

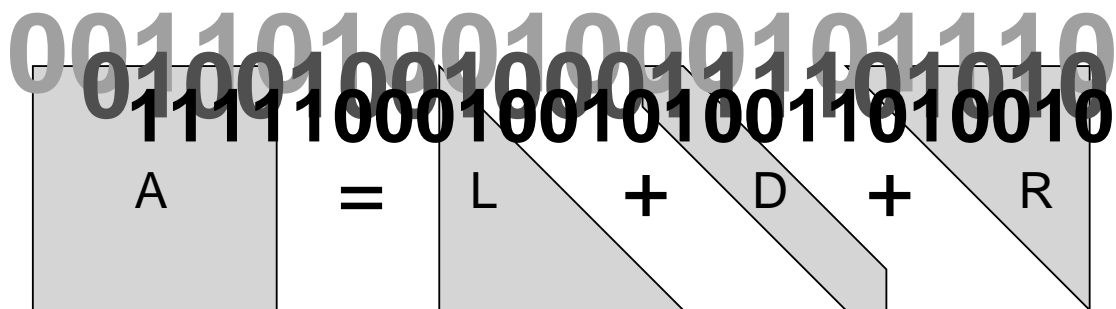
---

Stefan A. Funken, Karsten Urban

## Numerik 1

(Einführung in die Numerische Lineare Algebra)

---





**Vorwort.** Dieses Manuskript ist entstanden aus Mitschriften und Skripten verschiedener Vorlesungen, die wir seit 2002 an der Universität Ulm gehalten haben. Es ist der Sinn des vorliegenden Dokumentes, den Studierenden unserer Vorlesungen einen einheitlichen Stoffumfang für die Vorlesung **Numerik 1** zu geben, unabhängig davon, wer von uns tatsächlich die Vorlesung hält. In diesem Sinne bietet das vorliegende Manuskript einen Rahmen für die Nachbearbeitung der Vorlesungen und der Vorbereitung auf Prüfungen. Dieses Manuskript kann keinesfalls das Studium von Lehrbüchern ersetzen. Eine entsprechende Liste von Lehrbüchern findet sich im Literaturverzeichnis und auf der Internet-Seite der Vorlesung.

Jedes Manuskript weist Fehler auf, sicher auch dieses. Wenn Sie Fehler, Druckfehler, sprachliche Unzulänglichkeiten oder inhaltliche Flüchtigkeiten finden, würden wir uns über einen entsprechenden Hinweis per Email freuen. Sie helfen damit zukünftigen Studierenden. Vielen Dank im Voraus.

**Danksagung.** Einige Vorgängerversionen dieses Manuskriptes wurden aus Mitteln der Studiengebühren finanziert. Eine Reihe von Personen haben bei der Erstellung geholfen. Wir danken Theresa und Julia Springer, Markus Bantle und Judith Rommel für zahlreiche Hinweise. Frau Kristin Kirchner und Herrn Moritz Reinhard sind wir für das sorgfältige Lesen des Manuskripts zu besonderem Dank verpflichtet. Ihre zahlreichen Kommentare, Vorschläge, Korrekturen und Hinweise haben die Qualität des Textes wesentlich verbessert. Weiterhin möchten wir Katharina Becker-Steinberger, Sebastian Kestler, Dr. Michael Lehn und Moritz Reinhard für die Ausarbeitung der Aufgaben und zugehöriger Lösungen danken.

Ganz besonderer Dank gebührt auch Frau Petra Hildebrand, die unsere handschriftlichen Aufzeichnungen in  $\text{\LaTeX}$  umgesetzt und zahlreiche Grafiken erstellt hat. Frau Brandner, Frau Serbine und Herrn Weithmann möchten wir für das  $\text{\LaTeX}$ 'en der Lösungen danken.

**Copyright.** Alle Rechte, insbesondere das Recht auf Vervielfältigung und Verbreitung sowie der Übersetzung sind den Autoren vorbehalten. Kein Teil des Werkes darf in irgendeiner Form ohne schriftliche Genehmigung des Autors reproduziert oder unter Verwendung elektronischer Systeme oder auf anderen Wegen verarbeitet, vervielfältigt oder verbreitet werden.

**Stand.** Ulm, September 2016, Stefan A. Funken, Karsten Urban



# Inhaltsverzeichnis

<b>1</b>	<b>Einleitung: Was ist Numerik?</b>	<b>1</b>
<b>2</b>	<b>Zahlendarstellung, Fehlerarten und Kondition</b>	<b>3</b>
2.1	Zahlendarstellung	4
2.2	Rundung und Maschinengenauigkeit	14
2.3	Gleitkommaarithmetik	17
2.4	Fehlerverstärkung bei elementaren Rechenoperationen	20
2.5	Kondition eines Problems	21
2.6	Numerische Stabilität eines Algorithmus	24
<b>3</b>	<b>Direkte Lösung linearer Gleichungssysteme</b>	<b>29</b>
3.1	Einführung, Cramersche Regel	30
3.2	Gestaffelte Systeme	33
3.3	Gaußsche Eliminationsmethode	37
3.4	Pivot-Strategien	40
3.5	Nachiteration	47
3.6	Cholesky-Verfahren	47
3.7	Bandgleichungen	51
3.8	Vandermond-Matrizen	53
3.9	Fehlerabschätzung mittels Kondition	56
<b>4</b>	<b>Iterative Lösung linearer Gleichungssysteme</b>	<b>61</b>
4.1	Klassische Iterationsverfahren	63
4.2	Konvergenz iterativer Verfahren	66
4.3	Konvergenz des Jacobi-Verfahrens	69
4.4	Konvergenz des Gauß-Seidel-Verfahrens	72
4.5	Abbruchkriterien	78
4.6	Gradienten-Verfahren	79
4.7	Verfahren der konjugierten Gradienten	85
4.8	Vorkonditionierung, das pcg-Verfahren	92
<b>5</b>	<b>Lineare Ausgleichsprobleme</b>	<b>97</b>
5.1	Die Methode der kleinsten Quadrate	98
5.2	Die <b>QR</b> -Zerlegung	100
5.3	<b>Givens</b> -Rotationen	102
5.4	Update einer <b>QR</b> -Zerlegung	105
5.5	<b>Householder</b> -Spiegelungen	107
5.6	Die Singulärwertzerlegung	112
<b>A</b>	<b>Landau-Symbole</b>	<b>117</b>

	A.1	Definition	117
	A.2	Beispiele	117
<b>B</b>		<b>Normen</b>	<b>119</b>
	B.1	Vektornorm-Eigenschaften	119
	B.2	Matrixnormen	120
<b>C</b>		<b>Einführung in MATLAB</b>	<b>123</b>
	C.1	Grundlegende MATLAB-Befehle	123
	C.2	Mathematik mit Matrizen	127
	C.3	Datenverwaltung	132
	C.4	Ausgabe von Text	133
	C.5	Kontrollbefehle	134
	C.6	Graphische Darstellung	136
	C.7	Fortgeschrittenes	139
<b>D</b>		<b>Speicherformate für schwach besetzte Matrizen</b>	<b>143</b>
	D.1	Koordinatenformat (COO-Format)	143
	D.2	Komprimierte Zeilenspeicherung (CRS-Format)	148
	D.3	Modifizierte komprimierte Zeilenspeicherung (MRS-Format)	150
	D.4	Harwell-Boeing-Format (CCS-Format)	154
		<b>Literaturverzeichnis</b>	<b>154</b>
		<b>Stichwortverzeichnis</b>	<b>157</b>

# 1 EINLEITUNG: WAS IST NUMERIK?

Wenn man z.B. in der Öffentlichkeit über Mathematik diskutiert, stellt man oft fest, dass viele eine vollkommen falsche Interpretation des Begriffes „Numerik“ (oder besser „Numerische Mathematik“) haben. Das Wort scheint eine fachliche Nähe zur Zahlentheorie anzudeuten und man mag vermuten, dass die Numerik ein Teilgebiet der Reinen Mathematik ist. Dies ist jedoch nicht richtig.

Numerik ist ein Teilgebiet der Angewandten Mathematik und bezeichnet allgemein die mathematische Untersuchung von Berechnungsverfahren, Algorithmen oder Methoden zur näherungsweisen Berechnung bestimmter Größen auf dem Computer.

Zu berechnende Größen können z.B. sein:

- Die Auswertung von Funktionen wie z.B.  $\sin(1)$ ,  $e^2$  und ähnliche, die nicht unmittelbar angegeben werden können. Solche Berechnungen geschehen auch in jedem Taschenrechner. Dort sind die entsprechenden Berechnungsverfahren in der Regel durch Schaltkreise oder spezielle Microchips realisiert, also in der Hardware fest verdrahtet.
- Die Lösung von Gleichungen z.B. lineare  $Ax = b$  oder nichtlineare Gleichungen. d.h.  $f(x) = 0$ , wobei die gesuchte Lösung  $x^* \in \mathbb{R}^n$  oft aus sehr vielen Komponenten besteht, also  $n > 1.000.000$  gilt! Andere Typen von Gleichungen sind z.B. Differentialgleichungen etwa  $u''(x) = f(x)$ ,  $x \in (0, 1)$ , wobei die Funktion  $u : [0, 1] \rightarrow \mathbb{R}$  bei gegebenem  $f : [0, 1] \rightarrow \mathbb{R}$  gesucht ist. Ein Beispiel hierfür ist etwa die gesuchte Auslenkung  $u$  eines Werkstücks unter der äußeren Last  $f$ .
- Die Näherung von Größen, die nicht exakt berechnet werden können (z.B. Ableitungen – Sensitivitäten, Integrale – Mittel- oder Erwartungswerte). Manchmal will man solche Größen auch gar nicht exakt bestimmen, etwa, wenn die Berechnung zu lange dauern würde. Andere Beispiele wären optimale Steuerungen oder optimale Strategien.
- Computer-gestützte Simulationen komplexer Vorgänge.  
In vielen Bereichen sind Experimente sehr teuer, aufwändig, zu gefährlich, **ethisch nicht vertretbar** oder gar nicht möglich. Beispiele sind:
  - Wettervorhersage: Dies bedeutet u.a. die Simulation der turbulenten Wolkenströmungen.
  - Strömungsmechanik, also etwa Aerodynamik, Flugzeug-, Automobil- oder Schiffsbau.
  - Bauingenieurwesen, z.B. die Simulation der Statik oder der Eigenschwingung von Brücken und anderen Bauwerken.
  - Medizin: Simulation der Knochenheilung oder die Therapie von Gehirn-Tumoren.
  - Wirtschaftswissenschaften: Simulation von Aktienkursen, Bewertung von komplexen Finanz- oder Versicherungsprodukten, Bestimmung optimaler Anlage-Strategien.

Speziell beschäftigt sich die Numerik mit

- der Konstruktion „geeigneter“ Lösungsverfahren, die insbesondere
  - schnell („effizient“) sind, teilweise in oder sogar schneller als in „Echtzeit“,
  - zuverlässig, also mit **beweisbarer** Abschätzung z.B. der folgenden Form  $\|x_{\text{numerisch}} - x_{\text{exakt}}\| \leq \text{Toleranz}$  (wobei die exakte Größe  $x_{\text{exakt}}$  unbekannt ist), und
  - robust gegenüber Störungen wie z.B. Messfehlern, Modell-Unsicherheiten etc.

sind;

- mit der mathematischen Analyse dieser Verfahren (Konvergenz, Geschwindigkeit, Aufwand, Robustheit etc.) und
- deren effizienter Realisierung (Implementierung).

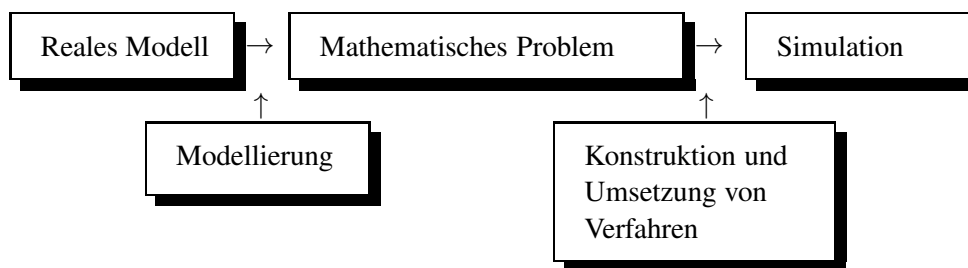
In diesem Sinne liegt die Numerik an der Schnittstelle von Mathematik und Informatik.

Die Numerische Mathematik selbst hat eine Reihe von Teilgebieten bzw. Gebiete, die unmittelbar an sie angrenzen, etwa:

- Numerische Lineare Algebra (Numerik 1),
- Numerische Analysis (Numerik 2),
- Numerische Optimierung (Numerik 3),
- Numerik von Differenzialgleichungen (Numerik 4 und Numerik von Partiellen Differenzialgleichungen)
- Numerical Finance
- Computational Physics, Computational Science
- CFD: Computational Fluid Dynamics
- Wissenschaftliches Rechnen
- ...

Insbesondere wird klar, dass die Numerik einen starken interdisziplinären Bezug aufweist.

Das Vorgehen bei der Erstellung einer numerischen Simulation kann man (vereinfacht) so darstellen:



Wir beschreiben dieses abstrakte Vorgehen an einem konkreten Beispiel.

**Beispiel 1.0.1 (Bestimmung des Abraums bei der Braunkohleförderung)** Nehmen wir an, dass wir den Abraum bestimmen wollen, der durch die Förderung von Braunkohle entstanden ist. Würde man die Form des entstandenen Lochs exakt kennen, so würde sich der Abraum als der Wert des Volumenintegrals dieses Loches ergeben. Da man aber die Form nicht exakt kennt, fliegt z.B. ein Flugzeug oder eine Drohne über die Braunkohle-Halde und macht Tiefenmessungen an einzelnen Punkten. Daraus ergibt sich dann folgendes Vorgehen:

- 1) **Reales Modell:** Volumen des Abraums; mit konkreten **Messungen (Experiment)** in Form von Tiefenmessungen durch Stereofotographien aus Flugzeugen, Satelliten;
- 2) **Mathematisches Problem:** Bestimme das Volumen des Lochs. Dies bedeutet die Berechnung von Volumenintegralen; verwende dabei die Mess-Ergebnisse als Daten (**Modellierung**);
- 3) **Konstruktion und Umsetzung von Verfahren:** Verwende sogenannte Kubaturformeln (Numerik 2) zur näherungsweise Berechnung der dreidimensionalen Integrale;  
**Simulation:** Programmiere das o.g. Verfahren und erzeuge so entsprechende Simulationen.



# 2 ZAHLENDARSTELLUNG, FEHLERARTEN UND KONDITION

Auf einer endlichen Maschine kann man niemals exakt rechnen, als Beispiel nehme man die Darstellung reeller Zahlen mit unendlicher Dezimaldarstellung. Deren Darstellung ist auf einer endlichen Maschine unmöglich. Wir werden noch sehen, dass viele Rechenoperationen ebenfalls nicht exakt ausgeführt werden können. Man hat also **immer** Fehler. Diese gilt es mathematisch zu analysieren und zu kontrollieren.

Nehmen wir an, wir wollten die Lösung  $x \in \mathbb{R}^n$  eines linearen Gleichungssystems  $Ax = b$  mit gegebener regulärer Matrix  $A \in \mathbb{R}^{n \times n}$  und gegebener rechten Seite  $b \in \mathbb{R}^n$  mittels eines Computerprogramms näherungsweise berechnen. Wir wollen also aus den Eingabegrößen  $(A, b)$  das Resultat  $f(A, b) := A^{-1}b$  berechnen. Etwas abstrakter formuliert, wir haben eine Abbildung  $f : U \subset X \rightarrow Y$  an einer Stelle  $x \in U$  ausgewertet, wobei  $X$  und  $Y$  geeignete Räume sind, die natürlich von der Problemstellung abhängen. Für das Beispiel der Lösung eines linearen Gleichungssystem gilt

$$X = \mathbb{R}^{n \times n} \times \mathbb{R}^n, \quad Y = \mathbb{R}^n, \quad U = \{(A, b) \in X : A \text{ ist regulär}\}.$$



Bei exakter Arithmetik würden wir mit dem Gauß-Verfahren (aus der Linearen Algebra, siehe auch Kapitel 3) ein exaktes Ergebnis erhalten. Im Allgemeinen treten jedoch sowohl Fehler in der Eingabe als auch im Algorithmus auf, z.B. lässt sich die Zeit nur auf eine Genauigkeit von  $10^{-14}s$  und die Masse auf  $10^{-8}g$  genau bestimmen<sup>1</sup>, Zahlen nur näherungsweise im Rechner als Gleitkommazahl (auch Fließkommazahl genannt, engl. floating point number, siehe unten) darstellen und z.B. auch die Darstellung

$$\sin(x) = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!} \quad (2.1)$$

als unendliche Reihe nur näherungsweise auswerten. Es stellt sich sofort folgende Frage: Welchen Einfluss haben nun beide Fehlerarten, d. h.

- Eingabefehler (lassen sich nicht vermeiden) und
- Fehler im Algorithmus (wobei keine algorithmischen Fehler gemeint sind)

auf den Fehler im Resultat? Die Unterscheidung beider Arten von Fehlern führt uns zu den zentralen Begriffen der Numerik

**Kondition** eines Problems und

**Stabilität** eines Algorithmus, vgl. 2.5 und 2.6.

---

<sup>1</sup>Siehe [www.ptb.de](http://www.ptb.de): Physikalisch-technische Bundesanstalt.

## 2.1 Zahlendarstellung

Ein erstes Beispiel soll motivieren, wieso wir uns mit der Frage, wie Zahlen im Rechner dargestellt werden, beschäftigen sollten.

**Beispiel 2.1.1** Eine einfache Überlegung zeigt, dass das Ergebnis von

$$4 \cdot 13860^4 - 19601^4 + 2 \cdot 19601^2$$

ungerade ist, eine erste Rechnung, dass an der letzten Stelle eine 1 oder 9 stehen muss. (Dabei könnte sich die 9 z.B. durch  $-1 \cdot 10 + 1$  ergeben.)

$$\begin{aligned} 4 \cdot 13860^4 - 19601^4 + 2 \cdot 19601^2 &= 4(1386 \cdot 10)^4 - (1960 \cdot 10 + 1)^4 + 2(1960 \cdot 10 + 1)^2 \\ &= 4 \cdot 1386^4 10^4 - (1960^4 10^4 + 4 \cdot 1960^3 10^3 + 6 \cdot 1960^2 10^2 + 4 \cdot 1960 10 + 1) + \\ &\quad + 2(1960^2 10^2 + 2 \cdot 1960 \cdot 10 + 1) \\ &= 10^4(\dots) + 10^3(\dots) + 10^2(\dots) + 10^1(\dots) + 10^0(-1 + 2). \end{aligned}$$

Rechnet man auch die anderen Stellen aus (per Hand oder mit Maple) so ergibt sich

$$4 \cdot 13860^4 - 19601^4 + 2 \cdot 19601^2 = 1.$$

### MATLAB-Beispiel:

Überraschenderweise liefert Matlab aber ein anderes Resultat.

```
>> myfun = @(x,y) 4*x^4-y^4+2*y^2;
>> myfun(13860,19601)
ans =
     2
```

Auch die folgende Realisierung mittels quadratischer Ergänzung ( $u = 4x^2 + 1$ ,  $v = y^2 - 1$ ,  $u - v^2 = 4x^4 - y^4 + 2y^2$ ) liefert ein (anderes) falsches Ergebnis.

```
>> x = 13860; y = 19601;
>> u = 4 * x * x * x * x + 1;
>> v = y * y - 1;
>> u - v * v
ans =
     0
```



Und Vorsicht beim Umgang mit Maple! Schreibt man dort versehentlich nach einer Zahl noch einen Punkt, so ergibt es das scheinbar überraschende folgende Ergebnis. (Man beachte dabei, dass Maple, wenn man den Punkt hinzufügt, nicht mehr mit exakter Arithmetik rechnet, sondern so genannte Gleitkommazahlen (Maschinenzahlen, wird später noch im Detail erklärt) verwendet, deren Genauigkeit man einstellen kann. Die Anweisung `Digits:=18` führt dazu, dass Maple Gleitkommazahlen mit 18 Stellen verwendet.

```
> restart:
4  *13860^4-19601^4+2*19601^2;

4. *13860^4-19601^4+2*19601^2;
Digits:=15:
4. *13860^4-19601^4+2*19601^2;
Digits:=16:
4. *13860^4-19601^4+2*19601^2;
Digits:=17:
4. *13860^4-19601^4+2*19601^2;
Digits:=18:
4. *13860^4-19601^4+2*19601^2;
```

```
1
68398402.
402.
2.
2.
1.
```

### MATLAB-Beispiel:

Aus dem Topf der mathematischen  
Mirakel ziehen wir noch folgendes  
Beispiel:

```
>> floor(114)
ans =
    114
>> floor(1.14*100)
ans =
    113
```

Die Routine `floor` rundet dabei auf die nächstkleinere ganze Zahl. Die Zahl 114 auf die nächstkleinere ganze Zahl gerundet sollte also in beiden Fällen 114 ergeben!

Ein Rechner verwendet als einfachste Form der Zahlendarstellung **Festkommazahlen**. Dabei wird eine (meistens) begrenzte Ziffernfolge gespeichert und an festgelegter Stelle das Komma angenommen. Bei vielen Anwendungen reicht der darstellbare Zahlenbereich jedoch nicht aus, so dass es naheliegend ist, bei jedem Wert die genaue Stelle des Kommas zusätzlich zu speichern. Das bedeutet mathematisch nichts anderes als die Darstellung der Zahl  $x$  mit zwei Werten, der **Mantisse**  $m$  und dem **Exponenten**  $e$ . Wir bekommen somit die Darstellung  $x = m \cdot 10^e$ . Diese Darstellung ist jedoch nicht eindeutig, siehe z.B.  $1014 = 1.014 \cdot 10^3 = 10140 \cdot 10^{-1}$ . Legt man jedoch den Wertebereich der Mantisse  $m$  geschickt fest, z.B.  $1 \leq m < 10$ , so erhält man eine eindeutige Darstellung (wenn man endlich viele Stellen hat). Diesen Schritt bezeichnet man als **Normalisierung** der Mantisse. Wir haben soeben stillschweigend vorausgesetzt, dass wir die Basis 10 verwenden<sup>2</sup>. Darauf ist man jedoch nicht festgelegt. Pascal<sup>3</sup> erkannte, dass man jede Basis  $b \geq 2$  verwenden kann. Heutige Rechner verwenden meist binäre Zahlendarstellungen, d.h. zur Basis 2. Betrachten wir nun die Zahlendarstellung noch etwas allgemeiner.

<sup>2</sup>**Dezimalsystem:** Unter dem Einfluss der Schrift „De Thiende“ (Leiden 1585) von Simon Stevin (1548-1620) setzte sich im 16. Jahrhundert das Rechnen mit Dezimalbrüchen in Westeuropa durch.

<sup>3</sup>**Blaise Pascal**, 1623-1662.

**Definition 2.1.2 (*b*-adischer Bruch)** Es sei  $b \geq 2$  eine natürliche Zahl,  $n \in \mathbb{N}_0$ ,  $a_k \in \mathbb{N}_0$  mit  $0 \leq a_k < b$  und  $a_n \neq 0$ . Eine Reihe der Gestalt

$$\pm \sum_{k=-\infty}^n a_k b^k$$

wird als ***b*-adischer Bruch** bezeichnet, *b* heißt **Basis**. Falls die Basis festgelegt ist, kann man einen *b*-adischen Bruch auch durch die Reihung der Ziffern  $a_k$  angeben:

$$\pm a_n a_{n-1} \cdots a_1 a_0 . a_{-1} a_{-2} a_{-3} \cdots$$

**Bemerkungen 2.1.3** i) Für  $b = 10$  spricht man von **Dezimalbrüchen**, für  $b = 2$  von **dyadischen Brüchen** und die Babylonier<sup>4</sup> verwendeten ein System zur Basis  $b = 60$ , was sich heute noch im Verhältnis von Sekunde, Minute und Stunde wieder findet.

ii) Die Basis wird auch als **Grundzahl** bezeichnet, weswegen man auch die Bezeichnung **g-adisch** findet.



iii) Vorsicht, nicht **b-adisch** mit **p-adisch** verwechseln ( $p \sim$  Primzahl)!

iv) Der Null kommt immer eine Sonderrolle zu. Sie wird gesondert betrachtet, um die Darstellung einfacher zu halten. Ohne den Sonderfall der Null jeweils aufzuführen, schließen wir ihn indirekt mit ein.

Die nächsten beiden Sätze besagen, dass sich jede reelle Zahl durch einen *b*-adischen Bruch darstellen lässt und umgekehrt.

**Satz 2.1.4 (Konvergenz *b*-adischer Bruch)** Jeder *b*-adische Bruch stellt eine Cauchy-Folge dar, d.h. dass jeder *b*-adische Bruch gegen eine reelle Zahl konvergiert.

*Beweis.* Wir beschränken uns auf den Fall eines nicht-negativen *b*-adischen Bruchs  $\sum_{k=-\infty}^n a_k b^k$ .

Wir definieren für  $-\ell \leq n$  die Partialsummen

$$s_\ell := \sum_{k=-\ell}^n a_k b^k.$$

Es ist nun zu zeigen, dass die Folge  $(s_\ell)_{-\ell \leq n}$  eine Cauchy-Folge ist. Sei  $\varepsilon > 0$  und  $L \in \mathbb{N}$  so groß, dass  $b^{-L} < \varepsilon$  gelte. Dann gilt für  $\ell \geq m \geq L$

$$\begin{aligned} |s_\ell - s_m| &= \sum_{k=-\ell}^n a_k b^k - \sum_{k=-m}^n a_k b^k = \sum_{k=-\ell}^{-m-1} a_k b^k \leq \sum_{k=-\ell}^{-m-1} (b-1)b^k \\ &= (b-1)b^{-m-1} \sum_{k=-\ell+m+1}^0 b^k = (b-1)b^{-m-1} \sum_{k=0}^{\ell-m-1} b^{-k} \\ &< (b-1)b^{-m-1} \frac{1}{1-b^{-1}} = b^{-m} \leq b^{-L} < \varepsilon \end{aligned}$$

und der Satz ist bewiesen. □

<sup>4</sup>**Sexagesimalsystem:** Zählt man mit den Fingern der linken Hand wie üblich von 1 bis 5 und registriert die an der linken Hand gezählten 5-er Gruppen durch das Drücken des Daumens der rechten Hand auf das passende Fingerglied an der rechten Hand, so kommt man zum  $5 \times 12 = 60$ -System. Diese Zählweise wurde schon von den Vorfahren der Sumerer im 4. Jahrtausend v. Chr. angewandt. Die Babylonier (1750 v. Chr.) verwendeten dann das Sexagesimalsystem in dem Sinne, dass sie 60 Herzschläge zu einer Minute zusammen fassten und 60 Minuten zu einer Stunde.

**Satz 2.1.5 (Reelle Zahl als  $b$ -adischer Bruch)** Sei  $b \geq 2$  eine natürliche Zahl. Dann lässt sich jede reelle von Null verschiedene Zahl in einen  $b$ -adischen Bruch entwickeln.

*Beweis.* Es genügt wieder, nur den Fall  $x \geq 0$  zu zeigen. Zuerst zeigen wir, dass es zu jeder Basis  $b \geq 2$  mindestens eine Zahl  $m \in \mathbb{N}$  gibt mit  $x < b^m$ . Nach dem Archimedischen Axiom gibt es ein  $m \in \mathbb{N}$  mit  $m > (x-1)/(b-1) \in \mathbb{R}$ . Mit diesem  $m$  und der Bernoullischen Ungleichung erhalten wir

$$b^m = (1 + (b-1))^m \geq 1 + m(b-1) > 1 + x - 1 = x.$$

Sei  $n := \min\{k \in \mathbb{N}_0 \mid 0 \leq x < b^{k+1}\}$ . Durch vollständige Induktion konstruieren wir nun eine Folge  $(a_{-\ell})_{\ell \geq -n}$  natürlicher Zahlen  $0 \leq a_{-\ell} < b$ , so dass für alle  $m \geq -n$  gilt

$$x = \sum_{\ell=-m}^n a_{\ell} b^{\ell} + \nu_{-m} \quad \text{mit } 0 \leq \nu_{-m} < b^{-m}.$$

Da  $\lim_{m \rightarrow \infty} \nu_{-m} = 0$  gilt, folgt  $x = \sum_{\ell=-\infty}^n a_{\ell} b^{\ell}$  und somit die Behauptung.

**Induktionsanfang  $-m = n$ :** Es gilt  $0 \leq x b^{-n} < b$ . Somit existiert ein  $a_n \in \{0, 1, \dots, b-1\}$  und ein  $\delta \in \mathbb{R}$  mit  $0 \leq \delta < 1$ , so dass  $x b^{-n} = a_n + \delta$  gilt. Mit  $\nu_n := \delta b^n$  gewinnt man

$$x = a_n b^n + \nu_n \quad \text{mit } 0 \leq \nu_n < b^n.$$

**Induktionsschritt  $(-m) \rightarrow (-m-1)$ :** Es gilt  $0 \leq \nu_{-m} b^{m+1} < b$ , also gibt es ein  $a_{-m-1} \in \{0, 1, \dots, b-1\}$  und ein  $\delta \in \mathbb{R}$  mit  $0 \leq \delta < 1$ , so dass  $\nu_{-m} b^{m+1} = a_{-m-1} + \delta$  gilt. Mit  $\nu_{-m-1} := \delta b^{-m-1}$  gewinnt man

$$x = \sum_{\ell=-m}^n a_{\ell} b^{\ell} + \nu_{-m} = \sum_{\ell=-m}^n a_{\ell} b^{\ell} + (a_{-m-1} + \delta) b^{-m-1} = \sum_{\ell=-m-1}^n a_{\ell} b^{\ell} + \nu_{-m-1},$$

wobei  $0 \leq \nu_{-m-1} < b^{-m-1}$  gilt. □

**Bemerkung 2.1.6** Die  $b$ -adische Darstellung ist in der gewählten Form nicht eindeutig, siehe z. B.  $1 = 0.\overline{9} = 0.999\dots$ , falls  $b = 10$  gewählt ist. Durch die zusätzliche Bedingung „ $a_k < b-1$  für unendlich viele  $k \leq n$ “ erhält man eine eindeutige Darstellung. Somit wäre z.B. für  $b = 10$  die Darstellung  $0.\overline{9} = 0.999\dots$  für  $x = 1$  nicht zugelassen.

Auf einem Rechner können wir nur eine endliche Entwicklung speichern, also werden wir den  $b$ -adischen Bruch durch eine endliche Entwicklung approximieren.

**Beispiel 2.1.7** i) Binärdarstellung ( $b = 2$ ) von  $x = (14.625)_{10} = (1110.101)_2$

$$\begin{array}{rclcl} 14 & = & 7 \cdot 2 & + 0 & \Rightarrow a_0 = 0 \\ 7 & = & 3 \cdot 2 & + 1 & \Rightarrow a_1 = 1 \\ & & 3 & = & 1 \cdot 2 + 1 \quad \Rightarrow a_2 = 1 \\ & & 1 & = & 0 \cdot 2 + 1 \quad \Rightarrow a_3 = 1 \end{array}$$

Im Folgenden bezeichnet  $\lfloor \cdot \rfloor$  die untere Gaußklammer, d.h.

$$\lfloor x \rfloor := \max \{j \in \mathbb{Z} : j \leq x\}.$$

$$a_{-1} = \lfloor 2 * 0.625 \rfloor = \lfloor 1.25 \rfloor = 1$$

$$a_{-2} = \lfloor 2 * 0.25 \rfloor = \lfloor 0.5 \rfloor = 0$$

$$a_{-3} = \lfloor 2 * 0.5 \rfloor = \lfloor 1 \rfloor = 1$$

ii) Binärdarstellung von  $x = (0.2)_{10} = (0.\overline{0011})_2$

$$\begin{aligned} a_{-1} &= \lfloor 2 * 0.2 \rfloor = \lfloor 0.4 \rfloor = 0 \\ a_{-2} &= \lfloor 2 * 0.4 \rfloor = \lfloor 0.8 \rfloor = 0 \\ a_{-3} &= \lfloor 2 * 0.8 \rfloor = \lfloor 1.6 \rfloor = 1 \\ a_{-4} &= \lfloor 2 * 0.6 \rfloor = \lfloor 1.2 \rfloor = 1 \\ a_{-5} &= \lfloor 2 * 0.2 \rfloor = \dots \\ &\vdots \end{aligned}$$



**Bemerkung 2.1.8** Da es unmöglich ist, reelle Zahlen als unendliche  $b$ -adische Brüche zu speichern, werden reelle Zahlen auf Rechnern durch eine endliche Entwicklung approximiert.

**Definition 2.1.9 (Festpunktzahl)** Zu gegebener Basis  $b$  ist

$$F(\ell, n) := \left\{ \pm \sum_{k=-\ell}^n a_k b^k = \pm (a_n \dots a_1 a_0 . a_{-1} \dots a_{-\ell})_b \right\}$$

die Menge der **Festpunktzahlen** mit  $(n+1)$  Vor- und  $\ell$  Nachkommastellen.

**Bemerkungen 2.1.10** i) Typischerweise werden ganze Zahlen (engl. integers) auf Computern dargestellt mit  $\ell = 0$  und  $b = 2$ .

ii) In modernen Rechnern werden negative Zahlen häufig als 2-er-Komplement<sup>5</sup> abgespeichert. Dabei ist  $x$  das Komplement zu  $y$ , wenn in der jeweiligen Arithmetik  $x+y = 0$  gilt. Beispiel: Für  $b = 2$  und  $n = 3$  ergibt sich

$ x $	$x$			$-x$		
0	0	0	0			
1	0	0	1	1	1	1
2	0	1	0	1	1	0
3	0	1	1	1	0	1
4	1	0	0	1	0	0

also darstellbarer Bereich:  $-4, \dots, 3$

**Bemerkung 2.1.11** Da im 2-er-Komplement der Wert 0 den positiven Zahlen zugeordnet wird,

<sup>5</sup>**Zweierkomplement** ist eine Möglichkeit, negative Zahlen im Dualsystem darzustellen. Dabei werden keine zusätzlichen Symbole wie + und - benötigt. Da im Zweierkomplement der Wert 0 den positiven Zahlen zugerechnet wird, umfasst der Wertebereich bei  $n$  binären Stellen allgemein den Bereich  $-2^{n-1}, \dots, 0, \dots, 2^{n-1} - 1$ . Zum Beispiel bei 4 Bit von -8 bis 7. Negative Zahlen gewinnt man aus positiven Zahlen, indem sämtliche binären Stellen negiert werden und zu dem Ergebnis der Wert 1 addiert wird. Somit wird aus  $(4)_{10} = (0100)_2$  durch Invertieren  $\text{Not}(0100) = 1011$  und Addition der 1  $(1011)_2 + (0001)_2 = (1100)_2$ . Addiert man nun 4 und -4, so gilt  $(4)_{10} + (-4)_{10} = (0100)_2 + (1100)_2 = (0000)_2 = (0)_{10}$ , wobei die vorderste Stelle verworfen wurde.

erhält man für 8-Bit<sup>6</sup>, 16-Bit und 32-Bit, bzw. 1-, 2- und 4-Byte<sup>7</sup> Darstellungen die Zahlenbereiche

8-Bit,  $n = 7$  :  $0, \dots, 255$  ohne Vorzeichen  
bzw.  $-128, \dots, 127$  (8 Stellen)

16-Bit,  $n = 15$  :  $0, \dots, 65535$  ohne Vorzeichen  
bzw.  $-32768, \dots, 32767$  (16 Stellen)

32-Bit,  $n = 31$  :  $0, \dots, 4.294.967.295$  ohne Vorzeichen  
bzw.  $-2.147.483.648, \dots, 2.147.483.647$  (32 Stellen)

**Bemerkung 2.1.12** Als Problem stellt sich dabei heraus, dass damit nur ein kleiner Zahlenbereich darstellbar ist.

Die Verwendung eines Festpunktes schränkt also die auf dem Rechner darstellbaren Werte stark ein. Gibt man stattdessen nur die Gesamtanzahl an Vor- und Nachkommastellen vor und lässt die Position des Punktes variabel, skaliert also mit unterschiedlichen Potenzen der Basis  $b$ , so ist die Menge der darstellbaren Zahlen wesentlich größer. Dies führt zur Definition der Gleitpunkt-Darstellung einer reellen Zahl  $x$ .

**Definition 2.1.13 (Gleitpunkt-Darstellung)** Die **normalisierte Gleitpunkt-Darstellung** einer reellen Zahl  $x \in \mathbb{R}$  hat die Form  $x = f \cdot b^\ell$ , wobei

- die **Mantisse**  $f$  ist eine feste Zahl, die  $m$  Stellen hat und

$$b^{-1} \leq |f| < 1, \text{ falls } x \neq 0 \text{ und } f = 0 \text{ falls, } x = 0$$

genügt, d.h.

$$f = \begin{cases} \pm (0.d_1 \dots d_m)_b = \pm \sum_{j=1}^m d_j b^{-j} & , d_1 \neq 0 \\ 0 & , d_1 = 0 \end{cases}$$

- der **Exponent**  $\ell$  ist eine ganze Zahl innerhalb fester Schranken

$$r \leq \ell \leq R,$$

welche sich in der Form

$$\ell = \pm (\ell_{n-1} \dots \ell_0)_b = \pm \sum_{j=0}^{n-1} \ell_j b^j$$

darstellen lässt. Hierbei braucht nicht  $\ell_{n-1} \neq 0$  zu gelten.

(Für 8-stelligen binären Exponenten wird z.B.  $r = -128$  und  $R = 127$  gewählt.)

Die Menge der Maschinenzahlen mit Basis  $b$ ,  $m$ -stelliger Mantisse und  $n$ -stelligem Exponenten wird als  $\mathbb{M}(b, m, n)$  bezeichnet.

<sup>6</sup>Bit  $\sim$  (engl. binary digit, lat. digitus  $\sim$  Finger).

<sup>7</sup>Byte  $\sim$  Maßeinheit für eine Datenmenge von 8 Bit.

**Beispiel 2.1.14**  $\mathbb{M}(2, 3, 1)$  besteht aus den Zahlen

$$x = \pm(d_{-1} \cdot 2^{-1} + d_{-2} \cdot 2^{-2} + d_{-3} \cdot 2^{-3}) \cdot 2^{\pm \ell_0},$$

wobei  $d_{-1} \neq 0 \vee x = d_{-1} = d_{-2} = d_{-3} = 0$  gilt. Somit

$$\mathbb{M}(2, 3, 1) = \left\{ 0, \pm \frac{1}{4}, \pm \frac{5}{16}, \pm \frac{3}{8}, \pm \frac{7}{16}, \pm \frac{1}{2}, \pm \frac{5}{8}, \pm \frac{3}{4}, \pm \frac{7}{8}, \pm 1, \pm \frac{5}{4}, \pm \frac{3}{2}, \pm \frac{7}{4} \right\}.$$

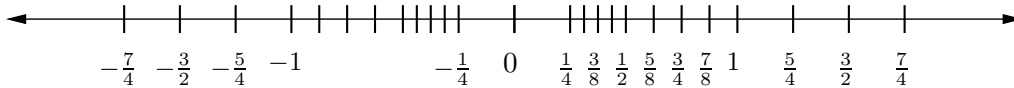


Abb. 2.1: Graphische Darstellung von  $\mathbb{M}(2, 3, 1)$ .

Man beachte den kleiner werdenden Abstand zur Null hin und den Abstand zwischen den betragsmäßig kleinsten Zahlen und der Null.

**Bemerkung 2.1.15** Wie man in Beispiel 2.1.14 sieht, sind Maschinenzahlen nicht gleichverteilt, d.h. der Fehler bei der Darstellung einer Zahl  $x \in \mathbb{R}$  auf einem Computer hängen von der Eingabe  $x$  ab.

**Bemerkung 2.1.16** Die betragsmäßig kleinste und größte Zahl in  $\mathbb{M}(b, m, n)$  bezeichnen wir mit  $x_{\min}$  bzw.  $x_{\max}$ ,

$$\begin{aligned} x_{\min} &:= \min \{ |z| \in \mathbb{M}(b, m, n) \}, & |x| < x_{\min} &\Rightarrow x = 0 \\ x_{\max} &:= \max \{ |z| \in \mathbb{M}(b, m, n) \}, & |x| > x_{\max} &\Rightarrow \text{overflow}. \end{aligned}$$

**Beispiel 2.1.17**

$$\begin{aligned} \mathbb{M}(10, 4, 1), & \quad x = -51.34 & \rightsquigarrow -0.5134 \cdot 10^2 \\ \mathbb{M}(2, 4, 2), & \quad x = 3.25 = (11.01)_2 & \rightsquigarrow (0.1101)_2 \cdot 2^{(10)_2} \\ \mathbb{M}(2, 4, 2), & \quad x = 14.625 = (1110.101)_2 & \rightsquigarrow (0.1110101)_2 \cdot 2^{(100)_2} \notin \mathbb{M}(2, 4, 2) \end{aligned}$$

**Bemerkung 2.1.18** Üblicherweise werden 8 Byte (64 Bit) Speicherplatz für eine Gleitpunktzahl wie folgt aufgeteilt:

- 1 Bit für das Vorzeichen;
- 11 Bit für den Exponenten, d.h. die Werte sind darstellbar im Bereich  $-1022 \dots 1023$  ( $1 \dots 2046 - \text{Bias}^8\text{-Wert } 1023$ ). Die Werte  $-1023$  und  $1024$ , bzw.  $0$  und  $2047 = 2^{11} - 1$  als Charakteristik sind reserviert für die speziellen Zahlenwerte „Null“, „Unendlich“ und „NaN“;
- 52 Bit für die Mantisse. Dies sind eigentlich Werte zwischen  $(0 \dots 01)_2$  und  $(1 \dots 1)_2$ , da die Null gesondert abgespeichert wird. Ist man in der normalisierten Darstellung, so ist die erste Null redundant und man gewinnt noch eine Stelle hinzu. Also muss auch noch irgendwie festgehalten werden, ob eine normalisierte Darstellung vorliegt oder nicht.

Also ergibt sich  $b = 2$ ,  $m = 52$ ,  $n = 11$  und damit  $r = -1023$  und  $R = 1024$ .

<sup>8</sup>Hier zu verstehen als ein konstanter Wert, der zu etwas hinzuaddiert bzw. subtrahiert wird. Wird auch als Offset bezeichnet.



Tab. 2.1: Interpretation des Zahlenformats nach IEEE-754

Exponent $e$	Mantisse	Bedeutung	Bezeichnung
$e = 0$	$m = 0$	$\pm 0$	Null
$e = 0$	$m > 0$	$\pm 0, m \times 2^{1-Bias}$	denormalisierte Zahl
$0 < e < e_{max}$	$m \geq 0$	$\pm 1, m \times 2^{e-Bias}$	normalisierte Zahl
$e = e_{max}$	$m = 0$	$\pm \infty$	Unendlich (Inf)
$e = e_{max}$	$m > 0$	keine Zahl	Not a Number (NaN)

Tab. 2.2: Zahlenformat nach IEEE-754

Typ	Größe	Exponent	Mantisse	Werte des Exponenten (e)
single	32 Bit	8 Bit	23 Bit	$-126 \leq e \leq 127$
double	64 Bit	11 Bit	52 Bit	$-1022 \leq e \leq 1023$
	rel. Abstand zweier Zahlen	Dezimalstellen	betragsmäßig kleinste Zahl	größte Zahl
single	$2^{-(23+1)}$	7-8	$2^{-23} \cdot 2^{-126}$	$(1 - 2^{-24})2^{128}$
double	$2^{-(52+1)}$	15-16	$2^{-52} \cdot 2^{-1022}$	$(1 - 2^{-53})2^{1024}$

**Bemerkung 2.1.19** Der Standardtyp in Matlab, wenn nichts Weiteres angegeben wird, ist double.

**Bemerkungen 2.1.20** i) Die Werte von  $b, m, \ell$  hängen in der Regel vom Typ des jeweiligen Computers ab, ebenso  $r, R$ . Der IEEE<sup>9</sup>-Standard versucht dies zu vereinheitlichen.

ii) Selbst bei der Umwandlung „exakter“ Eingaben in Maschinenzahlen können Fehler auftreten, z.B. bei

$$x = 0.1 = (0.000\overline{1100})_2.$$

<sup>9</sup>IEEE~ Institute of Electrical and Electronics Engineers.

Mit Hilfe der folgenden mex-Routine `digits.c` lassen sich auf einem 32-Bit Rechner unter Matlab die einzelnen Bits einer double precision-Variablen auslesen. (Kompilieren der c-Routine auf der Kommandozeilenebene von Matlab mit `mex digits.c`.)

### C-Funktion: `digits.c`

```

1  #include "mex.h"
2  bool get_bit_at(double y[], int pos)
3  {
4      unsigned long bitmask;
5      unsigned long *cy = (unsigned long *) y;
6
7      if (pos >= 0){
8          if (pos < 32){
9              bitmask = 1 << pos;          // bitmask setzen
10             return (cy[0] & bitmask) != 0;
11         } else if (pos < 64){
12             bitmask = 1 << (pos-32); // bitmask setzen
13                                     // unsigned long ist nur
14                                     // 4 Byte = 32 Bit lang
15             return (cy[1] & bitmask) != 0;
16         }
17     }
18     mexErrMsgTxt("No valid position for Bit.");
19 }
20
21 void mexFunction( int nlhs, mxArray *plhs[],
22                  int nrhs, const mxArray *prhs[] )
23 {
24     double *x,*y;
25     mwSize mrows,ncols;
26     int k;
27     /* Check for proper number of arguments. */
28     if(nrhs!=1) {
29         mexErrMsgTxt("One input required.");
30     } else if(nlhs>1) {
31         mexErrMsgTxt("Too many output arguments");
32     }
33     /* The input must be a noncomplex scalar double.*/
34     mrows = mxGetM(prhs[0]);
35     ncols = mxGetN(prhs[0]);
36     if( !mxIsDouble(prhs[0]) || mxIsComplex(prhs[0]) ||
37         !(mrows==1 && ncols==1) ) {
38         mexErrMsgTxt("Input must be a noncomplex scalar double.");
39     }
40     /* Create matrix for the return argument. */
41     plhs[0] = mxCreateDoubleMatrix(1,64, mxREAL);
42     /* Assign pointers to each input and output. */
43     x = mxGetPr(prhs[0]);
44     y = mxGetPr(plhs[0]);
45     /* Call the digit subroutine. */
46     for (k=0; k<64; k++) y[63-k] = (double) get_bit_at(x,k);
47 }

```

Versuchen Sie sich die Ergebnisse der folgenden Matlab-Ausgaben zu erklären bzw. testen Sie es auf Ihrem eigenen Rechner.

```
>> A=digits(-1);A(1)
ans =
    1
>> A=digits(1);A(1)
ans =
    0
>> A=digits(2^(1023)*(1.9999999999999998));
>> A(1),A(2:12),[A(13:29); A(30:46); A(47:63)],A(64)
ans =
    0
ans =
    1    1    1    1    1    1    1    1    1    1    1    0
ans =
    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
ans =
    1
>> A=digits(2^(-971)*(2-1.9999999999999995));
>> A(1),A(2:12),[A(13:29); A(30:46); A(47:63)],A(64)
ans =
    0
ans =
    0    0    0    0    0    0    0    0    0    0    0    1
ans =
    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
ans =
    0
>> A=digits(2^(-972)*(2-1.9999999999999995));
>> A(1),A(2:12),[A(13:29); A(30:46); A(47:63)],A(64)
ans =
    0
ans =
    0    0    0    0    0    0    0    0    0    0    0    0
ans =
    1    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
ans =
    0
>> A=digits(realmax)           % try by yourself
>> A=digits(realmin)
>> A=digits(realmin/2^52)
>> A=digits(2^0+2^(-1)); A(13:29)
>> A=digits(2^0+2^(-1)+2^(-2)); A(13:29)
>> A=digits(2^0+2^(-1)+2^(-2)+2^(-3)); A(13:29)
>> A=digits(2^0+2^(-1)+2^(-2)+2^(-3)+2^(-17))A(13:29)1
```

## 2.2 Rundung und Maschinengenauigkeit

**Aufgabe:** Approximiere  $x \in \mathbb{R}$  durch  $fl(x) \in \mathbb{M}(b, m, n)$  („float“)

$$fl : \mathbb{R} \rightarrow \mathbb{M}(b, m, n)$$

bzw.

$$fl : [-x_{\max}, x_{\max}] \rightarrow \mathbb{M}(b, m, n)$$

Dies wird durch geeignete Rundungsstrategien gewährleistet. Wir betrachten zwei konkrete Beispiele:

### 2.2.1 Standardrundung, kaufmännische Rundung

(entspricht „normalem Runden“, also  $1.4 \rightarrow 1$ ,  $1.5 \rightarrow 2$ ,  $1.6 \rightarrow 2$ )

Sei  $x \in \mathbb{R}$  mit  $|x| \in [x_{\min}, x_{\max}]$  und  $b$ -adischer Darstellung

$$x = \pm b^\ell \sum_{j=1}^{\infty} d_j b^{-j},$$

wobei  $\ell$  ein  $n$ -stelliger Exponent sei. Es gelte  $d_1 \neq 0$  sowie  $d_j < b-1$  für unendlich viele  $j \in \mathbb{N}$ . Für eine vorgegebene Mantissenlänge  $m$  ist die **Standardrundung** definiert durch

$$fl(x; b, m, n) = fl(x) := \pm \sum_{j=1}^m d_j b^{\ell-j} \pm \begin{cases} 0 & , \text{ falls } d_{m+1} < \frac{b}{2} \\ b^{\ell-m} & , \text{ falls } d_{m+1} \geq \frac{b}{2} \end{cases}$$

und für die anderen Fälle gilt

$$\begin{aligned} |x| < x_{\min} &\Rightarrow fl(x) := 0 \\ |x| > x_{\max} &\Rightarrow \text{overflow.} \end{aligned}$$

**Bemerkung 2.2.1** Man mache sich klar, dass die Rundung im Falle  $d_{m+1} = b/2$  ( $b$  gerade,  $b/2$  ungerade) rein willkürlich ist.

**Beispiel 2.2.2** Betrachte  $\mathbb{M}(10, 4, 2)$ . Dann gilt für die Eingabe  $x = 0.21576 \cdot 10^3$  die Rundung  $fl(x) = 0.2158 \cdot 10^3$  und für  $y = 216.53$  ergibt sich  $fl(y) = 0.2165 \cdot 10^3$ , entscheidend sind also die signifikanten Stellen, nicht die Größe der Eingabe.

### 2.2.2 „Round to even“ (oder auch Banker’s Rule oder mathematisch unverzerrte Rundung)

Die unverzerrte Rundung (engl. „round to even“), auch als Banker’s Rule<sup>10</sup> bekannt, ist eine Rundungsregel, welche sich dann von der Standardrundung unterscheidet, wenn der zu rundende Rest genau zwischen zwei Zahlen mit der gewählten Anzahl von Ziffern liegt. Voraussetzung zur Anwendung der Banker’s Rule ist, dass die Basis  $b$  gerade und  $b/2$  ungerade ist, dies ist z.B. für die interessanten Fälle  $b = 2, 10$  erfüllt. Die Standardrundung erzeugt statistische „Fehler“, da das Aufrunden um  $b/2$  vorkommt, das Abrunden um  $b/2$  jedoch nie. Rundet man z.B. bei großen Mengen von Daten, bei denen der Trend untersucht werden soll, immer auf, falls der zu rundende

<sup>10</sup>Diese Rundungsart ist in der IEEE Norm P754 verankert.

Rest genau  $b/2$  ist, so verschiebt sich der Trend aufwärts. Außerdem ist das Verhalten bei positiven und negativen Zahlen unterschiedlich, wenn die zu rundende Ziffer  $b/2$  ist. Das „round to even“ vermeidet dieses Verhalten. Es rundet von der genauen Mitte zwischen zwei Ziffern immer zur nächsten geraden Zahl auf oder ab. Ansonsten entspricht es der Standardrundung.

**Bemerkung 2.2.3** Man könnte genauso gut auch zur nächsten ungeraden Zahl runden, dann würde man jedoch nie auf Null runden.

**Banker's Rule:** Voraussetzung:  $b$  gerade,  $b/2$  ungerade.

Sei  $x \in \mathbb{R}$  mit  $|x| \leq x_{\max}$  und  $b$ -adischer Darstellung

$$x = \pm b^\ell \sum_{j=1}^{\infty} d_j b^{-j},$$

wobei  $\ell$  ein  $n$ -stelliger Exponent sei. Es gelte  $d_1 \neq 0$  sowie  $d_j < b-1$  für unendlich viele  $j \in \mathbb{N}$ . Für eine vorgegebene Mantissenlänge  $m$  ist die **Banker's Rule** definiert durch

$$fl(x) := \pm \sum_{j=1}^m d_j b^{\ell-j} \pm \begin{cases} 0 & , \text{ falls } d_{m+1} < b/2 \\ b^{\ell-m} & , \text{ falls } d_{m+1} > b/2 \text{ oder} \\ & d_{m+1} = b/2 \text{ und es ex. ein } k > m+1 \text{ mit } d_k \neq 0 \\ 0 & , \text{ falls } \sum_{k=m+1}^{\infty} d_k b^{\ell-k} = b^{\ell-m}/2 \text{ und } d_m \text{ gerade} \\ b^{\ell-m} & , \text{ falls } \sum_{k=m+1}^{\infty} d_k b^{\ell-k} = b^{\ell-m}/2 \text{ und } d_m \text{ ungerade.} \end{cases}$$

**Beispiel 2.2.4** Für  $b = 10$  und  $m = 2$  erhält man mit der Banker's Rule folgende Ergebnisse:

aus 2.33 wird 2.3,	aus $-2.33$ wird $-2.3$ ,
aus 2.35 wird 2.4,	aus $-2.45$ (exakt) wird $-2.4$ ,
aus 2.45 (exakt) wird 2.4,	aus $-2.4500001$ wird $-2.5$ ,
aus 2.4500001 wird 2.5 und	aus $-2.449$ wird $-2.4$ .

**Bemerkungen 2.2.5** i) Der Vorteil der Banker's Rule liegt in einer „besseren“ statistischen Verteilung der Rundungsfehler.

ii) Zur Basis 10 lautet die Regel: Folgt auf die letzte beizubehaltende Ziffer lediglich eine 5 (oder eine 5, auf die nur Nullen folgen), so wird derart gerundet, dass die letzte beizubehaltende Ziffer gerade wird.

**Beispiel 2.2.6 (Wiederholtes Runden)**

$b$	Standard	Fehler	Banker's Rule	Fehler
10	2.445 $\rightarrow$ 2.45 $\rightarrow$ 2.5 $\rightarrow$ 3	0.555	2.445 $\rightarrow$ 2.44 $\rightarrow$ 2.4 $\rightarrow$ 2	0.445
10	4.545 $\rightarrow$ 4.55 $\rightarrow$ 4.6 $\rightarrow$ 5	0.455	4.545 $\rightarrow$ 4.54 $\rightarrow$ 4.5 $\rightarrow$ 4	0.545
2	0.101 $\rightarrow$ 0.11 $\rightarrow$ 1.0 $\rightarrow$ 1	0.011	0.101 $\rightarrow$ 0.10 $\rightarrow$ 0.1 $\rightarrow$ 0	0.101
2	0.011 $\rightarrow$ 0.10 $\rightarrow$ 0.1 $\rightarrow$ 1	0.101	0.011 $\rightarrow$ 0.10 $\rightarrow$ 0.1 $\rightarrow$ 0	0.011

**Bemerkung 2.2.7** Es lassen sich also sowohl bei der Standardrundung als auch bei der Banker's Rule Beispiele finden, bei denen durch wiederholtes Runden auf eine ganze Zahl der Fehler größer ist als der maximale Fehler  $b^{-1}$ , den man durch einmaliges Runden erhalten könnte.

**Definition 2.2.8 (absoluter/relativer Fehler)** Es sei  $x \in \mathbb{R}$  und  $\tilde{x} \in \mathbb{R}$  eine Approximation.

(i) Die Größe  $|x - \tilde{x}|$  heißt **absoluter Fehler**.

(ii) Die Größe  $\frac{|x - \tilde{x}|}{|x|}$  heißt **relativer Fehler**.

Der relative Fehler ist normalerweise die interessante Größe; er ist dimensionslos, während die Größe des absoluten Fehlers von der gewählten Maßeinheit abhängt.

**Lemma 2.2.9** Die Basis  $b$  sei gerade. Es sei  $x = \pm b^\ell \sum_{k=1}^{\infty} d_k b^{-k}$  ( $0 < d_1 < b$ ) und  $fl(x)$  die Standardrundung von  $x$  auf  $m$  Stellen. Dann gilt

$$|fl(x) - x| \leq \frac{b^{1-m}}{2} \cdot b^{\ell-1} = \frac{1}{2} b^{\ell-m}$$

und

$$\frac{|fl(x) - x|}{|x|} \leq \frac{b^{1-m}}{2}.$$

*Beweis.* O.B.d.A sei  $x$  positiv. Sei  $n \in \mathbb{N}$ . Für  $b \geq 2$  und  $0 \leq d_k \leq b-1$  ( $k = n, n+1, \dots$ ) gilt

$$\sum_{k=n}^{\infty} d_k b^{-k} \leq \sum_{k=n}^{\infty} (b-1) b^{-k} = \sum_{k=n-1}^{\infty} b^{-k} - \sum_{k=n}^{\infty} b^{-k} = b^{-(n-1)}. \quad (2.2)$$

Betrachten wir zuerst die Situation des Abrundens bei der Standardrundung, was der Situation  $d_{m+1} < b/2$  entspricht. Da  $b$  nach Voraussetzung gerade ist, gilt somit  $b/2 \in \mathbb{Z}$ . Somit folgt aus  $d_{m+1} < b/2$ , dass auch  $d_{m+1} + 1 \leq b/2$  gilt. Da  $0 < d_1 < b$  vorausgesetzt wurde (was der normierten Gleitpunktdarstellung 2.1.13 entspricht), hat man  $b^{-1} \leq \sum_{k=1}^{\infty} d_k b^{-k} \leq 1$  und somit  $b^{\ell-1} \leq x \leq b^\ell$ . Beachtet man  $x - fl(x) = b^\ell \sum_{m+1}^{\infty} d_k b^{-k}$ , so schließt man nun für  $d_{m+1} < b/2$  mit (2.2)

$$\begin{aligned} |fl(x) - x| &= b^\ell \left[ d_{m+1} b^{-(m+1)} + \sum_{k=m+2}^{\infty} d_k b^{-k} \right] \leq b^\ell \left[ d_{m+1} b^{-(m+1)} + b^{-(m+1)} \right] \\ &= b^{\ell-(m+1)} (d_{m+1} + 1) \leq b^{\ell-(m+1)} \cdot \frac{b}{2} = \frac{b^{1-m}}{2} \cdot b^{\ell-1}. \end{aligned} \quad (2.3)$$

Für das Aufrunden, d.h. für den Fall  $d_{m+1} \geq \frac{b}{2}$  gilt nun  $d_{m+1} b^{-1} \geq \frac{1}{2}$  und analog zu (2.3)

$$\begin{aligned} |fl(x) - x| &= b^\ell \left[ b^{-m} - \sum_{k=m+1}^{\infty} d_k b^{-k} \right] \leq b^\ell \left[ b^{-m} - d_{m+1} b^{-(m+1)} \right] \\ &= b^{\ell-m} \underbrace{\left| 1 - \underbrace{d_{m+1} b^{-1}}_{\substack{\geq 1/2 \\ \leq 1/2}} \right|}_{\substack{\geq 1/2 \\ \leq 1/2}} \leq \frac{b^{1-m}}{2} \cdot b^{\ell-1}. \end{aligned}$$

Damit ist die Abschätzung für den absoluten Fehler bewiesen und die Abschätzung für den relativen Fehler erhält man sofort mit  $b^{\ell-1} \leq x$ .  $\square$

**Definition 2.2.10** Die Zahl  $\text{eps} := \frac{b^{1-m}}{2}$  heißt (relative) **Maschinengenauigkeit**.

**Definition 2.2.11** Wir definieren

$$\mathbb{M}' := \{x \in \mathbb{R} : x_{\min} \leq |x| \leq x_{\max}\} \cup \{0\}.$$

**Bemerkungen 2.2.12** i) Es gilt  $\text{eps} = \inf \{ \delta > 0 : fl(1 + \delta) > 1 \}$ .

ii) In Matlab liefert die Anweisung `eps` die Maschinengenauigkeit, die sich aber auch durch folgende Routine `mineps.m` bestimmen lässt. Unter anderen Programmiersprachen geht dies ähnlich, man muss sich jedoch vergewissern, dass der Compiler die Anweisung `while 1 < 1 + value` nicht zu `while 0 < value` optimiert!

iii)  $\forall x \in \mathbb{M}', x \neq 0 \exists \varepsilon \in \mathbb{R}$  mit  $|\varepsilon| < \text{eps}$  und  $\frac{|fl(x) - x|}{|x|} = \varepsilon$ , d.h.

$$fl(x) = x(1 + \varepsilon). \quad (2.4)$$

### MATLAB-Funktion: mineps.m

```
1 function value = mineps
2 value = 1;
3 while 1 < 1 + value
4     value = value / 2;
5 end
6 value = value * 2;
```

### MATLAB-Beispiel:

Die Funktion `mineps` liefert das gleiche Ergebnis wie die Matlab-Konstante `eps`. Ergänzend sei noch die Matlab-Konstante `realmin` wiedergegeben. Sie ist die kleinste positive darstellbare Maschinenzahl.

```
>> eps
ans =
    2.220446049250313e-016
>> mineps
ans =
    2.220446049250313e-016
>> realmin
ans =
    2.225073858507201e-308
```

## 2.3 Gleitkommaarithmetik

Ein Algorithmus ist eine Folge von arithmetischen Operationen, mit denen Maschinenzahlen verknüpft werden. Ein Computer rechnet also mit Maschinenzahlen durch die Realisierung (Implementierung) eines Algorithmus.

**Beispiel 2.3.1** Betrachte  $\mathbb{M}(10, 3, 1)$ . Dann gilt:

$$\underbrace{0.346 \cdot 10^3}_{\in \mathbb{M}(10,3,1)} + \underbrace{0.785 \cdot 10^3}_{\in \mathbb{M}(10,3,1)} = 0.1131 \cdot 10^4 \notin \mathbb{M}(10,3,1).$$

Aus diesem einfachen Beispiel sieht man, dass das Ergebnis einer arithmetischen Operation auf Maschinenzahlen **keine** Maschinenzahl sein muss!

**Voraussetzung 2.3.2** Wir nehmen an, dass für die Addition  $\oplus$  im Rechner gilt

$$x \oplus y = fl(x + y) \quad \text{für alle } x, y \in \mathbb{M} \text{ mit } |x + y| \in \mathbb{M}'$$

und Entsprechendes auch für die anderen Grundrechenarten:  $\nabla \in \{+, -, *, \div\}$  wird durch eine Gleitpunktoperation  $\mathring{\nabla}$  ersetzt mit

$$x \mathring{\nabla} y = fl(x \nabla y).$$

Man sagt  $\oplus$  (bzw.  $\mathring{\nabla}$ ) ist **exakt gerundet**.

**Folgerung 2.3.3** Für alle  $x, y \in \mathbb{M}$  mit  $x + y \in \mathbb{M}'$  gilt

$$x \oplus y = (x + y)(1 + \varepsilon) \quad (2.5)$$

für ein geeignetes  $|\varepsilon| \leq \text{eps}$ , und Entsprechendes auch für die anderen Grundrechenarten:  $x \mathring{\nabla} y = (x \nabla y)(1 + \varepsilon)$  mit  $|\varepsilon| \leq \text{eps}$ . Dies folgt aus  $fl(x) = x(1 + \varepsilon)$ .

**Bemerkung 2.3.4** (a) Maschinenoperationen sind nicht assoziativ, wie folgendes Beispiel in  $\mathbb{M}(10, 3, 1)$  zeigt. Für  $x = 6590 = 0.659 \cdot 10^4$ ,  $y = 1 = 0.100 \cdot 10^1$  und  $z = 4 = 0.400 \cdot 10^1$  ergibt sich bei exakter Rechnung  $(x + y) + z = x + (y + z) = 6595$ . Für die Maschinenoperationen gilt hingegen auf der einen Seite

$$x \oplus y = 0.659 \cdot 10^4 = 6590 \Rightarrow (x \oplus y) \oplus z = 6590,$$

aber

$$y \oplus z = 0.500 \cdot 10 = 5 \Rightarrow x \oplus (y \oplus z) = 6660$$

Dies zeigt, dass die Reihenfolge der Verknüpfungen wesentlich ist!

Als „Faustregel“ (die aber auch nicht immer hilft, s.u.) kann man sagen, dass man die Summation in der Reihenfolge aufsteigender Beträge ausführen sollte, um Rundungsfehler zu minimieren.

(b) Ebenso gilt das Distributivgesetz nicht mehr.

(c) Als Beispiel betrachte folgende Subtraktion:  $0.73563 - 0.73441 = 0.00122$ . Bei 3-stelliger Rechnung:  $0.736 - 0.734 = 0.002 = 0.2 \cdot 10^{-2}$ , d.h., der absolute Fehler ist in der Größenordnung des Rundungsfehlers, für den relativen Fehler gilt jedoch

$$\frac{0.002 - 0.00122}{0.00122} = 0.64 \hat{=} 64\%$$

also sehr groß. Dieser unangenehme Effekt der Gleitpunktarithmetik heißt **Auslöschung** (genauer später).

**Beispiel 2.3.5 (Auslöschung)** (a) Für  $b = 10$  und  $m = 2$  gilt

$$(100 \oplus 4) \oplus 4 = 100 \neq 110 = 100 \oplus (4 \oplus 4).$$

Man beachte dabei  $100 \oplus 4 = 100$ , da  $100 + 4$  auf 2 Stellen gerundet 100 ergibt, und dass  $100 + 8$  auf 2 Stellen gerundet 110 ergibt. Wir sehen also wieder den Effekt der Auslöschung.

(b) Wir betrachten die Rechnung

$$0.1236 + 1.234 - 1.356 = 0.0016 = 0.1600 \cdot 10^{-2}.$$

Gleitpunktrechnung mit  $b = 10, m = 4$  liefert

$$(0.1236 \oplus 1.234) \ominus 1.356 = 1.358 \ominus 1.356 = 0.2000 \cdot 10^{-2},$$

d.h. schon die erste Stelle des berechneten Ergebnisses ist falsch! Der Rundungsfehler der ersten Addition wird durch die nachfolgende Subtraktion extrem verstärkt.



**Bemerkung 2.3.6** Auch wenn alle Summanden positiv sind, kann etwas schiefgehen: Sei wie im Beispiel 2.3.5  $b = 10$  und  $m = 4$ . Wir wollen

$$\sum_{n=0}^N a_n \text{ mit } a_0 = 1 \quad (2.6)$$

berechnen. Falls  $0 \leq a_n < 10^{-4}$  für alle  $n > 0$ , und falls wir der Reihe nach summieren, ist das berechnete Ergebnis 1, egal wie groß die Summe wirklich ist. Eine erste Abhilfe ist es nach der obigen Faustregel, der Größe nach zu summieren (zuerst die kleinste Zahl). In unserem Beispiel (2.6) hilft das allerdings nicht wirklich, das berechnete Ergebnis ist dann höchstens 2. Besser ist jedoch folgendes Vorgehen: Man ersetzt die zwei kleinsten Zahlen durch ihre Summe und wiederholt diesen Vorgang so lange, bis nur noch eine Zahl, nämlich die gesuchte Summe, übrigbleibt.

**Beispiel 2.3.7 (Guard Digit)** Sei  $\mathbb{M} = \mathbb{M}(10, 3, 1)$  und betrachte  $\ominus$ . Sei weiter  $x = 0.215 \cdot 10^9$  und  $y = 0.125 \cdot 10^{-9}$ . Naive Realisierung von  $x \ominus y = fl(x - y)$  erfordert **schieben** von  $y$  auf den größeren Exponenten  $y = 0.125 \cdot 10^{-18} \cdot 10^9$  und **subtrahieren** der Mantissen:

$$\begin{array}{rcl} x & = & 0.21500000000000000000 \cdot 10^9 \\ y & = & 0.000000000000000000125 \cdot 10^9 \\ \hline x - y & = & 0.214999999999999999875 \cdot 10^9 \end{array} \quad (2.7)$$

**Runden** auf drei Stellen liefert dann  $x \ominus y = 0.215 \cdot 10^9$ . Dies erfordert einen Addierer mit  $2(b^n - 1) + m = 21$  Stellen. In diesem Fall hätten wir das Ergebnis auch durch die Abfolge **Schieben**, **Runde**  $y$ , **Subtrahiere** erhalten. Im Allgemeinen ist das aber nicht gut wie folgendes Beispiel zeigt:

$$\begin{array}{rcl} x & = & 0.101 \cdot 10^1 \\ y & = & 0.993 \cdot 10^0 \end{array} \longrightarrow \begin{array}{rcl} x & = & 0.101 \cdot 10^1 \\ y & = & 0.099 \cdot 10^1 \\ \hline x \ominus y & = & 0.002 \cdot 10^1 \end{array} .$$

Für den relativen Fehler im Ergebnis gilt dann

$$\frac{(x \ominus y) - (x - y)}{(x - y)} = \frac{0.02 - 0.017}{0.017} \approx 0.176 \approx 35 \text{ eps},$$

wobei  $\text{eps} = 1/2 \cdot 10^{-3+1} = 0.005$  sei.

Verwenden wir nun einen  $m + 1$ -stelligen Addierer, dann erhalten wir

$$\begin{array}{rcl} x & = & 0.1010 \cdot 10^1 \\ y & = & 0.0993 \cdot 10^1 \\ \hline x - y & = & 0.0017 \cdot 10^1 \end{array} .$$

Das Ergebnis  $x \ominus y = 1.7 \cdot 10^{-2}$  ist exakt!



**Bemerkung 2.3.8** Allgemein kann man zeigen (siehe [Knuth]): Mit einer zusätzlichen Stelle (so genannter **Schutzziffer** oder **Guard Digit**) gilt

$$\frac{(x \ominus y) - (x - y)}{(x - y)} \leq 2 \text{ eps} .$$

Mit nur 2 Guard Digits erhält man sogar genau das gerundete Ergebnis der exakten Rechnung, d.h. das gerundete Ergebnis aus (2.7) ohne einen  $2(b^n - 1) + m$ -stelligen Addierer.

Völlig unkalkulierbar sind Gleitkommaoperationen aber nicht. Es gilt z.B. folgendes Resultat:  
Sind  $u, v$  Gleitkommazahlen und

$$\begin{aligned} u' &= (u \oplus v) \ominus v, & v' &= (u \oplus v) \ominus u, \\ u'' &= (u \oplus v) \ominus v', & v'' &= (u \oplus v) \ominus u', \end{aligned}$$

dann folgt

$$u + v = (u \oplus v) + ((u \ominus u') \oplus (v \ominus v'')).$$

Dies erlaubt eine Berechnung des Fehlers mittels Gleitkommaarithmetik. (Siehe [Knuth, 4.2.2, Theorem B].)

## 2.4 Fehlerverstärkung bei elementaren Rechenoperationen

In einem Programm/Algorithmus werden **viele** Rechenoperationen durchgeführt. Es stellt sich also die Frage, was mit Fehlern im weiteren Programmverlauf passiert. Hier beschäftigen wir uns zunächst mit **exakten Operationen**, d.h. ohne weitere Rundung.

Seien also  $x, y$  die exakten Eingaben und  $\tilde{x}, \tilde{y}$  die gestörten (realen) Eingaben, mit den relativen Fehlern

$$\delta_x := \frac{\tilde{x} - x}{x}, \quad \delta_y := \frac{\tilde{y} - y}{y},$$

d.h.

$$\tilde{x} = x(1 + \delta_x), \quad \tilde{y} = y(1 + \delta_y), \quad (2.8)$$

und wir nehmen an, dass  $|\delta_x|, |\delta_y| \leq \varepsilon < 1$  gilt, also, dass die Eingabefehler „klein“ sind.

Da ein Algorithmus aus den Grundrechenarten zusammengesetzt ist, betrachten wir die Fehlerverstärkung bei diesen elementaren Rechenoperationen.

**Bemerkung 2.4.1 (Multiplikation)** Offenbar gilt

$$\tilde{x} * \tilde{y} = (x(1 + \delta_x)) * (y(1 + \delta_y)) = (x * y)(1 + \delta_x + \delta_y + \delta_x \delta_y) =: (x * y)(1 + \delta)$$

mit  $|\delta| = |\delta_x + \delta_y + \delta_x \delta_y| \leq \varepsilon(2 + \varepsilon)$ , d.h. bei Vernachlässigung des quadratischen Terms ist dies von der Ordnung  $2\varepsilon$ . Damit ergibt sich für den relativen Fehler (falls  $|\delta_x|, |\delta_y| \leq \varepsilon$  und  $\varepsilon < 1$ )

$$\left| \frac{\tilde{x} * \tilde{y} - x * y}{x * y} \right| = |\delta| = 2\varepsilon + \varepsilon \cdot \varepsilon < 3\varepsilon \leq 3\varepsilon.$$

Also bleibt der relative Fehler im Rahmen der Maschinengenauigkeit.

**Bemerkung 2.4.2 (Division)** Bei der Division gilt

$$\frac{\tilde{x}}{\tilde{y}} = \frac{x(1 + \delta_x)}{y(1 + \delta_y)}.$$

Mit Hilfe der geometrischen Reihe ( $\frac{1}{1-q} = \sum_{j=1}^{\infty} q^j$ , verwende  $q = -\delta_y$ ) gilt

$$\frac{1}{1 + \delta_y} = 1 + \sum_{j=1}^{\infty} (-1)^j \delta_y^j,$$

und damit folgt

$$\frac{1 + \delta_x}{1 + \delta_y} = 1 + \delta_x + (1 + \delta_x) \sum_{j=1}^{\infty} (-1)^j \delta_y^j =: 1 + \delta.$$

Für den Fehlerverstärkungsfaktor  $\delta$  gilt dann unter der Annahme  $|\delta_x|, |\delta_y| \leq \varepsilon$

$$|\delta| \leq \varepsilon + (1 + \varepsilon) \underbrace{\varepsilon \sum_{j=0}^{\infty} \varepsilon^j}_{= \frac{1}{1-\varepsilon}} = \varepsilon \left( 1 + \frac{1 + \varepsilon}{1 - \varepsilon} \right).$$

Falls z.B.  $\varepsilon \leq \frac{1}{2}$  gilt, folgt  $\frac{1+\varepsilon}{1-\varepsilon} \leq 3$  und damit  $|\delta| \leq 4\varepsilon$ , also zeigt die Division ein ähnliches Verhalten wie die Multiplikation, d.h. der relative Fehler bleibt im Rahmen der Maschinengenauigkeit.

**Bemerkung 2.4.3 (Addition)** Im Fall der Addition gilt

$$\begin{aligned} \tilde{x} + \tilde{y} &= x(1 + \delta_x) + y(1 + \delta_y) = x + y + x\delta_x + y\delta_y \\ &= (x + y) \left( 1 + \frac{x}{x+y}\delta_x + \frac{y}{x+y}\delta_y \right) =: (x + y)(1 + \delta) \end{aligned}$$

mit der folgenden Abschätzung für den Fehlerverstärkungsfaktor

$$|\delta| \leq (|\delta_x| + |\delta_y|) \max \left\{ \left| \frac{x}{x+y} \right|, \left| \frac{y}{x+y} \right| \right\} \leq 2\varepsilon \max \left\{ \left| \frac{x}{x+y} \right|, \left| \frac{y}{x+y} \right| \right\}$$

falls  $|\delta_x|, |\delta_y| \leq \varepsilon$ . Man beachte, dass der Term

$$\max \left\{ \left| \frac{x}{x+y} \right|, \left| \frac{y}{x+y} \right| \right\}$$

nicht gegen  $\varepsilon$  abgeschätzt werden kann. Insbesondere für  $x \not\approx -y$  kann dieser Faktor beliebig groß werden, der Ausgabefehler ist also **nicht** notwendigerweise in der Größe des Eingabefehlers. Dies ist die Erklärung der Auslöschung und kann durch folgende Überlegung verdeutlicht werden. Nehmen wir an, wir wollten  $x + y$  berechnen für zwei Eingaben mit folgender Darstellung

$$x = 0.d_1 \dots d_r d_{r+1} \dots d_m * b^e, \quad y = -0.d_1 \dots d_r \tilde{d}_{r+1} \dots \tilde{d}_m * b^e,$$

d.h., die ersten  $r$  Stellen von  $x$  und  $y$  sind identisch. Wir nehmen der Einfachheit halber an  $d_j \geq \tilde{d}_j$ ,  $j \geq r+1$ ,  $d_{r+1} > \tilde{d}_{r+1}$ . Daraus folgt

$$x + y = 0.0 \dots 0 \hat{d}_{r+1} \dots \hat{d}_m * b^e = 0.\hat{d}_{r+1} \dots \hat{d}_m * b^{e-r}$$

mit  $\hat{d}_j \geq 0$  und damit für den relativen Fehler

$$\left| \frac{x}{x+y} \right| = \left| \frac{0.d_1 \dots d_m * b^e}{0.\hat{d}_{r+1} \dots \hat{d}_m * b^{e-r}} \right| \leq \frac{b \cdot b^e}{b^{-1} b^{e-r}} = b^{r+2},$$

da  $0.d_1 \dots d_m \leq b$  und  $0.\hat{d}_{r+1} \dots \hat{d}_m * b^{-1}$ . Man verliert also mindestens  $r$  Stellen an Genauigkeit, diese werden „ausgelöscht“. Intern (im Computer) werden diese Stellen zufällig aufgefüllt mit Information, die sich in den entsprechenden Speicherzellen befindet. Egal, was dort steht, diese Stellen sind im Prinzip immer falsch, da bezüglich  $x$  und  $y$  keine Information jenseits der  $m$ -ten Stelle vorliegt.

## 2.5 Kondition eines Problems

Bislang haben wir einzelne Operationen untersucht, nun werden diese Operationen zu einem Algorithmus verkettet. Ein Algorithmus wird dann verwendet, um ein mathematisches Problem zu lösen. Wir wollen nun untersuchen, welche Probleme „einfach“ bzw. „schwer“ sind und welche Algorithmen „gut“ bzw. „schlecht“ sind. Wir beginnen mit der Untersuchung der Problemstellungen.

**Beispiel 2.5.1** Diskutieren wir am Anfang zuerst folgendes geometrische Problem: die zeichnerische Bestimmung des Schnittpunkts  $S$  zweier Geraden  $g$  und  $h$  in der Ebene. Schon beim Zeichnen haben wir Schwierigkeiten, die Geraden ohne Fehler darzustellen. Die Frage, die wir näher untersuchen wollen, lautet, wie stark der Schnittpunkt  $S$  (Output) von den Zeichenfehlern (Fehler im Input) abhängt.

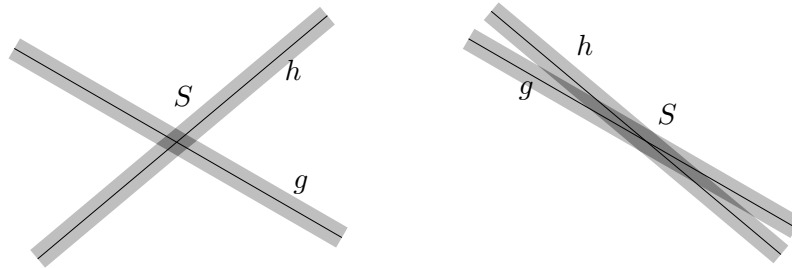


Abb. 2.2: Schnittpunkt  $S$  zweier Geraden  $g$  und  $h$ . Links gut konditioniert, rechts schlecht konditioniert.

Wie wir der Grafik direkt entnehmen können, hängt der Fehler in der Ausgabe stark davon ab, in welchem Winkel  $\angle(g, h)$  sich die Geraden schneiden. Stehen  $g$  und  $h$  annähernd senkrecht aufeinander, so variiert der Schnittpunkt  $S$  etwa im gleichen Maße wie der Fehler beim Zeichnen von  $g$  und  $h$ . In diesem Fall bezeichnet man das Problem, den Schnittpunkt  $S$  zeichnerisch zu bestimmen, als **gut konditioniert**.

Ist jedoch der Winkel  $\angle(g, h)$  sehr klein ( $g$  und  $h$  sind fast parallel), so kann man schon mit dem Auge keinen genauen Schnittpunkt  $S$  ausmachen, und eine kleine Lageänderung von  $g$  oder  $h$  liefert einen gänzlich anderen Schnittpunkt. Man spricht dann von einem **schlecht konditionierten** Problem.

Kommen wir nun zu einer mathematischen Präzisierung des Konditionsbegriffs. Dazu brauchen wir einen etwas allgemeineren Rahmen und werden später sehen, dass dieser Rahmen für die Untersuchung sehr gut geeignet ist.

**Problem 2.5.2** Seien  $X, Y$  Mengen und  $\varphi : X \rightarrow Y$ . Wir betrachten das Problem:

$$\text{Gegeben sei } x \in X, \text{ gesucht } y = \varphi(x).$$

Wir untersuchen also, wie sich Störungen in den Daten  $x$  auf das Ergebnis  $y$  auswirken. Man beachte, dass dies nichts mit der Realisierung auf dem Computer (dem Algorithmus) zu tun hat, sondern einzig eine Eigenschaft der Problemstellung ist.

**Beispiel 2.5.3 (Noch einmal der Geradenschnittpunkt)** Die beiden Geraden seien in folgender Form gegeben

$$G_1 = \{(x_1, x_2) \in \mathbb{R}^2 : a_{1,1}x_1 + a_{1,2}x_2 = b_1\}, \quad G_2 = \{(x_1, x_2) \in \mathbb{R}^2 : a_{2,1}x_1 + a_{2,2}x_2 = b_2\},$$

wobei  $b = (b_1, b_2)^T \in \mathbb{R}^2$  und die Koeffizienten  $a_{i,j}$  für  $i, j = 1, 2$  gegeben sind. Mit  $A := (a_{ij})_{i,j=1,2} \in \mathbb{R}^{2 \times 2}$  ist der gesuchte Schnittpunkt  $x = (x_1, x_2)^T$  von  $G_1, G_2$  gegeben durch

$$Ax = b,$$

also die Lösung eines linearen Gleichungssystems. Falls  $A$  regulär ist können wir schreiben  $x = A^{-1}b$ . Also sind  $A$  und  $b$  die Eingaben und  $x$  die Ausgabe, d.h.  $X = \mathbb{R}^{2 \times 2} \times \mathbb{R}^2$ ,  $Y = \mathbb{R}^2$  und  $\varphi(A, b) = A^{-1}b = x$ .

**Beispiel 2.5.4** Betrachten wir das Beispiel 1.0.1 des Kohleaushubs. Wir gehen davon aus, dass wir aus den Messungen bereits eine Höhenfunktion  $h(x) = z$ ,  $h : \mathbb{R}^2 \rightarrow \mathbb{R}$ , gewonnen haben, die jedem Punkt  $x \in \mathbb{R}^2$  einer Karte die Höhe zuweist. Weiterhin sei das Kohlerevier in einem Rechteck  $R := [a, b] \times [c, d] \subset \mathbb{R}^2$  enthalten. Dann lautet die Formel für den Kohleaushub

$$f(h) = \int_a^b \int_c^d h(x) dx_2 dx_1,$$

also ein Doppelintegral. Damit ist klar, dass  $Y = \mathbb{R}$ , aber für welche Eingaben ist  $\varphi$  definiert? Offenbar muss die Funktion  $h$ , die hier als Eingabe fungiert, integrierbar sein, etwa  $h \in \mathcal{R}(\mathbb{R}) = X$ , wobei  $\mathcal{R}(\mathbb{R})$  die Menge der Riemann-integrierbaren Funktionen auf  $\mathbb{R}$  bezeichnet. Man beachte, dass  $X$  hier (im Gegensatz zu den beiden ersten Beispielen) unendlich-dimensional ist.

Beide Beispiele zeigen, dass der in Problem 2.5.2 beschriebene Rahmen in der Tat angemessen ist.

**Beispiel 2.5.5** Wir betrachten den Spezialfall  $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}$  mit

$$\varphi(x) = \langle a, x \rangle + b, \quad a \in \mathbb{R}^n, \quad b \in \mathbb{R},$$

wobei  $\langle \cdot, \cdot \rangle$  das (euklidische) Skalarprodukt bezeichne. Seien  $x, \tilde{x} \in \mathbb{R}^n$ . Der relative Fehler von  $\varphi(\tilde{x})$  bezüglich  $\varphi(x)$  lässt sich folgendermaßen abschätzen (falls  $\varphi(x) \neq 0$  und  $x_j \neq 0$  für alle  $j$ ):

$$\left| \frac{\varphi(x) - \varphi(\tilde{x})}{\varphi(x)} \right| = \frac{|\langle a, x - \tilde{x} \rangle|}{|\varphi(x)|} \leq \frac{\sum_{j=1}^n |a_j| |x_j - \tilde{x}_j|}{|\varphi(x)|} = \sum_{j=1}^n \frac{|x_j|}{|\varphi(x)|} |a_j| \cdot \frac{|x_j - \tilde{x}_j|}{|x_j|}. \quad (2.9)$$

Setzen wir  $\tilde{x}_j = fl(x_j)$  in (2.9) ein, so ergibt sich mit Lemma 2.2.9, dass die Zahl

$$\text{eps} \cdot \sum_{j=1}^n \frac{|x_j|}{|\varphi(x)|} |a_j|$$

den unvermeidlichen Fehler (d.h. selbst bei exakter Rechnung) bei der Berechnung von  $\varphi$  repräsentiert. Sei nun  $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}$  differenzierbar in  $x$ . Dann gilt mit der Taylor-Entwicklung<sup>11</sup>

$$\varphi(x) - \varphi(\tilde{x}) = \langle \nabla \varphi(x), x - \tilde{x} \rangle + o(\|x - \tilde{x}\|),$$

und (2.9) wird zu

$$\frac{|\varphi(x) - \varphi(\tilde{x})|}{|\varphi(x)|} \leq \sum_{j=1}^n \frac{|x_j|}{|\varphi(x)|} \left| \frac{\partial \varphi}{\partial x_j}(x) \right| \cdot \frac{|x_j - \tilde{x}_j|}{|x_j|} + o(\|x - \tilde{x}\|).$$

Somit gibt der Faktor  $\frac{|x_j|}{|\varphi(x)|} \left| \frac{\partial \varphi}{\partial x_j}(x) \right|$  an, wie stark sich ein relativer Fehler in  $x_j$  auf den relativen Fehler von  $\varphi(x)$  auswirkt. Ist schließlich  $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^m$ , so können wir die bisherigen Überlegungen auf jede Komponente von  $\varphi$  einzeln anwenden.

Dies motiviert den Begriff der Konditionszahlen.

**Definition 2.5.6 (Konditionszahlen)** Sei  $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^m$  differenzierbar in  $x \in \mathbb{R}^n$  sowie  $\varphi_i(x) \neq 0$ ,  $1 \leq i \leq m$ . Die Zahlen

$$\kappa_{ij}(x) = \frac{|x_j|}{|\varphi_i(x)|} \left| \frac{\partial \varphi_i}{\partial x_j}(x) \right|, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n, \quad (2.10)$$

heißen die **Konditionszahlen** von  $\varphi$  in  $x$ .

<sup>11</sup>Die Landau-Symbole  $o$  und  $\mathcal{O}$  sind im Anhang A erklärt.

**Beispiel 2.5.7** 1. Multiplikation:  $\varphi : \mathbb{R}^2 \rightarrow \mathbb{R}, \varphi(x_1, x_2) = x_1 \cdot x_2$ ,

$$\kappa_1(x) = \frac{|x_1|}{|x_1 x_2|} \left| \frac{\partial \varphi}{\partial x_1}(x) \right| = 1, \quad \kappa_2(x) = 1.$$

Keine Verstärkung des relativen Fehlers; die Multiplikation ist „gut konditioniert“.

2. Addition:  $\varphi : \mathbb{R}^2 \rightarrow \mathbb{R}, \varphi(x_1, x_2) = x_1 + x_2$ ,

$$\kappa_1(x) = \frac{|x_1|}{|x_1 + x_2|} \left| \frac{\partial \varphi}{\partial x_1}(x) \right| = \frac{|x_1|}{|x_1 + x_2|}, \quad \kappa_2(x) = \frac{|x_2|}{|x_1 + x_2|}.$$

Im Falle der Auslöschung, d.h. wenn  $|x_j| \gg |x_1 + x_2|$ , große Verstärkung des relativen Fehlers; die Addition ist in diesem Fall „schlecht konditioniert“.

3. Lösen der quadratischen Gleichung  $x^2 + 2px - q = 0$  im Fall  $p, q > 0$ : Berechnung der größeren der beiden Nullstellen (Mitternachtsformel):

$$\varphi(p, q) = -p + \sqrt{p^2 + q}, \quad \varphi : \mathbb{R}^2 \rightarrow \mathbb{R}, \quad (2.11)$$

Eine Rechnung ergibt

$$\kappa_p = \frac{p}{\sqrt{p^2 + q}}, \quad \kappa_q = \frac{p + \sqrt{p^2 + q}}{2\sqrt{p^2 + q}}, \quad (2.12)$$

also  $\kappa_p \leq 1$ ,  $\kappa_q \leq 1$ ; das Problem ist ebenfalls „gut konditioniert“. Im Fall  $q < 0$ ,  $q \approx -p^2$  hingegen wäre das Problem „schlecht konditioniert“.

## 2.6 Numerische Stabilität eines Algorithmus

Jeder Algorithmus zur Lösung von Problem 2.5.2 lässt sich auffassen als eine Abbildung  $\tilde{\varphi} : X \rightarrow Y$ . Von einem guten Algorithmus wird man erwarten, dass er die relativen Fehler nur unwesentlich mehr verstärkt als die Konditionszahlen  $\kappa_{ij}(x)$  des Problems  $\varphi$  es erwarten lassen. Für  $\tilde{\varphi} : \mathbb{R}^n \rightarrow \mathbb{R}^m$  sollte also eine Abschätzung der Form

$$\left| \frac{\tilde{\varphi}_i(\tilde{x}) - \tilde{\varphi}_i(x)}{\tilde{\varphi}_i(x)} \right| \leq C_{i1} \underbrace{\sum_{j=1}^n \kappa_{ij}(x) \frac{|\tilde{x}_j - x_j|}{|x_j|}}_{\geq \frac{|\varphi_i(\tilde{x}) - \varphi_i(x)|}{|\varphi_i(x)|} + o(\|\tilde{x} - x\|)} + C_{i2} n \text{ eps} \quad (i = 1 \dots, m), \quad (2.13)$$

(oder so ähnlich) gelten mit Konstanten  $C_{i1}, C_{i2} \geq 0$ , welche nicht viel größer als 1 sind.

**Definition 2.6.1 (Numerische Stabilität eines Algorithmus)** Ein Algorithmus  $\tilde{\varphi} : \mathbb{R}^n \rightarrow \mathbb{R}^m$  zur Lösung eines Problems  $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^m$  heißt **numerisch stabil**, falls (2.13) oder etwas Ähnliches gilt mit vernünftigen Konstanten  $C_{i1}, C_{i2}$ . Andernfalls heißt der Algorithmus **numerisch instabil**.

**Bemerkung 2.6.2 (Konsistenz eines Algorithmus)** Aus (2.13) ist nicht ersichtlich, was das exakte Datum  $\varphi(x)$  mit  $\tilde{\varphi}(x)$  zu tun hat. (Z.B. ist  $\tilde{\varphi}(x) = 0$  ein stabiles Verfahren, aber im Allgemeinen nutzlos.) Für diesen Zusammenhang hat man in der Numerik den Begriff der **Konsistenz**. Ein numerisches Verfahren heißt **konsistent**, wenn der Algorithmus (in einer noch näher zu bestimmenden Art und Weise) tatsächlich das gegebene Problem löst und nicht ein anderes. Wir werden dies später (in Numerik 4) noch insbesondere im Zusammenhang mit der numerischen Lösung von gewöhnlichen Differentialgleichungen untersuchen, aber man mag schon jetzt festhalten, dass Stabilität und Konsistenz für die Konvergenz einer numerischen Näherungslösungsfolge gegen die exakte Lösung wichtig sind.

gut konditioniert entfernt  
schlecht konditioniert entfernt

**Bemerkung 2.6.3 (Standardfehler)** Typisch für das Entstehen numerischer Instabilität ist, dass man das Ausgangsproblem  $\varphi$  in zwei Teilschritte  $\varphi = \varphi^{(2)} \circ \varphi^{(1)}$  zerlegt, von denen einer erheblich schlechter konditioniert ist als das Ausgangsproblem.

**Beispiel 2.6.4 (Quadratische Gleichung)** Wir untersuchen zwei verschiedene Verfahren zur Lösung von (2.11) in Beispiel 2.5.7

$$\varphi(p, q) = -p + \sqrt{p^2 + q} \text{ löst } x^2 + 2px - q = 0.$$

**Methode 2.6.5**

$$u = \sqrt{p^2 + q}, \quad y = \chi_1(p, u) = -p + u. \quad (2.14)$$

Falls  $u \approx p$ , d.h. falls  $p \gg q$ , sind die Konditionszahlen von  $\chi_1$  erheblich größer als 1 (Auslöschung).

**Methode 2.6.6**

$$u = \sqrt{p^2 + q}, \quad y = \chi_2(p, q, u) = \frac{q}{p + u}. \quad (2.15)$$

Die Konditionszahlen von  $\chi_2$  sind kleiner als 1. (Das Verfahren beruht darauf, dass  $-p - u$  die andere Lösung und das Produkt der beiden Lösungen gleich  $-q$  ist (Satz von Viëta).)

Es stellt sich heraus, dass Algorithmus 2.6.5 numerisch instabil, aber Algorithmus 2.6.6 numerisch stabil ist.

**Bemerkung 2.6.7** Man beachte beim Algorithmus 2.6.5 in Beispiel 2.6.4, dass die numerische Auswertung der Funktion  $\chi_1$  für sich genommen (trotz Auslöschung) nicht numerisch instabil ist! Schlecht konditioniert ist hier das Problem,  $\chi_1(p, \sqrt{p^2 + q})$  zu berechnen.

Umgekehrt kann man daran sehen, dass das Zusammensetzen zweier gut konditionierter Algorithmen sehr wohl einen schlecht konditionierten Algorithmus für das Gesamtproblem ergeben kann. Diese Tatsache steht in unangenehmem, aber unvermeidlichem Kontrast zu dem Wunsch, ein Problem in unabhängig voneinander zu bearbeitende Teilprobleme zu zerlegen.

## 2.6.1 Vorwärtsanalyse

Die sogenannte Vorwärtsanalyse besteht darin, die Rundungsfehler im Laufe einer Rechnung durch Anwendung der Formeln (2.4) und (2.5) (und weiterer entsprechender) für jeden einzelnen Rechenschritt abzuschätzen. Man erhält dadurch eine obere Schranke für den gesamten durch die Rundungen bedingten Fehler. Kennt man außerdem die Konditionszahlen des Problems, so kann man auf diese Weise (theoretisch wenigstens) feststellen, ob der Algorithmus gut oder schlecht konditioniert ist.

**Beispiel 2.6.8** Berechnung des Skalarprodukts  $s = \langle x, y \rangle$  für  $x, y \in \mathbb{R}^n$  mit dem Verfahren

$$s := 0; \text{ for } k := 1 \text{ to } n \text{ do } s := s + x_k \cdot y_k. \quad (2.16)$$

**Satz 2.6.9 (Vorwärtsanalyse des Rundungsfehlers für das Skalarprodukt)**

Sei  $0 < \text{eps} < 1/n$ . Dann gilt

$$|s - \langle x, y \rangle| \leq \frac{n \text{eps}}{1 - n \text{eps}} \langle |x|, |y| \rangle. \quad (2.17)$$

*Beweis.* Wegen (2.4) lassen sich die Zwischenergebnisse  $s_k$  schreiben als

$$\begin{aligned} s_1 &= x_1 y_1 (1 + \delta_1), & |\delta_1| &\leq \text{eps}, \\ s_k &= [s_{k-1} + x_k y_k (1 + \delta_k)] (1 + \varepsilon_k), & |\delta_k|, |\varepsilon_k| &\leq \text{eps}, \quad k = 2, \dots, n, \end{aligned}$$

wobei  $x_i \odot y_i = x_i y_i (1 + \delta_i)$  sowie  $s_{i-1} \oplus (x_i \odot y_i) = (s_{i-1} + (x_i \odot y_i))(1 + \varepsilon_i)$  verwendet wurde. Also gilt für  $s = s_n$

$$s - \langle x, y \rangle = \sum_{k=1}^n x_k y_k \left[ (1 + \delta_k) \prod_{j=k}^n (1 + \varepsilon_j) - 1 \right], \text{ mit } \varepsilon_1 = 0. \quad (2.18)$$

Die Behauptung folgt aus (2.18) mit der Bernoullischen-Ungleichung

$$\begin{aligned} \left| (1 + \delta_k) \prod_{j=k}^n (1 + \varepsilon_j) - 1 \right| &\leq |(1 + \text{eps})^n - 1| \leq \left| \frac{1}{(1 - \text{eps})^n} - 1 \right| \\ &\leq \left| \frac{1}{1 - n \text{eps}} - 1 \right| = \left| \frac{n \text{eps}}{1 - n \text{eps}} \right|. \end{aligned} \quad (2.19)$$

□

**Bemerkung 2.6.10** Aus der Abschätzung (2.17) ergibt sich

$$\left| \frac{s - \langle x, y \rangle}{\langle x, y \rangle} \right| \leq \frac{\langle |x|, |y| \rangle}{|\langle x, y \rangle|} \left| \frac{n \text{eps}}{1 - n \text{eps}} \right|,$$

d.h. bei gegebenem  $n \text{eps}$  wird der relative Fehler groß, falls der Winkel zwischen  $x$  und  $y$  nahe bei 90 Grad ist.

**Bemerkung 2.6.11 (Intervallararithmetik)** In der Intervallararithmetik wird jede Zahl  $x \in \mathbb{M}'$  durch ein (mittels zweier Gleitpunktzahlen) definiertes Intervall dargestellt, welches  $x$  einschließt. Die arithmetischen Operationen werden für solche Intervalle so definiert, dass das resultierende Intervall das exakte Ergebnis der Operation einschließt (und zwar für jede Kombination von Operationen in den Ausgangsintervallen). Falls die Hardware die Intervallararithmetik nicht unterstützt, muss sie softwareseitig simuliert werden (was die Rechnung natürlich stark verlangsamt). In der eben beschriebenen einfachen Form ist sie allerdings für die meisten Probleme der Praxis unbrauchbar, da die Intervalle unrealistisch groß werden, d.h. das mit normaler Gleitpunktarithmetik berechnete Ergebnis ist erheblich genauer.

**Bemerkung 2.6.12 (Hochgenauigkeitsarithmetik)** Die Grundlage der sogenannten Hochgenauigkeitsarithmetik besteht darin, das Skalarprodukt bis auf Maschinengenauigkeit auszurechnen, d.h. zu erreichen, dass der berechnete Wert  $s$  des Skalarprodukts die Eigenschaft

$$s = fl(\langle x, y \rangle)$$

hat. Hierzu wird hardwareseitig ein langer Akkumulator benötigt, d.h. ein spezielles Register, welches einen festen Dezimalpunkt und so viele Stellen hat, dass jede Gleitkommazahl in  $\mathbb{M}$  dargestellt werden kann. Hält man die Zwischensummen  $s$  aus (2.16) in diesem Akkumulator, so kann  $\langle x, y \rangle$  exakt berechnet werden, falls  $x_i, y_i \in \mathbb{M}$ . Der Rundungsfehler entsteht dann beim Abspeichern von  $s$  als Gleitkommazahl.

In Verbindung mit Intervallararithmetik lassen sich Algorithmen konstruieren, die bereits für einen recht breiten Anwendungsbereich sehr gute Intervalle, die das exakte Ergebnis einschließen, und damit genaue und sichere numerische Ergebnisse liefern. Ein wesentliches Problem ist die benötigte Hardwareunterstützung. Die prinzipiellen Schwierigkeiten schlecht konditionierter Probleme werden natürlich nicht beseitigt, insbesondere auch nicht die Auswirkungen des „Standardfehlers“ (Zerlegung eines gut konditionierten Problems in schlecht konditionierte Teilschritte).



## 2.6.2 Rückwärtsanalyse

Während in der Vorwärtsanalyse die Kondition eines Algorithmus durch Vergleich der Fehlerverstärkung des Algorithmus (durch Abschätzung der akkumulierten Rundungsfehler) mit den Konditionszahlen des Problems ermittelt wird, geht man in der Rückwärtsanalyse folgendermaßen vor: Sei  $\varphi : X \rightarrow Y$  ein Problem und  $\tilde{\varphi} : X \rightarrow Y$  ein Algorithmus. Zu gegebenem  $x \in X$  sucht man ein  $\tilde{x} \in X$  mit

$$\tilde{\varphi}(x) = \varphi(\tilde{x}). \quad (2.20)$$

Gilt dann

$$\left| \frac{\tilde{x} - x}{x} \right| \leq C n \text{ eps}, \quad (2.21)$$

mit einer kleinen Konstanten  $C$  ( $n$  repräsentiert in einem geeigneten Sinn die Problemgröße), so ist man zufrieden, da sich das berechnete Ergebnis  $\tilde{\varphi}(x)$  interpretieren lässt als das Ergebnis einer exakten Rechnung auf mit kleinem relativen Fehler behafteten Daten.

**Definition 2.6.13 (Alternative Definition der numerischen Stabilität)** Ein Algorithmus heißt gut konditioniert (oder numerisch stabil), falls er (2.20), (2.21) erfüllt mit einer kleinen Konstanten  $C$ .

**Beispiel 2.6.14 (Rückwärtsanalyse des Skalarprodukts)** Die Rechnung im Beweis von Satz 2.6.9 zeigt, dass für das mit dem Verfahren (2.16) berechnete Skalarprodukt  $s$  gilt

$$s = \langle x, \tilde{y} \rangle, \text{ wobei } \tilde{y}_k = y_k(1 + \delta_k) \prod_{j=k}^n (1 + \varepsilon_j), \quad (2.22)$$

also mit (2.19)

$$\left| \frac{\tilde{y}_k - y_k}{y_k} \right| \leq \frac{n \text{ eps}}{1 - n \text{ eps}}, \text{ falls } \text{eps} \leq 1/n. \quad (2.23)$$

**Bemerkungen 2.6.15** i) Also ist das Verfahren (2.16) gut konditioniert (nochmal: das Problem, das Skalarprodukt zu berechnen, kann schlecht konditioniert sein, wenn die Konditionszahlen des Skalarprodukts groß sind).

ii) Man beachte, dass die Rückwärtsanalyse prinzipiell nichts über die Kondition des Problems aussagt.

iii) Die Rückwärtsanalyse hat sich als recht brauchbares Werkzeug der Rundungsfehleranalyse erwiesen, und zwar vor allem bei numerischen Verfahren der Linearen Algebra.

Die bislang eingeführten Konditionszahlen erlauben die Untersuchung der Fehlerverstärkung für jede einzelne Komponente von  $\varphi$ . Manchmal ist man aber nur an einem Gesamtfehler (im Sinne eines Fehlermaßes, einer Norm) interessiert. Dann sind folgende Begriffe sinnvoll.

**Definition 2.6.16** (a) Es sei  $\tilde{x} \in X$  eine Approximation von  $x \in X$ . Als **Fehler** bezeichnet man  $\Delta x := x - \tilde{x}$  und zu einer gegebenen Norm  $\|\cdot\|_X$  ist

$$\|x - \tilde{x}\|_X$$

der **absolute Fehler** von  $\tilde{x}$ . Es sei  $x \neq 0$ , dann beschreibt

$$\frac{\|x - \tilde{x}\|_X}{\|x\|_X} = \frac{\|\Delta x\|_X}{\|x\|_X}$$

den **relativen Fehler** von  $\tilde{x}$ .

(b) Seien  $\|\cdot\|_X$ ,  $\|\cdot\|_Y$  geeignete Normen auf  $X$  bzw.  $Y$  und

$$\delta_x := \frac{\|\Delta x\|_X}{\|x\|_X}, \quad \delta_y := \frac{\|\Delta y\|_Y}{\|y\|_Y}$$

die relativen Ein- bzw. Ausgabefehler mit  $\Delta x := x - \tilde{x}$ ,  $\Delta y := y - \tilde{y}$ . Dann heißt

$$\kappa_\varphi := \frac{\delta_y}{\delta_x} \quad (2.24)$$

die **(relative) Kondition** des Problems  $\varphi(x)$  beschrieben durch die Funktion  $\varphi : X \rightarrow Y$ . Die Größe

$$\kappa_{\varphi, \text{abs}} := \frac{\|\Delta y\|_Y}{\|\Delta x\|_X} \quad (2.25)$$

heißt **absolute Kondition** von  $y = \varphi(x)$ .

(c) Ein Problem heißt **gut konditioniert**, wenn „kleine“ Schranken für  $\kappa_\varphi$  für  $\delta_x \rightarrow 0$  existieren. Offenbar wäre  $\kappa_\varphi = 1$  (bzw.  $\kappa_\varphi \approx 1$ ) optimal.

**Bemerkung 2.6.17** Relative Fehler in der  $\infty$ -Norm können durch eine Aussage über die Anzahl korrekter Stellen von  $\tilde{x}$  ausgedrückt werden, d.h.

$$\frac{\|x - \tilde{x}\|_\infty}{\|x\|_\infty} \approx 10^{-p},$$

falls die betragsgrößte Komponente von  $\tilde{x}$  näherungsweise  $p$  korrekte signifikante Stellen hat.

# 3 DIREKTE LÖSUNG LINEARER GLEICHUNGSSYSTEME

Lineare Gleichungssysteme (LGS) sind zwar ein verhältnismäßig einfaches Problem, bedürfen aber gerade für große Dimensionen sehr guter numerischer Lösungsverfahren. Hinzu kommt, dass LGS in extrem vielen Anwendungen vorkommen. Oft kann man das Wissen über die „Herkunft“ eines Gleichungssystems zu dessen schnellem Lösen nutzen.

**Beispiel 3.0.1 (Schwingungsgleichung)** Gegeben sei eine elastische Saite der Länge 1, die an beiden Enden fixiert ist. Die Saite wird nun durch eine äußere Kraft  $f$  ausgelenkt (angezupft). Wir wollen die Auslenkung  $u$  der Saite aus ihrer Ruhelage als Funktion von  $x \in [0, 1]$  berechnen. Die

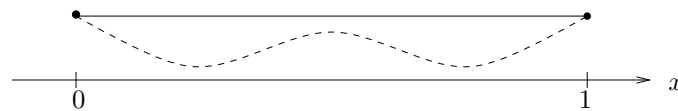


Abb. 3.1: Elastische, an beiden Enden fixierte Saite.

gesuchte Auslenkung  $u : [0, 1] \rightarrow \mathbb{R}$  ist Lösung des folgenden linearen Randwertproblems zweiter Ordnung

$$-u''(x) + \lambda(x)u(x) = f(x), \quad x \in (0, 1), \quad u(0) = u(1) = 0 \quad (3.1)$$

mit gegebenen  $f : (0, 1) \rightarrow \mathbb{R}$  und  $\lambda : (0, 1) \rightarrow \mathbb{R}$ . Genauer zur Modellierung findet man z.B. in [Arendt/Urban].

Wir wollen (3.1) näherungsweise mit Hilfe eines numerischen Verfahrens lösen. Dazu unterteilen wir  $[0, 1]$  in Teilintervalle gleicher Länge. Die Anzahl der Intervalle sei  $N > 1$ ,  $N \in \mathbb{N}$  und  $h = \frac{1}{N}$  die Schrittweite. Dann setzt man  $x_i := ih$ ,  $i = 0, \dots, N$  ( $x_0 = 0, x_N = 1$ ), die  $x_i$  werden

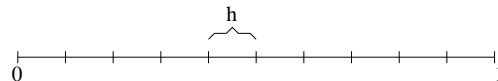


Abb. 3.2: Unterteilung des Intervalls in Teilintervalle der Länge  $h > 0$ .

als **Knoten** bezeichnet. Die **Schrittweite** ist  $h := x_{i+1} - x_i$  für alle  $i$ , man spricht von einem **äquidistanten Gitter**. Wir wollen die Lösung an den Knoten  $x_i$  approximieren und ersetzen hierzu (wie aus der Analysis bekannt) die zweite Ableitung durch den zentralen Differenzenquotienten

$$u''(x_i) \approx \frac{1}{h^2}(u(x_{i-1}) - 2u(x_i) + u(x_{i+1})) =: D_c^2 u(x_i).$$

Es gilt bekanntlich  $\|u'' - D_c^2 u\| = \mathcal{O}(h^2)$ , falls  $u \in C^4[0, 1]$ . Damit erhält man für die Näherung  $u_i \approx u(x_i)$  also folgende Bedingungen ( $\lambda_i = \lambda(x_i)$ ,  $f_i = f(x_i)$ ):

$$\begin{cases} \frac{1}{h^2}(-u_{i-1} + 2u_i - u_{i+1}) + \lambda_i u_i = f_i, & 1 \leq i \leq N-1, \\ u_0 = u_N = 0, \end{cases}$$

also ein lineares Gleichungssystem der Form

$$\underbrace{\begin{bmatrix} (2 + h^2 \lambda_1) & -1 & & 0 \\ -1 & (2 + h^2 \lambda_2) & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & (2 + h^2 \lambda_{n-1}) \end{bmatrix}}_{=:A_h} \underbrace{\begin{bmatrix} u_1 \\ \vdots \\ u_{n-1} \end{bmatrix}}_{=:u_h} = \underbrace{\begin{bmatrix} h^2 f_1 \\ \vdots \\ h^2 f_{n-1} \end{bmatrix}}_{=:f_h},$$

d.h.  $A_h u_h = f_h$ . Für  $N \rightarrow \infty$  ( $h \rightarrow 0$ ) konvergiert die „**diskrete Lösung**“  $u_h$  gegen die Lösung  $u$  von (3.1). Allerdings wächst die Dimension der Matrix  $A_h$  mit kleiner werdendem  $h$ . Bei mehrdimensionalen Problemen führt dies leicht zu sehr großen LGS. Wir nennen diese Matrix auch **Standardmatrix**.

### 3.1 Einführung, Cramersche Regel

Wir beginnen mit dem klassischen Verfahren zur Lösung eines linearen Gleichungssystems (LGS), der Gaußschen<sup>1</sup> Eliminationsmethode.

Zu lösen ist ein System von  $n$  linearen Gleichungen mit  $n$  Unbekannten  $x_1, \dots, x_n \in \mathbb{R}$

$$\begin{array}{cccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \cdots & + & a_{nn}x_n & = & b_n \end{array} \quad (3.2)$$

oder kurz

$$Ax = b, \quad (3.3)$$

wobei  $A \in \mathbb{R}^{n \times n}$  eine reelle  $n \times n$ -Matrix (also insbesondere eine quadratische Matrix) ist und  $b, x \in \mathbb{R}^n$  reelle (Spalten-)Vektoren sind.

Wann ist ein lineares Gleichungssystem überhaupt lösbar? Aus der Linearen Algebra (siehe z.B. [Wille]) kennen wir das folgende Resultat, das die Lösbarkeit mit Hilfe der Determinante der Matrix  $A$  charakterisiert.

**Satz 3.1.1 (Lösbarkeit)** *Sei  $A \in \mathbb{R}^{n \times n}$  mit  $\det A \neq 0$  und  $b \in \mathbb{R}^n$ . Dann existiert genau ein  $x \in \mathbb{R}^n$ , so dass  $Ax = b$ .*

Falls  $\det A \neq 0$ , so lässt sich die Lösung  $x = A^{-1}b$  mit der **Cramerschen Regel** berechnen, d.h.

$$x_i = \frac{1}{\det A} \begin{vmatrix} a_{11} & \cdots & b_1 & \cdots & a_{1n} \\ a_{21} & \cdots & b_2 & \cdots & a_{2n} \\ \vdots & & \vdots & & \vdots \\ a_{n1} & \cdots & b_n & \cdots & a_{nn} \end{vmatrix} = \frac{D_i}{D} \quad (i = 1, \dots, n).$$

Dabei geht die im Zähler stehende Determinante  $D_i$  dadurch aus  $D := \det A$  hervor, dass man die  $i$ -te Spalte der Matrix  $A$  durch den Vektor  $b$  der rechten Seite ersetzt.

<sup>1</sup>**Carl Friedrich Gauß**, 1777 - 1855. Lagrange hatte 1759 die Methode schon vorweggenommen und in China war sie schon vor dem ersten Jahrhundert bekannt. Näheres zu Gauß, Lagrange und weiteren Mathematikern findet man im Internet unter [www-groups.dcs.st-andrews.ac.uk/~history](http://www-groups.dcs.st-andrews.ac.uk/~history).

Man beachte hier die Verbindung von Existenz- und Eindeutigkeitsaussage mit dem Rechenverfahren, was einen „guten“ Algorithmus ausmacht. Dieser braucht dabei nicht unbedingt optimal zu sein!

Wenn wir die Lösung eines linearen Gleichungssystems mit Hilfe der Cramerschen Regel bestimmen wollen, müssen wir die verschiedenen auftretenden Determinanten berechnen.

Die Determinanten von  $n \times n$ -Matrizen lassen sich mittels der **Leibnizschen Darstellung** berechnen, d.h.

$$\det A := \sum_{\pi} (-1)^{j(\pi)} a_{1i_1} a_{2i_2} \cdots a_{ni_n},$$

wobei die Summe über alle möglichen  $n!$  Permutationen  $\pi$  der Zahlen  $1, 2, \dots, n$  zu berechnen ist. Der Wert des Ausdrucks  $(-1)^{j(\pi)}$  ergibt sich aus der Anzahl  $j(\pi)$  der Inversionen der Permutation  $\pi = \begin{pmatrix} 1 & 2 & \dots & n \\ i_1 & i_2 & \dots & i_n \end{pmatrix}$ .

**Bemerkung 3.1.2 (Rechenoperationen)** Im Folgenden werden die Verknüpfungen Multiplikation, Addition, Division und Subtraktion in ihrem Rechenaufwand nicht unterschieden und unter dem Begriff **Gleitkommaoperation** zusammengefasst (1 Gleitkommaoperation  $\simeq 1 \text{ FLOP}^2$ ). Systematische Multiplikationen mit  $\pm 1$  bleiben im Allgemeinen unberücksichtigt. Anderweitige Operationen wie z.B. Wurzelziehen werden gesondert betrachtet.

**Satz 3.1.3 (Aufwand Leibnizsche Darstellung)** Sei  $A \in \mathbb{R}^{n \times n}$ . Der Aufwand zur Berechnung von  $\det A$  mit der Leibnizschen Darstellung, d.h. als Summe über alle Permutationen der Menge  $\{1, \dots, n\}$ , beträgt

$$\text{FLOP}(\det A) = n n! - 1.$$

*Beweis.* Der Aufwand zur Berechnung von  $\det A$  für  $A \in \mathbb{R}^{n \times n}$  in der Leibnizschen Darstellung (unter Vernachlässigung der Multiplikation mit  $(-1)^{j(\pi)}$ ) ergibt sich wie folgt:

$$\begin{aligned} \text{FLOP}(\det A) &= \text{„}(n! - 1) \text{ Additionen} + n! \text{ Produkte mit } n \text{ Faktoren“} \\ &= n! - 1 + n!(n - 1) = n n! - 1. \end{aligned}$$

Damit ist die Aussage des Satzes bewiesen. □

Alternativ lässt sich die Determinante rekursiv mit Hilfe des **Laplaceschen Entwicklungssatzes** berechnen. Dieser lautet für eine quadratische Matrix  $A \in \mathbb{R}^{n \times n}$

$$\det A = \sum_{j=1}^n (-1)^{1+j} a_{1j} \det(A_{1j}),$$

wobei  $A_{1j}$  diejenige Matrix ist, die aus  $A$  durch Streichen der ersten Zeile und der  $j$ -ten Spalte entsteht.

**Satz 3.1.4 (Aufwand Laplacescher Entwicklungssatz)** Sei  $A \in \mathbb{R}^{n \times n}$  ( $n \geq 2$ ). Der Aufwand zur Berechnung von  $\det A$  mit dem Laplaceschen Entwicklungssatz beträgt

$$\text{FLOP}(\det A) = \sum_{k=0}^n \frac{n!}{k!} - 2 < e n! - 2,$$

wobei  $e = \exp(1)$  die Eulersche Zahl bezeichnet.

<sup>2</sup>**FLOP** (Floating point operation) ist nicht zu verwechseln mit **FLOP/s** oder **flops** (floating point operations per second), welches als Maßeinheit für die Geschwindigkeit von Computersystemen verwendet wird.

*Beweis.* Die Ungleichung ist klar. Die Gleichung lässt sich induktiv beweisen:

**Induktionsanfang ( $n = 2$ ):** Die Determinante der Matrix  $A = (a_{ij}) \in \mathbb{R}^{2 \times 2}$  lautet  $\det(A) = a_{11}a_{22} - a_{21}a_{12}$ , d.h. der Aufwand beträgt 2 Multiplikationen und 1 Addition, also

$$\text{FLOP}(\det(\mathbb{R}^{2 \times 2})) = 3 = \frac{2!}{0!} + \frac{2!}{1!} + \frac{2!}{2!} - 2.$$

**Induktionsschritt ( $n \rightarrow n + 1$ ):** Für  $n \geq 2$  gilt

$$\begin{aligned} \text{FLOP}(\det(\mathbb{R}^{(n+1) \times (n+1)})) &= \text{„}n \text{ Additionen} + (n+1) \text{ Multiplikationen} \\ &\quad + (n+1) \text{ Berechnungen von } \det(\mathbb{R}^{n \times n})\text{“} \\ &= n + (n+1) + (n+1) \cdot \text{FLOP}(\det(\mathbb{R}^{n \times n})) \\ &\stackrel{\text{IV}}{=} 2n + 1 + (n+1) \cdot \left( \sum_{k=0}^n \frac{n!}{k!} - 2 \right) \\ &= 2n + 1 + \sum_{k=0}^n \frac{(n+1)!}{k!} - 2(n+1) \\ &= \sum_{k=0}^n \frac{(n+1)!}{k!} - 1 = \sum_{k=0}^{n+1} \frac{(n+1)!}{k!} - 2, \end{aligned}$$

womit der Satz bewiesen ist. □

**Beispiel 3.1.5** Ein einfaches Beispiel soll zeigen, dass man die Determinante überhaupt nur für sehr kleine allgemeine Matrizen der Dimension  $n \ll 23$  mit dem Laplaceschen Entwicklungssatz berechnen kann.

Ein Jahr hat  $365 \cdot 24 \cdot 60 \cdot 60 \approx 3 \cdot 10^7$  Sekunden. Geht man nun von einem schnellen Rechner<sup>3</sup> mit 3000 TFlops aus, so benötigt man zur Berechnung der Determinante einer  $21 \times 21$ -Matrix mit dem Laplaceschen Entwicklungssatz

$$\frac{\text{Anzahl der Operationen}}{\text{Gleitkommaoperationen pro Sekunde}} [s] = \frac{\sum_{k=0}^{21} \frac{21!}{k!} - 2}{3000 \cdot 10^{12}} [s] \approx 46293 [s] \approx 12.9 [h]$$

bzw. für eine  $25 \times 25$ -Matrix

$$\frac{\sum_{k=0}^{25} \frac{25!}{k!} - 2}{3000 \cdot 10^{12}} [s] \approx 1.406 \cdot 10^{10} [s] \approx 445.7 \text{ Jahre}.$$

**Bemerkung 3.1.6** Die Steigerung der Leistungsfähigkeit moderner Computer (Moore'sches Gesetz) ist bei weitem geringer als das Wachstum der Problemgröße mit steigendem  $n$ . Das Hoffen auf immer schnellere Rechner hilft nicht, um immer größere Probleme lösen zu können. Zum Vergleich: Im Jahr 2008 war der Supercomputer JUGENE vom Forschungszentrum Jülich weltweit Platz sechs und in Europa Platz eins mit 180 Flops. Für  $n = 21$  brauchte man damals 214.3h, also immerhin eine Beschleunigung um den Faktor 16 in 4 Jahren. Trotzdem bleibt  $n = 25$  auch in naher Zukunft unerreichbar.

<sup>3</sup>Der weltweit auf Platz vier rangierende und zugleich schnellste europäische Supercomputer SuperMUC am Leibniz Rechenzentrum Garching bei München hat ca. 3100 TFlops (Stand Juni 2012). Ein aktueller PC hat eine Leistung von ca. 300 GFlops. T = Tera =  $10^{12}$ , G = Giga =  $10^9$ .

**Bemerkung 3.1.7** Bei der Cramerschen Regel ist zum Lösen eines linearen Gleichungssystems  $Ax = b$  mit  $A \in \mathbb{R}^{n \times n}$  und  $b \in \mathbb{R}^n$  die Berechnung von  $n + 1$  Determinanten und  $n$  Quotienten notwendig. Der Aufwand zur Lösung eines linearen Gleichungssystems lässt sich somit zusammenfassen, wobei wir auf die Komplexität des Gauß-Verfahrens (siehe Seite 39) erst später eingehen werden.

Cramersche Regel		Gauß-Elimination
Leibniz	Laplace	
$n(n+1)! - 1$	$\sum_{k=0}^n \frac{(n+1)!}{k!} - n - 2$	$\frac{4n^3 + 9n^2 - n}{6}$

Tab. 3.1: Aufwand für die Lösung eines linearen Gleichungssystems mit verschiedenen direkten Verfahren.

**Bemerkung 3.1.8** Für das Lösen eines LGS mit 21 Unbekannten mit der Cramerschen Regel und dem Laplaceschen Entwicklungssatz benötigt man auf einem der schnellsten Rechner mehr als 10 Tage, d.h.  $22 \cdot 12.9 [h] = 283.8 [h] = 11.825 [d]$  (22 verschiedene Determinanten im Zähler und eine im Nenner, vgl. Beispiel 3.1.5). Ein System mit 24 Unbekannten ist mit einem heutigen Supercomputer und der Cramerschen Regel nicht in einem Menschenleben zu lösen.



**Beispiel 3.1.9 (Vergleich Rechenaufwand)** Aufwand zum Lösen eines linearen Gleichungssystems  $Ax = b$  via Cramerscher Regel und Gauß-Verfahren.

Für einige  $n$  sei die Anzahl der notwendigen FLOPs wiedergegeben.

	Cramer/Leibniz	Cramer/Laplace	Gauß
$n$	$= n(n+1)! - 1$	$= \sum_{k=0}^n \frac{(n+1)!}{k!} - n - 2$	$= (4n^3 + 9n^2 - n)/6$
$n = 2$	11	11	11
$n = 3$	71	59	31
$n = 4$	479	319	66
$n = 5$	3599	1949	120
$n = 8$	2903039	986399	436
$n = 10$	399167999	108505099	815

Im Folgenden werden wir nun Verfahren beschreiben, die bereits für  $n \geq 3$  effektiver als die Cramersche Regel sind. Mit Hilfe dieser Verfahren lassen sich auch Determinanten in polynomieller Zeit bestimmen. Daher wird die Cramersche Regel im Allgemeinen nur für  $n = 2, 3$  verwendet. Der Vorteil der Cramerschen Regel ist jedoch die explizite Schreibweise, d.h. sie ist eine Formel für alle Fälle. Man spart sich bei ähnlichem Rechenaufwand (d.h.  $n = 2, 3$ ) aufwendige Fallunterscheidungen (z.B. Pivotwahl) im Programm.

## 3.2 Gestaffelte Systeme

Betrachten wir zuerst besonders einfach zu lösende Spezialfälle. Am einfachsten ist sicherlich der Fall einer diagonalen Matrix  $A$ .

## Diagonalmatrix

Man bezeichnet eine quadratische Matrix  $A = (a_{ij}) \in \mathbb{R}^{n \times n}$  als **Diagonalmatrix**, falls  $a_{ij} = 0$  für  $i \neq j$ ,  $i, j = 1, \dots, n$  gilt. Häufig schreibt man eine Diagonalmatrix  $A$  mit Diagonaleinträgen  $a_{11}, \dots, a_{nn}$  auch einfach  $A = \text{diag}(a_{11}, \dots, a_{nn})$ . Die Anwendung der Inversen einer Diagonalmatrix  $A$  mit Diagonalelementen  $a_{ii} \neq 0$ ,  $i = 1, \dots, n$ , auf einen Vektor  $b$  lässt sich als Pseudocode wie folgt schreiben:

---

### Algorithmus 3.2.1: Lösen von $Ax = b$ mit Diagonalmatrix $A$

---

Sei  $A \in \mathbb{R}^{n \times n}$  eine invertierbare Diagonalmatrix und  $b \in \mathbb{R}^n$

**Input**  $A \in \mathbb{R}^{n \times n}$ ,  $b \in \mathbb{R}^n$

for  $j = 1, \dots, n$

$x_j = b_j / a_{jj}$

end

**Output**  $x = (x_1, \dots, x_n)^T$

---

Eine einfache Matlab Realisierung ist im Folgenden dargestellt.

---

### MATLAB-Beispiel:

Man löse  $Ax = b$  mit

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4 \end{pmatrix}, \quad b = \begin{pmatrix} 5 \\ 3 \\ 2 \end{pmatrix}.$$

```
>> A=diag([1,2,4]);
>> b=[5;3;2];
>> for j=1:3, x(j)=b(j)/A(j,j); end
>> x
x =
    5.0000    1.5000    0.5000
```

---

## Dreiecksmatrix

Der nächstschwierigere Fall ist der einer **Dreiecksmatrix**  $A$ . Man spricht von einer (quadratischen) oberen Dreiecksmatrix  $A = (a_{ij})$ , falls  $a_{ij} = 0$  für  $i > j$ ,  $i, j = 1, \dots, n$  und von einer unteren Dreiecksmatrix  $A = (a_{ij})$ , falls  $a_{ij} = 0$  für  $i < j$ ,  $i, j = 1, \dots, n$ , gilt. Häufig verwenden wir auch  $R$  für eine obere Dreiecksmatrix und  $L$  für eine untere Dreiecksmatrix.

Betrachten wir nun das „gestaffelte“ Gleichungssystem

$$\begin{array}{ccccccc} r_{11}x_1 & + & r_{12}x_2 & + & \cdots & + & r_{1n}x_n & = & z_1 \\ & & r_{22}x_2 & + & \cdots & + & r_{2n}x_n & = & z_2 \\ & & & & \ddots & & \vdots & & \vdots \\ & & & & & & r_{nn}x_n & = & z_n \end{array} \quad (3.4)$$

oder in Matrix-Vektor-Schreibweise

$$Rx = z, \quad (3.5)$$

wobei  $R = (r_{ij}) \in \mathbb{R}^{n \times n}$  gilt. Offenbar erhalten wir  $x$  durch sukzessive Auflösung des „gestaffelten“ Gleichungssystems, beginnend mit der  $n$ -ten Zeile:



$$\begin{aligned}
x_n &:= z_n / r_{nn} && , \text{ falls } r_{nn} \neq 0 \\
x_{n-1} &:= (z_{n-1} - r_{n-1,n} x_n) / r_{n-1,n-1} && , \text{ falls } r_{n-1,n-1} \neq 0 \\
&\vdots && \vdots \\
x_k &:= (z_k - \sum_{i=k+1}^n r_{ki} x_i) / r_{kk} && , \text{ falls } r_{kk} \neq 0 \\
&\vdots && \vdots \\
x_1 &:= (z_1 - r_{12} x_2 - \cdots - r_{1n} x_n) / r_{11} && , \text{ falls } r_{11} \neq 0.
\end{aligned} \tag{3.6}$$

Für die obere Dreiecksmatrix  $R$  gilt, dass

$$\det R = r_{11} \cdot r_{22} \cdot \cdots \cdot r_{nn} \tag{3.7}$$

und daher


$$\det R \neq 0 \iff r_{ii} \neq 0 \quad (i = 1, \dots, n). \tag{3.8}$$

Der angegebene Algorithmus ist also wiederum genau dann anwendbar, wenn  $\det R \neq 0$  (d.h., wenn  $R$  regulär ist), also unter der Bedingung des Existenz- und Eindeutigkeitssatzes (Satz 3.1.1).

Analog zum obigen Vorgehen lässt sich auch ein gestaffeltes lineares Gleichungssystem der Form

$$Lx = z$$

mit einer unteren Dreiecksmatrix  $L \in \mathbb{R}^{n \times n}$  lösen. In diesem Fall beginnt man in der ersten Zeile mit der Berechnung von  $x_1$  und arbeitet sich dann bis zur letzten Zeile zur Bestimmung von  $x_n$  vor.

**Bemerkung 3.2.1 (Vorwärts-, Rückwärtssubstitution)** Das Lösen eines LGS mit oberer Dreiecksmatrix nennt man auch Rückwärtseinsetzen/-substitution (Index läuft „rückwärts“ von  $n$  nach 1), bzw. mit einer unteren Dreiecksmatrix auch Vorwärtseinsetzen/-substitution (Index läuft „vorwärts“ von 1 nach  $n$ ). 

**Satz 3.2.2 (Rechenaufwand  $Ax$  und  $A^{-1}b$  – Dreiecksmatrix)** Sei  $A \in \mathbb{R}^{n \times n}$  eine reguläre obere oder untere Dreiecksmatrix und  $x, b \in \mathbb{R}^n$ . Dann gilt 

$$\text{FLOP}(Ax) = n^2 \quad \text{und} \quad \text{FLOP}(A^{-1}b) = n^2.$$

Bei der Berechnung von  $A^{-1}b$  ist im Gegensatz zu  $Ax$  die Reihenfolge, in der die Zeilen „abgearbeitet“ werden, festgelegt.

**Beweis.** Es genügt den Fall einer regulären oberen Dreiecksmatrix  $R \in \mathbb{R}^{n \times n}$  zu betrachten. Für den Rechenaufwand zur Lösung von  $Rx = b$  (oder die Anwendung von  $R^{-1}$  auf einen Vektor  $b$ ) ergibt sich:

- i) für die  $i$ -te Zeile: je  $(n - i)$  Additionen und Multiplikationen sowie eine Division
- ii) insgesamt für die Zeilen  $n$  bis 1: Betrachten wir zuerst die Summenformel

$$\sum_{i=1}^n (n - i) = \sum_{i=1}^{n-1} i = \frac{n(n-1)}{2} = \frac{n^2 - n}{2}.$$

Mit i) erhält man insgesamt einen Aufwand von  $2 \frac{n^2 - n}{2} + n = n^2$  Operationen.

Der zweite Teil der Aussage bleibt Ihnen als Übungsaufgabe überlassen. □

Man beachte allerdings, dass man allgemein für die Berechnung von  $A^{-1}b$  zunächst die Inverse  $A^{-1}$  berechnen muß. Für ganz allgemeine reguläre Matrizen ist dies sehr aufwändig, weswegen wir unter der Schreibweise  $A^{-1}b$  stets die Lösung des LGS mit gegebener rechten Seite  $b$  verstehen wollen und für diesen Fall schnelle Verfahren konstruieren wollen.

**Aufgabe 3.2.3** Es sei  $A \in \mathbb{R}^{n \times n}$  eine obere (oder untere) Dreiecksmatrix und  $x \in \mathbb{R}^n$ . Man zeige, dass das Matrix-Vektor-Produkt  $Ax$  mit  $n^2$  Operationen berechnet werden kann.

Eine Matlab Realisierung zur Bestimmung von  $R^{-1}b$ , bei der die Matrix  $R$  zeilenweise durchlaufen wird, und ein Anwendungsbeispiel sind im Folgenden dargestellt.

#### MATLAB-Funktion: Rinvb.m

```
1 function x = Rinvb(R,b)
2 % compute solution of R * x = b with upper triangular matrix R
3 n = size(R,1);
4 x = zeros(n,1);
5 for j = n:-1:1
6     for k=j+1:n
7         b(j) = b(j) - R(j,k) * x(k);
8     end
9     x(j) = b(j) / R(j,j);
10 end
```

#### MATLAB-Beispiel:

Man löse  $Rx = b$  mit

$$R = \begin{pmatrix} 4 & 1 & 1 \\ 0 & 3 & 2 \\ 0 & 0 & 1 \end{pmatrix}, b = \begin{pmatrix} 9 \\ 12 \\ 3 \end{pmatrix}.$$

```
>> R = [4,1,1;0,3,2;0,0,1];
>> b = [9;12;3];
>> x = Rinvb(R,b);
>> x'
ans =
      1      2      3
```

Eine alternative Realisierung, bei der die Matrix  $R$  spaltenweise durchlaufen wird, ist mit `Rinvb2.m` gegeben.

**MATLAB-Funktion: Rinvb2.m**

```

1 function x = Rinvb2(R,b)
2 % compute solution of R * x = b with upper triangular matrix R
3 n = size(R,1);
4 x = zeros(n,1);
5 for j = n:-1:1
6     x(j) = b(j) / R(j,j);
7     for k=1:j-1
8         b(k) = b(k) - R(k,j) * x(j);
9     end
10 end

```

Was ist der Unterschied? Machen Sie sich dies klar!

### 3.3 Gaußsche Eliminationsmethode

Gäbe es nun zu einer beliebigen Matrix  $A$  eine Zerlegung

$$A = L \cdot R, \quad (3.9)$$

so könnte ein beliebiges lineares Gleichungssystem  $Ax = b$  durch eine Rückwärts- und Vorwärts-substitution mittels der beiden Schritte

- i) löse  $Lz = b$ ,
- ii) löse  $Rx = z$ ,

gelöst werden. Die folgende Gaußsche Eliminationsmethode liefert gerade eine solche Zerlegung. Betrachten wir ein allgemeines lineares Gleichungssystem  $Ax = b$  ( $A \in \mathbb{R}^{n \times n}$  regulär,  $b \in \mathbb{R}^n$ )

$$\begin{array}{ccccccc}
 a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1 \\
 a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2 \\
 \vdots & & \vdots & & & & \vdots & & \vdots \\
 a_{n1}x_1 & + & a_{n2}x_2 & + & \cdots & + & a_{nn}x_n & = & b_n
 \end{array} \quad (3.10)$$

und versuchen dies in ein gestaffeltes System umzuformen.

Durch die folgenden drei Äquivalenzoperationen

1. Vertauschung von Zeilen,
2. Multiplikation einer Zeile mit einem Skalar  $\neq 0$ ,
3. Addition eines Vielfachen einer Zeile zu einer anderen,

wird die Lösungsmenge des Gleichungssystems (3.10) nicht verändert.

Wir setzen zunächst  $a_{11} \neq 0$  voraus. Um (3.10) nun in ein gestaffeltes System umzuformen, muss die erste Zeile nicht verändert werden. Die restlichen Zeilen werden so modifiziert, dass in einem ersten Schritt die Koeffizienten vor  $x_1$  verschwinden, d.h. die Variable  $x_1$  aus den Gleichungen in

den Zeilen 2 bis  $n$  eliminiert wird.

So entsteht ein System der Art

$$\begin{aligned} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \cdots + a_{1n}^{(1)}x_n &= b_1^{(1)} \\ a_{22}^{(2)}x_2 + \cdots + a_{2n}^{(2)}x_n &= b_2^{(2)} \\ &\vdots \\ a_{n2}^{(2)}x_2 + \cdots + a_{nn}^{(2)}x_n &= b_n^{(2)}, \end{aligned} \quad (3.11)$$

wobei  $a_{11}^{(1)} := a_{11}, \dots, a_{1n}^{(1)} := a_{1n}$  und  $b^{(1)} := b_1$ . Haben wir dies erreicht, so können wir das selbe Verfahren auf die letzten  $(n-1)$  Zeilen anwenden und so rekursiv ein gestaffeltes System erhalten. Mit den Quotienten

$$l_{i1} = a_{i1}/a_{11} \quad (i = 2, 3, \dots, n) \quad (3.12)$$

sind die Elemente in (3.11) gegeben durch

$$\begin{aligned} a_{ik}^{(2)} &= a_{ik} - l_{i1}a_{1k}, \\ b_i^{(2)} &= b_i - l_{i1}b_1. \end{aligned} \quad (3.13)$$

Damit ist der erste Eliminationsschritt unter der Annahme  $a_{11} \neq 0$  ausführbar. Die Matrix, die sich aus diesem ersten Eliminationsschritt ergibt, bezeichnen wir mit  $A^{(2)}$ .

Wenden wir auf diese Restmatrix die Eliminationsvorschrift erneut an, so erhalten wir eine Folge

$$A = A^{(1)} \rightarrow A^{(2)} \rightarrow \cdots \rightarrow A^{(n)} =: R$$

von Matrizen der speziellen Gestalt

$$A^{(k)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ & & \ddots & \\ & & & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ & & & \vdots & & \vdots \\ & & & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix} \quad (3.14)$$

mit einer  $(n-k+1, n-k+1)$ -Restmatrix, auf die wir den Eliminationsschritt

$\begin{aligned} l_{ik} &:= a_{ik}^{(k)} / a_{kk}^{(k)} && \text{für } i = k+1, \dots, n \\ a_{ij}^{(k+1)} &:= a_{ij}^{(k)} - l_{ik}a_{kj}^{(k)} && \text{für } i, j = k+1, \dots, n \\ b_i^{(k+1)} &:= b_i^{(k)} - l_{ik}b_k^{(k)} && \text{für } i = k+1, \dots, n \end{aligned}$	(3.15)
---	--------

ausführen können, wenn das **Pivotelement** (Dreh- und Angelpunkt)  $a_{kk}^{(k)}$  nicht verschwindet. Da jeder Eliminationsschritt eine lineare Operation auf den Zeilen von  $A$  ist, lässt sich der Übergang von  $A^{(k)}$  und  $b^{(k)}$  zu  $A^{(k+1)}$  und  $b^{(k+1)}$  als Multiplikation mit einer Matrix  $L_k \in \mathbb{R}^{n \times n}$  von links darstellen, d.h.

$$A^{(k+1)} = L_k A^{(k)}, \quad b^{(k+1)} = L_k b^{(k)}. \quad (3.16)$$

Die Matrix

$$L_k = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -l_{k+1,k} & 1 & \\ & & \vdots & & \ddots \\ & & -l_{n,k} & & 1 \end{pmatrix} = I - \begin{pmatrix} 0 \\ \vdots \\ 0 \\ l_{k+1,k} \\ \vdots \\ l_{n,k} \end{pmatrix} \cdot e_k^T =: I - \ell_k \cdot e_k^T, \quad (3.17)$$

wobei  $e_k$  der  $k$ -te Einheitsvektor sei, hat die Eigenschaft, dass die Inverse  $L_k^{-1}$  aus  $L_k$  durch einen Vorzeichenwechsel in den Einträgen  $l_{ik}$  ( $i > k$ ) entsteht und für das Produkt der  $L_k^{-1}$  gilt

$$L := L_1^{-1} \cdot \dots \cdot L_{n-1}^{-1} = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ l_{21} & 1 & \ddots & & \vdots \\ l_{31} & l_{32} & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ l_{n1} & \dots & & l_{n,n-1} & 1 \end{pmatrix}. \quad (3.18)$$

Zusammengefasst erhalten wir auf diese Weise das zu  $Ax = b$  äquivalente gestaffelte System  $Rx = z$  mit der oberen Dreiecksmatrix

$$R = L^{-1}A \quad \text{und der rechten Seite} \quad z = L^{-1}b. \quad (3.19)$$

Das **Gaußsche Eliminationsverfahren** schreibt sich dann wie folgt:

Sei  $A \in \mathbb{R}^{n \times n}$  und  $b \in \mathbb{R}^n$ . Berechne nacheinander

- i)  $L \cdot R = A$  (Zerlegung mit  $R$  obere und  $L$  untere Dreiecksmatrix),
- ii)  $Lz = b$  (Vorwärtseinsetzen bzw. -substitution),
- iii)  $Rx = z$  (Rückwärtseinsetzen bzw. -substitution).

**Definition 3.3.1 (unipotente Matrix)** Eine untere oder obere Dreiecksmatrix, deren Diagonalelemente alle gleich eins sind, heißt *unipotent*.

**Definition 3.3.2 (Gaußsche Dreieckszerlegung, LR-Zerlegung)** Die oben genannte Darstellung  $A = L \cdot R$  der Matrix  $A$  als Produkt einer unipotenten unteren Dreiecksmatrix  $L$  und einer oberen Dreiecksmatrix  $R$  heißt **Gaußsche Dreieckszerlegung** oder **LR-Zerlegung** von  $A$ .

**Satz 3.3.3** Existiert die LR-Zerlegung einer regulären Matrix  $A \in \mathbb{R}^{n \times n}$ , so sind  $L$  und  $R$  eindeutig bestimmt.

*Beweis.* Übung. □

**Satz 3.3.4 (Rechenaufwand Gaußsche Dreieckszerlegung, Gaußsche Elimination)**

Sei  $A \in \mathbb{R}^{n \times n}$  regulär. Existiert die LR-Zerlegung, so gilt

$$\text{FLOP}(LR\text{-Zerlegung}) = \frac{4n^3 - 3n^2 - n}{6}.$$

Des Weiteren sei  $b \in \mathbb{R}^n$ . Dann gilt für die Gaußsche Elimination

$$\text{FLOP}(A^{-1}b) = \frac{4n^3 + 9n^2 - n}{6},$$

d.h. die Lösung von  $Ax = b$  kann mit  $(4n^3 + 9n^2 - n)/6$  FLOPs berechnet werden.



**Bemerkung 3.3.5** Die Aufwandsberechnung unterscheidet sich teilweise in der Literatur. Neuerdings werden  $+$  und  $\cdot$  Operationen nicht mehr unterschieden!

*Beweis von Satz 3.3.4.* Für den Rechenaufwand der Gaußschen Dreieckszerlegung gilt Folgendes:

- i) für den  $k$ -ten Eliminationsschritt:  
je  $(n - k)$  Divisionen zur Berechnung der  $\ell_{ik}$  und je  $(n - k)^2$  Multiplikationen und Subtraktionen zur Berechnung der  $a_{ij}^{(k+1)}$ , d.h.  $2(n - k)^2 + (n - k)$  Operationen
- ii) für alle  $n - 1$  Eliminationsschritte erhalten wir durch Umindizierung und mit Summenformeln aus Analysis 1

$$\begin{aligned} & \sum_{k=1}^{n-1} ((n - k) + 2(n - k)^2) = \sum_{k=1}^{n-1} (k + 2k^2) = \frac{n(n - 1)}{2} + \frac{(2n - 1)n(n - 1)}{3} \\ &= \frac{2(2n - 1)(n - 1)n + 3n(n - 1)}{6} = \frac{n(n - 1)(4n - 2 + 3)}{6} \\ &= \frac{n(n - 1)(4n + 1)}{6} = \frac{4n^3 - 3n^2 - n}{6} \end{aligned}$$

- iii) insgesamt benötigt man zum Lösen von  $Ax = b$  mittels Gauß-Elimination, d.h.  $LR$ -Zerlegung und Vorwärts-, Rückwärtssubstitution, d.h. mit dem Lösen von zwei Gleichungssystemen (siehe Satz 3.2.2)

$$\frac{4n^3 - 3n^2 - n}{6} + 2n^2 = \frac{4n^3 + 9n^2 - n}{6} \quad \text{Operationen.}$$

Somit sind die Behauptungen des Satzes bewiesen.  $\square$

Das Speicherschema für die Gauß-Elimination orientiert sich an der Darstellung von  $A^{(k)}$ . In die erzeugten Nullen des Eliminationsschritts können die  $\ell_{ik}$  eingetragen werden. Da die Elemente mit den Werten 0 oder 1 natürlich nicht eigens gespeichert werden müssen, kann die  $LR$ -Zerlegung mit dem Speicherplatz der Matrix  $A$  bewerkstelligt werden:

$$\begin{aligned} A &\rightarrow \begin{pmatrix} a_{11}^{(1)} & \dots & \dots & a_{1n}^{(1)} \\ l_{21} & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ \vdots & \vdots & & \vdots \\ l_{n1} & a_{n2}^{(2)} & & a_{nn}^{(2)} \end{pmatrix} \rightarrow \dots \rightarrow \begin{pmatrix} a_{11}^{(1)} & \dots & \dots & a_{1n}^{(1)} \\ l_{21} & a_{22}^{(2)} & \dots & a_{2n}^{(2)} \\ \vdots & \ddots & & \vdots \\ l_{n1} & \dots & \dots & l_{n,n-1} & a_{nn}^{(n)} \end{pmatrix} \\ \Rightarrow L &= \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ l_{21} & \ddots & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & 0 \\ l_{n1} & \dots & \dots & l_{n,n-1} & 1 \end{pmatrix}, \quad R = \begin{pmatrix} a_{11}^{(1)} & \dots & \dots & \dots & a_{1n}^{(1)} \\ 0 & \ddots & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ 0 & \dots & \dots & 0 & a_{nn}^{(n)} \end{pmatrix}. \end{aligned}$$

## 3.4 Pivot-Strategien

Bisher haben wir uns noch nicht mit der Frage beschäftigt, ob eine  $LR$ -Zerlegung immer existiert und ob eine Vertauschung von Zeilen bei  $Ax = b$  zu einem anderen Ergebnis führt, wenn der Algorithmus mit endlicher Rechengenauigkeit durchgeführt wird. (Bei exakter Arithmetik sind beide

Ergebnisse gleich, bzw. Vertauschung von Zeilen in  $A$  führt zur gleichen Lösung). Betrachten wir hierzu zwei Beispiele:

**Beispiel 3.4.1** i) **Permutation liefert  $LR$ -Zerlegung:** Sei  $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ . Wie wir leicht sehen, gilt für die Eigenwerte  $\lambda_{1,2} = \pm 1$  und somit  $\det A \neq 0$ . Falls eine Zerlegung existieren würde, gäbe es  $a, b, c, d \in \mathbb{R}$  mit

$$A = \begin{pmatrix} 1 & 0 \\ a & 1 \end{pmatrix} \begin{pmatrix} b & c \\ 0 & d \end{pmatrix} = \begin{pmatrix} b & c \\ ab & ac + d \end{pmatrix}.$$

Ein einfacher Vergleich der Koeffizienten zeigt jedoch (zuerst nach  $b$  und dann nach  $a$ ), dass keine  $LR$ -Zerlegung existiert, obwohl nach Vertauschen der Zeilen trivialerweise eine  $LR$ -Zerlegung sofort gegeben ist. Die Frage, die sich nun stellt, lautet: Kann man für jede reguläre Matrix  $A$  durch Vertauschen ihrer Zeilen eine Matrix finden, für die dann eine  $LR$ -Zerlegung existiert?

ii) **Permutation liefert höhere Genauigkeit:** Berechnen wir die Lösung des folgenden linearen Gleichungssystems ( $\epsilon > 0$ )

$$\epsilon x_1 + x_2 = 1 \quad (3.20)$$

$$x_1 + x_2 = 2. \quad (3.21)$$

Bei exakter Arithmetik erhalten wir

$$\frac{1}{\epsilon-1} \begin{pmatrix} 1 & -1 \\ -1 & \epsilon \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \frac{1}{1-\epsilon} \begin{pmatrix} 1 \\ 1-2\epsilon \end{pmatrix}$$

d.h.  $x_1 = \frac{1}{1-\epsilon}$ ,  $x_2 = \frac{1-2\epsilon}{1-\epsilon}$ . Für  $2\epsilon < \text{Rechengenauigkeit}$  (d.h.  $1 \pm 2\epsilon$  und 1 werden im Rechner durch dieselbe Zahl dargestellt) gilt nun

$$x_1 = 1, \quad x_2 = 1.$$

Für die Gauß-Elimination erhalten wir, wenn wir den Koeffizienten vor  $x_1$  in (3.21) eliminieren, d.h. das  $1/\epsilon$ -fache von (3.20) subtrahieren

$$\begin{aligned} \epsilon x_1 + x_2 &= 1 \\ (1 - \frac{1}{\epsilon}) x_2 &= 2 - \frac{1}{\epsilon}. \end{aligned}$$

Somit ergibt sich für  $2\epsilon < \text{Rechengenauigkeit}$  aus der letzten Gleichung  $x_2 = 1$  und damit aus der ersten Gleichung  $x_1 = 0$ . Vertauschen wir jedoch 1. und 2. Zeile (bzw. eliminieren den Koeffizienten vor  $x_1$  in (3.20)) so erhalten wir

$$\begin{aligned} x_1 + x_2 &= 2 \\ (1 - \epsilon) x_2 &= 1 - 2\epsilon. \end{aligned}$$

Dies liefert  $x_1 = x_2 = 1$  bei  $2\epsilon < \text{Rechengenauigkeit}$ .

**MATLAB-Beispiel:**

Dieses scheinbar theoretische Ergebnis aus Beispiel 3.4.1 können wir direkt in Matlab umsetzen. Da die Recheneinheiten heutiger Rechner meist 2 Stellen genauer rechnen als unsere bisherigen theoretischen Untersuchungen vermuten lassen, muss hier  $\epsilon \leq \text{Maschinengenauigkeit}/8$  gelten.

```
>> e = eps/8; % eps/2 gen\ugt nicht!
>> x(2) = (2 - 1/e) / (1 - 1/e);
>> x(1) = (1 - x(2)) / e
x =
    0    1
>> x(2) = (1 - 2*e) / (1 - e);
>> x(1) = 2 - x(2)
x =
    1    1
```

Das Vertauschen kann folglich nötig sein. Zum einen, damit eine  $LR$ -Zerlegung überhaupt existiert, zum anderen, um den „Rechenfehler“ bedingt durch Rundungsfehler möglichst klein zu halten. Daraus leiten sich besondere Strategien zur Wahl des Pivotelements ab.

Betrachten wir nun die Gauß-Elimination mit **Spaltenpivotstrategie**, welche theoretisch nur dann nicht zielführend sein kann, falls die Matrix  $A$  singular ist.

**Algorithmus 3.4.1: Gauß-Elimination mit Spaltenpivotstrategie**

- a) Wähle im Eliminationsschritt  $A^{(k)} \rightarrow A^{(k+1)}$  ein  $p \in \{k, \dots, n\}$ , so dass

$$|a_{pk}^{(k)}| \geq |a_{jk}^{(k)}| \quad \text{für } j = k, \dots, n.$$

Die Zeile  $p$  soll die Pivotzeile werden.

- b) Vertausche die Zeilen  $p$  und  $k$

$$A^{(k)} \rightarrow \tilde{A}^{(k)} \quad \text{mit} \quad \tilde{a}_{ij}^{(k)} = \begin{cases} a_{kj}^{(k)} & \text{falls } i = p \\ a_{pj}^{(k)} & \text{falls } i = k \\ a_{ij}^{(k)} & \text{sonst.} \end{cases}$$

Nun gilt

$$|l_{ik}| = \left| \frac{\tilde{a}_{ik}^{(k)}}{\tilde{a}_{kk}^{(k)}} \right| = \left| \frac{\tilde{a}_{ik}^{(k)}}{a_{pk}^{(k)}} \right| \leq 1.$$

- c) Führe den nächsten Eliminationsschritt angewandt auf  $\tilde{A}^{(k)}$  aus,

$$\tilde{A}^{(k)} \rightarrow A^{(k+1)}.$$



**Bemerkung 3.4.2** Anstelle der Spaltenpivotstrategie mit Zeilentausch kann man auch eine Zeilenpivotstrategie mit Spaltentausch durchführen. Beide Strategien benötigen im schlimmsten Fall  $\mathcal{O}(n^2)$  zusätzliche Operationen. In der Praxis werden im Gegensatz zum oben genannten Algorithmus die Zeile bzw. Spalte nicht umgespeichert, sondern besondere Indexvektoren verwendet (siehe Matlabfunktion `mylu.m`). Die Kombination von Spalten- und Zeilenpivotstrategie führt zur vollständigen Pivotsuche, bei der die gesamte Restmatrix nach dem betragsgrößten Eintrag durchsucht wird. Wegen des Aufwands  $\mathcal{O}(n^3)$  wird dies jedoch so gut wie nie angewandt.



### MATLAB-Funktion: mylu.m

```

1 function [L,R,P] = mylu(A)
2 n = size(A,1); % get leading dimension of A
3 p = 1 : n; % pivot element vector
4 for k = 1 : n-1 % consider k-th column
5     [m,mptr] = max(abs(A(p(k:end),k))); % find pivot element
6     tmp = p(k); % interchange in vector p
7     p(k) = p(k-1+mptr);
8     p(k-1+mptr) = tmp;
9
10    for j = k+1 : n % modify entries in
11        A(p(j),k) = A(p(j),k)/A(p(k),k); % compute l_jk, store in A
12        for i = k+1 : n % (n-k-1)*(n-k-1) submatrix
13            A(p(j),i) = A(p(j),i) ...
14                - A(p(j),k)*A(p(k),i);
15        end
16    end
17 end
18 L = tril(A(p,:),-1)+eye(n); % these lines could be
19 R = triu(A(p,:)); % neglected, all information
20 P = eye(n); % already in A and p
21 P(:,p) = P;
```

**Satz 3.4.3** Es sei  $A \in \mathbb{R}^{n \times n}$  regulär. Dann existiert vor dem  $k$ -ten Eliminationsschritt des Gauß-Algorithmus stets eine Zeilen-/Spaltenpermutation derart, dass das  $k$ -te Diagonalelement von Null verschieden ist. Bei Zeilenpermutation, d.h. Spaltenpivotisierung, sind alle Einträge von  $L$  vom Betrag kleiner oder gleich eins.

*Beweis.* Nach Voraussetzung gilt  $\det A \neq 0$ . Angenommen, es sei  $a_{11} = 0$ . Dann existiert eine Zeilenvertauschung, so dass das erste Pivotelement  $a_{11}^{(1)}$  von Null verschieden und das betragsgrößte Element in der Spalte ist, d.h.

$$0 \neq |a_{11}^{(1)}| \geq |a_{i1}^{(1)}| \quad \text{für } i = 1, \dots, n,$$

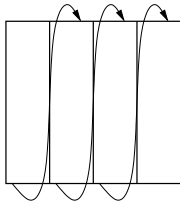
denn andernfalls wäre die Determinante von  $A$  in Widerspruch zur Voraussetzung gleich Null. Die Zeilenvertauschung hat dabei nur einen Vorzeichenwechsel in der Determinante zur Folge. Die Überlegungen für den ersten Eliminationsschritt übertragen sich sinngemäß auf die folgenden, reduzierten Systeme, bzw. ihre zugehörigen Determinanten. Diese Schlussfolgerungen gelten analog bei Zeilenpivotisierung.  $\square$

Eine unmittelbare Folge aus der Beweisführung des letzten Satzes ist das folgende Lemma.

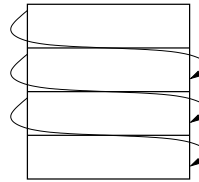
**Lemma 3.4.4** *Erfolgen im Verlauf des Gauß-Algorithmus total  $m$  Zeilenvertauschungen (Spaltenvertauschungen), so ist die Determinante von  $A$  gegeben durch*

$$\det A = (-1)^m \prod_{k=1}^n r_{kk}.$$

**Bemerkung 3.4.5** *Numerisch sind Spalten- und Zeilenpivotisierungen äquivalent. Die Auswahl hängt von der Speichermethode der Matrix  $A$  ab, d.h. man favorisiert die Spaltenpivotisierung, wenn die Einträge der Matrix spaltenweise im Speicher abgelegt sind, z.B. bei den Programmiersprachen Matlab, Fortran und die Zeilenpivotisierung u.a. bei C, denn hier sind die Matrizen als Folge von Zeilenvektoren abgelegt.*



Spaltenpivotisierung, bzw.



Zeilenpivotisierung.

Die genannten Pivotisierungen lassen sich formal durch die Multiplikation mit einer geeigneten Permutationsmatrix  $P$  beschreiben. Diese ist eine quadratische Matrix, welche in jeder Zeile und in jeder Spalte genau eine Eins und sonst Nullen enthält. Die Determinante ist  $\det P = \pm 1$  und die Inverse ist durch  $P^{-1} = P^T$  gegeben. Die Zeilenvertauschungen bei Spaltenpivotisierung entsprechen dabei der Multiplikation  $P \cdot A$ , analog lassen sich Spaltenpermutationen bei Zeilenpivotisierung durch  $A \cdot P$  ausdrücken. Aus dem letzten Satz ergibt sich folgendes Resultat.

**Satz 3.4.6 (Existenz einer Zerlegung  $PA = LR$ )** *Zu jeder regulären Matrix  $A$  existiert eine Permutationsmatrix  $P$ , so dass  $P \cdot A$  in ein Produkt  $L \cdot R$  zerlegbar ist, d.h.*

$$P \cdot A = L \cdot R.$$

Ist nun immer eine Pivotisierung notwendig? Wir nennen im Folgenden ein Kriterium, welches leicht zu überprüfen ist.

**Definition 3.4.7 (Diagonaldominanz)** *Eine Matrix heißt diagonaldominant, falls gilt*

$$|a_{ii}| \geq \sum_{\substack{k=1 \\ k \neq i}}^n |a_{ik}| \quad \forall i = 1, \dots, n.$$

*Sie heißt strikt diagonaldominant, falls für alle  $i$  „ $>$ “ anstatt „ $\geq$ “ gilt.*

**Satz 3.4.8** *Es sei  $A \in \mathbb{R}^{n \times n}$  eine diagonaldominante und reguläre Matrix. Dann existiert eine Zerlegung*

$$A = L \cdot R.$$

**Beweis.** Einen Beweis dieser Aussage (in leicht abgewandelter Form) findet man z.B. bei [Schwarz]. Nach Voraussetzung ist entweder  $|a_{11}| \geq \sum_{k=2}^n |a_{1k}| > 0$  oder es gilt  $\sum_{k=2}^n |a_{1k}| = 0$  und  $|a_{11}| > 0$ ; dies folgt aus der Regularität von  $A$ . Somit ist  $a_{11} \neq 0$  ein zulässiges Pivotelement für den ersten Eliminationsschritt. Man muss nun zeigen, dass sich die Eigenschaft der diagonalen Dominanz auf das reduzierte Gleichungssystem überträgt.  $\square$

Zum Schluss dieses Abschnittes betrachten wir noch folgendes Beispiel zur konkreten Berechnung der  $LR$ -Zerlegung einer Matrix  $A$ .

**Beispiel 3.4.9 (Berechnung von  $PA = LR$ )** Sei

$$A = \begin{pmatrix} 0 & 2 & -1 & -2 \\ 2 & -2 & 4 & -1 \\ 1 & 1 & 1 & 1 \\ -2 & 1 & -2 & 1 \end{pmatrix}.$$

Diese Matrix ist offensichtlich nicht diagonaldominant, also wenden wir die Gauß-Elimination mit Spaltenpivotisierung an. Hierzu sei  $P(i, j)$  diejenige Permutationsmatrix, die aus der Einheitsmatrix durch Vertauschen der  $i$ -ten und  $j$ -ten Zeile entsteht. Das Vertauschen der ersten mit der zweiten Zeile von  $A$  entspricht der Multiplikation der Matrix  $P_1 := P(1, 2)$  mit  $A$  von links, also

$$\tilde{A}^{(1)} = P_1 A = \begin{pmatrix} 2 & -2 & 4 & -1 \\ 0 & 2 & -1 & -2 \\ 1 & 1 & 1 & 1 \\ -2 & 1 & -2 & 1 \end{pmatrix}.$$

Der erste Schritt der  $LR$ -Zerlegung liefert

$$L_1 = \begin{pmatrix} 1 & & & \\ 0 & 1 & & \\ -\frac{1}{2} & & 1 & \\ 1 & & & 1 \end{pmatrix}, \quad L_1 P_1 A = \begin{pmatrix} 2 & -2 & 4 & -1 \\ 0 & 2 & -1 & -2 \\ 0 & 2 & -1 & \frac{3}{2} \\ 0 & -1 & 2 & 0 \end{pmatrix}.$$

Bei der nächsten Spaltenpivotisierung sind keine Zeilenvertauschungen notwendig, also  $P_2 = I$ . Es folgt

$$L_2 = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & -1 & 1 & \\ & \frac{1}{2} & & 1 \end{pmatrix}, \quad L_2 P_2 L_1 P_1 A = \begin{pmatrix} 2 & -2 & 4 & -1 \\ 0 & 2 & -1 & -2 \\ 0 & 0 & 0 & \frac{7}{2} \\ 0 & 0 & \frac{3}{2} & -1 \end{pmatrix}.$$

Um nun auf eine obere Dreiecksgestalt zu kommen, müssen lediglich noch die dritte und vierte Zeile vertauscht werden, also formal  $P_3 = P(3, 4)$ ,  $L_3 = I$  und damit

$$L_3 P_3 L_2 P_2 L_1 P_1 A = \begin{pmatrix} 2 & -2 & 4 & -1 \\ 0 & 2 & -1 & -2 \\ 0 & 0 & \frac{3}{2} & -1 \\ 0 & 0 & 0 & \frac{7}{2} \end{pmatrix} =: R.$$

Aber wie sehen nun  $P$  und  $L$  aus, sodass

$$P \cdot A = L \cdot R \quad ?$$

Wir betrachten hierzu im allgemeinen Fall mit  $A \in \mathbb{R}^{n \times n}$  invertierbar für  $k = 1, \dots, n-1$  die Matrizen

$$\tilde{L}_k := P_n \cdot \dots \cdot P_{k+1} L_k P_{k+1} \cdot \dots \cdot P_n,$$

wobei  $P_n := I$  gesetzt wird und  $P_1, \dots, P_{n-1}$  analog zu Beispiel 3.4.9 gewählt werden. Dann gilt

$$\begin{aligned} \tilde{L}_k &= P_n \cdot \dots \cdot P_{k+1} L_k P_{k+1} \cdot \dots \cdot P_n \\ &\stackrel{(3.17)}{=} P_n \cdot \dots \cdot P_{k+1} (I - \ell_k e_k^T) P_{k+1} \cdot \dots \cdot P_n \\ &= I - \underbrace{P_n \cdot \dots \cdot P_{k+1} \ell_k}_{=: \tilde{\ell}_k} \underbrace{e_k^T P_{k+1} \cdot \dots \cdot P_n}_{=: e_k^T} \\ &= I - \tilde{\ell}_k e_k^T. \end{aligned}$$

Da die Multiplikation mit den Matrizen  $P_{k+1}, \dots, P_n$  von links bzw. rechts nur Zeilen- bzw. Spaltenvertauschungen innerhalb der Zeilen bzw. Spalten  $k+1, \dots, n$  bewirkt, besitzt die Matrix  $\tilde{L}_k$  dieselbe Struktur wie  $L_k$ . Aus der Definition der  $\tilde{L}_k$  folgt nun

$$\tilde{L}_{n-1} \cdot \dots \cdot \tilde{L}_1 P_{n-1} \cdot \dots \cdot P_1 = L_{n-1} P_{n-1} \cdot \dots \cdot L_2 P_2 L_1 P_1.$$

Nach Konstruktion der  $L_k$  und  $P_k$  gilt schließlich

$$\tilde{L}_{n-1} \cdot \dots \cdot \tilde{L}_1 P_{n-1} \cdot \dots \cdot P_1 A = R.$$

Somit ergibt sich mit

$$L := \left( \tilde{L}_{n-1} \cdot \dots \cdot \tilde{L}_1 \right)^{-1} = \tilde{L}_1^{-1} \cdot \dots \cdot \tilde{L}_{n-1}^{-1}$$

und

$$P := P_{n-1} \cdot \dots \cdot P_1$$

eine Zerlegung  $PA = LR$ .

**Bemerkung 3.4.10** Man beachte, dass mit den obigen Überlegungen ein alternativer, konstruktiver Beweis des Satzes 3.4.6 vorliegt.

Führen wir nun Beispiel 3.4.9 weiter fort:

**Beispiel 3.4.11** Zu der Matrix  $A$  aus Beispiel 3.4.9 sollen die Matrizen  $L$  und  $P$  der Zerlegung  $PA = LR$  bestimmt werden. Wir erhalten

$$\begin{aligned} P &= P_3 P_2 P_1 = P(3,4) I P(1,2) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \\ \tilde{L}_1 &= P_3 P_2 L_1 P_2 P_3 = P(3,4) \begin{pmatrix} 1 & & & \\ 0 & 1 & & \\ -\frac{1}{2} & & 1 & \\ 1 & & & 1 \end{pmatrix} P(3,4) = \begin{pmatrix} 1 & & & \\ 0 & 1 & & \\ 1 & & 1 & \\ -\frac{1}{2} & & & 1 \end{pmatrix}, \\ \tilde{L}_2 &= P_3 L_2 P_3 = P(3,4) \begin{pmatrix} 1 & & & \\ & 1 & & \\ -1 & & 1 & \\ & \frac{1}{2} & & 1 \end{pmatrix} P(3,4) = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & \frac{1}{2} & 1 & \\ -1 & & & 1 \end{pmatrix}, \\ \tilde{L}_3 &= P_4 L_3 P_4 = I, \\ L &= \tilde{L}_1^{-1} \tilde{L}_2^{-1} \tilde{L}_3^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & -\frac{1}{2} & 1 & 0 \\ \frac{1}{2} & 1 & 0 & 1 \end{pmatrix} \end{aligned}$$

und somit

$$PA = \begin{pmatrix} 2 & -2 & 4 & -1 \\ 0 & 2 & -1 & -2 \\ -2 & 1 & -2 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & -\frac{1}{2} & 1 & 0 \\ \frac{1}{2} & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & -2 & 4 & -1 \\ 0 & 2 & -1 & -2 \\ 0 & 0 & \frac{3}{2} & -1 \\ 0 & 0 & 0 & \frac{7}{2} \end{pmatrix} = LR.$$

### 3.5 Nachiteration

Die oben diskutierten Pivotstrategien schließen offenbar nicht aus, dass die so berechnete Lösung  $x$  immer noch „ziemlich ungenau“ ist. Wie kann man nun  $x$  ohne großen Aufwand verbessern?

Häufig lässt sich dies durch eine **Nachiteration** mit Hilfe einer expliziten Auswertung des **Residuums**  $r := b - A\tilde{x}$  erreichen. Dabei bezeichne  $\tilde{x}$  die durch ein numerisches Verfahren berechnete Näherung an  $x$ . Ausgehend von  $\tilde{x}$  soll die exakte Lösung mittels des Korrekturansatzes

$$x = \tilde{x} + s$$

ermittelt werden. Der Korrekturvektor ist so zu bestimmen, dass die Gleichungen erfüllt sind, d.h.

$$Ax - b = A(\tilde{x} + s) - b = A\tilde{x} + As - b = 0.$$

Der Korrekturvektor  $s$  ergibt sich somit als Lösung von

$$As = b - A\tilde{x} = r. \quad (3.22)$$

Bei der numerischen Lösung dieser Korrekturgleichung erhalten wir im Allgemeinen eine wiederum fehlerhafte Korrektur  $\tilde{s} \neq s$ . Trotzdem erwarten wir, dass die Näherungslösung

$$\tilde{x} + \tilde{s}$$

„besser“ ist als  $\tilde{x}$ .

Das Gleichungssystem (3.22) unterscheidet sich nur in der rechten Seite von dem ursprünglichen Problem  $Ax = b$ , sodass die Berechnung des Korrekturvektors  $s$  relativ wenig Aufwand erfordert, da beispielsweise bei Anwendung einer Nachiteration in Kombination mit der Gauß-Elimination schon die  $LR$ -Zerlegung der Koeffizientenmatrix  $A$  bekannt ist.

**Bemerkung 3.5.1** *In der Praxis genügen bei Spalten- oder Zeilenpivotisierung meist wenige Nachiterationen, um eine Lösung auf eine dem Problem angepasste Genauigkeit zu erhalten.*

### 3.6 Cholesky-Verfahren

Wir wollen nun die Gauß-Elimination auf die eingeschränkte Klasse von Gleichungssystemen mit symmetrisch positiv definiten Matrizen anwenden. Es wird sich herausstellen, dass die Dreieckszerlegung in diesem Fall stark vereinfacht werden kann und dass dieser vereinfachte Algorithmus für große Matrizen nur den halben Aufwand an Operationen benötigt. Wir erinnern hier nochmals an die folgende Definition:

**Definition 3.6.1** Eine symmetrische Matrix  $A \in \mathbb{R}^{n \times n}$  heißt **positiv definit**, wenn  $x^T Ax > 0$  für alle  $x \in \mathbb{R}^n$  mit  $x \neq 0$  ist.

**Bemerkung 3.6.2** Die Bedingung in Definition 3.6.1 macht auch für nicht-symmetrische Matrizen Sinn. Deshalb fordert man oft zusätzlich zur positiven Definitheit die Symmetrie und nennt solche Matrizen **symmetrisch positiv definit**, abgekürzt **s.p.d.**

Positiv definite Matrizen haben folgende Eigenschaften.

**Satz 3.6.3 (Eigenschaften positiv definiter Matrizen)** Für jede s.p.d. Matrix  $A \in \mathbb{R}^{n \times n}$  gilt:

- i)  $A$  ist invertierbar,

- ii)  $a_{ii} > 0$  für  $i = 1, \dots, n$ ,
- iii)  $\max_{i,j=1,\dots,n} |a_{ij}| = \max_{i=1,\dots,n} a_{ii}$ ,
- iv) Bei der Gauß-Elimination ohne Pivotsuche ist jede Restmatrix wiederum symmetrisch und positiv definit.

*Beweis.* Wäre  $A$  nicht invertierbar, gäbe es einen Eigenwert  $\lambda = 0$ , da  $\det A = 0$  nach Annahme. Dies steht aber im Widerspruch zur Voraussetzung, dass  $A$  positiv definit ist, da es ansonsten einen Eigenvektor  $x \neq 0$  zu  $\lambda = 0$  mit  $Ax = \lambda x$  gäbe und somit dann auch  $x^T Ax = 0$  gälte. Somit ist i) bewiesen. Setzt man für  $x$  einen kanonischen Einheitsvektor  $e_i$  ein, so folgt gerade  $a_{ii} = e_i^T A e_i > 0$  und daher gilt die Aussage ii). Behauptung iii) sei als Hausübung überlassen. Um die verbleibende Behauptung iv) zu beweisen, schreiben wir  $A = A^{(1)}$  in der Form

$$A^{(1)} = \left( \begin{array}{c|c} a_{11} & z^T \\ \hline z & B^{(1)} \end{array} \right),$$

wobei  $z = (a_{12}, \dots, a_{1n})^T$  sei. Nach einem Eliminationsschritt ergibt sich

$$A^{(2)} = L_1 A^{(1)} = \left( \begin{array}{c|c} a_{11} & z^T \\ \hline 0 & B^{(2)} \end{array} \right) \quad \text{mit} \quad L_1 = \left( \begin{array}{cccc} 1 & & & \\ -l_{21} & 1 & & \\ \vdots & & \ddots & \\ -l_{n1} & & & 1 \end{array} \right).$$

Multipliziert man nun  $A^{(2)}$  von rechts mit  $L_1^T$ , so wird auch  $z^T$  in der ersten Zeile eliminiert und die Teilmatrix  $B^{(2)}$  bleibt unverändert, d.h.

$$L_1 A^{(1)} L_1^T = \left( \begin{array}{c|ccc} a_{11} & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & B^{(2)} & \\ 0 & & & \end{array} \right).$$

Es lässt sich einfach nachweisen, dass die Matrix  $B^{(2)}$  wieder symmetrisch ist. Können wir nun noch zeigen, dass  $B^{(2)}$  auch wiederum positiv definit ist, so können wir unsere Argumentation sukzessive fortsetzen. Es sei  $y \in \mathbb{R}^{n-1}$  mit  $y \neq 0$ . Da  $L_1$  regulär ist, gilt  $x^T := (0 \mid y^T) L_1 \neq 0$ . Nach Voraussetzung ist  $A$  positiv definit, also gilt

$$0 < x^T A x = (0 \mid y^T) L_1 A L_1^T \begin{pmatrix} 0 \\ y \end{pmatrix} = (0 \mid y^T) \left( \begin{array}{c|c} a_{11} & 0 \\ \hline 0 & B^{(2)} \end{array} \right) \begin{pmatrix} 0 \\ y \end{pmatrix} = y^T B^{(2)} y.$$

Somit haben wir gezeigt, dass auch  $B^{(2)}$  wieder positiv definit ist. □

**Aufgabe 3.6.4** Man beweise Aussage iii) in Satz 3.6.3.

Mit Hilfe des letzten Satzes über die  $LR$ -Zerlegung können wir jetzt die **rationale** Cholesky-Zerlegung für symmetrische, positiv definite Matrizen herleiten.

**Satz 3.6.5 (rationale Cholesky-Zerlegung)** Für jede symmetrisch positiv definite Matrix  $A$  existiert eine eindeutig bestimmte Zerlegung der Form

$$A = \bar{L} D \bar{L}^T,$$

wobei  $\bar{L}$  eine unipotente untere Dreiecksmatrix und  $D$  eine positive Diagonalmatrix ist.

*Beweis.* Wir setzen die Konstruktion im Beweis von Satz 3.6.3(iv) für  $k = 2, \dots, n-1$  fort und erhalten so unmittelbar  $\bar{L}$  als Produkt der  $L_1^{-1}, \dots, L_{n-1}^{-1}$  und  $D$  als Diagonalmatrix der Pivotelemente.  $\square$

**Bemerkung 3.6.6** Man beachte die Eigenschaften der  $L_k$  bei der tatsächlichen Berechnung von  $L$ , d.h. das Produkt  $L_1^{-1} \cdot \dots \cdot L_{n-1}^{-1}$  ist nicht wirklich durch Multiplikation auszurechnen (vgl. (3.18)).

**Definition und Bemerkung 3.6.7 (Cholesky-Zerlegung)** Da alle Einträge der Diagonalmatrix  $D$  positiv sind, existiert  $D^{1/2} = \text{diag}(\pm\sqrt{d_i})$  und daher die Cholesky-Zerlegung

$$A = L L^T,$$

wobei  $L$  die untere Dreiecksmatrix  $L := \bar{L} D^{1/2}$  mit  $\bar{L}$  und  $D$  aus der rationalen Cholesky-Zerlegung ist.

**Bemerkung 3.6.8** Die Cholesky-Zerlegung ist nicht eindeutig, da  $D^{1/2}$  nur bis auf Vorzeichen der Elemente in der Diagonalen eindeutig bestimmt ist und somit auch  $\bar{L} = L D^{1/2}$  nicht eindeutig ist.

Zur Herleitung des Algorithmus zur Berechnung einer Cholesky-Zerlegung betrachten wir folgende Gleichung

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & & & \vdots \\ \vdots & & & \vdots \\ a_{n1} & \cdots & \cdots & a_{nn} \end{pmatrix} \stackrel{!}{=} \begin{pmatrix} l_{11} & & & \\ l_{21} & l_{22} & & \\ \vdots & & \ddots & \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{pmatrix} \begin{pmatrix} l_{11} & l_{21} & \cdots & l_{n1} \\ & l_{22} & & l_{n2} \\ & & \ddots & \vdots \\ & & & l_{nn} \end{pmatrix} =$$

$$\begin{pmatrix} l_{11}^2 & l_{11}l_{21} & l_{11}l_{31} & \cdots & l_{11}l_{n1} \\ l_{11}l_{21} & l_{21}^2 + l_{22}^2 & l_{21}l_{31} + l_{22}l_{32} & \cdots & l_{21}l_{n1} + l_{22}l_{n2} \\ l_{11}l_{31} & l_{21}l_{31} + l_{22}l_{32} & l_{31}^2 + l_{32}^2 + l_{33}^2 & \cdots & l_{31}l_{n1} + l_{32}l_{n2} + l_{33}l_{n3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{11}l_{n1} & l_{21}l_{n1} + l_{22}l_{n2} & l_{31}l_{n1} + l_{32}l_{n2} + l_{33}l_{n3} & \cdots & \sum_{k=1}^n l_{nk}^2 \end{pmatrix}$$

d.h.

$$\text{für } i = k \quad \text{gilt } a_{kk} = \sum_{j=1}^k l_{kj}^2 \quad \text{und}$$

$$\text{für } i \neq k \quad \text{gilt } a_{ik} = \sum_{j=1}^k l_{ij} \cdot l_{kj} \quad (k+1 \leq i \leq n)$$

und werten diese in einer geschickten Reihenfolge aus.

### Cholesky-Verfahren zur Berechnung von $L$ mit $A = L L^T$

Spaltenweise berechnet man für  $k = 1, 2, \dots, n$

$$l_{kk} = \left( a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2 \right)^{1/2}$$

$$l_{ik} = \frac{1}{l_{kk}} \left( a_{ik} - \sum_{j=1}^{k-1} l_{ij} \cdot l_{kj} \right) \quad (k+1 \leq i \leq n).$$

**Bemerkungen 3.6.9** i) Aus der Cholesky-Zerlegung  $A = L \cdot L^T$  ergibt sich für  $1 \leq k \leq n$  die Abschätzung

$$\sum_{j=1}^k l_{kj}^2 \leq \max_{1 \leq j \leq n} |a_{jj}|.$$

Folglich sind alle Elemente der Matrix  $L$  betragsweise durch  $\max_{1 \leq j \leq n} \sqrt{|a_{jj}|}$  beschränkt. Die Elemente können damit nicht allzu stark anwachsen, was sich günstig auf die Stabilität des Verfahrens auswirkt.

ii) Da  $A$  symmetrisch ist, wird nur Information oberhalb und einschließlich der Hauptdiagonalen benötigt. Wenn man die Diagonalelemente  $l_{kk}$  separat in einem Vektor der Länge  $n$  speichert, und die Elemente  $l_{jk}$ ,  $k < j$  unterhalb der Diagonale, so kann man die Information der Matrix  $A$  bewahren.

iii) Bei der algorithmischen Durchführung der Cholesky-Zerlegung liefert das Verfahren auch die Information, ob die Matrix positiv definit ist. Man mache sich dies als Übungsaufgabe klar!

**Satz 3.6.10 (Rechenaufwand Cholesky-Zerlegung)** Sei  $A \in \mathbb{R}^{n \times n}$  positiv definit. Dann gilt

$$\text{FLOP}(\text{Cholesky-Zerl. von } A) = \frac{2n^3 + 3n^2 - 5n}{6} + n \text{ Wurzelberechnung} = \frac{1}{3}n^3 + \mathcal{O}(n^2).$$

*Beweis.* Untersuchen wir nun die Komplexität der Cholesky-Zerlegung zuerst für einen  $k$ -ten Zerlegungsschritt und bestimmen dann den Gesamtaufwand.

i) Für den  $k$ -ten Zerlegungsschritt, d.h. die Berechnung der Elemente  $l_{ik}$  für festen Spaltenindex, sind jeweils  $(k-1)$  Multiplikationen,  $(k-1)$  Subtraktionen und eine Wurzelberechnung zur Berechnung des Diagonalelements nötig. Für jedes Nebendiagonalelement werden  $(k-1)$  Multiplikationen,  $(k-1)$  Subtraktionen und eine Division benötigt, d.h. bei  $(n-k)$  Elementen unterhalb des  $k$ -ten Diagonalelements sind dies in der Summe

$$\begin{aligned} & (2(k-1) + (n-k)(2k-1)) \text{ FLOP und 1 Wurzelberechnung} \\ &= (-2k^2 + (3+2n)k - (n+2)) \text{ FLOP und 1 Wurzelberechnung.} \end{aligned}$$

ii) Insgesamt ergibt sich dann für die Summe über alle Zerlegungsschritte

$$\begin{aligned} & n \text{ Wurzelberechnungen} + \sum_{k=1}^n (-2k^2 + (3+2n)k - (n+2)) \\ &= n \text{ Wurzelberechnungen} + \left( -2 \frac{(2n+1)(n+1)n}{6} + (3+2n) \frac{(n+1)n}{2} - (n+2)n \right) \\ &= n \text{ Wurzelberechnungen} + \frac{-(4n^3 + 6n^2 + 2n) + (6n^3 + 15n^2 + 9n) - (6n^2 + 12n)}{6} \\ &= n \text{ Wurzelberechnungen} + \frac{2n^3 + 3n^2 - 5n}{6}. \end{aligned}$$

Da eine einzelne Wurzelberechnung an Aufwand ein konstantes Vielfaches einer Gleitkommaoperation benötigt, kann man den Term  $n \text{ Wurzelberechnungen} + (3n^2 - 5n)/6$  zu  $\mathcal{O}(n^2)$  Operationen zusammenfassen.  $\square$

**Bemerkung 3.6.11** Für große  $n$  ist der Aufwand des Cholesky-Verfahrens im Vergleich zum Gauß-Algorithmus ungefähr halb so groß.





### 3.7 Bandgleichungen

Eine weitere Klasse spezieller Matrizen, welche beim Lösen linearer Gleichungssysteme eine besondere Rolle spielen, sind Bandmatrizen. Man spricht von einer Bandmatrix  $A$ , falls alle von Null verschiedenen Elemente  $a_{ik}$  in der Diagonale und in einigen dazu benachbarten Nebendiagonalen stehen. Für die Anwendungen sind insbesondere die symmetrischen, positiv definiten Bandmatrizen wichtig.

**Definition 3.7.1** Unter der **Bandbreite**  $m$  einer symmetrischen Matrix  $A \in \mathbb{R}^{n \times n}$  versteht man die kleinste natürliche Zahl  $m < n$ , so dass gilt

$$a_{ik} = 0 \quad \text{für alle } i \text{ und } k \text{ mit } |i - k| > m.$$

**Bemerkung 3.7.2** Die Bandbreite  $m$  gibt somit die Anzahl der Nebendiagonalen unterhalb bzw. oberhalb der Diagonalen an, welche die im Allgemeinen von Null verschiedenen Matrixelemente enthalten.

**Bemerkung 3.7.3** In verallgemeinerter Form spricht man auch von unterer und oberer Bandbreite, welche sich auf suggestive Weise definieren.

**Satz 3.7.4** Die untere Dreiecksmatrix  $L$  der Cholesky-Zerlegung  $A = LL^T$  einer symmetrischen, positiv definiten Bandmatrix mit der Bandbreite  $m$  besitzt dieselbe Bandstruktur, denn es gilt

$$l_{ik} = 0 \quad \text{für alle } i, k \quad \text{mit } i - k > m.$$

*Beweis.* Ergibt sich direkt aus einer Hausübung zur Hülle einer Matrix. □

**Aufgabe 3.7.5** Man beweise Satz 3.7.4.

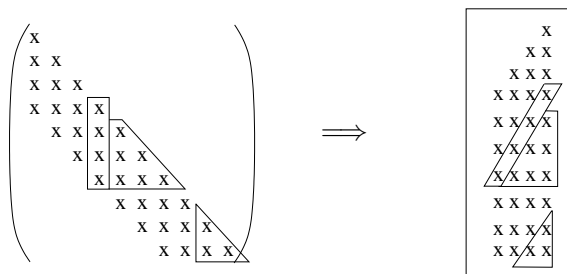


Abb. 3.3: Reduktion und Speicherung einer symmetrischen, positiv definiten Bandmatrix.

Analog zur Herleitung des Cholesky-Verfahrens auf Seite 49 erhalten wir die Rechenvorschrift des Cholesky-Verfahrens für Bandmatrizen, wobei nun die Koeffizienten zeilenweise bestimmt werden. Für  $n = 5$  und  $m = 2$  sieht die Situation wie folgt aus:

$$\begin{pmatrix} l_{11} & 0 & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 & 0 \\ 0 & l_{42} & l_{43} & l_{44} & 0 \\ 0 & 0 & l_{53} & l_{54} & l_{55} \end{pmatrix} = \begin{pmatrix} l_{11} & l_{21} & l_{31} & 0 & 0 \\ 0 & l_{22} & l_{32} & l_{42} & 0 \\ 0 & 0 & l_{33} & l_{43} & l_{53} \\ 0 & 0 & 0 & l_{44} & l_{54} \\ 0 & 0 & 0 & 0 & l_{55} \end{pmatrix} \quad \text{sym.}$$

$$= \begin{pmatrix} l_{11}^2 & & & & \\ l_{21}l_{11} & l_{21}^2 + l_{22}^2 & & & \\ l_{31}l_{11} & l_{31}l_{21} + l_{32}l_{22} & l_{31}^2 + l_{32}^2 + l_{33}^2 & & \\ & l_{42}l_{22} & l_{42}l_{32} + l_{43}l_{33} & l_{42}^2 + l_{43}^2 + l_{44}^2 & \\ & & l_{53}l_{33} & l_{53}l_{43} + l_{54}l_{44} & l_{53}^2 + l_{54}^2 + l_{55}^2 \end{pmatrix}.$$

**Cholesky-Verfahren** zur Berechnung von  $L$  mit  $A = LL^T$ , wobei  $A \in \mathbb{R}^{n \times n}$  mit Bandbreite  $0 \leq m < n$ .  
 Zeilenweise berechnet man für  $k = 1, 2, \dots, n$

$$l_{ki} = \frac{1}{l_{ii}} \left( a_{ki} - \sum_{j=\max\{1, k-m\}}^{i-1} l_{kj} \cdot l_{ij} \right) \quad (\max\{1, k-m\} \leq i \leq k-1)$$

$$l_{kk} = \left( a_{kk} - \sum_{j=\max\{1, k-m\}}^{k-1} l_{kj}^2 \right)^{1/2}$$

**Satz 3.7.6** Es sei  $A \in \mathbb{R}^{n \times n}$  eine positiv definite Bandmatrix mit Bandbreite  $1 \leq m \leq n$ . Dann beträgt der Aufwand zur Berechnung der Cholesky-Zerlegung von  $A$

$$\text{FLOP}(\text{Cholesky-Zerl. von } A) = n(m^2 + 2m) - \frac{4m^3 + 9m^2 + 5m}{6} + n \text{ Wurzelberechnungen}.$$

*Beweis.* Wendet man die Cholesky-Zerlegung auf eine positiv definite Bandmatrix  $A$  an, so ist die resultierende untere Dreiecksmatrix  $L$  mit  $A = LL^T$  wieder eine Bandmatrix mit der gleichen Bandbreite, wie  $A$  sie hat.

- i) Gehen wir davon aus, dass die Matrix  $L$  zeilenweise berechnet wird und betrachten zuerst den Aufwand für die ersten  $m$  Zeilen. Hier kann man das Ergebnis aus Satz 3.6.10 verwenden. Es sind hierfür somit

$$m \text{ Wurzelberechnungen} + \frac{2m^3 + 3m^2 - 5m}{6}$$

Operationen notwendig.

- ii) In den verbleibenden  $n - m$  Zeilen, in denen jeweils  $m + 1$  Koeffizienten zu bestimmen sind, ergibt sich Folgendes:

Um den  $j$ -ten von Null verschiedenen Nebendiagonaleintrag von  $L$  in der  $k$ -ten Zeile  $m + 1 \leq k \leq n$  zu berechnen, sind jeweils  $j - 1$  Multiplikationen,  $j - 1$  Subtraktionen und eine Division notwendig. Zur Berechnung des Diagonalelements werden  $m$  Multiplikationen,  $m$  Subtraktionen und eine Wurzelberechnung benötigt. Bei  $m$  Nebendiagonalelementen sind dies in der Summe

$$1 \text{ Wurzelberechnung} + 2m + \sum_{j=1}^m (2j - 1) = 1 \sqrt{\cdot} + m + 2 \sum_{j=1}^m j$$

$$= 1 \sqrt{\cdot} + m + m(m + 1) = 1 \sqrt{\cdot} + m(m + 2).$$

- iii) Der gesamte Aufwand beträgt also

$$n \text{ Wurzelberechnungen} + \frac{2m^3 + 3m^2 - 5m}{6} + (n - m)m(m + 2)$$

$$= n \text{ Wurzelberechnungen} + n(m^2 + 2m) - \frac{4m^3 + 9m^2 + 5m}{6}.$$

□

**Bemerkung 3.7.7** Ist für eine Klasse von positiv definiten Bandmatrizen aus  $\mathbb{R}^{n \times n}$  für beliebiges  $n$  die Bandbreite beschränkt, so wächst der Aufwand zur Berechnung der Cholesky-Zerlegung asymptotisch nur **linear** in  $n$ .

**Aufgabe 3.7.8** Man leite mit Hilfe des Cholesky-Verfahrens für symmetrische Bandmatrizen eine Rekursionsformel für s.p.d. Tridiagonalmatrizen her und gebe den Aufwand an.

### 3.8 Vandermond-Matrizen

Bei vielen Approximations- und Interpolationsproblemen treten sogenannte Vandermond-Matrizen auf, d.h. Matrizen  $V \in \mathbb{R}^{(n+1) \times (n+1)}$  von der Form

$$V = V(x_0, \dots, x_n) = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_0 & x_1 & \cdots & x_n \\ \vdots & \vdots & & \vdots \\ x_0^n & x_1^n & \cdots & x_n^n \end{pmatrix}.$$

Wir bezeichnen mit  $\mathbb{P}_n$  die Menge der Polynome vom Grad  $n$ . Sei z.B. zu  $(n+1)$ -Daten  $(x_i, f_i)$  ein Polynom  $n$ -ten Grades  $p \in \mathbb{P}_n$  gesucht mit der Eigenschaft  $p(x_i) = f_i$  ( $i = 0, \dots, n$ ). Wählt man den Ansatz

$$p(x) = \sum_{j=0}^n a_j x^j,$$

so führt dies zu einem LGS

$$V^T a = f$$

mit  $a = (a_0, \dots, a_n)^T$  und  $f = (f_0, \dots, f_n)^T$ . Man mache sich dies an einem einfachen Beispiel klar!

Diesen Zusammenhang zwischen dem Interpolationsproblem und dem System  $V^T a = f$  macht man sich auch beim Lösen zu Nutze. Die Bestimmung des Koeffizientenvektors  $a$  bzw. das Lösen des Gleichungssystems gliedert sich dabei in zwei Teile.

- Zuerst bestimmt man das Interpolationspolynom  $p$  in der Newton-Darstellung, d.h.

$$\begin{aligned} p(x) &= \sum_{k=0}^n c_k \left( \prod_{j=0}^{k-1} (x - x_j) \right) \\ &= c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + \dots + c_n(x - x_0) \cdots (x - x_{n-1}) \end{aligned}$$

Die Konstanten  $c_k$  werden als dividierte Differenzen bezeichnet, sie lösen das lineare Gleichungssystem

$$\begin{pmatrix} 1 & & & & \\ 1 & (x_1 - x_0) & & & \\ 1 & (x_2 - x_0) & (x_2 - x_0)(x_2 - x_1) & & \\ \vdots & & & \ddots & \\ 1 & (x_n - x_0) & (x_n - x_0)(x_n - x_1) & \dots & \prod_{j=0}^{n-1} (x_n - x_j) \end{pmatrix} \begin{pmatrix} c_0 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} f_0 \\ \vdots \\ f_n \end{pmatrix}$$

und können wie folgt bestimmt werden.

$$\begin{aligned} &(c_0, \dots, c_n)^T = (f_0, \dots, f_n)^T \\ &\text{for } k = 0 : n - 1 \\ &\quad \text{for } i = n : -1 : k + 1 \\ &\quad \quad c_i = (c_i - c_{i-1}) / (x_i - x_{i-k-1}) \\ &\quad \text{end} \\ &\text{end} \end{aligned} \tag{3.23}$$

- Im zweiten Schritt gewinnt man nun die  $a_i$ 's aus den  $c_i$ 's durch eine Basistransformation von  $(1, (x - x_0), (x - x_0)(x - x_1), \dots, \prod_{j=0}^{n-1} (x - x_j))$  nach  $(1, x, x^2, \dots, x^n)$ . Die Transformationsmatrix  $U$  ergibt sich aus entsprechendem Koeffizientenvergleich

$$\begin{pmatrix} 1 & -x_0 & x_0x_1 & \dots & \prod_{j=0}^{n-1} (-x_j) \\ & 1 & -(x_0 + x_1) & & \\ & & 1 & & \vdots \\ & & & \ddots & \\ & & & & 1 \end{pmatrix} c = a. \quad (3.24)$$

Anstatt die Matrix  $U$  zu bestimmen, kann  $a$  wie folgt bestimmt werden.

$$\begin{aligned} (a_0, \dots, a_n)^T &= (c_0, \dots, c_n)^T \\ \text{for } k &= n-1 : -1 : 0 \\ &\text{for } i = k : n-1 \\ &\quad a_i = a_i - x_k a_{i+1} \\ &\text{end} \\ \text{end} \end{aligned} \quad (3.25)$$

Die Kombination beider Schritte liefert den folgenden Algorithmus.

---

**Algorithmus 3.8.1: Polynominterpolation:**  $V^T a = f$

Gegeben seien  $x(0:n) \in \mathbb{R}^{n+1}$  mit paarweise verschiedenen Einträgen und  $f = f(0:n) \in \mathbb{R}^{n+1}$ . Der folgende Algorithmus überschreibt  $f$  mit der Lösung  $a = a(0:n)$  zu dem Vandermondschen System

$$V(x_0, \dots, x_n)^T a = f$$

```
for k = 0 : n-1
    for i = n : -1 : k+1
        f(i) = (f(i) - f(i-1))/(x(i) - x(i-k-1))
    end
end
for k = n-1 : -1 : 0
    for i = k : n-1
        f(i) = f(i) - x(k) f(i+1)
    end
end
```

---

**Satz 3.8.1** Das Verfahren 3.8.1 benötigt zum Lösen des Vandermondschen System

$$V(x_0, \dots, x_n)^T a = f$$

nur  $5(n^2 + n)/2$  Operationen.

*Beweis.* Die Anweisung in der ersten inneren FOR-Schleife von Algorithmus 3.8.1 wird für ein  $k \in \{0, \dots, n-1\}$   $(n-k)$ -mal ausgeführt, insgesamt also  $\sum_{k=0}^{n-1} (n-k) = \sum_{k=1}^n k = n(n+1)/2$

-mal. Hierbei werden jedesmal 3 Gleitkommaoperationen benötigt. Für die zweite Doppelschleife ergeben sich analog zu der obigen Betrachtung  $n(n+1)/2$  Aufrufe der inneren Anweisung, für die jeweils 2 Gleitkommaoperationen benötigt werden.  $\square$

**Bemerkung 3.8.2** Man beachte, dass für spezielle vollbesetzte Matrizen Verfahren existieren, die weniger als  $\mathcal{O}(n^3)$  Operationen zur Berechnung der Lösung benötigen. Auch die Matrix braucht nicht vollständig gespeichert werden.

### Das System $Vz = b$

Der Vollständigkeit wegen sei hier auch ein Verfahren für  $Vz = b$  aufgeführt. Es sei  $L_k(\alpha) \in \mathbb{R}^{(n+1) \times (n+1)}$  eine untere bidiagonale Matrix definiert durch

$$L_k(\alpha) = \left( \begin{array}{c|cccc} I_k & & & & 0 \\ \hline & 1 & \dots & & 0 \\ & -\alpha & 1 & & \\ & & \ddots & \ddots & \\ 0 & \vdots & & \ddots & \ddots & \vdots \\ & & & \ddots & 1 & \\ & 0 & \dots & & -\alpha & 1 \end{array} \right)$$

und die Diagonalmatrix  $D_k$  sei definiert durch

$$D_k = \text{diag}(\underbrace{1, \dots, 1}_{(k+1)\text{-mal}}, x_{k+1} - x_0, \dots, x_n - x_{n-k-1}).$$

Mit diesen Definitionen kann man leicht aus (3.23) nachvollziehen, falls  $f = f(0 : n)$  und  $c = c(0 : n)$  der Vektor der dividierten Differenzen ist, dass  $c = Lf$  gilt, wobei  $L$  die Matrix sei, die wie folgt definiert ist:

$$L = D_{n-1}^{-1} L_{n-1}(1) \cdots D_0^{-1} L_0(1).$$

Ähnlich erhält man aus (3.25), dass

$$a = Uc \tag{3.26}$$

gilt, wobei die durch (3.26) definierte Matrix  $U$  sich auch definieren lässt durch

$$U = L_0(x_0)^T \cdots L_{n-1}(x_{n-1})^T.$$

Mit diesen Definitionen ergibt sich sofort die Aussage  $a = ULf$  mit  $V^{-T} = UL$ . Die Lösung von  $Vz = b$  ergibt sich somit durch

$$\begin{aligned} z &= V^{-1}b = L^T(U^Tb) \\ &= (L_0(1)^T D_0^{-1} \cdots L_{n-1}(1)^T D_{n-1}^{-1}) (L_{n-1}(x_{n-1}) \cdots L_0(x_0)) b. \end{aligned}$$

Diese Erkenntnis führt zu dem folgenden Algorithmus.

---

### Algorithmus 3.8.2: Vandermondsches System: $Vz = b$

Gegeben seien  $x(0:n) \in \mathbb{R}^{n+1}$  mit paarweise verschiedenen Einträgen und  $b = b(0:n) \in \mathbb{R}^{n+1}$ . Der folgende Algorithmus überschreibt  $b$  mit der Lösung  $z = z(0:n)$  zu dem Vandermondschen System

$$V(x_0, \dots, x_n)z = b.$$

```

for k = 0 : n - 1
    for i = n : -1 : k + 1
        b(i) = b(i) - x(k) b(i - 1)
    end
end
for k = n - 1 : -1 : 0
    for i = k + 1 : n
        b(i) = b(i) / (x(i) - x(i - k - 1))
    end
    for i = k + 1 : n
        b(i) = b(i) - b(i + 1)
    end
end
end

```

**Satz 3.8.3** Das Verfahren 3.8.2 benötigt zum Lösen des Vandermondschen System

$$V(x_0, \dots, x_n)a = b$$

nur  $5(n^2 + n)/2$  Operationen.

*Beweis.* Analog zu den Betrachtungen im Beweis von Satz 3.8.1 erkennt man, dass die Anweisungen in den inneren FOR-Schleifen von Algorithmus 3.8.2 jeweils  $n(n+1)/2$ -mal aufgerufen werden. Jedesmal werden 2 bzw. 1 Gleitkommaoperation benötigt.  $\square$

### 3.9 Fehlerabschätzung mittels Kondition

Wir wollen nun zwei wichtige Fragestellungen untersuchen, welche bei dem numerischen Lösen von linearen Gleichungssystemen (LGS) auftreten. Zum einen betrachten wir eine Näherungslösung  $\tilde{x}$  zu der Lösung  $x$  von  $Ax = b$ . Welche Rückschlüsse kann man von der Größe des **Residuums**  $r := b - A\tilde{x}$  auf den Fehler  $e = x - \tilde{x}$  ziehen? Des Weiteren rechnet man im Allgemeinen mit endlicher Genauigkeit auf einem Rechner. Welche Einflüsse haben dabei Störungen der Ausgangsdaten auf die Lösung  $x$ ?

Zu einer Matrix  $A \in \mathbb{R}^{n \times n}$  und einem Vektor  $x \in \mathbb{R}^n$  sei  $\|A\|$  eine beliebige Matrixnorm<sup>4</sup> und  $\|x\|$  eine dazu verträgliche Vektornorm (d.h.  $\|Ax\| \leq \|A\| \|x\|$ ). Nach Definition des Fehlers  $e$  und Residuums  $r$  gilt

$$Ae = Ax - A\tilde{x} = b - A\tilde{x} = r.$$

Daraus folgt

$$\|e\| = \|A^{-1}r\| \leq \|A^{-1}\| \|r\|.$$

Mit

$$\|b\| = \|Ax\| \leq \|A\| \|x\|$$

<sup>4</sup>Vgl. hierzu Anhang B.

folgt für den relativen Fehler die Abschätzung

$$\frac{\|e\|}{\|x\|} \leq \frac{\|A^{-1}\| \|r\|}{\|b\| \cdot \frac{1}{\|A\|}} = \|A^{-1}\| \|A\| \cdot \frac{\|r\|}{\|b\|}. \quad (3.27)$$

Dies motiviert den Begriff der Konditionszahl.

**Definition 3.9.1 (Konditionszahl)** Man bezeichnet

$$\kappa(A) = \|A\| \|A^{-1}\| \quad (3.28)$$

als die **Konditionszahl** der Matrix  $A$  bezüglich der verwendeten Matrixnorm.

**Beispiel 3.9.2** Man betrachte das LGS  $Ax = b$  mit

$$A = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} \\ \frac{1}{5} & \frac{1}{6} & \frac{1}{7} & \frac{1}{8} & \frac{1}{9} \end{pmatrix} \quad b := (b_j), \quad b_j := \sum_{k=1}^5 \frac{1}{k+j-1} \quad (1 \leq j \leq 5).$$

Für die Matrixnormen und die Konditionszahlen erhält man in den Fällen  $p = 1, 2, \infty$

$$\begin{array}{lll} \|A\|_1 & = & 2.28\overline{3}, \quad \|A^{-1}\|_1 = 413280, \quad \kappa_1(A) = 943656, \\ \|A\|_2 & \approx & 1.5670507, \quad \|A^{-1}\|_2 \approx 304142.84, \quad \kappa_2(A) \approx 476607.25, \\ \|A\|_\infty & = & 2.28\overline{3}, \quad \|A^{-1}\|_\infty = 413280, \quad \kappa_\infty(A) = \kappa_1(A). \end{array}$$

Sei die exakte Lösung  $x = (1, \dots, 1)^T$ . Ist nun  $\tilde{x} = (1 - \varepsilon)x$ , dann gilt

$$\frac{\|e\|}{\|x\|} = \frac{\|\varepsilon x\|}{\|x\|} = |\varepsilon|$$

und

$$r = b - A\tilde{x} = Ax - A\tilde{x} = A(x - (1 - \varepsilon)x) = \varepsilon Ax = \varepsilon b,$$

d.h.

$$\frac{\|r\|}{\|b\|} = |\varepsilon|.$$

Obwohl die Konditionszahl sehr groß ist, verhält sich bei dieser Störung  $\frac{\|e\|}{\|x\|}$  wie  $\frac{\|r\|}{\|b\|}$ . Obige Abschätzung ist offensichtlich eine „worst case“ Abschätzung. Betrachten wir nun



$$\tilde{x} = (0.993826, 1.116692, 0.493836, 1.767191, 0.623754)^T. \quad (3.29)$$

$A$  habe die Eigenwerte  $0 \leq \lambda_1 < \dots < \lambda_5$  mit den zugehörigen Eigenvektoren  $\varphi_1, \dots, \varphi_5$ ,  $\|\varphi_j\|_2 = 1$ . Man beachte, dass  $\tilde{x}$  in (3.29) so gewählt ist, dass  $\tilde{x} - x \approx \varphi_1$  gilt und

$$x \approx -0.004 \varphi_1 - 0.042 \varphi_2 + 0.244 \varphi_3 + 0.972 \varphi_4 + 1.998 \varphi_5,$$

d.h.  $x$  von  $\varphi_5$  dominiert wird. Dann erhalten wir

$$\frac{\|e\|_1 \|b\|_1}{\|x\|_1 \|r\|_1} \approx 0.41 \kappa_1(A), \quad \frac{\|e\|_2 \|b\|_2}{\|x\|_2 \|r\|_2} \approx 0.89 \kappa_2(A), \quad \frac{\|e\|_\infty \|b\|_\infty}{\|x\|_\infty \|r\|_\infty} \approx 0.74 \kappa_\infty(A)$$

und schätzen dann in der richtigen Größenordnung ab!

**Aufgabe 3.9.3** Es sei  $A \in \mathbb{R}^{n \times n}$  eine positiv definite Matrix. Die Eigenwerte von Matrix  $A$  seien  $0 < \lambda_1 \leq \dots \leq \lambda_n$  mit den zugehörigen Eigenvektoren  $\varphi_1, \dots, \varphi_n$  ( $\|\varphi_j\|_2 = 1, j = 1, \dots, n$ ).

i) Es sei  $x \in \mathbb{R}^n \setminus \{0\}$ ,  $b := Ax$  und  $\tilde{x} = cx$  ( $0 \neq c \in \mathbb{R}$ ). Man zeige

$$\frac{\|x - \tilde{x}\|}{\|x\|} \frac{\|b\|}{\|b - A\tilde{x}\|} = 1.$$

ii) Man zeige

$$\kappa_2(A) = \lambda_n / \lambda_1.$$

iii) Es sei  $x = c\varphi_n$ ,  $c \neq 0$ ,  $b := Ax$  und  $\tilde{x} = x + \varphi_1$ . Man zeige

$$\frac{\|x - \tilde{x}\|}{\|x\|} \frac{\|b\|}{\|b - A\tilde{x}\|} = \kappa_2(A).$$

Untersuchen wir nun, welchen Einfluss kleine Störungen in den Ausgangsdaten  $A, b$  auf die Lösung  $x$  des linearen Gleichungssystems haben können, d.h. wir sind interessiert an der **Empfindlichkeit** der Lösung  $x$  auf Störungen in den Koeffizienten. Die genaue Frage lautet:

Wie groß kann die Änderung  $\delta x$  der Lösung  $x$  von  $Ax = b$  sein, falls die Matrix um  $\delta A$  und  $b$  durch  $\delta b$  gestört sind? Dabei seien  $\delta A$  und  $\delta b$  kleine Störungen, so dass  $A + \delta A$  immer noch regulär ist. Es sei  $x + \delta x$  die Lösung zu

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

Teilweises Ausmultiplizieren liefert

$$Ax + \delta A x + (A + \delta A)\delta x = b + \delta b$$

und somit mit  $Ax = b$

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

Für eine submultiplikative Matrixnorm und eine dazu verträgliche Vektornorm ergibt sich somit

$$\|\delta x\| \leq \|(I + A^{-1}\delta A)^{-1}\| \|A^{-1}\| (\|\delta b\| + \|\delta A\| \|x\|)$$

und wir erhalten die Abschätzung

$$\frac{\|\delta x\|}{\|x\|} \leq \|(I + A^{-1}\delta A)^{-1}\| \|A^{-1}\| \left( \frac{\|\delta b\|}{\|x\|} + \|\delta A\| \right). \quad (3.30)$$

Wir zeigen in Lemma 3.9.6, dass für  $\|B\| < 1$  die Abschätzung

$$\|(I + B)^{-1}\| \leq \frac{1}{1 - \|B\|} \quad \text{existiert.}$$

Setzen wir nun  $B = A^{-1}\delta A$  und  $\|A^{-1}\| \|\delta A\| < 1$  (d.h.  $\delta A$  ist eine kleine Störung von  $A$ ) voraus und erweitern auf der rechten Seite mit  $\|A\|$ , so erhalten wir unter Ausnutzung der Verträglichkeit der Normen und mit  $Ax = b$


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

Wir halten das Ergebnis nun in einem Satz fest.



**Satz 3.9.4** Es sei  $A \in \mathbb{R}^{n \times n}$  regulär und  $x \in \mathbb{R}^n$  die exakte Lösung von  $Ax = b$ . Die rechte Seite  $b$  sei um  $\delta b$  gestört und für die Störung von  $A$  gelte  $\kappa(A) \|\delta A\|/\|A\| = \|A^{-1}\| \|\delta A\| < 1$  und  $A + \delta A$  ist immer noch regulär. Dann gilt für die Lösung  $\tilde{x}$  des gestörten Systems mit Koeffizientenmatrix  $A + \delta A$  und rechter Seite  $b + \delta b$

$$\frac{\|x - \tilde{x}\|}{\|x\|} \leq \frac{\kappa(A)}{1 - \kappa(A) \frac{\|\delta A\|}{\|A\|}} \left( \frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right). \quad (3.31)$$

**Bemerkung 3.9.5** Die Konditionszahl  $\kappa(A)$  der Koeffizientenmatrix  $A$  ist folglich die entscheidende Größe, welche die Empfindlichkeit der Lösung bzgl. der Störungen  $\delta A$  und  $\delta b$  beschreibt. 

Es bleibt folgendes Lemma zu beweisen:

**Lemma 3.9.6 (Neumann-Reihe)** Sei  $A \in \mathbb{R}^{n \times n}$ ,  $\|\cdot\|$  eine submultiplikative Matrixnorm und  $\|A\| < 1$ . Dann ist  $(I - A)$  regulär und

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k \quad (\text{Neumann-Reihe}) \quad (3.32)$$

mit

$$\|(I - A)^{-1}\| \leq \frac{1}{1 - \|A\|}. \quad (3.33)$$

*Beweis.* Mit  $\|A\| < 1$ , der Submultiplikativität und der Summenformel für die geometrische Reihe erhält man

$$\sum_{k=0}^m \|A^k\| \leq \sum_{k=0}^m \|A\|^k \leq \sum_{k=0}^{\infty} \|A\|^k = \frac{1}{1 - \|A\|} < \infty \quad (m \in \mathbb{N}). \quad (3.34)$$

Der Raum  $\mathbb{R}^{n \times n}$  ist isomorph zu  $\mathbb{R}^{n^2}$  (siehe Anhang B.2). In [Analysis II, Beispiel 8.4.6] wurde gezeigt, dass  $\mathbb{R}^{n^2}$  vollständig ist bezüglich irgendeiner Norm auf  $\mathbb{R}^{n^2}$ . Somit ist  $\mathbb{R}^{n \times n}$  vollständig bezüglich  $\|\cdot\|$  und aus der absoluten Konvergenz folgt die Konvergenz von  $\sum_{k=0}^{\infty} A^k$ . Ebenso folgt aus  $\|A^k\| \leq \|A\|^k \rightarrow 0$  für  $k \rightarrow \infty$  die Konvergenz von  $\lim_{k \rightarrow \infty} A^k = 0$ . Weiter gilt die Gleichung („Teleskopsumme“)

$$\left( \sum_{k=0}^m A^k \right) (I - A) = I - A^{m+1} \quad (m \in \mathbb{N}). \quad (3.35)$$

Der Grenzübergang von (3.35) führt zur Gleichung

$$\left( \sum_{k=0}^{\infty} A^k \right) (I - A) = I. \quad (3.36)$$

Das bedeutet,  $I - A$  ist regulär und  $(I - A)^{-1} = \sum_{k=0}^{\infty} A^k$ . Mit der Summenformel für die geometrische Reihe erhält man schließlich

$$\|(I - A)^{-1}\| \leq \lim_{N \rightarrow \infty} \sum_{k=0}^N \|A^k\| \leq \lim_{N \rightarrow \infty} \sum_{k=0}^N \|A\|^k = \frac{1}{1 - \|A\|}, \quad (3.37)$$

womit der Satz bewiesen ist. □

Zum Ende des Kapitels wollen wir nun den praktischen Nutzen der Abschätzung (3.31) in einer Faustregel festhalten.

Bei einer  $d$ -stelligen dezimalen Gleitkommarechnung können die relativen Fehler der Ausgangsgrößen für beliebige, kompatible Normen von der Größenordnung

$$\frac{\|\delta A\|}{\|A\|} \approx 5 \cdot 10^{-d} \quad \frac{\|\delta b\|}{\|b\|} \approx 5 \cdot 10^{-d}$$

sein. Ist die Konditionszahl  $\kappa(A) \approx 10^\alpha$  mit  $5 \cdot 10^{\alpha-d} \ll 1$ , so ergibt die Abschätzung (3.31)

$$\frac{\|\delta x\|}{\|x\|} \leq 10^{\alpha-d+1}.$$

Das heißt, dass  $\|\delta x\|$  maximal in der Größenordnung der  $(d - \alpha - 1)$ -ten Dezimalstelle von  $\|x\|$  liegen kann und dies motiviert folgende Daumenregel:



**Bemerkung 3.9.7 (Daumenregel zur Genauigkeit)** Wird  $Ax = b$  mit  $d$ -stelliger dezimaler Gleitkommarechnung gelöst, und beträgt die Konditionszahl  $\kappa(A) \approx 10^\alpha$ , so sind, bezogen auf die betragsgrößte Komponente, nur  $(d - \alpha - 1)$  Dezimalstellen sicher.

### MATLAB-Beispiel:

Es ist bekannt, dass die Gauß-Elimination selbst mit Spaltenpivotstrategie zu überraschend ungenauen Ergebnissen beim Lösen von linearen Gleichungssystemen führen kann, obwohl die Matrix gut konditioniert ist.

Betrachten wir hierzu die von Wilkinson angegebene pathologische Matrix

$$A = \begin{pmatrix} 1 & & & 1 \\ -1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ -1 & \dots & -1 & 1 \end{pmatrix}.$$

```
>> A=toeplitz([1,-ones(1,59)], ...
               [1,zeros(1,59)]);
>> A(:,60)=1;
>> cond(A)
ans =
    26.8035 % rel. gut konditioniert
>> randn('state', 3383)
>> x=randn(60,1);
>> b=A*x;
>> x1=A\b;
>> norm(x-x1)/norm(x)
ans =
    0.3402 % großer rel. Fehler
```

**Bemerkung 3.9.8** Das Beispiel lässt vermuten, dass das Gauß-Verfahren über die ganze Menge der invertierbaren Matrizen betrachtet nicht stabil ist. Für die in der Praxis auftretenden Matrizen, ist das Gauß-Verfahren mit Spaltenpivotierung jedoch „in der Regel“ stabil. Für eine weitere Stabilitätsanalyse des Gauß-Verfahrens sei auf [Deuffhard/Hohmann], die grundlegenden Artikel von Wilkinson [Wilkinson65, Wilkinson69] sowie auf die Übersichtsartikel [Higham, Discroll/Maki] verwiesen.

# 4 ITERATIVE LÖSUNG LINEARER GLEICHUNGSSYSTEME

Die in Kapitel 3 beschriebenen direkten Verfahren gehen überwiegend von beliebigen vollbesetzten Matrizen aus. Viele praktische Probleme führen aber zu der Aufgabe, ein sehr großes lineares Gleichungssystem  $Ax = b$  zu lösen, bei dem  $A \in \mathbb{R}^{n \times n}$  nur schwachbesetzt (engl. sparse) ist, d.h. viele Nulleinträge besitzt (wie etwa Beispiel 3.0.1). Die bisherigen Verfahren nutzen diese spezielle Struktur nicht aus und führen beim Lösen des LGS teilweise sogar zu vollbesetzten Zwischenmatrizen. Man betrachte dazu folgendes Beispiel.

**Beispiel 4.0.1** Zu der Matrix

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & & & & \\ 1 & & 5 & & & \\ 1 & & & 10 & & \\ 1 & & & & 15 & \\ 1 & & & & & 10 \end{pmatrix}$$

lautet die  $LR$ -Zerlegung mit  $A = L \cdot R$

$$L = \begin{pmatrix} 1 & & & & & \\ 1 & 1 & & & & \\ 1 & -1 & 1 & & & \\ 1 & -1 & -\frac{2}{3} & 1 & & \\ 1 & -1 & -\frac{2}{3} & -\frac{1}{2} & 1 & \\ 1 & -1 & -\frac{2}{3} & -\frac{1}{2} & -\frac{1}{2} & 1 \end{pmatrix}, \quad R = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ & 1 & -1 & -1 & -1 & -1 \\ & & 3 & -2 & -2 & -2 \\ & & & \frac{20}{3} & -\frac{10}{3} & -\frac{10}{3} \\ & & & & 10 & -5 \\ & & & & & \frac{5}{2} \end{pmatrix}.$$

Obwohl  $A$  nur in der ersten Zeile und Spalte sowie auf der Diagonalen Nichtnulleinträge besitzt, sind  $L$  und  $R$  vollbesetzt.



**Bemerkung 4.0.2** Ist die Bandbreite einer Matrix groß und in jeder Zeile treten nur wenige Nichtnulleinträge auf, dann ist ein Bandlöser „teuer“ und die folgenden Verfahren liefern gute Alternativen.

**Beispiel 4.0.3 (2D-Standardmatrix)** Nun betrachten wir eine (zweidimensionale) quadratische Membran, die am Rand fest eingespannt ist und auf die eine äußere Kraft  $f$  wirkt. Die gesuchte Auslenkung  $u : [0, 1]^2 \rightarrow \mathbb{R}$  ergibt sich als Lösung des sogenannten **Dirichlet-Problems** auf dem Einheitsquadrat

$$\begin{aligned} -\Delta u(x, y) &:= -\frac{\partial^2}{\partial x^2} u(x, y) - \frac{\partial^2}{\partial y^2} u(x, y) = f(x, y), \quad (x, y) \in \Omega := (0, 1)^2, \\ u(x, y) &= 0, \quad (x, y) \in \Gamma := \partial\Omega. \end{aligned} \quad (4.1)$$

Den Differenzialoperator  $\Delta$  nennt man auch **Laplace-Operator**.

Wir verwenden die gleiche Idee wie im eindimensionalen Fall (Beispiel 3.0.1) und definieren ein äquidistantes Gitter

$$\Omega_h := \{(x, y) \in \overline{\Omega} : x = kh, \quad y = \ell h, \quad 0 \leq k, \ell \leq N + 1\} \quad (4.2)$$

für  $h := \frac{1}{N+1}$ ,  $N = N_h \in \mathbb{N}$ , wie in Beispiel 3.0.1. Der Rand besteht jetzt natürlich aus mehr als zwei Punkten, nämlich

$$\partial\Omega_h := \{(x, y) \in \Gamma : x = kh \text{ oder } y = \ell h, \quad 0 \leq k, \ell \leq N+1\}, \quad (4.3)$$

vgl. Abbildung 4.1. Wir definieren  $\mathring{\Omega}_h := \Omega_h \setminus \partial\Omega_h$ .

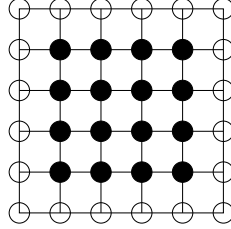


Abb. 4.1: Äquidistantes Gitter auf  $\overline{\Omega} = [0, 1]^2$ . Die inneren Punkte ( $\mathring{\Omega}_h$ ) sind ausgefüllt dargestellt,  $\partial\Omega_h$  besteht aus den nicht ausgefüllten Punkten ( $\circ$ ).

Die partiellen Ableitungen zweiter Ordnung approximieren wir wieder durch den zentralen Differenzenquotienten, also

$$\begin{aligned} \Delta u(x, y) &= \frac{\partial^2}{\partial x^2} u(x, y) + \frac{\partial^2}{\partial y^2} u(x, y) \\ &\approx \frac{1}{h^2} (u(x+h, y) - 2u(x, y) + u(x-h, y)) \\ &\quad + \frac{1}{h^2} (u(x, y+h) - 2u(x, y) + u(x, y-h)) \\ &= \frac{1}{h^2} (u(x+h, y) + u(x-h, y) + u(x, y+h) + u(x, y-h) - 4u(x, y)). \end{aligned} \quad (4.4)$$

Dazu beschreiben wir zunächst das lineare Gleichungssystem  $A_h u_h = f_h$ . Die genaue Gestalt der Matrix  $A_h \in \mathbb{R}^{N^2 \times N^2}$  hängt von der Nummerierung ab. Wir wählen die so genannte **lexikographische Nummerierung** (vgl. Abbildung 4.2)

$$z_k = (x_i, y_j), \quad x_i = ih, \quad y_j = jh, \quad k := (j-1)N + i.$$

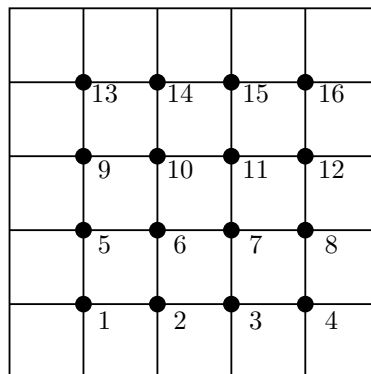


Abb. 4.2: Lexikographische Nummerierung der Gitterpunkte für  $\Omega = (0, 1)^2$ ,  $N = 5$ .

Dann erhält man für das lineare Gleichungssystem  $A_h u_h = f_h$  eine **Block-Tridiagonalmatrix**

$$A_h = \begin{bmatrix} B_h & C_h & & 0 \\ C_h & B_h & \ddots & \\ & \ddots & \ddots & \ddots \\ 0 & & & C_h & B_h \end{bmatrix} \in \mathbb{R}^{N^2 \times N^2}$$

mit den Blöcken

$$B_h = \begin{bmatrix} 4 & -1 & & 0 \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & \ddots \\ & & \ddots & 4 & -1 \\ 0 & & & -1 & 4 \end{bmatrix} \in \mathbb{R}^{N \times N}, \quad C_h = \text{diag}(-1) \in \mathbb{R}^{N \times N},$$

der rechten Seite  $f_h = (h^2 f(z_k))_{k=1, \dots, N^2}$ ,  $k = (j-1)N + i$ ,  $1 \leq i, j \leq N$ , sowie dem Lösungsvektor  $u_h = (u(z_k))_{k=1, \dots, N^2} \in \mathbb{R}^{N^2}$ . Wir nennen diese Matrix auch **2D-Standardmatrix**, sie ist die direkte Verallgemeinerung der Standardmatrix aus Beispiel 3.0.1.

Für dieses Gleichungssystem kann man keine einfache Rekursionsformel für die Cholesky-Zerlegung herleiten. Allerdings ist  $A_h$  dünnbesetzt, symmetrisch und positiv definit. Mehr zu diesem Beispiel in [Arendt/Urban].  $\square$

Aus den oben genannten Gründen wurden schon früh iterative Verfahren zur Lösung von LGS herangezogen. Bei diesen Verfahren wird ausgehend von einem Startvektor  $x^{(0)}$  eine Folge von Vektoren

$$x^{(0)} \rightarrow x^{(1)} \rightarrow x^{(2)} \rightarrow \dots$$

mittels einer Iterationsvorschrift

$$x^{(k+1)} = \phi(x^{(k)}), \quad k = 0, 1, \dots \quad (4.5)$$

erzeugt, die gegen die gesuchte Lösung  $x$  konvergiert. In den folgenden Abschnitten werden sowohl die klassischen Iterationsverfahren, die bereits Mitte des 19. Jahrhunderts entdeckt wurden, als auch das Gradienten-Verfahren sowie das 1952 von Hestenes und Stiefel entwickelte Verfahren der konjugierten Gradienten vorgestellt.

Allen diesen Verfahren ist gemein, dass ein einzelner Iterationsschritt  $x^{(k)} \rightarrow x^{(k+1)}$  einen Rechenaufwand erfordert, welcher vergleichbar ist mit der Multiplikation von  $A$  mit einem Vektor, d.h. insbesondere mit einem geringen Aufwand, sofern  $A$  schwachbesetzt ist. Im Gegensatz zu den direkten Verfahren liefern diese Iterationsverfahren die exakte Lösung  $x$  des LGS im Allgemeinen nicht mehr nach endlich vielen Schritten. Da man aber in der Regel an der Lösung  $x$  nur bis auf eine vorgegebene Genauigkeit  $\epsilon$  interessiert ist, die von der Genauigkeit der Eingabedaten abhängt (vgl. Kapitel 2 und 3.9), scheint dieses Vorgehen sinnvoll.

## 4.1 Klassische Iterationsverfahren

Gegeben sei eine reguläre Matrix  $A \in \mathbb{R}^{n \times n}$  und ein lineares Gleichungssystem

$$Ax = b$$

mit der exakten Lösung  $x$ . Mit Hilfe einer beliebigen regulären Matrix  $B \in \mathbb{R}^{n \times n}$  erhält man Iterationsvorschriften der Form (4.5) aus der Gleichung

$$Bx + (A - B)x = b,$$

indem man

$$Bx^{(k+1)} + (A - B)x^{(k)} = b$$

setzt und nach  $x^{(k+1)}$  auflöst

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - B^{-1}(Ax^{(k)} - b) \\ &= (I - B^{-1}A)x^{(k)} + B^{-1}b. \end{aligned} \quad (4.6)$$

Das bedeutet, die Iterationsvorschrift  $\phi$  aus (4.5) ist gegeben durch  $\phi(x^{(k)}) := x^{(k)} - B^{-1}(Ax^{(k)} - b)$ . Insbesondere gilt, dass  $\phi(x) = x$  äquivalent zu  $Ax = b$  ist.

Jede Wahl einer nichtsingulären Matrix  $B$  führt zu einem möglichen **Iterationsverfahren**. Es wird umso brauchbarer, je besser  $B$  die folgenden Kriterien erfüllt

- i)  $B$  ist leicht zu invertieren (**einfache Realisierbarkeit**);
- ii) die Eigenwerte von  $(I - B^{-1}A)$  sollen möglichst kleine Beträge haben (**Konvergenzeigenschaft**).

Wir wollen hier nun einige Beispiele angeben. Dazu verwenden wir folgende (additive) Standardzerlegung

$$A = L + D + R,$$

wobei  $D$  eine **Diagonalmatrix**,  $L$  eine strikte **untere Dreiecksmatrix** und  $R$  eine strikte **obere Dreiecksmatrix** seien. Die Wahl

- i)  $B = \gamma I$  liefert das **Richardson-Verfahren**;
- ii)  $B = D$  liefert das **Jacobi-Verfahren (Gesamtschrittverfahren)**;
- iii)  $B = L + D$  oder  $B = D + R$  liefert das **Gauß-Seidel-Verfahren (Einzelschrittverfahren)**.

Was zeichnet nun die einzelnen Verfahren aus? Betrachten wir dazu ein Beispiel.

**Beispiel 4.1.1** Zu gegebenem  $n \in \mathbb{N}$  und

$$A = \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{n \times n}, b = \vec{1} \in \mathbb{R}^n, x^{(0)} = \vec{0} \in \mathbb{R}^n$$

(vgl. Beispiel 3.0.1 mit  $\lambda = 0$ ) bestimmen wir die **Konvergenzrate**

$$c := \max_k \frac{\|x - x^{(k+1)}\|_2}{\|x - x^{(k)}\|_2},$$

für das Jacobi- und das Gauß-Seidel-Verfahren ( $B = D + R$ ), d.h. den Faktor, um den der **Fehler** in der 2-Norm in jedem Iterationsschritt **mindestens reduziert** wird.

**MATLAB-Funktion: runKonvergenz.m**

```

1  n = 10;
2  e = ones(n,1);
3  A = spdiags([e -2*e e], -1:1, n, n);
4  x_ex = rand(n,1);      % exakte Loesung
5  b = A * x_ex;          % exakte rechte Seite
6  x{1} = rand(n,1);      % zufaelliger Startv.
7  x{2}=x{1};
8  W{1} = triu(A);         % Gauss-Seidel
9  W{2} = diag(diag(A));   % Jacobi
10 for j = 1:length(x)
11     error_old = norm(x{j}-x_ex);
12     for k = 1 : 20
13         x{j} = x{j} + W{j} \ (b-A*x{j});
14         error_new = norm(x{j}-x_ex);
15         quot{j}(k) = error_new/error_old;
16         error_old = error_new;
17     end
18 end
19 plot(1:20,quot{1},'m-s',1:20,quot{2},'k:*');
20 xlabel('Anzahl der Iterationen'), ylabel('Kontraktion')
21 legend({'Gauss-Seidel-Verf.','Jacobi-Verf.'},4)

```

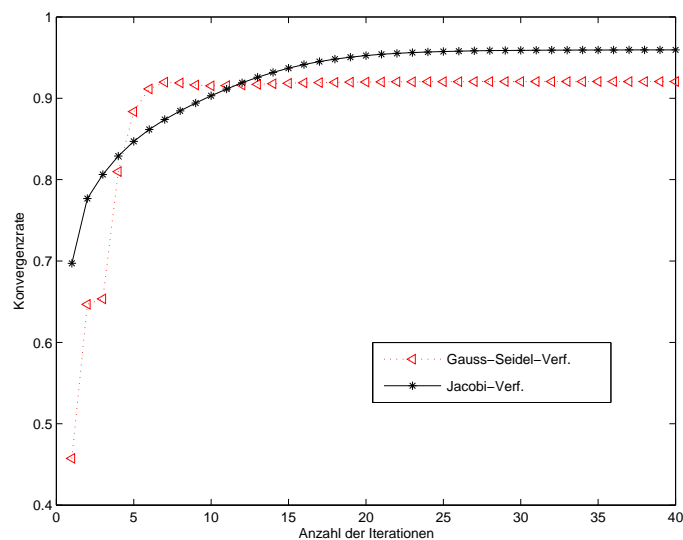


Abb. 4.3: Kontraktionszahlen für Gesamt- und Einzelschrittverfahren.

Dem numerischen Experiment kann man entnehmen, dass in diesem Fall beide Verfahren konvergieren, da die Konvergenzrate jeweils kleiner 1 ist, und dass das Gauß-Seidel-Verfahren schneller konvergiert, da die Konvergenzrate hier kleiner ist.

In der praktischen Anwendung kann man bei dem **Gauß-Seidel-Verfahren** mit  $B = L + D$  folgende Formulierung verwenden, um die Anzahl der Operationen zu reduzieren

$$x^{(k+1)} = (L + D)^{-1}(b - Rx^{(k)}),$$

da allgemein

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - B^{-1}(Ax^{(k)} - b) \\ &= x^{(k)} - B^{-1}([(A - B) + B]x^{(k)} - b) \\ &= x^{(k)} - B^{-1}(A - B)x^{(k)} - B^{-1}Bx^{(k)} + B^{-1}b \\ &= B^{-1}(b - (A - B)x^{(k)}) \end{aligned}$$

gilt und mit  $B = L + D$  und  $A - B = L + D + R - (L + D) = R$  folgt

$$x^{(k+1)} = (L + D)^{-1}(b - Rx^{(k)}).$$

Ein Schritt des Gauß-Seidel-Verfahrens ist also etwa so aufwendig wie eine Matrix-Vektor-Multiplikation. Ähnlich kann man auch für  $B = D + R$  vorgehen, d.h.

$$x^{(k+1)} = (D + R)^{-1}(b - Lx^{(k)}).$$

Schließlich vereinfacht man das **Jacobi-Verfahren** zu

$$x^{(k+1)} = D^{-1}(b - (L + R)x^{(k)})$$

sowie das **Richardson-Verfahren** zu

$$x^{(k+1)} = x^{(k)} + \frac{1}{\gamma}(b - Ax^{(k)}).$$

## 4.2 Konvergenz iterativer Verfahren

Es sei  $x$  Lösung von  $Ax = b$ . Mit (4.6) erhalten wir

$$\begin{aligned} x - x^{(k)} &= x - B^{-1}b - (I - B^{-1}A)x^{(k-1)} \\ &= Ix - B^{-1}Ax - (I - B^{-1}A)x^{(k-1)} \\ &= (I - B^{-1}A)(x - x^{(k-1)}) = \dots = (I - B^{-1}A)^k(x - x^{(0)}). \end{aligned}$$

Die Konvergenz des dargestellten Verfahrens hängt also nur von den Eigenschaften der Iterationsmatrix  $I - B^{-1}A$  ab. Es sei  $C$  eine beliebige komplexwertige  $(n \times n)$ -Matrix,  $\lambda_i := \lambda_i(C)$  ( $i = 1, \dots, n$ ) seien die Eigenwerte von  $C$ . Dann bezeichnen wir mit

$$\varrho(C) := \max_{1 \leq i \leq n} \{|\lambda_i(C)|\}$$

den **Spektralradius** und mit

$$\sigma(C) := \{\lambda_1(C), \dots, \lambda_n(C)\} = \{\lambda \in \mathbb{C} : \exists x \in \mathbb{R}^n \text{ sodass } Ax = \lambda x\} \quad (4.7)$$

das **Spektrum** von  $C$ . Bevor wir ein Konvergenzkriterium angeben, bereiten wir noch den Begriff der Jordanschen Normalform vor.



**Definition 4.2.1 (Jordan-, bzw. Elementarmatrix)** Eine Matrix  $E_k(\lambda) \in \mathbb{C}^{k \times k}$  heißt **Jordan-matrix** (oder **Elementarmatrix**) zum Eigenwert  $\lambda$ , wenn

$$E_k(\lambda) = \begin{pmatrix} \lambda & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{pmatrix}. \quad (4.8)$$

**Satz 4.2.2 (Jordansche Normalform (siehe z.B. [Fischer]))** Zu jeder Matrix  $A \in \mathbb{C}^{n \times n}$  existiert eine reguläre Matrix  $T \in \mathbb{C}^{n \times n}$ , so dass für  $A$  die Ähnlichkeitstransformation

$$A = T^{-1}JT$$

gilt, wobei  $J$ , die durch die Paare  $(\lambda_1, n_1), \dots, (\lambda_k, n_k)$  mit  $\lambda_i \in \mathbb{C}$ ,  $n_i \geq 1$  (eindeutig bis auf die Reihenfolge) bestimmte **Jordansche Normalform**

$$J = \begin{pmatrix} E_{n_1}(\lambda_1) & & 0 \\ & \ddots & \\ 0 & & E_{n_k}(\lambda_k) \end{pmatrix}$$

von  $A$  ist.

**Satz 4.2.3 (Konvergenzkriterium)** Es sei  $C \in \mathbb{C}^{n \times n}$ . Die Folge  $(C^k)_{k \in \mathbb{N}}$  ist genau dann eine Nullfolge, wenn  $\varrho(C) < 1$  gilt.

*Beweis.* Sei zunächst  $\varrho(C) \geq 1$ . Dann gibt es einen Eigenwert  $\lambda$  mit  $|\lambda| \geq 1$  und einen Vektor  $x \neq 0$  mit  $Cx = \lambda x$ . Wegen  $C^k x = \lambda^k x$  und  $\lim_{k \rightarrow \infty} \lambda^k \neq 0$  kann folglich  $(C^k)_{k \in \mathbb{N}}$  keine Nullfolge sein. Die Bedingung  $\varrho(C) < 1$  ist somit notwendig.

Sei nun  $\varrho(C) < 1$ . Weil  $(TCT^{-1})^k = TC^kT^{-1}$  für jede Ähnlichkeitstransformation  $T$  ( $T$  ist regulär) gilt, reicht es,  $\lim_{k \rightarrow \infty} (TCT^{-1})^k = 0$  zu zeigen. Die Matrix  $C$  lässt sich durch Ähnlichkeitstransformation auf die Jordansche Normalform  $J$  transformieren, d.h.  $C = T^{-1}JT$  was äquivalent zu  $TCT^{-1} = J$  ist. Da das Spektrum einer Matrix  $C$  invariant unter einer Ähnlichkeitstransformation ist, wollen wir zeigen, dass  $\lim_{k \rightarrow \infty} J^k = 0$  gilt, wenn alle Eigenwerte  $\lambda_1, \dots, \lambda_n$  dem Betrag nach kleiner Eins sind. Dazu sei  $\mu \in \{1, \dots, n\}$  beliebig und

$$E_\mu = E_{n_\mu}(\lambda_\mu) = \begin{pmatrix} \lambda_\mu & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_\mu \end{pmatrix} \in \mathbb{C}^{n_\mu \times n_\mu}$$

eine Elementarmatrix zum Eigenwert  $\lambda_\mu$  der Jordanschen Normalform  $J$  von  $C$ . Da offenbar

$$J^k = \begin{pmatrix} E_1^k & & \\ & E_2^k & \\ & & \ddots \\ & & & E_\ell^k \end{pmatrix}$$

mit  $1 \leq \ell \leq n$  gilt, genügt es, das Konvergenzverhalten einer Jordanmatrix  $E_\mu$  zu untersuchen. Wir schreiben  $E_\mu$  in der Form  $E_\mu = \lambda_\mu I + S$  mit

$$S = \begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{pmatrix} \in \mathbb{C}^{n_\mu \times n_\mu}$$

und bilden  $E_\mu^k = (\lambda_\mu I + S)^k$ . Wir untersuchen nun den Grenzwert  $\lim_{k \rightarrow \infty} E_\mu^k$ . Nach Anwendung der Binomialentwicklung und unter Beachtung von  $S^{n_\mu} = 0$  erhält man die Beziehung

$$E_\mu^k = \sum_{\nu=0}^k \binom{k}{\nu} \lambda_\mu^{k-\nu} S^\nu = \sum_{\nu=0}^{\min\{k, n_\mu-1\}} \binom{k}{\nu} \lambda_\mu^{k-\nu} S^\nu.$$

Für feste  $\nu$  hat man mit

$$\binom{k}{\nu} = \frac{k!}{\nu!(k-\nu)!} = \frac{k(k-1) \cdot \dots \cdot (k-\nu+1)}{1 \cdot \dots \cdot \nu} \leq k^\nu$$

die Abschätzung

$$\left| \binom{k}{\nu} \lambda_\mu^{k-\nu} \right| \leq |\lambda_\mu^{k-\nu} k^\nu| \leq \exp((k-\nu) \log |\lambda_\mu| + \nu \log k).$$

Hier haben wir verwendet, dass

$$x^a y^b = e^{\log(x^a y^b)} = e^{\log x^a + \log y^b} = e^{a \log x + b \log y}.$$

Da  $|\lambda_\mu| < 1$  ist, strebt

$$k \log |\lambda_\mu| + \nu \log(k) \rightarrow -\infty \text{ für } k \rightarrow \infty$$

und damit folgt die Konvergenz  $\lim_{k \rightarrow \infty} \left| \binom{k}{\nu} \lambda_\mu^{k-\nu} \right| = 0$ . Damit ist gezeigt, dass  $(E_\mu^k)_{k \in \mathbb{N}}$  eine Nullfolge ist und somit auch die Behauptung.  $\square$

Um die Konvergenz der Richardson-Iteration zu beweisen, muss der Spektralradius der Iterationsmatrix bestimmt werden. Während man diesen in dem Spezialfall einer symmetrischen, positiv definiten Matrix  $A$  exakt angeben und somit Konvergenzaussagen treffen kann, vgl. Aufgabe 4.2.4, so ist dies im allgemeinen Fall analytisch nicht möglich und numerisch sehr aufwendig. Es ist daher das Ziel der nächsten beiden Abschnitte, aus einfachen algebraischen Eigenschaften der Matrix  $A$  auf die Konvergenz der Jacobi- bzw. Gauß-Seidel-Iteration zu schließen. Anders als in Satz 4.2.3 sind die Ergebnisse in diesen beiden Abschnitten hinreichende Konvergenzkriterien, im Allgemeinen aber nicht notwendig.

**Aufgabe 4.2.4 (Konvergenz des Richardson-Verfahrens für positiv definite Matrizen)** Es sei  $A \in \mathbb{R}^{n \times n}$  symmetrisch und positiv definit mit  $\lambda_{\min}$  und  $\lambda_{\max}$  als kleinstem bzw. größtem Eigenwert. Man beweise folgende Aussagen:

(a) Für die Iterationsmatrix  $C_R(\gamma) = I - \gamma^{-1}A$  des Richardson-Verfahrens gilt

$$\varrho(C_R(\gamma)) = \max \{ |1 - \gamma^{-1} \lambda_{\max}|, |1 - \gamma^{-1} \lambda_{\min}| \} \quad \forall \gamma \in \mathbb{R} \setminus \{0\}.$$

(b) Das Richardson-Verfahren konvergiert genau dann, wenn  $\gamma > \frac{\lambda_{\max}}{2}$  gilt.

(c)  $\gamma^* := \frac{\lambda_{\max} + \lambda_{\min}}{2}$  minimiert den Spektralradius  $\varrho(C_G(\gamma))$  für  $\gamma \in \mathbb{R} \setminus \{0\}$ .

(d) Es gilt  $\varrho(C_G(\gamma^*)) = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}$ .

## 4.3 Konvergenz des Jacobi-Verfahrens

Bei dem **Jacobi-Verfahren** (auch Gesamtschrittverfahren genannt) werden alle Komponenten des Lösungsvektors in einem Iterationsschritt gleichzeitig korrigiert. Mit  $B = D$  lautet die Iterationsvorschrift

$$x^{(k+1)} = D^{-1}(b - (L + R)x^{(k)}),$$

d.h. die Iterationsmatrix ist  $C_J = I - D^{-1}A = -D^{-1}(L + R)$  und es gilt  $x^{(k+1)} - x = C_J(x - x^{(k)})$ .

**Satz 4.3.1 (Starkes Zeilen- und Spaltensummenkriterium)** Es sei  $A \in \mathbb{C}^{n \times n}$ . Das Jacobi-Verfahren konvergiert für jeden Startvektor  $x^{(0)} \in \mathbb{C}^n$ , wenn für die Matrix  $A$  das

i) starke Zeilensummenkriterium:

$$|a_{ii}| > \sum_{\substack{k=1 \\ k \neq i}}^n |a_{ik}|, \quad \text{für } i = 1, 2, \dots, n,$$

d.h.  $A$  ist strikt diagonaldominant, oder das

ii) starke Spaltensummenkriterium:

$$|a_{kk}| > \sum_{\substack{i=1 \\ i \neq k}}^n |a_{ik}|, \quad \text{für } k = 1, 2, \dots, n,$$

d.h.  $A^T$  ist strikt diagonaldominant,

erfüllt ist.

Für den Beweis von Satz 4.3.1 benötigen wir das folgende Lemma.

**Lemma 4.3.2** Es sei  $A \in \mathbb{C}^{n \times n}$ . Dann gilt für jede natürliche  $p$ -Matrixnorm  $\varrho(A) \leq \|A\|_p$ .

*Beweis.* Wir betrachten das Spektrum der Matrix  $A$ ,  $\sigma(A) = \{\lambda \in \mathbb{C} : \exists x \neq 0, Ax = \lambda x\}$ . Jeder Eigenwert  $\lambda \in \sigma(A)$  von  $A$  mit zugehörigem Eigenvektor  $x$  genügt für jede natürliche  $p$ -Matrixnorm  $\|\cdot\|_p$  der Beziehung  $\|Ax\|_p = \|\lambda x\|_p = |\lambda| \|x\|_p$  und damit  $\frac{\|Ax\|_p}{\|x\|_p} = |\lambda|$ . D.h., es gilt dass

$$\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} \geq \max_{\lambda \in \sigma(A)} |\lambda| = \varrho(A).$$

□

**Bemerkung 4.3.3** Lemma 4.3.2 gilt allgemeiner auch, wenn die  $p$ -Matrixnorm von  $A$ ,  $\|A\|_p$ , durch eine Operatornorm von  $A$ ,  $\|A\|$ , ersetzt wird (s.a. Bemerkung B.2.3).

*Beweis von Satz 4.3.1.* i) Die Iterationsmatrix des Jacobi-Verfahrens ist

$$C_J = I - D^{-1}A = -D^{-1}(L + R).$$

Wenn das starke Zeilensummenkriterium erfüllt ist, gilt die Abschätzung

$$\|C_J\|_\infty = \left\| \begin{pmatrix} 0 & -\frac{a_{1,2}}{a_{1,1}} & \dots & \dots & -\frac{a_{1,n}}{a_{1,1}} \\ -\frac{a_{2,1}}{a_{2,2}} & 0 & -\frac{a_{2,3}}{a_{2,2}} & \dots & -\frac{a_{2,n}}{a_{2,2}} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ -\frac{a_{n,1}}{a_{n,n}} & \dots & \dots & -\frac{a_{n,n-1}}{a_{n,n}} & 0 \end{pmatrix} \right\|_\infty = \max_{1 \leq i \leq n} \sum_{\substack{j=1 \\ j \neq i}}^n \frac{|a_{ij}|}{|a_{ii}|} < 1.$$

Lemma 4.3.2 liefert dann die Behauptung i).

ii) Ist für  $A$  das starke Spaltensummenkriterium (ii) erfüllt, so gilt (i) für  $A^T$ . Also konvergiert das Jacobi-Verfahren für  $A^T$  und es ist daher wegen Satz 4.2.3  $\varrho(C) < 1$  für  $C = I - D^{-1}A^T$ . Nun hat  $C$  die gleichen Eigenwerte wie  $C^T$  und wie  $D^{-1}C^TD = I - D^{-1}A = C_J$ . Also ist auch  $\varrho(C_J) < 1$ , d.h. das Jacobi-Verfahren ist auch für die Matrix  $A$  konvergent.  $\square$

**Bemerkung 4.3.4** Die Standardmatrix aus Beispiel 4.1.1 (s.a. Beispiel 3.0.1) erfüllt weder das starke Zeilensummenkriterium noch das starke Spaltensummenkriterium.

**Definition 4.3.5** Eine Matrix  $A \in \mathbb{R}^{n \times n}$  heißt **zerlegbar (reduzibel)**, wenn es nichtleere Teilmengen  $N_1$  und  $N_2$  der Indexmenge  $N := \{1, 2, \dots, n\}$  gibt mit den Eigenschaften

- i)  $N_1 \cap N_2 = \emptyset$ ;
- ii)  $N_1 \cup N_2 = N$ ;
- iii)  $a_{ij} = 0$  für alle  $i \in N_1$  und  $j \in N_2$ .

Eine Matrix, die nicht zerlegbar ist, heißt **unzerlegbar (irreduzibel)**.

**Beispiel 4.3.6** i)

$$A = \begin{pmatrix} a_{11} & \dots & a_{1k} & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{N1} & \dots & a_{NN} & 0 & \dots & 0 \\ a_{N+1,1} & \dots & a_{N+1,N} & a_{N+1,N+1} & \dots & a_{N+1,2N} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{2N,1} & \dots & a_{2N,N} & a_{2N,N+1} & \dots & a_{2N,2N} \end{pmatrix}$$

Die Teilmengen  $N_1 = \{1, 2, \dots, N\}$ ,  $N_2 = \{N+1, \dots, 2N\}$  haben alle in der Definition geforderten Eigenschaften. Somit ist  $A$  zerlegbar.

- ii) Eine Tridiagonalmatrix mit nicht verschwindenden Nebendiagonal- und Diagonalelementen, wie zum Beispiel die Standardmatrix, ist unzerlegbar.

**Bemerkung 4.3.7** Dass eine Matrix  $A$  unzerlegbar (irreduzibel) ist, kann man häufig leicht mit Hilfe des der Matrix  $A$  zugeordneten (gerichteten) Graphen  $G(A)$  zeigen. Wenn  $A$  eine  $n \times n$ -Matrix ist, so besteht  $G(A)$  aus  $n$  Knoten  $K_1, \dots, K_n$  und es gibt eine gerichtete Kante  $K_i \rightarrow K_j$  in  $G(A)$  genau dann, wenn  $a_{ij} \neq 0$ . Man zeigt leicht, dass  $A$  genau dann unzerlegbar ist, falls der Graph  $G(A)$  in dem Sinne zusammenhängend ist, dass es für jedes Knotenpaar  $(K_i, K_j)$  in  $G(A)$  einen gerichteten Weg  $K_i$  nach  $K_j$  gibt.

$$A = \begin{pmatrix} 2 & -1 & 0 \\ 1 & 4 & 0 \\ 0 & -1 & 3 \end{pmatrix}, \quad G(A) : \quad \begin{array}{ccccc} & & \curvearrowright & & \\ & & K_1 & \curvearrowright & K_2 & \curvearrowright & K_3 \\ & & & \curvearrowleft & & & \\ & & & & & \curvearrowleft & \end{array}$$

Abb. 4.4: Beispiel einer zerlegbaren Matrix  $A$  und ihres Graphen  $G(A)$ .

**Definition 4.3.8 (Schwachere Zeilen- und Spaltensummenkriterium)** Eine Matrix  $A \in \mathbb{R}^{n \times n}$  erfüllt das **schwache Zeilensummenkriterium**, falls

$$\sum_{\substack{\nu=1 \\ \nu \neq \mu}}^n |a_{\mu\nu}| \leq |a_{\mu\mu}|$$

für alle Zeilen  $\mu = 1, \dots, n$  gilt, d.h.  $A$  ist **diagonaldominant**, und

$$\sum_{\substack{\nu=1 \\ \nu \neq \mu_0}}^n |a_{\mu_0\nu}| < |a_{\mu_0\mu_0}|$$

für mindestens einen Index  $\mu_0 \in \{1, \dots, n\}$  erfüllt ist.

Eine Matrix  $A \in \mathbb{R}^{n \times n}$  erfüllt das **schwache Spaltensummenkriterium**, wenn  $A^T$  das schwache Zeilensummenkriterium erfüllt.

**Satz 4.3.9 (Schwachere Zeilensummenkriterium)** Es sei  $A \in \mathbb{R}^{n \times n}$  eine irreduzible Matrix, die das schwache Zeilensummenkriterium erfüllt. Dann ist das Jacobi-Verfahren konvergent.

**Bemerkung 4.3.10** Die Standardmatrix aus Beispiel 4.1.1 (s.a. Beispiel 3.0.1) erfüllt das schwache Zeilensummenkriterium sowie das schwache Spaltensummenkriterium.

Zum Beweis von Satz 4.3.9 werden wir direkt den Spektralradius der Iterationsmatrix abschätzen. Die wesentliche Beobachtung dabei ist, dass jede irreduzible Matrix, die das schwache Zeilensummenkriterium erfüllt, bereits regulär ist.

**Lemma 4.3.11** Jede irreduzible Matrix  $A \in \mathbb{R}^{n \times n}$ , die das schwache Zeilensummenkriterium erfüllt, ist regulär und für die Diagonalelemente gilt  $a_{jj} \neq 0$  ( $j = 1, \dots, n$ ).

*Beweis.* Wir nehmen an,  $A$  sei nicht regulär, d.h. es existiert ein  $x \in \mathbb{K}^n \setminus \{0\}$  mit  $Ax = 0$ . Insbesondere folgt aus der Dreiecksungleichung, dass für alle  $j = 1, \dots, n$  gilt:

$$|a_{jj}| |x_j| = |a_{jj}x_j| = \left| \sum_{\ell=1}^n a_{j\ell}x_\ell - \sum_{\substack{\ell=1 \\ \ell \neq j}}^n a_{j\ell}x_\ell \right| \leq \underbrace{\left| \sum_{\ell=1}^n a_{j\ell}x_\ell \right|}_{(Ax)_j=0} + \left| \sum_{\substack{\ell=1 \\ \ell \neq j}}^n a_{j\ell}x_\ell \right| \leq \sum_{\substack{\ell=1 \\ \ell \neq j}}^n |a_{j\ell}| |x_\ell|. \quad (4.9)$$

Wir definieren die Indextmengen  $J := \{j : |x_j| = \|x\|_\infty\}$  und  $K := \{k : |x_k| < \|x\|_\infty\}$ . Offensichtlich gilt  $J \cap K = \emptyset$ ,  $J \cup K = \{1, \dots, n\}$  und  $J \neq \emptyset$ . Wäre  $K = \emptyset$ , so könnte man in (4.9) die  $x_j$ - und  $x_\ell$ -Terme herauskürzen, denn  $|x_j| = |x_k|$  für all  $k = 1, \dots, n$ , und erhielte einen Widerspruch zum schwachen Zeilensummenkriterium. Also gilt  $K \neq \emptyset$ , und aufgrund der Irreduzibilität von  $M$  existieren Indizes  $j \in J$  und  $k \in K$  mit  $a_{jk} \neq 0$ . Mit (4.9) ergibt sich

$$|a_{jj}| \leq \sum_{\substack{\ell=1 \\ \ell \neq j}}^n |a_{j\ell}| \frac{|x_\ell|}{|x_j|} < \sum_{\substack{\ell=1 \\ \ell \neq j}}^n |a_{j\ell}|,$$

denn der Quotient ist stets  $\leq 1$  wegen  $|x_j| = \|x\|_\infty$  und er ist  $< 1$  für  $\ell \in K \neq \emptyset$  (also zumindest für  $\ell = k$ ). Also erhalten wir einen Widerspruch zum schwachen Zeilensummenkriterium von  $A$ , d.h.  $A$  ist regulär.

Gäbe es schließlich ein triviales Diagonalelement  $a_{jj} = 0$ , so folgte aus dem schwachen Zeilensummenkriterium, dass bereits die  $j$ -te Zeile die Nullzeile wäre. Da  $A$  regulär ist, folgt insbesondere  $a_{jj} \neq 0$  für alle  $j = 1, \dots, n$ .  $\square$

*Beweis von Satz 4.3.9.* Wegen  $a_{jj} \neq 0$  für alle  $j = 1, \dots, n$  ist  $C_J = -D^{-1}(A - D)$  wohldefiniert. Um  $\varrho(C_J) < 1$  zu zeigen, beweisen wir, dass  $M := C_J - \lambda I$  für  $\lambda \in \mathbb{C}$  mit  $|\lambda| \geq 1$  regulär ist.

Unter der Annahme, dass  $M$  regulär ist, gilt  $\varrho(M) > 0$ . Sei nun  $\mu \in \sigma(C_J)$  und  $x$  der zugehörige Eigenvektor. Dann gilt  $Mx = C_J x - \lambda x = (\mu - \lambda)x$ , d.h.,  $\mu - \lambda$  ist ein Eigenwert von  $M$ . Da  $\varrho(M) > 0$ , gilt also  $|\mu - \lambda| > 0$ . Würde nun  $\varrho(C_J) \geq 1$  gelten, dann gibt es wenigstens ein  $\mu \in \sigma(C_J)$  mit  $|\mu| \geq 1$  und wir können  $\lambda = \mu$  wählen, was zu einem Widerspruch führt. Also muss  $\varrho(C_J) < 1$  gelten.

Wir zeigen nun, dass  $M$  regulär ist. Da  $A$  irreduzibel ist, ist auch  $A - D$  irreduzibel, denn es wird lediglich die Diagonale verändert.  $C_J$  entsteht durch zeilenweise Multiplikation von  $A - D$  mit Werten  $\neq 0$ . Deshalb ist auch  $C_J$  irreduzibel. Da  $M$  und  $C_J$  sich nur auf der Diagonale unterscheiden, ist  $M$  irreduzibel. Aufgrund des schwachen Zeilensummenkriteriums von  $A$  gilt

$$\sum_{\substack{k=1 \\ k \neq j}}^n |m_{jk}| = \sum_{\substack{k=1 \\ k \neq j}}^n |c_{jk}^{(J)}| = \sum_{\substack{k=1 \\ k \neq j}}^n \frac{|a_{jk}|}{|a_{jj}|} \leq 1 \leq |\lambda| = |m_{jj}| \quad \text{für alle } j = 1, \dots, n.$$

und für mindestens einen Index  $j$  gilt diese Ungleichung strikt. Also erfüllt  $M$  auch das schwache Zeilensummenkriterium und ist nach Lemma 4.3.11 insgesamt regulär.  $\square$

**Aufgabe 4.3.12** Man zeige, dass das Jacobi-Verfahren auch konvergent ist unter der Annahme, dass  $A \in \mathbb{R}^{n \times n}$  irreduzibel ist und das schwache Spaltensummenkriterium erfüllt.

## 4.4 Konvergenz des Gauß-Seidel-Verfahrens

Die Iterationsvorschrift des **Gauß-Seidel-Verfahrens** (auch Einzelschrittverfahren genannt) für  $B = L + D$  lautet

$$x^{(k+1)} = (L + D)^{-1}(b - Rx^{(k)}),$$

d.h. die Iterationsmatrix ist  $C_{GS} := -(L + D)^{-1}R$ ;

die Iterationsvorschrift des Gauß-Seidel-Verfahrens für  $B = D + R$  lautet

$$x^{(k+1)} = (D + R)^{-1}(b - Lx^{(k)}),$$

mit Iterationsmatrix  $\tilde{C}_{GS} := -(D + R)^{-1}L$ .

**Satz 4.4.1 (Konvergenzsatz)** Es sei  $A \in \mathbb{R}^{n \times n}$  eine reguläre Matrix, die das starke Zeilensummenkriterium oder das starke Spaltensummenkriterium erfüllt. Dann sind beide Varianten des Gauß-Seidel-Verfahrens zur Lösung des linearen Gleichungssystems  $Ax = b$  konvergent.

*Beweis.* Sei das starke Zeilensummenkriterium erfüllt. Die Iterationsmatrizen des Gauß-Seidel-Verfahrens mit  $B = L + D$  bzw. des Jacobi-Verfahrens sind  $C_{GS} := -(L + D)^{-1}R$  bzw.

$C_J := -D^{-1}(L + R)$ . Wenn das starke Zeilensummenkriterium erfüllt ist, gilt die Abschätzung

$$\|C_J\|_\infty = \max_{1 \leq i \leq n} \sum_{\substack{j=1 \\ j \neq i}}^n \frac{|a_{ij}|}{|a_{ii}|} < 1.$$

Es sei jetzt  $y \in \mathbb{R}^n$  beliebig und  $z = C_{GS} y$ . Durch vollständige Induktion beweisen wir, dass alle Komponenten  $z_i$  des Vektors  $z$  der Abschätzung

$$|z_i| \leq \sum_{\substack{j=1 \\ j \neq i}}^n \frac{|a_{ij}|}{|a_{ii}|} \|y\|_\infty$$

genügen. Dazu schreiben wir die Gleichung  $z = C_{GS} y$  in  $-(L + D)z = Ry$  um und schätzen ab:

$$|z_1| \leq \sum_{j=2}^n \frac{|a_{1j}|}{|a_{11}|} |y_j| \leq \sum_{j=2}^n \frac{|a_{1j}|}{|a_{11}|} \|y\|_\infty.$$

Schreiben wir das Gauß-Seidel-Verfahren mit  $B = L + D$  in der Form

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^n a_{ij} x_j^{(k)} \right) \quad (1 \leq i \leq n),$$

so folgt daraus dann mit der Induktionsvoraussetzung

$$\begin{aligned} |z_i| &\leq \frac{1}{|a_{ii}|} \left( \sum_{j=1}^{i-1} |a_{ij}| |z_j| + \sum_{j=i+1}^n |a_{ij}| |y_j| \right) \\ &\stackrel{\text{IH}}{\leq} \frac{1}{|a_{ii}|} \left( \sum_{j=1}^{i-1} |a_{ij}| \|C_J\|_\infty + \sum_{j=i+1}^n |a_{ij}| \right) \|y\|_\infty \leq \sum_{\substack{j=1 \\ j \neq i}}^n \frac{|a_{ij}|}{|a_{ii}|} \|y\|_\infty. \end{aligned}$$

Hiermit erhält man die Abschätzung

$$\|C_{GS} y\|_\infty = \|z\|_\infty \leq \|C_J\|_\infty \|y\|_\infty \quad \forall y \in \mathbb{R}^n$$

und somit

$$\|C_{GS}\|_\infty \leq \|C_J\|_\infty < 1. \quad (4.10)$$

Da  $\varrho(-(L + D)^{-1}R) = \varrho(C_{GS}) \leq \|C_{GS}\|_\infty$ , folgt daraus die Konvergenz des Gauß-Seidel-Verfahrens für  $B = L + D$ .

Um die Konvergenz des Gauß-Seidel-Verfahrens für  $B = D + R$  mit  $\tilde{C}_{GS} = -(D + R)^{-1}L$  nachzuweisen, betrachten wir zunächst folgende Permutationsmatrix

$$P = \begin{pmatrix} 0 & & 1 \\ & \ddots & \\ 1 & & 0 \end{pmatrix} = P^T \in \mathbb{R}^{n \times n}.$$

Die Matrix  $\tilde{A} = PAP^T$  geht aus  $A$  durch simultane Zeilen- und Spaltenvertauschungen hervor, sodass die Gültigkeit des starken Zeilensummenkriteriums auch für  $\tilde{A}$  vorliegt. Mit dem oben Bewiesenen gilt somit  $\varrho(-(\tilde{L} + \tilde{D})^{-1}\tilde{R}) < 1$ , wobei

$$\begin{aligned} \tilde{L} &= PRP^T \\ \tilde{D} &= PDP^T \\ \tilde{R} &= PLP^T \end{aligned}$$

und daher

$$\begin{aligned} 1 &> \varrho(-(PRP^T + PDP^T)^{-1}PLP^T) = \varrho(-(P(R + D)P^T)^{-1}PLP^T) \\ &= \varrho(-P(R + D)^{-1}P^T PLP^T) = \varrho(-P(R + D)^{-1}LP^T) \\ &= \varrho(-(R + D)^{-1}L). \end{aligned}$$

Also ist  $\varrho(\tilde{C}_{GS}) = \varrho(-(D + R)^{-1}L) < 1$  und somit das Gauß-Seidel-Verfahren für die Wahl  $B = D + R$  konvergent.

Sei nun das starke Spaltensummenkriterium erfüllt. Dann erfüllt  $A^T$  das starke Zeilensummenkriterium und beide Varianten des Gauß-Seidel-Verfahrens konvergieren für  $A^T$ . Mit der Standardzerlegung von  $A^T$

$$A^T = L_T + D_T + R_T,$$

wobei  $L_T = R^T$ ,  $D_T = D$  und  $R_T = L^T$  ist, gilt somit

$$\begin{aligned} \varrho(-(R^T + D)^{-1}L^T) &= \varrho(-(L_T + D_T)^{-1}R_T) < 1, \\ \varrho(-(L^T + D)^{-1}R^T) &= \varrho(-(R_T + D_T)^{-1}L_T) < 1. \end{aligned}$$

Hieraus ergibt sich die Konvergenz des Gauß-Seidel-Verfahrens für  $B = L + D$

$$\begin{aligned} \varrho(C_{GS}) &= \varrho(-(L + D)^{-1}R) = \varrho(-(L + D)(L + D)^{-1}R(L + D)^{-1}) \\ &= \varrho(-R(L + D)^{-1}) = \varrho(-(L^T + D)^{-1}R^T) < 1 \end{aligned}$$

sowie für  $B = D + R$

$$\begin{aligned} \varrho(\tilde{C}_{GS}) &= \varrho(-(D + R)^{-1}L) = \varrho(-(D + R)(D + R)^{-1}L(D + R)^{-1}) \\ &= \varrho(-L(D + R)^{-1}) = \varrho(-(D + R^T)^{-1}L^T) < 1. \end{aligned}$$

Damit sind alle Aussagen des Satzes bewiesen. □



**Bemerkung 4.4.2** Häufig verleitet (4.10) zu der falschen Schlussfolgerung, dass das Gauß-Seidel-Verfahren schneller als das Jacobi-Verfahren konvergiert, wenn die Matrix strikt diagonaldominant ist.

**Beispiel 4.4.3** Dass bei strikter Diagonaldominanz einer regulären Matrix  $A \in \mathbb{R}^{n \times n}$  aus  $\|C_{GS}\|_\infty \leq \|C_J\|_\infty < 1$  nicht  $\varrho(C_{GS}) \leq \varrho(C_J)$  folgen muss, sieht man, wenn man die Matrix  $A$  folgendermaßen wählt:

$$A = \begin{pmatrix} 50 & -10 & -20 \\ -20 & 49 & -20 \\ 20 & -10 & 49 \end{pmatrix}.$$

Dann besitzen die zugehörigen Iterationsmatrizen

$$C_{GS} = -(L + D)^{-1}R = \frac{1}{12005} \begin{pmatrix} 0 & 2401 & 4802 \\ 0 & 980 & 6860 \\ 0 & -780 & -560 \end{pmatrix}, \quad C_J = \frac{1}{245} \begin{pmatrix} 0 & 49 & 98 \\ 100 & 0 & 100 \\ -100 & 50 & 0 \end{pmatrix}$$

nämlich die Spektralradien  $\varrho(C_{GS}) = (4\sqrt{5})/49 \approx 0.1825$  und  $\varrho(C_J) = 2/49 \approx 0.04082$  sowie die Maximumnormen  $\|C_{GS}\|_\infty = 32/49 \approx 0.6531$  und  $\|C_J\|_\infty = 40/49 \approx 0.8163$ .

**Bemerkung 4.4.4** Mit Bezug auf das letzte Beispiel halten wir fest: Ist eine Matrix  $A$  strikt diagonaldominant, gilt für die Iterationsmatrizen  $\|C_{GS}\|_\infty \leq \|C_J\|_\infty < 1$ , es folgt im Allgemeinen aber nicht  $\varrho(C_{GS}) \leq \varrho(C_J)$ .



**Bemerkung 4.4.5** Ebenfalls wäre die Schlussfolgerung aus Satz 4.4.1, dass eine Form des Gauß-Seidel-Verfahrens genau dann konvergent ist, wenn es die andere ist, falsch. Es gibt Beispiele regulärer Matrizen, für die  $\varrho(C_{GS}) < 1$ , aber  $\varrho(\tilde{C}_{GS}) \geq 1$  bzw.  $\varrho(\tilde{C}_{GS}) < 1$ , aber  $\varrho(C_{GS}) \geq 1$ .

Man betrachte dazu folgendes Beispiel.

**Beispiel 4.4.6** Gegeben sei die reguläre Matrix

$$A = \begin{pmatrix} 2 & 0 & 2 \\ 2 & 2 & 2 \\ 0 & 2 & -1 \end{pmatrix}.$$

Die zugehörigen Iterationsmatrizen

$$C_{GS} = -(L + D)^{-1}R = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$\tilde{C}_{GS} = -(D + R)^{-1}L = \begin{pmatrix} 0 & -2 & 0 \\ -1 & -2 & 0 \\ 0 & 2 & 0 \end{pmatrix}$$

besitzen die Spektralradien  $\varrho(C_{GS}) = 0 < 1$  sowie  $\varrho(\tilde{C}_{GS}) = 1 + \sqrt{3} > 1$ , d.h. das Gauß-Seidel-Verfahren für  $B = L + D$  ist konvergent, für  $B = D + R$  jedoch divergent.

**Satz 4.4.7 (Schwaches Zeilensummenkriterium)** Ist  $A \in \mathbb{R}^{n \times n}$  irreduzibel und erfüllt das schwache Zeilensummenkriterium, so sind beide Varianten des Gauß-Seidel-Verfahrens konvergent.

*Beweis.* Die Wohldefiniertheit von  $C_{GS} = -(L + D)^{-1}R$  und  $\tilde{C}_{GS} = -(D + R)^{-1}L$  ist wieder klar. Wir betrachten  $W := C_{GS} - \lambda I$  sowie  $\tilde{W} := \tilde{C}_{GS} - \lambda I$  für  $\lambda \in \mathbb{C}$  mit  $|\lambda| \geq 1$ . Durch Multiplikation mit  $-(L + D)$  sieht man, dass  $W$  genau dann regulär ist, wenn  $M := R + \lambda L + \lambda D$  regulär ist. Analog folgt durch Multiplikation mit  $-(D + R)$ , dass  $\tilde{W}$  genau dann regulär ist, wenn  $\tilde{M} := L + \lambda D + \lambda R$  es ist. Offensichtlich erben  $M$  und  $\tilde{M}$  die Irreduzibilität von  $A = D + L + R$ . Ferner erfüllen  $M$  und  $\tilde{M}$  das schwache Zeilensummenkriterium, denn es gilt

$$\sum_{\substack{k=1 \\ k \neq j}}^n |m_{jk}| = |\lambda| \sum_{k=1}^{j-1} |a_{jk}| + \sum_{k=j+1}^n |a_{jk}| \leq |\lambda| \sum_{\substack{k=1 \\ k \neq j}}^n |a_{jk}| \leq |\lambda| |a_{jj}| = |m_{jj}|$$

sowie

$$\sum_{\substack{k=1 \\ k \neq j}}^n |\tilde{m}_{jk}| = \sum_{k=1}^{j-1} |a_{jk}| + |\lambda| \sum_{k=j+1}^n |a_{jk}| \leq |\lambda| \sum_{\substack{k=1 \\ k \neq j}}^n |a_{jk}| \leq |\lambda| |a_{jj}| = |\tilde{m}_{jj}|$$

für  $j = 1, \dots, n$  mit strikter Ungleichung jeweils für mindestens einen Index  $j$ . Nach Lemma 4.3.11 sind  $M$  und  $\tilde{M}$  regulär. Insgesamt erhalten wir wie zuvor  $\varrho(C_{GS}) < 1$  und  $\varrho(\tilde{C}_{GS}) < 1$ .  $\square$

**Aufgabe 4.4.8** Man zeige, dass beide Formen des Gauß-Seidel-Verfahrens auch konvergent sind unter der Annahme, dass  $A \in \mathbb{R}^{n \times n}$  irreduzibel ist und das schwache Spaltensummenkriterium erfüllt.

**Beispiel 4.4.9** Es sei  $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$  mit  $a, b, c, d \in \mathbb{C}$ . Die zugehörige Iterationsmatrix zum Jacobi-Verfahren lautet somit

$$C_J = -D^{-1}(L + R) = -\begin{pmatrix} a^{-1} & 0 \\ 0 & d^{-1} \end{pmatrix} \begin{pmatrix} 0 & b \\ c & 0 \end{pmatrix} = \begin{pmatrix} 0 & -\frac{b}{a} \\ -\frac{c}{d} & 0 \end{pmatrix}.$$

Das charakteristische Polynom hierzu lautet  $p(\lambda) = \lambda^2 - \frac{bc}{ad}$  und hat Nullstellen  $\lambda_{1,2} = \pm \sqrt{\frac{bc}{ad}}$ . Entsprechend erhält man für die Gauß-Seidel-Verfahren

$$C_{GS} = -(L + D)^{-1}R = -\frac{1}{ad} \begin{pmatrix} d & 0 \\ -c & a \end{pmatrix} \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -\frac{bd}{ad} \\ 0 & \frac{bc}{ad} \end{pmatrix},$$

$$\tilde{C}_{GS} = -(D + R)^{-1}L = -\frac{1}{ad} \begin{pmatrix} d & -b \\ 0 & a \end{pmatrix} \begin{pmatrix} 0 & 0 \\ c & 0 \end{pmatrix} = \begin{pmatrix} \frac{bc}{ad} & 0 \\ -\frac{ac}{ad} & 0 \end{pmatrix}.$$

In beiden Fällen lautet das charakteristische Polynom  $p(\lambda) = \lambda(\lambda - bc/ad)$  und hat Nullstellen

$$\lambda_1 = 0, \quad \lambda_2 = \frac{bc}{ad}.$$

womit

$$\varrho(C_J) = \sqrt{\frac{|bc|}{|ad|}} \quad \text{und} \quad \varrho(C_{GS}) = \varrho(\tilde{C}_{GS}) = \frac{|bc|}{|ad|}$$

gilt. Man beachte  $\|C_J\|_1 = \|C_J\|_\infty = \max\{|b/a|, |c/d|\}$  sowie

$$\|C_{GS}\|_1 = \frac{|b|(|c| + |d|)}{|ad|}, \quad \|C_{GS}\|_\infty = \frac{|b| \max\{|c|, |d|\}}{|ad|},$$

$$\|\tilde{C}_{GS}\|_1 = \frac{|c|(|a| + |b|)}{|ad|}, \quad \|\tilde{C}_{GS}\|_\infty = \frac{|c| \max\{|a|, |b|\}}{|ad|}.$$

Wir können somit für  $A \in \mathbb{R}^{2 \times 2}$  festhalten: Konvergiert das Jacobi- oder Gauß-Seidel-Verfahren, so konvergiert auch das jeweilige andere Verfahren. Und im Falle der Konvergenz, konvergiert das Gauß-Seidel doppelt so schnell wie das Jacobi-Verfahren. Die Frage ist nun: Gilt dies immer, bzw. kann dies ggf. einfach charakterisiert werden?

**Aufgabe 4.4.10** Man nutze das letzte Beispiel, um zu zeigen, dass aus  $\|C_J\|_1 < 1$  im Allgemeinen

$$\|C_{GS}\|_1 \leq \|C_J\|_1 < 1$$

nicht folgt. (Z.B.  $a = d = 1$ ,  $b = c = 2/3$  in der Matrix von Beispiel 4.4.9 liefert ein Gegenbeispiel).

**Definition 4.4.11** Eine Matrix  $A \in \mathbb{R}^{m \times n}$  heißt **nichtnegativ**, wenn alle Koeffizienten  $a_{ij}$  von  $A$  nichtnegativ sind.

**Satz 4.4.12 (von Stein und Rosenberg)** Die Iterationsmatrix  $C_J \in \mathbb{R}^{n \times n}$  des Jacobi-Verfahrens sei nichtnegativ. Dann gilt genau eine der folgenden Aussagen:

- i)  $\varrho(C_J) = \varrho(C_{GS}) = 0$
- ii)  $\varrho(C_J) = \varrho(C_{GS}) = 1$
- iii)  $0 < \varrho(C_{GS}) < \varrho(C_J) < 1$
- iv)  $1 < \varrho(C_J) < \varrho(C_{GS})$

*Beweis.* Siehe [Hämmerlin/Hoffmann]. □

Die Voraussetzung  $C_J \geq 0$  ist insbesondere dann erfüllt, wenn die Matrix  $A$  positive Diagonalelemente und nichtpositive Nichtdiagonalelemente besitzt, d.h.  $a_{ii} > 0$ ,  $a_{ik} \leq 0$  für  $i \neq k$ . Dieser Fall liegt auch im folgenden Beispiel vor.

#### Beispiel 4.4.13

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix} \Rightarrow C_J = -D^{-1}(L + R) = \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 0 \end{pmatrix}$$

Die Iterationsmatrix ist nichtnegativ,  $\rho(C_J) = \frac{1+\sqrt{5}}{4} \approx 0.809 < 1$  und somit folgt, dass das Gauß-Seidel-Verfahren schneller ist als das Jacobi-Verfahren!

**Bemerkung 4.4.14** Dass das Gauß-Seidel-Verfahren nicht immer besser sein muss als das Jacobi-Verfahren oder aus der Divergenz des Jacobi-Verfahrens nicht auch die Divergenz des Gauß-Seidel-Verfahrens folgen muss, zeigen die folgenden beiden Beispiele.

#### Beispiel 4.4.15 (Jacobi- immer schlechter als Gauß-Seidel-Verfahren?)

i) Die Iterationsmatrizen  $C_J$  bzw.  $C_{GS}$  zur Matrix

$$A = \begin{pmatrix} 1 & 2 & -2 & 2 \\ 1 & 1 & 1 & 0 \\ 2 & 2 & 1 & 2 \\ -1 & -2 & 1 & 1 \end{pmatrix}$$

lauten

$$C_J = \begin{pmatrix} 0 & -2 & 2 & -2 \\ -1 & 0 & -1 & 0 \\ -2 & -2 & 0 & -2 \\ 1 & 2 & -1 & 0 \end{pmatrix} \quad \text{bzw.} \quad C_{GS} = -(L + D)^{-1}R = \begin{pmatrix} 0 & -2 & 2 & -2 \\ 0 & 2 & -3 & 2 \\ 0 & 0 & 2 & -2 \\ 0 & 2 & -6 & 4 \end{pmatrix}$$

mit den Spektralradien  $\rho(C_J) = 0$  und  $\rho(C_{GS}) \approx 7.3850$ .

ii) Die Iterationsmatrizen  $C_J$  bzw.  $C_{GS}$  zur Matrix

$$A = \frac{1}{3} \begin{pmatrix} 3 & -2 & -1 & 1 \\ 1 & 2 & -2 & 1 \\ 1 & -2 & 2 & 2 \\ 1 & -2 & 1 & 1 \end{pmatrix}$$

lauten

$$C_J = \frac{1}{6} \begin{pmatrix} 0 & 4 & 2 & -2 \\ -3 & 0 & 6 & -3 \\ -3 & 6 & 0 & -6 \\ -6 & 12 & -6 & 0 \end{pmatrix} \quad \text{bzw.} \quad C_{GS} = -(L + D)^{-1}R = \frac{1}{6} \begin{pmatrix} 0 & 4 & 2 & -2 \\ 0 & -2 & 5 & -2 \\ 0 & -4 & 4 & -7 \\ 0 & -4 & 4 & 5 \end{pmatrix}$$

mit den Spektralradien  $\rho(C_J) \approx 1.4527$  und  $\rho(C_{GS}) \approx 0.9287 < 1$ .

**Bemerkungen 4.4.16** i) Die obigen Iterationsverfahren ließen sich in der Form

$$x^{(k+1)} = B^{-1}(B - A)x^{(k)} + B^{-1}b = Cx^{(k)} + d$$

schreiben. Da die Iterationsmatrix  $C$  für alle  $k$  konstant ist, spricht man von **stationären Iterationsverfahren**.

ii) Das quantitative Verhalten solch stationärer Verfahren lässt sich durch die Einführung eines (Relaxations-) Parameters  $\omega$  verbessern:

$$x^{(k+1)} = \omega(Cx^{(k)} + d) + (1 - \omega)x^{(k)}.$$

Für  $0 < \omega < 1$  spricht man von einem **Unterrelaxationsverfahren** und für  $\omega > 1$  von einem **Überrelaxationsverfahren**. Man kann zeigen, dass der optimale Parameter für eine positiv definite Matrix  $A$  beim gedämpften Jacobi-Verfahren

$$\omega_{\text{opt}} = \frac{2}{\lambda_{\min}(D^{-1}A) + \lambda_{\max}(D^{-1}A)}$$

lautet und für das überrelaxierte Gauß-Seidel-Verfahren (SOR = successive overrelaxation method) angewandt auf eine positiv definite Matrix  $A = L + D + L^T$  ergibt sich der optimale Parameter zu

$$\omega_{\text{opt}} = \frac{2}{1 + \sqrt{\lambda_{\min}(D^{-1}A) + \lambda_{\max}((D + 2L)D^{-1}(D + 2L^T)A^{-1})}}.$$

Ergebnisse für allgemeinere Fälle findet man z.B. bei [Niethammer].

iii) Die Bedeutung der oben genannten Iterationsverfahren liegt heute weniger in deren unmittelbarem Einsatz zur Lösung von  $Ax = b$ , sondern auf Grund deren „**Glättungseigenschaften**“ als Beschleuniger anderer moderner Verfahren (vorkonditioniertes konjugiertes Gradienten-Verfahren, Mehrgitter).

## 4.5 Abbruchkriterien

Da ein Iterationsverfahren aufeinanderfolgende Näherungen der Lösung liefert, ist ein praktischer Test notwendig, um das Verfahren zu stoppen, wenn die gewonnene Approximation genau genug ist. Da es nicht möglich ist, den **Fehler**  $e^{(k)} := x - x^{(k)}$ , d.h. den Abstand zur eigentlichen (gesuchten) Lösung, zu bestimmen, müssen andere Quantitäten gewonnen werden, die meist auf dem Residuum  $r = b - Ax$  basieren.

Die hier vorgestellten Verfahren liefern eine Folge  $(x^{(k)})$  von Vektoren, die gegen den Vektor  $x$  streben, welcher Lösung des linearen Gleichungssystems  $Ax = b$  ist. Um effizient zu sein, muss die Methode wissen, wann sie abbrechen soll. Eine gute Methode sollte

- i) feststellen, ob der Fehler  $e^{(k)} := x - x^{(k)}$  klein genug ist,
- ii) abbrechen, falls der Fehler nicht weiter kleiner wird oder nur noch sehr langsam, und
- iii) den maximalen Aufwand, der zur Iteration verwendet wird, beschränken.

Das folgende **Abbruchkriterium** ist eine einfache, aber häufig genügende Variante. Man muss hierzu die Quantitäten  $\maxit$ ,  $\|b\|$ ,  $tol$  und wenn möglich auch  $\|A\|$  (und  $\|A^{-1}\|$ ) zur Verfügung stellen. Dabei ist

- die natürliche Zahl  $\maxit$  die **maximale Anzahl an möglichen Iterationen des Verfahrens**,

- die reelle Zahl  $\|A\|$  eine Norm von  $A$ , (jede einigermaßen vernünftige Approximation des betragsgrößten Eintrags in  $A$  genügt schon),
- die reelle Zahl  $\|b\|$  eine Norm von  $b$  (auch hier genügt eine einigermaßen vernünftige Approximation des betragsgrößten Eintrags in  $b$ ),
- die reelle Zahl  $tol$  eine **Schranke für die Größe des Residuums bzw. des Fehlers**.

### MATLAB-Beispiel: Beispiel eines vernünftigen Abbruchkriteriums

```
k = 0;
while 1
    k = k + 1;
    % Berechne die Approximation x^(k)
    % Berechne das Residuum r^(k) = b - A x^(k)
    % Berechne norm_ak = || A * x^(k) ||, norm_rk = || r^(k) ||
    % und norm_b = || b ||
    if (k >= maxit) | ( norm_rk <= tol * max( norm_ak, norm_b ) )
        break
    end
end
end
```

Da sich nach einigen Iterationen der Term  $\|Ax^{(k)}\|$  nicht mehr groß ändert, braucht man diesen nicht immer neu zu bestimmen. Zu bestimmen ist eigentlich  $\|e^{(k)}\| = \|A^{-1}r^{(k)}\| \leq \|A^{-1}\| \|r^{(k)}\|$ . Man beachte, dass man  $\|A^{-1}B\|$  bei den bisherigen Verfahren mit Lemma 3.9.6 zur Neumannschen-Reihe abschätzen kann. Es gilt  $x^{(k+1)} = B^{-1}(B - A)x^{(k)} + B^{-1}b = Cx^{(k)} + d$  und

$$B^{-1}A = I - B^{-1}(B - A) = I - C \quad \text{sowie} \quad \|A^{-1}B\| = \|(I - C)^{-1}\| \leq \frac{1}{1 - \|C\|}.$$

## 4.6 Gradienten-Verfahren

Im Folgenden nehmen wir stets an, dass

$$A \in \mathbb{R}^{n \times n} \quad \text{symmetrisch positiv definit (s.p.d.) ist.} \quad (4.11)$$

Wir ordnen nun dem Gleichungssystem  $Ax = b$ ,  $b \in \mathbb{R}^n$ , die Funktion

$$f : \mathbb{R}^n \rightarrow \mathbb{R}, \quad f(x) := \frac{1}{2}x^T Ax - b^T x$$

zu. Man sieht leicht, dass die Funktion aufgrund von (4.11) strikt konvex ist. Der Gradient von  $f$  ist  $f'(x) = \frac{1}{2}(A + A^T)x - b$ . Da  $A = A^T$  nach Voraussetzung (4.11), lautet die Ableitung

$$f'(x) = \nabla f(x) = Ax - b.$$

Notwendig für ein Minimum von  $f$  ist das Verschwinden des Gradienten, d.h.  $Ax = b$ . Da die Hesse-Matrix  $f''(x) = A$  positiv definit ist, liegt für die Lösung von  $Ax = b$  tatsächlich ein Minimum vor. Das Minimum ist eindeutig.

Mit  $\arg \min_{y \in \mathbb{R}^n} f(y)$  bezeichnen wir denjenigen Wert aus  $\mathbb{R}^n$ , der den Term  $f$  minimiert, d.h.

$$f(x) = \min_{y \in \mathbb{R}^n} f(y), \quad \text{falls } x := \arg \min_{y \in \mathbb{R}^n} f(y).$$

**Idee:**

$$Ax = b \iff x = \arg \min_{y \in \mathbb{R}^n} f(y)$$

mit  $f(y) := \frac{1}{2}y^T Ay - b^T y$ ,  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ .

Bewiesen haben wir soeben das folgende Lemma.

**Lemma 4.6.1** *Es sei  $A \in \mathbb{R}^{n \times n}$  symmetrisch positiv definit, dann gilt*

$$Ax = b \iff x = \arg \min_{y \in \mathbb{R}^n} f(y).$$

*Alternativer Beweis zu Lemma 4.6.1.* Ein zweiter Beweis folgt aus der Darstellung

$$f(x) = f(x^*) + \frac{1}{2}(x - x^*)^T A(x - x^*) \quad \text{mit } x^* := A^{-1}b. \quad (4.12)$$

Hieraus folgt  $f(x) > f(x^*)$  für  $x \neq x^*$ , d.h.  $x^* := A^{-1}b$  ist das eindeutige Minimum von  $f$ . Man beachte dabei, dass (4.12) ein Sonderfall der folgenden Entwicklung von  $f$  um einen beliebigen Wert  $\tilde{x} \in \mathbb{R}^n$  ist, welche sich durch ausmultiplizieren zeigen lässt:

$$f(x) = f(\tilde{x}) + \langle x - \tilde{x}, A\tilde{x} - b \rangle + \frac{1}{2}\langle x - \tilde{x}, A(x - \tilde{x}) \rangle$$

□

**Folgerung:** Man kann also Verfahren zur numerischen Optimierung/Minimierung verwenden, um das lineare Gleichungssystem zu lösen.

Der **Gradient** ist die Richtung des **steilsten Anstiegs**, also kann man  $-\nabla f$  als Abstiegsrichtung wählen und entlang dieser Geraden ein Minimum suchen.

**Gradienten-Verfahren (allgemein):**

Es sei  $\Omega \subseteq \mathbb{R}^n$ ,  $f : \Omega \rightarrow \mathbb{R}$ ,  $x^{(0)} \in \Omega$  Startwert, für  $k = 1, 2, 3, \dots$

- 1) **Bestimmung der Abstiegsrichtung:**  $d^{(k)} := -\nabla f(x^{(k)})$
- 2) **Linienuche:** Suche auf der Geraden  $\{x^{(k)} + td^{(k)} : t \geq 0\} \cap \Omega$  ein Minimum, d.h. bestimme  $\lambda_k \geq 0$  mit  $f(x^{(k)} + \lambda_k d^{(k)}) \leq f(x^{(k)})$  und setze

$$x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)}.$$

*Bemerkung:* Daraus folgt  $f(x^{(0)}) \geq f(x^{(1)}) \geq f(x^{(2)}) \geq \dots$

Für die quadratische Funktionen  $f(x) = \frac{1}{2}x^T Ax - b^T x$  und  $\Omega = \mathbb{R}^n$  kann man 1) und 2) leicht berechnen: Da  $\nabla f(x) = Ax - b$ , ergibt sich  $d^{(k)} = -\nabla f(x^{(k)}) = b - Ax^{(k)}$ . Sei  $p \in \mathbb{R}^n \setminus \{0\}$  und  $F(\lambda) := f(x + \lambda p)$ , dann gilt für die Linienuche

$$\begin{aligned} F(\lambda) &= f(x + \lambda p) \\ &= \frac{1}{2}\langle x + \lambda p, A(x + \lambda p) \rangle - \langle b, x + \lambda p \rangle \\ &= \frac{1}{2}\langle x, Ax \rangle - \langle b, x \rangle + \lambda \langle p, Ax - b \rangle + \frac{1}{2}\lambda^2 \langle p, Ap \rangle \\ &= f(x) + \lambda \langle p, Ax - b \rangle + \frac{1}{2}\lambda^2 \langle p, Ap \rangle. \end{aligned} \quad (4.13)$$

Da  $p \neq 0$  nach Voraussetzung, folgt  $\langle p, Ap \rangle > 0$ .  $F$  ist also eine quadratische Funktion mit positivem führenden Koeffizienten. Somit folgt aus

$$0 \stackrel{!}{=} F'(\lambda) = \langle p, Ax - b \rangle + \lambda \langle p, Ap \rangle, \quad (4.14)$$

dass der Parameter

$$\lambda_{opt}(x, p) = \frac{\langle p, b - Ax \rangle}{\langle p, Ap \rangle}. \quad (4.15)$$

zu gegebenem Vektor  $p \in \mathbb{R}^n \setminus \{0\}$  das Funktional  $F(\lambda) := f(x + \lambda p)$  minimiert.

Für allgemeine Funktionen  $f$  wird die Liniensuche angenähert, z.B. mit der **Schrittweitenregel von Armijo**.

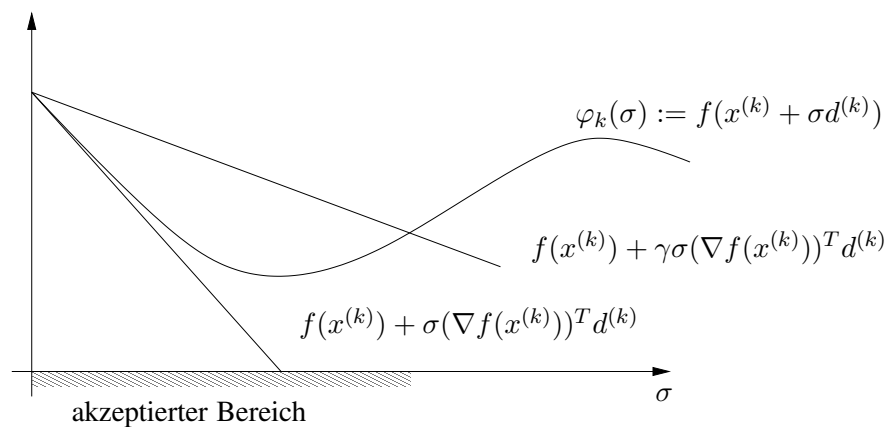


Abb. 4.5: Optimale Wahl des Dämpfungsparameters

#### Schrittweitenregel von Armijo:

Wähle  $\beta \in (0, 1)$  (z.B.  $\beta = \frac{1}{2}$ ) und  $\gamma \in (0, 1)$  (z.B.  $\gamma \in [10^{-3}, 10^{-2}]$ )

Bestimme die größte Schrittweite  $\sigma_k \in \{1, \beta, \beta^2, \beta^3, \dots\}$  mit

$$f(x^{(k)}) - f(x^{(k)} + \sigma_k d^{(k)}) \geq -\gamma \sigma_k \nabla f(x^{(k)})^T d^{(k)},$$

d.h.

$$f(x^{(k)} + \sigma_k d^{(k)}) \leq f(x^{(k)}) + \gamma \sigma_k \nabla f(x^{(k)})^T d^{(k)}.$$

Formulieren wir nun das Gradienten-Verfahren mit der optimalen Schrittweite (4.15) in folgendem Algorithmus.

#### Algorithmus 4.6.1: Gradienten-Verfahren: $Ax = b$

**Input:** Initial guess  $x^{(0)}$

$$r^{(0)} := b - Ax^{(0)}$$

**Iteration:**  $k = 0, 1, \dots$

$$a^{(k)} := Ar^{(k)}$$

$$\lambda_{opt} := \langle r^{(k)}, r^{(k)} \rangle / \langle r^{(k)}, a^{(k)} \rangle$$

$$x^{(k+1)} := x^{(k)} + \lambda_{opt} r^{(k)}$$

$$r^{(k+1)} := r^{(k)} - \lambda_{opt} a^{(k)}$$

Man beachte:

$$r^{(k+1)} = b - Ax^{(k+1)} = b - A(x^{(k)} + \lambda_{opt} r^{(k)}) = r^{(k)} - \lambda_{opt} Ar^{(k)}.$$

Wir untersuchen nun die Konvergenz des Verfahrens für quadratische Funktionen. Hierzu bietet sich die sogenannte **Energienorm** an

$$\|x\|_A := \sqrt{x^T A x}, \quad (A \in \mathbb{R}^{n \times n}).$$

Man beachte: alle Normen auf dem  $\mathbb{R}^n$  sind äquivalent.

**Lemma 4.6.2** *Es sei  $A \in \mathbb{R}^{n \times n}$  positiv definit und  $x^* \in \mathbb{R}^n$  erfülle  $Ax^* = b$ . Dann gilt für die durch das Gradienten-Verfahren erzeugten Iterierten  $x^{(k)}$  folgende Abschätzung:*

$$\|x^{(k+1)} - x^*\|_A^2 \leq \|x^{(k)} - x^*\|_A^2 \left( 1 - \frac{\langle r^{(k)}, r^{(k)} \rangle^2}{\langle r^{(k)}, Ar^{(k)} \rangle \langle r^{(k)}, A^{-1}r^{(k)} \rangle} \right).$$

*Beweis.* Es gilt

$$\begin{aligned} f(x^{(k+1)}) &= f(x^{(k)} + \lambda_{opt} r^{(k)}) = \frac{1}{2} \lambda_{opt}^2 \langle r^{(k)}, Ar^{(k)} \rangle - \lambda_{opt} \langle r^{(k)}, r^{(k)} \rangle + f(x^{(k)}) \\ &= f(x^{(k)}) + \frac{1}{2} \frac{\langle r^{(k)}, r^{(k)} \rangle^2}{\langle r^{(k)}, Ar^{(k)} \rangle} - \frac{\langle r^{(k)}, r^{(k)} \rangle^2}{\langle r^{(k)}, Ar^{(k)} \rangle} = f(x^{(k)}) - \frac{1}{2} \frac{\langle r^{(k)}, r^{(k)} \rangle^2}{\langle r^{(k)}, Ar^{(k)} \rangle}. \end{aligned} \quad (4.16)$$

Für die exakte Lösung  $x^*$  von  $Ax = b$  gilt für  $x \in \mathbb{R}^n$

$$\begin{aligned} f(x^*) + \frac{1}{2} \|x - x^*\|_A^2 &= \frac{1}{2} \langle x^*, Ax^* \rangle - \langle b, x^* \rangle + \frac{1}{2} \langle x - x^*, A(x - x^*) \rangle \\ &= \frac{1}{2} \langle x^*, Ax^* \rangle - \langle x^*, b \rangle + \frac{1}{2} \langle x, Ax \rangle - \langle x, Ax^* \rangle + \frac{1}{2} \langle x^*, Ax^* \rangle \\ &= \langle x^*, Ax^* - b \rangle + \frac{1}{2} \langle x, Ax \rangle - \langle x, b \rangle \\ &= f(x), \quad \text{d.h. } \|x - x^*\|_A^2 = 2(f(x) - f(x^*)) \end{aligned}$$

also mit (4.16)

$$\begin{aligned} \|x^{(k+1)} - x^*\|_A^2 &= 2f(x^{(k+1)}) - 2f(x^*) = 2f(x^{(k+1)}) - 2f(x^{(k)}) + \|x^{(k)} - x^*\|_A^2 \\ &= \|x^{(k)} - x^*\|_A^2 - \frac{\langle r^{(k)}, r^{(k)} \rangle^2}{\langle r^{(k)}, Ar^{(k)} \rangle}. \end{aligned}$$



Mit  $r^{(k)} = b - Ax^{(k)} = A(x^* - x^{(k)})$  folgt wegen

$$\begin{aligned}\|x^{(k)} - x^*\|_A^2 &= \|x^* - x^{(k)}\|_A^2 = \langle x^* - x^{(k)}, A(x^* - x^{(k)}) \rangle \\ &= \langle A^{-1}A(x^* - x^{(k)}), r^{(k)} \rangle = \langle r^{(k)}, A^{-1}r^{(k)} \rangle\end{aligned}$$

die Behauptung. □

**Frage:** Was sagt Lemma 4.6.2 bzgl. der Konvergenz und Kondition aus?

**Lemma 4.6.3 (Kantorowitsch-Ungleichung)** *Es sei  $A \in \mathbb{R}^{n \times n}$  symmetrisch positiv definit und  $\kappa := \kappa_2(A) := \|A\|_2 \|A^{-1}\|_2$ . Dann gilt für alle  $x \in \mathbb{R} \setminus \{0\}$*

$$\frac{\langle x, Ax \rangle \langle x, A^{-1}x \rangle}{\langle x, x \rangle^2} \leq \frac{1}{4} \left( \sqrt{\kappa} + \sqrt{\kappa^{-1}} \right)^2.$$

*Beweis.* Die Eigenwerte von  $A$  seien geordnet gemäß

$$0 < \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n, \quad \kappa = \frac{\lambda_n}{\lambda_1}.$$

Da  $A$  symmetrisch ist, existiert eine orthonormale Matrix  $Q$  mit  $Q^T A Q = \Lambda = \text{diag}(\lambda_i)$ . Für  $y = Q^T x$  gilt dann

$$x^T A x = x^T Q \Lambda Q^T x = y^T \Lambda y = \sum_{i=1}^n \lambda_i y_i^2, \quad x^T A^{-1} x = x^T Q \Lambda^{-1} Q^T x = \sum_{i=1}^n \lambda_i^{-1} y_i^2$$

sowie  $x^T x = x^T Q Q^T x = y^T y$ , also

$$\frac{\langle x, Ax \rangle \langle x, A^{-1}x \rangle}{\langle x, x \rangle^2} = \frac{\left( \sum_i \lambda_i y_i^2 \right) \left( \sum_i \lambda_i^{-1} y_i^2 \right)}{\|y\|_2^4} = \left( \sum_{i=1}^n \lambda_i z_i \right) \left( \sum_{i=1}^n \lambda_i^{-1} z_i \right) \quad (4.17)$$

mit  $z_i := \frac{y_i^2}{\|y\|_2^2}$ . Man beachte  $\sum_{i=1}^n z_i = 1$ . Für  $\lambda_1 \leq \alpha \leq \lambda_n$  folgt

$$0 \geq (\alpha - \lambda_1)(\alpha - \lambda_n) = \alpha^2 - \alpha(\lambda_1 + \lambda_n) + \lambda_1 \lambda_n$$

und somit

$$\lambda_1 \lambda_n + \lambda_k^2 \leq \lambda_k(\lambda_1 + \lambda_n) \quad \Rightarrow \quad \frac{\lambda_1 \lambda_n}{\lambda_k} + \lambda_k \leq \lambda_1 + \lambda_n \quad (k = 1, \dots, n).$$

Es gilt nun

$$\lambda_1 \lambda_n \sum_{i=1}^n \lambda_i^{-1} z_i + \sum_{i=1}^n \lambda_i z_i = \left( \frac{\lambda_1 \lambda_n}{\lambda_1} + \lambda_1 \right) z_1 + \left( \frac{\lambda_1 \lambda_n}{\lambda_2} + \lambda_2 \right) z_2 + \dots + \left( \frac{\lambda_1 \lambda_n}{\lambda_n} + \lambda_n \right) z_n \leq \lambda_1 + \lambda_n,$$

d.h.

$$\sum_{i=1}^n \lambda_i^{-1} z_i \leq \frac{\lambda_1 + \lambda_n - \lambda}{\lambda_1 \lambda_n}$$

mit  $\lambda := \sum_{i=1}^n \lambda_i z_i$ . Somit lässt sich (4.17) abschätzen durch

$$\frac{\langle x, Ax \rangle \langle x, A^{-1}x \rangle}{\langle x, x \rangle^2} \leq \lambda \cdot \underbrace{\frac{\lambda_1 + \lambda_n - \lambda}{\lambda_1 \lambda_n}}_{=: h(\lambda)}.$$

Für welches  $\lambda$  wird nun das Polynom  $h$  maximal?

$$\begin{aligned} h'(\lambda) &= \frac{\lambda_1 + \lambda_n - \lambda}{\lambda_1 \lambda_n} - \frac{\lambda}{\lambda_1 \lambda_n} = \frac{\lambda_1 + \lambda_n}{\lambda_1 \lambda_n} - \lambda \frac{2}{\lambda_1 \lambda_n} \stackrel{!}{=} 0 \Rightarrow \lambda^* = \frac{\lambda_1 + \lambda_n}{2} \\ h''(\lambda) &= -\frac{2}{\lambda_1 \lambda_n} < 0 \Rightarrow \lambda^* \text{ maximiert } h \end{aligned}$$

d.h.

$$\max_{\lambda \in [\lambda_1, \lambda_n]} h(\lambda) = h(\lambda^*) = \frac{(\lambda_1 + \lambda_n)^2}{4 \lambda_1 \lambda_n} = \frac{1}{4} \left( \sqrt{\frac{\lambda_1}{\lambda_n}} + \sqrt{\frac{\lambda_n}{\lambda_1}} \right)^2 = \frac{1}{4} \left( \sqrt{\kappa} + \sqrt{\kappa^{-1}} \right)^2.$$

□

**Satz 4.6.4** Es sei  $A \in \mathbb{R}^{n \times n}$  positiv definit und  $x^* \in \mathbb{R}^n$  erfülle  $Ax^* = b$ . Dann gilt für das Gradienten-Verfahren

$$\|x^{(k)} - x^*\|_A \leq \left( \frac{\kappa - 1}{\kappa + 1} \right)^k \|x^{(0)} - x^*\|_A$$

*Beweis.* Lemma 4.6.2 liefert die Abschätzung

$$\|x^{(k+1)} - x^*\|_A^2 \leq \|x^{(k)} - x^*\|_A^2 \left( 1 - \frac{\langle r^{(k)}, r^{(k)} \rangle^2}{\langle r^{(k)}, Ar^{(k)} \rangle \langle r^{(k)}, A^{-1}r^{(k)} \rangle} \right).$$

und mit Lemma 4.6.3 ergibt sich

$$\begin{aligned} 1 - \frac{\langle r^{(k)}, r^{(k)} \rangle^2}{\langle r^{(k)}, Ar^{(k)} \rangle \langle r^{(k)}, A^{-1}r^{(k)} \rangle} &\leq 1 - 4 \left( \sqrt{\kappa} + \sqrt{\kappa^{-1}} \right)^{-2} = \frac{\left( \sqrt{\kappa} + \sqrt{\kappa^{-1}} \right)^2 - 4}{\left( \sqrt{\kappa} + \sqrt{\kappa^{-1}} \right)^2} \\ &= \frac{\kappa + 2 + \kappa^{-1} - 4}{\kappa + 2 + \kappa^{-1}} = \frac{\kappa^2 - 2\kappa + 1}{\kappa^2 + 2\kappa + 1} = \left( \frac{\kappa - 1}{\kappa + 1} \right)^2. \end{aligned}$$

□

**Bemerkungen 4.6.5** i) Für große  $\kappa$  gilt

$$\frac{\kappa - 1}{\kappa + 1} = \underbrace{\frac{\kappa}{\kappa + 1}}_{\approx 1} - \frac{1}{\kappa + 1} \approx 1 - \frac{1}{\kappa + 1},$$

also sehr nahe bei 1, d.h. geringe Konvergenzgeschwindigkeit!

ii) Dies tritt auch schon bei einfachen Beispielen auf:

$$A = \begin{pmatrix} 1 & 0 \\ 0 & a \end{pmatrix}, \quad a \gg 1, \quad b = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{und} \quad x^{(0)} = \begin{pmatrix} a \\ 1 \end{pmatrix}$$

daraus folgt (**Übung**):

$$\begin{pmatrix} x^{(k+1)} \\ y_{k+1} \end{pmatrix} = \rho \begin{pmatrix} x^{(k)} \\ -y^{(k)} \end{pmatrix} \quad \text{mit} \quad \rho = \frac{a-1}{a+1}$$

wegen  $a = \kappa_2(A)$  ist das genau die Rate aus Satz 4.6.4!

iii) Anschaulich sieht man ein „Zick-Zack-Verhalten“ der Iteration, vgl. Abbildung 4.6.

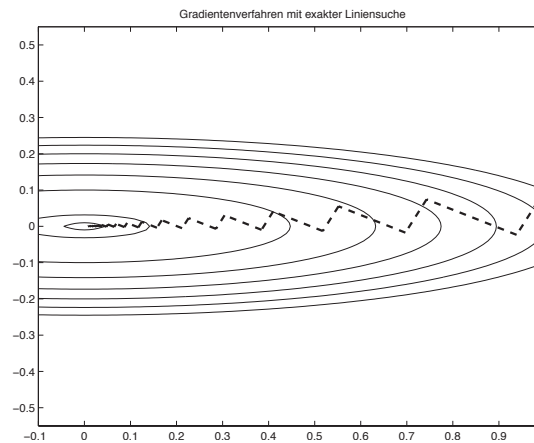


Abb. 4.6: Zick-Zack-Verhalten des Gradienten-Verfahrens mit exakter Liniensuche.

## 4.7 Verfahren der konjugierten Gradienten

Das Verfahren der konjugierten Gradienten (englisch: conjugate gradient, auch **cg-Verfahren** genannt) wurde 1952 von von Hestenes und Stiefel entwickelt. Man konnte zunächst zeigen, dass dieses Verfahren nach  $n$  Schritten die exakte Lösung liefert, wenn keine Rundungsfehler auftreten. In diesem Sinne ist das cg-Verfahren ein direktes Verfahren. Für große  $n$  ist diese Aussage aber wertlos. Erst 1971 gewann das cg-Verfahren durch die sogenannte **Vorkonditionierung** enorm an Bedeutung und heute gehören vorkonditionierte cg-Verfahren zu den schnellsten Verfahren, die sehr oft verwendet werden. Für das Beispiel der 2D-Standardmatrix (Beispiel 4.0.3) ist das cg-Verfahren ab einer Systemgröße von 2000–4000 Variablen deutlich besser als das Gauß-Verfahren bei zusätzlich erheblich geringerem Speicherbedarf.

**Idee:** Vermeide Zick-Zack-Verhalten durch Verwendung von Orthogonalität bzgl.

$$(x, y)_A := x^T A y,$$

dies ergibt „konjugierte Gradienten“, daher der Name cg-Verfahren. Das Skalarprodukt bzgl. dessen man Orthogonalität misst, ist also durch die Matrix selber bestimmt.

### Bemerkungen 4.7.1

i) Zwei Vektoren  $x, y \in \mathbb{R}^n$  heißen **konjugiert** oder **A-orthogonal**, falls  $(x, y)_A = 0$ .

ii) Sind die Vektoren  $\{x^{(1)}, \dots, x^{(k)}\}$  **paarweise konjugiert**, d.h.

$$(x^{(i)}, x^{(j)})_A = \delta_{ij} \|x^{(i)}\|_A^2, \quad x^{(i)} \neq 0 \quad (i, j \in \{1, \dots, k\}),$$

dann sind  $\{x^{(1)}, \dots, x^{(k)}\}$  **linear unabhängig**.

iii) Jeder Vektor  $x \in \mathbb{R}^n$  besitzt eine **eindeutige Entwicklung**

$$x = \sum_{k=1}^n \alpha_k d^{(k)} \tag{4.18}$$

bezüglich konjugierter Richtungen  $\{d^{(1)}, \dots, d^{(n)}\}$ . Aus (4.18) folgt

$$(x, d^{(i)})_A = \sum_{k=1}^n \alpha_k \underbrace{(d^{(k)}, d^{(i)})_A}_{=\delta_{ik} \|d^{(i)}\|_A^2} = \alpha_i \|d^{(i)}\|_A^2,$$

also

$$\alpha_k = \frac{(d^{(k)})^T A x}{(d^{(k)})^T A d^{(k)}} \quad (k = 1, \dots, n). \quad (4.19)$$

iv) Für die Lösung  $x^*$  von  $Ax = b$  gilt offenbar

$$\alpha_i = \frac{(d^{(i)})^T b}{(d^{(i)})^T A d^{(i)}}.$$

**Lemma 4.7.2** Seien  $\{d^{(1)}, \dots, d^{(n)}\}$  konjugierte Richtungen. Für jedes  $x^{(0)} \in \mathbb{R}^n$  und

$$x^{(k)} = x^{(k-1)} + \alpha_k d^{(k)}, \quad \alpha_k = \frac{(r^{(k-1)})^T d^{(k)}}{(d^{(k)})^T A d^{(k)}}, \quad r^{(k)} := b - Ax^{(k)} \quad (k \geq 1) \quad (4.20)$$

gilt nach (höchstens)  $n$  Schritten  $x^{(n)} = A^{-1}b$ .

*Beweis.* Aus (4.19), (4.20) folgt für  $x^* = A^{-1}b$ ,

$$x^* - x^{(0)} = \sum_{k=1}^n \tilde{\alpha}_k d^{(k)} \quad \text{mit} \quad \tilde{\alpha}_k = \frac{(d^{(k)})^T A(x^* - x^{(0)})}{(d^{(k)})^T A d^{(k)}} = \frac{(d^{(k)})^T (b - Ax^{(0)})}{(d^{(k)})^T A d^{(k)}}.$$

Da  $d^{(k)}$  zu  $d^{(i)}$ ,  $i \neq k$ , konjugiert ist, gilt

$$(d^{(k)})^T A(x^{(k-1)} - x^{(0)}) = (d^{(k)})^T A \left( \sum_{i=1}^{k-1} \alpha_i d^{(i)} \right) = \sum_{i=1}^{k-1} \alpha_i \underbrace{(d^{(k)})^T A d^{(i)}}_{=0, \text{ da } i \neq k} = 0,$$

also

$$\begin{aligned} (d^{(k)})^T A(x^* - x^{(0)}) &= (d^{(k)})^T A(x^* - x^{(k-1)}) + \underbrace{(d^{(k)})^T A(x^{(k-1)} - x^{(0)})}_{=0} \\ &= (d^{(k)})^T (b - Ax^{(k-1)}) = (d^{(k)})^T r^{(k-1)} \Rightarrow \tilde{\alpha}_k = \alpha_k, \end{aligned}$$

womit die Aussage bewiesen ist.  $\square$

**Bemerkungen 4.7.3** i)  $r^{(k)} := b - Ax^{(k)}$  wird als **Residuum** von  $Ax = b$  bzgl.  $x^{(k)}$  bezeichnet.

ii) Lemma 4.7.2 besagt, dass das Verfahren ein direktes Verfahren ist, welches nach  $n$  Iterationen konvergiert. Also:

$$A \text{ sparse} \Rightarrow \mathcal{O}(n^2)$$

(dies ist **nicht** optimal).

iii) Wie in (4.14) gilt

$$\frac{d}{d\lambda} f(x^{(k-1)} + \lambda d^{(k)}) = \lambda \langle d^{(k)}, A d^{(k)} \rangle + \langle d^{(k)}, (Ax^{(k-1)} - b) \rangle,$$

d.h.

$$f(x^{(k)}) = f(x^{(k-1)} + \alpha_k d^{(k)}) = \min_{\lambda \in \mathbb{R}} f(x^{(k-1)} + \lambda d^{(k)}),$$

dann ist

$$\lambda_{\text{opt}} = -\frac{\langle d^{(k)}, (Ax^{(k-1)} - b) \rangle}{\langle d^{(k)}, A d^{(k)} \rangle} = \frac{\langle r^{(k-1)}, d^{(k)} \rangle}{\langle d^{(k)}, A d^{(k)} \rangle} = \alpha_k.$$

**Satz 4.7.4** Seien  $\{d^{(1)}, \dots, d^{(n)}\}$  konjugierte Richtungen und  $r^{(k)}$  ( $k = 0, \dots, n-1$ ) die durch (4.20) definierten Residuen. Dann gilt

$$(r^{(k)})^T d^{(j)} = 0 \text{ bzw. } r^{(k)} \perp U_k := \text{span}\{d^{(1)}, \dots, d^{(k)}\} \quad (1 \leq k \leq n, 1 \leq j \leq k). \quad (4.21)$$

*Beweis.* Nach (4.20) gilt für  $k \in \{1, \dots, n\}$

$$\begin{aligned} r^{(k)} &= b - Ax^{(k)} = r^{(k-1)} - \alpha_k Ad^{(k)} = r^{(k-2)} - \alpha^{(k-1)} Ad^{(k-1)} - \alpha_k Ad^{(k)} \\ &= \dots = r^{(0)} - \sum_{j=1}^k \alpha_j Ad^{(j)}. \end{aligned}$$

Daraus folgt nun für  $1 \leq j \leq k$

$$\begin{aligned} (r^{(k)})^T d^{(j)} &= (r^{(0)})^T d^{(j)} - \sum_{\ell=1}^k \alpha_\ell (d^{(\ell)})^T Ad^{(j)} \\ &= (r^{(0)})^T d^{(j)} - \alpha_j (d^{(j)})^T Ad^{(j)} = (r^{(0)})^T d^{(j)} - \frac{(r^{(j-1)})^T d^{(j)}}{(d^{(j)})^T Ad^{(j)}} (d^{(j)})^T Ad^{(j)} \\ &= ((r^{(0)})^T - (r^{(j-1)})^T) d^{(j)} = \sum_{\ell=1}^{j-1} \alpha_\ell (d^{(\ell)})^T Ad^{(j)} = 0 \end{aligned}$$

womit der Satz bewiesen ist.  $\square$

**Frage:** Wie sind nun die  $d^{(k)}$  und damit erzeugten Teilräume  $\text{span}\{d^{(1)}, \dots, d^{(k)}\}$  zu wählen?

Falls  $r^{(0)} \neq 0$  gilt (sonst ist  $x^{(0)}$  schon die gesuchte Lösung) setzt man  $d^{(1)} = r^{(0)}$ .

Für  $k = 1, 2, 3, \dots$  verfahren wir wie folgt: Falls  $r^{(k)} \neq 0$  (sonst wäre ja  $x^{(k)}$  schon die gesuchte Lösung), gewinnt man formal  $d^{(k+1)}$  mittels des Gram-Schmidtschen-Orthogonalisierungsverfahren aus  $r^{(k)}$  und den schon bestimmten konjugierten Richtungen  $d^{(1)}, \dots, d^{(k)}$ , d.h.

$$d^{(k+1)} = r^{(k)} - \sum_{j=1}^k \frac{\langle r^{(k)}, Ad^{(j)} \rangle}{\langle d^{(j)}, Ad^{(j)} \rangle} d^{(j)}. \quad (4.22)$$

**Aufgabe 4.7.5** Man zeige induktiv, dass die so erzeugten  $A$ -orthogonalen Richtungen

$$d^{(k+1)} \notin \text{span}\{r^{(0)}, \dots, r^{(k-1)}\} \quad \text{und} \quad d^{(k+1)} \in \text{span}\{r^{(0)}, \dots, r^{(k)}\}$$

erfüllen für  $0 \leq k \leq n-1$  mit  $r^{(k)} \neq 0$ .

Damit das ganze Verfahren effizient wird, benötigen wir noch folgende Eigenschaft

$$Ad^{(k)} \in U_{k+1} := \text{span}\{d^{(1)}, \dots, d^{(k+1)}\} = \text{span}\{r^{(0)}, \dots, r^{(k)}\},$$

wenn  $r^{(k)} \neq 0$  gilt. Denn daraus ergibt sich  $\langle r^{(k)}, Ad^{(j)} \rangle = 0$  für  $1 \leq j \leq k-1$  und (4.22) verkürzt sich zu

$$d^{(k+1)} = r^{(k)} - \underbrace{\frac{\langle r^{(k)}, Ad^{(k)} \rangle}{\langle d^{(k)}, Ad^{(k)} \rangle}}_{=: \beta_{k+1}} d^{(k)}. \quad (4.23)$$

(Beweis: Aus der Definition von  $r^{(k)} = r^{(k-1)} - \alpha_k Ad^{(k)}$  folgt sofort  $Ad^{(k)} \in \{r^{(0)}, \dots, r^{(k)}\}$ , da  $r^{(k-1)} \in U_k$  und  $\alpha_k \neq 0$  gilt. Die Gleichheit von  $U_k$  und  $\{r^{(0)}, \dots, r^{(k)}\}$  ergibt sich aus Aufgabe 4.7.5.)

Für den Algorithmus schreiben wir nur noch  $\alpha_k, \beta_k$  um:  $\alpha_k = \frac{(r^{(k-1)})^T d^{(k)}}{(d^{(k)})^T A d^{(k)}}$  und

$$\begin{aligned} (r^{(k-1)})^T d^{(k)} &= (r^{(k-1)})^T r^{(k-1)} - \beta_k \underbrace{(r^{(k-1)})^T d^{(k-1)}}_{=0} \quad \text{also} \\ \alpha_k &= \frac{(r^{(k-1)})^T r^{(k-1)}}{(d^{(k)})^T A d^{(k)}} \end{aligned} \quad (4.24)$$

und wegen  $\alpha_k (r^{(k)})^T A d^{(k)} = (r^{(k-1)} - r^{(k)})^T r^{(k)} = -(r^{(k)})^T r^{(k)}$

$$\beta_{k+1} = \frac{(r^{(k)})^T A d^{(k)}}{(d^{(k)})^T A d^{(k)}} = -\frac{(r^{(k)})^T r^{(k)}}{\alpha_k (d^{(k)})^T A d^{(k)}} = -\frac{(r^{(k)})^T r^{(k)}}{(r^{(k-1)})^T r^{(k-1)}}. \quad (4.25)$$

**Bemerkung 4.7.6** Die Ausdrücke (4.24), (4.25) haben sich als numerisch stabiler und Speicherplatz-optimal herausgestellt.

**Algorithmus 4.7.1: Konjugiertes Gradienten-Verfahren (cg-Verfahren):**  $Ax = b$ **Input:** Initial guess  $x^{(0)}$ 

$$r^{(0)} := b - Ax^{(0)}$$

$$\rho_0 := \langle r^{(0)}, r^{(0)} \rangle$$

$$d^{(1)} := r^{(0)}$$

**Iteration:**  $k = 1, 2, \dots$  as long as  $k \leq n$  and  $r^{(k)} \neq 0$ 

$$a^{(k)} := Ad^{(k)}$$

$$\alpha_k := \rho^{(k-1)} / \langle d^{(k)}, a^{(k)} \rangle$$

$$x^{(k)} := x^{(k-1)} + \alpha_k d^{(k)}$$

$$r^{(k)} := r^{(k-1)} - \alpha_k a^{(k)}$$

$$\rho_k := \langle r^{(k)}, r^{(k)} \rangle$$

$$d^{(k+1)} := r^{(k)} + \frac{\rho^{(k)}}{\rho^{(k-1)}} d^{(k)}$$

**Satz 4.7.7** Es sei  $A$  symmetrisch positiv definit. Für das cg-Verfahren und  $x^* = A^{-1}b$  gilt folgende Abschätzung

$$\|x^{(k)} - x^*\|_A \leq 2 \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|x^{(0)} - x^*\|_A.$$

*Beweis.* Der Beweis gliedert sich in 4 Schritte. Es sei  $e^{(k)} := x^* - x^{(k)}$ .

(1) Zuerst zeigt man induktiv, dass Polynome  $p_k \in \mathbb{P}_k$ ,  $k = 0, \dots, n-1$ , existieren mit

$$e^{(k)} = p_k(A)e^{(0)}, \quad p_k(0) = 1. \quad (4.26)$$

$k = 0$ : klar!

$k-1 \Rightarrow k$ : Aus der Definition (4.20) folgt

$$\alpha_k d^{(k)} = x^{(k)} - x^{(k-1)} = x^* - x^{(k-1)} - (x^* - x^{(k)}) = e^{(k-1)} - e^{(k)},$$

bzw.  $e^{(k)} = e^{(k-1)} - \alpha_k d^{(k)}$ . Mit (4.23) ergibt sich nun

$$\begin{aligned} d^{(k)} &= r^{(k-1)} - \beta_k d^{(k-1)} = b - Ax^{(k-1)} - \beta_k \frac{1}{\alpha_{k-1}} (e^{(k-2)} - e^{(k-1)}) \\ &= A(x^* - x^{(k-1)}) + \frac{\beta_k}{\alpha_{k-1}} (e^{(k-1)} - e^{(k-2)}) \\ &= \left( \frac{\beta_k}{\alpha_{k-1}} I + A \right) e^{(k-1)} - \frac{\beta_k}{\alpha_{k-1}} e^{(k-2)} \\ &\stackrel{\text{IH}}{=} \underbrace{\left\{ \left( \frac{\beta_k}{\alpha_{k-1}} I + A \right) p^{(k-1)}(A) - \frac{\beta_k}{\alpha_{k-1}} p_{k-2}(A) \right\}}_{=: \tilde{q}_k(A), \tilde{q}_k \in \mathbb{P}_k} e^{(0)}, \end{aligned}$$

also

$$e^{(k)} = e^{(k-1)} - \alpha_k d^{(k)} \stackrel{\text{H}}{=} p^{(k-1)}(A)e^{(0)} - \alpha_k d^{(k)} = \underbrace{[p_{k-1}(A) - \alpha_k \tilde{q}_k(A)]}_{=: p_k(A)} e^{(0)}.$$

- (2) Nun zeigt man, dass für alle  $q_k \in \mathbb{P}_k$  mit  $q_k(0) = 1$  die Ungleichung  $\|e^{(k)}\|_A \leq \|q_k(A)e^{(0)}\|_A$  gilt.

Zuerst halten wir  $r^{(k)} = b - Ax^{(k)} = A(x^* - x^{(k)}) = Ae^{(k)}$  fest. Des Weiteren gilt für beliebige  $\sigma^{(0)}, \dots, \sigma^{(k-1)} \in \mathbb{R}$  (beachte:  $(r^{(j)})^T r^{(k)} = \delta_{jk}$ ):

$$\begin{aligned} \|e^{(k)}\|_A^2 &= (e^{(k)})^T Ae^{(k)} = (r^{(k)})^T e^{(k)} \\ &= (r^{(k)})^T \left( e^{(k)} + \sum_{j=0}^{k-1} \sigma_j r^{(j)} \right) = (r^{(k)})^T \left( e^{(k)} + \sum_{j=0}^{k-1} \sigma_j Ae^{(j)} \right) \\ &= (Ae^{(k)})^T \left( p_k(A) + \sum_{j=0}^{k-1} \sigma_j Ap_j(A) \right) e^{(0)}. \end{aligned}$$

Sei  $q_k \in \mathbb{P}_k$  mit  $q_k(0) = 1$  beliebig. Da die  $\{p^{(0)}, \dots, p^{(k-1)}\}$  linear unabhängig sind, folgt aus der letzten Umformung, dass es eindeutig bestimmte  $\tilde{\sigma}^{(0)}, \dots, \tilde{\sigma}^{(k-1)}$  existieren mit

$$q_k(t) = p_k(t) + \sum_{j=0}^{k-1} \tilde{\sigma}_j t p_j(t).$$

Mit  $\sigma_i = \tilde{\sigma}_i$  und der Ungleichung von Cauchy-Schwarz folgt also

$$\begin{aligned} \|e^{(k)}\|_A^2 &= \langle e^{(k)}, Aq_k(A)e^{(0)} \rangle = \langle A^{\frac{1}{2}}e^{(k)}, A^{\frac{1}{2}}q_k(A)e^{(0)} \rangle \\ &\stackrel{\text{CSU}}{\leq} \|A^{\frac{1}{2}}e^{(k)}\| \cdot \|A^{\frac{1}{2}}q_k(A)e^{(0)}\| = \|e^{(k)}\|_A \cdot \|q_k(A)e^{(0)}\|_A. \end{aligned}$$

Wir haben somit gezeigt, dass

$$\|e^{(k)}\|_A \leq \|q_k(A)e^{(0)}\|_A \quad \text{für alle } q_k \in \mathbb{P}_k, q_k(0) = 1 \quad (4.27)$$

gilt.

- (3) Die Eigenwerte von  $A$  seien  $0 < \lambda_{\min} = \lambda_1 \leq \dots \leq \lambda_n = \lambda_{\max}$ . Es gelte  $A = Q\Lambda Q^T$  mit  $QQ^T = Q^TQ = I$  und  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ . Beachtet man weiterhin  $q_k(A) = q_k(Q\Lambda Q^T) = Qq_k(\Lambda)Q^T$ , so schätzt man wie folgt ab

$$\begin{aligned} \|q_k(A)e^{(0)}\|_A^2 &= (e^{(0)})^T Q q_k(\Lambda) Q^T Q \Lambda Q^T Q q_k(\Lambda) Q^T e^{(0)} \\ &= (e^{(0)})^T Q q_k(\Lambda) \Lambda q_k(\Lambda) Q^T e^{(0)} \\ &\leq \max_{\lambda \in \{\lambda_1, \dots, \lambda_n\}} \{q_k(\lambda)^2\} (e^{(0)})^T Q \Lambda Q^T e^{(0)} \\ &\leq \left( \max_{\lambda \in [\lambda_{\min}, \lambda_{\max}]} |q_k(\lambda)| \right)^2 (e^{(0)})^T A e^{(0)}, \end{aligned} \quad (4.28)$$

d.h. es gilt

$$\|q_k(A)e^{(0)}\|_A \leq \|e^{(0)}\|_A \cdot \max_{\lambda \in [\lambda_1, \lambda_n]} |q_k(\lambda)|.$$



- (4) Man sucht nun Polynome  $p_k$ , die die rechte Seite minimieren. Man kann zeigen, dass die **Tschebyscheff–Polynome**  $T_k : \mathbb{R} \rightarrow \mathbb{R}$  ( $k = 0, 1, \dots$ ) definiert als

$$T_k(x) := \frac{1}{2} \left[ \left( x + \sqrt{x^2 - 1} \right)^k + \left( x - \sqrt{x^2 - 1} \right)^k \right]$$

auf  $[-1, 1]$  folgende Eigenschaften haben

$$T_k(1) = 1, \quad |T_k(x)| \leq 1 \quad \forall -1 \leq x \leq 1$$

und minimal bzgl.  $\|\cdot\|_\infty$  unter allen Polynomen  $p \in \mathbb{P}_k$  mit  $p(1) = 1$  sind (siehe Aufgabe 4.7.9). Gesucht ist nun eine Transformation, die  $[\lambda_{\min}, \lambda_{\max}]$  auf  $[-1, 1]$  abbildet. Somit ist  $\tilde{T}_k(z) := T_k\left(\frac{\lambda_{\max} + \lambda_{\min} - 2z}{\lambda_{\min} - \lambda_{\max}}\right)$  minimal bzgl.  $\|\cdot\|_\infty$  auf  $[\lambda_{\min}, \lambda_{\max}]$  unter allen Polynomen aus  $p \in \mathbb{P}_k$  mit  $p(\lambda_{\max}) = 1$ . Man wählt also

$$q_k(z) := T_k\left(\frac{\lambda_{\max} + \lambda_{\min} - 2z}{\lambda_{\min} - \lambda_{\max}}\right) / T_k\left(\frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\min} - \lambda_{\max}}\right)$$

auf  $[0, \lambda_{\max}]$ ,  $q_k(0) = 1$ .

Daraus folgt

$$\min_{q_k(0)=1} \max_{\lambda \in [\lambda_1, \lambda_n]} |q_k(\lambda)| \leq \frac{1}{\left| T_k\left(\frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\min} - \lambda_{\max}}\right) \right|}$$

und

$$\left| T_k\left(\frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\min} - \lambda_{\max}}\right) \right| = \left| T_k\left(\frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}}\right) \right| = \left| T_k\left(\frac{\kappa + 1}{\kappa - 1}\right) \right| \geq \frac{1}{2} \left( z + \sqrt{z^2 - 1} \right)^k$$

mit  $z = \frac{\kappa+1}{\kappa-1}$ , also

$$\begin{aligned} z + \sqrt{z^2 - 1} &= \frac{\kappa + 1}{\kappa - 1} + \sqrt{\frac{\kappa^2 + 2\kappa + 1 - \kappa^2 + 2\kappa - 1}{(\kappa - 1)^2}} \\ &= \frac{1}{\kappa - 1} (\kappa + 1 + 2\sqrt{\kappa}) = \frac{(\sqrt{\kappa} + 1)^2}{(\sqrt{\kappa} + 1)(\sqrt{\kappa} - 1)} = \frac{\sqrt{\kappa} + 1}{\sqrt{\kappa} - 1}, \end{aligned}$$

womit der Satz bewiesen ist. □

**Bemerkung 4.7.8** (i) Im letzten Beweis wurde unter (3) (siehe (4.28)) gezeigt, dass

$$\|q_k(A)e^{(0)}\|_A \leq \max_{\lambda \in \{\lambda_1, \dots, \lambda_n\}} |q(\lambda)| \|e^{(0)}\|_A$$

gilt. Daraus folgt sofort, dass das cg-Verfahren nach  $m$ -Schritten terminiert, wenn das Spektrum von  $A$  nur  $m$  verschiedene Eigenwerte besitzt.

- (ii) Man beachte, dass sich der Fehlerreduktionsfaktor im Vergleich zum Gradienten–Verfahren (vgl. Satz 4.6.4) durch die Wurzel reduziert.

**Aufgabe 4.7.9 (Tschebyscheff–Polynome)** Die **Tschebyscheff–Polynome**  $T_n$  sind orthogonal bezüglich des Skalarprodukts

$$(f, g)_\omega := \int_{-1}^1 \frac{f(x)g(x)}{\sqrt{1-x^2}} dx$$

und werden standardisiert durch  $T_n(1) = 1$ .

Man zeige, dass die **Tschebyscheff–Polynome** folgende Eigenschaften besitzen:

- (i) Sie haben stets ganzzahlige Koeffizienten;
- (ii) Der höchste Koeffizient von  $T_n$  ist  $a_n = 2^{n-1}$ ;
- (iii)  $T_n$  ist stets eine gerade Funktion, falls  $n$  gerade und eine ungerade, falls  $n$  ungerade ist;
- (iv)  $T_n(1) = 1, T_n(-1) = (-1)^n$ ;
- (v)  $|T_n(x)| \leq 1$  für  $x \in [-1, 1]$ ;
- (vi) Die Nullstellen von  $T_n(x)$  sind

$$x^{(k)} := \cos\left(\frac{2k-1}{2n}\pi\right), \quad (k = 1, \dots, n);$$

- (vii) Es gilt die Darstellung

$$T_k(x) = \begin{cases} \cos(k \cdot \arccos(x)), & -1 \leq x \leq 1; \\ \cosh(k \cdot \operatorname{arcosh}(x)), & x \geq 1; \\ (-1)^k \cosh(k \cdot \operatorname{arcosh}(-x)), & x \leq -1; \end{cases}$$

- (viii) Die **Tschebyscheff**-Polynome besitzen die globale Darstellung

$$T_k(x) = \frac{1}{2} \left[ \left( x + \sqrt{x^2 - 1} \right)^k + \left( x - \sqrt{x^2 - 1} \right)^k \right], \quad \text{wobei } x \in \mathbb{R};$$

- (ix)  $|T_n(x)|$  nimmt seinen maximalen Wert im Intervall  $[-1, 1]$  an den sogenannten **Tschebyscheff-Abszissen**  $\bar{x}_k = \cos\left(\frac{k\pi}{n}\right)$  für  $k = 0, \dots, n$  an, d.h.

$$|T_n(x)| = 1 \Leftrightarrow x = \bar{x}_k = \cos\left(\frac{k\pi}{n}\right), \quad k = 0, \dots, n.$$

- (x) Für  $x \in \mathbb{R}$  erfüllen die **Tschebyscheff**-Polynome die folgende **Drei-Term-Rekursion**

$$T_0(x) = 1, T_1(x) = x, T_k(x) = 2x T^{(k-1)}(x) - T_{k-2}(x), \quad k \geq 2.$$

Ergänzend seien für  $n = 0, \dots, 5$  die  $T_n$  explizit angegeben und diese grafisch dargestellt:

$$\begin{aligned} T_0 &= 1, & T_3 &= 4x^3 - 3x, \\ T_1 &= x, & T_4 &= 8x^4 - 8x^2 + 1, \\ T_2 &= 2x^2 - 1, & T_5 &= 16x^5 - 20x^3 + 5x, \end{aligned}$$

vgl. Abbildung 4.7.

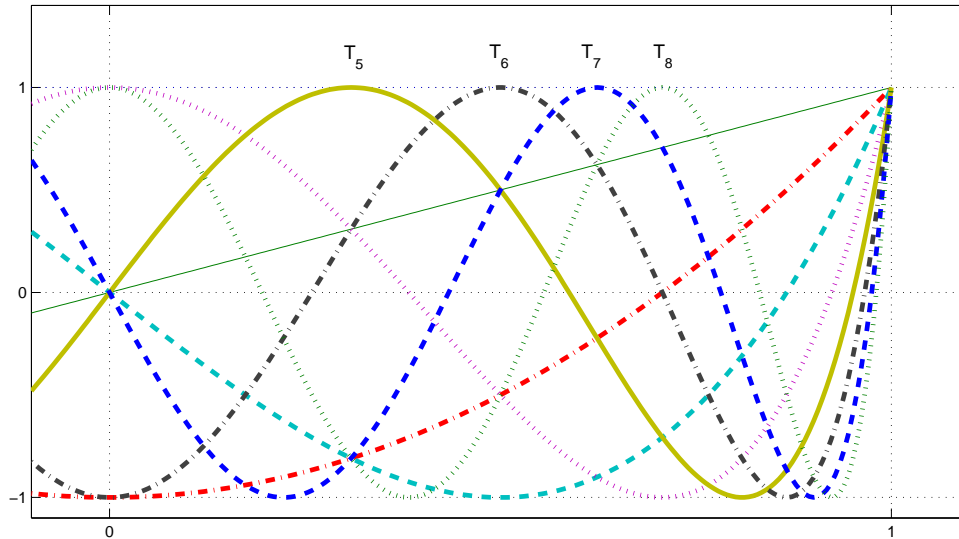
## 4.8 Vorkonditionierung, das pcg-Verfahren

Die Abschätzung der Konvergenzgeschwindigkeit des cg-Verfahrens hängt gemäß Satz 4.7.7 monoton von der Kondition  $\kappa$  der Matrix  $A$  ab. Ziel dieses Abschnittes ist es daher, das Problem  $Ax = b$  so zu transformieren, dass die entstehende Matrix möglichst „gut konditioniert“ (d.h.  $\kappa$  möglichst klein) ist.

**Idee:** Betrachte anstatt

$$Ax = b \text{ mit } A \text{ s.p.d.} \tag{4.29}$$

$$\bar{A}\bar{x} = \bar{b} \text{ mit } \bar{A} = P^{-\frac{1}{2}}AP^{-\frac{1}{2}}, \bar{x} = P^{\frac{1}{2}}x \text{ und } \bar{b} = P^{-\frac{1}{2}}b \tag{4.30}$$

Abb. 4.7: Die Tschebyscheff-Polynome  $T_1, \dots, T_8$ .

mit einer symmetrischen, positiv definiten Matrix  $P \in \mathbb{R}^{n \times n}$ , dem so genannten **Vorkonditionierer** (manchmal auch **Präkonditionierer** genannt), und wende auf  $\bar{A}\bar{x} = \bar{b}$  das cg-Verfahren an. Beachtet man, dass

$$\bar{r}^{(k)} = \bar{b} - \bar{A}\bar{x}^{(k)} = P^{-\frac{1}{2}}b - P^{-\frac{1}{2}}AP^{-\frac{1}{2}}P^{\frac{1}{2}}x^{(k)} = P^{-\frac{1}{2}}(b - Ax^{(k)}) = P^{-\frac{1}{2}}r^{(k)}$$

gilt und setzt

$$\bar{x}^{(k)} = P^{\frac{1}{2}}x^{(k)} \quad \text{sowie} \quad \bar{d}^{(k)} = P^{\frac{1}{2}}d^{(k)},$$

so erhält man

$$\begin{aligned} \bar{\alpha}_k &= \frac{\langle \bar{r}^{(k)}, \bar{d}^{(k)} \rangle}{\langle \bar{d}^{(k)}, \bar{A}\bar{d}^{(k)} \rangle} = \frac{\langle r^{(k)}, d^{(k)} \rangle}{\langle d^{(k)}, Ad^{(k)} \rangle} = \alpha_k & \Rightarrow \bar{\alpha}_k &= \alpha_k \\ \bar{x}^{(k+1)} &= \bar{x}^{(k)} + \bar{\alpha}_k \bar{d}^{(k)} = P^{\frac{1}{2}}x^{(k)} + \alpha_k P^{\frac{1}{2}}d^{(k)} & \Rightarrow x^{(k+1)} &= x^{(k)} + \alpha_k d^{(k)} \\ \bar{r}^{(k+1)} &= \bar{r}^{(k)} - \bar{\alpha}_k \bar{A}\bar{d}^{(k)} = P^{-\frac{1}{2}}r^{(k)} - \alpha_k P^{-\frac{1}{2}}Ad^{(k)} & \Rightarrow r^{(k+1)} &= r^{(k)} - \alpha_k Ad^{(k)} \\ \bar{d}^{(k+1)} &= \bar{r}^{(k+1)} - \frac{\langle \bar{r}^{(k+1)}, \bar{A}\bar{d}^{(k)} \rangle}{\langle \bar{d}^{(k)}, \bar{A}\bar{d}^{(k)} \rangle} \bar{d}^{(k)} \\ &= P^{-\frac{1}{2}}r^{(k+1)} - \frac{\langle P^{-1}r^{(k+1)}, Ad^{(k)} \rangle}{\langle d^{(k)}, Ad^{(k)} \rangle} d^{(k)} \\ &\Rightarrow d^{(k+1)} = P^{-1}r^{(k+1)} - \frac{\langle P^{-1}r^{(k+1)}, Ad^{(k)} \rangle}{\langle d^{(k)}, Ad^{(k)} \rangle} d^{(k)}. \end{aligned}$$

Insgesamt ergibt sich folgendes **pcg-Verfahren**.

**Algorithmus 4.8.1: Vorkonditionierter Konjugiertes Gradienten–Verfahren  
(pcg-Verfahren)**

**Input:** Initial guess  $x^{(0)}$

$$r^{(0)} := b - Ax^{(0)}$$

$$d^{(0)} := P^{-1}r^{(0)}$$

$$\rho_0 := \langle d^{(0)}, r^{(0)} \rangle$$

**Iteration:**  $k = 1, 2, \dots$  as long as  $k \leq n$  and  $r^{(k)} \neq 0$

$$a^{(k)} := Ad^{(k)}$$

$$\alpha_k := \frac{\rho_k}{\langle d^{(k)}, a^{(k)} \rangle}$$

$$x^{(k+1)} := x^{(k)} + \alpha_k d^{(k)}$$

$$r_{k+1} := r^{(k)} - \alpha_k a^{(k)}$$

$$q^{(k+1)} := P^{-1}r^{(k+1)}$$

$$\rho_{k+1} := \langle q^{(k+1)}, r^{(k+1)} \rangle$$

$$d^{(k+1)} := q^{(k+1)} + \frac{\rho_{k+1}}{\rho_k} d^{(k)}$$

Pro Iterationsschritt benötigt dieser Algorithmus gegenüber dem cg-Verfahren nur eine Multiplikation mit der Matrix  $P^{-1}$  mehr. Doch dieser zusätzliche Aufwand rentiert sich, sofern sich die Kondition  $\kappa(\bar{A})$  der Matrix  $\bar{A} = P^{-\frac{1}{2}}AP^{-\frac{1}{2}}$  gegenüber der Konditionszahl  $\kappa(A)$  des nicht vorkonditionierten Systems „verbessert“, d.h. abnimmt. Dies wollen wir genauer untersuchen:

Nach 4.7.7 gilt

$$\|\bar{x}^{(k)} - \bar{x}^*\|_{\bar{A}} \leq 2 \left( \frac{\sqrt{\bar{\kappa}} - 1}{\sqrt{\bar{\kappa}} + 1} \right)^k \|\bar{x}^{(0)} - \bar{x}^*\|_{\bar{A}}$$

mit  $\bar{\kappa} = \kappa_2(\bar{A})$  und  $\bar{A}\bar{x}^* = \bar{b}$ . Nun gilt zum Einen

$$\begin{aligned} \|\bar{x}^{(k)} - \bar{x}^*\|_{\bar{A}}^2 &= \langle P^{1/2}(x^{(k)} - x^*), P^{-1/2}AP^{-1/2}P^{1/2}(x^{(k)} - x^*) \rangle \\ &= \langle x^{(k)} - x^*, A(x^{(k)} - x^*) \rangle \\ &= \|x^{(k)} - x^*\|_A^2 \end{aligned}$$

und zum Anderen

$$\bar{\kappa} = \kappa_2(\bar{A}) = \frac{\lambda_{\max}(P^{-1/2}AP^{-1/2})}{\lambda_{\min}(P^{-1/2}AP^{-1/2})} = \kappa_2(P^{-1}A).$$

Denn sei  $\lambda \in \sigma(P^{-1}A)$  mit dem zugehörigen Eigenvektor  $v \in \mathbb{R}^n$ , d.h.

$$P^{-1}Av = \lambda v.$$

Setze  $\bar{v} := P^{1/2}v$ . Dann gilt

$$\bar{A}\bar{v} = P^{-1/2}AP^{-1/2}P^{1/2}v = P^{1/2}P^{-1}Av = P^{1/2}\lambda v = \lambda\bar{v}$$

und somit  $\sigma(\bar{A}) = \sigma(P^{-1}A)$ . Insgesamt erhalten wir daher die folgende Fehlerabschätzung:

$$\|x^{(k)} - x^*\| \leq 2 \left( \frac{\sqrt{\kappa_2(P^{-1}A)} - 1}{\sqrt{\kappa_2(P^{-1}A)} + 1} \right)^k \|x^{(0)} - x^*\|_A.$$

**Beispiel 4.8.1** Eine sehr einfache, aber häufig schon wirkungsvolle Vorkonditionierung einer s.p.d. Matrix  $A$  mit nichtverschwindenden Diagonalelementen liefert die Wahl  $P^{-1} := D^{-1}$ , also der Inversen der Diagonale von  $A$ . Man spricht in diesem Fall von **diagonaler Vorkonditionierung** (oder auch **Diagonalskalierung**).

**Beispiel 4.8.2** Man kann auch einige (wenige) Schritte des Jacobi- oder symmetrischen Gauß–Seidel–Verfahrens als Vorkonditionierer nutzen. Dabei entspricht Jacobi der Diagonalskalierung. Symmetrisches Gauß–Seidel bedeutet folgendes: Beim klassischen Gauß–Seidel–Verfahren ist die Iterationsmatrix  $Q_{GS} = D + L$  nicht symmetrisch und kann also nicht als Vorkonditionierer verwendet werden. Wähle also

$$Q_{SGS} := (D + L)D^{-1}(D + L)^T = (D + L)D^{-1}(D + R),$$

welches eine symmetrische (Iterations-)Matrix ist.

Man stellt die Iterationsmatrix **nicht** auf. Im pcg–Verfahren braucht man nur die Anwendung des Vorkonditionierers auf einen Vektor  $r^{(k)}$  zu kennen. Beim symmetrischen Gauß–Seidel–Vorkonditionierer bedeutet dies, dass  $r^{(k)}$  der Startwert für einige Schritte des iterativen Verfahrens ist.

**Bemerkung 4.8.3** Für Matrizen  $A_h$ , die aus der Diskretisierung der Gitterweite  $h > 0$  von elliptischen partiellen Differenzialgleichungen herrühren (wie unsere Standardmatrix in den Beispielen 3.0.1 und 4.0.3) gibt es (asymptotisch) **optimale** Vorkonditionierer:

$$\kappa(A_h B_h) = \mathcal{O}(1), \quad h \rightarrow 0$$

(Oswald 1988, Bramble–Pascial–Xu 1989). Dies bedeutet, dass man einen Anfangsfehler um einen Faktor reduzieren kann mit einer Anzahl von pcg–Schritten, die **unabhängig** von der Gitterweite  $h$  (und damit der Matrixdimension) ist.

**Bemerkung 4.8.4** Es ist klar, dass die Wahl des Vorkonditionierers  $P$  elementar von den Eigenschaften der Matrix  $A$  abhängt. Die Wahl eines geeigneten Vorkonditionierers ist daher oft nicht leicht. Hat man aber einen gefunden, so erhöht sich die Konvergenzgeschwindigkeit beträchtlich, was besonders für große lineare Gleichungssysteme von Bedeutung ist.



# 5 LINEARE AUSGLEICHSPROBLEME

**Beispiel 5.0.1** Bestimmung eines unbekannten Widerstands  $x$  aus Messungen für die Stromstärke  $t$  und die Spannung  $b$ . Angenommen, es liegen  $m$  Messungen  $(b_i, t_i)$ ,  $i = 1, \dots, m$ , mit  $m \gg 1$ ,

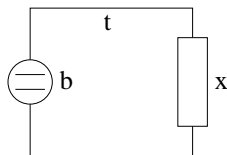


Abb. 5.1: Widerstandsbestimmung aus Messungen für Spannung und Stromstärke.

für Spannung und Stromstärke vor. Das **Ohm'sche Gesetz** aus der Physik besagt:  $b = tx$ , also

$$b_i = t_i x, \quad i = 1, \dots, m. \quad (5.1)$$

Da die Messdaten in der Regel (Mess-)Fehler beinhalten, kann man (5.1) nicht exakt für alle  $i$  erfüllen. Gesucht ist jetzt also ein Widerstand  $x$ , der „möglichst gut“ zu den Messdaten passt. Oder genauer: Wir suchen den Wert  $x \in \mathbb{R}$ , der im quadratischen Mittel möglichst nahe an den Messwerten liegt, also

$$F(x) := \sum_{i=1}^m (t_i x - b_i)^2 \rightarrow \min, \quad x \in \mathbb{R}.$$

Die Lösung lässt sich leicht bestimmen, denn wegen

$$F'(x) = 2 \sum_{i=1}^m t_i (t_i x - b_i) = x \cdot 2 \sum_{i=1}^m t_i^2 - 2 \sum_{i=1}^m t_i b_i,$$

gilt  $x = \frac{t^T b}{\|t\|^2}$ . Dass es sich hierbei tatsächlich um ein Minimum handelt zeigt

$$F''(x) = 2 \sum_{i=1}^m t_i^2 = 2\|t\|^2 > 0.$$

Man kann dieses Beispiel auch so interpretieren, dass man diejenige Ursprungsgerade

$$y(t) = x \cdot t, \quad t \in \mathbb{R},$$

mit Steigung  $x$  sucht, die im quadratischen Mittel am Nächsten zu den Messwerten

$$(t_i, b_i), \quad i = 1, \dots, m,$$

liegt.

Als nächstes - nur wenig komplizierteres - Beispiel, betrachten wir nun eine Gerade, die nicht durch den Ursprung gehen muss. Wir suchen also eine Gerade

$$g(t) = x_1 t + x_2,$$

die im quadratischen Mittel am Nächsten zu gegebenen Punkten (Messwerten)

$$(t_i, b_i), \quad i = 1, \dots, m,$$

liegt (vgl. Abbildung 5.2). Gemäß obiger Motivation gilt es zu gegebenen Punktepaaren  $(t_1, b_1), (t_2, b_2), \dots, (t_m, b_m)$  mit  $m \geq 2$ , die in der Geradengleichung  $g(t) = x_1 t + x_2$  freien Parameter  $x_1$  und  $x_2$  derart zu bestimmen, dass

$$F(x) := \sum_{i=1}^m (g(t_i) - b_i)^2 \rightarrow \min, \quad x \in \mathbb{R}^2.$$

Wegen  $F(x) = \sum_{i=1}^m (x_1 t_i + x_2 - b_i)^2$  erhalten wir die alternative Formulierung

$$\min_{x=(x_1, x_2)^T \in \mathbb{R}^2} \|Ax - b\|_2^2,$$

wobei wir nachfolgende Definitionen verwenden

$$A = \begin{pmatrix} t_1 & 1 \\ \vdots & \vdots \\ t_m & 1 \end{pmatrix} \in \mathbb{R}^{m \times 2}, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \in \mathbb{R}^m.$$

Die Minimierung im quadratischen Mittel ist ein allgemeines Prinzip, welches nun im Detail beschrieben werden soll.

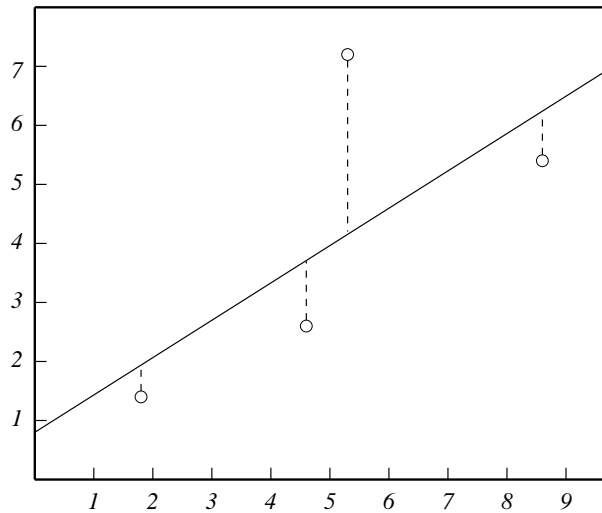


Abb. 5.2: Die Ausgleichsgerade, die die Quadratsumme der Abstände minimiert

## 5.1 Die Methode der kleinsten Quadrate

Im folgenden betrachten wir für  $A \in \mathbb{R}^{m \times n}$ ,  $m \geq n$  und  $b \in \mathbb{R}^n$  das etwas allgemeinere Problem

$$F(x) := \|Ax - b\|_2^2 \rightarrow \min, \quad x \in \mathbb{R}^n. \quad (5.2)$$

Zu gegebenen “Beobachtungen”  $b \in \mathbb{R}^m$  werden also die zugehörigen “optimalen” Parameter  $x \in \mathbb{R}^n$  gesucht, wobei wir einen linearen Zusammenhang zwischen Parametern und Beobachtungen vermuten, welches durch die Matrix  $A$  ausgedrückt wird. Für  $m = n$  ist die Lösung  $x \in \mathbb{R}^n$  von 5.2 durch das lineare Gleichungssystem  $Ax = b$  gegeben. Um 5.2 zu lösen, betrachte

$$\begin{aligned} F(x) &= \langle Ax - b, Ax - b \rangle \\ &= (Ax - b)^T (Ax - b) \\ &= x^T A^T A x - x^T A^T b - b^T A x + b^T b \\ &= x^T A^T A x - 2x^T A^T b + b^T b \end{aligned}$$



sowie

$$F'(x) = 2A^T Ax - 2A^T b \quad \text{und} \quad F''(x) = 2A^T A.$$

Man beachte hierbei, dass die zweite Ableitung symmetrisch und positiv definit ist, falls  $A$  den vollen Rang  $n$  besitzt. In diesem Fall ist die Nullstelle  $x^*$  von  $F'$  das gesuchte Minimum, also

$$A^T Ax = A^T b, \quad (5.3)$$

was als ‘‘Gaußsche Normalengleichung’’ bekannt ist.

**Bemerkung 5.1.1** Sei  $A \in \mathbb{R}^{m \times n}$ ,  $m \geq n$ , mit vollem Rang  $n$ . Dann löst  $x \in \mathbb{R}^n$  genau dann das lineare Ausgleichsproblem 5.2, wenn es die Gaußsche Normalengleichung 5.3 löst. Die Lösung ist eindeutig.  $\square$

Wir wollen die Lösung von 5.3 untersuchen und erinnern daran, dass die Kondition

$$\kappa(A) = \|A\| \cdot \|A^{-1}\| \quad (5.4)$$

nicht nur einen wesentlichen Einfluss auf die Konvergenzgeschwindigkeit des (p)cg-Verfahrens besitzt, sondern insbesondere die Fehlerverstärkung und Schwierigkeit eines Problems beschreibt. Offenbar ist 5.4 aber nur für reguläre Matrizen wohldefiniert und macht daher für  $A \in \mathbb{R}^{m \times n}$ ,  $m > n$ , keinen Sinn, da  $A$  nicht regulär sein kann.

**Bemerkung 5.1.2** Sei  $A \in \mathbb{R}^{n \times n}$  regulär. Unter Verwendung einer verträglichen Matrixnorm lässt sich zeigen, dass

$$\kappa(A) = \frac{\max_{\|x\|=1} \|Ax\|}{\min_{\|x\|=1} \|Ax\|} \quad (5.5)$$

*Beweis.* Übung.  $\square$

Offenbar ist der Ausdruck auf der rechten Seite von 5.5 auch für nicht-quadratische Matrizen sinnvoll. Wir können dies also als Definition für die Kondition verwenden.

Nun wieder zur Normalengleichung 5.3.

**Lemma 5.1.3** Sei  $A \in \mathbb{R}^{m \times n}$  mit vollem Rang  $n \leq m$ . Dann gilt

$$\kappa_2(A^T A) = \kappa_2(A)^2. \quad (5.6)$$

*Beweis.* Mit  $\|\cdot\| = \|\cdot\|_2$  gilt mit 5.5:

$$\begin{aligned} \kappa_2(A)^2 &= \frac{\max_{\|x\|_2=1} \|Ax\|_2^2}{\min_{\|x\|_2=1} \|Ax\|_2^2} = \frac{\max_{\|x\|_2=1} (Ax)^T Ax}{\min_{\|x\|_2=1} (Ax)^T Ax} \\ &= \frac{\max_{\|x\|_2=1} x^T (A^T A) x}{\min_{\|x\|_2=1} x^T (A^T A) x} = \frac{\lambda_{\max}(A^T A)}{\lambda_{\min}(A^T A)}, \end{aligned}$$

da  $A$  vollen Rang hat.  $\square$

Die Kondition der Normalengleichung beträgt demnach  $\kappa_2(A)^2$  und ist numerisch daher ungeeignet zur Lösung von 5.2.

**Bemerkung 5.1.4** Bei den einführenden Beispielen haben wir uns auf die Betrachtung von linear unabhängigen Basisvektoren des  $\mathbb{P}_1$  in der Darstellung von  $g$  beschränkt, d.h.  $1, x$ . Natürlich kann man auch zu allgemeineren Funktionen in der Funktionsvorschrift von  $g$  übergehen, wie etwa

$$g(t) = x_1 \exp(t) + x_2 \exp(-t) + x_3 \log(t).$$

Damit erhalten wir bei bekannten Daten  $(t_i, b_i)$  ( $i = 1, \dots, m$ ) folgendes Minimierungsproblem

$$\sum_{i=1}^m (g(t_i) - b_i)^2 \rightarrow \min, \quad x = (x_1, x_2, x_3)^T \in \mathbb{R}^3,$$

welches wir völlig analog zu obigem Verfahren als Normalengleichung in der Form

$$A^T A x = A^T b, \quad \text{mit } A = \begin{pmatrix} e^{t_1} & e^{-t_1} & \log t_1 \\ \vdots & \vdots & \vdots \\ e^{t_m} & e^{-t_m} & \log t_m \end{pmatrix} \quad \text{und } b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}$$

interpretieren können. Je komplexer also die Funktionsvorschrift von  $g$ , desto aufwendiger wird auch das Lösen der Gaußschen Normalengleichung 5.3. Im Folgenden werden wir ein Verfahren bereitstellen, das eine numerisch stabile und effiziente Lösung des Ausgleichsproblems garantiert.

## 5.2 Die QR-Zerlegung

**Definition 5.2.1** Eine quadratische Matrix  $Q \in \mathbb{R}^{n \times n}$  heißt **orthogonal**, falls gilt:

$$Q Q^T = Q^T Q = I.$$

Weiterhin bezeichne  $\|x\|$  die vom **Euklidischen** Skalarprodukt im  $\mathbb{R}^n$  induzierte Norm, vgl. Anhang B. Darüber hinaus halten wir noch fest, dass orthogonale Abbildungen stets längen- und winkeltreu sind, dass also gilt

$$\|Qx\|^2 = (Qx)^T Qx = x^T Q^T Qx = x^T x = \|x\|^2, \quad \text{d.h. } \|Qx\| = \|x\|.$$

**Motivation:** Zur Lösung des Minimierungsproblems  $\min_{x \in \mathbb{R}^n} \|Ax - b\|^2$  setzen wir voraus, dass zu  $A \in \mathbb{R}^{m \times n}$  ( $\text{rg } A = n < m$ ) stets eine Zerlegung  $A = QR$  existiert, wobei  $Q \in \mathbb{R}^{m \times m}$  orthogonal ist und  $R \in \mathbb{R}^{m \times n}$  folgende Gestalt besitzt

$$R = \begin{pmatrix} * & * & * \\ & * & * \\ & & * \\ 0 & & \end{pmatrix} = \begin{pmatrix} \hat{R} \\ \hline 0 \end{pmatrix},$$

$\hat{R}$  stellt dabei eine reguläre obere Dreiecksmatrix dar. Damit können wir unser bisheriges Minimierungsproblem wie folgt modifizieren (beachte die Faktorisierung  $A = QR$ ):

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|^2 = \min_{x \in \mathbb{R}^n} \|Q^T(Ax - b)\|^2 = \min_{x \in \mathbb{R}^n} \|Rx - Q^T b\|^2.$$

Hierbei weisen wir vor allen Dingen nochmals auf die Gestalt von  $R$  hin:

$$Rx = \begin{pmatrix} * & * & * \\ & * & * \\ & & * \\ 0 & & \end{pmatrix} \cdot x = \begin{pmatrix} b_1 \\ \hline b_2 \end{pmatrix} = Q^T b$$

Dieser Faktorisierungsansatz führt uns nun zu folgenden Fragen:

- Zu welchen Matrizen  $A \in \mathbb{R}^{m \times n}$  existiert eine derartige QR-Zerlegung?
- Wie bestimmen wir eine solche Zerlegung möglichst effizient?

**Satz 5.2.2** Zu jeder Matrix  $A \in \mathbb{R}^{m \times n}$  mit Maximalrang  $n < m$  existiert eine orthogonale Matrix  $Q \in \mathbb{R}^{m \times m}$  derart, dass

$$A = QR \text{ mit } R = \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix}, \hat{R} \in \mathbb{R}^{n \times n}$$

gilt, wobei  $\hat{R}$  eine reguläre obere Dreiecksmatrix darstellt.

*Beweis.* Der Kern des Beweises liegt in der Verwendung von orthogonalen Drehmatrizen der Form

$$\mathcal{U}(p, q; \varphi) := \begin{pmatrix} 1 & & & & & & & & 0 \\ & 1 & & & & & & & \\ & & \ddots & & & & & & \\ & & & \cos \varphi & & \sin \varphi & & & \\ & & & -\sin \varphi & & \cos \varphi & & & \\ & & & & \ddots & & & & \\ & & & & & 1 & & & \\ 0 & & & & & & & & 1 \end{pmatrix} \quad \begin{matrix} \downarrow p\text{-te Spalte} \\ \downarrow q\text{-te Spalte} \end{matrix} \quad \begin{matrix} \leftarrow p\text{-te Zeile} \\ \leftarrow q\text{-te Zeile} \end{matrix} \quad (5.7)$$

Die Orthogonalität von  $\mathcal{U}(p, q; \varphi)$  ergibt sich unmittelbar, denn es gilt  $\mathcal{U}^T(p, q; \varphi) \cdot \mathcal{U}(p, q; \varphi) = I$ . Das Produkt  $\tilde{A} := \mathcal{U}^T(p, q; \varphi) \cdot A$  liefert dagegen für  $j = 1, \dots, n$

$$\begin{pmatrix} 1 & & & & & & & & 0 \\ & \ddots & & & & & & & \\ & & \cos \varphi & & -\sin \varphi & & & & \\ & & \sin \varphi & & \cos \varphi & & & & \\ & & & \ddots & & & & & \\ 0 & & & & & & & & 1 \end{pmatrix} \cdot A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{p-1,1} & \cdots & a_{p-1,n} \\ a_{p1} \cos \varphi - a_{q1} \sin \varphi & \cdots & a_{pn} \cos \varphi - a_{qn} \sin \varphi \\ a_{p+1,1} & \cdots & a_{p+1,n} \\ \vdots & & \vdots \\ a_{q-1,1} & \cdots & a_{q-1,n} \\ a_{p1} \sin \varphi + a_{q1} \cos \varphi & \cdots & a_{pn} \sin \varphi + a_{qn} \cos \varphi \\ a_{q+1,1} & \cdots & a_{q+1,n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

Demzufolge werden die Zeilen  $p$  und  $q$  von  $A$  durch Linearkombinationen der cos- und sin-Terme ersetzt.

Die sukzessive Multiplikation der Matrix  $A$  von links mit Rotationsmatrizen  $\mathcal{U}^T(p, q; \varphi)$  ( $\varphi$  dabei so gewählt, dass  $\tilde{a}_{q,p} = 0$ ) mit den Rotationsindexpaaren

$$(1, 2), (1, 3), \dots, (1, m), (2, 3), (2, 4), \dots, (2, m), (3, 4), \dots, (n, m)$$

eliminiert dann die Matricelemente  $a_{2,1}, a_{3,1}, \dots, a_{m,n}$ . Sukzessive werden dabei unterhalb des Diagonalelements Nullen erzeugt. Nach  $k := \frac{1}{2}n(2m - n - 1)$  Transformationsschritten gilt dann  $A = QR$  mit

$$\mathcal{U}_k^T \cdot \dots \cdot \mathcal{U}_2^T \mathcal{U}_1^T A = Q^T A = R \quad \text{oder} \quad A = QR.$$

Da die orthogonale Matrix  $Q$  regulär ist, ist der Rang von  $A$  und der Rang von  $R$  gleich  $n$  und folglich ist die Rechtecksdreiecksmatrix  $\hat{R}$  regulär, da  $A$  nach Voraussetzung Maximalrang hat.  $\square$

Der soeben erbrachte Existenzbeweis ist rein konstruktiv und liefert damit eine erste Möglichkeit eine Zerlegung  $QR$  zu bestimmen. Wir unterscheiden die folgenden beiden Vorgehensweisen:

- **Givens**<sup>1</sup>-Rotation,
- **Householder**<sup>2</sup>-Spiegelung.

### 5.3 Givens-Rotationen

Es bezeichne im Folgenden stets  $c := \cos \varphi$  und  $s := \sin \varphi$ . Die Grundidee der **Givens**-Rotation besteht darin, durch die Multiplikation mit einer geeigneten orthogonalen (Dreh-)Matrix einen Vektor  $x \in \mathbb{R}^n$  so zu drehen, dass möglichst viele seiner Komponenten verschwinden.

Im Falle  $(a, b) \in \mathbb{R}^2$  werden wir also  $c, s \in \mathbb{R}$  so bestimmen, dass gilt

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} r \\ 0 \end{pmatrix}.$$

Setzen wir zuerst  $a, b \neq 0$  voraus. Wir erhalten damit folgende Gleichung

$$as = bc \tag{5.8}$$

und mit  $s^2 + c^2 = 1$  gilt

$$r^2 = (ac + bs)^2 = a^2 c^2 + abcs + abcs + b^2 s^2 \stackrel{(5.8)}{=} a^2 c^2 + a^2 s^2 + b^2 c^2 + b^2 s^2 = a^2 + b^2.$$

Weiter ergibt sich für  $a \neq 0$

$$\begin{aligned} ac + bs = r &\Rightarrow c = \frac{1}{a}(r - bs) &\Rightarrow as - \frac{b}{a}(r - bs) = 0 \\ &\Leftrightarrow a^2 s - br + b^2 s = 0 \\ &\Leftrightarrow br = (a^2 + b^2)s \\ &\Leftrightarrow s = \frac{br}{a^2 + b^2} \stackrel{(r^2=a^2+b^2)}{=} \frac{b}{r}. \end{aligned}$$

Und damit erhalten wir dann auch für  $b \neq 0$

$$bc = as \Rightarrow c = \frac{a}{b}s = -\frac{a}{r}.$$

Offen ist noch die Wahl des Vorzeichens des Vorzeichens  $r = \pm\sqrt{a^2 + b^2}$ . Setzen wir  $r := \text{sign}(a)\sqrt{a^2 + b^2}$ , so folgt  $c > 0$ , falls  $a \neq 0$ , und für  $a = 0$  setzen wir  $c = 0$  und  $s = -1$ . Somit sind auch  $s$  und  $c$  für die anderen Spezialfälle  $b = 0$  und  $a = b = 0$  wohldefiniert. Damit

<sup>1</sup>Givens, W.

<sup>2</sup>Householder, A.

gilt (obwohl der Winkel bisher gar nicht explizit genannt wurde)  $\varphi \in (-\frac{\pi}{2}, \frac{\pi}{2}]$  und somit lässt sich  $\cos \varphi$  eindeutig aus  $\sin \varphi$  mit  $\cos \varphi = \sqrt{1 - \sin^2 \varphi}$  bestimmen.

Seien  $p$  und  $q$  zwei Zeilenindizes von  $A$ . Zur Eliminierung des Matrixelements  $a_{q,j}$  gehen wir daher wie folgt vor:

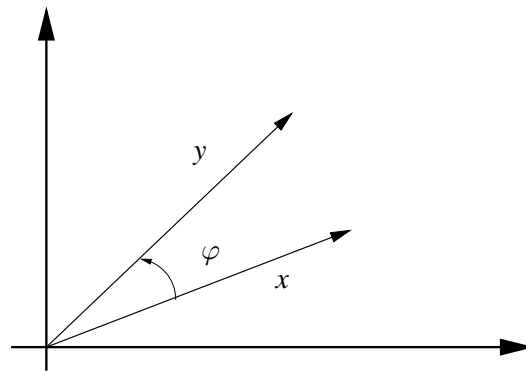
$$\begin{aligned} \text{Falls } a_{p,j} \neq 0: \quad \cos \varphi &= \frac{|a_{p,j}|}{\sqrt{a_{p,j}^2 + a_{q,j}^2}}, \\ \sin \varphi &= \frac{-\operatorname{sign}(a_{p,j}) a_{q,j}}{\sqrt{a_{p,j}^2 + a_{q,j}^2}}, \\ a_{p,j} = 0: \quad \cos \varphi &= 0, \quad \sin \varphi = 1. \end{aligned}$$

Da sich  $\cos \varphi$  aus  $\sin \varphi$  bestimmen lässt, ermöglicht das Verfahren eine effiziente Speicherung, denn es genügt  $\sin \varphi$  an den entsprechenden freiwerdenden Stellen in  $A$  abzulegen:

$$\begin{aligned} A &= \begin{pmatrix} a_{1,1}^{(0)} & \cdots & a_{1,n}^{(0)} \\ \vdots & & \vdots \\ a_{m,1}^{(0)} & \cdots & a_{m,n}^{(0)} \end{pmatrix} \rightsquigarrow \begin{pmatrix} a_{1,1}^{(1)} & \cdots & a_{1,n}^{(1)} \\ \sin \varphi_{2,1} & a_{2,2}^{(1)} & \cdots & a_{2,n}^{(1)} \\ a_{3,1}^{(0)} & a_{3,2}^{(0)} & \cdots & a_{3,n}^{(0)} \\ a_{4,1}^{(0)} & a_{4,2}^{(0)} & \cdots & a_{4,n}^{(0)} \\ \vdots & & & \vdots \\ a_{m,1}^{(0)} & \cdots & \cdots & a_{m,n}^{(0)} \end{pmatrix} \rightsquigarrow \begin{pmatrix} a_{1,1}^{(2)} & \cdots & a_{1,n}^{(2)} \\ \sin \varphi_{2,1} & a_{2,2}^{(1)} & \cdots & a_{2,n}^{(1)} \\ \sin \varphi_{3,1} & a_{3,2}^{(2)} & \cdots & a_{3,n}^{(2)} \\ a_{4,1}^{(0)} & a_{4,2}^{(0)} & \cdots & a_{4,n}^{(0)} \\ \vdots & & & \vdots \\ a_{m,1}^{(0)} & \cdots & \cdots & a_{m,n}^{(0)} \end{pmatrix} \\ &\rightsquigarrow \begin{pmatrix} a_{1,1}^{(3)} & \cdots & a_{1,n}^{(3)} \\ \sin \varphi_{2,1} & a_{2,2}^{(1)} & \cdots & a_{2,n}^{(1)} \\ \sin \varphi_{3,1} & a_{3,2}^{(2)} & \cdots & a_{3,n}^{(2)} \\ \sin \varphi_{4,1} & a_{4,2}^{(3)} & \cdots & a_{4,n}^{(3)} \\ \vdots & & & \vdots \\ a_{m,1}^{(0)} & \cdots & \cdots & a_{m,n}^{(0)} \end{pmatrix} \rightsquigarrow \begin{pmatrix} a_{1,1}^{(m-1)} & \cdots & a_{1,n}^{(m-1)} \\ \sin \varphi_{2,1} & a_{2,2}^{(1)} & \cdots & a_{2,n}^{(1)} \\ \sin \varphi_{3,1} & a_{3,2}^{(2)} & \cdots & a_{3,n}^{(2)} \\ \sin \varphi_{4,1} & a_{4,2}^{(3)} & \cdots & a_{4,n}^{(3)} \\ \vdots & & & \vdots \\ \sin \varphi_{m,1} & a_{m,2}^{(m-1)} & \cdots & a_{m,n}^{(m-1)} \end{pmatrix} \rightsquigarrow \begin{pmatrix} & & & \hat{R} \\ \ddots & & & \\ & \ddots & & \\ \text{„}Q\text{“} & & \ddots & \end{pmatrix}. \end{aligned}$$

Die **QR-Zerlegung** liefert also  $QR$  in kompakter Form, gespeichert in  $A$ . Der Aufwand dieser Vorgehensweise liegt bei ungefähr  $\mathcal{O}(n^3)$  Operationen.

**Bemerkung 5.3.1** Die Multiplikation eines gegebenen Vektors  $x \in \mathbb{R}^n$  mit der Drehmatrix  $\mathcal{U}(p, q; \varphi)$  ist äquivalent zur Drehung von  $x$  um einen Winkel  $\varphi$  entgegen dem Uhrzeigersinn in der Koordinatenebene. Vergleiche hierzu die nachfolgende Abbildung, die diesen geometrischen Sachverhalt verdeutlicht:

Abb. 5.3: Drehung um einen Winkel  $\varphi$  in der Ebene**MATLAB-Funktion: GivensRotMat.m**

```

1 function rot = GivensRotMat(ap, aq)
2 if ap == 0
3     s = 1; c = 0;
4 else
5     dist2 = sqrt(ap^2+aq^2);
6     c = abs(ap)/dist2;
7     s = -sign(ap)*aq/dist2;
8 end
9 rot = [c, s; -s, c];

```

**MATLAB-Funktion: Givens.m**

```

1 function [Q,A] = Givens(A)
2 Q = eye(size(A,1));
3 for j=1:size(A,2)
4     for i=j+1:size(A,1)
5         rot = GivensRotMat(A(j,j), A(i,j));
6         A([j,i], j:end) = rot' * A([j,i], j:end);
7         Q(:, [j,i]) = Q(:, [j,i]) * rot;
8     end
9     A(j+1:end, j) = 0;
10 end

```

**MATLAB-Beispiel:**

Alternativ zu der Matlab-Funktion `qr` können wir auch die Routine `Givens` verwenden, die eine  $QR$ -Zerlegung, wie oben diskutiert, berechnet.

```
>> A = [1,2;3,4;5,6];
> [Q,R] = Givens(A);
>> Q*R
ans =
    1.0000    2.0000
    3.0000    4.0000
    5.0000    6.0000
```

## 5.4 Update einer QR-Zerlegung

Bei verschiedenen numerischen Verfahren müssen wiederholt  $QR$ -Zerlegungen bestimmt werden, wobei sich die Matrizen nur wenig voneinander unterscheiden. Dies ist z.B. der Fall beim Broyden-Verfahren (siehe Numerik II, nichtlineare Gleichungen).

Sei die Zerlegung  $A_k = Q_k R_k$  gegeben und die  $QR$ -Zerlegung von  $A_{k+1} = A_k + st^T$  zu bestimmen. Wir werden zeigen, dass anstatt jeweils  $\mathcal{O}(n^3)$  Operationen für eine vollständige  $QR$ -Zerlegung zu verwenden, für jeden einzelnen Update-Schritt  $\mathcal{O}(n^2)$  Operationen genügen.

Wir benötigen noch einige Hilfsmittel, um diese Aussage zu zeigen.

**Definition 5.4.1 (Hessenberg-Matrix)** Eine Matrix  $A \in \mathbb{R}^{n \times n}$  der Form

$$A = \begin{pmatrix} * & \cdots & \cdots & \cdots & * \\ * & & & & \vdots \\ 0 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & & \vdots \\ 0 & \cdots & 0 & * & * \end{pmatrix},$$

d.h.  $a_{ij} = 0$  für  $i - j \geq 2$ , wird obere **Hessenberg-Matrizen** genannt. Gilt  $a_{ij} = 0$  für  $j - i \geq 2$ , so hat die Matrix  $A$  untere Hessenberg-Gestalt.

**Lemma 5.4.2** Es seien  $u, v \in \mathbb{R}^n$  und  $R \in \mathbb{R}^{n \times n}$  eine obere Dreiecksmatrix. Dann benötigt man  $n - 2$  Givens-Rotationen, um  $R + uv^T$  auf obere Hessenberg-Gestalt zu transformieren.

*Beweis.* Multipliziert man  $R$  sukzessive von links mit Drehmatrizen  $\mathcal{U}^T(p, q)$  (siehe (5.7)) zu Indexpaaren  $(n, n - 1), \dots, (4, 3), (3, 2)$  so hat  $QR$  obere Hessenberg-Gestalt, wobei  $Q := \mathcal{U}_{32} \cdots \mathcal{U}_{n, n-1}$  ist. Dabei lassen sich die Drehmatrizen so wählen, dass  $Qu = (*, *, 0, \dots, 0)^T$  gilt. Beachtet man nun, dass  $(Qu)v^T$  und somit auch  $Q(R + uv^T)$  obere Hessenberg-Form hat, so ist die Aussage bewiesen.  $\square$

**Bemerkung 5.4.3** Beachtet man, dass sich jede Givens-Rotation angewandt auf  $A \in \mathbb{R}^{n \times n}$  durch  $\mathcal{O}(n)$  Operationen realisieren lässt, so genügen  $\mathcal{O}(n^2)$  Operationen, um eine orthogonale Matrix  $\tilde{Q}$  und eine Hessenberg-Matrix  $\tilde{H}$  mit  $\tilde{Q}\tilde{H} = R + uv^T$  zu bestimmen.

**Bemerkung 5.4.4** Es genügen  $n - 1$  Givens-Rotation, um eine Hessenberg-Matrix  $A \in \mathbb{R}^{n \times n}$  auf Dreiecksgestalt zu transformieren.

Kommen wir nun zu unserer ursprünglichen Fragestellung zurück, d.h. der effizienten Update-Berechnung. Der Ansatz

$$A_{k+1} = Q_{k+1}R_{k+1} = Q_k(R_k + uv^T) = A_k + Q_k uv^T = A_k + st^T$$

liefert  $v = t$  und  $s = Q_k u$  bzw.  $u = Q_k^T s$ . Das Update besteht nun aus zwei Schritten.

1. Mittels  $n - 2$  Givens-Rotationen wird  $R_k + uv^T$  auf eine obere Hessenbergform gebracht, d.h. wir bestimmen eine orthogonale Matrix  $Q'_{k+1}$  und eine Hessenberg-Matrix  $H_{k+1}$  mit  $Q'_{k+1}H_{k+1} = R_k + uv^T = R_k + (Q_k^T s)t^T$ .
2. Mittels  $n - 1$  Givens-Rotationen wird diese obere Hessenbergform zur neuen Matrix  $R_{k+1}$  transformiert, d.h. wir bestimmen eine orthogonale Matrix  $Q''_{k+1}$  und eine obere Dreiecksmatrix  $R_{k+1}$  mit  $Q''_{k+1}R_{k+1} = H_{k+1}$ . Die Matrix  $Q_{k+1}$  ergibt sich dann als Produkt  $Q_k \cdot Q'_{k+1} \cdot Q''_{k+1}$ .

### MATLAB-Funktion: QR2Hessenberg.m

```
1 function [Q,R] = QR2Hessenberg(Q,R,s,t)
2 % computes QH-decomposition of Q(R+s*t), H upper Hessenberg form
3 for j=size(R,1):-1:3
4     rot = GivensRotMat(s(j),s(j-1));
5     s([j,j-1]) = rot * s([j,j-1]);
6     Q(:, [j,j-1]) = Q(:, [j,j-1]) * rot';
7     R([j,j-1], :) = rot * R([j,j-1], :);
8 end
9 R(1:2, :) = R(1:2, :) + s(1:2) * t';
```

### MATLAB-Funktion: GivensHessenberg.m

```
1 function [Q,H] = GivensHessenberg(Q,H)
2 % computes QR-decomposition of QH, H Hessenberg matrix
3 for j=1:min(size(H,1)-1, size(H,2))
4     rot = GivensRotMat(H(j+1,1),H(j,1));
5     H([j+1,j], j:end) = rot * H([j+1,j], j:end);
6     Q(:, [j+1,j]) = Q(:, [j+1,j]) * rot';
7     H(j+1,j) = 0;
8 end
```



**MATLAB-Beispiel:**

Ein einfaches Beispiel zeigt, dass man das  $\mathcal{O}(n^2)$ -Verhalten des Updates auch im Vergleich zur optimierten Matlab-Routine `qr` sieht und nutzen kann. Verdoppelt man z.B.  $n$  auf  $n = 4000$  im nebenstehenden Listing, so benötigt das Update nur  $2^2 = 4$ -mal länger, die interne Funktion jedoch  $2^3 = 8$ -mal. Testen Sie es einmal!

```
>> n = 2000;
>> A=rand(n); s=rand(n,1); t=rand(n,1);
>> An = A + s * t';
>> [Q,R] = qr(A);
>> tic
>> [Qtmp,H] = QR2Hessenberg(Q,R,Q'*s,t);
>> [Qn,Rn] = GivensHessenberg(Qtmp,H);
>> toc
Elapsed time is 1.078000 seconds.
>> tic, [Qn,Rn] = qr(An); toc
Elapsed time is 5.766000 seconds.
```

## 5.5 Householder–Spiegelungen

**Definition 5.5.1** Sei  $\omega \in \mathbb{R}^n$ . Die  $n \times n$ -Matrix

$$P := I - 2 \frac{\omega \omega^T}{\omega^T \omega} \quad (5.9)$$

wird als **Householder-Transformation** und  $\omega$  als **Householder-Vektor** bezeichnet.

**Satz 5.5.2** Eine **Householder-Transformation**  $P = I - 2(\omega \omega^T)/(\omega^T \omega)$  ist symmetrisch und orthogonal, d.h.  $P^{-1} = P^T$ .

*Beweis.* Da  $\omega \omega^T$  symmetrisch ist  $((\omega \omega^T)^T = \omega \omega^T)$ , folgt

$$P^T = \left( I - 2 \frac{\omega \omega^T}{\omega^T \omega} \right)^T = I - 2 \frac{\omega \omega^T}{\omega^T \omega} = P.$$

Des Weiteren ergibt sich

$$P P^T = \left( I - 2 \frac{\omega \omega^T}{\omega^T \omega} \right) \left( I - 2 \frac{\omega \omega^T}{\omega^T \omega} \right) = I - 2 \frac{\omega \omega^T}{\omega^T \omega} - 2 \frac{\omega \omega^T}{\omega^T \omega} + 4 \frac{\omega \omega^T \omega \omega^T}{(\omega^T \omega)^2}$$

und mit  $\omega \omega^T \omega \omega^T = (\omega^T \omega) (\omega \omega^T)$  somit die Behauptung.  $\square$

Betrachten wir nun folgende Frage: Sei  $x \in \mathbb{R}^n$ . Wie müsste ein  $\omega \in \mathbb{R}^n$  aussehen, so dass  $Px = \alpha e_1$  gelte ( $\alpha \in \mathbb{R}$ ,  $e_1$  erster kanonischer Einheitsvektor)?

Mit  $Px = x - 2 \frac{\omega \omega^T}{\omega^T \omega} x = x - \lambda \omega \stackrel{!}{=} \alpha e_1$  folgt  $\omega \in \text{span}\{x - \alpha e_1\}$ .

Wir setzen nun  $\omega = x - \alpha e_1$  in  $Px = \alpha e_1$  ein und erhalten

$$Px = x - 2 \frac{\omega^T x}{\omega^T \omega} \omega = \left( 1 - \frac{2(x - \alpha e_1)^T x}{\|x - \alpha e_1\|^2} \right) x + \alpha \frac{2(x - \alpha e_1)^T x}{\|x - \alpha e_1\|^2} e_1 = \alpha e_1.$$

Damit in der letzten Gleichung der Faktor vor  $x$  verschwindet, muss

$$1 = \frac{2(x - \alpha e_1)^T x}{\|x - \alpha e_1\|^2} \Leftrightarrow (x - \alpha e_1)^T (x - \alpha e_1) = 2x^T x - 2\alpha x_1 \Leftrightarrow \alpha = \pm \sqrt{x^T x}$$

gelten.

Wie ist nun das Vorzeichen zu wählen,  $\alpha = \pm\sqrt{x^T x}$ ? Die Wahl  $\alpha = \|x\|_2$  hat die schöne Eigenschaft, dass  $Px$  ein positives Vielfaches von  $e_1$  ist. Aber das Rezept ist gefährlich, wenn  $x$  annähernd ein positives Vielfaches von  $e_1$  ist, da Auslöschungen auftreten können. Berechnet man  $\omega_1$  mittels  $\omega_1 = x_1 - \|x\|_2$  treten für  $x_1 \leq 0$  keine Auslöschungen auf und mit der Umformung

$$\omega_1 = x_1 - \|x\|_2 = \frac{x_1^2 - \|x\|_2^2}{x_1 + \|x\|_2} = \frac{-(x_2^2 + \dots + x_n^2)}{x_1 + \|x\|_2}$$

treten auch für  $x_1 > 0$  keine Auslöschungen auf. Man beachte außerdem, dass  $(\omega_2, \dots, \omega_n) = (x_2, \dots, x_n)$ .

Normiert man den Vektor  $\omega$  so, dass  $\omega_1 = 1$  gilt, d.h.

$$\omega := \frac{x - \alpha e_1}{x_1 - \alpha} \quad \text{mit } \alpha^2 = \|x\|^2 \quad (\omega^T e_1 = 1),$$

dann folgt

$$\omega^T \omega = \frac{(x - \alpha e_1)^T (x - \alpha e_1)}{(x_1 - \alpha)^2} = \frac{\|x\|^2 - 2\alpha x^T e_1 + \alpha^2}{(x_1 - \alpha)^2} = \frac{2\alpha(\alpha - x_1)}{(x_1 - \alpha)^2} = \frac{2\alpha}{\alpha - x_1}.$$

Des Weiteren bleibt festzuhalten, dass bei der Speicherung der Matrix  $P = I - 2(\omega\omega^T)/(\omega^T\omega)$ ,  $P \in \mathbb{R}^{n \times n}$ , auch nur der Vektor  $(\omega_2, \dots, \omega_n)^T$  zu berücksichtigen ist, der sich auf den  $n - 1$  freigebliebenen Einträgen speichern lässt.

#### Algorithmus 5.5.1: Berechnung Householder-Vektor:

Zu gegebenem  $x \in \mathbb{R}^n$  berechnet die Funktion ein  $\omega \in \mathbb{R}^n$  mit  $\omega(1) = 1$  und  $\beta \in \mathbb{R}$ , so dass  $P = I_n - \beta\omega\omega^T$  orthogonal ist und  $Px = \|x\|_2 e_1$ .

```

function:  $[\omega, \beta] = \text{housevector}(x)$ 
 $n = \text{length}(x)$ 
 $\sigma = x(2:n)^T x(2:n)$ 
 $\omega = \begin{bmatrix} 1 \\ x(2:n) \end{bmatrix}$ 
if  $\sigma = 0$ 
     $\beta = 0$ 
else
     $\mu = \sqrt{x(1)^2 + \sigma}$ 
    if  $x(1) \leq 0$ 
         $\omega(1) = x(1) - \mu$ 
    else
         $\omega(1) = -\sigma / (x(1) + \mu)$ 
    end
     $\beta = 2\omega(1)^2 / (\sigma + \omega(1)^2)$ 
     $\omega = \omega / \omega(1)$ 
end

```

**MATLAB-Funktion: HouseholderVektor.m**

```

1 function [v,beta] = HouseholderVector(x)
2 n = length(x);
3 if n>1
4     sigma = x(2:end)'*x(2:end);
5     if sigma==0
6         beta = 0;
7     else
8         mu = sqrt(x(1)^2+sigma);
9         if x(1)<=0
10            tmp = x(1) - mu;
11        else
12            tmp = -sigma / (x(1) + mu);
13        end
14        beta = 2*tmp^2/(sigma + tmp^2);
15        x(2:end) = x(2:end)/tmp;
16    end
17    v = [1;x(2:end)];
18 else
19     beta = 0;
20     v = 1;
21 end

```

Wir haben gerade konstruktiv gezeigt, dass sich mit Hilfe der **Householder**-Transformation eine Matrix  $A \in \mathbb{R}^{m \times n}$  in eine Matrix der folgenden Form transformieren können:

$$H_1 A = \begin{pmatrix} * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \\ 0 & * & * & * & * \end{pmatrix}.$$

Gehen wir nun davon aus, dass wir nach einigen Schritten die Ausgangsmatrix  $A$  auf folgende Gestalt gebracht haben:

$$H_2 H_1 A = \begin{pmatrix} * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & \boxplus & * & * \\ 0 & 0 & \boxplus & * & * \\ 0 & 0 & \boxplus & * & * \\ 0 & 0 & \boxplus & * & * \end{pmatrix}.$$

Im nächsten Schritt soll nun die nächste Subspalte unterhalb der Diagonalen eliminiert werden. Es ist also eine Householder-Matrix zu bestimmen, so dass

$$\tilde{H}_3 \begin{pmatrix} \boxplus \\ \boxplus \\ \boxplus \\ \boxplus \end{pmatrix} = \begin{pmatrix} * \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

gilt. Definiert man  $H_3$  als Blockdiagonalmatrix

$$H_3 := \begin{pmatrix} I & 0 \\ 0 & \tilde{H}_3 \end{pmatrix}$$

so erhalten wir

$$H_3 H_2 H_1 A = \begin{pmatrix} * & * & * & * & * \\ 0 & * & * & * & * \\ 0 & 0 & * & * & * \\ 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & * & * \\ 0 & 0 & 0 & * & * \end{pmatrix}.$$

Mit dieser Vorgehensweise erhalten wir folgende Faktorisierung:

$$R = H_{n-1} H_{n-2} \cdots H_1 A \Leftrightarrow A = (H_1 \cdots H_{n-1}) R \rightsquigarrow Q = H_1 \cdots H_{n-1}.$$

Stellen wir nun dar, wie  $A$  überschrieben wird. Es sei

$$\omega^{(j)} = (\underbrace{0, \dots, 0}_{j-1}, 1, \omega_{j+1}^{(j)}, \dots, \omega_m^{(j)})^T$$

der  $j$ -te Householder-Vektor, dann erhält man nach Durchführung der  $QR$ -Zerlegung

$$A = \begin{pmatrix} r_{11} & r_{12} & r_{13} & r_{14} & r_{15} \\ \omega_2^{(1)} & r_{22} & r_{23} & r_{24} & r_{25} \\ \omega_3^{(1)} & \omega_3^{(2)} & r_{33} & r_{34} & r_{35} \\ \omega_4^{(1)} & \omega_4^{(2)} & \omega_4^{(3)} & r_{44} & r_{45} \\ \omega_5^{(1)} & \omega_5^{(2)} & \omega_5^{(3)} & \omega_5^{(4)} & r_{55} \\ \omega_6^{(1)} & \omega_6^{(2)} & \omega_6^{(3)} & \omega_6^{(4)} & \omega_6^{(5)} \end{pmatrix}.$$

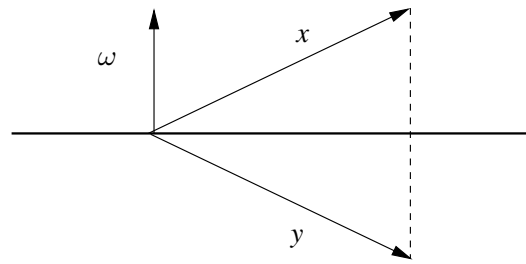
---

### Algorithmus 5.5.2: Berechnung Householder- $QR$ :

Gegeben sei  $A \in \mathbb{R}^{m \times n}$  mit  $m \geq n$ . Der folgende Algorithmus bestimmt die Householder-Matrizen  $H_1, \dots, H_n$ , so dass mit  $Q = H_1 \cdots H_n$  die Matrix  $R = Q^T A$  eine obere Dreiecksmatrix ist. Der obere Dreiecksteil von  $A$  wird mit  $R$  überschrieben und unterhalb der Diagonalen werden die nichttrivialen Komponenten der Householder-Vektoren gespeichert.

```
function:  $A = \text{housematrix}(A)$ 
for  $j = 1 : n$ 
     $[\omega, \beta] = \text{housevector}(A(j : m, j))$ 
     $A(j : m, j : n) = (I_{m-j+1} - \beta \omega \omega^T) A(j : m, j : n)$ 
    if  $j < m$ 
         $A(j+1 : m, j) = \omega(2 : m-j+1)$ 
    end
end
end
```

---

Abb. 5.4: Spiegelung an der zu  $\omega$  orthogonalen Ebene

**Bemerkung 5.5.3** Analog zur **Givens**-Rotation führen wir auch an dieser Stelle kurz die geometrische Anschauung der **Householder**-Spiegelung an. Für einen gegebenen Vektor  $x \in \mathbb{R}^n$  ist der Vektor  $y = Px$  die Spiegelung von  $x$  an der Ebene  $\text{span}\{\omega\}^\perp$ . Hierzu nachfolgende Abbildung:

**Aufgabe 5.5.4** Man zeige, dass die  $QR$ -Zerlegung mittels Givens-Rotation  $\approx 50\%$  mehr Rechenoperationen benötigt als mittels Householder-Transformation.

### MATLAB-Funktion: Householder.m

```

1 function [A,R] = Householder(A)
2 for j = 1:size(A,2)
3     [v,beta(j)] = HouseholderVector(A(j:end,j));
4     A(j:end,j:end) = A(j:end,j:end) - v * (beta(j) * v' * A(j:end,j:end))
5         ;
6     if j < size(A,1)
7         A(j+1:end,j) = v(2:end);
8     end
9 if nargin == 2
10    R = A;
11    A = eye(size(A,1));
12    for j = size(R,2):-1:1
13        v = [1;R(j+1:end,j)];
14        A(j:end,j:end) = A(j:end,j:end) - v*(beta(j)*v'*A(j:end,j:end))
15            ;
16    end
17    R = triu(R);
18 end

```

### MATLAB-Funktion: HouseholderMult.m

```

1 function y = prod_rx(A,x)
2 % extract R from A and multiply with x
3 y = triu(A) * x;
4
5 function y = prod_qx(A,x)
6 y = x;
7 for j = size(A,2)-1:-1:1

```

```

8     v = [1;A(j+1:end,j)];
9     beta = 2/(v'*v);
10    y(j:end) = y(j:end) - v*(beta*v'*y(j:end));
11 end
12
13 function y = prod_qtx(A,x)
14 y = x;
15 for j = 1:size(A,2)-1
16     v = [1;A(j+1:end,j)];
17     beta = 2/(v'*v);
18     y(j:end) = y(j:end) - v*(beta*v'*y(j:end));
19 end

```

### MATLAB-Beispiel:

Mit der Funktion `Householder` wird zum einen die Matrix  $A$  kompakt überschrieben, d.h. der Rückgabewert benötigt nur den Speicherplatz der ursprünglichen Matrix; zum anderen kann man sich auch  $Q$  und  $R$  mit  $A = QR$  ausgeben lassen. Die Berechnung von  $Rx$ ,  $Qx$  und  $Q^T x$ , wenn die kompakte Form vorliegt, ist in den Routinen `prod_qx`, etc. realisiert.

```

>> A = rand(3);
>> x = rand(3,1);
>> [Q,R] = Householder(A);
>> A-Q*R
ans =
    1.0e-015 *
    -0.1110         0    -0.1110
    -0.0035    -0.1110         0
    -0.1110         0     0.2220
>> C = Householder(A);
>> Q'*x - prod_qtx(C,x)
ans =
    1.0e-015 *
    0.2220
    0.0555
    0.5551

```

## 5.6 Die Singulärwertzerlegung

Die Singulärwertzerlegung (engl. Singular Value Decomposition, SVD) ist eine weitere Möglichkeit, ein lineares Ausgleichsproblem zu lösen. Sie hat aber auch zahlreiche andere Anwendungen z.B. in der Statistik (u.a. bei der Hauptkomponentenanalyse, Hauptachsentransformation).

**Satz 5.6.1** Zu jedem  $A \in \mathbb{R}^{m \times n}$  existieren orthogonale Matrizen  $U \in \mathbb{R}^{m \times m}$ ,  $V \in \mathbb{R}^{n \times n}$  und eine Diagonalmatrix

$$\Sigma := \text{diag}(\sigma_1, \dots, \sigma_p) \in \mathbb{R}^{m \times n}, \quad p = \min\{m, n\}$$

mit  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p \geq 0$ , (den **Singulärwerten**), so dass

$$A = U \Sigma V^T. \quad (5.10)$$

**Beweis:** Es genügt zu zeigen, dass

$$U^T A V = \begin{pmatrix} \sigma & 0 \\ 0 & B \end{pmatrix} \quad \text{mit } \sigma \geq 0, \quad B \in \mathbb{R}^{(m-1) \times (n-1)},$$

der Rest folgt dann induktiv. Sei  $\sigma = \|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2$ . Da  $B_1(0) = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$  kompakt ist, existiert ein  $v \in B_1(0)$  mit

$$\|Av\|_2 = \|A\|_2 \|v\|_2 = \sigma.$$

Also erfüllt  $u := \frac{1}{\sigma} Av$  die Gleichungen  $Av = \sigma u$ ,  $\|u\|_2 = \frac{1}{\sigma} \|Av\|_2 = 1$ . Ergänze nun  $u$  und  $v$  zu Orthogonalbasen

$$\{u = U^1, U^2, \dots, U^m\}, \{v = V^1, \dots, V^n\}$$

des  $\mathbb{R}^m$  bzw.  $\mathbb{R}^n$ . Daraus folgt, dass  $U = (U^1, \dots, U^m)$ ,  $V = (V^1, \dots, V^n)$  orthogonal sind und

$$\begin{aligned} A_1 &:= U^T A V = (U^T A V^1, \dots, U^T A V^n) \\ &= (\sigma e^1, U^T A V^2, \dots, U^T A V^n) = \begin{pmatrix} \sigma & w^T \\ 0 & B \end{pmatrix}, w \in \mathbb{R}^{n-1}. \end{aligned}$$

Nun gilt

$$\left\| A_1 \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_2^2 = \left\| \begin{pmatrix} \sigma^2 + w^T w \\ Bw \end{pmatrix} \right\|_2^2 \geq (\sigma^2 + \|w\|_2^2)^2 = \left\| \begin{pmatrix} \sigma \\ w \end{pmatrix} \right\|_2^2,$$

also  $\|A_1\|_2^2 \geq \sigma^2 + \|w\|_2^2$ .

Daraus folgt  $\sigma^2 = \|A\|_2^2 = \|U^T A V\|_2^2 = \|A_1\|_2^2 \geq \sigma^2 + \|w\|_2^2 \geq \sigma^2 > 0$ , was nur gelten kann, wenn  $w = 0$ .  $\square$

**Bemerkung 5.6.2** (a) Mit  $A = U \Sigma V^T$  folgt  $A^T = V \Sigma^T U^T$ , also

$$A^T A = V \Sigma^T U^T U \Sigma V^T = V \Sigma^2 V^T$$

mit  $\Sigma^2 = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ , also der Diagonalisierung von  $A^T A$ .

(b) Der SVD-Algorithmus ist etwas kompliziert. Er besteht

- aus einer Bi-Diagonalisierung von  $A$  z.B. mit Givens und
- einem symmetrischen QR für  $A^T A$ .

Details findet man z.B. in [Golub/Loan, 427-433].

Die Lösung des linearen Ausgleichsproblems mittels der SVD erfolgt nun über die sogenannte Pseudo-Inverse.

**Definition 5.6.3** Angenommen  $A \in \mathbb{R}^{m \times n}$  habe Rang  $r$  und die SVD  $A = U \Sigma V^T$ , dann heißt die Matrix

$$A^+ = V \Sigma^+ U^T \tag{5.11}$$

mit  $\Sigma^+ := \text{diag}(\frac{1}{\sigma_1}, \dots, \frac{1}{\sigma_r}, 0, \dots, 0)$  **Pseudo-Inverse** (oder **Moore-Penrose-Inverse**).

**Lemma 5.6.4** Es gilt  $A^+ = (A^T A)^{-1} A^T$ .

*Beweis.* Nach Bemerkung 5.6.2 (a) gilt

$$(A^T A)^{-1} = (V \Sigma^2 V^T)^{-1} = V \Sigma^{-2} V^T = V (\Sigma^T \Sigma)^{-1} V^T,$$

also

$$(A^T A)^{-1} A^T = V (\Sigma^T \Sigma)^{-1} V^T V \Sigma U^T = V (\Sigma^T \Sigma)^{-1} U^T$$

sowie  $\Sigma^T \Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_r^2)$  und damit  $(\Sigma^T \Sigma)^{-1} \Sigma = \Sigma^+$ .  $\square$

**Lemma 5.6.5** Die Pseudo-Inverse  $A^+$  von  $A$  besitzt folgende Eigenschaften

- (a)  $(A^+A)^T = A^+A$ ,  $(AA^+)^T = AA^+$ ,
- (b)  $A^+AA^+ = A^+$ ,
- (c)  $AA^+A = A$ ,
- (d) Falls  $A$  vollen Rang hat, gilt  $A^+A = Id$ .

Beweis. Übung. □

**Bemerkung 5.6.6** Sowohl die Eigenschaft in Lemma 5.6.4 als auch diejenigen in Lemma 5.6.5 (gemeinsam) können als äquivalente Definition von  $A^+$  verwendet werden.

**Satz 5.6.7** Sei  $A \in \mathbb{R}^{m \times n}$ ,  $m \geq n$ ,  $b \in \mathbb{R}^m$ , dann ist  $x = A^+b$  diejenige Lösung von

$$\|Ax - b\|_2 \longrightarrow \text{Min!}$$

mit der **kleinsten** euklidischen Norm.

Beweis. Für  $x = A^+b$  gilt nach Lemma 5.6.4

$$x = A^+b = (A^T A)^{-1} A^T b,$$

also  $A^T Ax = A^T b$ , also erfüllt  $x$  die Normalengleichungen und löst damit das lineare Ausgleichsproblem. Seien nun  $x_1, x_2 \in \mathbb{R}^n$  zwei Lösungen, dann gilt  $A^T A(x_1 - x_2) = 0$ , also

$$0 = (x_1 - x_2)^T A^T A(x_1 - x_2) = \|A(x_1 - x_2)\|_2^2,$$

d.h.  $A(x_1 - x_2)$  bzw.  $x_1 - x_2 \in \text{Ker}(A)$ . Damit hat also jede Lösung  $\tilde{x}$  die Form

$$\tilde{x} = x + y \text{ mit } x = A^+b, y \in \text{Ker}(A).$$

Sei nun  $y \in \text{Ker}(A)$ , dann gilt wegen Lemma 5.6.5 (b)

$$y^T A^+b = y^T A^+AA^+b = ((A^+A)^T y)^T A^+b = (A^+Ay)^T A^+b = 0,$$

da  $y \in \text{Ker}(A)$  und wegen Lemma 5.6.5 (a). Also gilt

$$x = A^+b \perp \text{Ker}(A).$$

Damit folgt nun für jede Lösung  $\tilde{x} = x + y$

$$\|\tilde{x}\|_2^2 = \|x + y\|_2^2 = \|x\|_2^2 + 2(x, y)_2 + \|y\|_2^2 = \|x\|_2^2 + \|y\|_2^2,$$

da  $(x, y)_2 = x^T y = 0$  und der obige Ausdruck wird minimal genau dann, wenn  $y = 0$ . □

**Bemerkung 5.6.8** Nun sei  $p := \text{Rang}(A) \leq \min\{m, n\}$ . Dann liefert die QR-Zerlegung nach  $p$  Schritten

$$QA = \left[ \begin{array}{c|c} R & S \\ \hline 0 & 0 \end{array} \right] \quad (5.12)$$

mit  $R \in \mathbb{R}^{p \times p}$  und  $S \in \mathbb{R}^{p \times (n-p)}$ . Damit kann man also die QR-Zerlegung auch zur numerischen Rangbestimmung verwenden.



**Lemma 5.6.9** (a)  $x \in \mathbb{R}^n$  ist genau dann Lösung des linearen Ausgleichsproblems, falls  $x = (x_1, x_2)^T$  mit  $x_1 \in \mathbb{R}^p$ ,  $x_2 \in \mathbb{R}^{n-p}$  und

$$x_1 = R^{-1}b_1 - R^{-1}Sx_2,$$

wobei  $(b_1, b_2)^T = Qb$ ,  $b_1 \in \mathbb{R}^p$ ,  $b_2 \in \mathbb{R}^{m-p}$ . Hier wird nichts über den Rang vorausgesetzt.

(b) Sei  $p < n$  (der sogenannte **Rang-defekte Fall**),

$$V := R^{-1}S \in \mathbb{R}^{p \times (n-p)}, \quad u := R^{-1}b_1 \in \mathbb{R}^p.$$

Dann ist die Lösung  $x$  des linearen Ausgleichsproblems mit der kleinsten Norm  $\|x\|_2$  gegeben durch  $x = (x_1, x_2) \in \mathbb{R}^p \times \mathbb{R}^{n-p}$  mit

$$(Id + V^T V)x_2 = V^T u, \quad x_1 = u - Vx_2.$$

*Beweis.* (a) Es gilt

$$\|b - Ax\|_2^2 = \|Qb - QAx\|_2^2 = \|Rx_1 + Sx_2 - b_1\|_2^2 + \|b_2\|_2^2,$$

also die Bedingung  $Rx_1 = b_1 - Sx_2$ .

(b) Nach (a) gilt  $x_1 = R^{-1}(b_1 - Sx_2) = u - Vx_2$  und damit

$$\begin{aligned} \|x\|_2^2 &= \|x_1\|_2^2 + \|x_2\|_2^2 = \|u - Vx_2\|_2^2 + \|x_2\|_2^2 \\ &= \|u\|_2^2 - 2\langle u, Vx_2 \rangle + \langle Vx_2, Vx_2 \rangle + \|x_2\|_2^2. \end{aligned}$$

Weiter gilt  $-2\langle u, Vx_2 \rangle + \langle Vx_2, Vx_2 \rangle = \langle x_2, V^T Vx_2 - 2V^T u \rangle$ , also

$$\begin{aligned} \|x\|_2^2 &= \|u\|_2^2 + \langle x_2, V^T Vx_2 - 2V^T u \rangle + \|x_2\|_2^2 \\ &= \|u\|_2^2 + \langle x_2, (Id + V^T V)x_2 - 2V^T u \rangle =: \varphi(x_2). \end{aligned}$$

Letztere Funktion gilt es zu minimieren. Es gilt

$$\begin{aligned} \varphi'(x_2) &= -2V^T u + 2(Id + V^T V)x_2 \\ \varphi''(x_2) &= 2(Id + V^T V) > 0, \end{aligned}$$

da diese Matrix s.p.d. ist. Also ist  $\varphi(x_2)$  genau dann minimal, wenn

$$(Id + V^T V)x_2 = V^T u,$$

also die Behauptung. □

Damit kann man die Pseudo-Inverse auch ohne SVD direkt aus  $QR$  bestimmen.

---

**Algorithmus 5.6.1: Pseudo-Inverse über QR**

Sei  $A \in \mathbb{R}^{m \times n}$ ,  $b \in \mathbb{R}^m$ . Bestimme  $x = A^+b$  gemäß:

- 1.) Bestimme die QR-Zerlegung von  $A$  gemäß (5.12) und berechne  $p, R$  und  $S$  sowie  $Q$ .
- 2.) Berechne  $V \in \mathbb{R}^{p \times (n-p)}$  aus  $RV = S$ , löse also  $n - p$  gestaffelte Gleichungssysteme der Dimension  $p$ .
- 3.) Bestimme die Cholesky-Zerlegung  $Id + V^TV = LL^T$ .
- 4.) Setze  $Qb =: (b_1, b_2)^T$  mit  $b_1 \in \mathbb{R}^p$ ,  $b_2 \in \mathbb{R}^{m-p}$ .
- 5.) Bestimme  $u$  aus dem gestaffelten System  $Ru = b_1$ .
- 6.) Berechne  $x_2 \in \mathbb{R}^{n-p}$  aus zwei gestaffelten Systemen  $LL^T x_2 = V^T u$ .
- 7.) Setze  $x_1 = u - Vx_2$ .

Ausgabe:  $x = (x_1, x_2)^T = A^+b$ .

---

# A LANDAU-SYMBOLLE

Landau-Symbole beschreiben das asymptotische Verhalten von Funktionen. Genauer vergleichen sie das gesuchte asymptotische Verhalten einer Funktion  $f$  mit dem bekannten asymptotischen Verhalten einer Referenzfunktion  $g$ .

## A.1 Definition

Ist  $x_0 \in \mathbb{R}$  oder  $x_0 = \infty$  und sind  $f$  und  $g$  in einer Umgebung von  $x_0$  definierte reellwertige Funktionen, wobei  $g(x) > 0$  vorausgesetzt wird, dann haben die Symbole  $\mathcal{O}(\cdot)$  und  $o(\cdot)$  die folgenden Bedeutungen:

**Definition A.1.1** (a)  $f(x) = \mathcal{O}(g(x))$  für  $x \rightarrow x_0$  bedeutet, dass  $\frac{f(x)}{g(x)}$  für  $x \rightarrow x_0$  beschränkt ist.

(b)  $f(x) = o(g(x))$  für  $x \rightarrow x_0$  bedeutet, dass  $\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = 0$ .

Wenn  $x_0$  aus dem Zusammenhang bekannt ist, schreibt man häufig nur  $f(x) = \mathcal{O}(g(x))$  bzw.  $f(x) = o(g(x))$ . Die gleichen Symbole werden sinngemäß auch für das Wachstum von Folgen für  $n \rightarrow \infty$  verwendet.

## A.2 Beispiele

**Beispiel A.2.1 (Landau-Symbole)** 1.  $f(x) = \mathcal{O}(1)$  bedeutet, dass  $f$  beschränkt ist, z.B.  $\sin(x) = \mathcal{O}(1)$ .

2.  $f(x) = \mathcal{O}(x)$  bedeutet, dass  $f$  für  $x \rightarrow \infty$  höchstens linear wächst. Beispiel:  $\sqrt{1+x^2} = \mathcal{O}(x)$ ,  $\log(x) = \mathcal{O}(x)$ .

3.  $f(x) = o(x)$  bedeutet, dass  $f$  für  $x \rightarrow \infty$  langsamer als jede lineare Funktion wächst. Beispiele:  $\sqrt{x} = o(x)$ ,  $\log(x) = o(x)$ .

4.  $\sum_{k=1}^n k = \frac{1}{2}n(n+1) = \mathcal{O}(n^2)$ .

In der Numerik werden Landau-Symbole häufig verwendet, um den Aufwand eines Algorithmus abzuschätzen.

**Beispiel A.2.2** Die Berechnung eines Skalarprodukts zweier Vektoren  $x, y \in \mathbb{R}^n$  erfordert  $n$  Produkte und  $n - 1$  Additionen:

$$x^T y = \sum_{i=1}^n x_i y_i.$$

Zählt man jede Addition und jedes Produkt als gleich aufwändige Gleitpunktoperationen, ergeben sich  $2n - 1$  Operationen. Diese genaue Zahl ist aber weniger interessant als die Tatsache, dass der Gesamtaufwand linear mit der Dimension steigt: verdoppelt man die Länge der Vektoren, verdoppelt sich (ungefähr!) der Aufwand. Man sagt: das Skalarprodukt besitzt die Komplexität  $\mathcal{O}(n)$ .

**Beispiel A.2.3** Bei der Multiplikation zweier Matrizen  $A, B \in \mathbb{R}^{n \times n}$  sind  $n^2$  Elementen des Produkts zu bestimmen. Jedes Element wird mit Hilfe eines Skalarprodukts berechnet. Insgesamt benötigt man also  $n^2(2n - 1) = \mathcal{O}(n^3)$  Gleitpunktoperationen.

# B NORMEN

Normen erfüllen den gleichen Zweck auf Vektorräumen, den Beträge auf der reellen Achse erfüllen. Genauer gesagt,  $\mathbb{R}^n$  mit einer Norm auf  $\mathbb{R}^n$  liefert einen metrischen Raum. Daraus ergeben sich bekannte Begriffe wie Umgebung, offene Menge, Konvergenz und Stetigkeit für Vektoren und vektorwertige Funktionen.

Wir beginnen mit der Vereinbarung einer Notation.

**Notation B.0.1** Für  $x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$  definieren wir  $|x| \in \mathbb{R}^n$  durch

$$|x| = (|x_1|, \dots, |x_n|)^T.$$

Für eine Matrix  $A \in \mathbb{R}^{m \times n}$  definieren wir die Matrix  $|A| \in \mathbb{R}^{m \times n}$  durch

$$|A|_{ij} = |a_{ij}|, \quad 1 \leq i \leq m, 1 \leq j \leq n.$$

**Definition B.0.2 (Vektornorm)** Eine Vektornorm auf  $\mathbb{R}^n$  ist eine Funktion  $\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}$  mit den Eigenschaften

$$\begin{aligned} \|x\| &\geq 0 & x &\in \mathbb{R}^n, & \|x\| = 0 &\Leftrightarrow x = 0, \\ \|x + y\| &\leq \|x\| + \|y\| & x, y &\in \mathbb{R}^n, \\ \|\alpha x\| &= |\alpha| \|x\| & \alpha &\in \mathbb{R}, x \in \mathbb{R}^n. \end{aligned} \tag{B.1}$$

Eine nützliche Klasse von Vektornormen sind die **p-Normen** ( $p \in \mathbb{N}$ ), definiert durch

$$\|x\|_p = \left( \sum_{k=0}^n |x_k|^p \right)^{\frac{1}{p}}. \tag{B.2}$$

Von diesen sind besonders die 1, 2 und  $\infty$  interessant:

$$\begin{aligned} \|x\|_1 &= |x_1| + \dots + |x_n|, \\ \|x\|_2 &= (|x_1|^2 + \dots + |x_n|^2)^{\frac{1}{2}} = (x^T x)^{\frac{1}{2}}, \\ \|x\|_\infty &= \max_{1 \leq j \leq n} |x_j|. \end{aligned} \tag{B.3}$$

Ein **Einheitsvektor** bzgl. der Norm  $\|\cdot\|$  ist ein Vektor  $x$  der Länge 1, d.h. mit  $\|x\| = 1$ .

## B.1 Vektornorm-Eigenschaften

Ein klassisches Ergebnis bzgl.  $p$ -Normen ist die **Hölder-Ungleichung**

$$|x^T y| \leq \|x\|_p \|y\|_q \quad \text{mit } \frac{1}{p} + \frac{1}{q} = 1. \tag{B.4}$$

Spezialfall der Hölder-Ungleichung ist die **Cauchy-Schwarz-Ungleichung** (CSU)

$$|x^T y| \leq \|x\|_2 \|y\|_2. \tag{B.5}$$

Alle Normen auf  $\mathbb{R}^n$  sind äquivalent, d.h. es seien  $\|\cdot\|_\alpha$  und  $\|\cdot\|_\beta$  Normen auf  $\mathbb{R}^n$ , dann existieren positive Konstanten  $c_1, c_2$ , so dass

$$c_1 \|x\|_\alpha \leq \|x\|_\beta \leq c_2 \|x\|_\alpha \quad (x \in \mathbb{R}^n). \tag{B.6}$$

Zum Beispiel sei  $x \in \mathbb{R}^n$ , dann gilt

$$\begin{aligned} \|x\|_2 &\leq \|x\|_1 \leq \sqrt{n} \|x\|_2, \\ \|x\|_\infty &\leq \|x\|_2 \leq \sqrt{n} \|x\|_\infty, \\ \|x\|_\infty &\leq \|x\|_1 \leq n \|x\|_\infty. \end{aligned} \quad (\text{B.7})$$

**Aufgabe B.1.1** Man beweise die Aussagen (B.4)–(B.7).

## B.2 Matrixnormen

Da  $\mathbb{R}^{m \times n}$  isomorph (d.h. es existiert eine bijektive Abb. mit  $\varphi(\lambda A + \mu B) = \lambda \varphi(A) + \mu \varphi(B)$ ) zu  $\mathbb{R}^{m \cdot n}$  ist, sollte die Definition einer Matrixnorm äquivalent sein zur Definition einer Vektornorm.

**Definition B.2.1 (Matrixnorm)** Es sei  $\|\cdot\| : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ . Dann ist  $\|\cdot\|$  eine Matrixnorm, wenn folgende Eigenschaften gelten:

$$\begin{aligned} \|A\| &\geq 0 & A &\in \mathbb{R}^{m \times n}, & \|A\| = 0 &\Leftrightarrow A = 0 \\ \|A + B\| &\leq \|A\| + \|B\| & A, B &\in \mathbb{R}^{m \times n} \\ \|\alpha A\| &= |\alpha| \|A\| & \alpha &\in \mathbb{R}, A \in \mathbb{R}^{m \times n}. \end{aligned} \quad (\text{B.8})$$

**Bemerkung B.2.2** Die mit am häufigsten verwendeten Normen in der Numerischen Linearen Algebra sind die **Frobenius-Norm**

$$\|A\|_F := \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2} \quad (\text{B.9})$$

und die **p-Normen**

$$\|A\|_p := \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}. \quad (\text{B.10})$$

Man beachte folgende äquivalente Definition

$$\|A\|_p = \sup_{x \neq 0} \left\| A \frac{x}{\|x\|} \right\|_p = \sup_{\|x\|_p=1} \|Ax\|_p. \quad (\text{B.11})$$

**Bemerkung B.2.3 (Operatornorm)** Etwas allgemeiner als die p-Matrixnorm ist die sogenannte **Operatornorm**. Ist  $\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}$  eine Vektornorm, so ist für  $A \in \mathbb{R}^{m \times n}$  die Operatornorm definiert als

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

**Bemerkung B.2.4 (Submultiplikativität)** Die Frobenius-Norm und die p-Normen erfüllen zusätzlich auch noch die Eigenschaft der **Submultiplikativität**, d.h.

$$\|A \cdot B\| \leq \|A\| \|B\|. \quad (\text{B.12})$$

**Bemerkung B.2.5** Es sei  $y$  ein Vektor mit  $\|y\|_p = 1$ , dann gilt

$$\begin{aligned} \|A \cdot B\|_p &= \max_{x \neq 0} \frac{\|ABx\|_p}{\|x\|_p} = \max_{x \neq 0} \frac{\|ABx\|_p \|Bx\|_p}{\|Bx\|_p \|x\|_p} \\ &\leq \max_{x \neq 0} \frac{\|ABx\|_p}{\|Bx\|_p} \cdot \max_{x \neq 0} \frac{\|Bx\|_p}{\|x\|_p} \\ &= \max_{y \neq 0} \frac{\|Ay\|_p}{\|y\|_p} \cdot \max_{x \neq 0} \frac{\|Bx\|_p}{\|x\|_p} = \|A\|_p \|B\|_p. \end{aligned}$$

**Bemerkungen B.2.6** i) Nicht alle Matrix-Normen erfüllen diese Eigenschaft, z.B. gilt für

$$\|A\|_A := \max |a_{ij}| \text{ und } A = B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \text{ die Ungleichung}$$

$$\|A \cdot B\|_A = 2 > \|A\|_A \|B\|_A = 1.$$

ii) Für die  $p$ -Normen haben wir die wichtige Eigenschaft, dass für alle  $A \in \mathbb{R}^{m \times n}$  und  $x \in \mathbb{R}^n$

$$\|Ax\|_p \leq \|A\|_p \|x\|_p. \quad (\text{B.13})$$

Die Frobenius und  $p$ -Normen (speziell  $p = 1, 2, \infty$ ) erfüllen gewisse Ungleichungen, welche in der Analysis von Matrixberechnungen verwendet werden. Es sei  $A \in \mathbb{R}^{m \times n}$ , dann gilt

$$\|A\|_2 \leq \|A\|_F \leq \sqrt{n} \|A\|_2, \quad (\text{B.14})$$

$$\max_{i,j} |a_{ij}| \leq \|A\|_2 \leq \sqrt{mn} \max_{i,j} |a_{ij}|, \quad (\text{B.15})$$

$$\frac{1}{\sqrt{n}} \|A\|_\infty \leq \|A\|_2 \leq \sqrt{m} \|A\|_\infty, \quad (\text{B.16})$$

$$\frac{1}{\sqrt{m}} \|A\|_1 \leq \|A\|_2 \leq \sqrt{n} \|A\|_1. \quad (\text{B.17})$$

Des Weiteren gilt für  $A \in \mathbb{R}^{m \times n}$

$$\begin{aligned} \|A\|_\infty &= \max_{\|x\|_\infty=1} \|Ax\|_\infty = \max_{\|x\|_\infty=1} \left\{ \max_{1 \leq j \leq m} \left| \sum_{k=1}^n a_{jk} x_k \right| \right\} \\ &= \max_{1 \leq j \leq m} \left\{ \max_{\|x\|_\infty=1} \left| \sum_{k=1}^n a_{jk} x_k \right| \right\} = \max_{1 \leq j \leq m} \sum_{k=1}^n |a_{jk}|. \end{aligned}$$

Aufgrund dieser Gleichung bezeichnet man  $\|\cdot\|_\infty$  auch als **Zeilensummennorm**. Diese ist mittels der letzten Gleichung auch leicht zu bestimmen, d.h.

$$\|A\|_\infty = \max_{1 \leq j \leq m} \sum_{k=1}^n |a_{jk}|. \quad (\text{B.18})$$

Analog gilt für die 1-Norm:

$$\begin{aligned} \|A\|_1 &= \max_{\|x\|_1=1} \|Ax\|_1 = \max_{\|x\|_1=1} \sum_{j=1}^m \left| \sum_{k=1}^n a_{jk} x_k \right| = \max_{\|x\|_1=1} \sum_{j=1}^m \sum_{k=1}^n |a_{jk}| |x_k| \text{sign}(a_{jk} x_k) \\ &= \max_{\|x\|_1=1} \sum_{k=1}^n |x_k| \sum_{j=1}^m |a_{jk}| \text{sign}(a_{jk} x_k) = \max_{\|x\|_1=1} \sum_{k=1}^n |x_k| \sum_{j=1}^m |a_{jk}| = \max_{1 \leq k \leq n} \sum_{j=1}^m |a_{jk}|. \end{aligned}$$

Also gilt:

$$\|A\|_1 = \max_{1 \leq k \leq n} \sum_{j=1}^m |a_{jk}|. \quad (\text{B.19})$$

Diese Norm bezeichnet man auch als **Spaltensummennorm**.

**Bemerkung B.2.7** Einfache Eselbrücke:  $\boxed{1}$ -Spaltensummen,  $\boxed{\infty}$ -Zeilensummen.

Auch für die 2-Norm (die sogenannte **Euklidische Norm**) lässt sich eine äquivalente Formulierung finden.

**Satz B.2.8** Es sei  $A \in \mathbb{R}^{m \times n}$ . Dann existiert ein Vektor  $z \in \mathbb{R}^n$  mit  $\|z\|_2 = 1$ , so dass

$$A^T A z = \mu^2 z, \quad \text{wobei } \mu = \|A\|_2.$$

**Bemerkung B.2.9** Der Satz impliziert, dass  $\|A\|_2^2$  eine Nullstelle des Polynoms  $p(z) = \det(A^T A - \lambda I)$  ist. Genauer betrachtet, ist die 2-Norm von  $A$  die Wurzel des größten Eigenwerts von  $A^T A$ .

*Beweis.* Es sei  $z \in \mathbb{R}^n$  mit  $\|z\|_2 = 1$  und  $\|Az\|_2 = \|A\|_2$ . Da  $z$  die Funktion

$$g(x) = \frac{1}{2} \frac{\|Ax\|_2^2}{\|x\|_2^2} = \frac{1}{2} \frac{x^T A^T A x}{x^T x} \quad (\text{B.20})$$

maximiert, folgt daraus, dass er  $\nabla g(z) = 0$  erfüllt, wobei  $\nabla g$  der Gradient von  $g$  ist  $\left(\nabla := \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \dots, \frac{\partial}{\partial x_n}\right)^T\right)$ . Die partiellen Ableitungen von  $g$  lauten für  $i = 1, \dots, n$

$$\frac{\partial g(z)}{\partial x_i} = \left[ x^T x \sum_{j=1}^n (A^T A)_{ij} x_j - (x^T A^T A x) x_i \right] / (x^T x)^2. \quad (\text{B.21})$$

In Vektornotation bedeutet dies für  $z$ , da  $\nabla g(z) = 0$  und  $\|z\|_2 = 1$ , dass  $A^T A z = (z^T A^T A z) z$ . Setzt man nun  $\mu = \|A\|_2$  so ergibt sich daraus die Behauptung.  $\square$

**Lemma B.2.10** Es sei  $A \in \mathbb{R}^{m \times n}$ . Dann gilt

$$\|A\|_2 \leq \sqrt{\|A\|_1 \|A\|_\infty}. \quad (\text{B.22})$$

*Beweis.* Es sei  $z \in \mathbb{R}^n$  mit  $A^T A z = \mu^2 z$  und  $\mu = \|A\|_2$ , d.h. es sei  $z \neq 0$  ein Vektor, für den das Maximum von  $\frac{\|Ax\|_2}{\|x\|_2}$  angenommen wird. Dann gilt

$$\mu^2 \|z\|_1 = \|A^T A z\|_1 \leq \|A^T\|_1 \|Az\|_1 \leq \|A\|_\infty \|A\|_1 \|z\|_1. \quad (\text{B.23})$$

Kondensation von  $\|z\|_1$  und Wurzelziehen liefert das gewünschte Ergebnis.  $\square$

**Definition B.2.11 (verträgliche Vektornorm)** Eine Matrixnorm  $\|A\|$  heißt **kompatibel** oder **verträglich** mit der Vektornorm  $\|x\|$ , falls folgende Ungleichung erfüllt ist

$$\|Ax\| \leq \|A\| \|x\|, \quad x \in \mathbb{R}^n, A \in \mathbb{R}^{m \times n}. \quad (\text{B.24})$$

Kombinationen von verträglichen Normen sind etwa

$\|A\|_G, \|A\|_\infty$  sind verträglich mit  $\|x\|_\infty$ ;

$\|A\|_G, \|A\|_1$  sind verträglich mit  $\|x\|_1$ ;

$\|A\|_G, \|A\|_F, \|A\|_2$  sind verträglich mit  $\|x\|_2$ , wobei

$$\|A\|_G := n \max_{i,j} |a_{ij}| \quad (A \in \mathbb{R}^{m \times n}).$$

**Aufgabe B.2.12** Man verifiziere selbständig an einigen Beispielen die Verträglichkeit von o.g. Normenpaaren.

Abschließend erhalten wir am Ende des Kapitels.

**Satz B.2.13** Die Matrix- $p$ -Normen sind unter allen mit der Vektornorm  $\|x\|_p$  verträglichen Matrixnormen die kleinsten.



# C EINFÜHRUNG IN MATLAB

## C.1 Grundlegende MATLAB-Befehle

Ruft man das Programm MATLAB mit dem Befehl `matlab` auf, so erscheinen auf dem Monitor einige Fenster. Auf den Linux-Rechnern des KIZ müssen Sie vorher die notwendigen Pfade ergänzen. Geben Sie dazu in einem Konsole-Fenster den Befehl `option matlab` ein.

Von diesen ist das Befehl-Fenster der primäre Ort um Befehle einzugeben und Fehler oder Ergebnisse abzulesen! Das Prompt-Zeichen `>>` ist im Befehl-Fenster dargestellt und dort findet man üblicherweise einen blinkenden Cursor. Der blinkende Cursor und der MATLAB-Prompt zeigen einem, dass MATLAB eine Eingabe erwartet.

### C.1.1 Einfache mathematische Operationen

Genauso wie mit einem simplen Taschenrechner kann man auch mit MATLAB einfache mathematische Operationen ausführen, z.B. ergibt die Eingabe

```
>> 3 + 4
```

die Ausgabe

```
ans =  
7
```

Man beachte, dass MATLAB im Allgemeinen keine Zwischenräume benötigt, um Befehle eindeutig zu verstehen. Alternativ zu dem obigen Beispiel können in MATLAB auch Variablen verwendet werden.

```
>> a = 3  
a =  
3  
>> b = 4  
b =  
4  
>> c = a + b  
c =  
7
```

MATLAB besitzt folgende einfache arithmetische Operationen

Operation	Symbol	Beispiel
Addition, $a + b$	+	$5 + 3$
Subtraktion, $a - b$	-	$23 - 12$
Multiplikation, $a \cdot b$	*	$13.3 * 63.13$
Division, $a \div b$	/ or \	$17/4 = 4 \backslash 17$
Potenz, $a^b$	^	$3^4$

Die Reihenfolge, in der eine Folge von Operationen abgearbeitet wird, lässt sich wie folgt beschreiben. Ausdrücke werden von links nach rechts ausgeführt, wobei die Potenzierung die höchste Priorität besitzt gefolgt von Punktoperation, sprich Multiplikation und Division. Die geringste Priorität haben Addition und Subtraktion. Mit Hilfe von Klammern kann diese Vorgehensweise geändert werden, wobei innere Klammern vor äußeren Klammern berechnet werden.



## C.1.2 Variablen



Wie in anderen Programmiersprachen hat auch MATLAB Regeln für Variablennamen. Eine Variable repräsentiert ein Datenelement, dessen Wert während der Programmausführung – gegebenenfalls mehrfach – geändert werden kann. Variablen werden anhand ihrer „Namen“ identifiziert. Namen bestehen aus ein bis neunzehn Buchstaben, Ziffern oder Unterstrichen, wobei das erste Zeichen ein Buchstabe sein muss. Man beachte, dass MATLAB Groß- und Kleinschreibung unterscheidet. (Windows ist im Gegensatz zu Linux nicht so restriktiv, da Sie jedoch Programme austauschen wollen, sollten Windows-Benutzer besondere Aufmerksamkeit walten lassen.)

Einer Variablen ist Speicherplatz zugeordnet. Wenn man eine Variable verwendet, dann meint man damit entweder den zugeordneten Speicherplatz oder den Wert, der dort augenblicklich abgespeichert ist. Einen Überblick über alle Variablen erhält man mit dem Befehl `who` oder `whos`, wobei letzterer die Angabe des benutzten Speicherplatzes beinhaltet.

Zusätzlich zu selbstdefinierten Variablen gibt es in MATLAB verschiedene spezielle Variablen. Diese lauten

spezielle Variablen	Wert
<code>ans</code>	standard Variablenname benutzt für Ergebnisse
<code>pi</code>	3.1415...
<code>eps</code>	Maschinengenauigkeit
<code>flops</code>	Zähler für die Anzahl der Fließkommaoperationen
<code>inf</code>	steht für Unendlich (eng. infinity). z.B. 1/0
<code>NaN</code>	eng. Not a Number, z.B. 0/0
<code>i (und) j</code>	$i = j = \sqrt{-1}$

In MATLAB kann der Speicherplatz, der durch Variablen belegt ist, durch den Befehl `clear` wieder freigegeben werden, z.B.

```
>> clear a b c
```

## C.1.3 Kommentare und Punktion

Der Text, der nach einem Prozentzeichen `%` folgt, wird in MATLAB als Kommentar verstanden

```
>> dummy = 4 % Wert von dummy
dummy =
    4
```

Mehrere Befehle können in eine Zeile geschrieben werden, wenn sie durch Kommata oder Semikola getrennt werden. Kommata veranlassen MATLAB, die Ergebnisse anzuzeigen. Bei einem Semikolon wird die Ausgabe unterdrückt. Durch eine Sequenz von drei Punkten kann man einen Befehl in der folgenden Zeile fortsetzen.

## C.1.4 Spezielle Funktionen

Eine unvollständige Liste von Funktionen, die MATLAB bereitstellt, ist im folgenden dargestellt. Die meisten Funktionen sind so definiert, wie man sie üblicherweise benutzt.

```
>> y = cos(pi)
y =
   -1
```



MATLAB bezieht sich im Zusammenhang mit Winkelfunktionen auf das Bogenmaß.

Funktion	Bedeutung
abs(x)	Absolutbetrag
cos(x)	Kosinus
exp(x)	Exponentialfunktion: $e^x$
fix(x)	rundet auf die nächste, vom Betrag her kleinere ganze Zahl
floor(x)	rundet auf die nächste, kleinere ganze Zahl
gcd(x,y)	größter gemeinsamer Teiler von x und y
lcm(x,y)	kleinstes gemeinsames Vielfaches von x und y
log(x)	natürlicher Logarithmus
rem(x,y)	Modulo (eng. remainder of division), z.B. rem(5,2)=1
sign(x)	Signum Funktion, z.B. sign(2.3) = 1, sign(0) = 0, sign(-.3) = -1
sin(x)	Sinus
sqrt(x)	Quadratwurzel
tan(x)	Tangens

### C.1.5 Skript-Dateien

Für einfache Probleme ist es schnell und effizient, die Befehle am MATLAB-Prompt einzugeben. Für größere und umfangreichere Aufgabenstellungen bietet MATLAB die Möglichkeit, sogenannte Skript-Dateien zu verwenden, in denen die Befehle in Form einer Textdatei aufgeschrieben sind und die man am Prompt übergibt. MATLAB öffnet dann diese Dateien und führt die Befehle so aus, als hätte man sie am Prompt eingegeben. Die Datei nennt man Skript-Datei oder M-Datei, wobei der Ausdruck M-Datei daher rührt, dass diese Dateien das Suffix `.m` haben, z.B. `newton.m`. Um eine M-Datei zu erstellen, ruft man einen Editor auf und speichert die Datei in dem Verzeichnis, von dem aus man MATLAB gestartet hat oder starten wird. Die Datei, z.B. `newton.m`, wird in MATLAB dann durch Eingabe von `newton` am Prompt aufgerufen.

Für die Benutzung von Skript-Dateien hat MATLAB unter anderem folgende hilfreichen Befehle

M-Datei-Funktionen	
disp(ans)	zeigt den Wert der Variablen <code>ans</code> , ohne ihren Namen auszugeben
input	erwartet vom Benutzer eine Eingabe
keyboard	übergibt zeitweise die Kontrolle an die Tastatur
pause	hält das Programm an, bis eine Taste betätigt wird

Die folgende Skript-Datei `beispiel1.m`

```
% beispiel1.m
% Beispiel fuer eine Skript-Datei
tmp = input('Geben Sie bitte eine Zahl an > ');
3 * tmp;
```

führt zu der Ausgabe

```
>> beispiel1
Geben Sie bitte eine Zahl an > 6
ans =
    18
```

## C.1.6 Dateiverwaltung

MATLAB unterstützt eine Vielzahl von Dateiverwaltungsbefehlen, welche es einem ermöglichen Dateien zu listen, Skript-Dateien anzusehen oder zu löschen und Verzeichnisse zu wechseln.

Datei-Management-Funktionen	
<code>cd path</code>	wechselt in das Verzeichnis <code>path</code>
<code>delete beispiel</code>	löscht die Datei <code>beispiel.m</code>
<code>ls</code>	zeigt alle Dateien im aktuellen Verzeichnis an
<code>pwd</code>	zeigt den aktuellen Verzeichnispfad an
<code>type beispiel</code>	zeigt den Inhalt der Datei <code>beispiel.m</code> im Befehl-Fenster
<code>what</code>	zeigt alle M-Dateien und MAT-Dateien im aktuellen Verzeichnis an

## C.1.7 Hilfe

**Online-Hilfe:** Da sich nicht jeder Benutzer alle MATLAB-Befehle merken kann oder auch von einigen auch nur die Syntax unklar ist, bietet MATLAB die Möglichkeit der Online-Hilfe. Dabei gibt es prinzipiell mehrere Möglichkeiten. Ist einem ein Befehl bekannt und man sucht Informationen über die Syntax, so gibt es den Befehl `help`.

```
>> help sqrt

SQRT    Square root.
        SQRT(X) is the square root of the elements of X. Complex
        results are produced if X is not positive.

        See also SQRTM.
```

Als Beispiel haben wir uns hier die Hilfe zu dem Befehl `sqrt` ausgeben lassen.

Die andere Möglichkeit der von MATLAB gelieferten Hilfe ist durch den Befehl `lookfor` gegeben. Hier durchsucht das Programm alle ersten Zeilen der MATLAB Hilfe-Kennwörter und Skript-Dateien die im MATLAB Suchpfad zu finden sind. Das Bemerkenswerte dabei ist, dass dieser Begriff kein Befehl zu sein braucht.

```
>> lookfor cholesky

CHOL    Cholesky factorization

>> CHOL    Cholesky factorization.

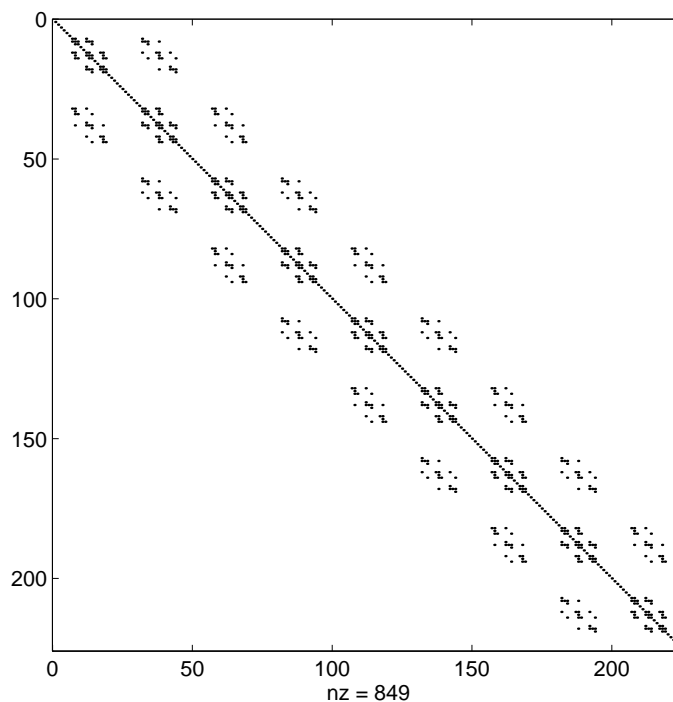
CHOL(X) uses only the diagonal and upper triangle of X.
The lower triangular is assumed to be the (complex conjugate)
transpose of the upper. If X is positive definite, then
R = CHOL(X) produces an upper triangular R so that R'*R = X.
If X is not positive definite, an error message is printed.

With two output arguments, [R,p] = CHOL(X) never produces an
error message. If X is positive definite, then p is 0 and R
is the same as above. But if X is not positive definite, then
p is a positive integer and R is an upper triangular matrix of
order q = p-1 so that R'*R = X(1:q,1:q).

>> lookfor factorization

CHOL    Cholesky factorization.
QRDELETE Delete a column from the QR factorization.
QRINSERT Insert a column in the QR factorization.
SYMBFACT Symbolic factorization analysis.
```

Eine weitere Möglichkeit, sich Hilfe zu verschaffen, besteht darin, das Helpdesk aufzurufen. Wenn Sie helpdesk am Prompt eingeben, öffnet sich die folgende Hilfsumgebung



## C.2 Mathematik mit Matrizen

### C.2.1 Matrixkonstruktion und Adressierung

einfache Matrix Konstruktionen	
$x=[1\ 4\ 2*\pi\ 4]$	erstelle einen Zeilenvektor $x$ mit genannten Einträgen
$x=anfang:ende$	erstelle einen Zeilenvektor $x$ beginnend mit $anfang$ , Inkrement 1 und endend mit $ende$
$x=anfang:inkrement:ende$	Ähnliches wie oben mit dem Inkrement $inkrement$
$x=linspace(anfang,ende,n)$	erzeugt einen Zeilenvektor der Dimension $n$ mit
	$x(i) = \frac{(n-i) \cdot anfang + (i-1) \cdot ende}{n-1}$

Im Folgenden sind einige charakteristische Beispiele aufgeführt.

```
>> B = [1 2 3 4; 5 6 7 8]
B =
     1     2     3     4
     5     6     7     8
```

Der Operator  $'$  liefert für reelle Matrizen die Transponierte.

```
>> C = B'
C =
     1     5
     2     6
     3     7
     4     8
```

Der Doppelpunkt `:` in der zweiten Komponente spricht alle vorhandenen Spalten an, d.h. er ist ein zu `1:4` äquivalenter Ausdruck.

```
>> C = B(1, :)
C =
    1    2    3    4
>> C = B(:, 3)'
C =
    3    7
```

Es lassen sich auch einzelne Komponenten neu definieren.

```
>> A = [1 2 3; 4 5 6; 7 8 9]
A =
    1    2    3
    4    5    6
    7    8    9
>> A(1, 3) = 9
A =
    1    2    9
    4    5    6
    7    8    9
```

Ist ein Eintrag noch nicht definiert, so verwendet MATLAB die minimale Erweiterung dieser Matrix und setzt undefinierte Einträge zu Null.

```
>> A(2, 5) = 4
A =
    1    2    9    0    0
    4    5    6    0    4
    7    8    9    0    0
```

Im Folgenden werden die Vektoren  $(3, 2, 1)$  und  $(2, 1, 3, 1, 5, 2, 4)$  dazu verwendet, die Matrix `C` zu indizieren, d.h. `C` hat die Struktur

```
A(3, 2) A(3, 1) A(3, 3) A(3, 1) A(3, 5) A(3, 2) A(3, 4)
A(2, 2) A(2, 1) A(2, 3) A(2, 1) A(2, 5) A(2, 2) A(2, 4)
A(1, 2) A(1, 1) A(1, 3) A(1, 1) A(1, 5) A(1, 2) A(1, 4)
```

In MATLAB erhält man nun

```
>> C=A(3:-1:1, [2 1 3 1 5 2 4])
C =
    8    7    9    7    0    8    0
    5    4    6    4    4    5    0
    2    1    9    1    0    2    0
```

Ein weiteres Beispiel für Indizierung ist

```
>> C=C(1:2, 2:3)
C =
    7    9
    4    6
```

Im nächsten Beispiel wird ein Spaltenvektor dadurch konstruiert, dass alle Elemente aus der Matrix `C` hintereinander gehängt werden. Dabei wird spaltenweise vorgegangen.

```
>> b=C(:)'  
b =  
    7    4    9    6
```

Das Löschen einer ganzen Zeile oder Spalte kann durch das Umdefinieren in eine  $0 \times 0$ -Matrix geschehen, z.B.

```
>> C(2,:)=[ ]  
C =  
    7    9
```

## C.2.2 Skalar-Matrix-Operationen

In MATLAB sind Skalar-Matrix-Operationen in dem Sinne definiert, dass Addition, Subtraktion, Division und Multiplikation mit einem Skalar elementweise durchgeführt werden. Es folgen zwei erklärende Beispiele.

```
>> B - 1  
ans =  
    0    1    2    3  
    4    5    6    7  
>> 9 + 3 * B  
ans =  
   12   15   18   21  
   24   27   30   33
```

## C.2.3 Matrix-Matrix-Operationen

Die Operationen zwischen Matrizen sind nicht so kanonisch zu definieren wie die zwischen Skalar und Matrix, insbesondere sind Operationen zwischen Matrizen unterschiedlicher Dimension schwer zu definieren. Des Weiteren sind die Operationen  $*$  und  $.*$ , bzw.  $/$  und  $./$  sowie  $\backslash$  und  $.\backslash$  zu unterscheiden. In nachfolgender Tabelle sind die Matrixoperationen beschrieben.



komponentenweise Matrixoperationen	
Beispieldaten	$a = [a_1, a_2, \dots, a_n]$ , $b = [b_1, b_2, \dots, b_n]$ , $c$ ein Skalar
komp. Addition	$a + c = [a_1 + c \ a_2 + c \ \dots \ a_n + c]$
komp. Multiplikation	$a * c = [a_1 * c \ a_2 * c \ \dots \ a_n * c]$
Matrix-Addition	$a + b = [a_1 + b_1 \ a_2 + b_2 \ \dots \ a_n + b_n]$
komp. Matrix-Multiplikationen	$a .* b = [a_1 * b_1 \ a_2 * b_2 \ \dots \ a_n * b_n]$
komp. Matrix-Div. von rechts	$a ./ b = [a_1 / b_1 \ a_2 / b_2 \ \dots \ a_n / b_n]$
komp. Matrix-Div. von links	$a .\backslash b = [b_1 / a_1 \ b_2 / a_2 \ \dots \ b_n / a_n]$
komp. Matrix-Potenz	$a.^c = [a_1^c \ a_2^c \ \dots \ a_n^c]$
	$c.^a = [c^{a_1} \ c^{a_2} \ \dots \ c^{a_n}]$
	$a.^b = [a_1^{b_1} \ a_2^{b_2} \ \dots \ a_n^{b_n}]$

Es folgen nun einige Beispiele zu Matrixoperationen

```

>>g=[1 2 3; 4 5 6]; % zwei neue Matrizen
>>h=[2 2 2; 3 3 3];
>>g+h % addiere g und h komponentenweise
ans =
     3     4     5
     7     8     9
>>ans-g % subtrahiere g von der vorherigen Antwort
ans =
     2     2     2
     3     3     3
>>h.*g % multipliziere g mit h komponentenweise
ans =
     2     4     6
    12    15    18
>>g*h' % multipliziere g mit h'
ans =
    12    18
    30    45

```

## C.2.4 Matrix-Operationen und -Funktionen

Matrixfunktionen	
<code>reshape(A,m,n)</code>	erzeugt aus den Einträgen der Matrix $A$ eine $m \times n$ -Matrix, wobei die Einträge spaltenweise aus $A$ gelesen werden.
<code>diag(A)</code>	ergibt die Diagonale von $A$ als Spaltenvektor
<code>diag(v)</code>	erzeugt eine Diagonalmatrix mit dem Vektor $v$ in der Diagonalen
<code>tril(A)</code>	extrahiert den unteren Dreiecksanteil der Matrix $A$
<code>triu(A)</code>	extrahiert den oberen Dreiecksanteil der Matrix $A$

Es folgen einige Beispiele

```

>>g=linspace(1,9,9) % ein neuer Zeilenvektor
g =
     1     2     3     4     5     6     7     8     9
>>B=reshape(g,3,3) % macht aus g eine 3 x 3 Matrix
B =
     1     4     7
     2     5     8
     3     6     9
>>tril(B)
ans =
     1     0     0
     2     5     0
     3     6     9

```



Funktion	Bedeutung
<code>R=chol(A)</code>	Choleskyzerlegung
<code>cond(A)</code>	Konditionszahl der Matrix $A$
<code>d=eig(A)</code>	Eigenwerte und -vektoren
<code>[V,d]=eig(A)</code>	
<code>det(A)</code>	Determinante
<code>hess(A)</code>	Hessenbergform
<code>inv(A)</code>	Inverse
<code>[L,U]=lu(A)</code>	Zerlegung gegeben durch Gauss-Algorithmus
<code>norm(A)</code>	euklidische-Norm
<code>rank(A)</code>	Rang der Matrix $A$

**Bemerkung:** Der `\` Operator ist auch für Matrizen definiert und liefert in Kombination mit Vektoren für reguläre Matrizen ihre Inverse, d.h.  $A^{-1}x = A \backslash x$ .



## C.2.5 Spezielle Matrizen

spezielle Matrizen	
<code>eye(n)</code>	erzeugt eine Einheitsmatrix der Dimension $n$
<code>ones(m,n)</code>	erzeugt eine $m \times n$ -Matrix mit den Einträgen 1
<code>zeros(m,n)</code>	erzeugt eine $m \times n$ -Matrix mit den Einträgen 0

## C.2.6 Spezielle Funktionen für schwachbesetzte Matrizen

Bei vielen numerischen Anwendungen treten schwachbesetzte Matrizen auf. MATLAB hat für solche Matrizen besondere Sparse-Funktionen, die dieser Eigenschaft Rechnung tragen.



Funktion	Bedeutung
<code>find(A)</code>	findet Indizes von Nichtnulleinträgen
<code>nnz(A)</code>	Anzahl an Nichtnulleinträgen
<code>spdiags(v)</code>	erzeugt eine Sparse-Diagonalmatrix mit dem Vektor $v$ als Diagonale
<code>speye(n)</code>	erzeugt eine Sparse-Einheitsmatrix
<code>spy(A)</code>	visualisiert die Struktur der Matrix $A$

Kurze Illustration der Funktionsweise obiger Befehle anhand einiger Beispiele.

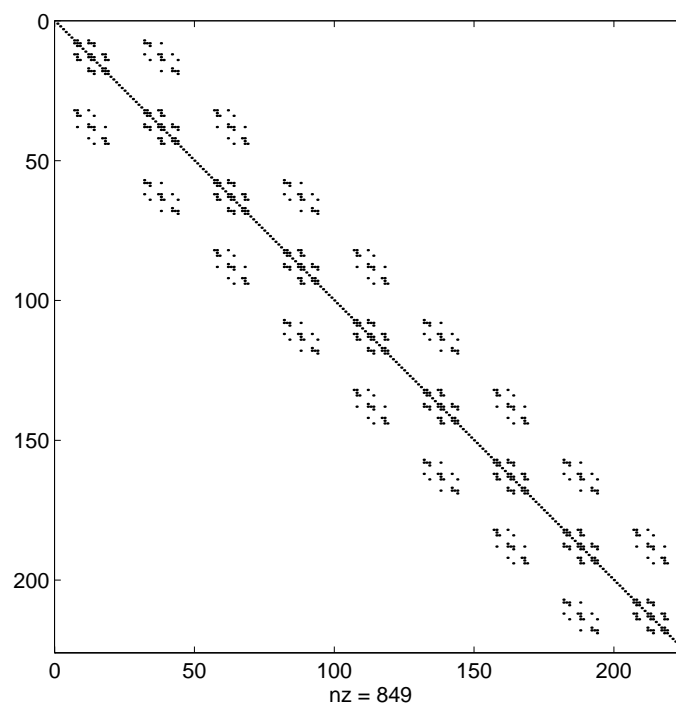
```

>> E=eye(100);      % vollbesetzte 100 x 100 Einheitsmatrix
>> Es=sparse(E);    % Sparse-Version von E
>> whos
      Name      Size      Elements      Bytes      Density      Comple
      E  100 by 100      10000      80000          Full        No
      Es  100 by 100         100       1600      0.0100        No
Grand total is 10100 elements using 81600 bytes

>> A=spdiags([7*ones(4,1),ones(4,1),2*ones(4,1)],[-1,0,1],4,4);
>> nnz(A)
ans =
      10
>> full(A)
ans =
      1      2      0      0
      7      1      2      0
      0      7      1      2
      0      0      7      1

```

Als Beispiel für `spy` sei hier die Besetzungsstruktur einer 3D-FEM Steifigkeitsmatrix gezeigt.



### C.3 Datenverwaltung

Für die meisten Anwendungen genügt es, Datenfelder in einem Format abzuspeichern und wieder laden zu können. Die Befehle `load` und `save` setzen voraus, dass die Daten in einem System unabhängigen, binären Format in einer Datei mit dem Suffix `.mat` gespeichert sind oder in einem einfachen ASCII-Format vorliegen.

### C.3.1 Daten speichern

Im Folgenden wird eine  $3 \times 5$ -Matrix im binär-Format in der Datei `A.mat` gespeichert. Diese Daten sind sehr kompakt gespeichert.

```
>> A=zeros(3,5);
>> save A
```

Gibt man sich aber den Inhalt dieser Datei auf dem Bildschirm aus, so gibt er wenig Sinn. Möchte man sich also z.B. einen Lösungsvektor sichern um ihn später „per Hand zu analysieren“ so speichere man die Daten als ASCII-Datei, dabei kann man wählen zwischen einer 8-stelligen oder 16-stelligen Abspeicherung.

```
>> save mat1.dat A -ascii           % 8-stellige Speicherung
>> save mat2.dat A -ascii -double  % 16-stellige Speicherung
```

Hier wurde die Matrix mit 8-stelligem Format in der Datei `mat1.dat` gespeichert, bzw. 16-stellig in der Datei `mat2.dat`.

### C.3.2 Daten laden

Mit dem Befehl `load A` versucht MATLAB, die in `A.mat` gespeicherten Daten in einem Datenfeld `A` zu speichern. Auch ASCII-Dateien kann MATLAB lesen. Da es hier jedoch keine Standardendung gibt, ist die Datei inklusive Endung anzugeben. Es ist darauf zu achten, dass ein rechteckiges Feld an Daten vorliegt, d.h. dass  $m$  Zeilen mit jeweils  $n$  numerischen Werten vorliegen. MATLAB erstellt dann eine  $m \times n$ -Matrix mit dem Namen der Datei ohne Suffix.

```
>> load mat1.dat
>> whos

      Name      Size      Elements      Bytes      Density      Complex
      mat1      3 by 5          15         120         Full         No

Grand total is 15 elements using 120 bytes
```

## C.4 Ausgabe von Text

Mit dem Befehl `fprintf` lassen sich Strings, d.h. Zeichenfolgen, auf dem Bildschirm ausgeben.

```
>> fprintf('\n Hello world %12.3e\n',4);

Hello world      4.000e+00
```

Man sieht, dass der auszugebende Text zusätzliche Zeichen enthält, die nicht mit ausgedruckt werden. Diese Zeichen nennt man Escape-Sequenzen. In obigem Beispiel ist die Sequenz `\n` eingebaut. `\n` steht für „newline“ und sorgt dafür, dass bei der Textausgabe an diesen Stellen eine neue Zeile begonnen wird. Der Ausdruck `%12.3e` dient als Platzhalter für einen reellen Wert, der durch Komma getrennt hinter der Zeichenkette folgt. Dabei sei die Zahl in Exponentialdarstellung auszugeben, wofür 12 Stellen mit 3 Nachkommastellen bereitgestellt. Im Beispiel ist dies der Wert 4. Anstatt eines expliziten Wertes können auch Variablen oder Ausdrücke, z.B. `3 * 4` stehen. Ein Platzhalter kann mehrfach in einem `printf`-Befehl vorkommen. In diesem Fall müssen hinter der Zeichenkette genau so viele Werte folgen, wie Platzhalter angegeben sind. Die Reihenfolge der Werte muss mit der Reihenfolge der Platzhalter übereinstimmen, da die Ausdrücke von links nach rechts bewertet werden. Die Escape-Sequenzen dürfen im Text an beliebiger Stelle stehen.

Ausgabeformate	
Befehl	Ausgabe
<code>fprintf('%0e\n',1.234567)</code>	1e00+
<code>fprintf('%0.2e\n',1.234567)</code>	1.23e00+
<code>fprintf('%0.5e\n',1.234567)</code>	1.23456e00+
<code>fprintf('%10.0e\n',1.234567)</code>	1e00+
<code>fprintf('%10.2e\n',1.234567)</code>	1.23e00+
<code>fprintf('%10.5e\n',1.234567)</code>	1.23456e00+
<code>fprintf('%10.2f\n',1.234567)</code>	1.23
<code>fprintf('%10.5f\n',1.234567)</code>	1.23457

MATLAB rundet numerische Werte bei der Ausgabe, wenn nötig!

## C.5 Kontrollbefehle

### C.5.1 For-Schleifen

1. Mit jeglicher gültigen Matrix-Darstellung lässt sich eine FOR-Schleife definieren, z.B.

```
>> data = [1 7 3 2; 5 4 7 2]
data =
     1     7     3     2
     5     4     7     2
>> for n=data
    x=n(1)-n(2)
end
x =
    -4
x =
     3
x =
    -4
x =
     0
```

2. FOR-Schleifen können nach Belieben geschachtelt werden.

```
>> for k=3:5
    for l=4:-1:2
        A(k,l)=k^2-l;
    end
end
>> A
A =
     0     0     0     0
     0     0     0     0
     0     7     6     5
     0    14    13    12
     0    23    22    21
```



3. FOR-Schleifen sollten vermieden werden, wann immer sie durch eine äquivalente Matrix-Darstellung ersetzt werden können. Der folgende Ausdruck ist darunten in optimierter Version aufgeführt.

```

>> for n=1:10
    x(n)=sin(n*pi/10);
end
>>x
x =
    Columns 1 through 7
    0.3090  0.5878  0.8090  0.9511  1.0000  0.9511  0.8090
    Columns 8 through 10
    0.5878  0.3090  0.0000

>> n=1:10;
>> x=sin(n*pi/10);
>> x
x =
    Columns 1 through 7
    0.3090  0.5878  0.8090  0.9511  1.0000  0.9511  0.8090
    Columns 8 through 10
    0.5878  0.3090  0.0000

```

**4. Um die Ausführungsgeschwindigkeit zu maximieren, sollte benötigter Speicherplatz vor Ausführung der FOR-Schleife alloziert werden.**



```

>> x=zeros(1,10);
>> x(1)=0.5;
>> for n=2:10
    x(n)=x(n-1)*sin(n*pi/10);
end
>> x
x =
    Columns 1 through 7
    0.5000  0.2939  0.2378  0.2261  0.2261  0.2151  0.1740
    Columns 8 through 10
    0.1023  0.0316  0.0000

```

## C.5.2 WHILE-Schleifen

WHILE-Schleifen sind wie folgt aufgebaut.

```

while Aussage
    Anweisungen
end

```

**Die Anweisungen zwischen while und end werden so lange ausgeführt, wie Aussage wahr ist, z.B.**

```

>> num=0; EPS=1;
>> while (1+EPS) > 1
    EPS=EPS/2;
    num=num+1;
end
>> num
num =
    53

```

### C.5.3 IF-ELSE-END Konstrukte

Eine IF-ELSE-END-Schleife enthält nach dem IF eine Aussage, die daraufhin überprüft wird, ob sie wahr oder falsch ist. Ist sie wahr, so werden die in den folgenden Zeilen stehenden Anweisungen ausgeführt und die Schleife beendet. Ist sie falsch, so erfolgen die Anweisungen, die dem ELSE folgen (ELSE ist optional). Solch ein Konstrukt kann erweitert werden um beliebig viele ELSIF Befehle, die dieselbe Funktion haben wie der am Anfang stehende IF Befehl, aber nur beachtet werden, falls alle vorher überprüften Aussagen falsch sind.

```
if Aussage1
    Anweisungen, wenn Aussage1 wahr
elseif Aussage2
    Anweisungen, wenn Aussage1 falsch und Aussage2 wahr
else
    Anweisungen, wenn Aussage1 und Aussage2 falsch
end
```

### C.5.4 Relationen und logische Operatoren

Relationen	
<	kleiner als
<=	kleiner als oder gleich
>	größer als
>=	größer als oder gleich
==	gleich
~=	ungleich

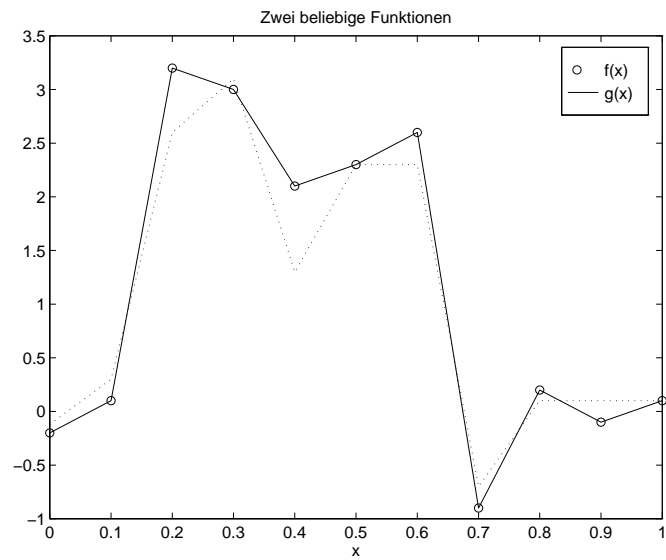
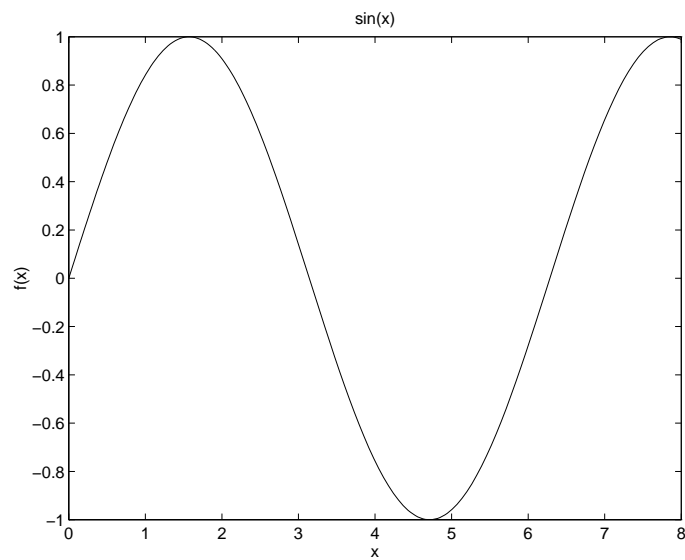
logische Operatoren	
&	UND
	ODER
~	NICHT

## C.6 Graphische Darstellung

### C.6.1 Zweidimensionale Graphiken

```
>> f='sin(x)';
>> fplot(f,[0 8]);
>> title(f),xlabel('x'),ylabel('f(x)');

>> x=0:0.1:1;
>> y=[-0.2 0.1 3.2 3 2.1 2.3 2.6 -0.9 0.2 -.1 .1];
>> z=[-0.12 0.3 2.6 3.1 1.3 2.3 2.3 -0.7 0.1 .1 .1];
>> plot(x,y,'o',x,y,x,z,':');
>> title('Zwei beliebige Funktionen'),xlabel('x');
>> legend('f(x)','g(x)')
```



### Linientypen und Farben

Symbol	Farbe	Symbol	Linientyp
y	gelb	·	Punkt
m	magenta	○	Kreis
c	cyan	x	x-Markierung
r	rot	+	+ -Markierung
g	grün	*	Sternchen
b	blau	—	durchgezogene Linie
w	weiß	:	gepunktete Linie
k	schwarz	—.	Strichpunkt-Linie
		— —	gestrichelte Linie

**2-D Graphikanweisung**

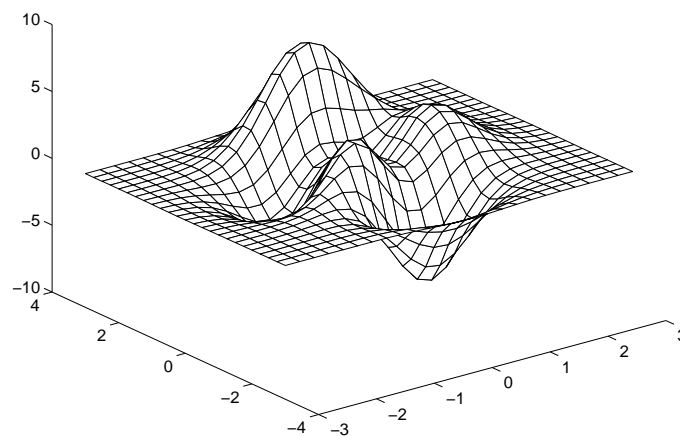
axis	modifiziert die Axen-Proportionen
clf	löscht die Graphik im Graphik-Fenster
close	schließt das Graphik-Fenster
grid	erzeugt ein achsenparalleles Gitter
hold	ermöglicht das Überlagern von Graphiken
subplot	erstellt mehrere Teilgraphiken in einem Fenster
text	gibt Text an vorgegebener Stelle aus
title	zeigt einen Titel an
xlabel	beschriftet die x-Achse
ylabel	beschriftet die y-Achse
colormap(white)	wechselt die Farbtabelle, für S/W-Monitore

**C.6.2 Dreidimensionale Graphiken**

```

>> [X,Y,Z] = peaks(25)
>> mesh(X,Y,Z)

```

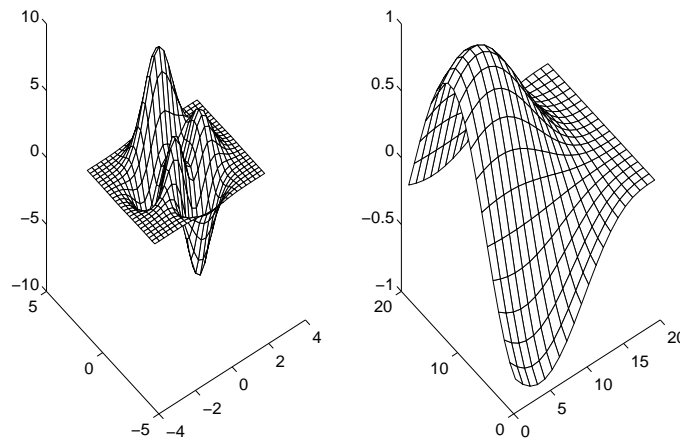


```

>> subplot(1,2,1);
>> [X,Y,Z] = peaks(25);
>> mesh(X,Y,Z);
>> subplot(1,2,2);
>> X=1:20;
>> Y=1:20;
>> Z(X,Y)=-(cos(X/4))'*(sin((20-Y)/10).^3);
>> mesh(X,Y,Z);

```





### C.6.3 Graphiken drucken

Graphik-Druckbefehl		
print [-dAusgabetyyp] [-Optionen] [Dateiname]		
Ausgabetyyp	-dps	Postscript für Schwarzweißdrucker
	-dpSC	Postscript für Farbdrucker
	-deps	Encapsulated Postscript
	-depSC	Encapsulated Color Postscript
Optionen	-P<Drucker>	Spezifiziert den zu benutzenden Drucker

Die Eingabe

```
>> print fig4 -deps
```

erzeugt die Datei fig4.eps.

## C.7 Fortgeschrittenes

### C.7.1 MATLAB-Skripte

Wie schon im Abschnitt C.1.5 erwähnt, lässt sich eine Abfolge von MATLAB-Befehlen auch in einer Datei speichern. Diese kann man dann am Befehl-Fenster aufrufen und die gespeicherte Folge von Befehlen wird ausgeführt. Solche Dateien werden als MATLAB-Skripte oder M-Files bezeichnet. Alle o.g. Befehle lassen sich in einer Datei z.B. mit dem Namen `abc.m` zusammenstellen. Für die Wahl des Dateinamens gelten dabei die folgenden Regeln:

- das erste Zeichen ist ein Buchstabe und
- die Datei hat die Endung **.m**.

Um ein solches M-File zu erstellen, ruft man den Editor auf, der es erlaubt Text einzugeben und diesen als Datei zu speichern. Wenn man den Editor gestartet hat, gebe man die folgenden Befehle ein und speichere die Datei unter dem Namen `abc.m`

```

a = 1;
b = 3;
c = -5;
x1 = (-b + sqrt(b^2 - 4*a*c)) / (2*a)
x2 = (-b - sqrt(b^2 - 4*a*c)) / (2*a)

```

Um nun die Befehle aus der Datei `abc.m` auszuführen, gibt man im MATLAB-Befehlsfenster

```
>> abc
```

ein. MATLAB sucht im aktuellen Pfad nach der Datei `abc.m` und führt die darin enthaltenen Befehle aus. Das aktuelle Verzeichnis wird angezeigt, wenn man

```
>> pwd
```

eingibt (`pwd`, engl. *print working directory*).

## C.7.2 Erstellen eigener Funktionen

Es wird sicherlich etwas komfortabler sein, eine eigene Funktion zu haben, der man  $a, b$  und  $c$  als Argument übergibt und die beiden Wurzeln als Ergebnis erhält, als jedesmal erneut ein eigenes M-File anzufertigen. Ein solches Programm könnte z.B. die folgende Datei `root.m` sein.

```

%-----
modified "abc.m"
%-----

a = input('Enter a: ');
b = input('Enter b: ');
c = input('Enter c: ');

x1 = (-b + sqrt(b^2 - 4*a*c)) / (2*a)
x2 = (-b - sqrt(b^2 - 4*a*c)) / (2*a)

```

Die Prozentzeichen in den ersten Zeilen dienen der Dokumentation. Alles was einem solchen Zeichen folgt, wird von MATLAB nicht ausgewertet, sprich interpretiert. Die Argumente werden in dieser Funktion jedoch nur bedingt übergeben und ausgegeben. Eine Funktion, die diese Aufgabe erfüllt, ist die folgende.

```

%-----
modified "abc.m"
%-----

function [x1, x2] = quadroot(a,b,c)

radical = sqrt(b^2 - 4*a*c);

x1 = (-b + radical) / (2*a)
x2 = (-b - radical) / (2*a)

```

Wenn eine Funktion keine Rückgabewerte hat, können die eckigen Klammern mit den Variablen und das Gleichheitszeichen fehlen. Fehlen die Eingabeparameter, so kann auch der Klammerausdruck nach `quadroot` fehlen. Auf die in der Funktion verwendeten Variablen, hier `radical`,

kann man vom Befehlsfenster nicht zugreifen. Ist die Funktion in der Datei `quadroot.m` gespeichert, so erhält man die Wurzeln, indem man

```
[x1, x2] = quadroot(1, 3, -5);
```

eingibt. Es kann vom Befehlsfenster oder einer anderen Datei immer nur die erste Funktion in einer Datei aufgerufen werden. Dies bedeutet, in einer Datei können mehrere Funktionen stehen, aber nur die erste Funktion kann extern aufgerufen werden. Alle weiteren Funktionen können nur von Funktionen in der gleichen Datei aufgerufen werden. Für den Anfang ist es einfacher, wenn jede Datei nur eine Funktion enthält. Entscheidend für den Funktionsnamen ist der Name der Datei.



# D SPEICHERFORMATE FÜR SCHWACH BESETZTE MATRIZEN

Im Folgenden werden wir einige Speicherformate für schwach besetzte Matrizen vorstellen, welche speziell im Zusammenhang mit Iterationsverfahren von Interesse sind. Den Speicherformaten für schwach besetzte Systeme kommen aus zweierlei Gründen eine besondere Bedeutung zu:

- Die Wahl der Speicherform hat direkte Auswirkungen auf die verwendeten Lösungsverfahren (spezielle Matrixform, z.B. Bandmatrix, symmetrische Matrix, etc.). Unter anderen wegen ihres Auffüllverhaltens (Erzeugung neuer Nichtnull-Elemente) der verschiedenen Algorithmen, der Zugriffshäufigkeit auf spezielle Einträge (z.B. der Diagonaleinträge) und somit ihrer Konsequenzen auf die Speicherstruktur.
- In den meisten höheren Programmiersprachen C/C++, Java, Fortran, etc. stehen für die meisten Standardprobleme der numerischen Linearen Algebra verschiedene Programmpakete (Lapack, Linpack, Eispack u.v.a.) mit fertig implementierten numerischen Algorithmen zur Verfügung. Diesen Bibliotheken müssen die Daten aber in bestimmten Datenstrukturen übergeben werden.

## D.1 Koordinatenformat (COO-Format)

Das einfachste Format zur Speicherung von schwach besetzten Matrizen für iterative Verfahren ist das Koordinatenformat (**C**oordinate **S**torage, COO-Format).

Es verwendet 3 Vektoren der Länge  $\text{nnz}(A)$  ( $\text{nnz} \simeq$  number of nonzero elements):

- ein Vektor „Wert“ vom Datentyp der Einträge  $a_{ij}$ , in dem die Nichtnull-Elemente von  $A$  in beliebiger Reihenfolge stehen,
- ein Index-Vektor „Zeilenindex“ und
- ein Index-Vektor „Spaltenindex“, in denen zu jedem Element von „Wert“ die entsprechenden Koordinaten, d.h. Zeilen- und Spaltenindices, aus der Matrix  $A$  eingetragen werden. Meist sind dies Vektoren vom `unsigned (long) integer`.

Dieses Format wird häufig wegen seiner Einfachheit verwendet (dient oft als eine von vielen möglichen Schnittstelle zu Bibliotheken). Der Umstand, dass die Nichtnull-Matrixelemente in beliebiger Reihenfolge im Feld „Wert“ stehen dürfen, ist sowohl ein Vorteil als auch ein Nachteil des COO-Formats. Einerseits können die bei der Lösung des Gleichungssystems neu auftretenden Nichtnull-Elemente einfach am Ende der Felder ohne irgendwelchen Umordnungsaufwand angehängt werden. Andererseits ist der Suchaufwand nach einem bestimmten Matrixelement enorm, da ja keinerlei Strukturinformation der Matrix  $A$  in das Speicherformat selbst eingeht (ist das gesuchte Element gleich Null, müssen die Felder komplett durchsucht werden, bis diese Information vorliegt). Ein weiterer Nachteil ist auch der immer noch recht hohe Speicherbedarf.

Ein Beispiel soll die Datenspeicherung illustrieren.

### Beispiel D.1.1 Die Matrix

$$A = \begin{pmatrix} 10 & 0 & 0 & 22 & 31 & 0 \\ -4 & 8 & 0 & 0 & 0 & 0 \\ 0 & 0 & -6 & 9 & 0 & 0 \\ 0 & 0 & 0 & 7 & 0 & 0 \\ 0 & 0 & 5 & 0 & 20 & 11 \\ 12 & 0 & 0 & 14 & 0 & 44 \end{pmatrix}$$

geht über zu

Wert $a_{i,j}$	44	-6	20	12	14	31	5	-4	10	8	7	9	22	11
Zeilenindex $i$	6	3	5	6	6	1	5	2	1	2	4	3	1	5
Spaltenindex $j$	6	3	5	1	4	5	3	1	1	2	4	4	4	6

In Matlab kann man einen eigenen Datentypen COO-Matrix definieren und sowohl die Operatoren  $+$ ,  $*$ ,  $-$ ,  $/$  überladen, als auch die Ein- und Ausgabe dem bekannten Verhalten anpassen. Im Folgenden beschränken wir uns auf die Wiedergabe einiger Methoden, die den Datentypen `coomatrix` definieren, Ein- und Ausgabe sowie Multiplikation ermöglichen, so dass man u.a. das CG-Verfahren testen könnte. Die Routinen bzw. Methoden zu einer Klasse stehen in Matlab üblicherweise in einem Unterverzeichnis gleichen Namens, d.h. die Methoden `coomatrix.m` etc. stehen im Verzeichnis `@coomatrix`. Für einen ersten Einstieg in Klassendefinitionen in Matlab, sei auf das Beispiel `polynom` in der Matlab-Dokumentation verwiesen.

#### MATLAB-Funktion: @coomatrix/coomatrix.m

```

1 function p = coomatrix(i,j,a)
2 %COOMATRIX Coordinate format class constructor.
3 %   A = COOMATRIX(i,j,a) creates a compressed matrix object from
4 %   the vectors i,j,a, containing the entries of a matrix A.
5 if nargin == 0
6     p.a = []; p.row = []; p.col = [];
7     p.m = 0;
8     p.n = 0;
9     p = class(p,'coomatrix');
10 elseif nargin == 1 && isa(i,'coomatrix')
11     p = i;
12 elseif nargin == 1 && isnumeric(i)
13     [p.row,p.col,p.a] = find(i);
14     p.row=p.row';p.col=p.col';p.a = p.a';
15     p.m = max(p.row); p.n = max(p.col);
16     p = class(p,'coomatrix');
17 else
18     p.row = i(:).';
19     p.col = j(:).';
20     p.a    = a(:).';
21     p.m = max(p.row);
22     p.n = max(p.col);
23     p = class(p,'coomatrix');
24 end

```

**MATLAB-Funktion: @coomatrix/display.m**

```

1 function display(p)
2 % COOMATRIX/DISPLAY Command window display of coomatrix
3 formattedspace
4 disp([inputname(1),' = '])
5 formattedspace
6 short = ~(max(p.row)>9999 || max(p.col)>9999);
7 for k=1:length(p.row)
8     i = num2str(p.row(k)); j = num2str(p.col(k));
9     if short
10         txt = '          ,          ';
11         txt(5-length(i):5) = ['(',i];
12         txt(7:7+length(j)) = [j,')'];
13     else
14         txt = '          ,          ';
15         txt(11-length(i):11) = ['(',i];
16         txt(13:13+length(j)) = [j,')'];
17     end
18     disp([txt,sprintf('%13.4f',p.a(k))])
19 end
20 formattedspace
21
22 function formattedspace
23 switch get(0,'FormatSpacing')
24     case 'loose'
25         disp(' ');
26 end

```

**MATLAB-Funktion: @coomatrix/subsasgn.m**

```

1 function p = subsasgn(p,s,b)
2 % SUBSASGN
3 switch s.type
4     case '()'
5         if size(s.subs,1)~=1 || size(s.subs,2)~=2
6             error('??? Index exceeds matrix dimensions.')
7         end
8         ii = s.subs{1,1}; jj = s.subs{1,2};
9         for i = 1:length(ii)
10             for j = 1:length(jj)
11                 p.row = [p.row,ii(i)]; p.col = [p.col,jj(j)];
12                 p.a = [p.a,b(i,j)];
13             end
14         end
15         p.m = max(p.row); p.n = max(p.col);
16     otherwise
17         error('Specify value for i and j as A(i,j)')
18 end

```

**MATLAB-Beispiel:**

Anlegen einer Matrix im COO-Format funktioniert mit der Anweisung `coomatrix`.

```
>> A=coomatrix([1,2],[2,3],[1.2,1.4])
A =
    (1,2)    1.2000
    (2,3)    1.4000
```

Durch die Funktion `subsref` können einzelne Einträge zu der Matrix hinzugefügt werden. Indexpaare  $(i,j)$  können dabei doppelt auftreten.

```
>> A(1,4)=2;
>> A(1,4)=3
A =
    (1,2)    1.2000
    (2,3)    1.4000
    (1,4)    2.0000
    (1,4)    3.0000
```

**MATLAB-Funktion: @coomatrix/subsref.m**

```
1 function b = subsref(a,s)
2 % SUBSREF
3 switch s.type
4 case '()'
5     if size(s.subs,1)~=1 || size(s.subs,2)~=2
6         error('??? Index exceeds matrix dimensions.')
7     end
8     ii = s.subs{1,1};
9     jj = s.subs{1,2};
10    b = zeros(length(ii),length(jj));
11    for i = 1:length(ii)
12        if ii(i) > a.m, error('??? Index exceeds matrix dimensions.')
13            ; end
14        for j = 1:length(jj)
15            if jj(j) > a.n, error('??? Index exceeds matrix dimensions.
16                '); end
17            b(i,j) = get_index(a,ii(i),jj(j));
18        end
19    end
20 otherwise
21     error('Specify value for i and j as A(i,j)')
22 end
23
24 function aij = get_index(mat,i,j)
25 aij = 0;
26 ii = find(mat.row==i);
27 if ~isempty(ii)
28     jj = find(mat.col(ii)==j);
29     if ~isempty(jj)
30         for k=1:length(jj)
31             aij = aij + mat.a(ii(jj(k)));
32         end
33     end
34 end
```



**MATLAB-Funktion: @coomatrix/mtimes.m**

```

1 function r = mtimes(A,q)
2 % COOMATRIX/MTIMES    Implement A * q for coordinate matrix,
3 % q vector or scalar
4 if isa(A,'coomatrix') && isnumeric(q)
5     if size(q,1) * size(q,2) == 1
6         A.a = q * A.a;
7         r = A;
8         return
9     elseif A.n ~= size(q,1),
10        error('Inner matrix dimensions must agree.')
11    end
12    r = zeros(A.m,size(q,2));
13    for j = 1:size(q,2)
14        for k = 1:length(A.a)
15            r(A.row(k)) = r(A.row(k)) + A.a(k) * q(A.col(k),j);
16        end
17    end
18 else
19     error('Multiplication not defined for these types.')
20 end

```

**MATLAB-Beispiel:**

Durch die Funktion subsasgn funktioniert auch der Zugriff auf einzelne Einträge der Matrix wie aus Matlab bekannt. Da ein Index  $(i,j)$  ggf. nicht eindeutig ist, muss die ganze Matrix durchsucht werden.

```

A(1,2)
ans =
    1.2000
>> A(1,3)
ans =
    0

```

Multiplikation einer Matrix im COO-Format mit einem Vektor lässt sich mittels des Überladens des  $*$ -Operators mit der Funktion `mtimes.m` wie gehabt realisieren.

```

>> A=coomatrix([1,2],[2,3],[1.2,1.4]);
>> b = A * [2;1;3]
b =
    1.2000
    4.2000

```

**Aufgabe D.1.2** Realisieren Sie eine Methode `ctranspose`, welches die Verwendung des Operators Transponiert-Operators `'` in Matlab wie bekannt ermöglicht.

## D.2 Komprimierte Zeilenspeicherung (CRS-Format)

Das CRS-Format (**C**ompressed **R**ow **S**torage-Format, Komprimierte Zeilenspeicherung) ist bereits eine relativ kompakte Speicherform, die aber nicht von einer bestimmten Struktur der Matrix  $A$  ausgeht. Sinn dieses Formates ist es, sehr schnell alle Nichtnull-Elemente einer bestimmten Zeile zu finden.

Es sei  $A \in \mathbb{R}^{m \times n}$ , dann benötigt das Format 3 Vektoren, 2 der Länge  $\text{nnz}(A)$  und 1 mit der Länge  $m + 1$ :

- ein Vektor „Wert“, in dem die Nichtnull-Elemente von  $A$  zeilenweise, von links nach rechts, stehen;
- einen Index-Vektor „Spaltenindex“, in dem zu jedem Element von „Wert“ der entsprechende Spaltenindex aus der Matrix  $A$  eingetragen wird;
- einen Index-Vektor „Zeilenzeiger“, in dem der Beginn jeder Matrixzeile im Feld „Wert“ gespeichert wird. Bei Index  $m + 1$  wird der Wert  $\text{nnz}(A) + 1$  abgelegt.

Das CRS-Format benötigt gegenüber dem COO-Format deutlich weniger Speicher, kein Eintrag kommt mehrfach vor, ein Index-Vektor ist i. Allg. deutlich kürzer. Auch das Suchen nach einem bestimmten Matrixelement geht hier deutlich schneller. Werden vom Lösungsalgorithmus aber neue Nichtnull-Elemente erzeugt, dann ist die Einordnung doch komplizierter und auch mit mehr Aufwand verbunden als beim COO-Format.

### Beispiel D.2.1 Die Matrix

$$A = \begin{pmatrix} 10 & 0 & 0 & 22 & 31 & 0 \\ -4 & 8 & 0 & 0 & 0 & 0 \\ 0 & 0 & -6 & 9 & 0 & 0 \\ 0 & 0 & 0 & 7 & 0 & 0 \\ 0 & 0 & 5 & 0 & 20 & 11 \\ 12 & 0 & 0 & 14 & 0 & 44 \end{pmatrix}$$

geht über zu

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Wert $a_{ij}$	10	22	31	-4	8	-6	9	7	5	20	11	12	14	44
Spaltenindex $j$	1	4	5	1	2	3	4	7	3	5	6	1	4	6

Zeilenzeiger	1	4	6	8	9	12	15
--------------	---	---	---	---	---	----	----

**MATLAB-Funktion: @crsmatrix/crsmatrix.m**

```

1 function p = crsmatrix(i,j,a)
2 %CRSMATRIX Compressed row storage format class constructor.
3 %   A = CRSMATRIX(i,j,a) creates a compressed matrix object from
4 %   the vectors i,j,a, containing the entries of a matrix A.
5 if nargin == 0
6     p.a = []; p.col = []; p.ptr = [];
7     p.m = 0; p.n = 0;
8     p = class(p,'crsmatrix');
9 elseif nargin == 1 && isa(i,'crsmatrix')
10    p = i;
11 elseif nargin == 1 && isnumeric(i)
12    [p.col,row,p.a] = find(i');
13    p.n = max(p.col); p.m = max(row);
14    p.col = p.col'; p.a = p.a';
15    p.ptr = find([row(:);row(end)+1]-[0;row(:)'])';
16    p = class(p,'crsmatrix');
17 else
18    p = crsmatrix(sparse(i,j,a));
19 end

```

**MATLAB-Funktion: @crsmatrix/display.m**

```

1 function display(p)
2 % CRSMATRIX/DISPLAY Command window display of crs-matrix
3 formattedspace, disp([inputname(1),' = ']), formattedspace
4 short = ~(max(p.col)>9999 || p.ptr(end)>9999+1);
5 for k = 2:length(p.ptr)
6     for j = p.ptr(k-1):p.ptr(k)-1
7         ii = num2str(k-1); jj = num2str(p.col(j));
8         if short
9             txt = '      ,      ';
10            txt(5-length(ii):5) = ['(',ii];
11            txt(7:7+length(jj)) = [jj,')']';
12        else
13            txt = '      ,      ';
14            txt(11-length(ii):11) = ['(',ii];
15            txt(13:13+length(jj)) = [jj,')']';
16        end
17        disp([txt,sprintf('%13.4f',p.a(j))])
18    end
19 end
20 formattedspace
21
22 function formattedspace
23 switch get(0,'FormatSpacing')
24 case 'loose'
25     disp(' ');
26 end

```

**MATLAB-Funktion: @crsmatrix/mtimes.m**

```

1 function r = mtimes(A,q)
2 % CRSMATRIX/MTIMES    Implement A * q for crs-format matrix,
3 % q full matrix of matching size
4 if isa(A,'crsmatrix') && isnumeric(q)
5     if A.n ~= size(q,1),
6         error('Inner matrix dimensions must agree.')
7     end
8     r = zeros(A.m, size(q,2));
9     for j = 1:A.m
10         for k = A.ptr(j):A.ptr(j+1)-1
11             r(j,:) = r(j,:) + A.a(k) * q(A.col(k),:);
12         end
13     end
14 else
15     error('Multiplication not defined for these types.')
16 end

```

**MATLAB-Beispiel:**

Anlegen einer Matrix im CRS-Format funktioniert mit der Anweisung `crsmatrix`

```

>> A = crsmatrix([3,2,1,2,3,1], ...
>>                [2,2,4,3,1,1], ...
>>                [1,2,3,4,5,6])
A =
    (1,1)      6.0000
    (1,4)      3.0000
    (2,2)      2.0000
    (2,3)      4.0000
    (3,1)      5.0000
    (3,2)      1.0000

```

**Aufgabe D.2.2** Realisieren Sie die Methoden `subsref`, `ctranspose`, welche den Zugriff auf einzelne Einträge und das Transponieren in Matlab wie bekannt ermöglichen.

### D.3 Modifizierte komprimierte Zeilenspeicherung (MRS-Format)

Mit dem MRS-Format (**M**odified **C**ompressed **R**ow **S**torage-Format, Modifiziertes CRS-Format) wird nochmals versucht, den ohnehin schon geringen Speicherplatzbedarf des CRS-Formats noch weiter zu senken und direkten Zugriff auf die Diagonalelemente zu haben. Denken Sie z.B. an das Jacobi-Verfahren, bei dem man nicht nur eine Matrix-Vektor-Multiplikation mit der Koeffizientenmatrix sondern auch mit der Inversen ihrer Diagonalen benötigt.

Beim MRS-Format geht man davon aus, dass die Hauptdiagonale voll (oder zumindest überdurchschnittlich hoch) besetzt ist, eine Bedingung, die fast alle in der Praxis vorkommenden Matrizen

erfüllen. Speichert man nun die Hauptdiagonale extra ab, so wird keinerlei zusätzliche Koordinateninformation dafür benötigt, denn man weiß ja, dass z.B. das dritte Hauptdiagonalelement in der dritten Zeile und der dritten Spalte steht.

Diese Grundidee wird im MRS-Format folgendermaßen umgesetzt:

Es sei  $A \in \mathbb{R}^{m \times n}$ . In dem Vektor „Wert“ werden zunächst sämtliche Hauptdiagonalelemente eingetragen, danach wird ein Feld freigelassen. Ab Index  $m + 2$  werden nun wie beim CRS-Format zeilenweise die Nichtnullelemente, natürlich ohne die Hauptdiagonalelemente, abgespeichert. Die ersten  $m + 1$  Elemente des Vektors „Index“ ersetzen das beim CRS-Format separate Feld „Zeilenzeiger“. Ab Index  $m + 2$  werden dann wieder die Spaltenindizes der entsprechenden Matrixelemente des Feldes „Wert“ abgespeichert.

Vor- und Nachteile dieser Speicherform sind ähnlich dem CRS-Format. Das MRS-Format ist relativ beliebt.

### Beispiel D.3.1

$$A = \begin{pmatrix} 10 & 0 & 0 & 22 & 31 & 0 \\ -4 & 8 & 0 & 0 & 0 & 0 \\ 0 & 0 & -6 & 9 & 0 & 0 \\ 0 & 0 & 0 & 7 & 0 & 0 \\ 0 & 0 & 5 & 0 & 20 & 11 \\ 12 & 0 & 0 & 14 & 0 & 44 \end{pmatrix}$$

geht über zu

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Wert $a_{ij}$	10	8	-6	7	20	44		22	31	-4	9	5	11	12	14
Index	8	10	11	12	12	14	16	4	5	1	4	3	5	1	4

### MATLAB-Funktion: @mrsmatrix/mrsmatrix.m

```

1 function p = coomatrix(i,j,a)
2 %COOMATRIX Coordinate format class constructor.
3 %   A = COOMATRIX(i,j,a) creates a compressed matrix object from
4 %   the vectors i,j,a, containing the entries of a matrix A.
5 if nargin == 0
6     p.a = []; p.ind = [];
7     p.m = 0; p.n = 0;
8     p = class(p,'mrsmatrix');
9 elseif nargin == 1 && isa(i,'mrsmatrix')
10    p = i;
11 elseif nargin == 1 && isnumeric(i)
12    % dirty, at least one off-diagonal entry necessary
13    [col,row,a] = find(i');
14    p.n = max(col); p.m = max(row);
15    ptr = find(col==row);
16    p.a = zeros(1,min(p.m,p.n));
17    p.a(col(ptr)) = a(ptr);

```

```

18     ptr = find(col~=row);    row = row(ptr);
19     p.ind = 1+min(p.m,p.n)+find([row(:);row(end)+1]-[0;row(:)])';
20     p.ind = [p.ind,col(ptr)]';
21     p.a = [p.a,0,a(ptr)]';
22     p = class(p,'mrsmatrix');
23 else
24     p = mrsmatrix(sparse(i,j,a));
25 end

```

### MATLAB-Funktion: @mrsmatrix/display.m

```

1 function display(p)
2 % MRSMATRIX/DISPLAY Command window display of mrs-matrix
3 formattedspace
4 disp([inputname(1),' = '])
5 formattedspace
6 short = ~(p.m>9999 || p.n>9999);
7 for k = 1:min(p.m,p.n)
8     if p.a(k)
9         printentries(k,k,p.a(k),short)
10    end
11 end
12 for k = 1:p.m
13     for j = p.ind(k):p.ind(k+1)-1
14         printentries(k,p.ind(j),p.a(j),short)
15     end
16 end
17 formattedspace
18
19 function formattedspace
20 switch get(0,'FormatSpacing')
21     case 'loose'
22         disp(' ');
23 end
24
25 function printentries(ii,jj,aij,short)
26 i = num2str(ii); j = num2str(jj);
27 if short
28     txt = '          ,          ';
29     txt(5-length(i):5) = ['(',i];
30     txt(7:7+length(j)) = [j,')']';
31 else
32     txt = '          ,          ';
33     txt(11-length(i):11) = ['(',i];
34     txt(13:13+length(j)) = [j,')']';
35 end
36 disp([txt,sprintf('%13.4f',aij)])

```

**MATLAB-Funktion: @mrsmatrix/mtimes.m**

```

1 function r = mtimes(A,q)
2 % MRSMATRIX/MTIMES    Implement A * q for mrs-format,
3 % q vector or matrix of matching size
4 if isa(A,'mrsmatrix') && isnumeric(q)
5     if A.n ~= size(q,1),
6         error('Inner matrix dimensions must agree.')
7     end
8     r = zeros(A.m,size(q,2));
9     for j = 1:A.m
10        r(j,:) = A.a(j) * q(j,:);
11        for k = A.ind(j):A.ind(j+1)-1
12            r(j,:) = r(j,:) + A.a(k) * q(A.ind(k),:);
13        end
14    end
15 else
16    error('Multiplication not defined for these types.')
17 end

```

**MATLAB-Beispiel:**

Anlegen einer Matrix im MRS-Format funktioniert z.B. mit der Anweisung `mrsmatrix`. Auch die Multiplikation lässt sich kanonisch realisieren.

```

>> A = mrsmatrix([3,2,1,2,3,1], ...
>>                [2,2,4,3,1,1], ...
>>                [1,2,3,4,5,6])
A =
    (1,1)      6.0000
    (2,2)      2.0000
    (1,4)      3.0000
    (2,3)      4.0000
    (3,1)      5.0000
    (3,2)      1.0000
>> x = [0,1;1,1;2,1;2 0];
>> A*x
ans =
     6     6
    10     6
     1     6

```

**Aufgabe D.3.2** Realisieren Sie die Methoden `subsref`, `ctranspose`, welche den Zugriff auf einzelne Einträge und das Transponieren in Matlab wie bekannt ermöglichen.

**Aufgabe D.3.3** Schreiben Sie eine Funktion, die das Jacobi-Verfahren zur Lösung von linearen Gleichungssystemen für in CRS-Format gespeicherte Matrizen umsetzt. (Tipp: Neben  $A \cdot x$  benötigen Sie noch eine Routine, die  $\text{diag}(\text{diag}(A)) \setminus x$  effizient umsetzt.

## D.4 Harwell-Boeing-Format (CCS-Format)

Das Harwell-Boeing-Format oder auch CCS (Compressed-Column-Storage-Format, Komprimierte Spaltenspeicherung) genanntes Format entspricht im wesentlichen dem CRS-Format, nur dass man hier nicht schnell auf eine Zeile, sondern eine komplette Spalte zugreifen möchte. Die Speicherung zeilenweise von links nach rechts wird durch spaltenweise Speicherung von oben nach unten ersetzt.



# Literaturverzeichnis

- [Arendt/Urban] W. ARENDT, K. URBAN, Partielle Differenzialgleichungen: Eine Einführung in analytische und numerische Methoden, Spektrum Akademischer Verlag 2010.
- [Deuflhard/Hohmann] P. DEUFLHARD, A. HOHMANN, Numerische Mathematik 1, de Gruyter-Verlag, Berlin, 1993.
- [Fischer] G. FISCHER, Lineare Algebra, Vieweg-Verlag, Wiesbaden, 1989.
- [Analysis I] S. FUNKEN, Skript zur Vorlesung „Analysis I“, gehalten im Wintersemester 07/08 an der Universität Ulm.
- [Analysis II] S. FUNKEN, Skript zur Vorlesung „Analysis II“, gehalten im Sommersemester 2008 an der Universität Ulm.
- [Discroll/Maki] T. A. DISCROLL, K. L. MAKI: Searching for Rare Growth Factors Using Multicanonical Monte Carlo Methods. SIAM Review. Vol. 49, No. 4, pp. 673-692. 2008.
- [Golub/Loan] G. H. GOLUB, C. F. VAN LOAN, Matrix Computations, 3. ed., Hopkins Univ. Press, 1996.
- [Hackbusch] W. HACKBUSCH, Iterative Lösung großer schwachbesetzter Gleichungssysteme, 2. Auflage, Teubner-Verlag, Stuttgart, 1993.
- [Hämmerlin/Hoffmann] G. HÄMMERLIN, K.-H. HOFFMANN, Numerische Mathematik, 4. Auflage, Springer-Verlag, Berlin, 1994.
- [Hanke] M. HANKE-BOURGEOIS, Grundlagen der Numerischen Mathematik und des wissenschaftlichen Rechnens, 1. Auflage, Teubner-Verlag, Stuttgart, 2002.
- [Higham] N. J. HIGHAM, How accurate is Gaussian elimination? Numerical Analysis 1989, Proceedings of the 13th Dundee Conference, Vol. 228 of Pitman Research Notes in Mathematics, 137–154.
- [Knuth] D.E. KNUTH The Art of Computer Programming, Band 2, Addison-Wesley, 3. Auflage, 1998.
- [Niethammer] W. NIETHAMMER, Relaxation bei nichtsymmetrischen Matrizen. Math. Zeitschr. **85** 319-327, 1964.
- [Plato] R. PLATO, Numerische Mathematik kompakt, Vieweg-Verlag.
- [Quarteroni et. al.] A. QUARTERONI, R. SACCO, F. SALERI, Numerische Mathematik, Band 1 & 2, Springer-Verlag, Berlin, 2002.
- [Schwarz] H. R. SCHWARZ, Numerische Mathematik, 4. Auflage, Teubner-Verlag, Stuttgart, 1997.
- [Stör/Bulirsch] J. STÖR, R. BULIRSCH, Numerische Mathematik, Band 1 & 2, Springer-Verlag, Berlin, 1994.
- [Törnig/Spellucci] W. TÖRNIG, P. SPELLUCCI, Numerische Mathematik für Ingenieure und Physiker, Band 1 & 2, Springer-Verlag, Berlin, 1988.

- [Wilkinson65] J. H. WILKINSON, The Algebraic Eigenvalue Problem, Oxford University Press, 1965.
- [Wilkinson69] J. H. WILKINSON, Rundungsfehler, Springer-Verlag, Berlin, Heidelberg, New York, 1969.
- [Wille] D. WILLE, Repetitorium der Linearen Algebra, Teil 1, 1. Auflage, Feldmann-Verlag, Springer, 1989.

# Index

- $A$ -orthogonal, 85
- Abbruchkriterien, 78
- absoluter Fehler, 27
- absoluter/relativer Fehler, 16
- Abstiegsrichtung, 80
- Äquivalenzoperationen, 37
- Arithmetik
  - Hochgenauigkeitsarithmetik, 26
  - Intervallarithmetik, 26
- Armijo, Schrittweitenregel, 81
- Auslöschung, 18
  
- $b$ -adisch, 6
- $b$ -adischer Bruch, 5
- Bandbreite, 51
- Bandgleichungen, 51
- Banker's Rule, 14
- Block-Tridiagonalmatrix, 62
- Block-Tridiagonalmatrixg, 62
- Bruch,  $b$ -adischer, 5
  
- Cauchy-Schwarz-Ungleichung, 119
- cg-Verfahren, 85, 89
- Cholesky, 47
  - Verfahren, 49, 52
  - Zerlegung, 50
  - Zerlegung, rationale, 48
- Cramersche Regel, 30
- CSU, 119
  
- dünn besetzt, **siehe** schwachbesetzt
- Darstellung, Gleitpunkt-, 9
- Darstellung, Leibnizsche, 31
- Daumenregel zur Genauigkeit, 60
- diagonaldominant, 44
- Diagonalmatrix, 34
- Drei-Term-Rekursion, 92
- Dreiecksmatrix, 34
- Dreieckszerlegung, Gaussche, 39
  
- einfache Realisierbarkeit, 64
- Einzelschrittverfahren, 64
- Eliminationsmethode, Gaußsche, 37
- Eliminationsschritt, 38
- Energienorm, 82
  
- Fehler, 27
  - absoluter, 27
  - absoluter/relativer, 16
  - relativer, 27
- Fehlerarten, 3
- Festpunktzahl], 8
- Frobenius-Norm, 120
  
- Gaußsches-Eliminations-Verfahren, 39
- Gauß-Seidel, 64
- Gaus-Elimination mit Spaltenpivotstrategie, 42
- Gaussche Dreieckszerlegung, 39
- Gaußsche Eliminationsmethode, 37
- Genauigkeit, Daumenregel zur, 60
- Gesamtschrittverfahren, 64
- Gestaffelte Systeme, 33
- Givens-Rotationen, 102
- Glättungseigenschaften, 78
- Gleitkommaarithmetik, 17
- Gleitkommaoperation, 31
- Gleitkommazahl, 3
- Gleitpunkt-Darstellung, 9
- Gradienten-Verfahren, 80
- Gradienten-Verfahren, Konjugiertes, 89
- Gradienten-Verfahren, Vorkonditionierter Konjugiertes, 94
- Gradienten-Verfahren:, 81
- gross-oh, 117
- Guard Digit, 19
  
- Hochgenauigkeitsarithmetik, 26
- Holder-Ungleichung, 119
- Householder-Spiegelungen, 107
- Householder-Transformation, 107
- Householder-Vektor, 107
  
- Interpolationspolynom, 53
- Intervallarithmetik, 26
- irreduzibel, 70
- Iterationsverfahren, 64
  
- Jacobi, 64
- Jacobi-Verfahren, 69
- Jordansche Normalform, 67
  
- Kantorowitsch-Ungleichung, 83
- kaufmännische Rundung, 14

- klein-oh, 117
- Kondition, 3, 21
- Konditionszahl, 57
- Konditionszahlen, 23
- konjugiert, 85
- Konjugiertes Gradienten–Verfahren, 89
- Konsistenz, 24
- Konvergenzeigenschaft, 64
- Konvergenzkriterium, 67
- Konvergenzrate, 64
  
- Laplace-Operator, 61
- Laplacescher Entwicklungssatz, 31
- Leibnizsche Darstellung, 31
- lexikographische Nummerierung, 62
- Linienuche, 80
- $LR$ -Zerlegung, 39
  
- Maschinengenauigkeit, 16
- mathematisch unverzerrte Rundung, 14
- MATLAB, 123
- Matrix
  - Diagonalmatrix, 34
  - Dreiecksmatrix, 34
  - Elementarmatrix, 67
  - Jordanmatrix, 67
  - positiv definite, 47
  - Rotationsmatrizen, 101
  - unipotent, 39
  - Vandermond, 53
- Matrixnorm, 120
- Methode der kleinsten Quadrate, 98
  
- Nachiteration, 47
- Neumann-Reihe, 59
- Newton–Darstellung, 53
- Normalisierung, 5
  
- Operationen
  - Aquivalenzoperationen, 37
  - Gleitkommaoperation, 31
  - Rechenoperationen, 31
- Operatornorm, 120
- orthogonal, 100
  
- $p$ -adisch, 6
- pcg-Verfahren, 94
- Permutation, 41
- Pivot-Strategien, 40
- Pivotelement, 38
- Polynominterpolation, 54
- positiv definite Matrix, 47
  
- Präkonditionierung **siehe**
  - Vorkonditionierung 94
  
- QR-Zerlegung, 100
  
- Rückwärtsanalyse, 27
- Rückwärtssubstitution, 35
- rationale Cholesky-Zerlegung, 48
- Realisierbarkeit, einfache Realisierbarkeit, 64
- Rechenoperationen, 31
- reduzibel, 70
- Regel, Cramersche, 30
- Reihe
  - Neumann-, 59
- relativer Fehler, 27
- Residuum, 47, 56, 86
- Richardson, 64
- Richardson-Verfahren, 68
- Rosenberg, 76
- Rotationsmatrizen, 101
- Round to even, 14
- Rundung, 14
  - kaufmännische, 14
  - mathematisch unverzerrte, 14
  - Standardrundung, 14
  
- Satz
  - Laplacescher Entwicklungssatz, 31
- Schrittweitenregel von Armijo, 81
- Schutzziffer, 19
- schwachbesetzt, 61
- Schwaches Zeilensummenkriterium, 71
- Spaltenpivotisierung, 44
- Spaltenpivotstrategie, 42
- Spaltenpivotstrategie, Gaus-Elimination mit, 42
- Spaltensummenkriterium, starkes , 69
- Spaltensummennorm, 121
- sparse, 61
- Spektralradius, 66
- Spektrum, 66
- Stabilität, 3, 24
- Standardfehler, 25
- Standardmatrix, 30, 61
- Starkes Spaltensummenkriterium, 69
- Starkes Zeilensummenkriterium, 69
- starkes Zeilensummenkriterium, 72
- Strategien, Pivot-, 40
- Submultiplikativität, 120
- Substitution, Vorwärts- Rückwärts-, 35
- Systeme, Gestaffelte, 33

- Tschebyscheff–Polynome, 91
- Überrelaxationsverfahren, 78
- Ungleichung, Cauchy-Schwarz-, 119
- Ungleichung, Holder-, 119
- unipotente Matrix, 39
- Unterrelaxationsverfahren, 78
- Vandermond-Matrizen, 53
- Vektornorm, 119
- Vektornorm, vertragliche, 122
- Verfahren
  - Überrelaxationsverfahren, 78
  - cg-Verfahren, 85, 89
  - Cholesky, 49
  - Cholesky-, 52
  - Einzelstufenverfahren, 64
  - Gaußsches-Eliminations-, 39
  - Gesamtstufenverfahren, 64
  - Gradienten–Verfahren, 80
  - Gradienten–Verfahren:, 81
  - Iterationsverfahren, 64
  - Jacobi, 69
  - Konjugiertes Gradienten–Verfahren, 89
  - pcg-Verfahren, 94
  - Richardson, 68
  - Unterrelaxationsverfahren, 78
  - Vorkonditionierter Konjugiertes Gradienten–Verfahren, 94
- vertragliche Vektornorm, 122
- von Stein, 76
- Vorkonditionierter Konjugiertes Gradienten–Verfahren, 94
- Vorkonditionierung, 92
- Vorwärtsanalyse, 25
- Vorwärtssubstitution, 35
- Zahlendarstellung, 3, 4
- Zahlenformat, 11
- Zeilenpivotisierung, 44
- Zeilenstufenkriterium, schwaches, 71
- Zeilenstufenkriterium, starkes, 69, 72
- Zeilenstufennorm, 121
- zerlegbar, 70
- Zerlegung, 44
- Zerlegung, Cholesky-, 50
- Zerlegung, Cholesky-, rationale, 48