 8 & 5 & 6 \end{bmatrix} x = \begin{bmatrix} 1 \\ 4 \\ 0 \end{bmatrix}$ using the LU decomposition

(a) without pivoting.

(b) with partial pivoting.

Definition 1 *Let $A \in \mathbb{R}^{n\times n}$ and let $A^{(k)} \in \mathbb{R}^{n\times n}$ define the converted system after the k-th step of the Gaussian elimination. Then the growth factor of A is defined by*

$$g_n = \frac{\max_{i,j,k} |a_{ij}^{(k)}|}{\max_{i,j} |a_{ij}|}.$$

Programming 1

(a) Write in Octave `function [LU,g] = LUfac(A)` that

- given a matrix $A \in \mathbb{R}^{n\times n}$
- computes the LU factorization of A and the growth factor g_n. (Hint: L and U are stored in place, i.e. in the same array as A)

(b) Write a second program `function x = forback(LU,b)` that

- using the LU factorization of A and the right hand side $b \in \mathbb{R}^n$
- computes the solution of the linear system $Ax = b$.

Test your program on the following examples:

$$i) \quad \begin{bmatrix} 5 & 3 & -1 & 0 \\ 2 & 5 & 0 & 1 \\ -1 & 0 & 5 & -2 \\ 0 & 1 & -2 & 5 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 27 \\ 24 \\ 4 \\ 4 \end{bmatrix} \qquad ii) \quad \begin{bmatrix} \varepsilon & 2 \\ 1 & 1 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{with} \quad \varepsilon = \texttt{1e-16}.$$

Compare your results with the one you get from $x = A\backslash b$. Explain numerically what happens in the case ii).

(c) Write an OCTAVE program `function [LU,p,g] = LUfac(A,pivot)` that

- given a matrix $A \in \mathbb{R}^{n \times n}$
- computes the LU factorization with or without partial pivoting of A and the growth factor g_n. (Hint: Include an option in your first function such that the user can decide whether partial pivoting is used (pivot=1) or not (pivot=0).)

Start the runme.m file. If your program is working correctly a little car is driving over a bridge. (Hint: It will take some seconds to see the result!)

Run your program on 10 random matrices of dimension $10, 30, 100, 300$ and 1000 and store the different growth factors in a 10 by 5 matrix.

- What is the maximum growth factor with and without partial pivoting?
- In the case of partial pivoting plot the growth factors as well as $f(n) = \sqrt{n}$. Explain the result. (Hint: Use a loglog scale plot.)

Exercise 1.2 How many operations (divisions and multiplications) are necessary to perform an LU decomposition without pivoting?

Exercise 1.3 Show that for LU factorization with partial pivoting applied to any matrix $A \in \mathbb{R}^{n \times n}$ the growth factor g_n satisfies $g_n \leq 2^{n-1}$.

Definition 2 *A matrix $A \in \mathbb{R}^{n \times n}$ is called diagonally dominant, if*

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \quad i = 1, \ldots, n.$$

Exercise 1.4 Let $A \in \mathbb{R}^{n\times n}$ be a nonsingular and diagonally dominant. Show that partial pivoting is not needed to perform the LU factorization.

Lemma 1 *(Perturbation Lemma) Let $P \in \mathbb{R}^{n\times n}$ and $\|P\| < 1$. Then $I-P$ is nonsingular and fulfills the estimation*

$$\|(I-P)^{-1}\| \leq \frac{1}{(1-\|P\|)}.$$

Exercise 1.5 Proof the following Theorem.

Theorem 1 *Let $Ax = b$ be a linear system with $A \in \mathbb{R}^{n\times n}$ and the solution $x = A^{-1}b$. Let $(A+\Delta A)(x+\Delta x) = b+\Delta b$ be the corresponding perturbed system with $\Delta A \in \mathbb{R}^{n\times n}$. If $\|\Delta A\| \leq \|A\|/\kappa(A)$, then $A+\Delta A$ is invertible and it holds*

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\kappa(A)}{1-\kappa(A)\frac{\|\Delta A\|}{\|A\|}}\left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|}\right).$$

Definition 3 *The matrix $A \in \mathbb{R}^{n\times n}$ is positive definite if and only if for all $x \in \mathbb{R}, x \neq 0$ it holds that $x^T Ax > 0$. The matrix A is symmetric if $A = A^T$.*

Exercise 1.6 Let $A \in \mathbb{R}^{n\times n}$ be symmetric and positive definite. Show that

(a) $a_{jj} > 0$ for $j = 1, 2, \ldots n$.

(b) $a_{jk}^2 < a_{jj}a_{kk}$ for $j, k = 1, 2, \ldots n, \quad j \neq k$.

(c) the largest entry of A in magnitude is on the main diagonal.

Exercise 1.7 Estimate the condition number of the the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & -2 & 0 & \ldots & & 0 \\ 0 & 1 & -2 & 0 & \ldots & 0 \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ 0 & & \ldots & 0 & 1 & -2 \\ 0 & & & \ldots & 0 & 1 \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

(Hint: Consider $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{x} = \begin{bmatrix} 2^{n-1} & \ldots & 8 & 4 & 2 & 1 \end{bmatrix}^T$.)

Additional exercises

Definition 4 *For $A \in \mathbb{R}^{m\times n}$*

$$\|A\|_1 := \max_{x\neq 0} \frac{\|Ax\|_1}{\|x\|_1} \quad \|A\|_\infty := \max_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty}$$
$$\|A\|_2 := \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} \quad \|A\|_F := \sqrt{\sum_{i=1}^{m}\sum_{k=1}^{n} |a_{ik}|^2}$$

Exercise 1.8* (norm properties)
Let $\|.\|$ be a vector norm in $\mathbb{R}^n$ and

$$\|A\| := \max_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|$$

the induced matrix norm. Let $A, B \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$. Show that:

(a) $\kappa(AB) \leq \kappa(A)\kappa(B)$

(b) $1 \leq \kappa(A)$

(c) $\frac{1}{\|A^{-1}b\|} \leq \frac{\|A\|}{\|b\|}$

(d) $\|A\|_1 = \max_{k=1,\dots,n} \sum_{i=1}^{m} |a_{ik}|$

(e) $\|A\|_\infty = \max_{i=1,\dots,m} \sum_{k=1}^{n} |a_{ik}|$

(f) $\|A\|_2 == \sqrt{\lambda_{\max}(A^T A)}$ (where $\lambda_{\max}(A)$ denotes the largest eigenvalue of A)

Definition 5 *A floating point numbers $\mathcal{F}(\beta, t, e_{\min}, e_{\max})$ are characterized by the base β, the precision t and the exponent range $[e_{\min}, e_{\max}]$. $\mathcal{F}$ consists of all numbers f of the form*

$$f = \pm .d_1 d_2 \dots d_t \times \beta^e, \quad 0 \leq d_i < \beta, \quad d_1 \neq 0 \quad e_{\min} \leq e \leq e_{\max}.$$

Exercise 1.9* (rounding)
Let $x_1, x_2 \in \mathbb{R}$ and $\bar{x} = \frac{x_1+x_2}{2}$. In exact arithmetic it holds the inequality

$$\min\{x_1, x_2\} \leq \bar{x} \leq \max\{x_1, x_2\} \tag{1}$$

In general this is not true in floating point arithmetic with $\bar{x} = fl(fl(x_1 + x_2)/2)$.
Find $x_1, x_2 \in \mathcal{F}(10, 2, -3, 3)$ such that (1) is violated.
How should $\bar{x}$ be modified in floating point arithmetic such that (1) holds?