Chapter 3
Error Analysis

3.1 Introduction

In the problems we are interested in algorithms, like the LU–decomposition, are applied to given data, like matrices and right hand sides, to compute solutions to the given problems,

$$\text{DATA} \longrightarrow \text{ALGORITHM} \longrightarrow \text{RESULT}.$$

In practical applications, the given data are usually contaminated by errors and the while applying algorithms errors are introduced. Inevitably, this leads to errors in the computed result.

$$\text{ERRORS in DATA} \longrightarrow \text{ERRORS in ALGORITHM} \longrightarrow \text{ERRORS in RESULT}$$

The purpose of this chapter is to study the influence of errors in the data and errors in the algorithms on the computed solutions.

Before we start with the analysis, let us discuss some sourses of errors. Errors in the input data are caused by the conversion real numbers into floating point numbers. If $x \in I\!R$ is the exact input, then the floating point number $fl(x)$ obtained by rounding of x is the used input. It holds that $|fl(x) - x| \leq \mathbf{u}|x|$, where $\mathbf{u}$ is the unit roundoff. There may be other sources of errors in the input data such as measurement errors. A source of errors in the algorithm are floating point operations. If x and y are two floating point numbers and if $\Box$ is one of the elementary operations $+, -, *, /$, then $fl(x \Box y) = (x \Box y)(1 + \epsilon)$, where $|\epsilon| \leq \mathbf{u}$. Another source of errors in an algorithm are approximations of quantities that cannot be computed exactly. For example, if an algorithm requires the evaluation of an integral $\int_0^1 g(x)dx$, then this usually cannot be done exactly, but the integral has to be approximated by, say, $\sum_{i=1}^{n} w_i g(x_i)$. Approximations are also needed in algorithms for the solution of differential equations.

Two notions are important for this analysis: the condition of a problem and the *(numerical) stablility of an algorithm.*

3.2 Conditioning versus Stability

Conditioning of a Problem A *problem* is *well–conditioned* if small variations in data produce small variations in associated solutions. It is *ill–conditioned* if small variations in data can produce large variations in associated solutions.

Stability of an Algorithm An *algorithm* for solving a problem is *stable* if the computed solution for given data is the exact solution to the problem after slightly perturbing the given data. It is *unstable* if the computed solution for given data is the exact solution to the problem only after a large perturbation in the given data.

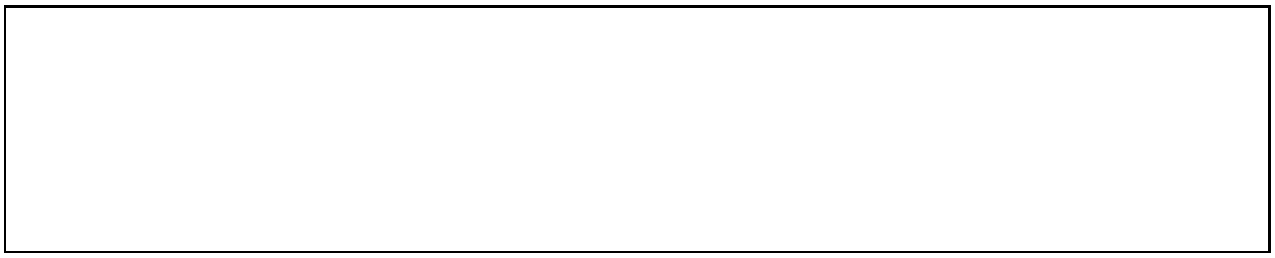

Figure 3.1: Changes in the solution x of $f(x) = y$ for small changes in the data.

Example 3.1 Let

$$f : I\!R \to I\!R.$$

and let $y \in I\!R$. Suppose we want to an x such that

$$f(x) = y. \tag{3.1}$$

i. The conditioning of the root finding problem. To simplify our presentation, we assume that the function f is monotone and that $f(x) = y$ has a solution x for every y. This implies that the inverse function f^{-1} exists. In particular, the unique solution of (3.1) is $x = f^{-1}(y)$. In general it will be sufficient that a solution x of $f(x) = y$ exists and that f is monotone in a neighborhood of x.

We are interested in how the solution x changes when the data y are perturbed by a small δy.

We can look at two quantities. We can compare the absolute error in the input data, $|(y+\delta y)-y| = |\delta y|$ with the absolute error in the solution, $|f^{-1}(y + \delta y) - f^{-1}(y)|$. This leads to

$$\frac{|f^{-1}(y + \delta y) - f^{-1}(y)|}{|\delta y|}. \tag{3.2}$$

We can also compare the relative error in the input data, $|(y + \delta y) - y|/|y| = |\delta y|/|y|$ with the relative error in the solution, $|f^{-1}(y + \delta y) - f^{-1}(y)|/|f^{-1}(y)|$. This leads to

$$\frac{|f^{-1}(y + \delta y) - f^{-1}(y)|/|f^{-1}(y)|}{|\delta y|/|y|}. \tag{3.3}$$

If f is sufficiently smooth, then its inverse function f^{-1} is also sufficiently smooth and the Taylor expansion of f^{-1} gives

$$f^{-1}(y + \delta y) = f^{-1}(y) + (f^{-1})'(y)\delta y + O(|\delta y|^2 \approx f^{-1}(y) + (f^{-1})'(y)\delta y$$

for small $|\delta y|$. From calculus we remember that

$$(f^{-1})'(y) = \frac{1}{f'(x)},$$

where $x = f^{-1}(y)$. Hence

$$f^{-1}(y + \delta y) \approx f^{-1}(y) + \frac{1}{f'(x)}\delta y \tag{3.4}$$

for small $|\delta y|$. If we insert (3.4) into (3.2) and (3.3), then we arrive at

$$\frac{|f^{-1}(y + \delta y) - f^{-1}(y)|}{|\delta y|} \approx \frac{1}{|f'(x)|} \tag{3.5}$$

and

$$\frac{|f^{-1}(y+\delta y)-f^{-1}(y)|/|f^{-1}(y)|}{|\delta y|/|y|} \approx \frac{|y|}{|f'(x)|\,|f^{-1}(y)|} = \frac{|y|}{|f'(x)|\,|x|}, \tag{3.6}$$

respectively. Equation (3.5) states that small changes δy in the data may lead to changes in the solution of size $|f'(x)|\,|\delta y|$. If $|f'(x)|$ is large the root finding problem (**??**) is well–conditioned; if $|f'(x)|$ is small the root finding problem (3.1) is well–conditioned. We call

$$\kappa_{\rm abs} = \frac{1}{|f'(x)|}$$

the *absolute condition number of the root finding problem* (3.1). The scalar

$$\kappa_{\rm rel} = \frac{|y|}{|f'(x)|\,|x|}$$

is called the *relative condition number of the root finding problem* (3.1). Equation (3.6) tells us that a small change δy in the input data may lead to a relative error $|f^{-1}(y+\delta y)-f^{-1}(y)|/|f^{-1}(y)|$ in the solution of size $\kappa_{\rm rel}|\delta y|/|y|$.

For example consider $f(x) = ax^2 + bx$ with $a \neq 0$. Let y be given such that $b^2 > -4ay$. If we set $c = -y$, then (3.1) is equivalent to

$$ax^2 + bx + c = 0 \tag{3.7}$$

and the two solutions of this equation are given by

$$x_\pm = \left(-b \pm \sqrt{b^2-4ac}\right)/(2a).$$

Clearly, $f'(x) = 2ax + b$ and

$$f'(x_\pm) = \pm\sqrt{b^2-4ac}.$$

We calculate that

$$\kappa^\pm_{\rm abs} = \frac{1}{\sqrt{b^2-4ac}}.$$

and

$$\kappa^\pm_{\rm rel} = \frac{c}{\sqrt{b^2-4ac}x_\pm}.$$

Thus, the problem of solving the quadratic equation (3.7) with $b^2 > 4ac$ is well–conditioned if $b^2 - 4ac \gg 0$; it is ill–conditioned if $b^2 - 4ac$ is small. In particular, if $a = 5 * 10^{-4}, b = 100$, and $c = 5 * 10^{-3}$ then $x_+ \approx -5 * 10^{-5}$, $x_- \approx -2 * 10^5$ and

$$\kappa^\pm_{\rm abs} \approx 10^{-2}.$$

and

$$\kappa^+_{\rm rel} \approx 1,$$
$$\kappa^-_{\rm rel} \approx -2.510^{-10}.$$

Thus, solution of this quadratic equation is a well–conditioned problem.

ii. The stability of an algorithm for solving the quadratic equation. We want to compute the roots of the quadratic equation (3.7). We assume $b^2 \gg 4ac$. The roots are given by

$$x_\pm = \left(-b \pm \sqrt{b^2 - 4ac}\right)/(2a). \tag{3.8}$$

Let $a = 5*10^{-4}, b = 100$, and $c = 5*10^{-3}$. We have seen that solving this quadratic equation is a well–posed problem.

Application of the formula for x_+ using the single precision Fortran on a SUN SparcStation 10 gives the computed value

$$x_+^{\text{comp}} = 0.$$

(see example 2.9). The exact root is $x_+ \approx -0.5E-04$. Hence the absolute error and relative error are given by $|x_+^{\text{comp}} - x_+| \approx 0.5E-04$ and $|x_+^{\text{comp}} - x_+|/|x_+| = 1$.

The reason for this large error in the solution is that the straight forward implementation of (3.8) leads to an unstable algorithm. To see this, note that $x_+^{\text{comp}} = 0$ is the exact solution of the quadratic equation

$$ax^2 + bx + \underbrace{0}_{=c+\delta c} = 0.$$

The straight forward implementation of (3.8) for computing x_+ gives a computed solution x_+^{comp} that is the exact solution to the root finding problem only after the data c have been perturbed by $\delta c = -c$. Thus the relative size of this data perturbation is $|\delta c|/|c| = 1$.

A stable formula for the computation of x_+ in the case $b > 0$ and $b^2 \geq 4ac$ is

$$x_+ = \frac{2c}{-b - \sqrt{b^2 - 4ac}}$$

(see Example 3.1).

iii. Using stability and condition number to analyse the error in the solution.

Let us consider (3.1). Suppose we apply an algorithm to solve (3.1) that insteasd of the exact solution $x = f^{-1}(y)$ returns a computed solution x^{comp}. We interpret this computed solution as the exact solution of a root finding problem with perturbed data $y + \delta y$, i.e.,

$$f(x^{\text{comp}}) = y + \delta y.$$

Case 1: The algorithm is stable and the problem is well–conditioned. The stability of the algorithm guarantees that $|\delta y|/|y|$ is small. Since our problem is well–conditioned, κ_{rel} is small. Hence (3.6) implies that

$$\frac{|x^{\text{comp}} - x|}{|x|} \leq \underbrace{\kappa_{\text{rel}}}_{\text{small}} \underbrace{\frac{|\delta y|}{|y|}}_{\text{small}}.$$

Case 2: The algorithm is stable and the problem is ill–conditioned. Again, the stability of the algorithm guarantees that $|\delta y|/|y|$ is small. Since our problem is ill–conditioned, κ_{rel} is large. Hence (3.6) implies that

$$\frac{|x^{\text{comp}} - x|}{|x|} \leq \underbrace{\kappa_{\text{rel}}}_{\text{large}} \underbrace{\frac{|\delta y|}{|y|}}_{\text{small}}.$$

Hence the relative error $|x^{\text{comp}} - x|/|x|$ could be large.

Case 3: The algorithm is unstable and the problem is well–conditioned. Since the algorithm is unstable $|\delta y|/|y|$ is large. Since our problem is well–conditioned, $\kappa_{\rm rel}$ is small. Hence (3.6) implies that

$$\frac{|x^{\rm comp}-x|}{|x|} \leq \underbrace{\kappa_{\rm rel}}_{\rm small} \underbrace{\frac{|\delta y|}{|y|}}_{\rm large}.$$

Again, the relative error $|x^{\rm comp}-x|/|x|$ could be large. This is what we have seen in part ii.

Case 4: The algorithm is unstable and the problem is ill–conditioned. Since the algorithm is unstable $|\delta y|/|y|$ is large. Since our problem is ill–conditioned, $\kappa_{\rm rel}$ is also large. Hence (3.6) implies that

$$\frac{|x^{\rm comp}-x|}{|x|} \leq \underbrace{\kappa_{\rm rel}}_{\rm large} \underbrace{\frac{|\delta y|}{|y|}}_{\rm large}.$$

Again, the relative error $|x^{\rm comp}-x|/|x|$ could be large.

◇

3.3 Vector and Matrix Norms

Many problems involve data and results that are vectors or matrices. To investigate the conditioning of such problems and the stability of algorithms for their solution, we need a measure for the 'size' of a vector x and a measure for the 'size' of a matrix A. Such measures, the so–called vector norms and matrix norms will be introduced in this section.

We begin with vector norms.

Definition 3.2. *A (vector–) norm on $\mathbb{R}^n$ is a function*

$$\begin{array}{rccc} \|\cdot\|: & \mathbb{R}^n & \to & \mathbb{R} \\ & x & \mapsto & \|x\| \end{array}$$

which for all $x, y \in \mathbb{R}^n, \alpha \in \mathbb{R}$ satisfies

i. $\|x\| \geq 0, \quad \|x\| = 0 \Leftrightarrow x = 0,$

ii. $\|\alpha x\| = |\alpha|\ \|x\|,$

iii. $\|x+y\| \leq \|x\| + \|y\|.$ *(triangle inequality)*

The most frequently used norms on $\mathbb{R}^n$ are given by

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}, \tag{3.9}$$

where $p \in [1, \infty)$, and

$$\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|. \tag{3.10}$$

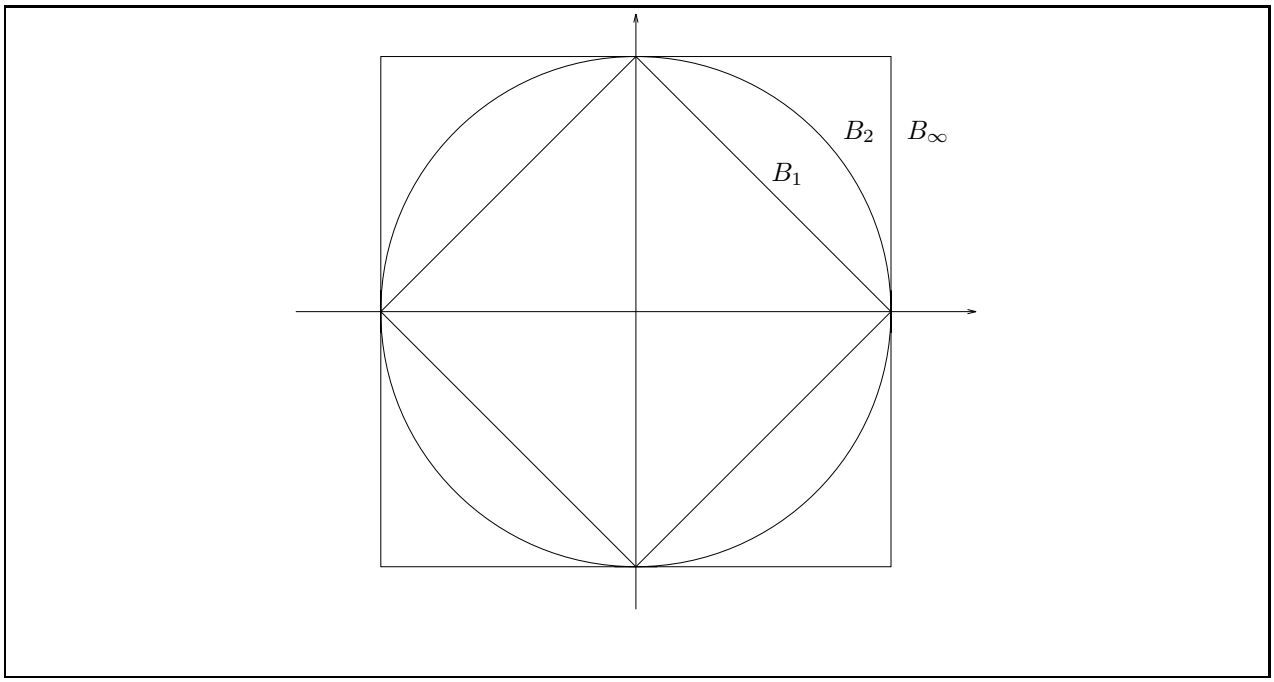

Figure 3.2: The unit "balls" B_1, B_2, B_∞.

In particular for $p = 1$ and $p = 2$ we have that

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|, \quad \|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}.$$

The norm (3.9) is called the *p–(vector–)norm* and (3.10) is called the *maximum–(vector–)norm.*

Example 3.3 Consider the vector $x = (1, -2, 3, -4)^T$. Then

$$\begin{aligned} \|x\|_1 &= 1 + 2 + 3 + 4 = 10, \\ \|x\|_2 &= \sqrt{1 + 4 + 9 + 16} = \sqrt{30} \approx 5.48, \\ \|x\|_\infty &= \max\{1, 2, 3, 4\} = 4. \end{aligned}$$

◇

The boundaries of the unit "balls" defined by

$$B_p = \{x \in I\!R^n \ : \ \|x\|_p \leq 1\}$$

are plotted in Figure 3.2.

Theorem 3.4. *Vector norms on $I\!R^n$ are equivalent, i.e. for every two vector norms $\|\cdot\|_a$ and $\|\cdot\|_b$ on $I\!R^n$ there exist constants c, C such that*

$$c\|x\|_b \leq \|x\|_a \leq C\|x\|_b \quad \forall x \in I\!R^n.$$

In particular, for any $x \in I\!R^n$ we have the inequalities

$$\begin{array}{rcccl} \frac{1}{\sqrt{n}}\|x\|_1 & \leq & \|x\|_2 & \leq & \|x\|_1, \\ \|x\|_\infty & \leq & \|x\|_2 & \leq & \sqrt{n}\|x\|_\infty, \\ \|x\|_\infty & \leq & \|x\|_1 & \leq & n\|x\|_\infty. \end{array}$$

Lemma 3.5. *Let $\|\cdot\|$ be a vector norm on $\mathbb{R}^n$. Then*

$$\|x+y\| \geq \big|\, \|x\| - \|y\| \,\big| \qquad \forall\, x, y.$$

Proof. The triangle inequality yields

$$\|x\| = \|-y+(x+y)\| \leq \|-y\| + \|x+y\| = \|y\| + \|x+y\|\,.$$

This gives $\|x+y\| \geq \|x\| - \|y\|$. The inequality $\|x+y\| \geq \|y\| - \|x\|$ can be proven by interchanging x and y. □

Lemma 3.6 (Cauchy-Schwarz Inequality). *For any $x, y \in \mathbb{R}$,*

$$|x^T y| \leq \|x\|_2 \|y\|_2. \tag{3.11}$$

Now we consider matrices. Since any matrix $A \in \mathbb{R}^{m\times n}$ can be identified with a vector in $\mathbb{R}^{mn}$, which is obtained by stacking the columns of the matrix into a long vector, the definition of a vector norm can immediately be extended to matrices.

Definition 3.7. *A (matrix–) norm on $\mathbb{R}^{n\times n}$ is a function*

$$\begin{array}{rccc} \|\cdot\|: & \mathbb{R}^{m\times n} & \to & \mathbb{R} \\ & A & \mapsto & \|A\| \end{array}$$

which for all $A, B \in \mathbb{R}^{n\times n}, \alpha \in \mathbb{R}$ satisfies

i. $\|A\| \geq 0, \quad \|A\| = 0 \Leftrightarrow A = 0,$

ii. $\|\alpha A\| = |\alpha|\, \|A\|,$

iii. $\|A+B\| \leq \|A\| + \|B\|.$ *(triangle inequality)*

Warning: Matrix– and vector–norms are denoted by the same symbol $\|\cdot\|$. However, as we will see shortly, vector–norms and matrix–norms are computed very differently. Thus, before computing a norm we need to examine carefully whether it is applied to a vector or to a matrix. It should be clear from the context which norm, a vector–norm or a matrix–norm, is used.

As mentioned above, each matrix $A \in \mathbb{R}^{m\times n}$ can be identified with a vector in $\mathbb{R}^{mn}$, which is obtained by stacking the columns of the matrix into a long vector. In particular, we obtain concrete matrix norms if we apply (3.9) or (3.10) to the long vector generated by the matrix. For example, if we consider a matrix $A \in \mathbb{R}^{m\times n}$ as a vector of length mn, the 2–vector norm of this long vector is

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

and it is called the *Frobenius–norm*. You may ask why we call this norm the Frobenius–norm and not the 2–norm. The reason is that there is another view of matrices that leads to a class of matrix-norms. These matrix–norms are generated by the p–vector–norms (3.9) or (3.10), they are called p–matrix–norms (or p–operator–norms) and they are denoted by the subscript p. The announced alternative view of matrices, does

not consider matrices to be long vectors, but rather views a matrix $A \in \mathbb{R}^{m\times n}$ as a linear mapping, which maps a vector $x \in \mathbb{R}^n$ into a vector $Ax \in \mathbb{R}^m$:

$$\begin{array}{cccc} A & \mathbb{R}^n & \to & \mathbb{R}^m \\ & x & \mapsto & Ax \end{array}$$

Now we compare the size of the image $Ax \in \mathbb{R}^m$ with the size of x. This leads to the following definition.

Definition 3.8. *The* p–matrix–norm, $p \in [1,\infty)$ *or* $p = \infty$, *is defined by*

$$\|A\|_p = \sup_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p}. \tag{3.12}$$

Note that on the left hand side in (3.12) the symbol $\|\cdot\|_p$ refers to the p–matrix–norm, while on the right hand side in (3.12) the symbol $\|\cdot\|_p$ refers to the p–vector–norm applied to the vectors $Ax \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$, respectively. *Warning: The same symbol* $\|\cdot\|_p$ *is used to denote the* p*–vector–norm and the* p*–matrix–norm, respectively. However, the* p*–vector–norm and the* p*–matrix–norm are computed very differently. Thus, before computing a norm we need to examine carefully whether it is applied to a vector or to a matrix. It should be clear from the context which norm, the* p*–vector–norm or the* p*–matrix–norm, is used.*

At this point it is not clear that (3.12) in fact defines a matrix norm, i.e., that (3.12) satisfies the conditions in Definition 3.7. We will convince ourselves of this fact in Theorem 3.10.

Note that the p–matrix–norm of the identity is always equal to one,

$$\|I\|_p = \max_{x\neq 0} \frac{\|Ix\|_p}{\|x\|_p} = 1.$$

In particular, $\|I\|_2 = 1$. On the other hand, if we compute Frobenius norm, which is obtained by viewing the matrix as a long vector, then $\|I\|_F = \sqrt{n}$.

The following identities are useful for theoretical investigations.

Lemma 3.9. *Let* $p \in [1,\infty)$ *or* $p = \infty$. *The following identities are valid*

$$\sup_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \sup_{\|x\|_p=1} \|Ax\|_p = \max_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \max_{\|x\|_p=1} \|Ax\|_p.$$

Theorem 3.10. *Definition* 3.8 *in fact defines a matrix norm, i.e.,* (3.12) *satisfies the properties in Definition* 3.7.

Proof. i. From the definition of the matrix norm we immediately find that $\|A\|_p \geq 0$ and $\|A\|_p = 0$ if $A = 0$. Now suppose that $\|A\|_p = 0$ and that $A \neq 0$. Then there exists $\bar{x} \neq 0$ such that $A\bar{x} \neq 0$. Thus,

$$0 = \|A\|_p = \max_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p} \geq \frac{\|A\bar{x}\|_p}{\|\bar{x}\|_p} > 0,$$

a contradiction. Hence $\|A\|_p = 0$ if and only if $A = 0$.
ii. From the second property of vector norms we find that

$$\|\lambda A\|_p = \max_{x\neq 0} \frac{\|\lambda Ax\|_p}{\|x\|_p} = \max_{x\neq 0} \frac{|\lambda|\ \|Ax\|_p}{\|x\|_p} = |\lambda|\ \max_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p} = |\lambda|\ \|A\|_p .$$

iii. The triangle inequality of vector norms implies that

$$\begin{aligned}\|A+B\|_p &= \max_{x\neq 0}\frac{\|(A+B)x\|_p}{\|x\|_p} \le \max_{x\neq 0}\frac{\|Ax\|_p}{\|x\|_p}+\frac{\|Bx\|_p}{\|x\|_p}\\ &\le \max_{x\neq 0}\frac{\|Ax\|_p}{\|x\|_p}+\max_{x\neq 0}\frac{\|Bx\|_p}{\|x\|_p} = \|A\|_p+\|B\|_p\,.\end{aligned}$$

□

The formula (3.12) looks very impractical. However, for the most commonly used matrix–norms (3.12) with $p=1$, $p=2$, or $p=\infty$, there exist rather simple representations.

Theorem 3.11. *Let $\|\cdot\|_p$ be the matrix norm defined in* (3.12)*. Then*

$$\begin{array}{rcll}\|A\|_1 &=& \max_{j=1,\dots,n}\sum_{i=1}^m |a_{ij}| & \textit{(maximum column norm)}\,,\\ \|A\|_\infty &=& \max_{i=1,\dots,m}\sum_{j=1}^n |a_{ij}| & \textit{(maximum row norm)}\,,\\ \|A\|_2 &=& \sqrt{\lambda_{max}(A^TA)} & \textit{(spectral norm)}\,,\end{array}$$

where $\lambda_{max}(A^TA)$ is the largest eigenvalue of A^TA.

Proof. For any $x\in\mathbb{R}^n$ it holds that

$$\begin{aligned}\|Ax\|_1 &= \sum_{i=1}^m |\sum_{j=1}^n a_{ij}x_j| \le \sum_{i=1}^m\sum_{j=1}^n |a_{ij}|\,|x_j| = \sum_{j=1}^n |x_j|\sum_{i=1}^m |a_{ij}|\\ &\le \sum_{j=1}^m |x_j| \max_{k=1,\dots,n}\sum_{i=1}^m |a_{ik}|\\ &= \|x\|_1 \max_{k=1,\dots,n}\sum_{i=1}^m |a_{ik}|.\end{aligned}$$

This shows that $\|A\|_1 \le \max_{j=1,\dots,n}\sum_{i=1}^m |a_{ij}|$.

To show equality, we have to construct a vector x with $\|x\|_1 = 1$ such that $\|Ax\|_1 = \max_{j=1,\dots,n}\sum_{i=1}^m |a_{ij}|$. Let j_0 be an index such that $\sum_{i=1}^m |a_{ij_0}| = \max_{j=1,\dots,n}\sum_{i=1}^m |a_{ij}|$. Then $\|Ae_{j_0}\|_1 = \sum_{i=1}^m |a_{ij_0}|$ and $\|e_{j_0}\|_1 = 1$. Thus, $\|A\|_1 = \max_{j=1,\dots,n}\sum_{i=1}^m |a_{ij}|$.

The proofs of the other identities are left as an exercise. □

Corollary 3.12. *Let $\|\cdot\|_p$, $p=1,2$ or $p=\infty$, be the matrix norm defined in* (3.12)*. Then*

$$\|A\|_1 = \|A^T\|_\infty$$

and

$$\|A\|_2 = \|A^T\|_2.$$

If A is symmetric, then

$$\begin{aligned}\|A\|_1 = \|A\|_\infty &= \max_{i=1,\dots,n}\sum_{j=1}^n |a_{ij}|,\\ \|A\|_2 &= \max_{i=1,\dots,n}|\lambda_i(A)|\,,\end{aligned}$$

where $\lambda_i(A)$ is the i–th eigenvalue of A.

Proof. The assertion concerning the norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$ follow immediately from Theorem 3.11.

If $\lambda = \lambda_{\max}(A^TA) > 0$, then $A^TAx = \lambda x$, $x \neq 0$, implies $Ax \neq 0$. Thus, $AA^T(Ax) = \lambda(Ax)$, $Ax \neq 0$. This shows that $\lambda = \lambda_{\max}(A^TA)$ is an eigenvalue of AA^T. From this we can conclude that $\|A\|_2 \leq \|A^T\|_2$.

The reverse inequality $\|A^T\|_2 \leq \|A\|_2$ can be established by interchanging A and A^T.

If A is symmetric with eigenvalues $\lambda_1, \ldots, \lambda_n$, then $\lambda_1^2, \ldots, \lambda_n^2$ are the eigenvalues of $A^2 = A^TA$. This yields $\|A\|_2 = \max_{i=1,\ldots,n} |\lambda_i(A)|$. ∎

Example 3.13 i. Let

$$A = \begin{pmatrix} 1 & 3 & -6 \\ -2 & 4 & 2 \\ 2 & 1 & -1 \end{pmatrix}.$$

Then

$$\begin{aligned} \|A\|_1 &= \max\{5, 8, 9\} = 9, \\ \|A\|_\infty &= \max\{10, 8, 4\} = 10, \\ \|A\|_2 &\approx \sqrt{\max\{3.0749, 23.8627, 49.0624\}} \approx 7.0045, \\ \|A\|_F &= \sqrt{76} \approx 8.718. \end{aligned}$$

(The eigenvalues of A^TA were computed using Matlab.)

ii. Consider the so–called Hilbert matrix $H \in \mathbb{R}^{n\times n}$ with elements h_{ij} given by

$$h_{ij} = \frac{1}{i+j-1}$$

For $n = 4$ the Hilbert matrix is given by

$$H = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \end{pmatrix}.$$

The Hilbert matrix is symmetric and therefore

$$\|H\|_\infty = \|H\|_1 = \max_{i=1,\ldots,n} \sum_{j=1}^{n} \frac{1}{i+j-1} = \sum_{j=1}^{n} \frac{1}{j}.$$

For $n = 4$ we find that $\|H\|_\infty = \|H\|_1 = 25/12$. ◇

Theorem 3.14. *For any $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times k}$ and $x \in \mathbb{R}^n$, the inequalities*

$$\|Ax\|_p \leq \|A\|_p \, \|x\|_p \qquad \textit{(compatibility of matrix and vector norm)}$$

and

$$\|AB\|_p \leq \|A\|_p \|B\|_p \qquad \textit{(submultiplicativity of matrix norms)}$$

are valid.

Proof. i. If $x = 0$, then $\|Ax\|_p = 0 = \|A\|_p \, \|x\|_p$. If $x \neq 0$, then the definition (3.12) of the p–matrix norm implies that

$$\frac{\|Ax\|_p}{\|x\|_p} \leq \max_{\bar{x}\neq 0} \frac{\|A\bar{x}\|_p}{\|\bar{x}\|_p} = \|A\|_p \, .$$

Thus, $\|Ax\|_p \leq \|A\|_p \, \|x\|_p$.

ii. If $Bx = 0$ for all x, then $\|AB\|_p = 0 = \|A\|_{m,k} \, \|B\|_{k,n}$. Otherwise, we use the definition of the matrix norm and the compatibility of the matrix norm to show that

$$\begin{aligned}
\|AB\|_p &= \max_{x\neq 0} \frac{\|ABx\|_p}{\|x\|_p} \\
&= \max_{Bx\neq 0} \frac{\|ABx\|_p}{\|Bx\|_p} \, \frac{\|Bx\|_k}{\|x\|_p} \\
&\leq \max_{Bx\neq 0} \frac{\|ABx\|_p}{\|Bx\|_p} \, \max_{x\neq 0} \frac{\|Bx\|_p}{\|x\|_p} \\
&= \max_{\bar{x}\neq 0} \frac{\|A\bar{x}\|_p}{\|\bar{x}\|_p} \, \max_{x\neq 0} \frac{\|Bx\|_p}{\|x\|_p} \; = \; \|A\|_p \, \|B\|_p \, .
\end{aligned}$$

□

Theorem 3.15. *For any $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{k\times m}$, and $x \in \mathbb{R}^n$,*

$$\|Ax\|_2 \leq \|A\|_F \, \|x\|_2$$

and

$$\|AB\|_F \leq \|A\|_F \|B\|_F.$$

Proof. The inequalities can be proven using the Cauchy-Schwarz inequality (3.11). □

Analogously to Lemma 3.5 we can prove the following result.

Lemma 3.16. *Let $\|\cdot\|$ be a matrix norm on $\mathbb{R}^{n\times n}$. Then*

$$\|A + B\| \geq \big| \, \|A\| - \|B\| \, \big| \quad \forall \, A, B \in \mathbb{R}^{m\times n}.$$

3.4 Error Analysis for the Solution of Linear Systems

3.4.1 The Condition Number of a Matrix (With Respect to Inversion)

Given $A \in \mathbb{R}^{n\times n}$, and $b \in \mathbb{R}^n$ we are interested in the solution $x \in \mathbb{R}^n$ of

$$Ax = b. \tag{3.13}$$

Usually the input data A and b are not given exactly, but are perturbed due to rounding errors (floating point representation) and measurement errors. Thus, instead of solving (3.13) we are actually solving a perturbed linear system

$$(A + \Delta A)x = b + \Delta b, \tag{3.14}$$

where $\Delta A \in \mathbb{R}^{n\times n}$ and $\Delta b \in \mathbb{R}^n$ represent the perturbations in A and b, respectively.

Let x denote the solution of (3.13) and let $x + \Delta x$ denote the solution of (3.14). Since we really want the solution x, but can only compute $x + \Delta x$ we are interested to know how good the computed solution is. Thus we want to have an estimate for the *absolute error in the solution*

$$\|x - (x + \Delta x)\|_p = \|\Delta x\|_p$$

and an estimate for the *relative error in the solution*

$$\frac{\|x - (\Delta x + x)\|_p}{\|x\|_p} = \frac{\|\Delta x\|_p}{\|x\|_p}.$$

In particular we want to investigate the dependence of the relative error in the solution upon the *relative errors in the input data*

$$\frac{\|\Delta A\|_p}{\|A\|_p}, \quad \frac{\|\Delta b\|_p}{\|b\|_p}.$$

Before we can study the sensitivity of the solution of the linear system with respect to perturbations in the input data we have to investigate the invertibility of the perturbed matrix $(A + \Delta A)$. If we multiply by A^{-1}, then we obtain the matrix $(I + A^{-1}\Delta A)$. The following result investigates the invertibility of such matrices:

Lemma 3.17. *Let $\|\cdot\|_p$ be an operator norm. If $B \in \mathbb{R}^{n\times n}$ is a matrix with $\|B\|_p < 1$, then the inverse of $I + B$ exists and it holds that*

$$\|(I + B)^{-1}\|_p \le \frac{1}{1 - \|B\|_p}.$$

Proof. Let $x \neq 0$, then

$$\|(I + B)x\|_p = \|x + Bx\|_p \ge \|x\|_p - \|Bx\|_p \ge \|x\|_p - \|B\|_p \|x\|_p = (1 - \|B\|_p)\|x\|_p > 0.$$

Hence, the linear system $(I + B)x = 0$ only has the trivial solution $x = 0$. Therefore $I + B$ is invertible.

Since $\|\cdot\|$ is an operator norm we find that

$$\begin{aligned} 1 = \|I\|_p &= \|(I + B)(I + B)^{-1}\|_p = \|(I + B)^{-1} + B(I + B)^{-1}\|_p \\ &\ge \|(I + B)^{-1}\|_p - \|B(I + B)^{-1}\|_p \\ &\ge \|(I + B)^{-1}\|_p - \|B\|_p\|(I + B)^{-1}\|_p \\ &\ge \big(1 - \|B\|_p\big)\|(I + B)^{-1}\|_p. \end{aligned}$$

This gives the assertion. □

Remark 3.18. *Lemma 3.17 can also be proven using the Neumann series. If x is a scalar with $|x| < 1$, then*

$$\frac{1}{1 - x} = \sum_{i=0}^{\infty} x^i.$$

A similar result holds true for matrices: If $B \in \mathbb{R}^{n\times n}$ is a matrix, then $I - B$ is invertible if and only if the Neumann series $\sum_{i=0}^{\infty} B^{-i}$ *is convergent. In this case*

$$(I - B)^{-1} = \sum_{i=0}^{\infty} B^i.$$

If $\|B\|_p < 1$, then the Neumann series $\sum_{i=0}^{\infty} B^i$ is convergent.

Theorem 3.19. *Let $\|\cdot\|$ be a vector norm on $\mathbb{R}^n$ and let $\|\cdot\|$ be a submultiplicative matrix norm which is compatible with this vector norm. Moreover, let $A \in \mathbb{R}^{n\times n}$ be nonsingular and let $\Delta A \in \mathbb{R}^{n\times n}$ be such that $\|A^{-1}\|_p \, \|\Delta A\|_p < 1$.*

If x is the solution of

$$Ax = b$$

and if $x + \Delta x$ is the solution of

$$(A + \Delta A)x = b + \Delta b,$$

then

$$\frac{\|\Delta x\|_p}{\|x\|_p} \leq \frac{\kappa_p(A)}{1 - \kappa_p(A)\frac{\|\Delta A\|_p}{\|A\|_p}} \left(\frac{\|\Delta A\|_p}{\|A\|_p} + \frac{\|\Delta b\|_p}{\|b\|_p} \right), \tag{3.15}$$

where

$$\kappa_p(A) = \|A\|_p \, \|A^{-1}\|_p.$$

Definition 3.20. *The (p–) condition number $\kappa_p(A)$ of a matrix (with respect to inversion) is defined by*

$$\kappa_p(A) = \|A\|_p \, \|A^{-1}\|_p.$$

We set $\kappa_p(A) = \infty$ if A is not invertible.

Proof. (Proof of Theorem 3.19)
It holds that

$$\begin{aligned} Ax &= b, \\ (A + \Delta A)(x + \Delta x) &= b + \Delta b. \end{aligned}$$

Hence

$$(A + \Delta A)\Delta x = -\Delta A x + \Delta b.$$

Multiplication with A^{-1} yields

$$(I + A^{-1}\Delta A)\Delta x = -A^{-1}(\Delta A x + \Delta b).$$

Note that $\|A^{-1}\Delta A\|_p \leq \|A^{-1}\|_p \|\Delta A\|_p < 1$. If we apply the previous lemma with $B = A^{-1}\Delta A$, then we find that $I + A^{-1}\Delta A$ is invertible and

$$\|(I + A^{-1}\Delta A)^{-1}\|_p \leq \frac{1}{1 - \|A^{-1}\Delta A\|_p} \leq \frac{1}{1 - \|A^{-1}\|_p \|\Delta A\|_p}.$$

Hence

$$\Delta x = -(I + A^{-1}\Delta A)^{-1} A^{-1}(\Delta A x + \Delta b)$$

and

$$\|\Delta x\|_p \leq \frac{1}{1 - \|A^{-1}\|_p \, \|\Delta A\|_p} \|A^{-1}\|_p \big(\|\Delta A\|_p \|x\|_p + \|\Delta b\|_p \big).$$

If we divide by $\|x\|_p$ and use $\|b\|_p = \|Ax\|_p \leq \|A\|_p\|x\|_p$, we find that

$$\begin{aligned}\frac{\|\Delta x\|_p}{\|x\|_p} &\leq \frac{\|A^{-1}\|_p\|A\|_p}{1-\|A^{-1}\|_p\|\Delta A\|_p}\left(\frac{\|\Delta A\|_p}{\|A\|_p}+\frac{\|\Delta b\|_p}{\|A\|_p\|x\|_p}\right)\\ &\leq \frac{\kappa_p(A)}{1-\kappa_p(A)\frac{\|\Delta A\|_p}{\|A\|_p}}\left(\frac{\|\Delta A\|_p}{\|A\|_p}+\frac{\|\Delta b\|_p}{\|b\|_p}\right).\end{aligned}$$

□

Theorem 3.21. *The condition number is invariant with respect to scaling of the matrix by a scalar, i.e.*

$$\kappa_p(\alpha A) = \kappa_p(A) \qquad \forall A \in I\!R^{n\times n}, \alpha \in I\!R.$$

The condition number is submultiplicative, i.e.,

$$\kappa_p(AB) \leq \kappa_p(A)\kappa_p(B) \qquad \forall A, B \in I\!R^{n\times n}.$$

The condition number is greater equaal to one

$$\kappa_p(A) \geq 1 = \kappa_p(I) \qquad \forall A \in I\!R^{n\times n}.$$

Proof. The first property follows immediately from $(\alpha A)^{-1} = \alpha^{-1}A^{-1}$. The second property is an immediate consequence of the submultiplicativity of the norm, and the third property follows from

$$1 = \|I\|_p\|I\|_p = \kappa_p(I) = \|AA^{-1}\|_p \leq \|A\|_p\|A^{-1}\|_p = \kappa_p(A).$$

□

The first property in the previous theorem shows that the condition number is invariant with respect to multiplication by a scalar. Notice that the determinant satisfies $det(\alpha A) = \alpha^n det(A)$. By multiplication with small scalars the determinant can be made arbitrary small. Therefore the determinant is a bad indicator of the conditioning of a matrix.

The condition number of A depends on the matrix norm that is used. We use the following notations

$$\begin{aligned}\kappa_1(A) &= \|A\|_1\,\|A^{-1}\|_1,\\ \kappa_2(A) &= \|A\|_2\,\|A^{-1}\|_2,\\ \kappa_\infty(A) &= \|A\|_\infty\,\|A^{-1}\|_\infty,\end{aligned}$$

where $\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_\infty$ are the operator norms induced by the vector norms $\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_\infty$, respectively.

Example 3.22 Consider

$$A = \begin{pmatrix} 1 & 3 & -6\\ -2 & 4 & 2\\ 2 & 1 & -1\end{pmatrix}, \quad A^{-1} = \begin{pmatrix} -\frac{1}{10} & \frac{1}{20} & \frac{1}{2}\\ \frac{1}{30} & \frac{11}{60} & \frac{1}{6}\\ -\frac{1}{6} & \frac{1}{12} & \frac{1}{6}\end{pmatrix},$$

cf. Example 3.13. Then

$$\begin{array}{lll}\|A\|_1 = 9, & \|A^{-1}\|_1 = \frac{5}{6}, & \kappa_1(A) = \frac{15}{2},\\ \|A\|_\infty = 10, & \|A^{-1}\|_\infty = \frac{13}{20}, & \kappa_\infty(A) = \frac{13}{2},\\ \|A\|_2 \approx 7.0045, & \|A^{-1}\|_2 \approx 0.5703, & \kappa_2(A) \approx 3.9947.\end{array}$$

(The eigenvalues of A^TA and $(A^{-1})^TA^{-1}$ were computed using Matlab.) ◇

Example 3.23 A standard example of an ill–conditioned matrix is the Hilbert matrix. This matrix arises in the least squares polynomial approximation and its entires are given by

$$h_{ij} = \int_0^1 x^{i+j-2}dx = \frac{1}{i+j-1}.$$

For $n = 4$ the Hilbert matrix and its inverse are given by

$$H = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \end{pmatrix}, \quad H^{-1} = \begin{pmatrix} 16 & -120 & 240 & -140 \\ -120 & 1200 & -2700 & 1680 \\ 240 & -2700 & 6480 & -4200 \\ -140 & 1680 & -4200 & 2800 \end{pmatrix}.$$

Hilbert matrices are examples of ill–conditioned matrices. In fact, the condition number of a Hilbert matrix grows very fast with n. For $n = 4$ we find that

$$\begin{array}{lll} \|H\|_1 = 25/12, & \|H^{-1}\|_1 = 13620, & \kappa_1(H) = 28375\,, \\ \|H\|_\infty = \|H\|_1, & \|H^{-1}\|_\infty = \|H^{-1}\|_1, & \kappa_\infty(H) = \kappa_1(H)\,, \\ \|H\|_2 \approx 1.5, & \|H^{-1}\|_2 \approx 1.03 * 10^4, & \kappa_2(H) \approx 1.55 * 10^4. \end{array}$$

(Again, the eigenvalues of H and H^{-1} were computed using Matlab.)

We consider the linear systems

$$Hx = b.$$

For given n we set $x_{\rm ex} = (1, \ldots, 1)^T \in I\!R^n$, and compute $b = Hx_{\rm ex}$. Then we compute the solution of the linear system $Hx = b$ using the LU–decomposition and compute the relative error between exact solution $x_{\rm ex}$ and computed solution x. The results are shown in the Table 3.1 and Figure 3.3. All computations are done in Matlab on a SUN Ultra10 workstation. ◇

Table 3.1. *Condition Number of the Hilbert Matrix and Relative Error in System Solution*

n	$\kappa_\infty(H)$	$\frac{\|x_{\rm ex}-x\|_\infty}{\|x_{\rm ex}\|_\infty}$
4	$2.837500E+04$	$2.327027E-13$
5	$9.436560E+05$	$4.896639E-12$
6	$2.907028E+07$	$8.405362E-10$
7	$9.851949E+08$	$1.479009E-08$
8	$3.387279E+10$	$8.561445E-07$
9	$1.099651E+12$	$2.231209E-05$
10	$3.535372E+13$	$9.362458E-04$

Example 3.24 (See [**?**, Example 6, Chapter 4]) We consider the linear system $Ax = b$ with

$$A = \begin{pmatrix} 0.780 & 0.563 \\ 0.913 & 0.659 \end{pmatrix}, \quad b = \begin{pmatrix} 0.217 \\ 0.254 \end{pmatrix}.$$

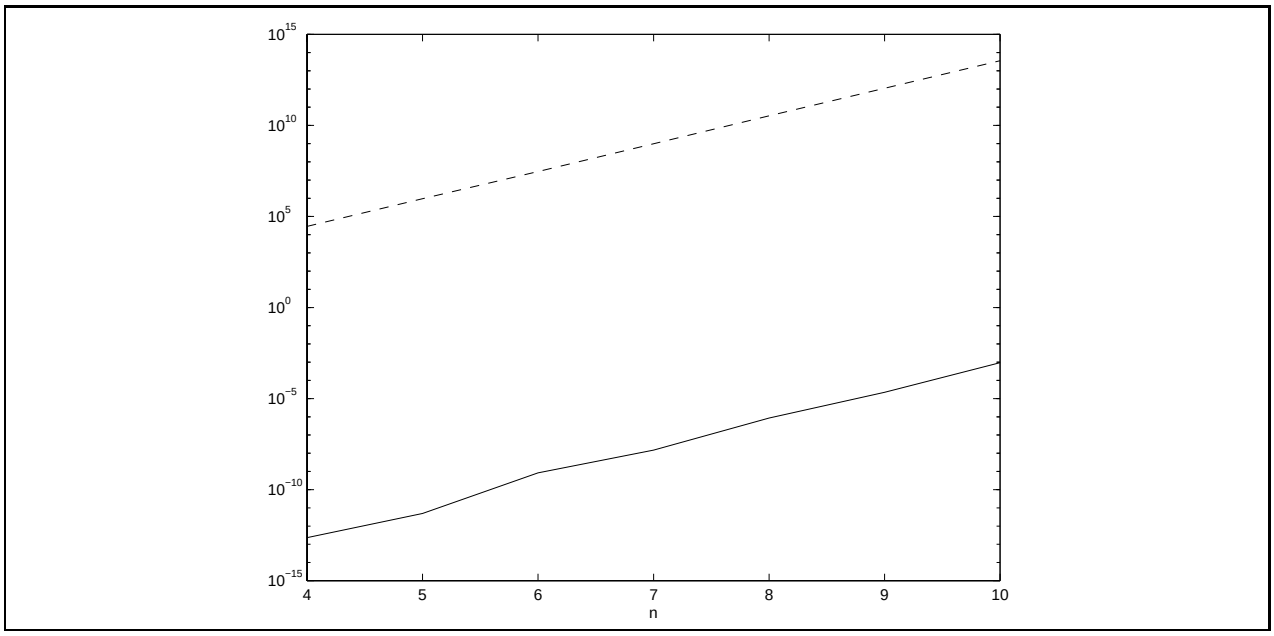

Figure 3.3: Condition Number κ_∞ of the Hilbert Matrix (dashed) and Relative Error in System Solution Measured in the ∞–Norm (solid)

The exact solution is given by $x = (1, -1)^T$. Now we consider the perturbations

$$\Delta b_1 = \begin{pmatrix} -1.343 * 10^{-3} \\ -1.572 * 10^{-3} \end{pmatrix}, \quad \Delta b_2 = \begin{pmatrix} -1 * 10^{-6} \\ 0 \end{pmatrix}.$$

The solutions x_1, x_2 of $Ax = b + \Delta b_1$, $Ax = b + \Delta b_2$ are given by

$$x_1 = \begin{pmatrix} 0.9990 \\ -1.0010 \end{pmatrix}, \quad x_2 = \begin{pmatrix} 0.341 \\ -0.087 \end{pmatrix}.$$

The inverse of A is given by

$$A^{-1} = \begin{pmatrix} 659000 & -563000 \\ -913000 & 780000 \end{pmatrix},$$

and the condition number of A is

$$\kappa_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty = 1.572 * 1693000 = 2661396.$$

The relative errors in the approximate solutions are

$$\frac{\|x - x_1\|_\infty}{\|x\|_\infty} = 0.001, \quad \frac{\|x - x_2\|_\infty}{\|x\|_\infty} = 0.913.$$

We now apply Theorem 3.19. For the first approximation we find that

$$0.001 = \frac{\|x - x_1\|_\infty}{\|x\|_\infty} \leq \kappa_\infty(A) \frac{\|\Delta b_1\|_\infty}{\|b\|_\infty} = 2661396 * \frac{1.572 * 10^{-3}}{0.254} \approx 16471.67$$

and for the second approximation we compute

$$0.913 = \frac{\|x - x_2\|_\infty}{\|x\|_\infty} \leq \kappa_\infty(A) \frac{\|\Delta b_2\|_\infty}{\|b\|_\infty} = 2661396 * \frac{10^{-6}}{0.254} \approx 10.47794\,.$$

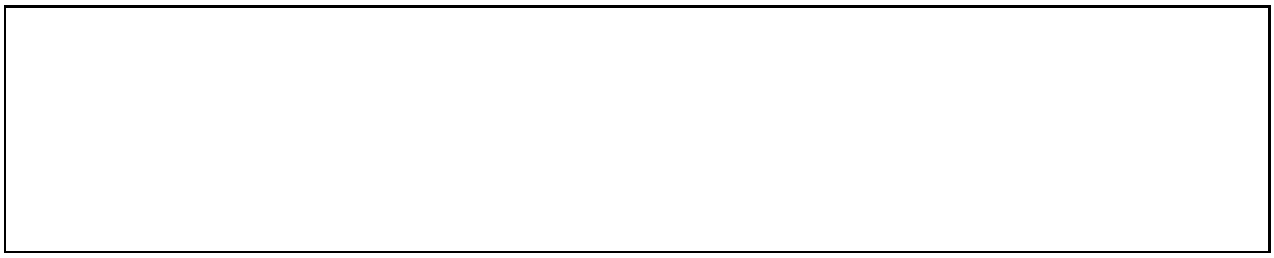

Figure 3.4: A Well–Conditioned Linear System

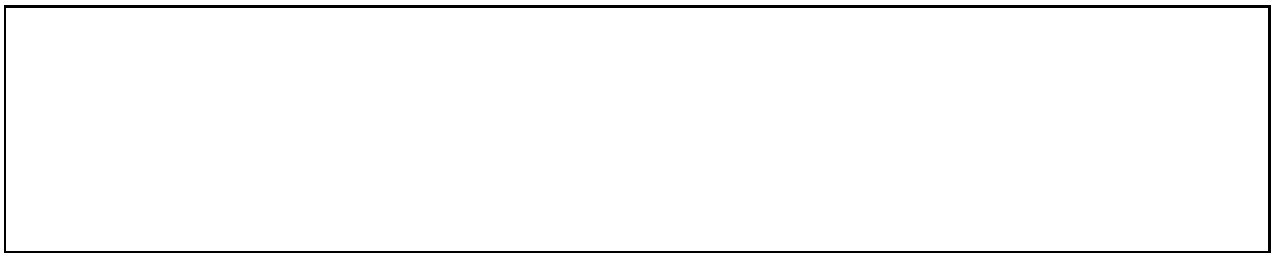

Figure 3.5: An Ill–Conditioned Linear System

We see that in the first case the error estimate gives a far too pessimistic bound on the accuracy of x_1, whereas in the second case the error bound established in Theorem 3.19 is quite satisfactory.

The so–called residuals $Ax_1 - b$ and $Ax_2 - b$ are given by

$$Ax_1 - b = \Delta b_1 = \begin{pmatrix} -1.343 * 10^{-3} \\ -1.572 * 10^{-3} \end{pmatrix}, \quad Ax_2 - b = \Delta b_2 = \begin{pmatrix} -1 * 10^{-6} \\ 0 \end{pmatrix}.$$

We observe that the more accurate solution x_1 has a larger residual than the less accurate approximation x_2. Thus if the condition number of A is large, then a small residual $Ax - b$ does not necessarily mean that the solution is accurate. ◇

We can summarize the observations from the previous example as follows.

- If the condition number of a matrix A is large, then small errors in the data may lead to large errors in the solution.
- If the condition number of a matrix A is large, then the residual is not necessarily a good measure for the quality of an approximate solution.

The conditioning of a linear system can also be illustrated graphically. Each equation in $Ax = b$ represents a hyperplane in $\mathbb{R}^n$. The row vector a_i is orthogonal to the hyperplane. In the case $n = 2$ this is illustrated in Figures 3.4, Figures 3.5. The solution of the linear system is the intersection of these n hyperplanes. If the row vectors are almost linearly dependent, then small perturbations lead to large deviations in the point of intersection. See Figure 3.5.

If we use finite precision arithmetic, then rounding causes errors in the input data. Using t–digit floating point arithmetic it holds that

$$\frac{|x - fl(x)|}{|x|} \leq 0.5 * 10^{-t+1}.$$

Thus, if we solve the linear system in t–digit floating point arithmetic, then, as rule of thumb, we may approximate the the input errors due to rounding by

$$\frac{\|\Delta A\|}{\|A\|} \approx 0.5 * 10^{-t+1}, \quad \frac{\|\Delta b\|}{\|b\|} \approx 0.5 * 10^{-t+1}.$$

(This is only a heuristic argument since in general we cannot norm estimates for the relative error from the relative errors of the components.) If the condition number of A is $\kappa(A) = 10^{\alpha}$, then

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{10^{\alpha}}{1 - 10^{\alpha-t+1}} \left(0.5 * 10^{-t} + 0.5 * 10^{-t+1}\right) \approx 10^{\alpha-t},$$

provided $10^{\alpha-t+1} \ll 1$.

> **Rule of thumb:** *If the linear system is solved in t–digit floating point arithmetic and if the condition number of A is of the order 10^{α}, then only $t-\alpha-1$ digits in the solution can be trusted.*

Example 3.25 We consider the linear system

$$Ax = b,$$

where $A \in I\!R^{n\times n}$ has the form

$$A = \begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{n-1} \\ 1 & t_2 & t_2^2 & \cdots & t_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & t_n & t_n^2 & \cdots & t_n^{n-1} \end{pmatrix}. \tag{3.16}$$

A matrix of the form (3.16) is called a *Vandermonde matrix*. We choose the so called nodes t_i to be

$$t_i = -1 + 2\frac{i-1}{n-1}, \quad i = 1, \ldots, n,$$

The right hand side is constructed so that the exact solution is known. This is done by setting

$$x = (1, 1, \ldots, 1)^T$$

and choosing

$$b = Ax.$$

Thus,

$$b_i = \sum_{j=1}^{n} a_{ij} = \sum_{j=1}^{n} t_i^{j-1} = \begin{cases} \frac{1-t_i^n}{1-t_i} & \text{if } t_i \neq 1, \\ n & \text{if } t_i = 1. \end{cases}$$

We compute the solution of the linear system $Ax = b$ using the LU–decomposition and we compute the absolute and the relative errors between exact solution x_{ex} and computed solution x. The results are shown in the Table 3.2 and Figure 3.6. The computations were performed using Matlab program on a SUN Ultra10.

$\diamond$

3.4.2 The Stability of the LU–Decomposition

Suppose we have computed a solution $\widehat{x}$ of $A\widehat{x} = b$. The computed $\widehat{x}$ can mot expected to be the exact solution of $A\widehat{x} = b$. Instead it is the exact solution of a perturbed system

$$(A + \Delta A)\widehat{x} = b. \tag{3.17}$$

Table 3.2. *Condition Number κ_2 of the Vandermonde Matrix and Relative Error in System Solution Measured in the 2–Norm.*

n	$\|x - x_{comp}\|_2$	$\frac{\|x-x_{comp}\|_2}{\|x\|_2}$	$\kappa_2(A)$
2	0	0	$1.0000E+00$
4	$3.3307E-16$	$1.6653E-16$	$8.0116E+00$
6	$8.6069E-15$	$3.5138E-15$	$6.3827E+01$
8	$3.5742E-14$	$1.2637E-14$	$5.3535E+02$
10	$6.8944E-13$	$2.1802E-13$	$4.6264E+03$
12	$2.5120E-12$	$7.2515E-13$	$4.0755E+04$
14	$9.0459E-12$	$2.4176E-12$	$3.6383E+05$
16	$4.1435E-11$	$1.0359E-11$	$3.2800E+06$
18	$2.2795E-09$	$5.3729E-10$	$2.9794E+07$
20	$7.6731E-09$	$1.7157E-09$	$2.7224E+08$
22	$8.5664E-08$	$1.8264E-08$	$2.4997E+09$
24	$3.2457E-06$	$6.6253E-07$	$2.3043E+10$
26	$2.6801E-05$	$5.2561E-06$	$2.1314E+11$
28	$3.2168E-04$	$6.0792E-05$	$1.9772E+12$
30	$6.8107E-04$	$1.2435E-04$	$1.8385E+13$
32	$1.7475E-03$	$3.0892E-04$	$1.7136E+14$
34	$7.4378E-01$	$1.2756E-01$	$1.5842E+15$
36	$1.3912E+00$	$2.3186E-01$	$1.4187E+16$
38	$1.6026E+01$	$2.5997E+00$	$1.0424E+17$
40	$5.2901E+01$	$8.3644E+00$	$7.9772E+17$

We say that $\widehat{x}$ was computed stably if $\|\Delta A\|_p/\|A\|_p$ is not too large. How can we find out if $\widehat{x}$ was computed stably? Define the residual

$$r = b - A\widehat{x}.$$

We can write

$$\left(A + \frac{r\,\widehat{x}^T}{\|\widehat{x}\|_2^2}\right)\widehat{x} = A\widehat{x} + \frac{r\,\widehat{x}^T\widehat{x}}{\|\widehat{x}\|_2^2} = A\widehat{x} + r = b.$$

Hence if we set

$$\Delta A = \frac{r\,\widehat{x}^T}{\|\widehat{x}\|_2^2},$$

then $\widehat{x}$ is the exact solution of the perturbed system (3.17). The 2–norm (3.12) of ΔA is given by

$$\|\Delta A\|_2 = \sup_{x\neq 0}\frac{\|\Delta A x\|_2}{\|x\|_2} = \sup_{x\neq 0}\frac{\left\|\frac{r\widehat{x}^T x}{\|\widehat{x}\|_2^2}\right\|_2}{\|x\|_2} = \sup_{x\neq 0}\frac{\|r\|_2|\widehat{x}^T x|}{\|\widehat{x}\|_2^2\|x\|_2}$$

By the Cauchy–Schwarz inequality,

$$\frac{\|r\|_2|\widehat{x}^T x|}{\|\widehat{x}\|_2^2\|x\|_2} \leq \frac{\|r\|_2\|\widehat{x}\|_2\|x\|_2}{\|\widehat{x}\|_2^2\|x\|_2} = \frac{\|r\|_2}{\|\widehat{x}\|_2}$$

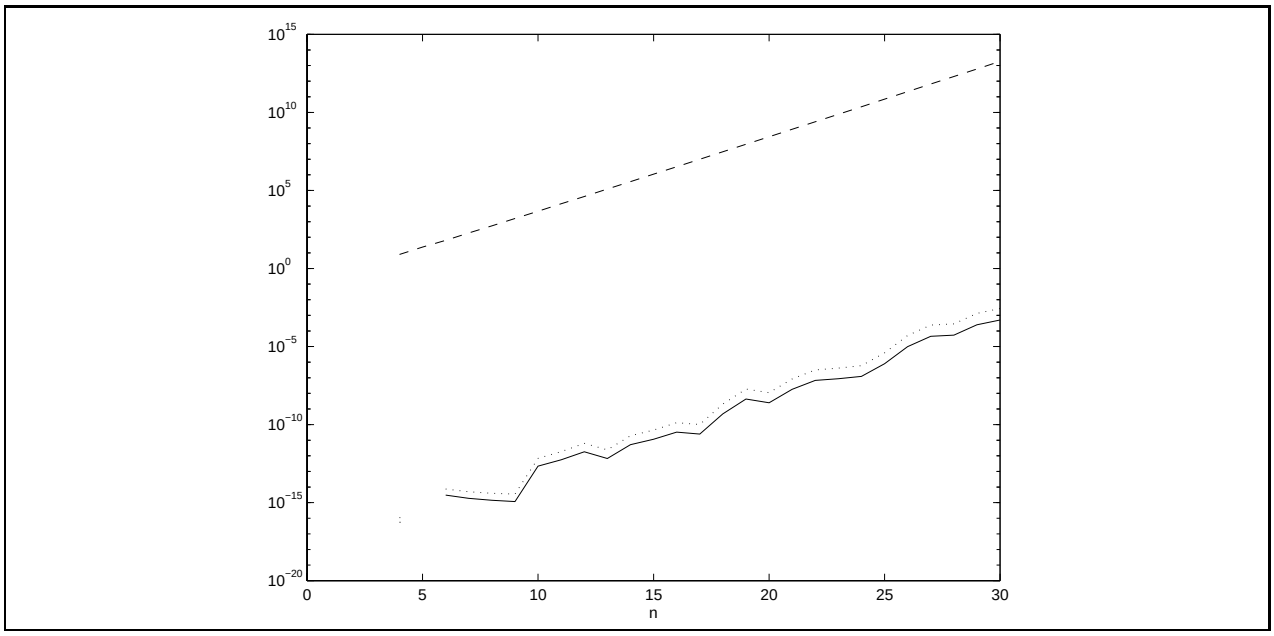

Figure 3.6: Condition Number κ_2 of the Vandermonde Matrix (dashed), Absolute Error in System Solution Measured in the 2–Norm (dotted), and Relative Error in System Solution Measured in the 2–Norm (solid).

and for $x = \widehat{x}$,

$$\frac{\|r\|_2 |\widehat{x}^T x|}{\|\widehat{x}\|_2^2 \|x\|_2} = \frac{\|r\|_2}{\|\widehat{x}\|_2}.$$

Hence,

$$\|\Delta A\|_2 = \frac{\|r\|_2}{\|\widehat{x}\|_2}. \tag{3.18}$$

Hence, the relative error of the perturbation ΔA in (3.17) is

$$\frac{\|\Delta A\|_2}{\|A\|_2} = \frac{\|r\|_2}{\|A\|_2 \|\widehat{x}\|_2}. \tag{3.19}$$

If $\|r\|_2/(\|A\|_2\|\widehat{x}\|_2)$ is small, then $\widehat{x}$ is computed stably. The right hand side (**??**) is our *stability indicator*. Notice that in our derivation of (**??**) we have assumed that once $\widehat{x}$ all computations, such as the computation of $r = b - A\widehat{x}$ can be computed exactly. This is certainly not true in general and therefore (**??**) can only be used as an indicator. More generally, we may use

$$\frac{\|r\|_p}{\|A\|_p \|\widehat{x}\|_p} \tag{3.20}$$

with any $p \in [1, \infty)$, $p = \infty$ as a stability indicator.

Note that the previous analysis did not use any information about the algorithm with which the computed solution $\widehat{x}$ was obtained. The stability indicator (3.20) can only be computed after the computed solution $\widehat{x}$ is obtained; (3.20) is an *a–posteriori stability indicator*.

Now we turn to the stability of the LU–decomposition. We want to give an *a–priori estimate* of the size of $\|\Delta A\|_p/\|A\|_p$. We use use $p = \infty$. Given $A \in I\!R^{n\times n}$, let $\widehat{P}, \widehat{L}, \widehat{U}$ be the permutation matrix, the

lower triangular matrix and the upper triangular matrix computed by the LU–decomposition with partial pivoting, Algorithm 1.6.4. Because of rounding errors

$$\widehat{P}A \neq \widehat{L}\widehat{U}.$$

Thus, if we use the computed factors $\widehat{P}, \widehat{L}, \widehat{U}$ in Algorithm 1.6.4 to compute the solution of the linear system $Ax = b$, then we do not obtain the exact solution $x = A^{-1}b$, but a vector $\widehat{x}$.

We interpret the computed decomposition $\widehat{P}, \widehat{L}, \widehat{U}$ of A as the exact decomposition of a perturbed matrix $A + \Delta A$, i.e.,

$$\widehat{L}\widehat{U} = \widehat{P}(A + \Delta A_1).$$

Furthermore, we interpret the computed solution $\widehat{x}$ the exact decomposition of a perturbed system $A + \Delta A$, i.e.,

$$(A + \Delta A)\widehat{x} = b.$$

The perturbations ΔA_1 and ΔA are not the same, since to get $\widehat{x}$ we have to apply Algorithm 1.6.4 with the computed factors $\widehat{P}, \widehat{L}, \widehat{U}$ and additional rounding error will occur in Algorithm 1.6.4.

We want to characterize how big the perturbations ΔA_1 and ΔA are. This is done in the following theorem, whose proof can be found in, e.g., [**?**, Thm. 9.3, 9.4]. Actually, the following theorem applies not only to the LU–decomposition with partial pivoting. Recall that in step k the LU–decomposition with partial pivoting interchanges rows k and i_0 of $A^{(k)} = \widehat{M}_{k-1}\widehat{P}_{k-1} \dots \widehat{M}_1\widehat{P}_1 A$, where $i_0 \geq k$ is a row index that satisfies

$$|a^{(k)}_{i_0,k}| = \max_{i=k,\dots n} |a^{(k)}_{i,k}|.$$

Here $a^{(k)}_{i,j}$ denote the entries of the matrix $A^{(k)} = \widehat{M}_{k-1}\widehat{P}_{k-1} \dots \widehat{M}_1\widehat{P}_1 A$. The following result is valid if we replace the partial pivoting rule by any pivoting rule that guarantees that $a^{(k)}_{i_0,k} \neq 0$, if $\max_{i=k,\dots n} |a^{(k)}_{i,k}| > 0$. In particular, we can choose i_0 to be the first row index greater equal to k that contains a nonzero entry.

Theorem 3.26. *Let $A \in \mathbb{R}^{n\times n}$ and let $b \in \mathbb{R}^n$. Suppose that $n\mathbf{u} < 1$, where $\mathbf{u}$ is the unit-roundoff.*

i. *The LU–decomposition with pivoting computes $\widehat{P}, \widehat{L}, \widehat{U}$ so that*

$$\widehat{P}(A + \Delta A_1) = \widehat{L}\widehat{U}, \quad \text{where } |\Delta A_1| \leq \frac{n\mathbf{u}}{1 - n\mathbf{u}} |\widehat{L}|\,|\widehat{U}|. \tag{3.21}$$

ii. *Suppose the LU–decomposition with pivoting computes $\widehat{P}, \widehat{L}, \widehat{U}$ and a computed solution $\widehat{x}$, then there exists ΔA such that*

$$\widehat{P}(A + \Delta A)\widehat{x} = b, \quad \text{where } |\Delta A| \leq 2\frac{n\mathbf{u}}{1 - n\mathbf{u}} |\widehat{L}|\,|\widehat{U}|. \tag{3.22}$$

In (3.21), (3.22), the absolute value $|B|$ of a matrix B is the matrix whose entries are obtained by taking the absolute values of entries of B.

The previous result is incomplete. We want to estimate the size of ΔA_1, ΔA in terms of the original problem data A and, possibly, b. Thus, we need to estimate the sizes of $\widehat{L}$ and $\widehat{U}$ in terms of A and, possibly, b. This is where the pivoting rule comes in. First, observe that if we use partial pivoting, then all entries in $\widehat{L}$ below the diagonal have an absolute value less than one,

$$|\widehat{l}| \leq 1$$

Therefore, if the LU–decomposition with partial pivoting is used,

$$\| \, |\widehat{L}| \, |\widehat{U}| \, \|_\infty \leq n \|\widehat{U}\|_\infty$$

independent of A. What about $\|\widehat{U}\|_\infty / \|A\|_\infty$? If we define the so-called *growth factor* ρ_n to be the smallest scalar such that

$$|\widehat{u}_{ij}| \leq \rho_n \|A\|_\infty,$$

for all i, j, then

$$\| \, |\widehat{L}| \, |\widehat{U}| \, \|_\infty \leq n \|\widehat{U}\|_\infty \leq n^2 \rho_n \|A\|_\infty.$$

Thus we have

Theorem 3.27. *Let $A \in \mathbb{R}^{n \times n}$ and let $b \in \mathbb{R}^n$. Suppose that $n\mathbf{u} < 1$, where $\mathbf{u}$ is the unit-roundoff. If $\widehat{x}$ is the computed solution of $Ax = b$ using the LU–decomposition* with partial pivoting*, then there exists ΔA such that*

$$\widehat{P}(A + \Delta A)\widehat{x} = b, \quad \textit{where } \|\Delta A\| \leq 2n^2 \frac{n\mathbf{u}}{1 - n\mathbf{u}} \rho_n \|A\|_\infty. \tag{3.23}$$

where ρ_n is the smallest number such that

$$|u_{ij}| \leq \rho_n \|A\|_\infty.$$

How big is the growth factor? An upper bound for the growth factor is $\rho_n \leq 2^{n-1}$ and there are matrices A for which this growth factor is actually attained. However in most cases the growth factor is small. According to W. M. Kahan[1],

> *Intolerable pivot growth [with partial pivoting] is a phenomenon that happens only to numerical analysts who are looking for that phenomenon.*

For some matrices with special structure one can prove that ρ_n is small. Theorem 3.27 is only applicable to the LU–decomposition with partial pivoting. It is not valid for the LU–decomposition with simple pivoting, in which we choose the pivot index i_0 simply to be the first row index greater equal to k that contains a nonzero entry. This leads to the following algorithm (compare with Algorithm 1.6.4).

[1] Cited from[**?**, p. 169].

Algorithm 3.4.1 LU–Decomposition with Simple Pivoting

Input: $A \in \mathbb{R}^{n\times n}$.
Output: $L, U \in \mathbb{R}^{n\times n}$, $ipivt \in \mathbb{N}^n$ (A is successively overwritten with L and U.)
The pivoting information is stored in $ipiv$.

```
1     For k = 1, ... , n − 1 do
2        (* find pivot index i_0 *)
3        a_max = |a_kk|
4        i_0 = k
5        while (a_max = 0 and i_0 < n) do
6              i_0 = i_0 + 1
7              a_max = |a_{i_0 k}|
8        End
9        ipivt_k = i_0
10       (* interchange rows if necessary *)
11       If i_0 ≠ k then
12          For j = k, ... , n do
13             t = a_{i_0 j}
14             a_{i_0 j} = a_kj
15             a_kj = t
16          End
17       Endif
18       (* If a_kk = 0 then all entries are zero and we do not need to eliminate *)
19       If a_kk ≠ 0 then
20          (* compute the Gauss transformation matrix M_k,
21             i.e., compute the multipliers −l_ik and store them in a_ik *)
22          For i = k + 1, ... , n do
23             a_ik = −a_ik/a_kk
24          End
25          (* row elimination *)
26          For i = k + 1, ... , n do
27             For j = k + 1, ... , n do
28                a_ij = a_ij + a_ik a_kj
29             End
30          End
31       Endif
32    End
```

3.4.3 Putting Everything Together

Suppose we are interested in the solution x of

$$Ax = b, \tag{3.24}$$

where $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$. However, we are not given the exact matrices A and b but only perturbations $A + \Delta A$ and $b + \Delta b$. together with estimates of the relative errors $\|A\|_p/\|\Delta A\|_p$ and $\|b\|_p/\|\Delta b\|_p$, where $p \in [1, \infty)$ or $p = \infty$. The errors ΔA and Δb could be due to rounding that occurs upon entry of the exact data A and b into the computer. Errors also arise if A and b are obtained from measurements. Now, we

solve

$$(A + \Delta A)x = b + \Delta b. \tag{3.25}$$

using, e.g., the LU–decomposition. The computed solution $\widehat{x}$ of (3.25) does not satisfy (3.25) exactly, but

$$(A + \Delta A + \Delta A_1)\widehat{x} = b + \Delta b, \tag{3.26}$$

where ΔA_1 reflects the errors introduced by the numerical algorithm for the solution of (3.25). We can estimate

$$\frac{\|\Delta A_1\|_p}{\|A + \Delta A\|_p}$$

using the error indicator (3.20) applied to (3.25), i.e.,

$$\frac{\|\Delta A_1\|_p}{\|A + \Delta A\|_p} \approx \frac{\|b + \Delta b - (A + \Delta A)\widehat{x}\|_p}{\|A + \Delta A\|_p \|\widehat{x}\|_p}.$$

Theorem 3.19 gives an estimate of the relative error between the computed solution $\widehat{x}$, i.e., the solution of (3.26), and the desired solution x of (3.24). Estimate (3.15) applied in this context reads

$$\frac{\|\widehat{x} - x\|_p}{\|x\|_p} \leq \frac{\kappa_p(A)}{1 - \kappa_p(A)\frac{\|\Delta A + \Delta A_1\|_p}{\|A\|_p}} \left(\frac{\|\Delta A + \Delta A_1\|_p}{\|A\|_p} + \frac{\|\Delta b\|_p}{\|b\|_p} \right),$$

To express the relative error $\|\Delta A + \Delta A_1\|_p / \|A\|_p$ in known quantities, we use

$$\begin{aligned} \frac{\|\Delta A + \Delta A_1\|_p}{\|A\|_p} &\leq \frac{\|\Delta A\|_p}{\|A\|_p} + \frac{\|\Delta A_1\|_p}{\|A\|_p} \\ &\leq \frac{\|\Delta A\|_p}{\|A\|_p} + \frac{\|\Delta A_1\|_p}{\|A + \Delta A\|_p} \frac{\|A + \Delta A\|_p}{\|A\|_p} \\ &= \frac{\|\Delta A\|_p}{\|A\|_p} + \frac{\|\Delta A_1\|_p}{\|A + \Delta A\|_p} \left(1 + \frac{\|\Delta A\|_p}{\|A\|_p} \right) \end{aligned}$$

3.5 Problems

Problem 3.1

i. Let

$$A = \begin{pmatrix} 1 & -1 & -1 & -1 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Compute the matrix norms $\|A\|_p$ and the condition numbers $\kappa_p(A) = \|A\|_p \|A^{-1}\|_p$ for $p = 1, 2, \infty$.

ii. Let $b = (5, 1.02, 1.04, 1.1)^T$. Compute $\widehat{b}$ by rounding the entries of b to the nearest integers. Compute the solution $\widehat{x}$ of $A\widehat{x} = \widehat{b}$.

iii. Use Theorem 3.19 to compute upper bounds for the relative errors $\|\widehat{x} - x\|_p / \|x\|_p$, $p = 1, 2, \infty$. Do not compute the solution x of $Ax = b$.

iv. Compute the solution x of $Ax = b$ and the relative errors $\|\widehat{x} - x\|_p / \|x\|_p$, $p = 1, 2, \infty$.

Problem 3.2 ([?, p. 197]) Let

$$A = \begin{pmatrix} 1.000 & 1.001 \\ 1.000 & 1.000 \end{pmatrix}.$$

i. The exact solution of $Ax = b_1$ with $b_1 = (2.001, 2.000)^T$ is $x_1 = (1, 1)^T$.
Compute the solution $\widehat{x}_1$ of $Ax = \widehat{b}_1$, where

$$\widehat{b}_1 = \begin{pmatrix} 2.002 \\ 2.000 \end{pmatrix}.$$

Compute $\|x_1 - \widehat{x}_1\|_\infty$ and $\|b_1 - \widehat{b}_1\|_\infty$.

ii. The exact solution of $Ax = b_2$ with $b_2 = (1, 0)^T$ is $x_2 = (-1000, 1000)^T$.
Compute the solution $\widehat{x}_2$ of $Ax = b_2$, where

$$\widehat{b}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Compute $\|x_2 - \widehat{x}_2\|_\infty$ and $\|b_2 - \widehat{b}_2\|_\infty$.

iii. Note that the residual $\|b_1 - \widehat{b}_1\|_\infty$ is small relative to $\|b_1\|_\infty$ while $\|x_1 - \widehat{x}_1\|_\infty$ is large relative to $\|x_1\|_\infty$. On the other hand, $\|b_2 - \widehat{b}_2\|_\infty$ is small relative to $\|b_2\|_\infty$ while $\|x_2 - \widehat{x}_2\|_\infty$ is large relative to $\|x_2\|_\infty$.
Do your results agree with Theorem 3.19?

Problem 3.3

i. Modify `lu_pp.m` to implement the LU–decomposition with simple decomposition, Algorithm 3.4.2, as a MATLAB function `lu_sp.m`

ii. Generate m random $n \times n$ linear systems `A = rand(n,n)`, `x = rand(n,1)` and `b = A*x`. Use $m = 10$, $n = 100$.

- Use `lu_pp.m` to compute the LU–decomposition with partial pivoting of A and apply `lu_pp_sl.m` to solve the linear system. Let x^{comp} be the computed solution. For each system compute $\|A - P^T LU\|_\infty$, $\kappa_\infty(A)\|b - Ax^{\text{comp}}\|_\infty / \|b\|_\infty$, $\|b - Ax^{\text{comp}}\|_\infty / (\|A\|_\infty \|x^{\text{comp}}\|_\infty)$, and $\|x - x^{\text{comp}}\|_\infty / \|x\|_\infty$.
  Report the results in form of a table or in form of a table or a graph.
- Repeat the computations with `lu_pp.m` replaced by `lu_sp.m`.

iii. Repeat the computations in ii. using the linear systems in Example 3.25 with $n = 4, \ldots, 30$.

iv. Interpret the results you have obtained in ii. and iii.

Problem 3.4 Consider the truss in Problem 1.10

i. Compute $\|B|_1, \|B\|_2, \|B^T\|_2, \|B\|_\infty$.

ii. Suppose the manufacturer of the bars guarantees that the bars delivered have cross sectional areas a_i, Young's moduli E_i and lengths ℓ_i that are within 5% of their requested values. (The requested areas and Young's moduli are the ones specified in Problem 1.10 and the requested lengths are the lengths determined from Figure 1.9.) Let $D = \text{ diag } (\ldots, a_i E_i/\ell_i, \ldots)$ be the diagonal matrix computed with the requested values and let $\widehat{D}$ be the corresponding diagonal matrix with the actual values for a_i, E_i, ℓ_i. Compute an upper bound for

$$\|D - \widehat{D}\|_1, \quad \|D - \widehat{D}\|_2, \quad \|D - \widehat{D}\|_\infty.$$

iii. Let $K = BDB^T$ and $\widehat{K} = B\widehat{D}B^T$. Use your results in i. and ii. to compute an upper bound for

$$\|K - \widehat{K}\|_1, \quad \|K - \widehat{K}\|_2, \quad \|K - \widehat{K}\|_\infty.$$

iv. Let $u = K^{-1}f$ and $\widehat{u} = \widehat{K}^{-1}f$. Apply Theorem 3.19 to compute upper bounds for

$$\|u - \widehat{u}\|_1/\|u\|_1, \quad \|u - \widehat{u}\|_2/\|u\|_2, \quad \|u - \widehat{u}\|_\infty/\|u\|_\infty.$$