MAT 460 Numerical Differential Equations

Version 28.10.14

Michael Karow, Andrea Dziubek

Lecture 5

Topics: Error Formulas for Linear Equation Systems, Condition of a Matrix, Matrix Norms

Inaccurate Input Data for Linear Equation Systems

Problem: Solve the equation $Ax = b$.

However, the input data A, b can be inaccurate due to the following reasons:

- A, b are measured values and not exactly known.
- A, b are results from earlier defective computations.
- The entries of A, b are not machine numbers and cannot be saved exactly in the computer.

Moreover, the algorithms to solve $Ax = b$ produce errors, which can be interpreted as errors in the data A, b. If we perform an LR-factorization

$$A = LR,$$

then numerically we get $\widetilde{L}, \widetilde{R}$ instead of L, R. The product is

$$\widetilde{L}\widetilde{R} = A + \Delta A.$$

In the best case we solve the following equation system when we do forward and backward substitution.

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

Problem: What is the effect of inaccurate input data on the solution of the linear equation system $Ax = b$?

Situation:

Exact Data: Solution:

$$(A, b) \longmapsto x = A^{-1}b$$

Inaccurate Data: Solution:

$$(\widetilde{A}, \widetilde{b}) \longmapsto \widetilde{x} = \widetilde{A}^{-1}\widetilde{b}$$

Question: How big is the difference between $\widetilde{x}$ and x?

Let $A\,x = b$ and $\widetilde{A}\,\widetilde{x} = \widetilde{b}$, where $det(A) \neq 0 \neq det(\widetilde{A})$.
Then we can estimate the error as follows.

The **absolute error**:

$$\|\widetilde{x} - x\| \leq (\|\widetilde{A}^{-1}\|\,\|x\|)\,\|\widetilde{A} - A\| + \|\widetilde{A}^{-1}\|\,\|\widetilde{b} - b\| \qquad (*)$$

The **relative error**:

$$\frac{\|\widetilde{x} - x\|}{\|x\|} \leq \|A\|\,\|\widetilde{A}^{-1}\| \left(\frac{\|\widetilde{A} - A\|}{\|A\|} + \frac{\|\widetilde{b} - b\|}{\|b\|} \right) \qquad (**)$$

$$\leq \frac{\text{cond}(A)}{1 - \frac{\|\widetilde{A} - A\|}{\|A\|}\text{cond}(A)} \left(\frac{\|\widetilde{A} - A\|}{\|A\|} + \frac{\|\widetilde{b} - b\|}{\|b\|} \right) \qquad (***)$$

Here $\text{cond}(A) = \|A\|\,\|A^{-1}\|$ is the **condition number** of A.

The inequalities $(*)$ and $(**)$ are always true when A and $\widetilde{A}$ are invertible, and when the inequality $\|My\| \leq \|M\|\,\|y\|$ is true with respect to a matrix norm for all matrices M and all vectors y.

The inequality $(***)$ is valid only when additionally $\frac{\|\widetilde{A} - A\|}{\|A\|}\text{cond}(A) < 1$.

Aim of this Lecture: Understand and 'prove' these error formulas.

Vector-Norms

Norm=measure for the size of the entries of a vector.

Definition:
A norm in $\mathbb{R}^n$ is a function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ with the following characteristics:

1) $\|x\| > 0$ for all $x \in \mathbb{R}^n$, $x \neq 0$,

2) $\|\lambda\, x\| = |\lambda|\, \|x\|$ for all $x \in \mathbb{R}^n$, $\lambda \in \mathbb{R}$

3) $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{R}^n$ (triangle inequality)

Often used norms:

a) Euklidean norm: $\|x\|_2 := \sqrt{\sum_{k=1}^n x_k^2}$

b) Sum-norm: $\|x\|_1 := \sum_{k=1}^n |x_k|$

c) Maximum-norm: $\|x\|_\infty := \max_{k=1}^n |x_k|$.

These norms belong to the family of Hölder-p-norms:

$$\|x\|_p := \left(\sum_{k=1}^n |x_k|^p \right)^{1/p}, \qquad 1 \leq p < \infty.$$

We write shortly

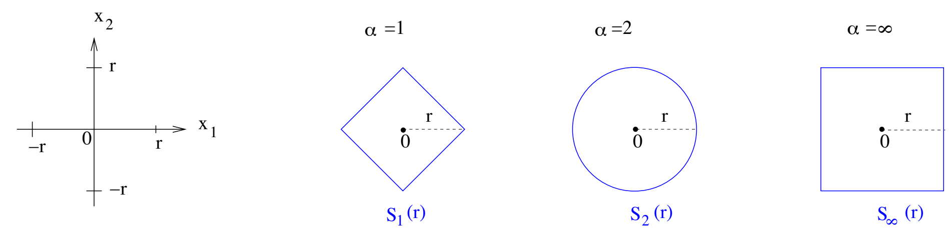
$$\|x\|_\infty = \lim_{p \to \infty} \|x\|_p.$$

Norm Spheres

The sphere to the norm $\|\cdot\|_\alpha$ and the radius $r > 0$ about the origin is the set of all vectors $x \in \mathbb{R}^n$ with $\|x\|_\alpha = r$. Formally:

$$S_\alpha(r) := \{\ x \in \mathbb{R}^n \mid\ \|x\|_\alpha = r\ \}.$$

Illustration: For the case $n = 2$ we have



For $n = 3$

- $S_1(r)$ is the surface of an oktahedron
- $S_2(r)$ is the surface of a sphere
- $S_\infty(r)$ is the surface of a cube

Note: Equivalence of Norms

Definitions: Two norms $\|\cdot\|$ and $|\cdot|$ are called equivalent if constants $c_1, c_2 > 0$ exist, such that for all $x \in \mathbb{R}^n$,

$$c_1\, \|x\| \leq |x| \leq c_2\, \|x\|.$$

If this is the case, then for all $x \in \mathbb{R}^n$

$$(1/c_2)\, |x| \leq \|x\| \leq (1/c_1)\, |x|.$$

Most important application of the norm-equivalence:
Let x_k be a sequence in $\mathbb{R}^n$, which converges to a point x_0 with respect to the norm $\|\cdot\|$, i. e.

$$\lim_{k \to \infty} \|x_k - x_0\| = 0.$$

Then this sequence also converges to x_0 with respect to any other norm $|\cdot|$ which is equivalent to the norm $\|\cdot\|$, i. e.

$$\lim_{k \to \infty} |x_k - x_0| = 0.$$

Theorem: All norms on $\mathbb{R}^n$ are equivalent.

Examples: for all $x \in \mathbb{R}^n$ we have

$$\|x\|_\infty \leq \|x\|_1 \leq n\, \|x\|_\infty, \qquad \|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\, \|x\|_\infty.$$

Induced Matrix Norms

Definition:
Let $A \in \mathbb{R}^{m \times n}$, and let $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$ be norms in $\mathbb{R}^n$ and $\mathbb{R}^m$ resp. Then the number

$$\|A\|_{\alpha,\beta} \ :=\ \max_{\|x\|_\alpha = 1} \|Ax\|_\beta = \max_{x \neq 0} \frac{\|Ax\|_\beta}{\|x\|_\alpha}$$

is the **matrix norm induced** by $\|\cdot\|_\alpha$ and $\|\cdot\|_\beta$.

Alternative definition:
$\|A\|_{\alpha,\beta}$ is the smallest number $c \geq 0$, such that $\|Ax\|_\beta \leq c\, \|x\|_\alpha$ for all $x \in \mathbb{R}^n$.

Interpretation:
$\|A\|_{\alpha,\beta}$ is the factor by which a vector x which is multiplied by the matrix A can be maximally stretched.

Characteristics:
1) $\|A\|_{\alpha,\beta} > 0$ for all $A \in \mathbb{R}^{m \times n}$, $A \neq 0$.
2) $\|\lambda\, A\|_{\alpha,\beta} = |\lambda|\, \|A\|_{\alpha,\beta}$ for all $A \in \mathbb{R}^{m \times n}$, $\lambda \in \mathbb{R}$
3) $\|A_1 + A_2\|_{\alpha,\beta} \leq \|A_1\|_{\alpha,\beta} + \|A_2\|_{\alpha,\beta}$ for all $A_1, A_2 \in \mathbb{R}^{m \times n}$.

4) $\|Ax\|_\beta \leq \|A\|_{\alpha,\beta}\, \|x\|_\alpha$ for all $x \in \mathbb{R}^n$. (special characteristic)

If $\|\cdot\|_\alpha = \|\cdot\|_\beta$, we write short: $\|A\|_\alpha := \|A\|_{\alpha,\alpha}$.

Usually the index α is dropped when it is clear which norm is meant.

Computation of $\|A\|_\infty$

To a matrix $A = [a_{ik}] \in \mathbb{R}^{m \times n}$ we define the row sums:

$$Z_i(A) = \sum_{k=1}^{n} |a_{ik}|, \quad i = 1, \ldots, m.$$

Theorem: We always have

$$\|A\|_\infty = \max_{\|x\|_\infty = 1} \|Ax\|_\infty = \max_i Z_i(A).$$

Proof: Let $y = Ax$. Then y has the components

$$y_i = a_{i1}\,x_1 + a_{i2}\,x_2 + \ldots + a_{ik}\,x_k + \ldots + a_{in}\,x_n,$$

and

$$\|Ax\|_\infty = \|y\|_\infty = \max\{\,|y_1|,\ |y_2|, \ldots |y_m|\,\}.$$

If $\|x\|_\infty = 1$, then $|x_k| \leq 1$ for all k and we estimate:

$$\begin{aligned}
|y_i| &= |a_{i1}\,x_1 + a_{i2}\,x_2 + \ldots + a_{ik}\,x_k + \ldots + a_{in}\,x_n| \\
&\leq |a_{i1}\,x_1| + |a_{i2}\,x_2| + \ldots + |a_{ik}\,x_k| + \ldots + |a_{in}\,x_n| \qquad (*) \\
&\leq |a_{i1}| + |a_{i2}| + \ldots + |a_{ik}| + \ldots + |a_{in}| \\
&= Z_i(A).
\end{aligned}$$

From this it follows that $\|A\|_\infty \leq \max_i Z_i(A)$.

The maximal possible value of $|y_i|$ under the condition $\|x\|_\infty = 1$ we get obviously when $x_k \in \{-1, 1\}$ and when the x_k have the same sign as the a_{ik}, $k = 1, \ldots, n$.
Then $a_{ik} x_i = |a_{ik}|$ and it follows that $|y_i| = y_i = Z_i(A)$. Doing this for the row i_0 that has maximal row sum we get $\|Ax\|_\infty = Z_{i_0}(A) = \max_i Z_i(A)$.

For the quantity $\max_{\|x\|_\alpha=1} \|Ax\|_\alpha$ we earlier introduced the shorter notation $\|A\|_\alpha$:

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

For the equally important quantity $\min_{\|x\|_\alpha=1} \|Ax\|_\alpha$ no similar (widely accepted) notation exist. This has the following reason.

Theorem: Let $A \in \mathbb{R}^{n \times n}$ be invertible (non singular). Then:

$$\frac{1}{\|A^{-1}\|_\alpha} = \min_{\|x\|_\alpha=1} \|Ax\|_\alpha = \min_{x \neq 0} \frac{\|Ax\|_\alpha}{\|x\|_\alpha}.$$

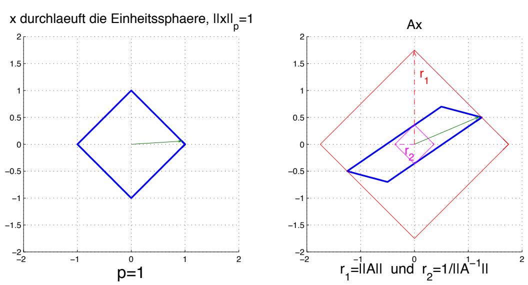
Proof:

$$\begin{aligned}
\|A^{-1}\|_\alpha &= \max_{y \neq 0} \frac{\|A^{-1}y\|_\alpha}{\|y\|_\alpha} \\
&= \max_{x \neq 0} \frac{\|A^{-1}(Ax)\|_\alpha}{\|Ax\|_\alpha} \qquad \text{let } y = Ax \\
&= \max_{x \neq 0} \frac{\|x\|_\alpha}{\|Ax\|_\alpha} \\
&= \frac{1}{\min_{x \neq 0} \frac{\|Ax\|_\alpha}{\|x\|_\alpha}}
\end{aligned}$$

In the last step we used the following straightforward fact:

Let M be a set of positive numbers and let M^{-1} be the set of the inverse of all numbers of M. Then $\max M = \frac{1}{\min M^{-1}}$.

Images to illustrate $\|A\|$ and $1/\|A^{-1}\|$



Explanation:

The thick blue curve in the left picture is the sphere

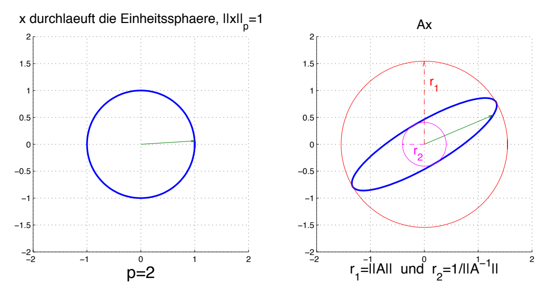
$$S_1(1) = \{\, x;\ \ \|x\|_1 = 1 \,\}.$$

The thick blue curve in the right picture is the A-image of the sphere:

$$\{\, Ax;\ \ \|x\|_1 = 1 \,\}, \qquad \text{where } A = \begin{bmatrix} 1.25 & 0.5 \\ 0.5 & 0.7 \end{bmatrix}.$$

The thin curves on the right are the spheres $S_1(r_1)$ and $S_1(r_2)$.

Images to illustrate $\|A\|$ and $1/\|A^{-1}\|$



Explanation:

The thick blue curve in the right picture is the sphere

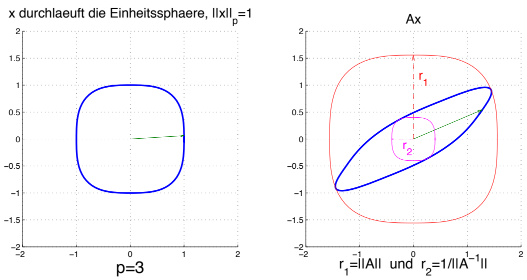
$$S_2(1) = \{\, x;\ \ \|x\|_2 = 1 \,\}.$$

The thick blue curve in the right picture is the A-image of the sphere:

$$\{\, Ax;\ \ \|x\|_2 = 1 \,\}, \qquad \text{where } A = \begin{bmatrix} 1.25 & 0.5 \\ 0.5 & 0.7 \end{bmatrix}.$$

The thin curves on the right are the spheres $S_2(r_1)$ and $S_2(r_2)$.

Images to illustrate $\|A\|$ and $1/\|A^{-1}\|$



Explanation:

The thick blue curve in the left picture is the sphere

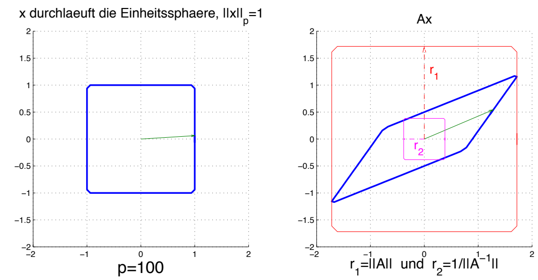
$$S_3(1) = \{\, x;\ \ \|x\|_3 = 1 \,\}.$$

The thick blue curve in the right picture is the A-image of the sphere:

$$\{\, Ax;\ \ \|x\|_3 = 1 \,\}, \qquad \text{where } A = \begin{bmatrix} 1.25 & 0.5 \\ 0.5 & 0.7 \end{bmatrix}.$$

The thin curves on the right are the spheres $S_3(r_1)$ and $S_3(r_2)$.

Images to illustrate $\|A\|$ and $1/\|A^{-1}\|$



Explanation:

The thick blue curve in the left picture is the sphere

$$S_{100}(1) = \{\, x;\ \ \|x\|_{100} = 1 \,\}.$$

The thick blue curve in the right picture is the A-image of the sphere:

$$\{\, Ax;\ \ \|x\|_{100} = 1 \,\}, \qquad \text{where } A = \begin{bmatrix} 1.25 & 0.5 \\ 0.5 & 0.7 \end{bmatrix}.$$

The thin curves on the right are the spheres $S_{100}(r_1)$ and $S_{100}(r_2)$.

Condition numbers of matrices

For an invertible matrix $A \in \mathbb{R}^{n\times n}$ we have

$$\|A\|_\alpha = \max_{\|x\|_\alpha=1} \|Ax\|_\alpha, \qquad \frac{1}{\|A^{-1}\|_\alpha} = \min_{\|x\|_\alpha=1} \|Ax\|_\alpha.$$

From this it follows:

$$\|A\|_\alpha \, \|A^{-1}\|_\alpha \;=\; \frac{\max_{\|x\|_\alpha=1} \|Ax\|_\alpha}{\min_{\|x\|_\alpha=1} \|Ax\|_\alpha}.$$

This quantity is called the condition number of A with respect to the vector norm $\|\cdot\|_\alpha$.
Notation:

$$\text{cond}_\alpha(A) \;:=\; \|A\|_\alpha \, \|A^{-1}\|_\alpha \;=\; \frac{\max_{\|x\|_\alpha=1} \|Ax\|_\alpha}{\min_{\|x\|_\alpha=1} \|Ax\|_\alpha}.$$

We note that the condition number is the quotient of the maximal and the minimal scale factor, when we multiply a vector x with the matrix A. We always have that

$$\text{cond}_\alpha(A) \ge 1 \quad \text{and} \quad \text{cond}_\alpha(A) = 1 \text{ if and only if } \|Ax\|_\alpha = \|x\|_\alpha \text{ for all } x \in \mathbb{R}^n.$$

The MATLAB-command to compute the condition number with respect to $\|\cdot\|_p$, $p = 1, 2, \infty$ is:

```
cond(A,p)
```

Extremal values of a quadratic form

Theorem: Let $S \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Let $\lambda_{\text{min}}, \lambda_{\text{max}} \in \mathbb{R}$ be the maximal and minimal eigenvalues of S and let $\underline{v}$, $\overline{v} \in \mathbb{R}^n$ be the corresponding normalized eigenvectors, i. e.:

$$S\underline{v} = \lambda_{\text{min}}\, \underline{v}, \qquad S\overline{v} = \lambda_{\text{max}}\, \overline{v}, \qquad \|\underline{v}\|_2 = \|\overline{v}\|_2 = 1.$$

Then:

$$\min_{x\neq 0} \frac{x^T S x}{\|x\|_2^2} = \min_{\|x\|_2=1} x^T S x = \underline{v}^T S \underline{v} = \lambda_{\text{min}}$$

$$\max_{x\neq 0} \frac{x^T S x}{\|x\|_2^2} = \max_{\|x\|_2=1} x^T S x = \overline{v}^T S \overline{v} = \lambda_{\text{max}}.$$

Notation: The quotient $\frac{x^T S x}{\|x\|_2^2}$ is called Rayleigh-quotient.

Proof: Let $\lambda_{\text{max}} = \lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_n = \lambda_{\text{min}}$ be the eigenvalues of S and let $\overline{v} = v_1, v_2, \ldots, v_n = \underline{v}$ be an orthonormal basis of eigenvalues, $Sv_k = \lambda_k v_k$. Every vector $x \in \mathbb{R}^n$ can be written as a linear combination of the eigenvectors:

$$x = x_1 v_1 + x_2 v_2 + \ldots + x_n v_n, \qquad x_k \in \mathbb{R}.$$

We compute

$$\frac{x^T S x}{\|x\|_2^2} = \frac{\lambda_1 x_1^2 + \lambda_2 x_2^2 + \ldots + \lambda_n x_n^2}{x_1^2 + x_2^2 + \ldots + x_n^2}.$$

This quotient is maximal e. g. when $x_1 = 1$ and $x_2 = \ldots = x_n = 0$.
This quotient is minimal e. g. when $x_n = 1$ and $x_1 = \ldots = x_{n-1} = 0$.

The 2-norm of a matrix

Theorem: For every matrix $A \in \mathbb{R}^{m\times n}$ we have

$$\|A\|_2 = \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sqrt{\lambda_{\max}(A^TA)},$$

$$\min_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sqrt{\lambda_{\min}(A^TA)}.$$

Where $\lambda_{\max}$ is the largest eigenvalue and $\lambda_{\min}$ is the smallest eigenvalue of the positive semi-definite symmetric matrix A^TA.

Proof: We have $\|Ax\|^2 = (Ax)^T(Ax) = x^TA^TAx$ and so

$$\|A\|_2^2 = \max_{x\neq 0} \frac{\|Ax\|_2^2}{\|x\|_2^2} = \max_{x\neq 0} \frac{x^TA^TAx}{\|x\|_2^2} = \lambda_{\max}(A^TA).$$

In the last equation we used the theorem of the maximal Rayleigh-quotient.
The proof for the minimum is analogous.

Notation:

1. The squares of the eigenvalues of A^TA are called **singular values** of A.

   Notation: $\sigma_k(A) := \sqrt{\lambda_k(A^TA)}$,

   In particular: $\sigma_{\max}(A) = \sqrt{\lambda_{\max}(A^TA)}$, $\quad \sigma_{\min}(A) = \sqrt{\lambda_{\min}(A^TA)}$.

2. The 2-norm $\|A\|_2$ is also called **spectral norm**
   (spectrum=set of the eigenvalues of a matrix)

With this notations we have

$$\|A\|_2 = \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sigma_{\max}(A),$$

$$\min_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sigma_{\min}(A).$$

Recall: If $A \in \mathbb{R}^{n\times n}$ is invertible, then

$$\min_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \frac{1}{\|A^{-1}\|_2}.$$

And so:

$$cond_2(A) = \|A\|_2\,\|A^{-1}\|_2 = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}.$$

Summary: The most important matrix norms

Let $A = [a_{ik}] \in \mathbb{R}^{m\times n}$.

- $\|A\|_\infty = \max_{i=1}^m \sum_{k=1}^n |a_{ik}|$ (row sum norm)
- $\|A\|_2 = \sqrt{\lambda_{\max}(A^TA)} = \sigma_{\max}(A)$ (spectral norm)
- $\|A\|_1 = \max_{k=1}^n \sum_{i=1}^m |a_{ik}|$ (column sum norm)

Induced matrix norms are sub-multiplicative.

Let $\|\cdot\|_\alpha$ be any vector norm in $\mathbb{R}^n$. Then the induced matrix norm satisfies

$$\|AB\|_\alpha \leq \|A\|_\alpha\,\|B\|_\alpha, \qquad A, B \in \mathbb{R}^{n\times n}.$$

Proof: According to the definition of induced matrix norms we have

$$\|ABx\|_\alpha \leq \|A\|_\alpha\,\|Bx\|_\alpha.$$

And so

$$\|AB\|_\alpha = \max_{x\neq 0} \frac{\|ABx\|_\alpha}{\|x\|_\alpha} \leq \max_{x\neq 0} \frac{\|A\|_\alpha\,\|Bx\|_\alpha}{\|x\|_\alpha} = \|A\|_\alpha \max_{x\neq 0} \frac{\|Bx\|_\alpha}{\|x\|_\alpha} = \|A\|_\alpha\,\|B\|_\alpha.$$

Derivation of error formulas

Derivation of error formulas I

Let $A, \widetilde{A} \in \mathbb{R}^{n\times n}$ be invertible, and $Ax = b$, $\widetilde{A}\widetilde{x} = \widetilde{b}$. Then it follows

$$\begin{aligned}\widetilde{x} - x &= \widetilde{A}^{-1}\widetilde{b} - x \\ &= \widetilde{A}^{-1}b - x + \widetilde{A}^{-1}(\widetilde{b} - b) \\ &= \widetilde{A}^{-1}(A - \widetilde{A})x + \widetilde{A}^{-1}(\widetilde{b} - b)\end{aligned}$$

$\Rightarrow$

$$\begin{aligned}\|\widetilde{x} - x\| &= \|\widetilde{A}^{-1}(A - \widetilde{A})x + \widetilde{A}^{-1}(\widetilde{b} - b)\| \\ &\le \|\widetilde{A}^{-1}\|\,\|A - \widetilde{A}\|\,\|x\| + \|\widetilde{A}^{-1}\|\,\|\widetilde{b} - b\| \\ &= \|\widetilde{A}^{-1}\|\,(\|A - \widetilde{A}\|\,\|x\| + \|\widetilde{b} - b\|)\end{aligned}$$

$\Rightarrow$

$$\begin{aligned}\frac{\|\widetilde{x} - x\|}{\|x\|} &\le \|\widetilde{A}^{-1}\|\,\|A\|\left(\frac{\|(A - \widetilde{A}\|}{\|A\|} + \frac{\|b\|}{\|A\|\,\|x\|}\,\frac{\|\widetilde{b} - b\|}{\|b\|}\right) \\ &\le \|\widetilde{A}^{-1}\|\,\|A\|\left(\frac{\|(A - \widetilde{A}\|}{\|A\|} + \frac{\|\widetilde{b} - b\|}{\|b\|}\right). \qquad (\|b\| = \|Ax\| \le \|A\|\,\|x\|)\end{aligned}$$

By this we proved the error formulas $(*)$ and $(**)$ from the beginning of the lecture.

To prove the error formula $(***)$ we have to estimate the factor $\|\widetilde{A}^{-1}\|\,\|A\|$ by its condition number. We do this on the next slide.

Derivation of error formulas II

for all $y \in \mathbb{R}^n$ is

$$\|Ay\| = \|\widetilde{A}y + (A - \widetilde{A})y\| \le \|\widetilde{A}y\| + \|(A - \widetilde{A})y\|.$$

$\Rightarrow$

$$\|\widetilde{A}y\| \ge \|Ay\| - \|(A - \widetilde{A})y\|.$$

$\Rightarrow$

$$\frac{\|\widetilde{A}y\|}{\|y\|} \ge \frac{\|Ay\|}{\|y\|} - \frac{\|(A - \widetilde{A})y\|}{\|y\|} \ge \frac{\|Ay\|}{\|y\|} - \|A - \widetilde{A}\|$$

$\Rightarrow$

$$\min_{y\neq 0}\frac{\|\widetilde{A}y\|}{\|y\|} \ge \min_{y\neq 0}\frac{\|Ay\|}{\|y\|} - \|A - \widetilde{A}\|$$

$\Rightarrow$

$$\frac{1}{\|\widetilde{A}^{-1}\|} \ge \frac{1}{\|A^{-1}\|} - \|A - \widetilde{A}\|$$

$\Rightarrow$

$$\|\widetilde{A}^{-1}\| \le \frac{1}{\frac{1}{\|A\|} - \|A - \widetilde{A}\|} = \frac{\|A^{-1}\|}{1 - \|A - \widetilde{A}\|\,\|A^{-1}\|}$$

$\Rightarrow$

$$\|A\|\,\|\widetilde{A}^{-1}\| \le \frac{\|A\|\,\|A^{-1}\|}{1 - \frac{\|A - \widetilde{A}\|}{\|A\|}\|A\|\,\|A^{-1}\|} = \frac{\text{cond}(A)}{1 - \frac{\|A - \widetilde{A}\|}{\|A\|}\text{cond}(A)}.$$

From this and from the last inequality on the last slide we get the error formula $(***)$.

By this we showed that

$$\frac{\|\widetilde{x} - x\|}{\|x\|} \le \frac{\text{cond}(A)}{1 - \frac{\|\widetilde{A} - A\|}{\|A\|}\text{cond}(A)}\left(\frac{\|\widetilde{A} - A\|}{\|A\|} + \frac{\|\widetilde{b} - b\|}{\|b\|}\right) \qquad (***)$$

Special case:

$$A = \widetilde{A} \qquad \Rightarrow \qquad \frac{\|\widetilde{x} - x\|}{\|x\|} \le \text{cond}(A)\,\frac{\|\widetilde{b} - b\|}{\|b\|}$$

Once again: Inaccurate Input Data for Linear Equation Systems

Problem: Solve the equation $Ax = b$.

However, the input data A, b can be inaccurate due to the following reasons:

- A, b are measured values and not exactly known.
- A, b are results from earlier defective computations.
- The entries of A, b are not machine numbers and cannot be saved exactly in the computer.

Moreover, the algorithms to solve $Ax = b$ produce errors, which can be interpreted as errors in the data A, b. If we perform an LR-factorization

$$A = LR,$$

numerically we get $\widetilde{L}, \widetilde{R}$ instead of L, R. The product is

$$\widetilde{L}\widetilde{R} = A + \Delta A.$$

In the best case we solve the following equation system when we do forward and backward substitution.

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

Practial consequence of the error formulas

From the error formula

$$\frac{\|\widetilde{x} - x\|}{\|x\|} \leq \frac{\text{cond}(A)}{1 - \frac{\|\widetilde{A}-A\|}{\|A\|}\text{cond}(A)} \left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|} \right)$$

we can derive the following rule :

A condition number $cond(A) = 10^q$ costs q digits accuracy when solving $Ax = b$.

What is th use of checking for ill-conditioned problems?

Problem:	$Ax = b$
Exact Solution:	$x = A^{-1}b$
Numerical Solution:	$\widetilde{x}$
Check:	$A\widetilde{x} = \widetilde{b}$

Assume, the product $A\widetilde{x}$ was computed exact and the relative error $\|b - \widetilde{b}\|/\|b\|$ is small. Can we conclude then that also the relative error $\|x - \widetilde{x}\|/\|x\|$ is small?

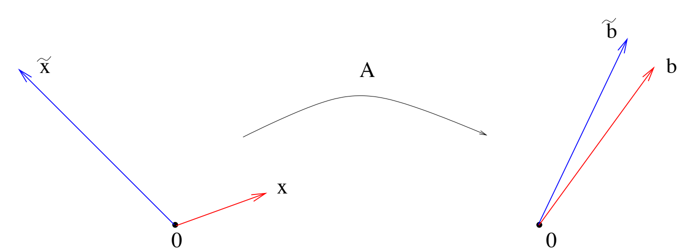
Answer: It depends on the condition number. The error formula

$$\frac{\|x - \widetilde{x}\|}{\|x\|} \leq \text{cond}(A)\frac{\|b - \widetilde{b}\|}{\|b\|}$$

tells us that the relative error in b in the worst case increases about the factor cond(A).

Illustration of the situation when a matrix is ill-conditioned:

Vectors which are relative far from each other $x, \widetilde{x}$ are transformed to vectors which are relative close to each other $b = Ax, \widetilde{b} = A\widetilde{x}$.



Situation when the right side is inaccurate:

We only know $\widetilde{b}$, the exact right side b is not known. If the equation system is solved exact, we get $\widetilde{x}$. But the solution x to the exact right side b can not be be too far away.

Situation when we check:

The right side is b. We have a defective computed solution $\widetilde{x}$. Checking gives $A\widetilde{x} = \widetilde{b}$. Even if b and $\widetilde{b}$ are almost equal, exact and computed solution can be very different.

The condition number of a matrix is large, when its rows and columns are linearly dependent.

Example: Let $A_\epsilon = \begin{bmatrix} 1+\epsilon & 3 \\ 2 & 6 \end{bmatrix}$.

The rows and columns of A_0 are linearly dependent. For $\epsilon \neq 0$:

$$A_\epsilon^{-1} = \frac{1}{det(A_\epsilon)} \begin{bmatrix} 6 & -3 \\ -2 & 1+\epsilon \end{bmatrix} = \frac{1}{6\,\epsilon} \begin{bmatrix} 6 & -3 \\ -2 & 1+\epsilon \end{bmatrix}.$$

The condition number of A_ϵ with respect to the row sum norm for $\epsilon \in [-6, 4]$ is:

$$\text{cond}_\infty(A_\epsilon) = \|A_\epsilon\|_\infty \, \|A_\epsilon^{-1}\|_\infty = (2+6)\,\frac{6+3}{6\,|\epsilon|} = \frac{12}{|\epsilon|} \to \infty \qquad \text{für } \epsilon \to 0.$$

Note: In this example the condition number is large for small ϵ because $det(A_\epsilon)$ is small. A small determinant implicates not necessarily a small condition number. Example :

$$\text{cond}(\epsilon\, I) = \|\epsilon I\| \, \|(\epsilon\, I)^{-1}\| = 1 \qquad \text{for all } \epsilon > 0.$$

Improve the condition number by pre-conditioning

Problem: Solve

$$A\,x = b. \qquad (*)$$

By multiplication of the equation with a non singular matrix $D \in \mathbb{R}^{n\times n}$ we get the equivalent equation

$$DA\,x = Db. \qquad (**)$$

If the condition number of A is large, then we search a matrix D with

$$\text{cond}(DA) << \text{cond}(A)$$

and solve $(**)$ instead of $(*)$.

Simplest option:

Chose D as diagonal matrix, such that all rows of DA have the same 1-norm **(row equilibration)**.