Vorlesung: Numerik 1 für Ingenieure

Version 28.10.14

Michael Karow

5. Vorlesung

Thema: Fehlerformeln für lineare Gleichungssysteme

Kondition einer Matrix, Matrixnormen

Ungenaue Eingangsdaten bei linearen Gleichungssystemen

Aufgabe: Es soll die Gleichung $Ax = b$ gelöst werden.

Die Ausgangsdaten A, b können aber z.B. aus folgenden Gründen ungenau sein:

- A, b sind Messwerte und daher nicht genau bekannt.
- A, b sind Ergebnisse früherer (fehlerhafter) Rechnungen
- Die Einträge von A, b sind keine Maschinenzahlen und können nicht exakt im Rechner abgespeichert werden.

Außerdem machen Algorithmen zur Lösung von $Ax = b$ Fehler, die man als Fehler in den Daten A, b interpretieren kann. Macht man z.B. eine LR-Zerlegung

$$A = LR,$$

dann bekommt man statt L, R numerisch $\widetilde{L}, \widetilde{R}$ heraus. Das Produkt ist

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

Beim Vorwärts-Rückwärtseinsetzen löst man also im besten Fall das Gleichungssystem

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

Problem: Wie wirken sich ungenaue Eingangsdaten auf die Lösung eines linearen Gleichungssystems $Ax = b$ aus?

Situation:

Exakte Daten: Lösung:

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

Ungenaue Daten: Lösung:

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

Frage: Wie stark unterscheidet sich $\widetilde{x}$ von x ?

Sei $A\,x = b$ und $\widetilde{A}\,\widetilde{x} = \widetilde{b}$, wobei $det(A) \neq 0 \neq det(\widetilde{A})$.
Dann gelten die folgenden Fehlerabschätzungen.

Für den **absoluten Fehler**:

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

Für den **relativen Fehler**:

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

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

Dabei ist $\mathrm{cond}(A) = \|A\|\,\|A^{-1}\|$ die **Konditionszahl** von A.

Die Ungleichungen $(*)$ und $(**)$ gelten stets, wenn A und $\widetilde{A}$ invertierbar sind, und wenn für die zugrunde liegenden Normen die Ungleichung $\|My\| \leq \|M\|\,\|y\|$ für alle Matrizen M und alle Vektoren y erfüllt ist.

Die Ungleichung $(***)$ gilt nur unter der zusätzlichen Bedingung, dass $\frac{\|\widetilde{A} - A\|}{\|A\|}\,\mathrm{cond}(A) < 1$.

Ziel der Vorlesung: Diese Fehlerformeln erklären und 'beweisen'.

Vektor-Normen

Norm=Maß für die Größe der Einträge in einem Vektor.

Definition:

Eine Norm auf $\mathbb{R}^n$ ist eine Funktion $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ mit den folgenden Eigenschaften:

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

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

3) $\|x + y\| \leq \|x\| + \|y\|$ für alle $x, y \in \mathbb{R}^n$ (Dreiecksungeichung)

Häufig verwendete Normen:

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

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

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

Diese Normen gehören zur Familie der Hölder-p-Normen:

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

Es ist

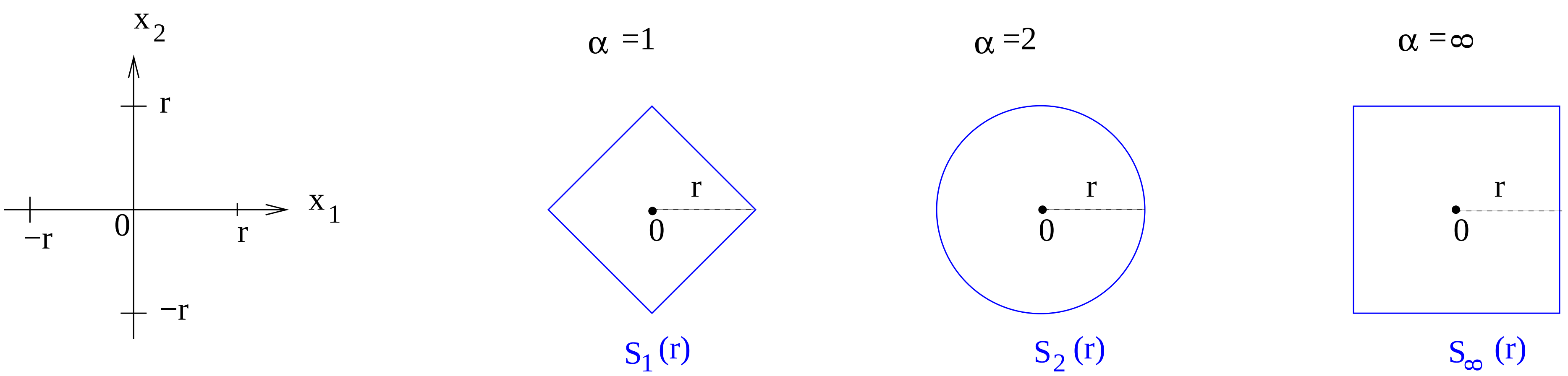
$$\|x\|_\infty = \lim_{p \to \infty} \|x\|_p.$$

Die Normsphären

Die Sphäre zur Norm $\|\cdot\|_\alpha$ und zum Radius $r > 0$ um den Nullpunkt ist die Menge aller Vektoren $x \in \mathbb{R}^n$ mit $\|x\|_\alpha = r$. Formal:

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

Anschauung: Für den Fall $n = 2$ hat man



Im Fall $n = 3$ ist

- $S_1(r)$ eine Oktaederoberfläche
- $S_2(r)$ eine Kugeloberfläche (Sphäre im engeren Sinne)
- $S_\infty(r)$ eine Würfeloberfläche.

Bemerkung: Äquivalenz von Normen

Definition: Zwei Normen $\|\cdot\|$ und $|\cdot|$ heissen äquivalent, falls es Konstanten $c_1, c_2 > 0$ gibt, so dass für alle $x \in \mathbb{R}^n$,

$$c_1 \|x\| \le |x| \le c_2 \|x\|.$$

Ist dies der Fall, dann folgt für alle $x \in \mathbb{R}^n$

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

Wichtigste Anwendung der Norm-Äquivalenz:
Sei x_k eine Folge in $\mathbb{R}^n$, die bzgl. der Norm $\|\cdot\|$ gegen den Punkt x_0 konvergiert, d.h.

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

Dann konvergiert diese Folge auch bzgl. jeder zu $\|\cdot\|$ äquivalenten Norm $|\cdot|$ gegen x_0, d.h. man hat

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

Satz: Alle Normen auf $\mathbb{R}^n$ sind äquivalent.

Beispiele: Für alle $x \in \mathbb{R}^n$ gilt

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

Induzierte Normen von Matrizen

Definition:

Sei $A \in \mathbb{R}^{m\times n}$, und seien $\|\cdot\|_\alpha$ und $\|\cdot\|_\beta$ Normen auf $\mathbb{R}^n$ bzw. $\mathbb{R}^m$. Dann ist die Zahl

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

die durch $\|\cdot\|_\alpha$ und $\|\cdot\|_\beta$ **induzierte Matrixnorm**.

Alternative Definition:

$\|A\|_{\alpha,\beta}$ ist die kleinste Zahl $c \geq 0$, so daß für alle $x \in \mathbb{R}^n$ gilt $\quad \|Ax\|_\beta \leq c\,\|x\|_\alpha$.

Interpretation:

$\|A\|_{\alpha,\beta}$ ist der Faktor, um den ein Vektor x bei Multiplikation mit A maximal gestreckt werden kann.

Eigenschaften:

1) $\|A\|_{\alpha,\beta} > 0$ für alle $A \in \mathbb{R}^{m\times n}$, $A \neq 0$.
2) $\|\lambda\, A\|_{\alpha,\beta} = |\lambda|\,\|A\|_{\alpha,\beta}$ für alle $A \in \mathbb{R}^{m\times n}$, $\lambda \in \mathbb{R}$
3) $\|A_1 + A_2\|_{\alpha,\beta} \leq \|A_1\|_{\alpha,\beta} + \|A_2\|_{\alpha,\beta}$ für alle $A_1, A_2 \in \mathbb{R}^{m\times n}$.
4) $\|Ax\|_\beta \leq \|A\|_{\alpha,\beta}\,\|x\|_\alpha$ für alle $x \in \mathbb{R}^n$. (Besondere Eigenschaft)

Ist $\|\cdot\|_\alpha = \|\cdot\|_\beta$, dann schreibt man kurz: $\quad \|A\|_\alpha := \|A\|_{\alpha,\alpha}$.

Wenn klar ist, welche Norm gemeint ist, dann lässt man den Index α weg.

Berechnung von $\|A\|_\infty$

Zu einer Matrix $A = [a_{ik}] \in \mathbb{R}^{m \times n}$ definieren wir die Zeilensummen:

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

Satz: Es gilt stets

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

Beweis: Sei $y = Ax$. Dann hat y die Komponenten

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

und es ist

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

Wenn $\|x\|_\infty = 1$, dann ist $|x_k| \leq 1$ für alle k und man hat die Abschätzung:

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

Daraus folgt schon mal, dass $\quad \|A\|_\infty \leq \max_i Z_i(A)$.

Den maximal möglichen Wert von $|y_i|$ unter der Nebenbedingung $\|x\|_\infty = 1$ bekommt man offensichtlich dann, wenn $x_k \in \{-1, 1\}$ und x_k dasselbe Vorzeichen hat wie a_{ik}, $k = 1, \ldots, n$. Dann ist $a_{ik} x_i = |a_{ik}|$ und folglich $|y_i| = y_i = Z_i(A)$. Macht man dies für eine Zeile i_0 mit maximaler Zeilensumme, dann bekommt man $\|Ax\|_\infty = Z_{i_0}(A) = \max_i Z_i(A)$.

Für die Größe $\max_{\|x\|_\alpha=1} \|Ax\|_\alpha$ hatten wir die Bezeichnung $\|A\|_\alpha$ eingeführt:

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

Für die ebenfalls wichtige Größe $\min_{\|x\|_\alpha=1} \|Ax\|_\alpha$ gibt es keine (allgemein anerkannte) Notation. Dies hat folgenden Grund.

Satz: Sei $A \in \mathbb{R}^{n\times n}$ invertierbar (nicht singulär). Dann gilt:

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

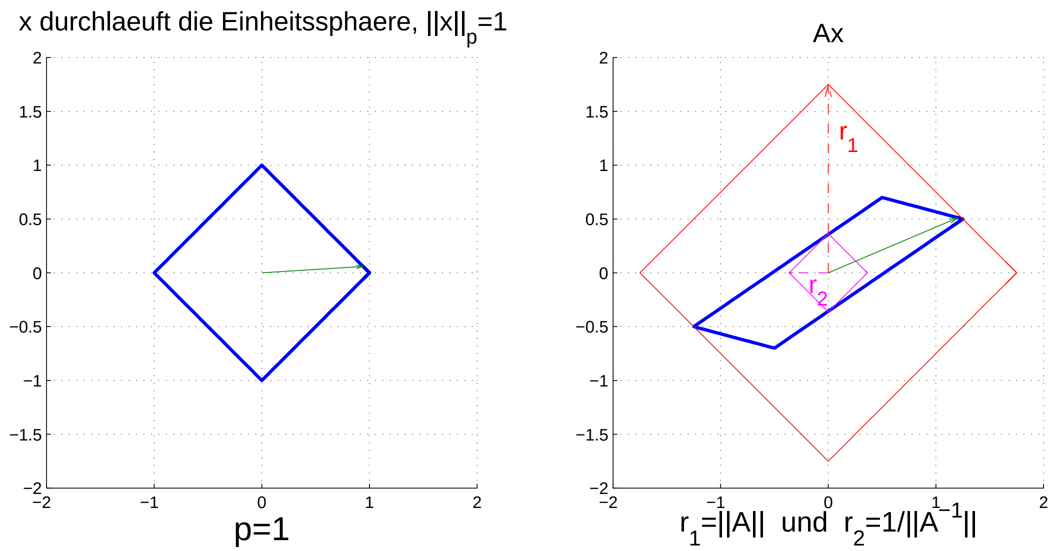
Beweis:

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

Beim letzten Schritt wurde folgende leicht einsehbare Tatsache benutzt:

Sei M eine Menge positiver Zahlen und M^{-1} die Menge der Kehrwerte aller Zahlen aus M. Dann ist $\max M = \frac{1}{\min M^{-1}}$.

Bilder zur Veranschaulichung von $\|A\|$ und $1/\|A^{-1}\|$



Erklärung:

Die dicke blaue Kurve im linken Bild ist die Sphäre

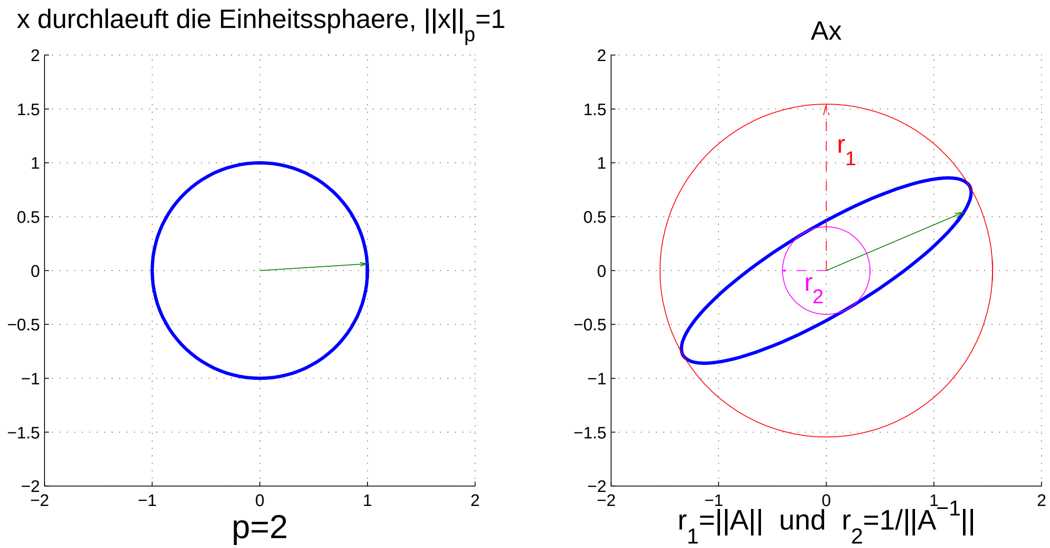
$$S_1(1) = \{\ x;\ \ \|x\|_1 = 1\ \}.$$

Die dicke blaue Kurve im rechten Bild ist das A-Bild der Sphäre:

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

Die dünnen Kurven rechts sind die Sphären $S_1(r_1)$ und $S_1(r_2)$.

Bilder zur Veranschaulichung von $\|A\|$ und $1/\|A^{-1}\|$



Erklärung:

Die dicke blaue Kurve im linken Bild ist die Sphäre

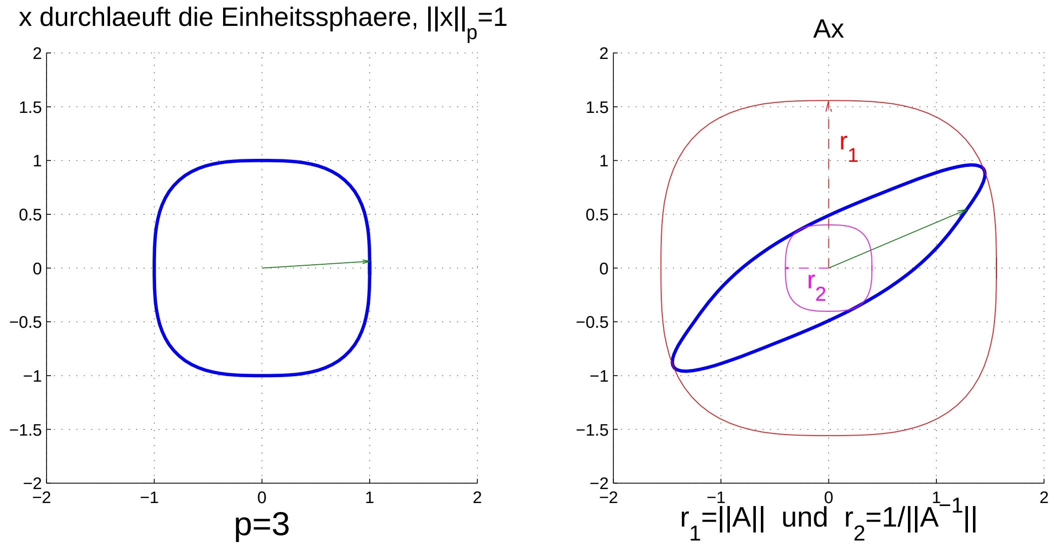
$$S_2(1) = \{\ x;\ \ \|x\|_2 = 1\ \}.$$

Die dicke blaue Kurve im rechten Bild ist das A-Bild der Sphäre:

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

Die dünnen Kurven rechts sind die Sphären $S_2(r_1)$ und $S_2(r_2)$.

Bilder zur Veranschaulichung von $\|A\|$ und $1/\|A^{-1}\|$



Erklärung:

Die dicke blaue Kurve im linken Bild ist die Sphäre

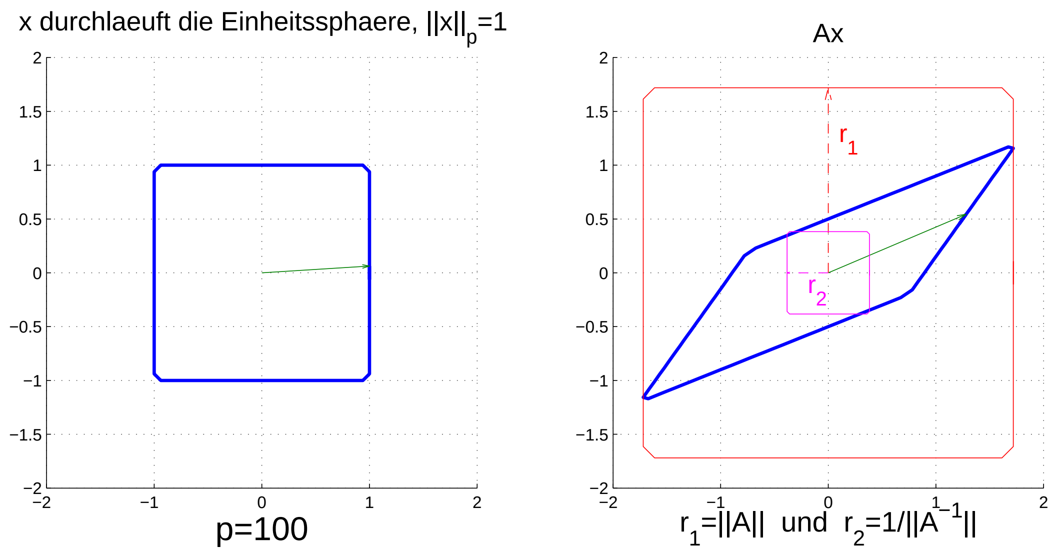
$$S_3(1) = \{ \; x; \;\; \|x\|_3 = 1 \; \}.$$

Die dicke blaue Kurve im rechten Bild ist das A-Bild der Sphäre:

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

Die dünnen Kurven rechts sind die Sphären $S_3(r_1)$ und $S_3(r_2)$.

Bilder zur Veranschaulichung von $\|A\|$ und $1/\|A^{-1}\|$



Erklärung:

Die dicke blaue Kurve im linken Bild ist die Sphäre

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

Die dicke blaue Kurve im rechten Bild ist das A-Bild der Sphäre:

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

Die dünnen Kurven rechts sind die Sphären $S_{100}(r_1)$ und $S_{100}(r_2)$.

Konditionszahlen von Matrizen

Für eine invertierbare Matrix $A \in \mathbb{R}^{n\times n}$ haben wir

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

Daraus folgt:

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

Diese Größe heißt Konditionszahl von A bezüglich der Vektornorm $\|\cdot\|_\alpha$. Notation:

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

Die Konditionszahl ist also der Quotient aus dem maximalen und dem minimalen Streckfaktor, wenn man einen Vektor x mit der Matrix A multipliziert. Es gilt stets

$\mathrm{cond}_\alpha(A) \geq 1$ und $\mathrm{cond}_\alpha(A) = 1$ genau dann, wenn $\|Ax\|_\alpha = \|x\|_\alpha$ für alle $x \in \mathbb{R}^n$.

MATLAB-Anweisung zur Berechnung der Konditionszahl bzgl. $\|\cdot\|_p$, $p = 1, 2, \infty$:

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

Extremwerte einer quadratischen Form

Satz: Sei $S \in \mathbb{R}^{n\times n}$ eine symmetrische Matrix. Seien λ_{min}, $\lambda_{\text{max}} \in \mathbb{R}$ der maximale und der minimale Eigenwert von S und seien $\underline{v}$, $\overline{v} \in \mathbb{R}^n$ zugehörige normierte Eigenvektoren, d.h.:

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

Dann gilt:

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

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

Terminologie: Der Quotient $\frac{x^T S x}{\|x\|_2^2}$ heißt Rayleigh-Quotient.

Beweis: Seien $\lambda_{\text{max}} = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n = \lambda_{\text{min}}$ die Eigenwerte von S und sei $\overline{v} = v_1, v_2, \ldots, v_n = \underline{v}$ eine Orthonormalbasis von Eigenvektoren, $Sv_k = \lambda_k\, v_k$. Jeder Vektor $x \in \mathbb{R}^n$ ist eine Linearkombination der Eigenvektoren:

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

Eine direkte Rechnung ergibt

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

Das Maximum dieses Quotienten wird z.B. angenommen, wenn $x_1 = 1$ und $x_2 = \ldots = x_n = 0$. Das Minimum wird z.B. angenommen, wenn $x_n = 1$ und $x_1 = \ldots = x_{n-1} = 0$.

Die 2-Norm einer Matrix

Satz: Für jede Matrix $A \in \mathbb{R}^{m\times n}$ gilt

$$\begin{aligned} \|A\|_2 = \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} &= \sqrt{\lambda_{\mathsf{max}}(A^TA)}, \\ \min_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} &= \sqrt{\lambda_{\mathsf{min}}(A^TA)}. \end{aligned}$$

Dabei sind λ_{max} und λ_{min} der größte und der kleinste Eigenwert der positiv semidefiniten symmetrischen Matrix A^TA.

Beweis: Man hat $\|Ax\|^2 = (Ax)^T(Ax) = x^TA^TAx$ und daher

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

Bei der letzten Gleichung wurde der Satz über das Maximum des Rayleigh-Quotienten benutzt. Der Beweis für das Minimum ist analog.

Terminologie:

1. Die Wurzeln aus den Eigenwerten von $A^T A$ nennt man **Singulärwerte** von A.

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

   Insbesondere: $\sigma_{\text{max}}(A) = \sqrt{\lambda_{\text{max}}(A^T A)}$, $\quad \sigma_{\text{min}}(A) = \sqrt{\lambda_{\text{min}}(A^T A)}$.

2. Die 2-Norm $\|A\|_2$ nennt man auch **Spektralnorm**
   (Spektrum=Menge der Eigenwerte einer Matrix)

Mit den obigen Bezeichungen hat man

$$\begin{aligned} \|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} &= \sigma_{\text{max}}(A), \\ \min_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} &= \sigma_{\text{min}}(A). \end{aligned}$$

Erinnerung: Wenn $A \in \mathbb{R}^{n \times n}$ invertierbar, dann ist

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

Daher:

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

Überblick: Die wichtigsten Matrixnormen

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

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

- $\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)} = \sigma_{\max}(A)$ (Spektralnorm)

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

Induzierte Matrixnormen sind submultiplikativ.

Sei $\|\cdot\|_\alpha$ irgendeine Vektornorm auf $\mathbb{R}^n$. Dann gilt für die induzierte Matrixnorm

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

Beweis: Nach Definition der induzierten Matrixnorm hat man

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

Somit

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

Herleitung der Fehlerformeln

Herleitung der Fehlerformeln I

Seien $A, \widetilde{A} \in \mathbb{R}^{n\times n}$ invertierbar, $Ax = b$, $\widetilde{A}\widetilde{x} = \widetilde{b}$. Dann folgt

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

$\Rightarrow$

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

$\Rightarrow$

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

Damit sind die ersten beiden Fehlerformeln $(*)$ und $(**)$ vom Anfang der Vorlesung bewiesen. Zum Beweis der Fehlerformel $(***)$ müssen wir den Faktor $\|\widetilde{A}^{-1}\|\,\|A\|$ durch die Konditionszahl abschätzen. Dies geschieht auf der nächsten Seite.

Herleitung der Fehlerformeln II

Für alle $y \in \mathbb{R}^n$ ist

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

$\Rightarrow$

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

$\Rightarrow$

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

$\Rightarrow$

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

$\Rightarrow$

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

$\Rightarrow$

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

$\Rightarrow$

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

Hieraus und aus der letzten Unlgleichung auf der vorherigen Seite folgt die Fehlerformel $(***)$.

Wir haben eben gezeigt, dass

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

Spezialfall:

$$A=\tilde{A} \qquad \Rightarrow \qquad \frac{\|\widetilde{x}-x\|}{\|x\|} \leq \mathrm{cond}(A)\,\frac{\|\widetilde{b}-b\|}{\|b\|}$$

Noch einmal: Ungenaue Eingangsdaten bei linearen Gleichungssystemen

Aufgabe: Es soll die Gleichung $Ax = b$ gelöst werden.

Die Ausgangsdaten A, b können aber z.B. aus folgenden Gründen fehlerhaft sein:

- A, b sind Messwerte und daher nicht genau bekannt.
- A, b sind Ergebnisse früherer (fehlerhafter) Rechnungen
- Die Einträge von A, b sind keine Maschinenzahlen und können nicht exakt im Rechner abgespeichert werden.

Außerdem machen Algorithmen zur Lösung von $Ax = b$ Fehler, die man als Fehler in den Daten A, b interpretieren kann. Macht man z.B. eine LR-Zerlegung

$$A = LR,$$

dann bekommt man statt L, R numerisch $\widetilde{L}, \widetilde{R}$ heraus. Das Produkt ist

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

Beim Vorwärts-Rückwärtseinsetzen löst man also im besten Fall das Gleichungssystem

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

Praktische Konsequenz der Fehlerformeln

Aus der Fehlerformel

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

lässt sich folgende Faustregel ableiten:

Eine Konditionszahl $cond(A) = 10^q$ kostet q Stellen Genauigkeit bei der Lösung von $Ax = b$.

Was nützt eine Probe bei schlecht konditionierten Problemen?

Problem:	$Ax = b$
Exakte Lösung:	$x = A^{-1}b$
Numerische Lösung:	$\widetilde{x}$
Probe ergibt:	$A\widetilde{x} = \widetilde{b}$

Angenommen, das Produkt $A\widetilde{x}$ wurde exakt berechnet und der relative Fehler $\|b-\widetilde{b}\|/\|b\|$ ist klein. Kann man daraus schließen, dass auch der relative Fehler $\|x-\widetilde{x}\|/\|x\|$ klein ist?

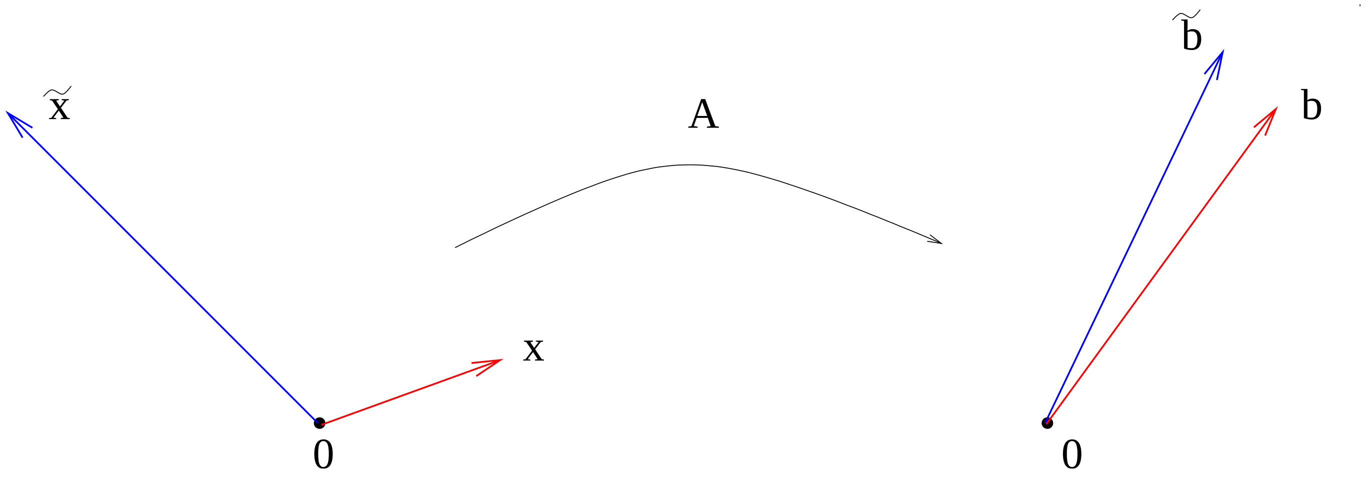
Antwort: Das hängt von der Konditionszahl ab. Die (nicht verbesserbare) Fehlerformel

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

besagt, dass sich der relative Fehler in b im ungünstigsten Fall um den Faktor $\mathrm{cond}(A)$ verstärkt.

Veranschaulichung der Situation bei einer schlecht konditionierten Matrix:

Relativ weit voneinander entfernte Vektoren $x, \widetilde{x}$ werden auf relativ nahe beieinander liegende Vektoren $b = Ax, \widetilde{b} = A\widetilde{x}$ abgebildet.



Situation bei ungenauer rechter Seite:

Man kennt nur $\widetilde{b}$, die exakte rechte Seite b ist unbekannt. Falls das Gleichungssystem exakt gelöst wird, kommt $\widetilde{x}$ heraus. Die Lösung x zur exakten rechten Seite b kann aber weit entfernt davon liegen.

Situation bei einer Proberechnung:

Rechte Seite ist b. Berechnet wurde die fehlerhafte Lösung $\widetilde{x}$. Die Probe ergibt $A\widetilde{x} = \widetilde{b}$. Auch wenn b und $\widetilde{b}$ nahezu übereinstimmen, können die wahre und die berechnete Lösung sich stark unterscheiden.

Die Konditionszahl einer Matrix ist hoch, wenn die Zeilen und Spalten fast linear abhängig sind

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

Die Zeilen und Spalten von A_0 sind linear abhängig. Für $\epsilon \neq 0$:

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

Die Konditionszahl von A_ϵ bzgl. Zeilensummennorm ist für $\epsilon \in [-6, 4]$:

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

Bemerkung: In diesem Beispiel ist die hohe Konditionszahl für kleine ϵ darauf zurückzuführen, dass $det(A_\epsilon)$ klein ist. Eine kleine Determinante impliziert aber nicht notwendig eine kleine Konditionszahl. Beispiel:

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

Verbesserung der Konditionszahl durch Präkonditionierung

Aufgabe: Löse

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

Durch Multiplikation der Gleichung mit einer nicht singulären Matrix $D \in \mathbb{R}^{n\times n}$ bekommt man die äquivalente Gleichung

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

Wenn A große Konditionszahl hat, dann sucht man eine Matrix D mit

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

und löst dann $(**)$ statt $(*)$.

Einfachste Möglichkeit:

Wähle D als Diagonalmatrix, und zwar so, dass alle Zeilen von DA die gleiche 1-Norm haben **(Zeilenäquilibrierung)**.