universität wien

MASTERARBEIT

Titel der Masterarbeit

„Enclosures for solutions of overdetermined linear systems using a directed QR-decomposition“

verfasst von

Armin Spazierer, BSc

angestrebter akademischer Grad

Master of Science (MSc)

Wien, im November 2014

Studienkennzahl lt. Studienblatt:	A 066 821
Studienrichtung lt. Studienblatt:	Masterstudium Mathematik
Betreut von:	ao. Univ.-Prof. Dipl.-Ing. Dr. Hermann Schichl

ABSTRACT

As the title predicts, in this masterthesis we are looking for a method which provides enclosures for solutions of overdetermined linear systems of equations, allowing for inaccuracies in the input data, i.e., for errors in the parameters. Assuming the system to have a solution, the (useable) system with perturbated parameters is generally not solveable. Therefore we have to consider the least squares problem with those. Knowing bounds for the perturbations, they can be translated into so-called hybrid norms. Using those and assuming exact arithmetic, we show that theoretic bounds can be computed for the solution by a reduced QR-decomposition.

Since we have to take into account roundoff errors in floating point arithmetic, we need stronger tools for enclosure. Computing a QR-decomposition, basing on the Householder method, in a specific way, we can control these errors during the factorization, and combine them with the initial errors to hybrid norms, so that it will also be possible to obtain enclosures for the existing solutions.

In addition, Matlab code, which perform the upcoming concept, will be presented. Entering inaccurate parameters and bounds for the size of the perturbations provides an interval vector containg the solution of the overdetermined system.

Finally, we will analyze the algorithm and compare it to the evaluation of a (floating point) solution, using the Householder method.

Contents

1 Mathematical background

In this section we recapitulate basic mathematical concepts which are important for our topic. We will consider mainly statements and results, found in Stoer/Bulirsch [1] resp. Schwarz, Köckler [2], and abstain from illustrating all details and proofs.
At first, we start with some definitions:

Definition 1.1. *For $m, n \in \mathbb{N}$, we denote the unit matrix of size m by I_m and the matrix of size $m \times n$ which contains only zeros by $0_{m\times n}$. Furthermore we define $0_m := 0_{m\times m}$, $o_m := 0_{m\times 1}$, the null vector of size m and we denote the vector of length m which contains only ones by 1_m.*

Definition 1.2. *Let $e_i^{(m)} \in \mathbb{R}^m$ define the ith unit vector of size m, i.e.,*

$$(e_i^{(m)})_j := \delta_{ij}, \qquad \textit{where} \qquad \delta_{ij} := \begin{cases} 1, & \textit{if } i = j, \\ 0, & \textit{otherwise} \end{cases}$$

for $j \in \{1, \ldots, m\}$.

In the following we take absolute values and inequalities for matrices componentwise, i.e., $\forall A, B \in \mathbb{R}^{m\times n}$ with $A = (a_{ij})_{1\le i\le m, 1\le j\le n}$ and $B = (b_{ij})_{1\le i\le m, 1\le j\le n}$:

$$|A| = (|a_{ij}|)_{ij} \quad \text{and} \quad A \le B \Leftrightarrow a_{ij} \le b_{ij}$$

for all $(i, j) \in \{1, \ldots, m\} \times \{1, \ldots, n\}$.

A matrix $A \in \mathbb{R}^{m\times n}$ is said to be *rectangular*, if $m \neq n$ for $m, n \in \mathbb{N}$. Considering linear systems of equations with rectangular matrices, then it

is not a priori clear, what is ment by the solution $x \in \mathbb{R}^n$ of the equation $Ax = b$ for a given $b \in \mathbb{R}^n$. If the system is overdetermined, i.e., $m > n$, the set of solutions might be empty, and if the system is underdetermined, which means that $m < n$, then in general there exist infinitely many solutions. In this thesis, we will focus on overdetermined systems.
Defining the *residual* (with respect to $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^n$) by

$$\mathrm{res}(x) : \mathbb{R}^n \to \mathbb{R}^m, \quad \mathrm{res}(x) := Ax - b,$$

the linear system of equations is satisfied with solution x, if and only if $\mathrm{res}(x) = o_n$. Now, if there exists no solution of the system of equations, then there is no $x \in \mathbb{R}^n$ for which $\mathrm{res}(x) = o_n$. In that case it is reasonable to minimize the residual in the following sense:

1.1 Least squares problems

Definition 1.3. *Let $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ and $m \geq n$. Then x^* is called a least squares solution of the system $Ax = b$ if*

$$x^* = \arg\min_{x\in\mathbb{R}^n} \|\mathrm{res}(x)\|_2 = \arg\min_{x\in\mathbb{R}^n} \|Ax - b\|_2,$$

whereas $\min \|Ax - b\|_2$ is called least squares problem.

Before we describe, how to solve a least squares problem, we need the following result:

Proposition 1.4. *Let $A \in \mathbb{R}^{m\times n}$ and $m \geq n$. Then $\mathrm{rk}(A) = \mathrm{rk}(A^TA)$.*

Hence $\mathrm{rk}(A) = n$ implies $\mathrm{rk}(A^TA) = n$, so that $A^TA \in \mathbb{R}^{n\times n}$ is regular.

Theorem 1.5. *Let $A \in \mathbb{R}^{m\times n}$, $m \geq n$ and $\mathrm{rk}(A) = n$. Then the (unique) solution of the least squares problem*

$$x^* = \arg\min_{x\in\mathbb{R}^n} \|Ax - b\|_2$$

satisfies the normal equation

$$A^T Ax^* = A^T b.$$

The solution is unique, since by Proposition 1.4 the matrix $A^T A$ has full rank. By Theorem 1.5, we obtain the least squares solution by solving the normal equation. For example, this can be done by Cholesky factorization of $A^T A$:
Since $A^T A$ is symmetric and positive semidefinite, there exists a Cholesky factorization $A^T A = R^T R$ and the normal equation changes to $R^T Rx^* = A^T b$. Then the solution x^* is obtained by solving $R^T y = A^T b$ and $Rx^* = y$ by forward respectively backward substitution. But the fact that $A^T A$ is often poorly conditioned makes this method potentially unstable:

Example 1.6. *Assume rounding to 5 significant digits and consider the least squares problem with parameters*

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 10^{-3} \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 2 \\ 10^{-3} \\ 3 \end{pmatrix}.$$

Then $x^* = (1, 1)^T$. But computing $A^T A$ yields

$$A^T A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},$$

which is singular, so that there is no unique solution and this method fails.

Thus, we look for a better suited method to solve the normal equation, which lead us to the QR-decomposition:

Definition 1.7. *A matrix $R = (r_{ij})_{1 \leq i \leq m, 1 \leq j \leq n} \in \mathbb{R}^{m \times n}$ is called an upper triangular matrix if $r_{ij} = 0$ for all $i > j$.*

Theorem 1.8. *For every matrix $A \in \mathbb{R}^{m \times n}$ (with $m \geq n$) there exists a reduced QR-decomposition (resp. QR-factorization) $A = QR$ into a matrix $Q \in \mathbb{R}^{m \times n}$ of orthonormal columns and an upper triangular matrix $R \in \mathbb{R}^{n \times n}$. If A has full rank, the factorization is unique, provided that $R_{ii} > 0$ for all $i \in \{1, \ldots, n\}$.*

Completing Q to a matrix $\widetilde{Q} \in \mathbb{R}^{m \times m}$, whose columns form an orthonormal basis of $\mathbb{R}^m$ and defining the upper triangular matrix $\widetilde{R} \in \mathbb{R}^{m \times n}$ by adding $m - n$ zero-rows to R, we get a so called *full* QR-decomposition $A = \widetilde{Q}\widetilde{R}$.

Proposition 1.9. *Let $A \in \mathbb{R}^{m \times n}$, $\mathrm{rk}(A) = n$ and $A = QR$ be a reduced QR-decomposition of A. Then $\mathrm{rk}(R) = n$.*

Proof: $n = \mathrm{rk}(A) = \mathrm{rk}(QR) \leq \min\{\mathrm{rk}(Q), \mathrm{rk}(R)\} = \min\{n, \mathrm{rk}(R)\} \Rightarrow \mathrm{rk}(R) = n$. □

Therefore, if A has full rank, then R (and therefore R^T) is regular and

using the (reduced) factorization $A = QR$, the normal equation changes to

$$R^T Q^T QRx = R^T Q^T b \Leftrightarrow R^T Rx = R^T Q^T b \Leftrightarrow Rx = Q^T b,$$

which again can be solved by backward substitution. Summarizing, we can find a least squares solution by computing a QR-factorization of A and solving the linear system of equations $Rx = Q^T b$.

In Example 1.6, a (reduced) QR-decomposition of the matrix A is found easily, since it already has upper triangular form: $A = QR$ holds for

$$Q = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad R = \begin{pmatrix} 1 & 1 \\ 0 & 10^{-3} \end{pmatrix}.$$

Hence, computing $Rx = Q^T b$ leads to the system

$$\begin{pmatrix} 1 & 1 \\ 0 & 10^{-3} \end{pmatrix} \begin{pmatrix} x_1^* \\ x_2^* \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 2 \\ 10^{-3} \\ 3 \end{pmatrix} = \begin{pmatrix} 2 \\ 10^{-3} \end{pmatrix},$$

with rounded solution $x^* = (1, 1)^T$.

Knowing how to calculate the solution of a least squares problem for a given matrix A and vector b, we can consider the following problem, drafted in Neumaier [4]:

In applications it is often the case that A and b are measurements of unknown parameters $\hat{A}$ and $\hat{b}$ of an overdetermined system of equations $\hat{A}\hat{x} = \hat{b}$. If x^* denotes the solution of the least squares problem with parameters A

and b, then x^* lies "close" to $\hat{x}$ and we will show that a bound for the error $|\hat{x} - x^*|$ can be found.

Since the accuracy of the components might differ, it is sensible to bound this error componentwise. Therefore it would be helpful to have columnwise bounds for the deviations $\hat{A} - A$ and $\hat{b} - b$. We assume that such columnwise bounds are given or can be determined (e.g. from the accuracy of the measurements).

A proper way to handle componentwise bounds is the concept of hybrid norms; main results can be found in Neumaier [4]:

1.2 Hybrid norms

Definition 1.10. *Let $A \in \mathbb{R}^{m\times n}$ and let $A_{i:}$ denote the ith row of A. Similarly, the jth column of A is denoted by $A_{:j}$. Then we define the vector-valued functions ν_p and μ_p by*

$$\nu_p : \mathbb{R}^{m\times n} \to \mathbb{R}^m, \quad \nu_p(A) := \begin{pmatrix} \|A_{1:}\|_p \\ \vdots \\ \|A_{m:}\|_p \end{pmatrix},$$

$$\mu_p : \mathbb{R}^{m\times n} \to \mathbb{R}^n, \quad \mu_p(A) := \begin{pmatrix} \|A_{:1}\|_p \\ \vdots \\ \|A_{:n}\|_p \end{pmatrix}.$$

These functions ν_p and μ_p are called hybrid norms (with respect to the p-norm).

1.2.1 Properties of hybrid norms

In this subsection we present important properties of hybrid norms, which are fundamental for our topic.

Proposition 1.11. *Let $A \in \mathbb{R}^{m\times n}$. Then*

(i) $\nu_p(A) = \mu_p(A^T)$ *and*

(ii) $\mu_p(A) = \nu_p(A^T)$.

Proof: (i) Obviously, $A_{i:} = A_{:i}^T$ for all $i \in \{1, \ldots, m\}$. Therefore

$$\nu_p(A) = \begin{pmatrix} \|A_{1:}\|_p \\ \vdots \\ \|A_{m:}\|_p \end{pmatrix} = \begin{pmatrix} \|A_{:1}^T\|_p \\ \vdots \\ \|A_{:m}^T\|_p \end{pmatrix} = \mu_p(A^T).$$

(ii) By (i) we obtain $\mu_p(A) = \mu_p((A^T)^T) = \nu_p(A^T)$. □

Theorem 1.12. *Let $A, B \in \mathbb{R}^{m\times n}$, $\alpha \in \mathbb{R}$ and $p \in \mathbb{N} \cup \{\infty\}$. Then*

(i) $\nu_p(A) = o_m \Leftrightarrow A = 0_{m\times n}$,

(ii) $\nu_p(\alpha A) = |\alpha|\nu_p(A)$,

(iii) $\nu_p(A + B) \leq \nu_p(A) + \nu_p(B)$,

(iv) (i) - (iii) *apply for μ_p too.*

Proof: (i) $\nu_p(A) = o_m \Leftrightarrow A_{i:} = o_n^T$ for all $i \in \{1, \ldots, m\} \Leftrightarrow A = 0_{m\times n}$.

(ii) $$\nu_p(\alpha A) = \begin{pmatrix} \|\alpha A_{1:}\|_p \\ \vdots \\ \|\alpha A_{m:}\|_p \end{pmatrix} = \begin{pmatrix} |\alpha|\|A_{1:}\|_p \\ \vdots \\ |\alpha|\|A_{m:}\|_p \end{pmatrix} = |\alpha|\nu_p(A).$$

(iii) $\nu_p(A+B) = \begin{pmatrix} \|A_{1:}+B_{1:}\|_p \\ \vdots \\ \|A_{m:}+B_{m:}\|_p \end{pmatrix} \leq \begin{pmatrix} \|A_{1:}\|_p + \|B_{1:}\|_p \\ \vdots \\ \|A_{m:}\|_p + \|B_{m:}\|_p \end{pmatrix}$

$= \nu_p(A) + \nu_p(B).$

(iv) By Proposition 1.11 (ii), the same results can be proved for μ_p inserting A^T and B^T instead of A and B in the equalities:

(i) $\mu_p(A) = o_n \Leftrightarrow \nu_p(A^T) = o_n \Leftrightarrow A^T = 0_{n\times m} \Leftrightarrow A = 0_{m\times n}$,

(ii) $\mu_p(\alpha A) = \nu_p(\alpha A^T) = |\alpha|\nu_p(A^T) = |\alpha|\mu_p(A)$ and

(iii) $\mu_p(A+B) = \nu_p(A^T+B^T) \leq \nu_p(A^T) + \nu_p(B^T) = \mu_p(A) + \mu_p(B)$.

□

Lemma 1.13. *Let $A, B \in \mathbb{R}^{m\times n}$ and $x \in \mathbb{R}^n$. Then*

(i) $\|Ax\|_p \leq \mu_p(A)^T|x|$, for every matrix norm $\|.\|_p$,

(ii) $\mu_p(AB) \leq \|A\|_p\mu_p(B)$, for every submultiplicative $\|.\|_p$,

(iii) $\nu_p(AB) \leq \|B\|_p\nu_p(A)$, for every submultiplicative $\|.\|_p$,

(iv) $\|AB\|_p \leq \|\mu_p(A)^T|B|\|_p$, for every submultiplicative $\|.\|_p$.

Proof: We denote the ith entry of $\nu_2(A)$ in the following by $\nu_2(A)_i$, i.e., $\nu_2(A)_i = \|A_{i:}\|_2$, $(1 \leq i \leq m)$ and equivalently $\mu_2(A)_j = \|A_{:j}\|_2$, $(1 \leq j \leq n)$. Then we obtain (i) by:

$$\|Ax\|_p = \left\|\sum_{j=1}^{n} A_{:j}x_j\right\|_p \leq \sum_{j=1}^{n} \|A_{:j}x_j\|_p = \sum_{j=1}^{n} \|A_{:j}\|_p|x_j| = \mu_p(A)^T|x|.$$

Furthermore, (ii) follows from

$$\mu_p(AB)_j = \|AB_{:j}\|_p \leq \|A\|_p\|B_{:j}\|_p = \|A\|_p\mu_p(B)_j$$

and analogously (iii) by

$$\nu_p(AB)_i = \|A_{i:}B\|_p \leq \|A_{i:}\|_p \|B\|_p = \|B\|_p \nu_p(A)_i.$$

Finally, we have (iv) by (i), since $\|Ax\|_p \leq \mu_p(A)^T|x| \Rightarrow \|ABy\|_p \leq \mu_p(A)^T|By|$ where $x = By \Rightarrow \|ABy\|_p \leq \mu_p(A)^T|By| \leq \mu_p(A)^T|B||y| \Rightarrow \|ABy\|_p \leq \|\mu_p(A)^T|B||y|\|_p \leq \|\mu_p(A)^T|B|\|_p \||y|\|_p$, whereby

$$\sup_{y \neq 0} \frac{\|ABy\|_p}{\|y\|_p} \leq \sup_{y \neq 0} \|\mu_p(A)^T|B|\|_p = \|\mu_p(A)^T|B|\|_p.$$

□

In addition, we consider statements for the particular case $p = 2$, which will be especially important:

Lemma 1.14. *Let $A, B \in \mathbb{R}^{m \times n}$ and $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$. Then the following statements hold:*

(i) $|Ax| \leq \nu_2(A)\|x\|_2$

(ii) $|y^T A| \leq \mu_2(A)^T \|y\|_2$

(iii) $\|A\|_2 \leq \|\nu_2(A)\|_2$

(iv) $\|\mu_2(A)\|_2 = \sqrt{tr\,(A^T A)}$ $(= \|A\|_F)$

(v) $\|\nu_2(A)\|_2 = \|\mu_2(A)\|_2$

(vi) $|A| \leq |B| \Rightarrow \nu_2(A) \leq \nu_2(B)$

(vii) $|A| \leq |B| \Rightarrow \mu_2(A) \leq \mu_2(B)$

Proof: Applying the Cauchy-Schwarz inequality we obtain (i) by

$$|Ax|_i = |A_{i:}x| \leq \|A_{i:}\|_2 \|x\|_2 = \nu_2(A)_i \|x\|_2$$

and analogously (ii) from

$$|y^T A|_j = |y^T A_{:j}| \leq \|y\|_2 \|A_{:j}\|_2 = \mu_2(A)_j^T \|y\|_2.$$

Making use of (i) provides (iii):

$$|Ax| \leq \nu_2(A)\|x\|_2 \Rightarrow \|Ax\|_2 \leq \|\nu_2(A)\|x\|_2\|_2 = \|\nu_2(A)\|_2 \|x\|_2$$

and thus

$$\sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} \leq \sup_{x \neq 0} \|\nu_2(A)\|_2 = \|\nu_2(A)\|_2.$$

Moreover, (iv) is equivalent to $\mu_2(A)^T \mu_2(A) = tr\left(A^T A\right)$ which follows from

$$\mu_2(A)^T \mu_2(A) = \sum_{j=1}^{n} \|A_{:j}\|_2^2 = \sum_{j=1}^{n} A_{:j}^T A_{:j} = tr\left(A^T A\right).$$

Using (iv) and Proposition 1.11 (i) we obtain (v):

$$\|\mu_2(A)\|_2 = \sqrt{tr\left(A^T A\right)} = \sqrt{tr\left(A A^T\right)} = \|\mu_2(A^T)\|_2 = \|\nu_2(A)\|_2.$$

Finally, for (vi) and (vii) we denote $A = (a_{ij})_{ij}$ and $B = (b_{ij})_{ij}$. Then

$$|A| \leq |B| \Rightarrow |a_{ij}| \leq |b_{ij}| \Rightarrow \sum_{j=1}^{n} a_{ij}^2 \leq \sum_{j=1}^{n} b_{ij}^2 \Rightarrow \nu_2(A)_i \leq \nu_2(B)_i,$$

and using again Proposition 1.11 (ii),

$$|A| \leq |B| \Rightarrow |A^T| \leq |B^T| \Rightarrow \nu_2(A^T) \leq \nu_2(B^T) \Rightarrow \mu_2(A) \leq \mu_2(B).$$

□

Remark 1.15. *Let $A \in \mathbb{R}^{m\times n}$. There are also similar inequalities which hold for differing but compatible norms, e.g.*

$$|Ax| \leq \nu_1(A)\|x\|_\infty.$$

Proof: Let $i \in \{1, \ldots, m\}$. Then $|Ax|_i = |A_{i:}x| \leq |A_{i:}||x| =$

$$\sum_{j=1}^{n} |a_{ij}||x_j| \leq \sum_{j=1}^{n} |a_{ij}|\|x\|_\infty = \|A_{i:}\|_1\|x\|_\infty = \nu_1(A)_i\|x\|_\infty.$$

□

1.3 Interval arithmetic

Since we are interested in bounding errors, we will consider *interval arithmetic*, which is a very useful concept for that problem. In the following we will give definitions, relations and results concerning interval arithmetic. Although there are many more, we will regard only those few that are used later on. This subsection is based upon Neumaier [3], notations also found in Neumaier, Domes [5].

Definition 1.16. *Let $\mathbb{IR}$ denote the set of all nonempty, connected and compact subsets $\mathbf{a}$ of $\mathbb{R}$ so that*

$$\mathbf{a} := [\underline{a}, \overline{a}] = \{a \in \mathbb{R} \mid \underline{a} \leq a \leq \overline{a}\},$$

for some $\underline{a}, \overline{a} \in \mathbb{R}$ with $\underline{a} \leq \overline{a}$, i.e., the closed intervals.
Analogously, let $\overline{\mathbb{IR}}$ define the set of all nonempty, connected and closed subsets of $\mathbb{R}$.

Definition 1.17. *Let* $\mathbf{a}_{ij} \in \mathbb{IR}$ *for* $(i,j) \in \{1,\ldots,m\} \times \{1,\ldots,n\}$. *Then we define an interval matrix* $\mathbf{A} \in \mathbb{IR}^{m\times n}$ *by*

$$\mathbf{A} := (\mathbf{a}_{ij})_{1\le i\le m, 1\le j\le n} = \{(\mathbf{a}_{ij})_{ij} \mid \underline{(a_{ij})} \le a_{ij} \le \overline{(a_{ij})}\} = \{A \mid \underline{A} \le A \le \overline{A}\}.$$

Definition 1.18. *Let* $\mathbf{a} = [\underline{a}, \overline{a}] \in \mathbb{IR}$. *Then we call*

$$\mathrm{wid}(\mathbf{a}) : \mathbb{IR} \to \mathbb{R}, \quad \mathrm{wid}(\mathbf{a}) := \overline{a} - \underline{a}$$

the width of $\mathbf{a}$. *This definition can be extended componentwise on interval matrices so that* wid*:* $\mathbb{IR}^{m\times n} \to \mathbb{R}^{m\times n}$, $\mathrm{wid}(\mathbf{A}) := (\mathrm{wid}(\mathbf{a}_{ij}))_{ij}$. *Similarly, we can define the radius of an interval matrix by*

$$\mathrm{rad}(\mathbf{A}) : \mathbb{IR}^{m\times n} \to \mathbb{R}^{m\times n}, \quad \mathrm{rad}(\mathbf{A}) := \frac{1}{2}(\overline{A} - \underline{A}) = \frac{1}{2}\mathrm{wid}(\mathbf{A}).$$

Now we can prove the following results:

Proposition 1.19. *Let* $\mathbf{a} = [\underline{a}, \overline{a}] \in \mathbb{IR}$. *Then*

$$a_1, a_2 \in \mathbf{a} \quad \Rightarrow \quad |a_1 - a_2| \le \mathrm{wid}(\mathbf{a}).$$

Proof: W.l.o.g. let $a_1 \ge a_2$. Then

$$|a_1 - a_2| = a_1 - a_2 \le \overline{a} - a_2 \le \overline{a} - \underline{a} = \mathrm{wid}(\mathbf{a}).$$

□

Theorem 1.20. *Let* $\mathbf{A} \in \mathbb{IR}^{m\times n}$. *Then*

$$B, C \in \mathbf{A} \quad \Rightarrow \quad |B - C| \leq \operatorname{wid}(\mathbf{A}).$$

Proof: For $(i, j) \in \{1, \ldots, m\} \times \{1, \ldots, n\}$ let $\mathbf{A} = (\mathbf{a}_{ij})_{ij}$, $B = (b_{ij})_{ij}$ and $C = (c_{ij})_{ij}$. Then for all (i, j) we have $b_{ij}, c_{ij} \in \mathbf{a}_{ij}$ and therefore by Proposition 1.19

$$|b_{ij} - c_{ij}| \leq \operatorname{wid}(\mathbf{a}_{ij}),$$

hence $|B - C| \leq \operatorname{wid}(\mathbf{A})$. □

In conclusion of this short insertion we consider the following definition: Since it is not possible to represent an irrational number on a computer, we usually calculate with approximations of real numbers in practise.

Definition 1.21. *Let* $\mathbb{M} \subseteq \overline{\mathbb{R}}$ *denote the machine representible numbers. For a real number* $x \in \mathbb{R}$, *a machine representible number* $m \in \mathbb{M}$, *obtained by some kind of rounding, is called floating point number of* x *and we write* $m = \mathrm{fl}(x)$. *Analogously, a floating point matrix* $\mathrm{fl}(A)$ *of a matrix* $A = (a_{ij})_{ij} \in \mathbb{R}^{m\times n}$ *is defined as* $\mathrm{fl}(A) := (\mathrm{fl}(a_{ij}))_{1\leq i\leq m, 1\leq j\leq n}$.
$|x-\mathrm{fl}(x)|$ *is called roundoff error (resp. rounding error).*

For example, we define the rounding modes *downward rounding* $\bigtriangledown$ and *upward rounding* $\bigtriangleup$ as functions $\bigtriangledown, \bigtriangleup : \overline{\mathbb{R}} \to \mathbb{M}$ with

$$\bigtriangledown x := \sup\{m \in \mathbb{M} : m \leq x\} \qquad \text{and} \qquad \bigtriangleup x := \inf\{m \in \mathbb{M} : m \geq x\}.$$

Remark 1.22. *Clearly, for* $\eta \in \mathbb{R}$ *an* $\tilde{\eta} \in \mathbb{M}$ *with* $\tilde{\eta} \approx \eta$ *and* $\tilde{\eta} \leq \eta$ *(resp.* $\tilde{\eta} \geq \eta$*) can be achieved by setting* $\tilde{\eta} := \bigtriangledown(\eta)$ *(resp.* $\tilde{\eta} := \bigtriangleup(\eta)$*).*

Furthermore, using the function *outward rounding*, defined by $\diamond : \mathbb{R} \to \mathbb{IR}$, $\diamond a := [\bigtriangledown a, \bigtriangleup a]$, one can represent a real number by an interval with machine representible bounds. Clearly, $a \in \diamond a$, i.e., the produced interval contains the real number. Now, calculations with intervals can be used to make error estimations. Therefore, we will need these functions resp. intervals at the implementation of our program (see Section 5).

For practical calculations we will use Intlab [9], a Matlab [8] toolbox. Intlab allows calculations with intervals and interval matrices, which will be necessary in the implementation of our concept. For built-in codes and general handling of Intlab, see Moore [6] resp. Hargreaves [7].

2 Bounds for overdetermined linear systems

In the following, we show that bounds for the solution of overdetermined linear systems of equations can be found, knowing bounds for the perturbations in the parameters. In addition to our main program, we will discuss a second, more general case. Since we work in exact arithmetic in this section, the results remain theoretical and can be considered as a basis for the next sections.

To recapitulate, we assume to know approximations A and b of parameters of an overdetermined system $\hat{A}\hat{x} = \hat{b}$. We suppose that the "underlying" system has at least one solution $\hat{x}$ and try to find a bound for the error $|\hat{x}-x^*|$, where $x^* = \arg\min \|Ax - b\|_2$. Moreover we suppose to know bounds for the errors $\hat{A} - A$ and $\hat{b} - b$, more precise for the norms of the columns of those, which can be described by hybrid norms. So we assume the following setting:

(A) Let $m \geq n$, $\hat{A}, A \in \mathbb{R}^{m\times n}$, $\hat{b}, b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$ satisfy $\mathrm{rk}(A) = n$,

$$\|\hat{A}_{:j} - A_{:j}\|_2 \leq c_j, \ \ (1 \leq j \leq n) \quad \Leftrightarrow \quad \mu_2(\hat{A} - A) \leq c \quad \text{and}$$
$$\|\hat{b} - b\|_2 \leq \beta \quad \Leftrightarrow \quad \mu_2(\hat{b} - b) \leq \beta.$$

Moreover, let x^* be the least squares solution of $\min \|Ax - b\|_2$, i.e.,

$$x^* = \arg\min_{x\in\mathbb{R}^n} \|Ax - b\|_2 \quad \Leftrightarrow \quad A^T Ax^* = A^T b.$$

(B) There exists a vector $\hat{x} \in \mathbb{R}^n$ which satisfies $\hat{A}\hat{x} = \hat{b}$.

Assuming exact arithmetic, we can find error bounds for this setting in Neumaier [4]:

Theorem 2.1. *Assume* (A) *and* (B) *and let* $A = QR$ *be a reduced QR-decomposition of* A. *We define*

$$r := Ax^* - b, \quad \rho := \|r\|_2, \quad \sigma := \beta + c^T|x^*| \quad \textit{and} \quad f := \nu_2(R^{-1}).$$

If $c^T f < 1$ *and* $\sigma^2 \geq \rho^2(1 - (c^T f)^2)$, *then for*

$$\gamma_B := \frac{\sigma c^T f + \sqrt{\sigma^2 - \rho^2\left(1 - (c^T f)^2\right)}}{1 - (c^T f)^2}$$

we have

$$|\hat{x} - x^*| \leq \gamma_B f.$$

Proof: Defining $\delta := \hat{x} - x^*$, we will first show

$$\|r + A\delta\|_2^2 = \|r\|_2^2 + \|R\delta\|_2^2. \tag{1}$$

$\|r + A\delta\|_2^2 = (r + A\delta)^T(r + A\delta) = (r^T + \delta^T A^T)(r + A\delta) = r^T r + 2\delta^T A^T r + \delta^T A^T A\delta$ and since $A^T r = A^T Ax^* - A^T b = 0$ and $A^T A = R^T Q^T QR = R^T R$, we obtain (1):

$$\|r + A\delta\|_2^2 = r^T r + \delta^T R^T R\delta = \|r\|_2^2 + \|R\delta\|_2^2.$$

Defining furthermore $\gamma := \|R\delta\|_2$ and the vector $d \in \mathbb{R}^n$ by

$$d := \hat{b} - b + (A - \hat{A})\hat{x},$$

we have

$$|\delta| = |R^{-1}R\delta| \leq \nu_2(R^{-1})\|R\delta\|_2 = \gamma f, \tag{2}$$

using Lemma 1.14 (i) and by $\hat{A}\hat{x} = \hat{b}$, Lemma 1.13 (i) and (2)

$\|d\|_2 = \|\hat{b} - b + (A - \hat{A})\hat{x}\|_2 \leq \|\hat{b} - b\|_2 + \|(A - \hat{A})\hat{x}\|_2 \leq \beta + \mu_2(A - \hat{A})^T|\hat{x}| \leq \beta + c^T|\hat{x}| = \beta + c^T|x^* + \delta| \leq \beta + c^T|x^*| + c^T|\delta| \leq \sigma + \gamma c^T f$ so that

$$\|d\|_2 \leq \sigma + \gamma c^T f \tag{3}$$

Therefore, we get the inequality

$$\rho^2 + \gamma^2 \leq (\sigma + \gamma c^T f)^2 \tag{4}$$

using (1) and (3):

$$\begin{aligned}\rho^2 + \gamma^2 &= \|r\|_2^2 + \|R\delta\|_2^2 = \|r + A\delta\|_2^2 = \|Ax^* - b + A(\hat{x} - x^*)\|_2^2 = \|A\hat{x} - b\|_2^2 \\ &= \|\hat{b} - b + A\hat{x} - \hat{A}\hat{x}\|_2^2 = \|d\|_2^2 \leq (\sigma + \gamma c^T f)^2.\end{aligned}$$

The relation (4) is equivalent to

$$(1 - (c^T f)^2)\gamma^2 - 2\sigma c^T f\gamma + \rho^2 - \sigma^2 \leq 0.$$

Now, if $h : \mathbb{R} \to \mathbb{R}$,

$$h(\gamma) := (1 - (c^T f)^2)\gamma^2 - 2\sigma c^T f\gamma + \rho^2 - \sigma^2,$$

then we need to find γ with $h(\gamma) \leq 0$. Computing the zeros of $h(\gamma)$, we obtain

$$\gamma_{1,2} = \frac{2\sigma c^T f \pm \sqrt{(-2\sigma c^T f)^2 - 4(1-(c^T f)^2)(\rho^2 - \sigma^2)}}{2(1-(c^T f)^2)}$$

$$= \frac{2\sigma c^T f \pm \sqrt{4\sigma^2 (c^T f)^2 - 4(\rho^2 - \rho^2 (c^T f)^2 - \sigma^2 + \sigma^2 (c^T f)^2)}}{2(1-(c^T f)^2)}$$

$$= \frac{\sigma c^T f \pm \sqrt{\sigma^2 - \rho^2 (1-(c^T f)^2)}}{1-(c^T f)^2}.$$

Thus, the solutions exist in $\mathbb{R}$ if $\sigma^2 \geq \rho^2(1-(c^T f)^2)$ and the largest zero of $h(\gamma)$ equals

$$(\gamma_1 =) \frac{\sigma c^T f + \sqrt{\sigma^2 - \rho^2 (1-(c^T f)^2)}}{1-(c^T f)^2} = \gamma_B.$$

Setting

$$\gamma_0 := \frac{\sigma c^T f}{1-(c^T f)^2},$$

then $c^T f < 1$ implies that $(\gamma_2 \leq) \gamma_0 \leq \gamma_B$, provided that $\sigma^2 \geq \rho^2(1-(c^T f)^2)$. Since

$$h(\gamma_0) = \frac{\sigma^2 (c^T f)^2}{1-(c^T f)^2} - \frac{2\sigma^2 (c^T f)^2}{1-(c^T f)^2} + \rho^2 - \sigma^2 = \frac{(1-(c^T f)^2)\rho^2 - \sigma^2}{1-(c^T f)^2} \leq 0,$$

we have

$$h(\gamma_0) \leq 0.$$

Therefore, we can conclude, that γ satisfies (4), if

$$\sigma^2 \geq \rho^2(1-(c^T f)^2) \quad \text{and} \quad (\gamma_2 \leq) \gamma \leq \gamma_B.$$

Hence, by (2) we find the bounds $|\hat{x} - x^*| \leq \gamma_B f$, which are sensible since $c^T f < 1$ guarantees $\gamma_B \geq 0$. □

The above setting (A) and (B) is mainly in our interest. Although, we can also compute bounds if we suppose instead of (B) the more general assumption (C):

(C) $\hat{x} \in \mathbb{R}^n$ is a solution of the least squares problem with parameters $\hat{A}$ and $\hat{b}$, i.e., $\hat{x}$ minimizes $\|\hat{A}x - \hat{b}\|_2$ over $x \in \mathbb{R}^n$.

Theorem 2.2. *Assume setting* (A) *and* (C) *and let* $A = QR$ *be a reduced QR-decomposition of* A. *We define (as before)*

$$r := Ax^* - b, \quad \rho := \|r\|_2, \quad \sigma := \beta + c^T|x^*|, \quad f := \nu_2(R^{-1})$$

and additionally

$$\omega := \frac{\beta + \|b\|_2}{1 - c^T f}, \qquad \tau := \||c^T|R^{-1}|\|_2.$$

If $c^T f < 1$, *then for*

$$\gamma_C := \frac{\sigma + \frac{\omega\tau^2}{2} + \tau\sqrt{\frac{\omega^2\tau^2}{4} + \rho^2 + \beta\rho + \omega\sigma}}{1 - c^T f}$$

we have

$$|\hat{x} - x^*| \leq \gamma_C f.$$

Proof: As in the proof of Theorem 2.1, let

$$\delta := \hat{x} - x^*, \quad \gamma := \|R\delta\|_2 \quad \text{and} \quad d := \hat{b} - b + (A - \hat{A})\hat{x},$$

such that (1)−(3) hold. Now in this setting the residual $\hat{r} := \hat{A}\hat{x} - \hat{b}$ can be nonzero. We only know $A^T r = 0$ and similarly $\hat{A}^T \hat{r} = \hat{A}^T \hat{A}\hat{x} - \hat{A}^T \hat{b} = 0$. By $A\delta = A(\hat{x} - x^*) = (A - \hat{A})\hat{x} + \hat{A}\hat{x} - Ax^* = (A - \hat{A})\hat{x} + \hat{r} + \hat{b} - r - b$ we obtain the equation

$$A\delta = d + \hat{r} - r, \tag{5}$$

which provides $R^T R\delta = R^T Q^T QR\delta = A^T A\delta = A^T d + A^T \hat{r} = A^T d + A^T \hat{r} - \hat{A}^T \hat{r} = R^T Q^T d - (\hat{A} - A)^T \hat{r}$ such that

$$R\delta = Q^T d - R^{-T}(\hat{A} - A)^T \hat{r}. \tag{6}$$

Taking norms in this equation yields $\|R\delta\|_2 = \|Q^T d + (-R^{-T}(\hat{A} - A)^T \hat{r})\|_2 \leq \|Q^T d\|_2 + \|R^{-T}(\hat{A} - A)^T \hat{r}\|_2 = \|Q^T d\|_2 + \|((\hat{A} - A)R^{-1})^T \hat{r}\|_2 \leq \|d\|_2 + \|(\hat{A} - A)R^{-1}\|_2 \|\hat{r}\|_2$. Using Lemma 1.13 (iv), we obtain

$$\|(\hat{A} - A)R^{-1}\|_2 \leq \|\mu_2(\hat{A} - A)^T |R^{-1}|\|_2 \leq \|c^T |R^{-1}|\|_2 = \tau. \tag{7}$$

Hence for $\gamma' := \|d\|_2 + \tau\|\hat{r}\|_2$ we have

$$\gamma \leq \gamma'. \tag{8}$$

Using (3) we obtain by (8) $\gamma' = \|d\|_2 + \tau\|\hat{r}\|_2 \leq \sigma + \gamma c^T f + \tau\|\hat{r}\|_2 \leq \sigma + \gamma' c^T f + \tau\|\hat{r}\|_2$, such that

$$\gamma' \leq \frac{\sigma + \tau\|\hat{r}\|_2}{1 - c^T f} \tag{9}$$

holds. Now from $\rho^2 = \|r\|_2^2 = r^T r = x^{*T} A^T r - b^T r = -b^T r$ and analogously

$\|\hat{r}\|_2^2 = \hat{r}^T\hat{r} = \hat{x^*}^T\hat{A}^T\hat{r} - \hat{b}^T\hat{r} = -\hat{b}^T\hat{r}$ follows that

$$\|\hat{r}\|_2^2 = -\hat{b}^T\hat{r} = -\hat{b}^T\hat{r} + \rho^2 + b^Tr + \hat{b}^Tr - \hat{b}^Tr = \rho^2 + (b-\hat{b})^Tr + \hat{b}^T(r-\hat{r}).$$

Using the Cauchy-Schwarz inequality, we have

$$(b-\hat{b})^Tr \leq \|b-\hat{b}\|_2\|r\|_2 \leq \beta\rho \quad \text{and} \quad \hat{b}^T(r-\hat{r}) \leq \|\hat{b}\|_2\|(r-\hat{r})\|_2,$$

whereby $\|\hat{r}\|_2^2 \leq \rho^2+\beta\rho+\|\hat{b}\|_2\|(r-\hat{r})\|_2$ and $\|\hat{b}\|_2 \leq \|\hat{b}-b\|_2+\|b\|_2 \leq \beta+\|b\|_2$, hence

$$\|\hat{r}\|_2^2 \leq \rho^2 + \beta\rho + (\beta + \|b\|_2)\|(r-\hat{r})\|_2. \tag{10}$$

The equations (5), (6) and (7) yield
$\|r-\hat{r}\|_2 = \|d-A\delta\|_2 = \|d-QR\delta\|_2 = \|d-Q(Q^Td - R^{-T}(\hat{A}-A)^T\hat{r})\|_2 =$
$\|QR^{-T}(\hat{A}-A)^T\hat{r}\|_2 \leq \|(\hat{A}-A)R^{-1}\|_2\|\hat{r}\|_2 \leq \tau\|\hat{r}\|_2 \leq \|d\|_2 + \tau\|\hat{r}\|_2 = \gamma'$.
Thus (10) and (9) provide $\|\hat{r}\|_2^2 \leq \rho^2+\beta\rho+(\beta+\|b\|_2)\|(r-\hat{r})\|_2 \leq \rho^2+\beta\rho+$
$(\beta+\|b\|_2)\gamma' \leq \rho^2+\beta\rho+(\beta+\|b\|_2)\frac{\sigma+\tau\|\hat{r}\|_2}{1-c^Tf} \leq \rho^2+\beta\rho+\omega(\sigma+\tau\|\hat{r}\|_2)$, which implies

$$\|\hat{r}\|_2^2 - \omega\tau\|\hat{r}\|_2 - (\rho^2+\beta\rho+\omega\sigma) \leq 0. \tag{11}$$

If we again define $h : \mathbb{R} \to \mathbb{R}$,

$$h(y) := y^2 - \omega\tau y - (\rho^2+\beta\rho+\omega\sigma)$$

then the zeros of $h(y)$ are

$$y_1 = \frac{\omega\tau}{2} + \sqrt{\frac{\omega^2\tau^2}{4} + \rho^2 + \beta\rho + \omega\sigma}, \quad y_2 = \frac{\omega\tau}{2} - \sqrt{\frac{\omega^2\tau^2}{4} + \rho^2 + \beta\rho + \omega\sigma}.$$

Since all terms in y_1 are non-negative we have that $y_2 \leq 0 \leq y_1$ and $h(0) = -(\rho^2 + \beta\rho + \omega\sigma) \leq 0$, so that $h(y) \leq 0$, if $\|\hat{r}\|_2 \leq y_1$, i.e., $\|\hat{r}\|_2$ satisfies (11), if

$$\|\hat{r}\|_2 \leq \frac{\omega\tau}{2} + \sqrt{\frac{\omega^2\tau^2}{4} + \rho^2 + \beta\rho + \omega\sigma}.$$

Inserting this in (9), then by (8) we have

$$(0 \leq)\, \gamma \leq \gamma' \leq \frac{\sigma + \frac{\omega\tau^2}{2} + \tau\sqrt{\frac{\omega^2\tau^2}{4} + \rho^2 + \beta\rho + \omega\sigma}}{1 - c^T f}.$$

Finally by (2) we obtain

$$|\hat{x} - x^*| \leq \gamma_C f.$$

□

The bounds obtained for the setting (A) and (C) are of course weaker than the bounds for (A) and (B) (if it is consistent, i.e., if $\sigma^2 \geq \rho^2(1 - (c^T f)^2)$):

Corollary 2.3. *Assume that* (A) *and* (B) *hold and let* ρ, σ *and* f *be defined as in Theorem 2.1. If* $c^T f < 1$ *and* $\sigma^2 \geq \rho^2(1 - (c^T f)^2)$ *hold, then* $\gamma_B \leq \gamma_C$.

Proof: Let ω and τ be defined as in Theorem 2.2. From $0 \leq c^T f < 1$ we get that $(c^T f)^2 < 1$ and therefore

$$1 - (c^T f)^2 > 0. \tag{12}$$

Hence $\rho^2(1 - (c^T f)^2) \geq 0$ which implies

$$0 \leq \sigma^2 - \rho^2(1 - (c^T f)^2) \leq \sigma^2 \Rightarrow \sqrt{\sigma^2 - \rho^2(1 - (c^T f)^2)} \leq \sigma$$

and therefore

$$\sigma c^T f + \sqrt{\sigma^2 - \rho^2(1-(c^T f)^2)} \leq \sigma + \sigma c^T f = \sigma(1 + c^T f).$$

By (12), we obtain

$$\frac{\sigma c^T f + \sqrt{\sigma^2 - \rho^2(1-(c^T f)^2)}}{1-(c^T f)^2} \leq \frac{\sigma(1+c^T f)}{1-(c^T f)^2} = \frac{\sigma}{1-c^T f}. \tag{13}$$

Obviously $\frac{\omega\tau^2}{2} + \tau\sqrt{\frac{\omega^2\tau^2}{4} + \rho^2 + \beta\rho + \omega\sigma} \geq 0$, whereby

$$\sigma \leq \sigma + \frac{\omega\tau^2}{2} + \tau\sqrt{\frac{\omega^2\tau^2}{4} + \rho^2 + \beta\rho + \omega\sigma}$$

and consequently

$$\frac{\sigma}{1-c^T f} \leq \frac{\sigma + \frac{\omega\tau^2}{2} + \tau\sqrt{\frac{\omega^2\tau^2}{4} + \rho^2 + \beta\rho + \omega\sigma}}{1-c^T f}.$$

Together with (13), this inequality provides

$$\frac{\sigma c^T f + \sqrt{\sigma^2 - \rho^2(1-(c^T f)^2)}}{1-(c^T f)^2} \leq \frac{\sigma + \frac{\omega\tau^2}{2} + \tau\sqrt{\frac{\omega^2\tau^2}{4} + \rho^2 + \beta\rho + \omega\sigma}}{1-c^T f},$$

hence $\gamma_B \leq \gamma_C$. □

For both settings, we had to assume the crucial property $c^T f < 1$ to compute bounds for $|\hat{x} - x|$. In the following subsection we can give a requirement on the one hand and we can find a consequence of this important property on the other hand.

2.1 The criterion $c^T f < 1$

We start with the following requirement: If the errors in A, i.e., the size of the entries of the matrix $|\hat{A} - A|$ are "small enough", then the criterion holds (and Theorem 2.2 applies). Lemma 2.4 gives a bound for the size of these errors:

Lemma 2.4. *Let $\hat{A}, A \in \mathbb{R}^{m \times n}$, $(a'_{kl})_{1 \le k \le m, 1 \le l \le n} := (\hat{A} - A)$, $A = QR$ be a reduced QR-factorization of A and $(r'_{ij})_{1 \le i,j \le n} := R^{-1}$. If*

$$a'_{kl} < (\sqrt{mn^3} \max_{1 \le i,j \le n} |r'_{ij}|)^{-1}$$

holds for all $k \in \{1, \ldots, m\}$, $l \in \{1, \ldots, n\}$ and $f := \nu_2(R^{-1})$, then there exists a vector $\tilde{c} \in \mathbb{R}^n$ with

$$\mu_2(\hat{A} - A) \le \tilde{c} \quad \text{and} \quad \tilde{c}^T f < 1.$$

Proof: Let

$$\bar{r} := \max_{1 \le i,j \le n} |r'_{ij}| \quad \text{and} \quad \bar{\mu} := \max_{1 \le j \le n} \mu_2(\hat{A} - A)_j.$$

Since for all $i \in \{1, \ldots, n\}$

$$\mu_2(\hat{A} - A)_i = \|\hat{A}_{:i} - A_{:i}\|_2 = \sqrt{\sum_{k=1}^{m} {a'_{ki}}^2} < \sqrt{\sum_{k=1}^{m} \frac{1}{mn^3 \bar{r}^2}} = \sqrt{\frac{1}{n^3 \bar{r}^2}} = \frac{1}{\sqrt{n^3} \bar{r}},$$

clearly

$$\bar{\mu} < \frac{1}{\sqrt{n^3} \bar{r}}.$$

Defining $\tilde{c} \in \mathbb{R}^n$ by $\tilde{c} := \bar{\mu} 1_n$, we have

$$\mu_2(\hat{A} - A) \leq \tilde{c} \quad \text{and} \quad \tilde{c}_i < \frac{1}{\sqrt{n^3}\bar{r}} \quad \forall i \in \{1, \ldots, n\}.$$

Finally

$$f_i = \sqrt{\sum_{j=1}^{n} {r'_{ij}}^2} \leq \sqrt{\sum_{j=1}^{n} \bar{r}^2} = \sqrt{n}\bar{r} \quad \forall i \in \{1, \ldots, n\}$$

implies

$$\tilde{c}^T f = \sum_{i=1}^{n} \tilde{c}_j f_j < \sum_{i=1}^{n} \frac{1}{\sqrt{n^3}\bar{r}} \sqrt{n}\bar{r} = \sum_{i=1}^{n} \frac{1}{n} = 1.$$

□

Furthermore, there is a consequence of the criterion:

Theorem 2.5. *Assume* (A) *and* (C), *let* $A = QR$ *be a reduced QR-decomposition of* A *and* $f := \nu_2(R^{-1})$. *If* $c^T f < 1$, *then* $\mathrm{rk}(\hat{A}) = n$.

Proof: Suppose $\mathrm{rk}(\hat{A}) < n$. Then there exists a vector $z \in \mathbb{R}^n$, $z \neq o_n$ with $\hat{A}z = o_m$. By (C) we have $\hat{A}^T\hat{A}\hat{x} = \hat{A}^T\hat{b}$ for a $\hat{x} \in \mathbb{R}^n$, hence $\hat{A}^T\hat{A}(\hat{x} + \lambda z) = \hat{A}^T\hat{b}$ for all $\lambda \in \mathbb{R}$. Now we can apply Theorem 2.2 for $\hat{x} + \lambda z$, which provides

$$|\hat{x} + \lambda z - x^*| \leq \gamma_C f \quad \forall \lambda \in \mathbb{R}.$$

Hence, for $\lambda \to \infty$ we must obtain $z = o_n$, a contradiction. □

Thus, by Theorem 2.5 we know, if $c^T f < 1$ and $\hat{b} \in \mathrm{im}(\hat{A})$ hold, there is a vector $\hat{x}$, satisfying $\hat{A}\hat{x} = \hat{b}$.

3 QR-decomposition

The QR-decomposition is a fundamental task in our topic. Up to now, we always assumed to have a factorization of the input matrix A. In this section we describe how to compute a QR-decomposition, using the Householder method, compare Stoer/Bulirsch [1].
At first we consider some definitions, necessary to formulate the method. In the second part of the section, we will modify the Householder method so that the produced reflections have a certain property. As we will see later on in Section 4, this property will be essential for controlling roundoff errors and evaluating enclosures for solutions of overdetermined systems in floating point arithmetic.

3.1 Householder method

There are different algorithms to compute a QR-decomposition of a matrix A. We will consider the *Householder method*, which iteratively uses orthogonal matrices, so-called *Householder reflections*, $P_1, \dots P_n$ to transform the matrix A into an upper triangular matrix $\widetilde{R}$, i.e., $P_n \cdots P_1 A = \widetilde{R}$.
Especially for this section, we consider the following definitions:

Definition 3.1. *For $a, e \in \mathbb{R}^m$ with $e \neq o_m$ and $\eta \in \mathbb{R}$ we define $\alpha \in \mathbb{R}$ by*

$$\alpha := \pm\sqrt{\frac{a^T a}{e^T e}} = \pm\frac{\|a\|_2}{\|e\|_2},$$

the vector $p \in \mathbb{R}^m$ by

$$p := a - \alpha e$$

and the functions

$$\phi : \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^{m\times m}, \quad \phi(a, e, \eta) := I_m - \eta p p^T$$

and for fixed $p \neq o_m$,

$$\Phi : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^{m\times m}, \quad \Phi(a, e) := \phi\left(a, e, \frac{2}{p^T p}\right).$$

The symbol $\pm$ *in the definition of* α *means that the sign of* α *is freely selectable.*

Theorem 3.2. *Let* $a, e \in \mathbb{R}^m$ *with* $e \neq o_m$, $p \neq o_m$ *and* $P := \Phi(a, e)$. *Then*

$$Pa = \alpha e.$$

Proof: For $z \in \mathbb{R}$, defined by $z := \|a\|_2^2 - \alpha e^T a$,

$$pp^T a = zp, \tag{14}$$

since $pp^T a = (a - \alpha e)(a - \alpha e)^T a = aa^T a - \alpha e a^T a - \alpha a e^T a + \alpha^2 e e^T a = \|a\|_2^2 a - \alpha \|a\|_2^2 e - \alpha e^T a a + \alpha^2 e^T a e = (\|a\|_2^2 - \alpha e^T a)(a - \alpha e) = zp$. Moreover, we have

$$p^T p = 2z, \tag{15}$$

by $p^T p = (a - \alpha e)^T (a - \alpha e) = a^T a - \alpha e^T a - \alpha a^T e + \alpha^2 e^T e = \|a\|_2^2 - 2\alpha e^T a + \alpha^2 \|e\|_2^2 = \|a\|_2^2 - 2\alpha e^T a + \frac{\|a\|_2^2}{\|e\|_2^2}\|e\|_2^2 = 2\|a\|_2^2 - 2\alpha e^T a = 2z$.

Inserting (14) and (15) into

$$Pa = (I_m - \frac{2}{p^T p} pp^T)a = a - \frac{2}{p^T p} pp^T a$$

yields

$$a - \frac{2}{2z} zp = a - p = a - a + \alpha e = \alpha e,$$

hence $Pa = \alpha e$. □

Proposition 3.3. *Let $a, e \in \mathbb{R}^m, \eta \in \mathbb{R}$ with $e \neq o_m$. Then the matrix $P := \phi(a, e, \eta)$ is symmetric.*

Proof: $P^T = \left(I_m - \eta pp^T\right)^T = I_m^T - \eta(pp^T)^T = I_m - \eta pp^T = P$. □

Proposition 3.4. *Let $a, e \in \mathbb{R}^m$ with $e \neq o_m$, $p \neq o_m$. Then the matrix $P := \Phi(a, e)$ is symmetric and orthogonal.*

Proof: $P^T = P$ follows directly from Proposition 3.3. Therefore $P^T = P^{-1}$ follows from:

$$PP^T = P^2 = \left(I_m - \frac{2}{p^T p} pp^T\right)\left(I_m - \frac{2}{p^T p} pp^T\right)$$

$$= I_m - \frac{4}{p^T p} pp^T + \frac{4}{(p^T p)^2} p(p^T p)p^T = I_m.$$

□

Proposition 3.5. *Let $k, m \in \mathbb{N}$ with $1 \leq k < m$ and $a, e \in \mathbb{R}^k$ with $e \neq o_k$, $p \neq o_k$. Then for $P := \Phi(a, e)$, the matrix $\tilde{P} \in \mathbb{R}^{m \times m}$, defined by*

$$\tilde{P} := \begin{pmatrix} I_{m-k} & 0_k \\ 0_k & P \end{pmatrix}$$

is symmetric and orthogonal.

Proof: By Proposition 3.3 we have

$$\tilde{P}^T = \begin{pmatrix} I_{m-k} & 0_k \\ 0_k & P \end{pmatrix}^T = \begin{pmatrix} I_{m-k} & 0_k \\ 0_k & P \end{pmatrix} = \tilde{P}.$$

Moreover

$$\tilde{P}\tilde{P}^T = \tilde{P}^2 = \begin{pmatrix} I_{m-k} & 0_k \\ 0_k & P \end{pmatrix} \begin{pmatrix} I_{m-k} & 0_k \\ 0_k & P \end{pmatrix} = \begin{pmatrix} I_{m-k} & 0_k \\ 0_k & I_k \end{pmatrix} = I_m.$$

□

Now we can formulate the Householder method, which computes a (full) QR-decomposition of A:

Theorem 3.6 (Householder Method)**.** *Let $A \in \mathbb{R}^{m\times n}$ and $m \geq n$. Then there exist symmetric and orthogonal matrices $\tilde{P}_i$, $i \in \{1, \ldots, n\}$ so that for $\widetilde{Q} \in \mathbb{R}^{m\times m}$ defined by*

$$\widetilde{Q} := \prod_{i=1}^{n} \tilde{P}_i$$

we have $A = \widetilde{Q}\widetilde{R}$ with $\widetilde{Q}$ orthogonal and $\widetilde{R} \in \mathbb{R}^{m\times n}$ an upper triangular matrix.

Proof: For $i \in \{1, \ldots, n\}$ we define

$$A^{(0)} := A \quad \text{and} \quad A^{(i)} := \tilde{P}_i A^{(i-1)}$$

and the vectors $a^{(i)} \in \mathbb{R}^{m-i+1}$ by

$$a^{(i)} := \begin{pmatrix} A_{ii}^{(i-1)} \\ \vdots \\ A_{mi}^{(i-1)} \end{pmatrix}.$$

Furthermore, if $p^{(i)} \in \mathbb{R}^{m-i+1}$ defined by

$$p^{(i)} := a^{(i)} - \alpha_i e_1^{(m-i+1)} \neq o_{m-i+1},$$

(where $\alpha_i = \pm\|a^{(i)}\|_2$,) we define the matrices $P_i \in \mathbb{R}^{m-i+1 \times m-i+1}$ by

$$P_i := \Phi(a^{(i)}, e_1^{(m-i+1)})$$

and $\tilde{P}_i \in \mathbb{R}^{m \times m}$ by

$$\tilde{P}_1 := P_1, \quad \tilde{P}_i := \begin{pmatrix} I_{i-1} & 0_{m-i+1} \\ 0_{m-i+1} & P_i \end{pmatrix} \quad \text{for} \quad i \in \{2, \ldots, n\}.$$

By Theorem 3.2, $P_i a^{(i)} = \alpha_i e_1^{(m-i+1)}$ holds. Thus

$$A^{(1)} = P_1 A = (P_1 A_{:1} | P_1 A_{:2} | \ldots | P_1 A_{:n}) = \begin{pmatrix} * & * & \cdots & * \\ 0 & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & * & \cdots & * \end{pmatrix},$$

$$A^{(i)} = \tilde{P}_i A^{(i-1)} = \left(\begin{array}{c|c} I_{i-1} & 0_{m-i+1} \\ \hline 0_{m-i+1} & P_i \end{array}\right) \left(\begin{array}{ccc|ccc} * & \cdots & * & * & \cdots & * \\ 0 & \ddots & \vdots & \vdots & & \vdots \\ \vdots & \ddots & * & * & \cdots & * \\ \hline \vdots & & 0 & * & \cdots & * \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & * & \cdots & * \end{array}\right)$$

$$= \left(\begin{array}{ccc|cccc} * & \cdots & * & * & \cdots & \cdots & * \\ 0 & \ddots & \vdots & \vdots & & & \vdots \\ \vdots & \ddots & * & * & \cdots & \cdots & * \\ \hline \vdots & & 0 & * & \cdots & \cdots & * \\ \vdots & & \vdots & 0 & * & \cdots & * \\ \vdots & & \vdots & \vdots & & & \vdots \\ 0 & \cdots & 0 & 0 & * & \cdots & * \end{array}\right) \quad \text{for } i \in \{2, \ldots, n-1\}$$

and

$$A^{(n)} = \tilde{P}_n A^{(n-1)} = \left(\begin{array}{c|c} I_{n-1} & 0_{m-n+1} \\ \hline 0_{m-n+1} & P_n \end{array}\right) \left(\begin{array}{ccc|c} * & \cdots & * & * \\ 0 & \ddots & \vdots & \vdots \\ \vdots & \ddots & * & * \\ \hline \vdots & & 0 & * \\ \vdots & & \vdots & \vdots \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & * \end{array}\right)$$

$$= \left(\begin{array}{ccc|c} * & \cdots & * & * \\ 0 & \ddots & \vdots & \vdots \\ \vdots & \ddots & * & * \\ \hline \vdots & & 0 & * \\ \vdots & & \vdots & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \end{array}\right) =: \widetilde{R}.$$

Now, if $p^{(i)} = a^{(i)} - \alpha_i e_1^{(m-i+1)} = o_{m-i+1}$, then

$$a_1^{(i)} = \alpha_i = \|a^{(i)}\|_2 \quad \text{and} \quad a_j^{(i)} = 0 \quad \text{for} \quad j \in \{i+1, \ldots, m\}.$$

But in that case, the ith column of $A^{(i-1)}$ has already the desired form (only zeros below the diagonal element) and we can proceed to the next step (optionally setting $\tilde{P}_i = I_m$). Therefore, we can always compute a decomposition (i.e., for arbitrary matrices A).

Hence $\tilde{P}_n \cdots \tilde{P}_1 A = \widetilde{R}$ with $\widetilde{R}$ an upper triangular matrix. By Proposition 3.5, we have that the matrices $\tilde{P}_i$ are symmetric and orthogonal and therefore

$$A = \tilde{P}_1^T \cdots \tilde{P}_n^T \tilde{P}_n \cdots \tilde{P}_1 A = \tilde{P}_1^T \cdots \tilde{P}_n^T \widetilde{R} = \tilde{P}_1 \cdots \tilde{P}_n \widetilde{R} = \widetilde{Q}\widetilde{R}$$

and

$$\widetilde{Q}^T \widetilde{Q} = (\tilde{P}_1 \cdots \tilde{P}_n)^T (\tilde{P}_1 \cdots \tilde{P}_n) = \tilde{P}_n^T \cdots \tilde{P}_1^T \tilde{P}_1 \cdots \tilde{P}_n = I_m,$$

implying $\widetilde{Q}$ to be orthogonal. □

Taking $Q \in \mathbb{R}^{m\times n}$ as the matrix which consists of the first n columns of $\widetilde{Q}$ and $R \in \mathbb{R}^{n\times n}$ as the matrix which remains by leaving out the last $m-n$ rows of $\widetilde{R}$, we get a reduced QR-decomposition $A = QR$ of A.

Remark 3.7. *Of course, for the computation of least squares solutions, the overdetermined case is important. However, for general $A \in \mathbb{R}^{m\times n}$ a (full) QR-decomposition of $A \in \mathbb{R}^{m\times n}$ can be computed by the Householder method in the same way as in the proof of Theorem 3.6, iterating for $i \in \{1, \ldots, \min\{m, n\}\}$.*

3.2 Directed QR-decomposition

Since roundoff errors occur in floating point calculations, the orthogonality of the matrices P_i can not be guaranteed and we generally obtain only an approximate factorization in practise. The errors in this factors of course affect the bounds, computed in Section 2. Thus, we have to do a modified construction of the reflections, which allows us to control these unavoidable errors:

3.2.1 Modified reflections

Theorem 3.8. *Let $a, e \in \mathbb{R}^m$ with $e \neq o_m$, $p \neq o_m$, $\tilde{\eta} \in \mathbb{R}$, $H := \phi(a, e, \tilde{\eta})$ and define $z \in \mathbb{R}$ by*

$$z := \|a\|_2^2 - \alpha e^T a.$$

If $\tilde{\eta} \leq \frac{2}{p^T p}$, so that there exists a $\delta \geq 0$ with $\tilde{\eta} + \delta = \frac{2}{p^T p}$ then

$$Ha = \alpha e + \delta p p^T a$$

and $\delta z \in [0,1]$ *imply*

$$\|H\|_2 \leq 1.$$

Proof. The first assertion follows from Theorem 3.2:

$$Ha = (I_m - \tilde{\eta} pp^T)a = a - \frac{2}{p^T p} pp^T a + \delta pp^T = \alpha e + \delta pp^T a.$$

Now, (14) implies

$$Ha = \alpha e + \delta pp^T a = \alpha e + \delta zp = \alpha e + \delta za - \delta z\alpha e = \delta za + (1 - \delta z)\alpha e,$$

and therefore

$$\begin{aligned}
\|H\|_2 &= \max_{\|a\|_2=1} \|Ha\|_2 = \max_{\|a\|_2=1} \|\delta za + (1-\delta z)\alpha e\|_2 \\
&\leq \max_{\|a\|_2=1} (\|\delta za\|_2 + \|(1-\delta z)\alpha e\|_2) \leq \max_{\|a\|_2=1} (|\delta z|\|a\|_2 + |1-\delta z||\alpha|\|e\|_2) \\
&\leq \max_{\|a\|_2=1} \left(|\delta z|\|a\|_2 + |1-\delta z| \frac{\|a\|_2}{\|e\|_2} \|e\|_2 \right) = |\delta z| + |1-\delta z|.
\end{aligned}$$

For $\delta z \in [0,1]$ we have

$$|\delta z| + |1 - \delta z| = \delta z + 1 - \delta z = 1$$

and thereby $\|H\|_2 \leq 1$. □

Corollary 3.9. *Let* $a \in \mathbb{R}^m$. *If* e *equals an unit vector* $e_k^{(m)}$, $(1 \leq k \leq m)$ *then*

$$z = \|a\|_2^2 - \alpha e^T a \geq 0.$$

Proof: By definition $\alpha = \pm\|a\|_2$. For all $i \in \{1, \ldots, m\}$

$$|a_i| \leq \sqrt{\sum_{i=1}^{n} a_i^2} = \|a\|_2 \Rightarrow \|a\|_2|a_i| \leq \|a\|_2^2,$$

hence $z = \|a\|_2^2 - \alpha e_k^T a = \|a\|_2^2 \mp \|a\|_2 a_k \geq \|a\|_2^2 - \|a\|_2|a_k| \geq 0.$ □

So, if we apply the Theorem 3.8 with $e = e_k^{(m)}$, where $k \in \{1, \ldots, m\}$, it is sufficient to assume $\delta z \leq 1$ instead of $\delta z \in [0, 1]$ to make sure that $\|H\|_2 \leq 1$, since for $\delta \geq 0$, we have always $\delta z \geq 0$ by Corollary 3.9.

3.2.2 Main property of modified reflections

In the following we present the important property of the modified reflections indicated in the beginning of this chapter. In Subsection 4.1 we will see how we can use it to control roundoff errors occuring during a QR-factorization.

Lemma 3.10. *Let $k, m \in \mathbb{N}$ with $1 \leq k < m$ and $H \in \mathbb{R}^{k\times k}$ with $\|H\|_2 \leq 1$. Then for $\widetilde{H} \in \mathbb{R}^{m\times m}$ defined by*

$$\widetilde{H} := \begin{pmatrix} I_{m-k} & 0_k \\ 0_k & H \end{pmatrix},$$

we have

$$\|\widetilde{H}\|_2 \leq 1.$$

Proof: Since $\|H\|_2 \leq 1$, we have for all that $w \in \mathbb{R}^k$

$$\|Hw\|_2 \leq \|H\|_2\|w\|_2 \leq \|w\|_2.$$

Thus, defining $\tilde{w} \in \mathbb{R}^k$ by $\tilde{w} := Hw$, we have $\|\tilde{w}\|_2 \leq \|w\|_2$ and thereby

$$\|\tilde{w}\|_2^2 \leq \|w\|_2^2. \tag{16}$$

Now, define

$$u \in \mathbb{R}^m \quad \text{by} \quad u := \begin{pmatrix} v \\ w \end{pmatrix}$$

for arbitrary $v \in \mathbb{R}^{m-k}$ and $w \in \mathbb{R}^k$. Then by (16)

$$\begin{aligned} \|\widetilde{H}\|_2 = \max_{\|u\|_2=1} \|\widetilde{H}u\|_2 = \max_{\|u\|_2=1} \left\| \begin{pmatrix} v \\ Hw \end{pmatrix} \right\|_2 &= \max_{\|u\|_2=1} \left\| \begin{pmatrix} v \\ \tilde{w} \end{pmatrix} \right\|_2 \\ = \max_{\|u\|_2=1} \sqrt{\sum_{i=1}^{m-k} v_i^2 + \sum_{i=1}^{k} \tilde{w}_i^2} &\leq \max_{\|u\|_2=1} \sqrt{\sum_{i=1}^{m-k} v_i^2 + \sum_{i=1}^{k} w_i^2} \\ = \max_{\|u\|_2=1} \|u\|_2 &= 1, \end{aligned}$$

hence $\|\widetilde{H}\|_2 \leq 1$. □

Now by Theorem 3.8, an $\tilde{\eta} \in \mathbb{R}$ with $\tilde{\eta} \approx \frac{2}{p^T p}$ and $\tilde{\eta} \leq \frac{2}{p^T p}$, implies $H \approx P$ for any $a, e \in \mathbb{R}^m$ (w.l.o.g. $p \neq o_m$), where

$$H := \phi(a, e, \tilde{\eta}) \quad \text{and} \quad P := \Phi(a, e),$$

but $\|H\|_2 \leq 1$! Thus, for a reflection $\widetilde{H}$ (obtained from H) we have also $\widetilde{H} \approx \tilde{P}$ with $\|\widetilde{H}\|_2 \leq 1$ by Lemma 3.10.

Applying the Householder method with modified reflections, we obtain a

factorization

$$\widetilde{H}_n \dots \widetilde{H}_1 A \approx \widetilde{R},$$

into an approximate upper triangular matrix and $\|\widetilde{H}_i\|_2 \leq 1$ for all $i \in \{1, \dots, n\}$. (In fact, $\widetilde{H}_n \cdots \widetilde{H}_1 A$ is not an upper triangular matrix, but the lower off-diagonal elements are of the size of roundoff errors.) From Proposition 3.3 it follows that also all H_i and therefore all $\widetilde{H}_i$ are symmetric.
Since all reflections $\widetilde{H}_i$ are required to satisfy $\|\widetilde{H}_i\|_2 \leq 1$ which can be assured by directed rounding of $\frac{2}{p^T p}$, it is reasonable to call a factorization of A, obtained by the Householder method which uses such modified reflections, a *directed QR-decomposition*.

4 Error control

Additionally to errors in the QR-decomposition, in practical calculations an error $|\tilde{P}A - \widetilde{H}A|$ can increase due to roundoff errors in computing a matrix product, which has the consequence that the bounds computed in Theorem 2.1 and Theorem 2.2 may be too optimistic, since this roundoff errors affect the QR-decomposition of A (and therefore also the vector $f = \nu_2(R^{-1})$). Assuming setting (A) and (B), our goal in this section is to find a bound for $|\hat{x} - x^*|$, which takes into account this additional error source (see Lemma 4.5). To provide the requirements of Lemma 4.5 we need several preparations:

By Theorem 1.20 it is possible to bound the error $|A - \mathrm{fl}(A)|$ by $\mathrm{wid}(\mathbf{A})$ for a given interval matrix $\mathbf{A}$ with $A, \mathrm{fl}(A) \in \mathbf{A}$. Relating to our topic, we consider such bounds for vectors:

Theorem 4.1. *Let $a_i \in \mathbb{R}^m$ and $\mathbf{a}_i \in \mathbb{IR}^m$ such that $a_i, \mathrm{fl}(a_i) \in \mathbf{a}_i$ for $i \in \{1, \ldots, n\}$. Then by Theorem 1.20 for $\omega_i' := \mathrm{wid}(\mathbf{a}_i) \in \mathbb{R}^m$ we have that*

$$|a_i - \mathrm{fl}(a_i)| \leq \omega_i'$$

and therefore $\omega_i := \|\omega_i'\|_2 \in \mathbb{R}$ satisfies

$$\|a_i - \mathrm{fl}(a_i)\|_2 \leq \omega_i.$$

Thus we can bound $\mu_2(A - \mathrm{fl}(A))$ for a matrix $A \in \mathbb{R}^{m \times n}$ with $A =$

$(a_1|\dots|a_n)$ by the vector $(\omega_i)_{1\leq i\leq n} =: \omega \in \mathbb{R}^n$ so that

$$\mu_2(A - \mathrm{fl}(A)) \leq \omega.$$

Theorem 4.1 will be important to control the error propagation in the QR-factorization of a matrix A in (A).

Furthermore, for an upper triangular matrix $R \in \mathbb{R}^{n\times n}$ we can bound $\nu_2(R^{-1})$ by some vector $u \in \mathbb{R}^n$: Theorem 4.2 states that the columns of R^{-1} can be obtained iteratively by backward substitution:

Theorem 4.2. *Let $R \in \mathbb{R}^{n\times n}$ be a regular upper triangular matrix. Then $R^{-1} = (x_1|\dots|x_n)$, where $x_k \in \mathbb{R}^n$ satisfies*

$$Rx_k = e_k^{(n)} \quad \text{for all} \quad k \in \{1,\dots,n\}.$$

Proof: Since R is regular, $\mathrm{rk}(R) = n$. Thereby all x_k are uniquely determined and $R(x_1|\dots|x_n) = (Rx_1|\dots|Rx_n) = (e_1^{(n)}|\dots|e_n^{(n)}) = I_n$. Hence $R^{-1} = (x_1|\dots|x_n)$. □

Remark 4.3. *For an upper triangular matrix $R = (r_{ij})_{ij} \in \mathbb{R}^{n\times n}$ and $e \in \mathbb{R}^n$, backwards substitution $Rx = e$ leads to the recursive formula:*

$$x_n = \frac{e_n}{r_{nn}}, \quad x_k = \frac{1}{r_{kk}}\left(e_k - \sum_{l=k+1}^{n} r_{kl}x_l\right) \text{ for } k = n-1,\dots,1.$$

Applying this formula iteratively to the columns of I_n provides the entries of R^{-1}. In this way, one can bound $\nu_2(R^{-1})$ using Intlab.

4.1 Bounds including roundoff errors

Suppose that a given interval matrix $\widetilde{\mathbf{H}}\mathbf{A}$ satisfies $\widetilde{H}A$, $\mathrm{fl}(\widetilde{H}A) \in \widetilde{\mathbf{H}}\mathbf{A}$ (this situation can be achieved using Intlab). Then we can apply Theorem 4.1 on the columns of A to obtain a vector c_r with $\mu_2(\widetilde{H}A - \mathrm{fl}(\widetilde{H}A)) \leq c_r$, a bound for the error generated by (matrix) calculations in floating point arithmetic. Now, the next theorem shows how to compute an overall error estimate for $\mu_2(\widetilde{H}\hat{A} - \mathrm{fl}(\widetilde{H}A))$, allowing for an initial error c ($\geq \mu_2(\hat{A} - A)$) and using c_r.

Theorem 4.4. *Let $\hat{A}, A \in \mathbb{R}^{m\times n}$, $\widetilde{H} \in \mathbb{R}^{m\times m}$ and $c, c_r \in \mathbb{R}^n$ such that $\|\widetilde{H}\|_2 \leq 1$,*

$$\mu_2(\hat{A} - A) \leq c \qquad \textit{and} \qquad \mu_2(\widetilde{H}A - \mathrm{fl}(\widetilde{H}A)) \leq c_r.$$

Then for $\tilde{c} \in \mathbb{R}^n$ defined by

$$\tilde{c} := c + c_r$$

we have

$$\mu_2(\widetilde{H}\hat{A} - \mathrm{fl}(\widetilde{H}A)) \leq \tilde{c}.$$

Proof: By Lemma 1.13 (ii)

$$\mu_2(\widetilde{H}(\hat{A} - A)) \leq \|\widetilde{H}\|_2 \mu_2(\hat{A} - A) \leq \mu_2(\hat{A} - A) \leq c$$

and therefore using Theorem 1.12 (iv) and (iii) we have $\mu_2(\widetilde{H}\hat{A} - \mathrm{fl}(\widetilde{H}A)) = \mu_2(\widetilde{H}\hat{A} - \widetilde{H}A + \widetilde{H}A - \mathrm{fl}(\widetilde{H}A)) \leq \mu_2(\widetilde{H}(\hat{A} - A)) + \mu_2(\widetilde{H}A - \mathrm{fl}(\widetilde{H}A)) \leq c + c_r = \tilde{c}$. □

Now it became clear why we requested the modified reflections $\widetilde{H}_i$ to satisfy $\|\widetilde{H}_i\|_2 \leq 1$ (see Lemma 3.10): Iterating Theorem 4.4 at the evaluation of a directed QR-decomposition (Section 3.2), we obtain a bound for $\mu_2(\widetilde{H}_n \cdots \widetilde{H}_1 \hat{A} - \widetilde{H}_n \cdots \widetilde{H}_1 A)$, a requirement of the main lemma of this section.

4.2 Evaluation of enclosures

Using Lemma 4.5, we will be able to enclose solutions of overdetermined systems of equations in floating point arithmetic. Even for arbitrary $x \in \mathbb{R}^n$, we can compute a bound for $|\hat{x} - x|$:

Lemma 4.5. *Assume setting* (A) *and* (B) *and* $Q \in \mathbb{R}^{m \times n}$, $R \in \mathbb{R}^{n \times n}$ *regular,* $r \in \mathbb{R}^m$, $\tilde{c} \in \mathbb{R}^n$ *and* $\tilde{\beta} \in \mathbb{R}$ *such that*

$$\mu_2(Q^T \hat{A} - R) \leq \tilde{c} \quad \text{and} \quad \mu_2(Q^T \hat{b} - r) \leq \tilde{\beta}$$

and $u \in \mathbb{R}^n$ *satisfies*

$$\nu_2(R^{-1}) \leq u.$$

If $\tilde{c}^T u < 1$, *then for* $\tilde{\gamma} : \mathbb{R}^n \to \mathbb{R}$ *defined by*

$$\tilde{\gamma}(x) := \frac{\tilde{c}^T |x| + \tilde{\beta} + \|r - Rx\|_2}{1 - \tilde{c}^T u}$$

we have that

$$|\hat{x} - x| \leq \tilde{\gamma}(x) u.$$

Proof: $\mu_2(Q^T \hat{A} - R) \leq \tilde{c}$ and $\mu_2(Q^T \hat{b} - r) \leq \tilde{\beta}$ can be written together as

$$\mu_2(Q^T(\hat{A}|\hat{b}) - (R|r)) \leq \begin{pmatrix} \tilde{c} \\ \tilde{\beta} \end{pmatrix} (\in \mathbb{R}^{n+1}).$$

Since (B) holds, we have $\hat{A}\hat{x} = \hat{b}$ which implies $\hat{A}\hat{x} - \hat{b} = o_m$ and

$$(\hat{A}|\hat{b}) \begin{pmatrix} \hat{x} \\ -1 \end{pmatrix} = o_m \quad \text{and therefore} \quad Q^T(\hat{A}|\hat{b}) \begin{pmatrix} \hat{x} \\ -1 \end{pmatrix} = o_m.$$

Thus by Lemma 1.13 (i)

$$\left\| (R|r) \begin{pmatrix} \hat{x} \\ -1 \end{pmatrix} \right\|_2 = \left\| Q^T((\hat{A}|\hat{b}) - (R|r)) \begin{pmatrix} \hat{x} \\ -1 \end{pmatrix} \right\|_2 \leq (\tilde{c}^T|\tilde{\beta}) \begin{pmatrix} |\hat{x}| \\ 1 \end{pmatrix},$$

which yields

$$\|R\hat{x} - r\|_2 \leq \tilde{c}^T|\hat{x}| + \tilde{\beta}. \tag{17}$$

Now by Lemma 1.13 (iii), every $x \in \mathbb{R}^n$ satisfies $|\hat{x} - x| \leq |R^{-1}R(\hat{x} - x)| \leq \nu_2(R^{-1})\|R\hat{x} - Rx\|_2 = \nu_2(R^{-1})\|R\hat{x} - r + r - Rx\|_2 \leq \nu_2(R^{-1})(\|R\hat{x} - r\|_2 + \|r - Rx\|_2)$, hence by (17), $|\hat{x} - x| \leq \nu_2(R^{-1})(\tilde{c}^T|\hat{x}| + \tilde{\beta} + \|r - Rx\|_2)$ and therefore

$$|\hat{x} - x| \leq u(\tilde{c}^T|\hat{x}| + \tilde{\beta} + \|r - Rx\|_2). \tag{18}$$

This inequality implies

$$|\hat{x}| - |x| \leq |\hat{x} - x| \leq u(\tilde{c}^T|\hat{x}| + \tilde{\beta} + \|r - Rx\|_2),$$

whereby

$$|\hat{x}| - u(\tilde{c}^T|\hat{x}| + \tilde{\beta} + \|r - Rx\|_2) \leq |x|$$

and since $\tilde{c} \geq o_n$, we obtain

$$\tilde{c}^T|\hat{x}| - \tilde{c}^T u(\tilde{c}^T|\hat{x}| + \tilde{\beta} + \|r - Rx\|_2) \leq \tilde{c}^T|x|.$$

Adding $\tilde{\beta} + \|r - Rx\|_2$ yields

$$\tilde{c}^T|\hat{x}| + \tilde{\beta} + \|r - Rx\|_2 - \tilde{c}^T u(\tilde{c}^T|\hat{x}| + \tilde{\beta} + \|r - Rx\|_2) \leq \tilde{c}^T|x| + \tilde{\beta} + \|r - Rx\|_2$$

so that

$$(1 - \tilde{c}^T u)(\tilde{c}^T|\hat{x}| + \tilde{\beta} + \|r - Rx\|_2) \leq \tilde{c}^T|x| + \tilde{\beta} + \|r - Rx\|_2.$$

Because of $\tilde{c}^T u < 1$ we have

$$\tilde{c}^T|\hat{x}| + \tilde{\beta} + \|r - Rx\|_2 \leq \tilde{\gamma}(x) \tag{19}$$

and finally by (18) and (19)

$$|\hat{x} - x| \leq \tilde{\gamma}(x)u.$$

□

The following corollary is a consequence of Lemma 4.5:

Corollary 4.6. *Assume setting* (A) *and* (B). *If $A = QR$ is a reduced QR-decomposition of A, then define $r := Q^T b$. If $c^T u < 1$, for a vector $u \in \mathbb{R}^n$*

with $\nu_2(R^{-1}) \leq u$, *then for* $\gamma : \mathbb{R}^n \to \mathbb{R}$ *defined by*

$$\gamma(x) := \frac{c^T|x| + \beta + \|r - Rx\|_2}{1 - c^T u}$$

we have that

$$|\hat{x} - x| \leq \gamma(x)u.$$

Proof: By Lemma 1.13 (ii) we find
$\mu_2(Q^T\hat{A} - R) = \mu_2(Q^T(\hat{A} - A)) \leq \|Q^T\|_2\mu_2(\hat{A} - A) \leq c$ and $\mu_2(Q^T\hat{b} - r) \leq \mu_2(Q^T(\hat{b} - b)) \leq \|Q^T\|_2\mu_2(\hat{b} - b) \leq \beta$.
Therefore, we can apply Lemma 4.5 with $\tilde{c} = c$ and $\tilde{\beta} = \beta$ and obtain

$$|\hat{x} - x| \leq \gamma(x)u, \qquad \text{where} \qquad \gamma(x) := \frac{c^T|x| + \beta + \|r - Rx\|_2}{1 - c^T u}.$$

□

Remark 4.7. *Of course, the most interesting case is* $x = x^*$, *where* x^* *is the least squares solution of* $Ax = b$. *E.g., in Corollary 4.6,* x^* *produces the bound*

$$|\hat{x} - x^*| \leq \gamma(x^*)u, \quad \textit{where} \quad \gamma(x^*) = \frac{c^T|x^*| + \beta}{1 - c^T u},$$

since x^* *satisfies* $Rx^* = Q^T b\,(= r)$.

The bound $\tilde{\gamma}(x)$ obtained by Lemma 4.5 is generally weaker than γ_B, obtained from Theorem 2.1, but it has a crucial advantage.

Theorem 4.8. *Assume* (A) *and* (B) *hold,* $A = QR$ *is a reduced QR-decomposition of* A *and let* ρ, σ *and* f *be defined as in Theorem 2.1. If* $\sigma^2 \geq \rho^2(1 - (c^T f)^2)$ *and* $c^T u < 1$ *for a vector* $u \geq \nu_2(R^{-1})$, *then* $\gamma_B \leq \gamma(x^*)$.

Proof: We can use (13) from the proof of Corollary 2.3 which states:

$$\gamma_B = \frac{\sigma c^T f + \sqrt{\sigma^2 - \rho^2(1-(c^T f)^2)}}{1-(c^T f)^2} \leq \frac{\sigma}{1-c^T f}.$$

Since $o_n \leq f = \nu_2(R^{-1}) \leq u$ and $c \geq o_n$ imply $c^T f \leq c^T u < 1$, we have

$$0 \leq 1 - c^T u \leq 1 - c^T f,$$

providing

$$\frac{\sigma}{1-c^T f} \leq \frac{\sigma}{1-c^T u} = \frac{c^T|x^*| + \beta}{1-c^T u} = \gamma(x^*).$$

Hence $\gamma_B \leq \gamma(x^*)$. □

As mentioned above, the reason why we consider Lemma 4.5 is that it has an essential benefit, namely it produces bounds for arbitrary vectors x (and matrices R, satisfying the requirement). Therefore, we can apply the lemma with

$$R = R_f := \mathrm{fl}(Q^T A), \quad r = r_f := \mathrm{fl}(Q^T b), \quad \text{and} \quad x = x_f^* := \mathrm{fl}(x^*).$$

If there exists a vector $\tilde{c} \in \mathbb{R}^m$ and a $\tilde{\beta} \in \mathbb{R}$ with

$$\mu_2(Q^T \hat{A} - R_f) \leq \tilde{c} \quad \text{and} \quad \mu_2(Q^T \hat{b} - r_f) \leq \tilde{\beta}$$

and $\tilde{c}^T u < 1$ for $u \geq \nu_2(R_f^{-1})$, then we obtain a bound

$$|\hat{x} - x_f^*| \leq \tilde{\gamma}_f u,$$

where

$$\tilde{\gamma}_f := \tilde{\gamma}(x_f^*) = \frac{\tilde{c}^T|x_f^*| + \tilde{\beta} + \|r_f - R_f x_f^*\|_2}{1 - \tilde{c}^T u}.$$

How to compute such quantities $\tilde{c}, \tilde{\beta}$ and u for given A, b and c, β with $\mu_2(\hat{A} - A) \leq c$ and $\mu_2(\hat{b} - b) \leq \beta$ has been described before in Theorem 4.4. Furthermore u can be computed since $\mathrm{rk}(A) = n \Rightarrow \mathrm{rk}(R) = n$, by Proposition 1.9.

Remark 4.9. *By the same assumptions as in Lemma 4.5, one can show that*

$$\|R(\hat{x} - x)\|_2 \leq \tilde{\gamma}(x)$$

holds for any $x \in \mathbb{R}^n$.

Proof: Making use of (17) and (19) provides $\|R(\hat{x} - x)\|_2 \leq \|R\hat{x} - r\|_2 + \|r - Rx\|_2 \leq \tilde{c}^T|\hat{x}| + \tilde{\beta} + \|r - Rx\|_2 \leq \tilde{\gamma}(x)$. □

5 Implementations

This section contains programs written in Matlab- resp. Intlab code, which perform the basic concept of the previous section(s).

For simplicity, we assume in the following algorithms, that the command 'intval' produces *rigorous* intervals. In fact, that can be wrong in practise (e.g. for numbers which are not expressible in binary floating point arithmetic), see Hargreaves [7], p. 8.

Our goal is to produce a method, based on Lemma 4.5. In preparation for that, we start with two algorithms used later: The first program performs backward substitution for a given regular upper triangular matrix $R \in \mathbb{R}^{n\times n}$ and a given vector $b \in \mathbb{R}^n$, using the formula of Remark 4.3., i.e., it computes a vector $x \in \mathbb{R}^n$ with $Rx = b$:

Algorithm 5.1 Backward substitution

```
1 function [x]=bwsub(R,b)
2 [~,n]=size(R);
3 x=zeros(n,1);
4 for j=n:-1:1
5     x(j,1)=(b(j,1)-R(j,:)*x)/R(j,j);
6 end
7 end
```

Another task in producing enclosures is the computation of a vector $u \in \mathbb{R}^n$ with $\nu_2(R^{-1}) \leq u$ for an upper triangular matrix $R \in \mathbb{R}^{n\times n}$, as described in Section 4. Such a vector can be computed by an interval matrix $\mathbf{S}$ which contains R^{-1}: Taking an $S \in \mathbf{S}$, with $|R^{-1}| \leq |S|$, then by Lemma 1.14 (vi) we have $\nu_2(S) \geq \nu_2(R^{-1})$. A matrix S with this property

has to exist always. In our program, we provide the bound for $\nu_2(R^{-1})$ as the componentwise supremum of $\nu_2(\mathbf{S})$. (For definitions and details, see Neumaier [3].)

The next algorithm, based on backward substitution and Theorem 4.2, provides such an interval matrix $\mathbf{S}$:

Algorithm 5.2 Computing an interval matrix containing R^{-1}

```
function [S]=invbws(R)
[~,n]=size(R);
I=eye(n);
x=intval(zeros(n,1));
for i=n:-1:1
    for j=n:-1:1
        x(j,1)=intval((I(j,i)-R(j,:)*x)/R(j,j));
    end
    S(:,i)=x;
    x=intval(zeros(n,1));
end
end
```

Now assume the setting (A) and (B). Then the following algorithm takes a matrix $A \in \mathbb{R}^{m\times n}$, vectors $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ and $\beta \in \mathbb{R}$ as initial values and produces an interval vector containing $\hat{x}$:

Algorithm 5.3 Computing an interval vector $\mathbf{x}$ with $\hat{x} \in \mathbf{x}$

```
function [box]=encl(A,b,c,beta)
[m,n]=size(A);
for i=1:n
    a=A(i:m,i);
    z=sup(norm(a)*(norm(a)-a(1,1)));
    if eps*z>1
```

```
7           error('Bad input')
8       else
9           w=a+intval(sign(a(1,1))*norm(a)*eye(m-i+1,1));
10          p=mid(w);
11          e=w'*w;
12          e=2/e;
13          M=A(i:m,i:n)-e*p*(p'*A(i:m,i:n));
14          A(i:m,i:n)=sup(M);
15          q=b(i:m,1)-e*p*(p'*b(i:m,1));
16          b(i:m,1)=sup(q);
17          M=2*rad(M);
18          k=n-i+1;
19          y=zeros(k,1);
20          y(1:k,1)=sup(sqrt(sum(M(:,1:k).^2)));
21          c(i:n,1)=c(i:n,1)+y;
22          q=2*rad(q);
23          q=sup(norm(q));
24          beta=beta+q;
25      end
26  end
27  S=invbws(A(1:n,1:n));
28  u=zeros(n,1);
29  for j=1:n
30      u(j,1)=sup(norm(S(j,:)));
31  end
32  if c'*u>=1
33      error('Bad bound(s) c or u')
34  end
35  x=bwsub(A(1:n,1:n),b);
36  g=sup((c'*abs(x)+beta+norm(b-A*x))/(1-c'*u));
37  box=midrad(x,g*u);
38  end
```

For better comprehension, we will give a short explanation of the particular parts of the algorithm:

5–7 Based on Theorem 3.8 resp. Corollary 3.9, the necessity of the relation $\delta z \leq 1$ is checked here.

9–16 In these lines the directed QR-decomposition of the matrix A is computed, so that we are able to generate the system $R_f x^* = r_f$.

17–24 This part performs the idea of Theorem 4.4, i.e., it controls the roundoff errors during the QR-decomposition. Allowing for bounds c and β for the initial errors $\mu_2(\hat{A} - A)$ and $\mu_2(\hat{b} - b)$, it produces the required bounds $\tilde{c}$ and $\tilde{\beta}$ for $\mu_2(Q^T\hat{A} - R_f)$ and $\mu_2(Q^T\hat{b} - r_f)$.

27–31 Using Algorithm 5.2., the evaluation of a bound u for the vector $\nu_2(R^{-1})$ is done here.

32–34 Since all requirements of Lemma 4.5 are provided, the only thing which has to be checked, is the criterion $\tilde{c}^T u < 1$.

35 If the criterion holds, x_f^* can be computed by backward substitution since A was transformed into an upper triangular matrix.

36 Now, making use of the floationg point solution x_f^*, we can simply evaluate the number $\tilde{\gamma}_f$. This number is rounded upwards, to make sure to have rigorous bounds.

37 Finally in this line, an enclosure for the solution $\hat{x}$ which is assumed to exist, is produced and denoted by box.

Moreover for comparison, we consider an implementation of the method, which solves a least squares problem using the Householder method. The output of the following algorithm is a floating point vector of the least squares solution.

Algorithm 5.4 Computing a least squares solution by Householder method

```
function [x]=househ(A,b)
[m,n]=size(A);
for i=1:min(m,n)
    a=A(i:m,i);
    z=sign(a(1,1))*norm(a)*eye(m-i+1,1);
    v=z+a;
    v=v/norm(v);
    A(i:m,i:n)=A(i:m,i:n)-2*v*(v'*A(i:m,i:n));
    b(i:m,1)=b(i:m,1)-2*v*(v'*b(i:m,1));
end
x=bwsub(A(1:n,1:n),b);
```

Since it is not necessary to compute the ith column of A, this algorithm can be implemented in a way so that it runs faster. But because we want to compare it with our method (Algorithm 5.3), we have to construct both programs similarly.

Among others, in the next chapter we will compare the Algorithms 5.3 and 5.4.

6 Numerical Tests

In this section we will do several classes of tests for Algorithm 5.3 (and Algorithm 5.4) and note important properties and results. The main characteristics we test for these algorithms are running time on the one hand and overestimation of the results on the other hand. This overestimation is measured by the 2-norm of the radius of the enclosures. The two measurements should be considered, depending on the dimension of the input paramters and the size of the perturbations. Another important property of the input data, which should not be omitted, is the condition of the matrix A.
We will assure a well-conditioned matrix using the built-in QR-decomposition of Matlab:

```
function [A]=condl(m,n)
B=2*rand(m,n)-1;
C=2*rand(n,n)-1;
[B ~]=qr(B);
[C ~]=qr(C);
A=B(1:m,1:n)*C;
end
```

The output of this algorithm is a matrix A of dimension $m \times n$, which can be used as input for our method. Since both factors of A have ("almost") orthonormal columns, we obtain $\text{cond}(A) \approx 1$. Converesly, we produce ill-conitioned matrices of size $m \times n$ by increasing their largest singular value:

```
function [A]=condh(m,n,l)
A=2*rand(m,n)-1;
[U S V]=svd(A);
S(1,1)=10^l;
A=U*S*V';
end
```

By Corollary 3.9, Algorithm 5.3 can only be executed if $\delta z \leq 1$ holds. To see, how far we can increase the condition of a matrix such that this relation holds, we consider the following test, which uses the function condh and counts the number of times the inequality is violated, depending on the parameter l in condh (but independent from the dimension of the matrix). The parameter t determines the number of repetitions of the test in which the dimension of the testmatrices runs up to 200.

```
function [u] = testcond(t)
u=zeros(12,1);
s=0;
for i=1:t
    for k=1:200
        for j=1:k
            for l=1:12
                v=0;
                A=condh(k,j,l);
                for m=1:j
                    if v==0
                        a=A(m:k,m);
                        z=norm(a)*(norm(a)-a(1,1));
                        if eps*z>1
                        u(l,1)=u(l,1)+1;
                        v=1;
                        end
                    end
                end
            end
            s=s+1;
        end
    end
end
u=u./s;
plot(u)
end
```

The number of violations for matrices, generated by condh(_,_,l), is stored in the lth entry of the vector u. Since each entry is divided by s (the number of tests), the output vector contains the percentage of failures of the inequality. Figure 1 shows the increament of failures of the inequality for increasing highest singular value (10^l, where $1 \leq l \leq 12$) of the testmatrices, in a logarithmic scale.

```
>> testcond(100)

ans =

         0
         0
         0
         0
         0
         0
         0
    0.1951
    0.9993
    0.9994
    0.9995
    0.9993
```

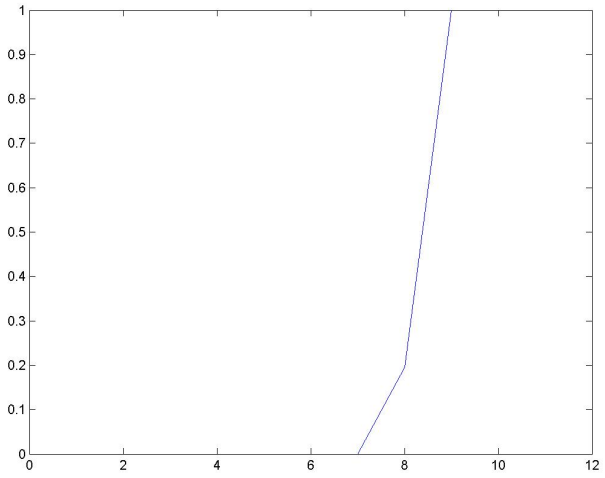

Figure 1: testcond(100)

Therefore, after making sure that $\delta z \leq 1$ holds for matrices up to condh(_,_,7), whose condition may exceed 10^9, we will only use at most $l = 7$ for condh, if this function is used in the following tests. Otherwise, we likely cannot use Algorithm 5.3.

As mentioned in the first section, the most interesting case in applications is that $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ are perturbated parameters of a system $\hat{A}\hat{x} = \hat{b}$.

If the accuracy of the measurements ϵ is known, e.g. $\epsilon = 10^{-k}$ for a $k \in \mathbb{N}$, then for each row of the matrix $|\hat{A} - A|$, we have $|(\hat{A} - A)_{:j}| \leq \epsilon 1_m$, whereby

$$\|(\hat{A} - A)_{:j}\|_2 \leq \sqrt{m}\epsilon.$$

Hence, $c \in \mathbb{R}^n$ defined by $c := 1_n\sqrt{m}\epsilon$, satisfies

$$\mu_2(\hat{A} - A) \leq c.$$

Analogously, $\mu_2(\hat{b} - b) \leq \beta$ holds for $\beta := \sqrt{m}\epsilon$. In this way, we can translate a given accuracy into bounds c and β for the hybrid norms in the following tests.

6.1 Tests

We start with a test that plots the running time of Algorithm 5.3 for random parameters A and b, depending on the dimensions of those. The threee input numbers are a given maximal (row-)dimension m, the accuracy $\epsilon = 10^{-\mathrm{ep}}$ ($\mathrm{ep} \in \mathbb{N}$) of the parameters and the number l of repetitions of the test:

```
function []=test1a(m,ep,l)
ep=10^-ep;
u=zeros(m,m);
s=zeros(m,1);
for z=1:l
    for k=1:m
        for j=1:k
            beta=ep*sqrt(k);
            A=condl(k,j);
            x=2*rand(j,1)-1;
            b=A*x;
```

```
12                  A=A+(2*rand(k,j)-1)*ep;
13                  b=b+(2*rand(k,1)-1)*ep;
14                  c=ones(j,1)*beta;
15                  tic;
16                  encl(A,b,c,beta);
17                  u(j,k)=u(j,k)+toc;
18              end
19              s(k,1)=k;
20          end
21  end
22  u=u./l;
23  surf(s,s,u)
24  xlabel('m');
25  ylabel('n');
26  end
```

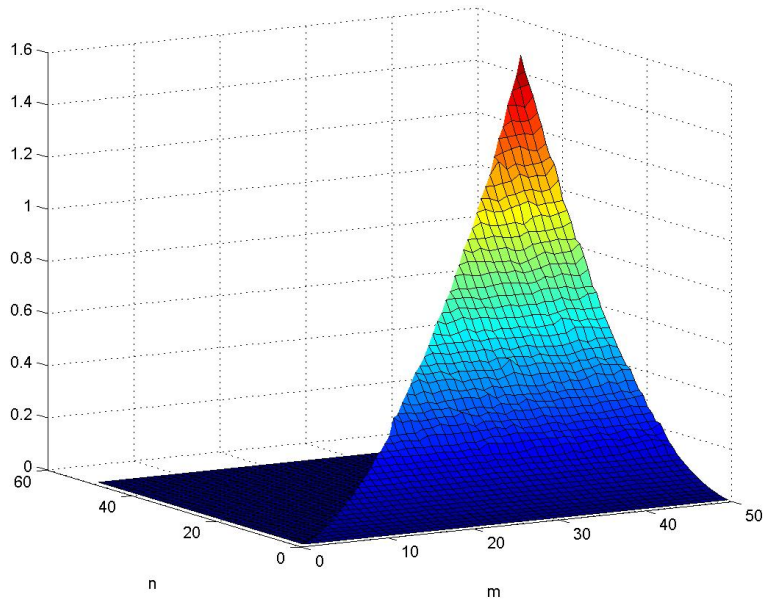

Figure 2: test1a(50,10,5)

As expected, at the part where $m \geq n$, we can see the cubic trend of the running time for increasing parameter n, due to the (directed) QR-decomposition. The lefthand side in Figure 2 remains flat, because we consider only overdetermined systems in this thesis. Hence, there is produced nothing if $m < n$ for $A \in \mathbb{R}^{m \times n}$.

Since the number of operations executed by a computer does not depend on the condition of the input matrix of Algorithm 5.3, of course the running time of the program does not change, applying test1a with the function condh, i.e., for ill-conditioned matrices. The same holds for Algorithm 5.4.

As announced above, we also consider the running time of this method, which computes a floating point least squares solution:

```
function []=test2(m,l)
u=zeros(m,m);
s=zeros(m,1);
for i=1:l
    for k=1:m
        for j=1:k
        A=condl(k,j);
        x=2*rand(j,1)-1;
        b=A*x;
        tic;
        househ(A,b);
        u(j,k)=u(j,k)+toc;
        end
        s(k,1)=k;
    end
end
u=u./l;
surf(s,s,u)
xlabel('m');
ylabel('n');
end
```

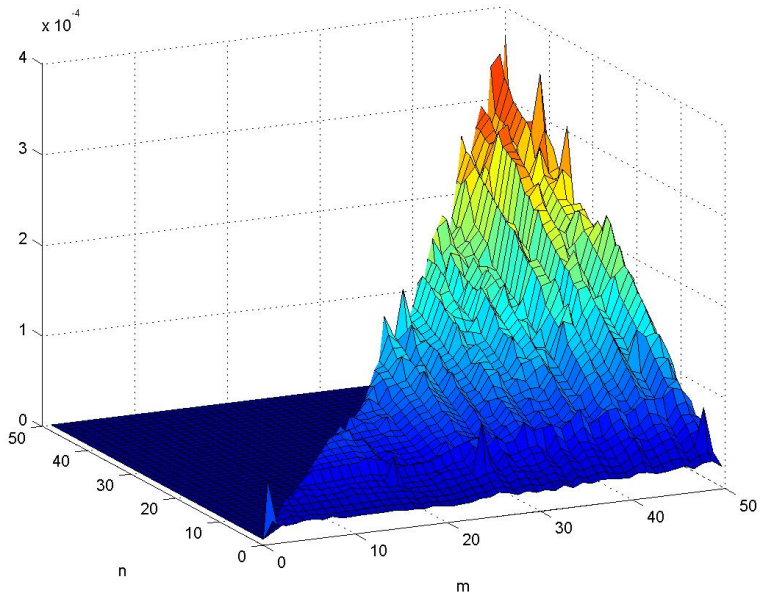

Figure 3: test2(50,10)

As before, Figure 3 is now the result of test2, which maps the running time of Algorithm 5.4, depending on the dimensions of parameter A. Clearly, this algorithm runs much faster than Algorithm 5.3, since it requires less operations.

The program test3(m,ep,k) shows the proportion of the running times of both methods up to dimension m, repeating the test k times. (The parameter ep does not have an influence on this test, but is a necessary input for our mehtod.)

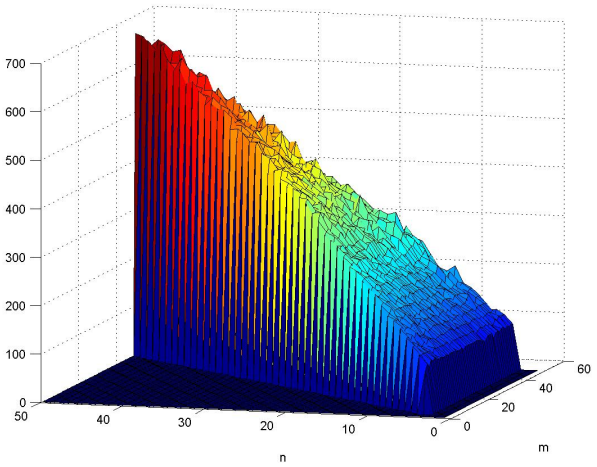

Figure 4: test3(50,10,3)

As we see in Figure 4, the proportion of the algorithms is linear in n, causing additional operations for error control in Algorithm 5.3, line 20 (and line 23). Repeating this test with the alternation of Algorithm 5.3 to omit these lines, the resulting figure shows the remaining factor:

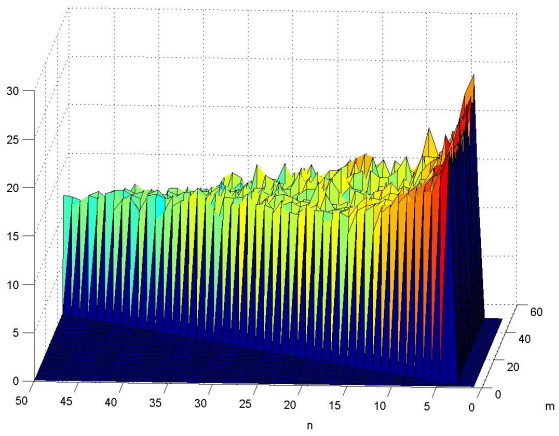

Figure 5: test3(50,10,3)

From the Figures 4 and 5 we now can conclude that the proportion of the algorithms is asymptotically

$$\sim 13n.$$

The next function, test1b, is construced similarly to test1a and tests the criterion $\tilde{c}^T u < 1$. To be able to make conclusions, the running time of Algorithm 5.3 is only mapped if the criterion holds. This test should depend on dimension, accuracy and condition of the input matrix of the least squares problem. Therefore we start with test1b(m,ep) for well-conditioned matrices, where the input numbers are m and ep, denoting dimension and accuracy once again.

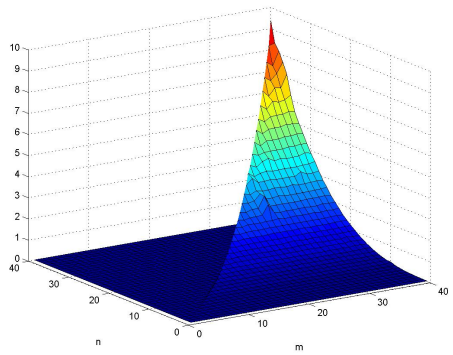

Figure 6: test1b(40,4)

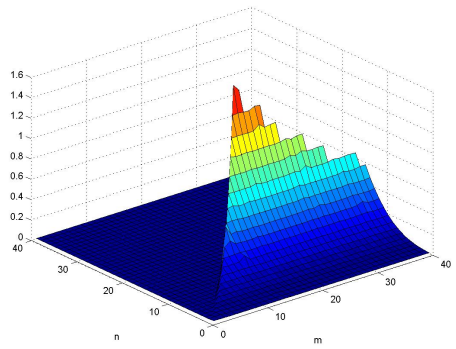

Figure 7: test1b(40,2)

For test1b(40,1), i.e., for well-conditioned matrices and measurement accuracy $\epsilon = 10^{-1}$, the criterion fails always if the number of columns of the matrix is larger than 4 (compare with the table below).

Of course, the condition depends on the bounds $\tilde{c}$ and u for the hybrid norms of $|\hat{A} - A|$ and R^{-1}. From the Figures 6 and 7 we see that for an error tol-

erance of 10^{-4}, which is relatively large, the criterion holds for all matrices with dimension up to 40, whereas for increasing error tolerance the inequality fails earlier. Increasing the dimension of the input matrix, depending on a given error tolerance $10^{-\mathsf{ep}}$, the following table contains the (row-)dimension, denoted by dim, where this inequality is violated for the first time:

ep	dim
1	5
2	22
3	100
4	465
5	2155

To be able to consider this test for ill-conditioned matrices, we use a modified version test1b(m,ep,l), where the additional parameter l is selectable to set the largest singular value of the testmatrix to 10^l. In this way we can increase the condition (see condh) for fixed error bounds. The Figures 8–13 show that a high condition causes the criterion to fail slightly earlier.

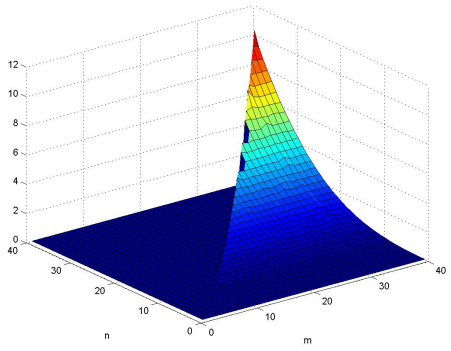

Figure 8: test1b(40,4,2)

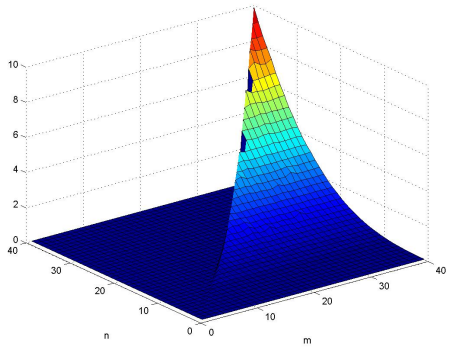

Figure 9: test1b(40,4,5)

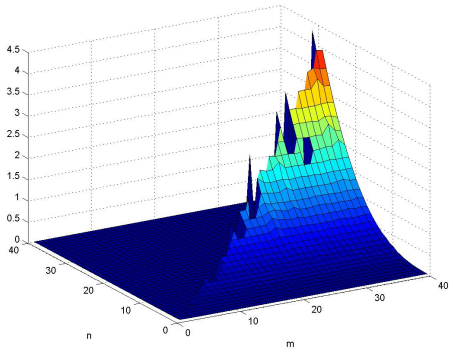

Figure 10: test1b(40,2,2)

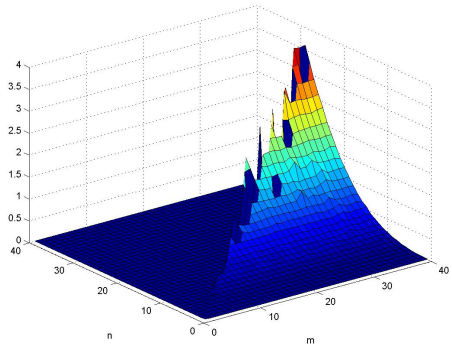

Figure 11: test1b(40,2,5)

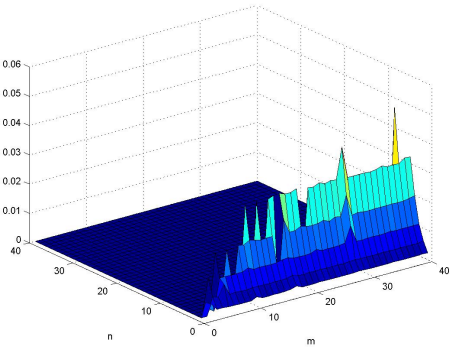

Figure 12: test1b(40,1,2)

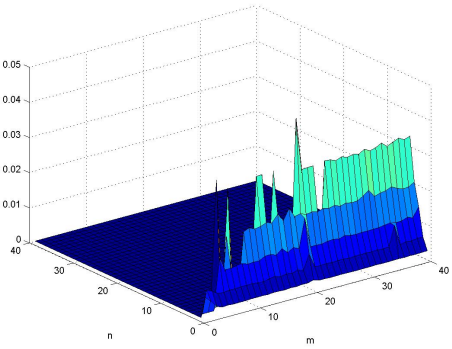

Figure 13: test1b(40,1,5)

Finally, we consider the function `test1c(m,ep)`, which is also similarly constructed as `test1a`, but measures the size of (the radius of) the enclosures by the 2-norm, depending once more on dimensions $m \times n$ of the matrix and accuracy $10^{-\mathsf{ep}}$. Plotting a result, whenever the criterion $\tilde{c}^T u < 1$ is satisfied, we start again with well-conditioned matrices:

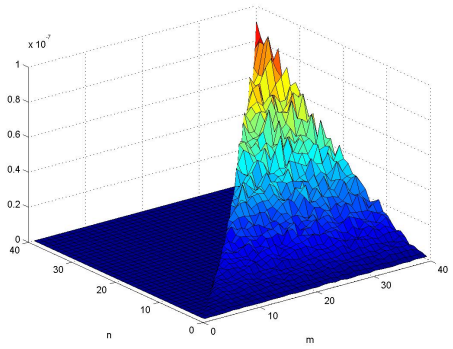

Figure 14: test1c(40,10)

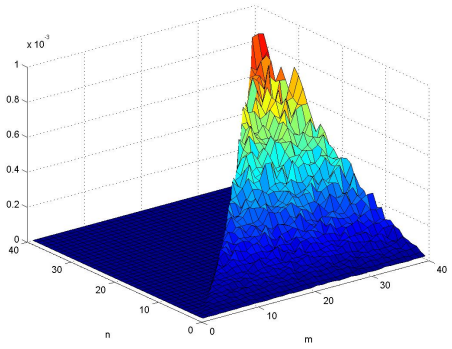

Figure 15: test1c(40,6)

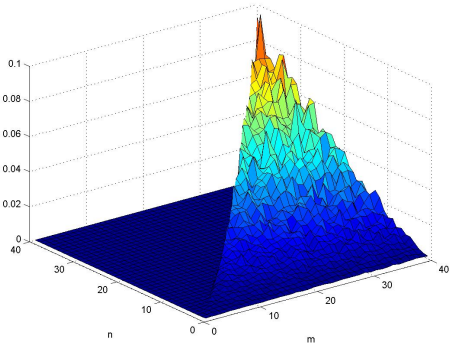

Figure 16: test1c(40,4)

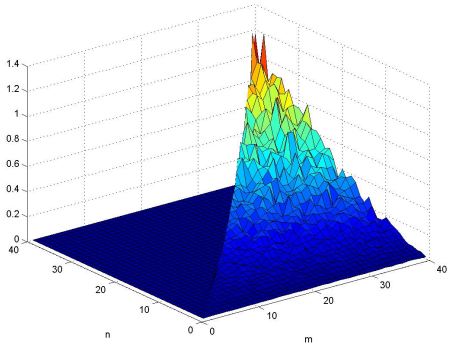

Figure 17: test1c(40,3)

For test1c(40,2), which means that the accuracy of the measurements amounts 10^{-2}, there are cases where this norms blows up so that the enclosures become meaningless resp. useless.

Finally, test1c(m,ep,l) tests the size of the errors for ill-conditioned matrices, where the condition of the testmatrix can be increased with parameter l.

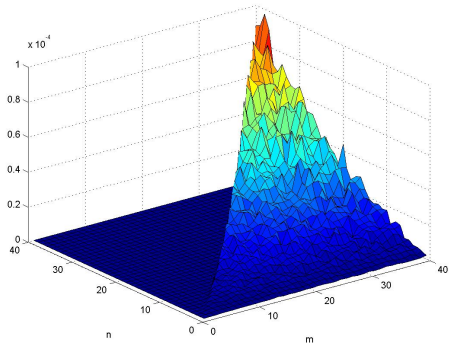

Figure 18: test1c(40,7,1)

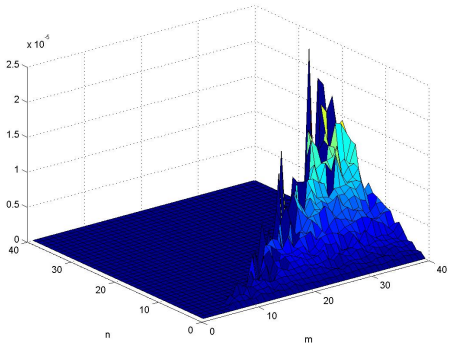

Figure 19: test1c(40,7,4)

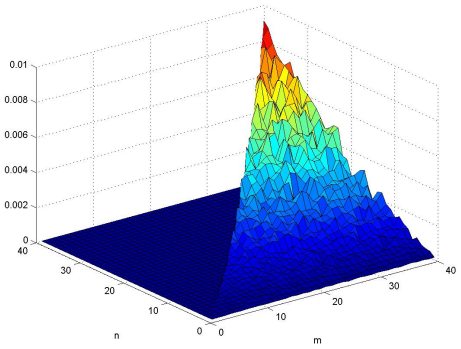

Figure 20: test1c(40,5,1)

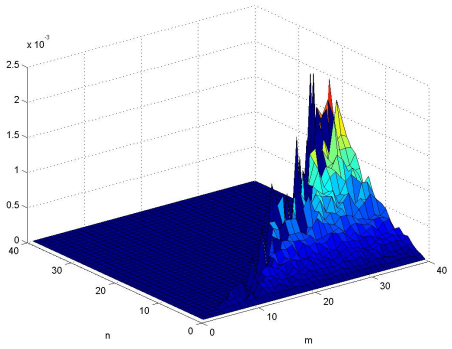

Figure 21: test1c(40,5,4)

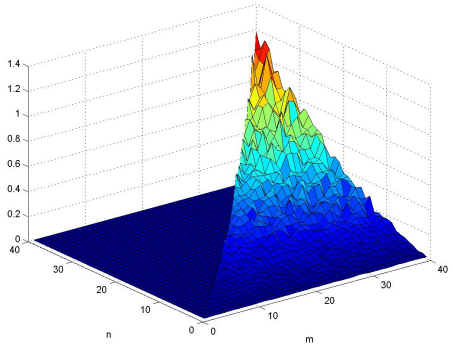

Figure 22: test1c(40,3,1)

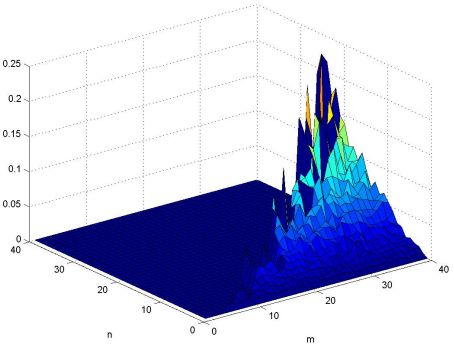

Figure 23: test1c(40,3,4)

Similar properties as for `test1b` hold for `test1c`, which analyzes the quality of the enclosures: For decreasing accuracy of the measurements, the 2-norm of the radius of the resulting interval vector increases, i.e., the quality gets worse and as before, a high condition of the input matrix induces the inequality $\tilde{c}^T u < 1$ to fail earlier.

7 Conclusion

After all, we were successful in developing a method, which produces an enclosure for the existing solution $\hat{x}$ of an overdetermined linear system of equations $\hat{A}\hat{x} = \hat{b}$, depending on perturbated parameters A and b and bounds ($\tilde{c}$ and $\tilde{\beta}$) for the perturbations. By Theorem 3.8 (resp. Corollary 3.9) and Lemma 4.5, this procedure is only possible, if several relations are satisfied ($\delta z \leq 1$ and $\tilde{c}^T u < 1$).

Summarizing the properties of the implementation of our program, Alogirthm 5.3, a high accuracy of the parameters respectively "small" bounds for the hybrid norms are essential for tight enclosures. But since the evaluation of those is very expensive for high dimensions, they should be really required. Moreover, a high (but small enough to run the algorithm) condition does not really have an influence on the results of the tests, causing the method to base on the directed QR-decomposition, which is numerically stable, since it only uses (almost) orthogonal matrices, the modified reflections.

Progressing further in this topic, one could look for a similar method for enclosing least squares solutions, i.e., developing a method, which computes bounds for $|\hat{x} - x|$, assuming the setting (A) and (C) from Section 2.

8 References

[1] Freund, R., Hoppe, R., Stoer/Bulirsch: Numerische Mathematik 1, 10. Auflage, Springer, 2007

[2] Schwarz, H., Köckler, N., Numerische Mathematik, 8. Auflage, Springer, 2011

[3] Neumaier, A.: Interval Methods for Systems of Equations, Cambridge University Press, 1990

[4] Neumaier, A.: Hybrid Norms and Bounds for Overdetermined Linear Systems, Linear Algebra and its Applications, Volume 216, Elsevier Science Inc., 1995, URL: `http://www.sciencedirect.com/science/article/pii/002437959300152P` [18.6.2014]

[5] Neumaier, A., Domes, F.: Directed modified Cholesky factorizations and convex quadratic relaxations, Universität Wien, 2014, URL: `http://www.mat.univie.ac.at/~neum/ms/Modchol.pdf` [14.7.2014]

[6] Moore, R., Baker Kearfott, R., Cloud, M.J., Introduction in Interval Arithmetic, Siam, 2009, URL: `http://www.sbras.ru/interval/Library/InteBooks/IntroIntervAn.pdf` [11.7.2014]

[7] Hargreaves, G.I.: Interval Analysis in Matlab, Masters thesis, Department of Mathematics, University of Manchester, 2002, URL: `http://www.ti3.tuhh.de/rump/intlab/narep416.pdf` [18.6.2014]

[8] Matlab, `http://www.mathworks.de/products/matlab/`

[9] Intlab, http://www.ti3.tu-harburg.de/rump/intlab/

Zusammenfassung

Wie der Titel vermuten lässt, ist Gegenstand dieser Masterarbeit die Einschließung von Lösungen überbestimmter Gleichungssysteme. Dabei werden Fehler in den Eingangsdaten erlaubt. Unter Annahme der Lösbarkeit des zugrundeliegenden überbestimmten Systems ist das verwendbare, "gestörte" System im Allgemeinen nicht mehr lösbar. Deshalb betrachtet man nun das Ausgleichsproblem mit diesen fehlerbehafteten Parametern. Sind (Schranken für) die Eingangsfehler bekannt, können diese in sogenannte Hybridnormen übersetzt werden, mit deren Hilfe man unter Verwendung einer reduzierten QR-Zerlegung in exakter Arithmetik Einschließungen finden kann.
Da in der Praxis jedoch Rundungsfehler berücksichtig werden müssen, werden stärkere Hilfsmittel benötigt. In Kapitel 3 und 4 wird beschrieben, wie eine QR-Faktorisierung berechnet werden kann, bei welcher Rundungsfehler kontrolliert werden können. Führt man diese gerichtete QR-Zerlegung, basierend auf dem Householder-Verfahren, an der fehlerbehafteten Matrix des Ausgleichsproblems durch, gelingt es, die Fortpflanzung der Rundungsfehler in der Zerlegung, zusätzlich zu den Eingangsfehlern, in Hybridnormen zu vereinen. Auf diese Weise ist es möglich, auch in Gleitkommaarithmetik Einschließungen zu berechnen.

Desweiteren werden Matlab-Programme vorgestellt, die das Kozept dieser Arbeit umsetzen. Durch Eingabe fehlerhafter Parameter eines zugrunde liegenden überbestimmten Gleichungssystems und Schranken für die Größe der Fehler, wird ein Intervalvektor ausgegeben, der die Lösung des Gleichungssystems enthält. In Kapitel 6 werden grundlegende Eigenschaften wie Laufzeit und Größe der Einschließungen dieser Methode analysiert.

Curriculum vitae

Personal data

Name:	Spazierer Armin
Title:	Bachelor of Science, BSc
Date of Birth:	3. September 1989
Place of Birth:	Vienna, Austria
Nationality:	Austria
Email:	armin.spazierer@gmx.at

Education

2011 - 2014	Master program (AMaSciCo), Faculty of Mathematics, University of Vienna
July, 2011	Bachelor of Science (BSc) in Mathematics
2008 - 2011	Bachelor program, Faculty of Mathematics, University of Vienna
1999 - 2007	Wirtschaftskundliches Realgymnasium Mater Salvatoris, 1070 Vienna
1995 - 1999	Volksschule Mater Salvatoris, 1070 Vienna