Numerik I
Wintersemester 2011/12

Anita Schöbel

Contents

1 Entwurf und Analyse von Algorithmen **7**
- 1.1 Aufwandsabschätzung von Algorithmen 8
- 1.2 Fehlerabschätzung 12

2 Lineare Gleichungssysteme: Eliminationsverfahren **20**
- 2.1 Begriffe und Grundlagen 20
- 2.2 Gauß-Verfahren und LU-Zerlegung 23
- 2.3 Das Cholesky-Verfahren 37
- 2.4 Schwach besetzte Matrizen 42

3 Funktionalanalytische Grundlagen: Normierte Räume und lineare Operatoren **45**
- 3.1 Metrische und normierte Räume 45
- 3.2 Normen für Abbildungen und Matrizen 54
- 3.3 Kondition 62

4 Lineare Gleichungssysteme: Orthogonalisierungsverfahren **67**
- 4.1 Die QR-Zerlegung 67
- 4.2 Lineare Ausgleichsprobleme 78
- 4.3 Singulärwertzerlegung 84
- 4.4 Anwendung der Singulärwertzerlegung auf lineare Ausgleichsprobleme 85

5 Iterationsverfahren für Gleichungssysteme **89**
- 5.1 Das Verfahren der sukzessiven Approximation 89
- 5.2 Der Banach'sche Fixpunktsatz 93
- 5.3 Iterative Verfahren für lineare Gleichungssysteme 98
- 5.4 Iterative Verfahren für nichtlineare Gleichungssysteme 115

6 Eigenwertprobleme **125**
- 6.1 Grundlagen und Eigenschaften 125
- 6.2 Der Spezialfall hermitescher Matrizen 127
- 6.3 Lokalisierungssätze 131

6.4 Potenzmethode zur Bestimmung des Spektralradius 134
6.5 Das QR-Verfahren zur Bestimmung aller Eigenwerte 138
6.6 Das Jakobiverfahren 145

Einleitung

In der Wikipedia ist der Begriff *Numerik* folgendermaßen definiert:

> Die numerische Mathematik, kurz Numerik genannt, beschäftigt sich als Teilgebiet der Mathematik mit der Konstruktion und Analyse von Algorithmen für kontinuierliche mathematische Probleme.

Interesse an solchen Algorithmen besteht meist aus einem der beiden folgenden Gründe:

- Es gibt zu dem Problem keine explizite Lösungsdarstellung (so zum Beispiel bei den Navier-Stokes-Gleichungen oder bei Integralen ohne Stammfunktion).
- Die Lösungsdarstellung existiert zwar, ist jedoch nicht geeignet, um die Lösung schnell auszurechnen, beziehungsweise sie liegt in einer Form vor, in der Rechenfehler sich stark bemerkbar machen (zum Beispiel bei vielen Potenzreihen).

In der angewandten Mathematik setzt man dabei noch einen Schritt früher an: Beschäftigt man sich mit Anwendungen, so liegt zunächst noch kein mathematisches Problem vor, sondern einzig eine von Praktikern formulierte Problembeschreibung. Die erste Aufgabe besteht also in der *Modellierung* d.h. in der Formalisierung eines in der Natur beobachteten Phänomens oder eines ökonomischen Problems durch ein sogenanntes *mathematisches Modell*. Mathematische Modelle sind Systeme von Gleichungen oder Ungleichungen, durch die Beziehungen zwischen bekannten und unbekannten Größen dargestellt werden. Dabei können algebraische Gleichungen (oder Ungleichungen), Differentialgleichungen oder Integralgleichungen verwendet werden und es dürfen dabei alle Arten von Formeln oder Grenzwerten auftreten. Gibt es zusätzlich noch ein Kriterium zur Beurteilung der Lösung (eine *Zielfunktion*), so liegt ein Problem der mathematischen Optimierung vor und das Modell wird auch als *mathematisches Programm* bezeichnet.

Klassische Disziplinen der "reinen" Mathematik wie die Algebra und Analysis beschäftigen sich mit Fragen der Existenz und Eindeutigkeit der Lösungen mathematischer Modelle. Dagegen besteht das Ziel der numerischen Mathematik darin,

Verfahren zu entwickeln, mit denen sich die Lösungen mathematischer Modelle praktisch (auf derzeit verfügbaren Rechenanlagen) ermitteln lassen. Typische Beispiele dafür sind:

- Der Fundamentalsatz der Algebra besagt, dass ein reelles Polynom vom Grad n auch n Nullstellen in der Menge der komplexen Zahlen besitzt. Der Existenzbeweis ist jedoch nicht konstruktiv, d.h. man erhält kein Verfahren, wie man die entsprechenden Nullstellen bestimmen kann. Das liefert die numerische Mathematik.

- Die Lösung linearer, nicht singulärer Gleichungssysteme kann durch die Cramersche Regel aufgeschrieben werden. Für die praktische Berechnung ist sie aber bei mehr als drei Variablen unbrauchbar.

- Der Satz von Weierstrass liefert die Aussage, dass stetige Funktionen auf kompakten Mengen ihre Minima (oder Maxima) annehmen. Wie aber soll man sie berechnen?

- Für Anfangswertprobleme einer gewöhnlichen Differentialgleichung liefert der Existenzbeweis von Picard-Lindelöf unter bestimmten Glattheitsvoraussetzungen ein konstruktives Iterationsverfahren. Bei der Realisierung auf Computern ist dieses Verfahren aber nicht sonderlich effektiv.

In der vorliegenden Vorlesung Numerik I sollen Verfahren für die Berechnung mathematischer Modelle vorgestellt und diskutiert werden. Dabei möchte und kann die Vorlesung keine Rezeptsammlung sein, die für jedes Problem der numerischen Mathematik ein fertiges Lösungsverfahren liefert. Vielmehr soll die Methodik der numerischen Mathematik anhand ausgewählter Problemstellungenen deutlich werden. Um effiziente numerische Verfahren zu entwickeln, die schlussendlich auch geeignet sind, praktische Probleme zu lösen, ist in der angewandten Mathematik genau wie in der "reinen" Mathematik eine saubere Theorie vonnöten. Diese basiert für viele weiterführende Bereiche auf der Funktionalanalysis, die in dieser Vorlesung daher (aus den Grundvorlesungen wiederholt) und fortgesetzt werden soll. (Lineare) Funktionalanalysis beschäftigt sich mit der Untersuchung von (linearen) Abbildungen zwischen Räumen mit topologischer Struktur. Hier werden wir uns vor allem mit der Lösung von Gleichungssystemen in normierten Räumen beschäftigen. Konkret werden in dieser Vorlesung Numerik I Lösungsverfahren für die folgenden Themen behandelt:

- Lineare Gleichungssysteme

- Ausgleichsprobleme

- Nullstellensuche (eine nichtlineare Gleichung oder ein System nichtlinearer Gleichungen)

- Eigenwertprobleme

Alle diese Probleme verlangen das Lösen eines Gleichungssystems und legen die Grundlagen für die meisten Veranstaltungen in der Angewandten Mathematik. So wird die Lösung von linearen Gleichungssystemen beispielsweise bei der Polynominterpolation benötigt und es werden die in der Vorlesung Numerik II zu besprechenden Integralgleichungen und Differentialgleichungen wieder als Gleichungssysteme (in Funktionenräumen) formuliert. In der Optimierung werden ebenfalls Gleichungssysteme behandelt, die in der Regel aber unterbestimmt sind, so dass viele Lösungen existieren, unter denen dann anhand eines Zielfunktionals eine möglichst gute ausgewählt werden soll.

Die klassische Numerik beschäftigt sich vorrangig mit kontinuierlichen Problemen. Dagegen sind *diskrete Probleme* durch eine endliche Menge an möglichen Lösungen gekennzeichnet. Auch sie sind ein wichtiger Bestandteil der angewandten Mathematik, insbesondere bei ökonomischen Fragestellungen. In diesem Skript werden wir auf diskrete Probleme aber nicht eingehen; bei Interesse an der Modellierung von Fragestellungen als diskrete Probleme kann die diesjährige Veranstaltung "Algorithmische Ergänzung zur Numerik I" besucht werden, in der Knobelaufgaben mit Hilfe von ganzzahligen Variablen modelliert und bearbeitet werden sollen.

Die Vorlesung "Numerik I" richtet sich an Studierende der Mathematik ab dem dritten Semester, und gerne auch an interessierte Physik oder Informatik-Studierende. Die in der Vorlesung verwendeten Grundlagen sind so ausgewählt, dass sie auch aus der “Mathematik für Informatik-Anfänger” Vorlesung bekannt sein sollten. Die Vorlesung bietet den Einstieg in den Bereich numerische Mathematik, wissenschaftliches Rechnen und Optimierung und ist als Grundlage der meisten Vorlesungen aus diesem Bereich zu verstehen. Die Vorlesung kann mit *Numerik II* oder mit *Optimierung* (oder mit beiden Vorlesungen!) fortgesetzt werden.

Um die numerische Mathematik richtig zu verstehen, sollte man natürlich auch einiges selbst programmiert und implementiert haben. Es werden in den Übungen daher theoretische und praktische Aufgaben gestellt, wobei die Programmieraufgaben mit MATLAB zu bearbeiten sind. MATLAB ist eine Skriptsprache, in der viele numerische Verfahren, Operationen und die entsprechenden Datenstrukturen bereits zur Verfügung stehen. Auf andere Schwierigkeiten trifft man, wenn man ein Verfahren in einer "klassischen" Programmiersprache von Grund auf neu implementieren möchte — es sei jedem empfohlen, das zumindest an einem der besprochenen Verfahren einmal auszuprobieren.

Es gibt zahlreiche Lehrbücher und Skripten über numerische Mathematik, von denen die folgenden Quellen erwähnt werden sollen:

- J. Stoer, Numerische Mathematik I, Springer, 1989.

- J. Werner. Numerische Mathematik 1, Vieweg, Braunschweig, 1992.
- P. Deuflhard und A. Hohmann. Numerische Mathematik I. Walter de Gruyter, Berlin, New York, 2nd edition, 1993.
- R. Kreß. Numerical Analysis. Springer, New York, 1998.
- M. Hanke-Bourgeois. Grundlagen der Numerischen Mathematik und des wissenschaftlichen Rechnens. Teubner, Stuttgart, 2002.
- R. Schaback, H. Wendland. Numerische Mathematik. Springer, Berlin, 2004.

Chapter 1

Entwurf und Analyse von Algorithmen

Ein Algorithmus für ein Problem ist ein durch eine Abfolge von (Rechen-)Vorschriften beschriebenes Verfahren, das zu einer "Lösung" des Problems führt. Je nach Qualität der erzielten Lösung, unterscheidet man zwischen exakten Verfahren, konstruktiven Verfahren und Heuristiken.

Exakte Verfahren sind streng genommen nur bei diskreten mathematischen Aufgaben möglich, in denen unter endlich vielen Möglichkeiten eine Lösung auszuwählen ist. Dazu gehört beispielsweise das Gebiet der ganzzahligen Programmierung sowie die meisten Aufgaben in Netzwerken. Dagegen ist bei kontinuierlichen Problemen aufgrund der beschränkten Genauigkeit der Darstellung reeller Zahlen durch den Computer jede Lösung eine Näherungslösung.
Ein **konstruktives oder direktes Verfahren** ist eine Rechenvorschrift, mit deren Hilfe die numerische Lösung einer mathematischen Aufgabe in endlich vielen Rechenschritten beliebig genau ermittelt werden kann.
Lässt sich gar keine Genauigkeit angeben oder sind dieser Genauigkeit Grenzen gesetzt, so spricht man von einer **Heuristik**. Kann wie im letzteren Fall zwar eine Genauigkeit angegeben werden, diese ist aber beschränkt, so hat die Heuristik eine *Gütegarantie*; man spricht dann auch von einer *Approximation*. Heuristiken werden vor allem bei sehr schweren Problemen der diskreten Optimierung verwendet und führen dort häufig zu empirisch sinnvollen Ergebnissen.

Hat man einen Algorithmus entwickelt, so interessieren die folgenden Fragestellungen:

- Aufwand
- Fehleranalyse
- Stabilität.

1.1 Aufwandsabschätzung von Algorithmen

Ein Verfahren kann nur dann sinnvoll eingesetzt werden, wenn es auch in praktikabler Zeit eine Lösung ermittelt. Daher ist es wichtig, den Aufwand verschiedener Verfahren für die gleiche Aufgabenstellung vergleichend zu diskutieren. Man spricht auch von der **Komplexität** eines Verfahrens und bezeichnet damit den Aufwand an wesentlichen Rechenoperationen in Abhängigkeit einer sinnvoll gewählten Eingangsgröße.

Als wesentliche Rechenoperationen zählen wir Additionen, Subtraktionen, Multiplikationen, Divisionen, Vergleiche und davon getrennt Funktionsauswertungen. (Zuweisungen werden nicht gezählt.)

Die Eingangsgröße kann über Turing-Maschinen in der theoretischen Informatik sauber definiert werden. Sie repräsentiert die Größe des Problems, gegeben als die Länge der zur Beschreibung des Problems nötigen Daten. Meistens wird sie vereinfacht durch einzelne Werte der Problembeschreibung repräsentiert.

Beispiele:

- Bestimme das Minimum von n Zahlen $x_1, \ldots, x_n$: Hier bestimmt man die Anzahl der Rechenoperationen in Abhängigkeit der Zahl n. Für den einfachen Algorithmus

  Min $:= \infty$; For $i := 1$ to n do: If $x_i <$ Min then Min $:= x_i$; Output: Min

  ergeben sich n Vergleiche, also eine Anzahl von $A_1(n) = n$ wesentlichen Rechenoperationen. Der Aufwand des Verfahrens ist also linear in n.

- Bei der Addition von zwei Vektoren x und y der Dimension k bietet sich als sinnvolle Eingangsgröße die Dimension k an. Das Verfahren

  For $i := 1$ to k do: $z_i := x_i + y_i$; Output: $z_1, \ldots, z_n$

  benötigt k Additionen, hat also ebenfalls einen linearen Aufwand von $A_2(k) = k$.

- Addition von zwei Matrizen A, B der Dimension $k \times k$ (mit Elementen $a_{ij}, b_{ij}, i, j = 1, \ldots, k$): Hier kann man als Eingabegröße k oder k^2 wählen. Das kanonische Verfahren ist das folgende.

  For $i := 1$ to k do:

  For $j = 1$ to k do: $c_{ij} := a_{ij} + b_{ij}$;

  Output: c_{ij}, $i, j = 1, \ldots, n$

Die Anzahl der wesentlichen Rechenoperationen beträgt k^2. Normalerweise wird man den Aufwand in Abhängigkeit der Anzahl der Matrixelemente $m = k^2$ mit $A_3(m) = m$ als linear angeben. In vielen Anwendungen mit quadratischen Matrizen macht es aber Sinn, die Dimension k als Eingabegröße zu wählen, was zu einem quadratischen Aufwand $A_4(k) = k^2$ führt.

Bei komplizierteren Problemen sind meist verschiedene Verfahren mit jeweils unterschiedlichem Aufwand möglich. Hat man also z.B. einen Algorithmus A und einen Algorithmus B zur Auswahl, und ist der Aufwand beider Verfahren durch Funktionen $A(n)$ beziehungsweise $B(n)$ bekannt so wird man für die Problemgröße n das Verfahren mit dem jeweils kleineren Aufwand wählen.

Um für steigende Problemgrößen den Aufwand von zwei Verfahren auf einfache Methode zu vergleichen, bieten sich die *Landau-Symbole* an. Diese sind nicht nur in der Analyse von Aufwandsabschätzungen sondern allgemeiner zur quantitativen Beschreibung von Grenzprozessen ein wichtiges Hilfsmittel.

Die Landau-Symbole geben dabei an, wie sich die Größe von zwei Funktionen $a(n), b(n) : \mathbb{N} \to \mathbb{R}$ im Verhältnis zueinander entwickelt, wenn $n \to \infty$ geht.

Definition 1.1 *Seien $(a_n), (b_n)$ reelle Zahlenfolgen. Die* **Landau-Symbole** *sind wie folgt definiert.*

1. $a_n = \mathrm{O}(b_n)$ *falls es ein $C \in \mathbb{R}, C > 0$ und ein $N \in \mathbb{N}$ gibt mit*
$$|a_n| \leq C|b_n| \text{ für alle } n \geq N.$$
2. $a_n = \mathrm{o}(b_n)$ *falls es zu jedem $\varepsilon > 0$ ein $N \in \mathbb{N}$ gibt mit*
$$|a_n| \leq \varepsilon|b_n| \text{ für alle } n \geq N.$$
3. $a_n = \Theta(b_n)$ *falls $a_n = \mathrm{O}(b_n)$ und $b_n = \mathrm{O}(a_n)$ und*

Um die Bedeutung dieser Symbole zu verdeutlichen, formulieren wir die Aussagen um, indem wir die Entwicklung des Quotienten $\frac{a_n}{b_n}$ betrachten, um die Wachstumsraten der beiden Folgen zu vergleichen. Nehmen wir dazu an, dass $b_n = 0$ für höchstens endlich viele Folgenglieder ist. Dann erhält man:

$$\begin{aligned}
a_n = \mathrm{O}(b_n) &\iff \left|\frac{a_n}{b_n}\right| \leq C \text{ für alle } n \geq N \text{ und ein } C > 0.\\
a_n = \mathrm{o}(b_n) &\iff \left|\frac{a_n}{b_n}\right| \to 0\\
a_n = \Theta(b_n) &\iff C_1 \leq \left|\frac{a_n}{b_n}\right| \leq C_2 \text{ für alle } n \geq N \text{ und } C_1, C_2 > 0.
\end{aligned}$$

Die Bedeutung der Landau-Symbole kann hier nun abgelesen werden: Ist $a_n = \mathrm{O}(b_n)$ so wächst a_n nicht schneller als b_n, ist $a_n = \Theta(b_n)$, dann wachsen beide Folgen annähernd gleich schnell und bei $a_n = \mathrm{o}(b_n)$ wächst b_n viel schneller als a_n.

Einfache Beispiele:

$$\begin{aligned} n^2 &= \mathrm{O}(n^3) \\ n^2 &= \mathrm{O}(\tfrac{1}{1000}n^3) \\ n^2 &= \mathrm{O}(\tfrac{1}{3}n^2) \\ \tfrac{1}{3}n^2 &= \mathrm{O}(n^2) \\ n^2 &= \Theta(\tfrac{1}{3}n^2) \\ n^2 &= \mathrm{o}(\tfrac{1}{1000}n^3) \\ n^2 &\neq \mathrm{o}(\tfrac{1}{3}n^2) \end{aligned}$$

Lemma 1.2 *Die folgenden Aussagen gelten:*

1. *Alle drei Begriffe sind transitiv, d.h.*

   $$a_n = \mathrm{O}(b_n), b_n = \mathrm{O}(c_n) \Longrightarrow a_n = \mathrm{O}(c_n),$$

   analog für o *und* Θ.

2. Θ *ist eine Äquivalenzrelation*

3. $a_n = \mathrm{o}(b_n) \Longrightarrow a_n = \mathrm{O}(b_n)$ *und* $b_n \neq \mathrm{O}(a_n)$

Beweis: Lässt sich leicht nachrechnen, Übungen!

Weiterhin sollte man sich klar machen, dass

$$\begin{aligned} a_n = \mathrm{O}(b_n) &\iff a_n = \mathrm{O}(\alpha b_n) \text{ für alle } \alpha \in \mathbb{R} \setminus \{0\} \\ a_n = \mathrm{O}(b_n) \text{ und } a'_n = \mathrm{O}(b_n) &\Longrightarrow a_n + a'_n = \mathrm{O}(b_n). \end{aligned}$$

Diese Aussagen gelten ebenfalls für o, Θ.

Von großer praktischer Bedeutung sind die folgenden Aussagen:

- Logarithmisches Wachstum langsamer ist als polynomiales, in Formeln:

  $$(\log_\beta(n))^\gamma = \mathrm{o}(n^\alpha) \text{ für alle } \alpha > 0, \beta > 1, \gamma > 0.$$

- Polynomiales Wachstum schwächer ist als exponentielles,

  $$n^\alpha = \mathrm{o}(\beta^n) \text{ für alle } \alpha > 0, \beta > 1.$$

- Exponentielles Wachstum schwächer ist als fakultatives,

$$\beta^n = \mathrm{o}(n!) \text{ für alle } \beta > 1.$$

Diese Aussagen lassen sich nun auf die Analyse von Algorithmen anwenden. Dazu betrachten wir die folgenden beiden schematischen Algorithmen-Bruchstücke:

- Algorithmus 1:

  Schritt 1: Führe Verfahren A aus

  Schritt 2: Führe Verfahren B aus

Hat Verfahren A einen Aufwand von $\mathrm{O}(a_n)$ und Verfahren B einen Aufwand von $\mathrm{O}(b_n)$, und gilt $a_n = \mathrm{O}(b_n)$, so ergibt sich für Algorithmus 1 ein Aufwand von $\mathrm{O}(b_n)$, das heißt, bei der Hintereinanderausführung von Algorithmenteilen ist immer der größere Aufwand maßgebend.

- Algorithmus 2:

  Schritt 1: Für $m = 1, \ldots, M$ führe Verfahren A aus

Hat Verfahren A einen Aufwand von $\mathrm{O}(a_n)$, und lässt sich die Größe der Zahl M in Abhängigkeit von der Eingabegröße n durch $M = \mathrm{O}(c_n)$ abschätzen, so ergibt sich für Algorithmus 2 ein Aufwand von $\mathrm{O}(a_n \cdot c_n)$.

Abschließend erweitern wir Definition 1.1 auf beliebige Funktionen. Vor allem O und o werden in dieser Formulierung auch häufig für die Abschätzung von Restgliedern verwendet.

Definition 1.3 *Seien $f, g : \mathbb{R} \to \mathbb{R}$. Dann definiert man*

- *$f = \mathrm{O}(g)$ für $x \to x_0$ falls $f(x_n) = \mathrm{O}(g(x_n))$ für jede Folge $x_n \to x_0$.*
- *Analog für* o, Θ.

Es lässt sich leicht zeigen, dass obige Definition äquivalent ist zu den folgenden Aussagen:

- $f = \mathrm{O}(g)$ falls es eine Zahl $C > 0$ und eine Umgebung $U = U(x_0)$ von x_0 gibt, so dass $|f(x)| \leq C|g(x)|$ für alle $x \in U$.
- $f = \mathrm{o}(g)$ falls es zu jedem $\varepsilon > 0$ eine Umgebung $U = U(x_0)$ von x_0 gibt, so dass $|f(x)| \leq \varepsilon|g(x)|$ für alle $x \in U$.
- $f = \Theta(g)$ falls es Zahlen $C_1, C_2 > 0$ und eine Umgebung $U = U(x_0)$ von x_0 gibt, so dass $C_1|g(x)| \leq |f(x)| \leq C_2|g(x)|$ für alle $x \in U$.

Übung: *Sei $f(x) = 1+x+x^2+x^3+x^4+\ldots$. Zeigen Sie, dass $f(x) = 1+x+O(x^2)$ für $x \to 0$. Gilt auch $f(x) = 1 + x + o(x^2)$ für $x \to 0$?*

1.2 Fehlerabschätzung

Hat man ein Verfahren zur Lösung eines mathematischen Problems entwickelt, so sollte man den Fehler in der Lösung abschätzen. Um zu formalisieren, was man unter *Fehler* versteht, betrachten wir ein mathematisches Problem mit Eingangsdaten ξ. Die exakte Lösung des Problems bei Vorliegen der Eingangsdaten ξ sei $x = f(\xi)$. Für das Problem liege weiterhin ein Algorithmus Alg vor, der bei Vorliegen der Eingangsdaten ξ die Lösung $\mathrm{Alg}(\xi)$ ausgibt. Wir unterscheiden dann die folgenden Fehler.

Definition 1.4 *Sei $\tilde{x} = \mathrm{Alg}(\tilde{\xi})$ die von einem Verfahren* Alg *bei (fehlerhaften) Eingangsdaten $\tilde{\xi}$ ermittelte Lösung. Sei $x = f(\xi)$ die exakte Lösung des Problems mit exakten Eingangsdaten ξ. Sei $\|\cdot\|$ eine Norm auf dem Raum der Eingangsdaten und auf dem Raum der Lösungen.*

- $\|\tilde{x} - x\|$ *bezeichnet den* **absoluten Fehler** *der Lösung. Im Fall $x \neq 0$ heißt $\frac{\|\tilde{x}-x\|}{\|x\|}$ der* **relative Fehler** *der Lösung.*

- *Der* **absolute Verfahrensfehler** *eines Verfahrens* Alg *ist $\|\mathrm{Alg}(\xi) - f(\xi)\|$, der* **relative Verfahrensfehler** *für $f(\xi) \neq 0$ ist $\frac{\|\mathrm{Alg}(\xi)-f(\xi)\|}{\|f(\xi)\|}$.*

  Das Verfahren Alg *nennt man* **K-Approximation**, *wenn es für alle möglichen Instanzen (d.h. für alle möglichen Eingangsdaten) ξ eine Lösung ermittelt, deren relativer Fehler maximal K ist, wenn also für alle ξ mit $f(x_i) \neq 0$ gilt:*

$$\frac{\|\mathrm{Alg}(\xi) - f(\xi)\|}{\|f(\xi)\|} \leq K.$$

- *Der durch fehlerhafte Eingangsdaten übertragene absolute Fehler ist $\|f(\tilde{\xi}) - f(\xi)\|$, der durch fehlerhafte Eingangsdaten übertragene relative Fehler ist $\frac{\|f(\tilde{\xi})-f(\xi)\|}{\|f(\xi)\|}$ (für $f(\xi) \neq 0$).*

Verfahrensfehler: Konsistenz

Wir beschäftigen uns zunächst mit Verfahrensfehlern. Diese werden von dem vorliegenden Verfahren an sich verursacht, selbst wenn man exakte Eingangsdaten hat und exakt rechnen kann. Bei der Analyse von Verfahrensfehlern geht man davon aus, dass sich das Verfahren Alg_h durch einen Parameter h steuern lässt. Solche Parameter können z.B. die Abbruchbedingung bezeichnen oder Diskretisierungsparameter sein, die die Schrittweite bestimmen. Wir geben für beide Typen ein Beispiel:

Wir schauen uns zunächst die Abschätzung des Abbruchfehlers am Beispiel der Berechnung der Exponentialfunktion $\exp(\xi) = \sum_{j=0}^{\infty} \frac{\xi^j}{j!}$ an. Ein mögliches Verfahren zur Berechnung von $\exp(\xi)$ für $\xi \in \mathbb{R}$ besteht in der Auswerten der n-ten

Partialsumme

$$P_n(\xi) = \sum_{j=0}^{n} \frac{\xi^j}{j!}.$$

(In den Bezeichnungen von Definition 1.4 ist $\mathrm{Alg}_n(\xi) = P_n(\xi)$ und $f(\xi) = \exp(\xi)$.)
Wir nehmen an, dass $n \geq |\xi|$ gewählt wurde.
Für $\xi < 0$ erhält man den absoluten Abbruchfehler

$$\begin{aligned} |\exp(\xi) - P_n(\xi)| &= \left| \sum_{j=n+1}^{\infty} \frac{\xi^j}{j!} \right| \\ &\leq \frac{|\xi|^{n+1}}{(n+1)!} \underbrace{-\frac{|\xi|^{n+2}}{(n+2)!} + \frac{|\xi|^{n+3}}{(n+3)!}}_{\leq 0} \underbrace{-\frac{|\xi|^{n+4}}{(n+4)!} + \frac{|\xi|^{n+5}}{(n+5)!}}_{\leq 0} \cdots \\ &\leq \frac{|\xi|^{n+1}}{(n+1)!} \end{aligned}$$

und für $\xi \geq 0$ kann man

$$|\exp(\xi) - P_n(\xi)| = \sum_{j=n+1}^{\infty} \frac{\xi^j}{j!} \leq \sum_{j=0}^{\infty} \frac{\xi^j}{j!} \cdot \frac{\xi^{n+1}}{(n+1)!} = \exp(\xi) \frac{|\xi|^{n+1}}{(n+1)!}$$

abschätzen. Das Verfahren, $\exp(\xi)$ durch Auswertung der n-ten Partialsumme zu bestimmen, ist also für positive reelle Zahlen ξ eine $K = \frac{|\xi|^{n+1}}{(n+1)!}$-Approximation. Weil $|\xi|^{n+1} = \mathrm{o}\left((n+1)!\right)$, kann man die gewünschte Zahl $\exp(\xi)$ beliebig genau approximieren, indem man n wachsen lässt.

Der zweite Typ von Verfahrensfehler betrifft den Diskretisierungsfehler und führt zum Begriff der **Konsistenz** des vorliegenden Verfahrens.

Definition 1.5 *Sei* Alg_h *ein von einem Diskretisierungsparameter* h *abhängiges Verfahren. Bezeichne* $\mathrm{Alg}_h(\xi)$ *die von dem Verfahren ermittelte Lösung und* $f(\xi)$ *die exakte Lösung des Problems mit denselben Eingangsdaten* ξ*. Das Verfahren* Alg_h *heißt* **konsistent** *falls für alle Eingangsdaten* ξ

$$\|f(\xi) - \mathrm{Alg}_h(\xi)\| \to 0 \text{ für } h \to 0.$$

Das Verfahren hat die **Konsistenzordnung** p *falls für alle* ξ

$$\|f(\xi) - \mathrm{Alg}_h(\xi)\| = \mathrm{O}(h^p).$$

Als Beispiel betrachten wir die numerische Berechnung der Ableitung einer (differenzierbaren) Funktion $g(x)$ an einem gegebenen Punkt $\xi = x_0$. Dazu können wir folgende Verfahren wählen:

$$\begin{aligned} &\text{einfacher Differenzenquotient:} && \mathrm{Alg}_h^1(\xi) = \frac{g(\xi+h) - g(\xi)}{h} \\ &\text{zentraler Differenzenquotient:} && \mathrm{Alg}_h^2(\xi) = \frac{g(\xi+h) - g(\xi-h)}{2h} \end{aligned}$$

Die Taylorentwicklung

$$g(\xi + h) = g(\xi) + hg'(\xi) + \frac{h^2 g''(\eta)}{2}$$

mit einer Zwischenstelle $\eta \in [\xi, \xi + h]$ liefert für den einfachen Differenzenquotienten

$$\left|\mathrm{Alg}_h^1(\xi) - g'(\xi)\right| = h \left|\frac{g''(\eta)}{2}\right| = O(h).$$

Dagegen ergibt sich unter Zuhilfenahme der Taylorentwicklung

$$g(\xi + h) = g(\xi) + hg'(\xi) + \frac{h^2 g''(\xi)}{2} + \frac{h^3 g'''(\eta)}{6}$$

mit $\eta \in [\xi, \xi + h]$ für den zentralen Differenzenquotienten

$$\left|\mathrm{Alg}_h^2(\xi) - g'(\xi)\right| = h^2 \left|\frac{g'''(\eta_1) + g'''(\eta_2)}{12}\right| = O(h^2)$$

In beiden Fällen geht der Fehler für $h \to 0$ gegen Null, aber der zentrale Differenzenquotient hat eine höhere Konsistenzordnung als der einfache Differenzenquotient.

Eingangsfehler

Ein wichtiger Bestandteil der Fehleranalyse ist es, zu untersuchen, wie sich Fehler in den Eingangsdaten auf das vorliegende Verfahren auswirken. Eingangsfehler beziehen sich auf die Qualität der Eingabedaten, die aufgrund von Messfehlern oder aufgrund statistischer Schwankungen ungenau vorliegen können. Oft sind Eingabedaten auch nicht hinreichend bekannt und man ist auf Schätzwerte (z.B. über das Wetter oder das Kundenverhalten) angewiesen. Eingangsfehler liegen insbesondere vor, wenn die Eingangsdaten durch die jeweilige Maschinengenauigkeit bedingt bei der Eingabe gerundet werden müssen. Bevor wir hier die Auswirkung von Eingangsfehlern auf das Verfahren untersuchen, betrachten wir die Größe von Rundungsfehlern, um zu sehen, mit welcher Abweichung der Eingangsdaten wir zumindest rechnen müssen.
In Rechnern verwendet man in der Regel die Darstellung durch Gleitkommazahlen mit einer Basis von $B = 2$ und eine Stellenzahl von $m = 52$, um mit einem weiteren Vorzeichenbit und 11 Bits für die Exponentendarstellung mit insgesamt 64 Bits pro Zahl aus zukommen. Gleitkommazahlen garantieren eine feste relative Genauigkeit der Zahldarstellung.

Satz 1.6 *Für die nach m Stellen abgeschnittene B-adische Darstellung $rd_m(x)$ von x gilt das* **Rundungsgesetz**

$$|x - rd_m(x)| \leq |x| eps$$

mit $eps = B^{1-m}$, d.h. der relative Fehler ist kleiner als $\frac{1}{B^{m-1}}$.

Heutige Rechner stellen also alle reelle Zahlen mit einem maximalen relativen Fehler von $eps = 2^{-51} \approx 4.4409 \cdot 10^{-16}$ dar, d.h. die 16te Dezimalstelle ist bis auf 5 Einheiten genau. Erfreulicherweise ist auf den meisten Rechenanlagen gewährleistet, dass auch alle Einzeloperationen ($+.-, \cdot, /$ aber auch $\sin, \sqrt{}, \exp \ldots$) auf Gleitkommazahlen mit einem maximalen relativen Fehler $eps = 2^{-51}$ ausgeführt werden. Deshalb bezeichnet man *eps* auch als **Maschinengenauigkeit**. Dennoch können sich die entstehenden Rundungsfehler im Verlauf eines Verfahrens vergrößern.

Fortpflanzung von Eingangsfehlern: Kondition und Stabilität

Eingangsfehler und bei der Eingabe entstehende Rundungsfehler sind zunächst unabhängig von der gewählten Rechenmethode, aber man muss besonders für konstruktive Verfahren abschätzen, wie sich solche Fehler im Verlauf des Verfahrens weiterentwickeln. Dabei werden schon vorhandene Fehler in jedem Schritt des Verfahrens übertragen. (Rundungsfehler können zusätzlich bei jeder numerischen Operation neu entstehen.)
Das Ziel ist, abzuschätzen, wie schlimm sich Eingangsfehler auf die Qualität des Ergebnisses auswirken. Dazu sei $x = f(\xi)$ die exakte Lösung eines Problems in Abhängigkeit einer Eingangsgröße y und es sei Alg der numerische Algorithmus, sowie $\tilde{\xi}$ die gestörten Eingangsdaten. Nach Definition 1.4 interessieren wir uns für den folgenden (absoluten) Fehler:

$$\|\mathrm{Alg}(\tilde{\xi}) - f(\xi)\|.$$

Mit der Dreiecksungleichung gilt:

$$\begin{aligned} \|\mathrm{Alg}(\tilde{\xi}) - f(\xi)\| &= \|\mathrm{Alg}(\tilde{\xi}) - f(\tilde{\xi}) + f(\tilde{\xi}) - f(\xi)\| \\ &\leq \underbrace{\|\mathrm{Alg}(\tilde{\xi}) - f(\tilde{\xi})\|}_{\text{Stabilität}} + \underbrace{\|f(\tilde{\xi}) - f(\xi)\|}_{\text{Kondition}}. \end{aligned}$$

Hierbei sagt der erste Fehler-Term aus, wie gut sich das Verfahren $\mathrm{Alg}(\tilde{\xi})$ im Vergleich mit der exakten Lösung $f(\tilde{\xi})$ des Problems bei gestörten Eingangsdaten $\tilde{\xi}$ verhält. Dieser Term ist klein, wenn das Verfahren **stabil** ist. Der zweite Term hängt dagegen nicht von dem Verfahren ab, sondern ausschließlich von dem Problem. Er ist klein, wenn das Problem **gut konditioniert** ist. Die Stabilität ist also eine Eigenschaft des Algorithmus und die Kondition eine Eigenschaft des Problems.

Ein *gut konditioniertes Problem* liegt vor, wenn kleine Änderungen der Eingangsdaten auch nur kleine Änderungen der berechneten Lösung bewirken. Das Problem wird dann auch *robust* genannt. Bei schlecht konditionierten Problemen muss man spezielle Verfahren entwickeln, um gute Lösungen zu erhalten.

Notation 1.7 *Sei $f(\xi)$ die exakte Lösung eines Problems und $f(\tilde{\xi})$ die sich aus fehlerhaften Eingangsdaten ergebende Lösung desselben Problems. Die* **Kondition** *des Problems ist der im ungünstigsten Fall auftretende Vergrößerungsfaktor für den Einfluss von relativen Eingangsfehlern auf relative Ergebnisfehler, d.h. gibt es ein $C > 0$ so dass*

$$\frac{\|f(\tilde{\xi})-f(\xi)\|}{\|f(\xi)\|} \leq C \frac{\|\xi-\tilde{\xi}\|}{\|\xi\|}$$

für alle Eingangsdaten $\tilde{\xi}, \xi$, so hat das Problem die Kondition C.
Ist die Kondition eines Problems groß, so spricht man von einem **schlecht konditionierten** *Problem.*

Im folgenden analysieren wir die Kondition der Operationen $+, -, \cdot, /$, das heißt, wie sich relative Fehler (die z.B. durch Rundung entstanden sein können und dann wie in Satz 1.6 ausgeführt eine relative Größe von bis zu *eps* haben können) durch $+, -, \cdot, /$ fortpflanzen, wenn diese exakt ausgeführt werden.

Lemma 1.8 *Seien $x, y \in \mathbb{R} \setminus \{0\}$ Eingangsdaten mit relativen Fehlern*

$$\varepsilon_x = \left|\frac{\tilde{x}-x}{x}\right|, \quad \varepsilon_y = \left|\frac{\tilde{y}-y}{y}\right|.$$

Für den relativen Fehler bei der Addition gilt

$$\left|\frac{\tilde{x}+\tilde{y}-(x+y)}{x+y}\right| \leq \varepsilon_x \left|\frac{x}{x+y}\right| + \varepsilon_y \left|\frac{y}{x+y}\right|.$$

Beweis: Nachrechnen zeigt, dass

$$\left|\frac{\tilde{x}+\tilde{y}-(x+y)}{x+y}\right| \leq \frac{|\tilde{x}-x|+|\tilde{y}-y|}{|x+y|} = \frac{|x|\varepsilon_x + |y|\varepsilon_y}{|x+y|} = \varepsilon_x \left|\frac{x}{x+y}\right| + \varepsilon_y \left|\frac{y}{x+y}\right|.$$

QED

Haben x und y das gleiche Vorzeichen, so ergibt sich also bei der Addition der beiden Zahlen ein relativer Fehler von höchstens $\varepsilon_x + \varepsilon_y$, die Kondition der Addition ist (in der $\|\cdot\|_1$-Norm) also 1.
Dagegen kann der relative Fehler bei der Subtraktion von zwei Zahlen x, y gleichen Vorzeichens (also der Addition von x und $-y$) den möglicherweise sehr großen Wert

$$\varepsilon_x \left|\frac{x}{x-y}\right| + \varepsilon_y \left|\frac{y}{x-y}\right|$$

erreichen.

Lemma 1.9 *Seien $x, y \in \mathbb{R} \setminus \{0\}$ mit relativen Fehlern*

$$\varepsilon_x = \left|\frac{\tilde{x} - x}{x}\right|, \quad \varepsilon_y = \left|\frac{\tilde{y} - y}{y}\right|.$$

Unter Vernachlässigung von Produkten von Fehlern lässt sich der relative Fehler bei der Multiplikation abschätzen durch

$$\left|\frac{\tilde{x}\tilde{y} - xy}{xy}\right| \leq \varepsilon_x + \varepsilon_y$$

und der relative Fehler bei der Division durch

$$\left|\frac{\frac{\tilde{x}}{\tilde{y}} - \frac{x}{y}}{\frac{x}{y}}\right| \leq \varepsilon_x + \varepsilon_y.$$

Beweis: Auch hier rechnen wir nach:

$$\begin{aligned}\left|\frac{\tilde{x}\tilde{y} - xy}{xy}\right| &= \left|\frac{(\tilde{x} - x)\tilde{y} + x(\tilde{y} - y)}{xy}\right| \leq \frac{|\tilde{x} - x||\tilde{y}| + |x||\tilde{y} - y|}{|xy|} \\ &= \varepsilon_x \left|\frac{\tilde{y}}{y}\right| + \varepsilon_y \leq \varepsilon_x\varepsilon_y + \varepsilon_x + \varepsilon_y,\end{aligned}$$

wobei im letzten Schritt ausgenutzt wurde, dass

$$\left|\frac{\tilde{y}}{y}\right| \leq \frac{|\tilde{y} - y|}{|y|} + \frac{|y|}{|y|} = \varepsilon_y + 1.$$

Vernachlässigen wir nun Produkte von Fehlern, erhalten wir das gewünschte Ergebnis.
Es fehlt noch die Fehlerübertragung bei der Division:

$$\begin{aligned}\left|\frac{\frac{\tilde{x}}{\tilde{y}} - \frac{x}{y}}{\frac{x}{y}}\right| &= \left|\frac{\tilde{x}y - x\tilde{y}}{y\tilde{y}} \cdot \frac{y}{x}\right| = \left|\frac{(\tilde{x} - x)y + x(y - \tilde{y})}{x \cdot y \cdot \tilde{y}}\right| \cdot |y| \\ &\leq \varepsilon_x \left|\frac{y}{\tilde{y}}\right| + \varepsilon_y \left|\frac{y}{\tilde{y}}\right|.\end{aligned}$$

Für den letzten Schritt schätzen wir ab, dass

$$\begin{aligned}\left|\frac{y}{\tilde{y}}\right| &\leq \frac{|y - \tilde{y}x| + |\tilde{y}|}{|\tilde{y}|} = 1 + \frac{|\tilde{y} - y|}{|y|} \cdot \frac{|y|}{|\tilde{y}|} \\ &= 1 + \varepsilon_y \frac{|y|}{|\tilde{y}|} \leq 1 + \varepsilon_y \left(1 + \varepsilon_y \frac{|y|}{|\tilde{y}|}\right) \\ &\leq 1 + \varepsilon_y + \varepsilon_y^2 \frac{|y|}{|\tilde{y}|} \leq 1 + \varepsilon_y + \varepsilon_y^2 + \varepsilon_y^3 \frac{|y|}{|\tilde{y}|} \leq \ldots \\ &= 1 + \varepsilon_y + O(\varepsilon_y^2) \text{ für } \varepsilon_y \to 0.\end{aligned}$$

Vernachlässigen der Produkte von Fehlern ergibt das gewünschte Ergebnis. QED

Wir fassen die Ergebnisse zusammen: Ist das zu betrachtende Problem die Multiplikation oder Division von zwei Zahlen (d.h. $f(x,y) = x \cdot y$ oder $f(x,y) = \frac{x}{y}$) so haben wir gezeigt, dass $\left|\frac{f(\tilde{x},\tilde{y})-f(x,y)}{f(x,y)}\right|$ klein ist: Im schlimmsten Fall ist für den relativen Fehler eine Addition der Beträge der relativen Fehler $\varepsilon_y + \varepsilon_y$ der Eingangsgrößen x und y zu erwarten. Beide Probleme sind also gut konditioniert. Betrachten wir nun die Addition: Haben die beiden zu addierenden Zahlen das gleiche Vorzeichen, so sind die Faktoren $|\frac{x}{x+y}|$ und $|\frac{y}{x+y}|$ aus Lemma 1.8 beide durch 1 beschränkt, so dass wir wieder eine gute Kondition erhalten. Bei der Addition von Zahlen verschiedenen Vorzeichens (also der Subtraktion von Zahlen gleichen Vorzeichens) können die beiden Faktoren $|\frac{x}{x-y}|$ und $|\frac{y}{x-y}|$ dagegen beliebig groß werden. Dieses Problem ist also schlecht konditioniert. Wir demonstrieren das an einem Beispiel:

Rechnen wir mit 6-stelliger Genauigkeit und subtrahieren die beiden 6-stelligen Zahlen $x = 1234.00$ und $y = 1233.99$. Angenommen, der relative Fehler von x beträgt $\varepsilon_x = 0.1$, also z.B. $\tilde{x} = 1357.40$ und der relative Fehler von y ist sogar Null ($\tilde{y} = y$). Dennoch ergibt sich ein relativer Fehler ε_{x-y} der Differenz von x und y von

$$\varepsilon_{x-y} = \frac{\tilde{x} - \tilde{y} - (x-y)}{x-y} = \frac{123.41 - 0.01}{0.01} = 12340.00$$

und das, obwohl wir exakt gerechnet haben!

Das Beispiel demonstriert, dass die Subtraktion von zwei Zahlen schlecht konditioniert ist, wenn die beiden zu subtrahierenden Zahlen fast gleich groß sind. Der Effekt wird als **Auslöschung** bezeichnet. Er ist ein ernst zu nehmendes Problem im wissenschaftlichen Rechnen. Man sollte daher, wenn es irgendwie möglich ist, die Differenzenbildung von fast gleich großen Zahlen vermeiden, oder zumindest möglichst zum Schluss eines Verfahrens ausführen.

Konsistenz und Kondition

Betrachten wir abschließend noch einmal ein Verfahren Alg_h, das von einem Diskretisierungsparameter h abhängt. Oft ist es nicht möglich, h so zu wählen, dass gleichzeitig eine gute Kondition und ein kleiner Verfahrensfehler erreicht werden. So möchte man h häufig klein wählen, um den Diskretisierungsfehler klein zu halten. Andererseits können kleine Werte von h aber zu einer schlechten Kondition führen.

Wir demonstrieren das erneut am Beispiel der numerischen Berechnung der Ableitungen einer differenzierbaren Funktion $g : \mathbb{R} \to \mathbb{R}$ mittels des einfachen Differen-

zenquotienten

$$\frac{g(x+h)-g(x)}{h}.$$

Nach dem schon besprochenen Beispiel auf Seite 14 sollte man h möglichst klein wählen, um einen kleinen Diskretisierungsfehler zu erreichen. Andererseits führen kleine Werte von h, $(h > 0)$ zu einer Auslöschung bei der Differenzenbildung im Zähler. Ein relativer Fehler von maximal ε in der Berechnung der Werte $g(x)$ und $g(x+h)$ hat nach Lemma 1.8 bei der Differenzenbildung im Zähler einen relativen Fehler von bis zu

$$\left|\frac{g(x+h)}{g(x+h)-g(x)}\right|\varepsilon + \left|\frac{g(x)}{g(x+h)-g(x)}\right|\varepsilon = \frac{|g(x+h)|+|g(x)|}{|g(x+h)-g(x)|}\varepsilon \approx \frac{2\varepsilon|g(x)|}{|hg'(x)|}$$

zur Folge. Das Problem ist also für kleine Werte von h schlecht konditioniert. Man ist also in einer Zwickmühle: Für kleine Werte von h ist die Auslöschung groß, für große h ist dagegen der Diskretisierungsfehler groß. Einige weitere Betrachtungen führen zu der Faustregel $h \approx \sqrt{\varepsilon}$, die z.B. in [Schaback und Wendland, 2004] nachgelesen werden kann.

Chapter 2

Lineare Gleichungssysteme: Eliminationsverfahren

2.1 Begriffe und Grundlagen

Wir wollen zunächst die nötigen Notationen einführen und dabei einige Begriffe und Ergebnisse aus der Linearen Algebra wiederholen.

Notation 2.1 *$A \in \mathbb{K}^{m,n}$ bezeichne eine reelle oder komplexe $m \times n$* **Matrix**, *d.h. eine Matrix mit m Zeilen und n Spalten. Wir schreiben*

$$A = (a_{ij})_{i=1,\dots,m,j=1,\dots,n} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} = \begin{pmatrix} A_1 & A_2 & \dots & A_n \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix}.$$

Dabei bezeichnen a_{ij} die Elemente der Matrix A, A_j die Spalten der Matrix und a_i ihre Zeilen, $i = 1, \dots, m, j = 1, \dots, n$. Gilt $m = n$ so nennt man die Matrix **quadratisch**.

Matrizen kann man miteinander multiplizieren, allerdings ist die Matrixmultiplikation nicht kommutativ. Die **Einheitsmatrix** $I \in \mathbb{K}^{n,n}$ bezüglich der Multiplikation von quadratischen Matrizen hat die Elemente $e_{ij} = 0$ für alle $i \neq j$, $e_{ii} = 1, i = 1, \dots, n$. Die Spalten von I sind die **Einheitsvektoren**, die wir (abweichend von Notation 2.1) wie in der Literatur sonst auch üblich mit $e_1, \dots, e_n$ bezeichnen. Gibt es zu einer Matrix A eine Matrix A^{-1} mit $A \cdot A^{-1} = A^{-1} \cdot A = I$ so nennt man A invertierbar.
Wir können jetzt definieren, was ein lineares Gleichungssystem ist:

Definition 2.2 *Ein* **lineares Gleichungssystem**

$$Ax = b$$

ist gegeben durch eine Matrix $A \in \mathbb{K}^{m,n}$, *einen Vektor* $b = (b_1, \ldots, b_m)^T \in \mathbb{K}^m$ *und* n *Variable* $x_1, \ldots, x_n$, *geschrieben als Vektor* $x = (x_1, \ldots, x_n)^T$. *Ausgeschrieben erhält man* m *Gleichungen*

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ \vdots & \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= b_m \end{aligned}$$

Falls $b = 0$ *nennt man das Gleichungssystem homogen.*
Ist $m < n$ *so heißt das Gleichungssystem* **unterbestimmt**.

Lineare Gleichungssysteme haben ausgesprochen viele Anwendungen. Einerseits tauchen sie direkt als praktische Probleme auf, andererseits sind sie ein wichtiger Baustein für viele numerische Verfahren z.B. zur numerischen Lösung von Differentialgleichungen.

Wir wiederholen einige Begriffe aus der linearen Algebra. Seien $A_1, \ldots, A_p \in \mathbb{K}^n$ Vektoren. Dann bezeichne

$$span\{A_1, \ldots, A_p\} = \{\sum_{i=1}^{p} \alpha_i A_i : \alpha_i \in \mathbb{K}\}$$

die Menge der von $A_1, \ldots, A_p$ erzeugten Linearkombinationen (die **lineare Hülle** von $A_1, \ldots A_p$).
Sicherheitshalber erinnern wir noch an den Begriff der linearen Unabhängigkeit: Die Vektoren $A_1, \ldots, A_p$ heißen linear unabhängig, falls aus $\sum_{i=1}^{p} \alpha_i A_i = 0$ folgt, dass $\alpha_i = 0$ für $i = 1, \ldots, p$. Die Anzahl der linear unabhängigen Spalten einer Matrix A definiert den Spaltenrang der Matrix, und dieser entspricht ihrem Zeilenrang, d.h. der Anzahl der linear unabhängigen Zeilen von A.

Satz 2.3 *Sei* $A \in \mathbb{K}^{n,n}$. *Die folgenden Aussagen sind äquivalent:*

(i) *A ist invertierbar*

(ii) $\det(A) \neq 0$

(iii) *Die Spalten* $A_1, \ldots, A_n$ *von A sind linear unabhängig.*

(iv) *Die Zeilen* $a_1, \ldots, a_n$ *von A sind linear unabhängig.*

Die Matrix A nennt man in obigem Fall auch *regulär* oder *nicht-singulär*. Eine $m \times n$ Matrix A kann man als lineare Abbildung

$$\begin{aligned} A : \mathbb{K}^n &\rightarrow \mathbb{K}^m \\ x &\rightarrow Ax \end{aligned}$$

auffassen und man kann dementsprechend z.B. vom *Kern*

$$Kern(A) = \{x \in \mathbb{K}^n : Ax = 0\}$$

der Matrix sprechen.

Bevor wir numerische Verfahren zur Lösung eines linearen Gleichungssystems entwickeln, fassen wir einige Ergebnisse (die alle schon bekannt sein sollten) über die Lösbarkeit linearer Gleichungssysteme zusammen.

Satz 2.4

- *Das Gleichungssystem $Ax = b$ hat mindestens eine Lösung genau dann wenn $b \in span\{A_1, \ldots, A_n\}$.*
- *Das Gleichungssystem $Ax = b$ hat höchstens eine Lösung genau dann wenn $A_1, \ldots, A_n$ linear unabhängig sind.*
- *Das Gleichungssystem $Ax = b$ ist eindeutig lösbar genau dann wenn die Matrix A nicht singulär ist. In diesem Fall ist $x = A^{-1}b$ die eindeutige Lösung.*

Aufgabe: *Beweisen Sie Satz 2.4!*

Für unterbestimmte Gleichungssysteme gilt, dass sie – wenn sie überhaupt lösbar sind – niemals eindeutig lösbar sein können: Sei $\bar{x}$ eine Lösung des Gleichungssystems, also $A\bar{x} = b$. Betrachten wir nun das entsprechende homogene System

$$Ax = 0.$$

Weil $m < n$ sind die Vektoren $A_1, \ldots, A_n \in \mathbb{K}^m$ linear abhängig, also gibt es ein $y \neq 0$ mit $Ay = 0$. Dementsprechend gilt $\bar{x} + y \neq \bar{x}$, aber wegen

$$A(\bar{x} + y) = A\bar{x} + Ay = b + 0 = b$$

ist auch $\bar{x} + y$ eine Lösung des Gleichungssystems. Genauer lässt sich die Lösungsmenge durch $\{\bar{x} + y : y \in Kern(A)\}$ angeben.

Das Problem $Ax = b$ heißt *schlecht gestellt*, wenn es nicht eindeutig lösbar ist.

Übersicht über Verfahren zum Lösen von linearen Gleichungssystemen: Man unterscheidet zunächst zwischen *direkten* und *iterativen* Verfahren. Bei den direkten Verfahren erhält man nach endlich vielen Schritten eine Lösung des Problems. Die bekanntesten hiervon sind die sogenannten *Eliminationsverfahren*, bei denen in jedem Schritt eine der n Unbekannten eliminiert wird. Dazu gehören das Gauß-Verfahren (siehe Abschnitt 2.2) und das Cholesky-Verfahren (Abschnitt 2.3). Das QR-Verfahren ist ein *Orthogonalisierungsverfahren* zur Lösung linearer Gleichungssysteme oder zur Behandlung von linearen Ausgleichsproblemen. Es wird in Kapitel 4 besprochen. *Iterative Verfahren* starten mit einer Näherungslösung, die in jedem Schritt verbessert wird. Sie sind vor allem bei großen Gleichungssystemen oder bei Gleichungssystemen mit spezieller Struktur der Matrix A sinnvoll. Mit ihnen werden wir uns später beschäftigen.

2.2 Gauß-Verfahren und LU-Zerlegung

Idee: Betrachten wir als Beispiel ein Gleichungssystem mit

$$A = \begin{pmatrix} 1 & 3 & 2 \\ 0 & 5 & 4 \\ 0 & 0 & 6 \end{pmatrix}, b = \begin{pmatrix} 9 \\ 14 \\ 6 \end{pmatrix}.$$

Ausgeschrieben erhält man das folgende *gestaffeltes* Gleichungssystem

$$\begin{aligned} 1x_1 + 3x_2 + 2x_3 &= 9 \\ 5x_2 + 4x_3 &= 14 \\ 6x_3 &= 6 \end{aligned}$$

Die dritte Gleichung $6x_3 = 6$ enthält nur eine Unbekannte; entsprechend lässt sich der Wert $x_3 = 1$ bestimmen. Setzt man diesen in die zweite Gleichung ein erhält man $5x_2 = 10$, also $x_2 = 2$. Setzt man abschließend die beiden gefundenen Werte in die erste Gleichung ein, ergibt sich $x_1 = 1$.

Die Idee des Gauß-Verfahrens nutzt nun diese einfache Lösbarkeit gestaffelter Gleichungssysteme aus: Ein gegebenes Gleichungssystem wird in ein gestaffeltes Gleichungssystem transformiert, und dann gelöst. Wir formalisieren dazu zunächst, wie man solche gestaffelten Gleichungssysteme beschreiben und lösen kann.

Definition 2.5 *Eine quadratische Matrix $A \in \mathbb{K}^{n,n}$ heißt* **untere Dreiecksmatrix** *falls $a_{ij} = 0$ für alle $i < j$. A heißt* **obere Dreiecksmatrix**, *falls $a_{ij} = 0$ für alle $i > j$. Eine Dreiecksmatrix heißt* **normiert** *falls $a_{ii} = 1$ für $i = 1, \ldots, n$. Ist A eine Dreiecksmatrix, so bezeichnet man $Ax = b$ als* **gestaffeltes** *Gleichungssystem.*

Bemerkung: Eine $n \times n$-Dreiecksmatrix ist regulär genau dann, wenn $a_{ii} \neq 0$ für alle $i = 1, \ldots, n$.

Lemma 2.6 (Lösen durch Rückwärts- oder Vorwärtsselimination)

- *Sei A eine obere Dreiecksmatrix mit Diagonalelementen $a_{ii} \neq 0$ für $i = 1, \ldots, n$. Die Lösung von $Ax = b$ lässt sich dann sukzessive durch*

$$x_j = \frac{1}{a_{jj}} \left(b_j - \sum_{k=j+1}^{n} a_{jk} x_k \right) j = n, \ldots, 1$$

  bestimmen.

- *Sei A eine untere Dreiecksmatrix mit Diagonalelementen $a_{ii} \neq 0$ für $i = 1, \ldots, n$. Die Lösung von $Ax = b$ lässt sich dann sukzessive durch*

$$x_j = \frac{1}{a_{jj}} \left(b_j - \sum_{k=1}^{j-1} a_{jk} x_k \right) j = 1, \ldots, n$$

  bestimmen.

Beweis: Die Gültigkeit der Formeln überprüft man schnell (ausgehend von $j = n$ bei der Rückwärtselimination und von $j = 1$ bei der Vorwärtselimination). QED

Das für obere Dreiecksmatrizen beschriebene Verfahren heißt *Lösen durch Rückwärtseinsetzen* oder *Rückwärtselimination* weil man mit der letzten Gleichung beginnt. Analog nennt man das für untere Dreiecksmatrizen beschriebene Verfahren *Lösen durch Vorwärtseinsetzen* oder *Vorwärtselimination.*

Aufwand: Beim Lösen durch Rückwärtseinsetzen (oder Lösen durch Vorwärtseinsetzen) benötigt man n Divisionen, $\frac{1}{2}n(n-1)$ Multiplikationen und $\frac{1}{2}n(n-1)$ Subtraktionen, also einen Gesamtaufwand von $O(n^2)$.

Definition 2.7 *Eine Faktorisierung einer Matrix $A \in \mathbb{K}^{n,n}$ der Form $A = LU$ mit einer regulären unteren Dreiecksmatrix L und einer regulären oberen Dreiecksmatrix U heißt LU-Zerlegung von A.*

Bevor wir uns ansehen, was für Eigenschaften eine LU-Zerlegung hat und wie man sie für eine gegebene Matrix A bestimmen kann, beschreiben wir drei Anwendungen.

Anwendung 1: Die wichtigste Anwendung der LU-Zerlegung betrifft das Lösen von Gleichungssystemen.
Ist eine LU-Zerlegung von A bekannt, so lässt sich die Lösung des Gleichungssystems $Ax = b$ durch das Lösen von zwei gestaffelten Gleichungssystemen effizient bestimmen: Durch Vorwärtselimination löst man zuerst das Gleichungssystem

$$Lz = b$$

und anschließend durch Rückwärtselimination

$$Ux = z.$$

Die so erhaltene Lösung x erfüllt dann

$$Ax = LUx = Lz = b$$

und ist somit eine Lösung von $Ax = b$.

Anwendung 2: Bestimmung der Inversen A^{-1} einer Matrix A.
Sei A^{-1} gegeben durch ihre Spalten $A^{-1} = (B_1, B_1, \ldots, B_n)$. Dann gilt

$$AB_k = e_k,$$

und B_k ergibt sich als Lösung x von $Ax = e_k$. Kennt man die LU-Zerlegung der Matrix A so bestimmt man also für $k = 1, \ldots, n$ zunächst die Lösung y_k des Gleichungssystems $Ly_k = e_k$ und löst anschließend das Gleichungssystem $Ux_k = y_k$ zur Bestimmung von $B_k := x_k$.

Anwendung 3: Bestimmung der Determinante von A.
Ist $A = LU$ eine LU-Zerlegung von A, so gilt

$$\det(A) = \det(L) \cdot \det(U) = l_{11} \cdot l_{22} \cdot \ldots \cdot l_{nn} \cdot u_{11} \cdot \ldots \cdot u_{nn}.$$

Die Determinante lässt sich also als Produkt der Diagonalelemente von U und L direkt berechnen.

Wir leiten nun ein paar Eigenschaften der LU-Zerlegung her. Dazu brauchen wir zunächst die folgende Aussagen.

Satz 2.8 *Folgende Mengen sind Gruppen bezüglich der Matrixmultiplikation: Die Menge der regulären oberen Dreiecksmatrizen, die Menge der oberen normierten Dreiecksmatrizen, die Menge der regulären unteren Dreiecksmatrizen, die Menge der unteren normierten Dreiecksmatrizen.*

Beweis: Sei Δ eine der oben genannten Mengen. Zu zeigen ist

(0) $A, B \in \Delta \Rightarrow A \cdot B \in \Delta$.

(1) $A \cdot (B \cdot C) = (A \cdot B) \cdot C$ für alle $A, B, C \in \Delta$.

(2) Die Einheitsmatrix $I \in \Delta$

(3) $A \in \Delta \Rightarrow A^{-1} \in \Delta$.

(1) ist für alle Matrizen richtig und (2) ist klar. Wir zeigen also (0) und (3) für die Menge Δ der (normierten) oberen Dreiecksmatrizen. (Für untere Dreiecksmatrizen verläuft der Beweis analog.)

ad (0): Seien $A, B \in \Delta$ und $C = (c_{ij}) = AB$. Dann ist

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} = \sum_{k=i}^{j} a_{ik} b_{kj},$$

weil $a_{ik} = 0$ für $i > k$ und $b_{kj} = 0$ für $k > j$. Für $i > j$ gilt also $c_{ij} = 0$ und damit ist C eine obere Dreiecksmatrix. (Man beachte, dass die Regularität der Matrizen A und B hierfür nicht nötig ist.)

Sind A, B weiterhin normiert, so auch C, denn

$$c_{ii} = a_{ii} b_{ii} = 1.$$

ad (3): Sei $A \in \Delta$ regulär und $A^{-1} = B = (b_{ij})$ die Inverse von A. Dann gilt für die Spalten $B_1, B_1, \ldots, B_n$ von der Inversen

$$AB_k = e_k.$$

Für jedes $k \in \{1, \ldots, n\}$ kann $B_k = (b_{1k}, b_{2k}, \ldots, b_{nk})^T$ also als Lösung von $Ax = e_k$ aufgefasst werden. Nach Lemma 2.6 folgt dass $b_{kj} = 0$ für $j = n, n-1, \ldots, k+1$ und $b_{kk} = \frac{1}{a_{kk}}$, also ist B eine obere Dreiecksmatrix, die für eine normierte Matrix A auch wieder normiert ist.

QED

Man sieht hier schon direkt die folgende Aussage:

Lemma 2.9 *Hat eine reguläre Matrix A eine LU-Zerlegung mit normierter unterer Dreiecksmatrix L, so ist diese eindeutig.*

Beweis: Weil $\det(A) \neq 0$ sind auch die Matrizen L, U mit $A = LU$ regulär. Sei nun $A = L_1 U_1 = L_2 U_2$. Das ist wegen der Regularität aller beteiligten Matrizen äquivalent zu

$$U_1 U_2^{-1} = L_1^{-1} L_2.$$

Wegen Satz 2.8 steht links eine obere und rechts eine untere Dreiecksmatrix. Um Gleichheit zu gewähren, muss also

$$U_1 U_2^{-1} = I = L_1^{-1} L_2$$

gelten, und dementsprechend folgern wir $L_1 = L_2$ und $U_1 = U_2$. QED

Im folgenden beschäftigen wir uns mit dem Gauß-Verfahren, mit dessen Hilfe man die LU-Zerlegung einer Matrix A (sofern sie existiert) bestimmen kann. Die Idee des Gauß-Verfahrens besteht darin, die Matrix durch *elementare Zeilenoperationen* in eine Matrix in Dreiecksform zu transformieren, aus der man die LU-Zerlegung schließlich ablesen kann. Das kann man mit Hilfe der folgenden Matrizen formulieren:

Definition 2.10 *Für einen Vektor* $l^{(k)} = (0, \ldots, 0, t_{k+1}, \ldots, t_n)^T \in \mathbb{K}^n$ *mit* $1 \leq k \leq n$ *und dem k.ten Einheitsvektor* $e_k \in \mathbb{K}^n$ *ist die* **Gauß-Matrix** M_k *definiert durch*

$$M_k := I_n - l^{(k)} e_k^T = \begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ & & \ddots & & & \\ & & & 1 & & \\ & & & -t_{k+1} & 1 & \\ & & & \vdots & & \ddots \\ & & & -t_n & & 1 \end{pmatrix}$$

Sammeln wir zunächst einige Eigenschaften der Gauß-Matrizen.

Lemma 2.11 *Sei* M_k *die Gauß-Matrix bezüglich eines Vektors* $l^{(k)} = (0, \ldots, 0, t_{k+1}, \ldots, t_n)^T$.

1. $\det(M_k) = 1$
2. $M_k^{-1} = I_n + l^{(k)} e_k^T$

Beweis: Da die Gauß-Matrizen untere Dreiecksmatrizen sind, folgt der erste Teil des Lemmas. Für den zweiten Teil rechnet man nach

$$\begin{aligned} M_k M_k^{-1} &= (I_n - l^{(k)} e_k^T)(I_n + l^{(k)} e_k^T) \\ &= I_n + l^{(k)} e_k^T - l^{(k)} e_k^T - l^{(k)} e_k^T l^{(k)} e_k^T = I_n, \end{aligned}$$

wobei im letzten Schritt ausgenutzt wurde, dass $e_k^T l^{(k)} = 0$ gilt. Analog erhält man $M_k^{-1} M_k = I_n$. QED

Multipliziert man eine Gauß-Matrix M_k links an eine Matrix A, so erhält man als Ergebnis eine Matrix A', die aus A entsteht, indem man das $t_j$te Vielfache der $k$ten Zeile a_k von A von der $j$ten Zeile abzieht, für $j = k+1, \ldots, n$. In Formeln erhält man also:

$$M_k \; A = \begin{pmatrix} a_1 \\ \vdots \\ a_k \\ a_{k+1} - t_{k+1} a_k \\ \vdots \\ a_n - t_n a_k \end{pmatrix}.$$

Man nennt diese Operation auch die *Anwendung elementarer Zeilenoperationen.* Lemma 2.11 besagt dabei, dass die Anwendung von elementaren Zeilenoperationen die Determinante der Matrix nicht verändert.

Aufgabe: *Was passiert wenn man eine Gauß-Matrix rechts an eine Matrix A multipliziert?*

Setzt man für einen Vektor $a = (a_1, \ldots, a_n)^T$ und eine Zahl $k \in \{1, \ldots, n\}$ mit $a_k \neq 0$

$$l^{(k)} = (0, \ldots, 0, \tfrac{a_{k+1}}{a_k}, \ldots, \tfrac{a_n}{a_k})^T$$

so erhält man

$$M_k a = (a_1, a_2, \ldots, a_k, 0, \ldots, 0)^T. \tag{2.1}$$

Genau das wird im Gauß-Verfahren zur Transformation einer Matrix auf Dreiecksform ausgenutzt. Das folgende Verfahren ist in der angegebenen Form zur Implementierung allerdings ungeeignet, weil Matrixoperationen rechenzeitmäßig einen hohen Aufwand bedeuten. Eine effizientere Variante wird in Algorithmus 3 auf Seite 36 beschrieben.

Algorithmus 1: Gauß-Verfahren ohne Spaltenpivotsuche (Matrixversion)

Input: $A \in \mathbb{K}^{n,n}$

Schritt 1: $A^{(1)} := A$

Schritt 2: **For** $k = 1, \ldots, n-1$ **do**

$$\begin{aligned} l^{(k)} &:= \left(\underbrace{0, \ldots, 0}_{k \texttt{ mal}}, \frac{a^{(k)}_{k+1,k}}{a^{(k)}_{kk}}, \ldots, \frac{a^{(k)}_{n,k}}{a^{(k)}_{kk}} \right)^T \\ M_k &:= I_n - l^{(k)} e_k^T \\ A^{(k+1)} &:= M_k A^{(k)} \end{aligned}$$

Ergebnis: LU `Zerlegung von` A `mit`

$$\begin{aligned} U &:= A^{(n)} \\ L &:= M_1^{-1} \cdot M_2^{-1} \cdots M_{n-1}^{-1} \end{aligned}$$

Wir müssen nun zeigen, dass obiger Algorithmus hält, was er verspricht, d.h. dass wirklich $A = LU$ gilt, und L eine untere und U eine obere Dreiecksmatrix ist. Außerdem muss die **Durchführbarkeit** des Verfahrens untersucht werden, die nur dann gewährleistet ist, wenn $a^{(k)}_{kk} \neq 0$ für alle $k = 1, \ldots, n-1$. Dazu betrachten wir die **Hauptminoren** der Matrix A.

Satz 2.12 (Korrektheit des Gauß-Verfahrens) *Sei für $k = 1, \ldots, n-1$:*

$$\det \begin{pmatrix} a_{11} & \ldots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \ldots & a_{kk} \end{pmatrix} \neq 0. \tag{2.2}$$

Dann ist Algorithmus 1 korrekt. Genauer:

1. *Algorithmus 1 ist durchführbar, d.h.*

$$a_{kk}^{(k)} \neq 0 \text{ für alle } k = 1, \ldots, n-1. \tag{2.3}$$

2. *Für die Matrizen $A^{(k)}$, $k = 1, \ldots, n$ gilt:*

$$a_{ij}^{(k)} = 0 \text{ für alle } i, j \text{ mit } j < k \text{ und } i > j. \tag{2.4}$$

   Insbesondere ist U eine obere Dreiecksmatrix.

3. *L ist eine untere Dreiecksmatrix.*

4. *$A = LU$.*

Beweis:

ad 1. und 2. Wir zeigen zuerst, dass für jedes feste k (2.3) aus (2.4) folgt. Danach beweisen wir (2.4) für alle k per Induktion.

Sei also $a_{ij}^{(k)} = 0$ für alle i, j mit $j < k$ und $i > j$. Wegen Lemma 2.11 wissen wir, dass

$$\begin{aligned} \det(A^{(k)}) &= \det(M_{k-1}A^{(k-1)}) = \det(M_{k-1})\det(A^{(k-1)}) = \det(A^{(k-1)}) \\ &= \ldots = \det(A). \end{aligned}$$

Wendet man die elementaren Zeilenoperationen ausschließlich auf Submatrizen der Form (2.2) an, gilt diese Gleichung weiterhin, d.h.

$$\begin{aligned} \det \begin{pmatrix} a_{11} & \ldots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \ldots & a_{kk} \end{pmatrix} &= \det \begin{pmatrix} a_{11}^{(k)} & \ldots & a_{1k}^{(k)} \\ \vdots & & \vdots \\ a_{k1}^{(k)} & \ldots & a_{kk}^{(k)} \end{pmatrix} = \det \begin{pmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \ldots & a_{1k}^{(k)} \\ 0 & a_{22}^{(k)} & \ldots & a_{2k}^{(k)} \\ \vdots & 0 & \ddots & \vdots \\ 0 & \ldots & 0 & a_{kk}^{(k)} \end{pmatrix} \\ &= a_{11}^{(k)} \cdot a_{22}^{(k)} \cdot \ldots \cdot a_{kk}^{(k)}, \end{aligned}$$

wobei wir im zweiten Schritt ausgenutzt haben, dass die Matrix $A^{(k)}$ (2.4) erfüllt. Nach Voraussetzung unseres Satzes ist

$$\det \begin{pmatrix} a_{11} & \ldots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \ldots & a_{kk} \end{pmatrix} \neq 0,$$

insbesondere gilt $a_{kk}^{(k)} \neq 0$. Damit ist (2.3) für dieses k gezeigt.

Um (2.4) zu zeigen, nutzen wir diese Aussage in einem Induktionsbeweis. Für den Anfang $k = 1$ ist nichts zu zeigen. Für den Induktionsschritt

$k \to k+1$ nehmen wir an, dass (2.4) für k richtig ist. Insbesondere gilt dann nach dem ersten Teil dieses Beweises, dass $a_{kk}^{(k)} \neq 0$. Der Vektor $l^{(k)}$ ist also definiert. Anwendung von (2.1) ergibt die geforderte Eigenschaft $a_{ik}^{(k+1)} = 0$ für alle $i > k$ für die $k$te Spalte. Zusammen mit der Induktionsannahme folgt (2.4) für $A^{(k+1)}$.

ad 3. L ist definiert als Produkt der M_k^{-1}. Da die M_k alle untere Dreiecksmatrizen sind, sind nach Satz 2.8 auch ihre Inversen Dreiecksmatrizen, ebenso die Produkte ihrer Inversen, also auch L.

ad 4. Nach Algorithmus 1 gilt

$$U = A^{(n)} = M_{n-1}A^{(n-1)} = M_{n-1}M_{n-2} \cdot \ldots \cdot M_1 A,$$

Mit $L = M_1^{-1} \cdot M_2^{-1} \cdots M_{n-1}^{-1}$ folgt $L \cdot U = A$.

Aufgabe: *Eine $n \times n$ Matrix heißt streng diagonal-dominant, falls für alle $i = 1, \ldots, n$ gilt:*

$$2|a_{ii}| > \sum_{j=1}^{n} |a_{ij}|.$$

Zeigen Sie, dass jede streng diagonal-dominante Matrix invertierbar ist und dass Algorithmus 1 auch in diesem Fall korrekt ist. Das heißt, die Aussagen von Satz 2.12 bleiben richtig, auch wenn man die bisherige Voraussetzung (2.2) durch die Forderung nach strenger Diagonal-Dominanz ersetzt.

Lemma 2.13 *Ist Algorithmus 1 durchführbar, so gilt für die Matrix L:*

$$L = I + \sum_{k=1}^{n-1} l^{(k)} e_k^T.$$

Beweis: Nach Definition von L und Lemma 2.11 ist

$$\begin{aligned} L &= M_1^{-1} \cdot M_2^{-1} \cdots M_{n-1}^{-1} \\ &= (I + l^{(1)} e_1^T)(I + l^{(2)} e_2^T) \cdots (I + l^{(n-1)} e_{n-1}^T). \end{aligned}$$

Wir beweisen, dass für alle m gilt:

$$I + \sum_{k=1}^{m} l^{(k)} e_k^T = (I + l^{(1)} e_1^T)(I + l^{(2)} e_2^T) \cdots (I + l^{(m)} e_m^T).$$

Für $m = 1$ sieht man die Behauptung direkt. Per Induktion leitet man sie dann für beliebige m her: Gelte die Behauptung also für $m-1$. Dann betrachte

$$\begin{aligned}
& (I + l^{(1)} e_1^T)(I + l^{(2)} e_2^T) \cdots (I + l^{(m)} e_m^T) \\
= & \left(I + \sum_{k=1}^{m-1} l^{(k)} e_k^T \right) (I + l^{(m)} e_m^T) \\
= & I + l^{(m)} e_m^T + \sum_{k=1}^{m-1} l^{(k)} e_k^T + \sum_{k=1}^{m-1} l^{(k)} \underbrace{e_k^T l^{(m)}}_{=0} e_m^T \\
= & I + \sum_{k=1}^{m} l^{(k)} e_k^T .
\end{aligned}$$

QED

Diese Beobachtung hilft uns, das Gauß-Verfahren effizient zu organisieren: Man speichert die Vektoren $l^{(1)}, l^{(2)}, \ldots, l^{(n-1)}$ über die erzeugten Nullen im unteren Teil der Matrix A, während der obere Teil die Matrix U enthält.

Bevor wir aber die effizientere Variante des Gauß-Verfahrens angeben, möchten wir das Verfahren so erweitern, dass wir es für alle regulären Matrizen anwenden können. Das ist in der Variante aus Algorithmus 1 leider nicht der Fall — sie scheitert schon an einer so einfachen regulären Matrix wie

$$\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

Ein weiteres Problem besteht darin, dass bei kleinen, aber von Null verschiedenen Elementen $a_{kk}^{(k)}$ große Rundungsfehler auftreten können, wie das folgende Beispiel zeigt:
Sei das Gleichungssystem

$$\begin{pmatrix} 0.001 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

gegeben. Die einzige nötige Umformung im Gauß-Verfahren führt zu dem System

$$\begin{pmatrix} 0.001 & 1 \\ 0 & -999 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -998 \end{pmatrix}$$

und entsprechend zu der exakten Lösung von

$$x_1 = \frac{1000}{999} \approx 1, x_2 = \frac{998}{999} \approx 1.$$

Angenommen, wir arbeiten mit zweistelliger Gleitkomma-Arithmetik. Dann erhält man nach der ersten Umformung das auf zwei Stellen gerundete Gleichungssystem

$$\begin{pmatrix} 0.10 \cdot 10^{-2} & 0.10 \cdot 10^{1} \\ 0 & -0.10 \cdot 10^{4} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0.10 \cdot 10^{1} \\ -0.10 \cdot 10^{4} \end{pmatrix},$$

dessen Lösung sich (sogar bei exakter Rechnung) zu $x_1 = 0$ und $x_2 = 1$ ergibt, also weit von der echten Lösung entfernt liegt.

Erfreulicherweise lassen sich die beiden aufgeführten Schwierigkeiten durch das nun zu beschreibende Verfahren der *Pivotisierung* vermeiden. Im einfachsten Fall der *Zeilenpivotisierung* vertauscht man während des $k$ten Schritts des Gauß-Verfahrens die $k$te Zeile mit einer darunter liegenden, und zwar mit der, die den betragsmäßig größten Eintrag in der $k$ten Spalte aufweist. Das Ziel dabei ist, dass nach der Vertauschung das neue Element $a_{kk}^{(k)}$ so groß wie möglich wird. Formal wählt man im $k$ten Schritt ein $j \in \{k, k+1, \ldots, n\}$ so dass

$$|a_{jk}^{(k)}| \geq |a_{lk}^{(k)}| \text{ für alle } l = k, \ldots, n.$$

In diesem Fall nennt man $a_{jk}^{(k)}$ das **Pivotelement**. Zur formalen Beschreibung dieser Vertauschungen benötigen wir die folgenden Matrizen.

Definition 2.14 *Eine bijektive Abbildung* $\Pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ *heißt* **Permutation** *der Menge* $\{1, \ldots, n\}$. *Eine* $n \times n$ *Matrix* P *heißt* **Permutationsmatrix** *falls es eine Permutation* Π *gibt so dass*

$$Pe_i = e_{\Pi(i)} \text{ für alle } i = 1, \ldots, n.$$

P entsteht also durch Permutation der Spalten der Einheitsmatrix. Wir sammeln Eigenschaften von Permutationsmatrizen.

Satz 2.15 *Sei die* $n \times n$*-Matrix* P *eine Permutationsmatrix zur Permutation* Π. *Dann gilt:*

1. P *ist invertierbar.*
2. P^{-1} *ist die Permutationsmatrix, die zu der Permutation* Π^{-1} *gehört. (Dabei ist* Π^{-1} *die Umkehrabbildung von* Π.*)*
3. P *ist orthogonal, das heißt, es gilt* $P^{-1} = P^T$.

Beweis: *Übung*

Eine Erweiterung der zweiten Aussage des Satzes soll noch erwähnt werden: Für zwei Permutationen Π_1, Π_2 mit zugehörigen Permutationsmatrizen P_1,P_2 gilt wegen

$$\begin{aligned} P_1 P_2(e_i) &= P_1(e_{\Pi_2(i)}) \\ &= e_{\Pi_1(\Pi_2(i))} = e_{\Pi_1 \circ \Pi_2(i)} \end{aligned}$$

dass P_1P_2 die Permutationsmatrix ist, die zu der Permutation $\Pi_1 \circ \Pi_2$ gehört, das heißt, die Verkettung von zwei Permutationen entspricht dem Produkt der entsprechenden Permutationsmatrizen.

Um das Gauß-Verfahren zu verbessern, benötigen wir spezielle Permutationen, nämlich solche, die genau zwei Elemente $r < s$ vertauschen. Zu so einer Permutation

$$\begin{aligned} \Pi(r) &= s \\ \Pi(s) &= r \\ \Pi(i) &= i \quad \text{für alle } i \notin \{r, s\} \end{aligned}$$

gehört entsprechend die Matrix

$$P_{rs} = (e_1, \ldots, e_{r-1}, e_s, e_{r+1}, \ldots, e_{s-1}, e_r, e_{s+1}, \ldots, e_n).$$

Solche Matrizen sind symmetrisch, das heißt $P_{rs} = P_{rs}^T$ und man kann sie auch als $P_{rs} = I - (e_r - e_s)(e_r - e_s)^T$ schreiben.

Man sollte sich das folgende einprägen:

- Die Linksmultiplikation einer Matrix A mit P_{rs} vertauscht die $r$te mit der $s$ten Spalte.
- Die Rechtsmultiplikation einer Matrix A mit P_{rs} vertauscht die $r$te mit der $s$ten Zeile.

Formal können wir nun die Matrixversion des Gauß-Verfahrens mit Spaltenpivotisierung folgendermaßen beschreiben.

Algorithmus 2: Gauß-Verfahren mit Spaltenpivotsuche (Matrixversion)

Input: $A \in \mathbb{K}^{n,n}$

Schritt 1: $\tilde{A}^{(1)} := A$

Schritt 2: **For** $k = 1, \ldots, n-1$ **do**

`Bestimme einen Pivotindex` $r \in \{k, \ldots, n\}$ `mit` $|\tilde{a}_{rk}^{(k)}| = \max_{i=k,\ldots,n} |\tilde{a}_{ik}^{(k)}|$.

$$\begin{aligned} P^{(k)} &:= P_{kr} \\ A^{(k)} &:= P^{(k)} \tilde{A}^{(k)} \\ l^{(k)} &:= \left(\underbrace{0, \ldots, 0}_{k \text{ mal}}, \frac{a_{k+1,k}^{(k)}}{a_{kk}^{(k)}}, \ldots \frac{a_{n,k}^{(k)}}{a_{kk}^{(k)}} \right)^T \\ M_k &:= I_n - l^{(k)} e_k^T \\ \tilde{A}^{(k+1)} &:= M_k A^{(k)} \end{aligned}$$

Ergebnis: $PA = LU$ `mit`

$$\begin{aligned} P &= P^{(n-1)} \cdot \ldots \cdot P^{(1)} \text{ eine Permutationsmatrix} \\ U &:= \tilde{A}^{(n)} \text{ ist eine obere Dreiecksmatrix} \\ L &:= I + \sum_{k=1}^{n-1} \Theta^{(k)} e_k^T \text{ ist eine untere Dreiecksmatrix mit} \\ \Theta^{(k)} &:= P^{(n-1)} \cdot \ldots \cdot P^{(k+1)} l^{(k)}. \end{aligned}$$

Satz 2.16 *Für eine reguläre $n \times n$ Matrix A existiert eine Permutationsmatrix $P \in \mathbb{R}^{n,n}$, eine normierte untere Dreiecksmatrix $L \in \mathbb{K}^{n,n}$ und eine obere Dreiecksmatrix $U \in \mathbb{K}^{n,n}$ so dass $PA = LU$, und diese Zerlegung wird von Algorithmus 2 gefunden.*

Beweis: Im Beweis zeigen wir zunächst die folgenden Eigenschaften:

1. Algorithmus 2 ist durchführbar, d.h.
$$a_{kk}^{(k)} \neq 0 \text{ für } k = 1, \ldots, n-1. \tag{2.5}$$
2. Für die Matrizen $A^{(k)}, k = 1, \ldots, n-1$ zeigen wir:
$$\begin{aligned} a_{ij}^{(k)} &= 0 \text{ für alle } i, j \text{ mit } j < k, i > j. && (2.6) \\ A^{(k)} &= P^{(k)} M_{k-1} P^{(k-1)} \cdot \ldots \cdot P^{(2)} M_1 P^{(1)} A && (2.7) \end{aligned}$$
3. Für die Matrizen $\tilde{A}^{(k)}, k = 1, \ldots, n$ zeigen wir:
$$\tilde{a}_{ij}^{(k)} = 0 \text{ für alle } i, j \text{ mit } j < k, i > j. \tag{2.8}$$

Insbesondere folgt aus 2.8, dass U eine obere Dreiecksmatrix ist.

Ähnlich wie im Beweis zu Satz 2.12 zeigen wir, dass für jedes feste $k = 1, \ldots, n-1$ aus den Eigenschaften (2.6) und (2.7) die erstgenannte Aussage $a_{kk}^{(k)} \neq 0$ folgt und beweisen anschließend (2.6) und (2.7) für alle k per Induktion.

Gelte also (2.6) und (2.7) für k. Wir nehmen an, dass $a_{kk}^{(k)} = 0$. Nach der Definition von $P^{(k)}$ erfüllt die Matrix $A^{(k)} = P^{(k)} \tilde{A}^{(k)}$

$$|a_{kk}^{(k)}| \geq |a_{lk}^{(k)}| \text{ für alle } l = k, \ldots, n.$$

Wegen $a_{kk}^{(k)} = 0$ folgt daraus dass die erste Spalte der Submatrix

$$C = \begin{pmatrix} a_{kk}^{(k)} & \ldots & a_{kn}^{(k)} \\ \vdots & & \vdots \\ a_{nk}^{(k)} & \ldots & a_{nn}^{(k)} \end{pmatrix}$$

eine Nullspalte ist und entsprechend $\det(C) = 0$ gilt. Wegen (2.6) gilt weiter dass

$$A^{(k)} = \left(\begin{array}{cccc|c} a_{11}^{(k)} & \dots & & a_{1k-1}^{(k)} & \\ 0 & a_{22}^{(k)} & \dots & a_{2k-1}^{(k)} & \\ \vdots & 0 & \ddots & \vdots & * \\ 0 & \dots & & a_{k-1k-1}^{(k)} & \\ \hline & & 0 & & C \end{array}\right),$$

also folgt

$$\det(A^{(k)}) = a_{11}^{(k)} \cdot a_{22}^{(k)} \cdot \ldots \cdot a_{k-1k-1}^{(k)} \cdot \det(C) = 0.$$

Wegen (2.7) folgt daraus $\det(A) = 0$, ein Widerspruch zur Regularität von A.

Jetzt zeigen wir (2.6), (2.7) und (2.8) per Induktion.
Für alle drei Aussagen ist der Induktionsanfang $k = 1$ klar. Für den Übergang $k \to k+1$ nehmen wir an, dass die Aussagen für k schon gelten. Wegen dem ersten Teil des Beweises gilt $a_{kk}^{(k)} \neq 0$, also ist die Matrix $A^{(k+1)}$ definiert.
(2.8) gilt dann für $\tilde{A}^{(k+1)}$ wegen der Induktionsannahme für $A^{(k)}$ und wegen der alten Aussage (2.1) auf Seite 28. Für (2.6) nutzen wir die Aussage für $\tilde{A}^{(k+1)}$ zusammen mit dem Argument, dass die Transformation $P^{(k+1)}$ die ersten k Zeilen von $\tilde{A}^{(k+1)}$ unberührt lässt. (2.7) ergibt sich schließlich durch Einsetzen

$$A^{(k+1)} = P^{(k)}\tilde{A}^{(k+1)} = P^{(k)}M_k A^{(k)}$$

und der Induktionsannahme.

Damit wissen wir, dass das Verfahren durchführbar ist und $U = \tilde{A}^{(n)}$ eine obere Dreiecksmatrix ist. L ist eine untere Dreiecksmatrix, da $\Theta^{(k)}$ nur Permutationen beinhaltet, die Zeilen mit Index $i \geq k$ vertauschen und damit die Eigenschaft der unteren Dreiecksmatrix im Vergleich zu Satz 2.12 nicht verändern.

Es bleibt noch, nachzuweisen, dass $PA = LU$ gilt. Dazu verwenden wir (2.7) und erhalten

$$U = \tilde{A}^{(n)} = M_{n-1}A^{n-1} = M_{n-1}P^{(n-1)} \cdot \ldots \cdot P^{(2)}M_1P^{(1)}A.$$

Wir nutzen, dass $M_j^{-1} = I + l^{(j)}e_j^T$ (Lemma 2.11) und $(P^{(k)})^{-1} = P^{(k)}$ (Lemma 2.15). Somit ergibt sich

$$A = P^{(1)}(I + l^{(1)}e_1^T)P^{(2)}(I + l^{(2)}e_2^T)P^{(3)} \cdot \ldots \cdot P^{(n-1)}(I + l^{(n-1)}e_{n-1}^T)U \qquad (2.9)$$

Wir möchten nun beide Seiten der Gleichung von links mit $P = P^{(n-1)} \cdot \ldots \cdot P^{(2)}P^{(1)}$ multiplizieren. Um den entstehenden Term zu vereinfachen, überlegen wir uns zunächst, dass für beliebige Vektoren l und alle $j > i$

$$P^{(j)}(I + le_i^T)P^{(j)} = (I + P^{(j)}l(P^{(j)}e_i)^T) = (I + P^{(j)}le_i^T)$$

gilt, weil $P^{(j)}e_i = e_i$, falls $j > i$.
Diese Aussage nutzen wir, um bei der Multiplikation von (2.9) mit $P = P^{(n-1)} \cdot \ldots \cdot P^{(2)}P^{(1)}$ die Permutationsmatrizen $P^{(i)}$ für $i \geq 3$ durch das Einfügen von Identitäten $I = P^{(i)}P^{(i)}$ bis zum entsprechenden Faktor $(I+l^{(i-1)}e_{i-1}^T)P^{(i)}$ "durchrutschen" zu lassen:

$$\begin{aligned}
P^{(1)}A &= (I + l^{(1)}e_1^T)P^{(2)}(I + l^{(2)}e_2^T)P^{(3)} \cdot \ldots \cdot P^{(n-1)}(I + l^{(n-1)}e_{n-1}^T)U \\
P^{(2)}P^{(1)}A &= P^{(2)}(I + l^{(1)}e_1^T)P^{(2)}(I + l^{(2)}e_2^T)P^{(3)} \cdot \ldots \cdot P^{(n-1)}(I + l^{(n-1)}e_{n-1}^T)U \\
&= (I + P^{(2)}l^{(1)}e_1^T)(I + l^{(2)}e_2^T)P^{(3)} \cdot \ldots \cdot P^{(n-1)}(I + l^{(n-1)}e_{n-1}^T)U \\
P^{(3)}P^{(2)}P^{(1)}A &= \underbrace{P^{(3)}(I + P^{(2)}l^{(1)}e_1^T)P^{(3)}}_{I+P^{(3)}P^{(2)}l^{(1)}e_1^T}\underbrace{P^{(3)}(I + l^{(2)}e_2^T)P^{(3)}}_{I+P^{(3)}l^{(2)}e_2^T}(I + l^{(3)}e_3^T) \cdot \ldots \\
&\qquad \ldots \cdot P^{(n-1)}(I + l^{(n-1)}e_{n-1}^T)U \\
&= (I + P^{(3)}P^{(2)}l^{(1)}e_1^T)(I + P^{(3)}l^{(2)}e_2^T) \cdot \ldots \cdot P^{(n-1)}(I + l^{(n-1)}e_{n-1}^T)U \\
&\vdots \qquad\qquad\qquad\qquad\qquad\qquad \vdots \\
\Longrightarrow PA &= (I + \Theta^{(1)}e_1^T)(I + \Theta^{(2)}e_2^T) \cdot \ldots \cdot (I + \Theta^{(n-1)}e_{n-1}^T)U \\
&= (I + \textstyle\sum_{k=1}^{n-1} \Theta^{(k)}e_k^T)U,
\end{aligned}$$

wobei der letzte Schritt per Induktion analog zu dem entsprechenden Schritt im Beweis von Lemma 2.13 (auf S. 30) gezeigt wird. QED

Für die praktische Implementierung empfiehlt es sich, auf Matrixoperationen zu verzichten, da diese aufwändig sind. Die folgende Variante ist effizienter.

Algorithmus 3: Gauß-Verfahren mit Spaltenpivotsuche

Input: $A \in \mathbb{K}^{n,n}$

Schritt 1: `For` $k = 1$ `to` $n-1$ `do`

> **Schritt 1.1:** `Finde Pivotelement` a_{rk}.
>
> **Schritt 1.2:** `Vertausche Zeilen` k `und` r `in` A `sowie` b_k `und` b_r.
>
> **Schritt 1.3:** `For` $i = k+1$ `to` n `do`
>
> > **Schritt 1.3.1.** $a_{ik} = \frac{a_{ik}}{a_{kk}}$
> >
> > **Schritt 1.3.2.** `For` $j = k+1$ `to` n `do` $a_{ij} = a_{ij} - a_{ik} \cdot a_{kj}$

Ergebnis: L `und` U `sind gegeben durch`

$$l_{ij} = \begin{cases} a_{ij} & \text{für } i > j \\ 1 & \text{für } i = j \\ 0 & \text{für } i < j \end{cases} \qquad u_{ij} = \begin{cases} a_{ij} & \text{für } i \leq j \\ 0 & \text{für } i > j \end{cases}$$

Aufwand der LU-Zerlegung nach Algorithmus 3:
Wir zählen hier noch einmal gründlich. Die äußere for-Schleife wird für jedes $k = 1, \ldots n-1$ durchlaufen. Darin werden folgende Operationen durchgeführt:

- Maximumsuche bei der Bestimmung des Pivotindexes: $n-1$ Vergleiche
- Vertauschungen sind Zuweisungen, die wir nicht mit zählen
- Innere for-Schleifen: $n-k$ Divisionen, $(n-k)(n-k)$ Multiplikationen, $(n-k)(n-k)$ Additionen

Zusammen beträgt die Anzahl der benötigten Operationen also

$$(n-1)(n-1) + \sum_{k=1}^{n-1}(n-k) + 2(n-k)^2 = \mathrm{O}(n^3)$$

Aufgabe: *Rechnen Sie die Anzahl der Operationen exakt (also ohne Abschätzung durch* O*) aus. Bestimmen Sie außerdem die Anzahl der in der Matrixversion (Algorithmus 2) benötigten Operationen exakt und durch* O*. Vergleichen Sie!*

2.3 Das Cholesky-Verfahren

Wir betrachten auch in diesem Abschnitt Gleichungssysteme $Ax = b$, allerdings nehmen wir nun an, dass die Matrix A eine symmetrische und positiv definite Matrix ist.

Definition 2.17 *Eine Matrix $A \in \mathbb{R}^{n,n}$ heißt* **symmetrisch** *falls $A = A^T$. Eine symmetrische Matrix A heißt*

- **positiv definit** *falls $x^T Ax > 0$ für alle $x \in \mathbb{R}^n \setminus \{0\}$,*
- **positiv semi-definit** *falls $x^T Ax \geq 0$ für alle $x \in \mathbb{R}^n \setminus \{0\}$.*

Lemma 2.18 *Die folgenden Aussagen gelten:*

1. *Eine symmetrische Matrix ist genau dann positiv definit, wenn alle ihre Eigenwerte echt positiv sind.*
2. *Eine symmetrische Matrix ist genau dann positiv semi-definit, wenn alle ihre Eigenwerte größer oder gleich Null sind.*

3. *Eine symmetrische Matrix ist genau dann positiv definit, ihre Hauptminoren positiv sind, d.h. wenn für alle ihre linken oberen $k \times k$-Teilmatrizen*

$$A^{[k]} := \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \dots & a_{kk} \end{pmatrix}, \quad k = 1, \dots, n$$

gilt: $\det(A^{[k]}) \geq 0$.

Positiv definite Matrizen sind also regulär (weil $\det(A^{[n]}) = \det(A) \neq 0$). Wir betrachten die folgenden beiden Zerlegungen:

Definition 2.19 *Eine Faktorisierung einer symmetrischen Matrix $A \in \mathbb{R}^{n,n}$ der Form $A = LL^T$ mit einer regulären unteren Dreiecksmatrix $L \in \mathbb{R}^{n,n}$ heißt* **Cholesky-Zerlegung** *von A .*

Definition 2.20 *Eine Faktorisierung einer symmetrischen Matrix $A \in \mathbb{R}^{n,n}$ der Form $A = LDL^T$ mit einer normierten unteren Dreiecksmatrix L und einer Diagonalmatrix D heißt* **LDL-Zerlegung** *von A .*

Im folgenden werden wir uns u.a. mit Diagonalmatrizen beschäftigen, die wir wie folgt bezeichnen.

Notation 2.21 *Für einen Vektoren $a \in \mathbb{K}^n$ mit ist die* **Diagonalmatrix** *bezüglich a gegeben durch*

$$diag(a) = \begin{pmatrix} a_1 & & & & \\ & a_2 & & & \\ & & \ddots & & \\ & & & a_{n-1} & \\ & & & & a_n \end{pmatrix}$$

Für positive definite symmetrische Matrizen sind die folgenden Aussagen bekannt.

Satz 2.22 *Sei $A \in \mathbb{R}^{n,n}$ eine positiv definite symmetrische Matrix. Dann existiert eine eindeutig bestimmte LDL-Zerlegung von A.*

Beweis: Zunächst bestätigen wir, dass A eine LU-Zerlegung mit normierter unterer Dreiecksmatrix L hat: Nach Satz 2.12 gilt das, wenn die Teilmatrizen

$$A^{[k]} := \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \dots & a_{kk} \end{pmatrix}$$

die Bedingung (2.2) erfüllen, d.h. wenn $\det(A^{[k]}) \neq 0$, $k = 1, \ldots, n$. Weil A positiv definit ist, gilt nach Lemma 2.18 sogar $\det(A^{[k]}) > 0$, die Bedingung ist also erfüllt. (Das heißt, die LU-Zerlegung von A kann ohne Pivotisierung gefunden werden.)

Sei daher

$$A = LU \tag{2.10}$$

mit normierter unterer Dreiecksmatrix L und oberer Dreiecksmatrix U. Wir setzen $D = diag(u_{11}, \ldots, u_{nn})$ als die Diagonalmatrix mit den Einträgen aus der Hauptdiagonalen von U. Da U regulär ist, ist auch D regulär, so dass wir

$$\tilde{U} := D^{-1}U$$

definieren können. Es gilt $LD\tilde{U} = LU = A$. Wir möchten zeigen, dass $\tilde{U} = L^T$: Betrachte dazu

$$A = A^T = (LD\tilde{U})^T = \tilde{U}^T D^T L^T = \tilde{U}^T \cdot (D^T L^T). \tag{2.11}$$

$\tilde{U}$ ist nach Konstruktion eine normierte obere Dreiecksmatrix, also ist $\tilde{U}^T$ eine normierte untere Dreiecksmatrix. Weiter ist $D^T L^T$ eine obere Dreiecksmatrix, also ist (2.11) auch eine LU-Zerlegung von A mit normierter unterer Dreiecksmatrix. Wegen Lemma 2.9 ist die LU-Zerlegung von A eindeutig, also folgern wir aus dem Vergleich von (2.10) und (2.11) dass

$$L = \tilde{U}^T$$

und haben damit die LDL-Zerlegung von A gefunden.

Sei nun $A = L'D'(L')^T$ eine weitere LDL-Zerlegung von A mit normierter unterer Dreiecksmatrix L, so kann man wiederum

$$A = L' \cdot (D'(L')^T)$$

als LU-Zerlegung auffassen. Da die LU-Zerlegung nach Lemma 2.9 eindeutig ist, folgt

$$L' = L \quad \text{und} \quad D'(L')^T = DL^T,$$

wobei sich aus letzterem wegen der Invertierbarkeit von $L = L'$ auch $D' = D$ ergibt. QED

Der Beweis des Satzes zeigt außerdem, dass die LU-Zerlegung einer symmetrischen und positiv definiten Matrix A ohne Pivotisierung gefunden werden kann. Wir kommen nun auf die Cholesky-Zerlegung zurück.

Satz 2.23 *Sei $A \in \mathbb{R}^{n,n}$ eine positiv definite symmetrische Matrix. Dann existiert eine Cholesky-Zerlegung von $A = LL^T$ mit positiven Diagonalelementen von L. Unter dieser Nebenbedingung ist L eindeutig bestimmt.*

Beweis: Nach Satz 2.22 gibt es eine eindeutige LDL-Zerlegung

$$A = LDL^T$$

von A. Bezeichnen wir mit $A^{[k]}$ und $D^{[k]}$ wieder die linken oberen $k \times k$ Teilmatrizen von A und D. Weil L eine untere Dreiecksmatrix ist, gilt

$$A^{[k]} = L^{[k]} D^{[k]} (L^{[k]})^T. \tag{2.12}$$

Wegen der positiven Definitheit von A (siehe Lemma 2.18) gilt $\det(A^{[k]}) > 0$. Zusammen mit (2.12) erhalten wir

$$d_{11} \cdot \ldots \cdot d_{kk} = \det(D^{[k]}) = \det(L^{[k]}) \cdot \det(A^{[k]}) \cdot \det(L^{[k]})^T = \det(A^{[k]}) > 0.$$

Diese Aussage gilt für alle k, also sind die Diagonalelemente von D positiv. Jetzt setzen wir

$$\tilde{L} = L \cdot diag(\sqrt{d_{11}}, \ldots, \sqrt{d_{nn}}) \tag{2.13}$$

und erhalten aus der normierten unteren Dreiecksmatrix L eine untere Dreiecksmatrix $\tilde{L}$ mit positiven Diagonalelementen, für die

$$\tilde{L}\tilde{L}^T = L \cdot diag(\sqrt{d_{11}}, \ldots, \sqrt{d_{nn}}) \cdot diag(\sqrt{d_{11}}, \ldots, \sqrt{d_{nn}}) \cdot L^T = LDL^T = A$$

gilt. Die entsprechende Cholesky-Zerlegung ist also gefunden.

Um die Eindeutigkeit zu zeigen, sei neben $A = \tilde{L}\tilde{L}^T$

$$A = \tilde{L}'(\tilde{L}')^T$$

eine weitere Cholesky-Zerlegung mit Diagonalelementen $\lambda_1, \lambda_2, \ldots, \lambda_n > 0$. Mit $D' := diag(\lambda_1^2, \ldots, \lambda_n^2)$ und

$$L' := \tilde{L}' \cdot diag\left(\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_n}\right)$$

erhält man $A = L'D'(L')^T$, also eine weitere LDL-Zerlegung von A mit normierter unterer Dreiecksmatrix L'. Aus der Eindeutigkeit der LDL-Zerlegung (nach Satz 2.22) folgt $L = L'$ und $D = D'$. Letzteres bedeutet $d_{ii} = \lambda_i^2$ für $i = 1, \ldots, n$ und wegen der Positivität der λ_i also

$$\lambda_i = \sqrt{d_{ii}} \text{ für } i = 1, \ldots, n.$$

Zusammen erhält man

$$\begin{aligned} \tilde{L}' &= L' \cdot diag(\lambda_1, \ldots, \lambda_n) \\ &= L \cdot diag(\sqrt{d_{11}}, \ldots, \sqrt{d_{nn}}) = \tilde{L} \text{ nach (2.13)} \end{aligned}$$

Um eine Cholesky-Zerlegung effizient ausrechnen zu können, betrachten wir die Gleichung $A = LL^T$ komponentenweise. Das ergibt ein Gleichungssystem mit Unbekannten l_{ij} für $i \geq j$. Bezeichnen wir dazu im folgenden mit l_{ij}^T die Elemente der Matrix L^T. Dann ergibt sich

$$a_{ik} = \sum_{j=1}^{n} l_{ij} l_{jk}^T = \sum_{j=1}^{n} l_{ij} l_{kj} = \sum_{j=1}^{k} l_{ij} l_{kj} \text{ für } k = 1, \ldots, n,\ i = k+1, \ldots, n \quad (2.14)$$

und

$$a_{kk} = \sum_{j=1}^{n} l_{kj} l_{jk}^T = \sum_{j=1}^{n} l_{kj} l_{kj} = \sum_{j=1}^{k} l_{kj}^2 \text{ für } K = 1, \ldots, n. \quad (2.15)$$

Wählt man die Reihenfolge geschickt aus, lassen sich die Werte l_{ij} effizient berechnen: Zunächst ergibt sich l_{11} aus (2.15) für $k = 1$ zu $l_{11} = \sqrt{a_{11}}$. Danach lassen sich nacheinander die Werte $l_{21}, \ldots, l_{n1}$ der ersten Spalte von L durch (2.14) bestimmen, dann das Diagonalelement der zweiten Spalte durch (2.15) und so weiter. Es ergibt sich das folgende Verfahren, in dem wir nur das untere Dreieck der Matrix A benutzen und die Elemente von L gleich über die Werte von A schreiben.

Algorithmus 4: Cholesky-Verfahren

Input: $A \in \mathbb{K}^{n,n}$ `symmetrisch und positiv definit gegeben durch Werte` a_{ij} `für` $i \geq j$.

Schritt 1: `For` $k = 1$ `to` $n - 1$ `do`

Schritt 1.1: $a_{kk} = \sqrt{a_{kk} - \sum_{j=1}^{k-1} |a_{kj}|^2}$

Schritt 1.2: `For` $i = k + 1$ `to` n `do`

$$a_{ik} = \frac{1}{a_{kk}} (a_{ik} - \sum_{j=1}^{k-1} a_{ij} a_{kj})$$

Ergebnis: L `ist gegeben durch` $l_{ij} = \begin{cases} a_{ij} & \text{für } i \geq j \\ 0 & \text{für } i < j \end{cases}$

2.4 Schwach besetzte Matrizen

Die bisher beschriebenen Verfahren sind bei sehr großen Matrizen leider ineffizient. Daher versucht man, die LU-Zerlegung an Matrizen mit spezieller Struktur anzupassen. Einen ersten Ansatz haben wir im letzten Abschnitt bei symmetrischen Matrizen kennengelernt. In Anwendungen treten oft *schwach besetzte* Matrizen auf, in denen für die meisten Elemente $a_{ij} = 0$ gilt. Leider sind die bei der LU-Zerlegung von schwach besetzten Matrizen entstehenden Dreiecksmatrizen L und U im allgemeinen nicht auch wieder schwach besetzt. Als Beispiel sei die Matrix

$$A = \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 1 & 0 & 0 & 0 \\ 0.1 & 0 & 1 & 0 & 0 \\ 0.1 & 0 & 0 & 1 & 0 \\ 0.1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

aus dem Skriptum von G. Lube genannt, bei der die Dreiecksmatrizen ihrer LU-Zerlegung voll besetzt sind. Es gibt aber eine Klasse von Matrizen, bei der sich die Struktur der Matrix A auf die Struktur der Matrizen L und U ihrer LU-Zerlegung überträgt. Dazu gehören sogenannte *Bandmatrizen*.

Definition 2.24 *Eine Matrix $A = (a_{ij}) \in \mathbb{K}^{n,n}$ ist eine (p,q)-***Bandmatrix***, falls für alle $i > j + p$ und für alle $j > i + q$ gilt: $a_{ij} = 0$. Die* **Bandbreite** *von A ist dann $p + q + 1$.*

Die folgende Matrix ist ein Beispiel für eine $(2,1)$-Bandmatrix:

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

Jede untere Dreiecksmatrix ist eine $(n-1,0)$-Bandmatrix, jede obere Dreiecksmatrix ist eine $(0,n-1)$-Bandmatrix.

Satz 2.25 *Sei $A = LU$ die LU-Zerlegung einer (p,q)-Bandmatrix A mit oberer Dreiecksmatrix U und normierter unterer Dreiecksmatrix L. Dann ist L eine $(p,0)$-Bandmatrix und U eine $(0,q)$-Bandmatrix.*

Beweis: Wir beweisen den Satz für feste p, q mittels vollständiger Induktion nach n. Für $n = 1$ ist nichts zu zeigen. Für den Induktionsschritt $n \to n+1$ nehmen wir also an, dass die Aussage für Matrizen der Dimension $n \times n$ richtig ist. Betrachte nun eine (p,q)-Bandmatrix $A \in \mathbb{K}^{n+1,n+1}$ mit LU-Zerlegung $A = LU$. Wir partitionieren $L = \begin{pmatrix} 1 & 0 \\ v & L_1 \end{pmatrix}$ und $U = \begin{pmatrix} \alpha & w^T \\ 0 & U_1 \end{pmatrix}$ mit $\alpha \in \mathbb{K}$, $v, w \in \mathbb{K}^n$ und

$L_1, U_1 \in \mathbb{K}^{n,n}$, wobei L_1 eine normierte untere Dreiecksmatrix und U_1 eine obere Dreiecksmatrix ist. Wir erhalten:

$$\begin{pmatrix} 1 & 0 \\ v & L_1 \end{pmatrix} \cdot \begin{pmatrix} \alpha & w^T \\ 0 & U_1 \end{pmatrix} = \begin{pmatrix} \alpha & w^T \\ \alpha v & vw^T + L_1 U_1 \end{pmatrix} = A.$$

Weil A eine (p,q)-Bandmatrix ist, folgt

$$v_i = 0 \text{ für alle } i > p, \tag{2.16}$$
$$w_j = 0 \text{ für alle } j > q, \tag{2.17}$$
$$vw^T + L_1 U_1 \text{ ist eine } (p,q)\text{-Bandmatrix.} \tag{2.18}$$

Aus (2.16) und (2.17) folgt, dass vw^T eine (p,q)-Bandmatrix ist, zusammen mit (2.18) ist also auch $\tilde{B} := L_1 U_1$ eine (p,q)-Bandmatrix. Da $\tilde{B}$ eine LU-Zerlegung besitzt, können wir die Induktionsannahme anwenden und folgern, dass L_1 eine $(p,0)$-Bandmatrix und U_1 eine $(0,q)$-Bandmatrix ist. Zusammen mit (2.16) folgt schließlich, dass L_1 eine $(p,0)$-Bandmatrix und analog folgt aus (2.17), dass U_1 eine $(0,q)$-Bandmatrix ist. QED

Mit folgendem Algorithmus kann man die LU-Zerlegung einer (p,q)-Bandmatrix bestimmen (falls sie existiert).

Algorithmus 5: LU-Zerlegung einer Bandmatrix

Input: (p,q)-Bandmatrix $A \in \mathbb{K}^{n,n}$, für die eine LU-Zerlegung existiert.

Schritt 1: For $k = 1$ to $n-1$, for $i = k+1$ to $\min\{k+p, n\}$ do

Schritt 1.1: $a_{ik} := \frac{a_{ik}}{a_{kk}}$

Schritt 1.2: For $j = k+1$ to $\min\{k+q, n\}$ do $a_{ij} := a_{ij} - a_{ik} a_{kj}$

Ergebnis: LU-Zerlegung von A wobei L und U gegeben sind durch

$$l_{ij} = \begin{cases} a_{ij} & \text{für } i > j \\ 1 & \text{für } i = j \\ 0 & \text{für } i < j \end{cases} \qquad u_{ij} = \begin{cases} a_{ij} & \text{für } i \le j \\ 0 & \text{für } i > j \end{cases}$$

Natürlich sind auch Vorwärts- und Rückwärtselimination für Bandmatrizen einfacher. Abschließend betrachten wir noch den Spezialfall von Tridiagonalmatrizen. Dazu führen wir die folgende Notation ein.

Notation 2.26 *Für drei Vektoren $a, b, c \in \mathbb{K}^n$ mit $b_1 = c_n = 0$ ist die* **Tridiagonalmatrix** *bezüglich a, b, c gegeben durch*

$$tridiag(b, a, c) = \begin{pmatrix} a_1 & c_1 & & & & \\ b_2 & a_2 & c_2 & & & \\ & \ddots & \ddots & & & \\ & & b_{n-1} & a_{n-1} & c_{n-1} & \\ & & & b_n & a_n & \end{pmatrix}$$

Nach Satz 2.25 wissen wir, dass (falls sie existieren) die Matrizen L und U der LU-Zerlegung das folgende Aussehen haben

$$L = \begin{pmatrix} 1 & & & & \\ l_2 & 1 & & & \\ & \ddots & \ddots & & \\ & & l_{n-1} & 1 & \\ & & & l_n & 1 \end{pmatrix} \quad U = \begin{pmatrix} u_1 & c_1 & & & \\ & u_2 & c_2 & & \\ & & \ddots & \ddots & \\ & & & u_{n-1} & c_{n-1} \\ & & & & u_n \end{pmatrix} \tag{2.19}$$

wobei durch einen ersten Koeffizientenvergleich schon ausgenutzt wurde, dass die Werte $c_1, \ldots, c_{n-1}$ der oberen Nebendiagonale von A in der oberen Nebendiagonalen von U erhalten bleiben. Es sind also die Unbekannten $u_1, \ldots, u_n$ und $l_1, \ldots, l_n$ zu bestimmen. Durch Multiplikation der Matrizen L und U und erneutem Koeffizientenvergleich mit A ergeben sich die folgenden Berechnungsvorschriften:

$$\begin{aligned} \text{Start:} \quad u_1 &:= a_1 \\ \text{Für } i = 2, \ldots, n: \quad l_i &:= \frac{b_i}{u_{i-1}} \\ u_i &:= a_i - l_i c_{i-1}. \end{aligned}$$

Man kommt also mit einer in n linearen Anzahl an Operationen aus. Bei n Unbekannten ist das das beste, was man erreichen kann. Allerdings lässt sich nicht für jede Tridiagonalmatrix eine LU-Zerlegung finden. Das folgende Lemma gibt eine hinreichende Bedingung für die Durchführbarkeit der LU-Zerlegung für Tridiagonalmatrizen.

Lemma 2.27 *Für $A = tridiag(b, a, c)$ mit $B_1 = c_n = 0$ sei für $j = 1, \ldots, n$ $|c_j| < |a_j|$ und $|b_j| + |c_j| \leq |a_j|$. Dann gibt es eine LU-Zerlegung von A mit Matrizen wie in (2.19).*

Aufgabe: *Beweisen Sie das Lemma durch Induktion!*

Chapter 3

Funktionalanalytische Grundlagen: Normierte Räume und lineare Operatoren

3.1 Metrische und normierte Räume

Bevor wir uns mit der Fehleranalyse bei linearen Gleichungssystemen beschäftigen können, benötigen wir einige Begriffe aus der Funktionalanalysis. Dazu gehört insbesondere, dass wir messen können, um wie viel sich ein Vektor x von einem gestörten Vektor $\tilde{x}$ unterscheidet. Den Unterschied

$$\tilde{x} - x$$

als Vektor anzugeben, hilft uns nicht weiter, da wir zwei verschiedene gestörte Vektoren $\tilde{x}$ und x' mangels einer Ordnung im $\mathbb{K}^n$ nicht vergleichen können. Wir suchen also eine Funktion, die die Differenz zwischen zwei Vektoren durch eine reelle, positive Zahl ausdrückt. Solche Funktionen nennt man auch *Distanzfunktionen*. Mit beliebigen Distanzfunktionen geben wir uns aber nicht zufrieden, sondern betrachten *Metriken* als spezielle Distanzfunktionen.

Definition 3.1 *Sei $\mathcal{R}$ eine nichtleere Menge. Eine Abbildung $d : \mathcal{R} \times \mathcal{R} \to \mathbb{R}$ heißt* **Metrik** *auf $\mathcal{R}$ falls sie die folgenden Bedingungen erfüllt:*

(M1) $d(x, y) = 0 \iff x = y$ *für alle* $x, y \in \mathcal{R}$

(M2) $d(x, y) = d(y, x)$ *für alle* $x, y \in \mathcal{R}$ *(Symmetrie)*

(M3) $d(x, y) \leq d(x, z) + d(z, y)$ *für alle* $x, y, z \in \mathcal{R}$ *(Dreiecksungleichung)*

$(\mathcal{R}, d)$ *heißt dann* **metrischer Raum**.

Man beachte, dass aus den Metrik-Eigenschaften sofort folgt, dass

$$d(x,y) \geq 0 \text{ für alle } x,y \in \mathcal{R},$$

denn

$$d(x,y) = \frac{1}{2}(d(x,y)+d(x,y)) = \frac{1}{2}(d(x,y)+d(y,x)) \geq \frac{1}{2}d(x,x) = 0.$$

Eine Metrik ist z.B. der sogenannte *Hamming-Abstand* d_H, der für $x,y \in \mathbb{K}^n$ gegeben ist durch

$$d_H(x,y) = \#\{i = \{1,\ldots,n\} : x_i \neq y_i\}.$$

Wir wiederholen zunächst einige Begriffe, die auf jedem metrischen Raum definiert sind.

Definition 3.2 *Sei $(\mathcal{R},d)$ ein metrischer Raum.*

- *Eine Folge $(x_n) \subseteq \mathcal{R}$* **konvergiert bezüglich der Metrik** *d, falls es ein Element $\bar{x} \in \mathcal{R}$ gibt, das folgendes erfüllt: Zu jedem $\epsilon > 0$ existiert eine natürliche Zahl $N(\epsilon)$, so dass*

  $$d(\bar{x},x_n) < \epsilon \text{ für alle } n \geq N(\epsilon).$$

  In diesem Fall nennt man $\bar{x}$ den **Grenzwert** *der Folge (x_n). Eine nicht-konvergente Folge heißt* **divergent**.

- *Eine Folge $(x_n) \subseteq \mathcal{R}$ heißt* **Cauchy-Folge** *falls es zu jedem $\epsilon > 0$ eine natürliche Zahl $N(\epsilon)$ gibt, so dass*

  $$d(x_n,x_m) < \epsilon \text{ für alle } n,m \geq N(\epsilon).$$

- *Ein metrischer Raum $(\mathcal{R},d)$ heißt vollständig, falls jede Cauchy-Folge konvergiert.*

Lemma 3.3

- *Sei (x_n) eine konvergente Folge. Dann ist ihr Grenzwert eindeutig bestimmt.*

- *Jede konvergente Folge ist eine Cauchy-Folge.*

- *Es gibt metrische Räume, in denen nicht jede Cauchy-Folge konvergiert.*

Übung: *Beweisen Sie Lemma 3.3!*

Auf metrischen Räumen lassen sich weitere Strukturen erarbeiten. So reichen die Begriffe *Folge* und *Konvergenz einer Folge* insbesondere aus, um offene und abgeschlossene Mengen zu definieren. Das bedeutet, dass jeder metrische Raum auch ein topologischer Raum ist.

Die wichtigsten Beispiele für metrische Räume sind *normierte Räume*, für die wir allerdings als Grundmenge einen Vektorraum V voraussetzen.

Definition 3.4 *Sei V ein Vektorraum über einem Körper $\mathbb{K}$. Eine Abbildung $\|\cdot\| : V \to \mathbb{R}_0^+$ heißt* **Norm** *auf V falls sie die folgenden drei Bedingungen erfüllt:*

(N1) $\|x\| = 0 \iff x = 0$ *für alle $x \in V$. (Definitheit)*

(N2) $\|\alpha x\| = |\alpha|\|x\|$ *für alle $\alpha \in \mathbb{K}, x \in V$. (Skalierbarkeit)*

(N3) $\|x + y\| \leq \|x\| + \|y\|$ *für alle $x, y \in V$. (Dreiecksungleichung)*

Die Menge

$$B_{\|\cdot\|} = \{x \in V : \|x\| \leq 1\}$$

nennt man den **Einheitskreis** *der Norm $\|\cdot\|$.*
Der Raum $(V, \|\cdot\|)$ heißt **normierter Raum** *oder* **Minkowski-Raum**. *Einen vollständigen normierten Raum nennt man* **Banachraum**.

Bemerkung: Ersetzt man im Fall $\mathbb{K} = \mathbb{R}$ die Bedingung (N2) durch $\|\alpha x\| = \alpha\|x\|$ für alle $\alpha \in \mathbb{R}^+, x \in V$, so erhält man ein reelles **Minkowski-Funktional** oder einen **Gauge**.

Für Normen gilt (ähnlich wie im Fall von Metriken) dass

$$\|x\| \geq 0 \text{ für alle } x \in V,$$

denn aus (N2) folgt für $\alpha = -1$ insbesondere $\|(-1)x\| = \|x\|$, und daraus

$$\|x\| = \frac{1}{2}(\|x\| + \|x\|) = \frac{1}{2}(\|x\| + \|-x\|) \geq \frac{1}{2}(\|x + (-x)\|) = \frac{1}{2}(\|0\|) = 0.$$

Wichtige Normen auf dem $\mathbb{K}^n$ sind die folgenden:

$$\begin{aligned}
\text{Manhattan-Norm:} \quad & \|x\|_1 = \sum_{i=1}^{n} |x_i| \\
\text{Maximum-Norm:} \quad & \|x\|_\infty = \max_{i=1}^{n} |x_i| \\
\text{Euklidische Norm:} \quad & \|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{x^T x} \\
p\text{–Norm:} \quad & \|x\|_p = \sqrt[p]{\sum_{i=1}^{n} |x|_i^p}, \text{ für } 1 \leq p \leq \infty,
\end{aligned}$$

wobei die p–Norm die drei erstgenannten Normen als Spezialfälle ($p = 1$, $p = \infty$, $p = 2$) enthält.

Um einzusehen, dass es sich bei diesen Abbildungen tatsächlich um Normen handelt, sind die Bedingungen (N1),(N2) und (N3) zu zeigen. Dabei sind (N1) und (N2) direkt klar. (N3) kann man für die Fälle $p = 1$ und $p = \infty$ leicht nachrechnen; in beiden Fällen folgt die Bedingung aus der Dreiecksungleichung für Beträge. Für $p = 2$ benötigt ergibt sich (N2) aus der Cauchy-Schwarz'schen Ungleichung. Dazu führen wir Skalarprodukte ein.

Definition 3.5 *Sei V ein Vektorraum über einem Körper $\mathbb{K}$. Eine Abbildung $(\cdot,\cdot) : V \times V \to \mathbb{K}$ heißt* **Skalarprodukt** *auf V falls sie die folgenden Bedingungen erfüllt:*

(H1) $(x, x) \geq 0$ *für alle $x \in V$. (Positivität)*

(H2) $(x, x) = 0$ *genau dann wenn $x = 0$ für alle $x \in V$. (Definitheit)*

(H3) $(x, y) = \overline{(y, x)}$ *für alle $x, y \in V$. (Symmetrie)*

(H4) $(\alpha x + \beta y, z) = \alpha(x, z) + \beta(y, z)$ *für alle $x, y, z \in V$ und alle $\alpha, \beta \in \mathbb{K}$. (Linearität)*

Ein mit einem Skalarprodukt versehener Vektorraum heißt heißt **Prä-Hilbert Raum**. *Einen vollständigen Prä-Hilbert Raum nennt man* **Hilbertraum**.

Ein Skalarprodukt erfüllt die sogenannte **Antilinearität**, nämlich

$$(x, \alpha y + \beta z) = \overline{\alpha}(x, y) + \overline{\beta}(x, z) \text{ für alle } x, y, z \in V \text{ und alle } \alpha, \beta \in \mathbb{K},$$

was man direkt durch

$$\begin{aligned}(x, \alpha y + \beta z) &= \overline{(\alpha y + \beta z, x)} = \overline{\alpha(y, x) + \beta(z, x)} \\ &= \overline{\alpha}\overline{(y, x)} + \overline{\beta}\overline{(z, x)} = \overline{\alpha}(x, y) + \overline{\beta}(x, z)\end{aligned}$$

sieht. Weiterhin gilt:

$$(0, y) = (0 - 0, y) = 1(0, y) - 1(0, y) = 0 \text{ und ebenso } (x, 0) = 0.$$

Das Standard-Skalarprodukt auf dem $\mathbb{K}^n$ ist gegeben durch

$$(x, y) := \sum_{i=1}^{n} x_i \overline{y}_i.$$

Lemma 3.6 *Sei V ein Vektorraum mit einem Skalarprodukt. Dann gilt die Cauchy-Schwarz'schen Ungleichung für alle $x, y \in V$, d.h.*

$$|(x, y)|^2 \leq (x, x)(y, y).$$

Gleichheit gilt genau dann, wenn x und y linear abhängig sind.

Beweis: Für $x = 0$ folgt die Gleichheit aus $(0, y) = 0$. Sein nun $x \neq 0$. Dann rechnet man, dass

$$\begin{aligned} 0 &\leq (\alpha x + \beta y, \alpha x + \beta y) \\ &= \alpha\overline{\alpha}(x,x) + \alpha\overline{\beta}(x,y) + \overline{\alpha}\beta(y,x) + \beta\overline{\beta}(y,y) \\ &= |(x,y)|^2 - 2|(x,y)|^2 + (x,x)(y,y) = (x,x)(y,y) - |(x,y)|^2 \end{aligned}$$

wobei wir im letzten Schritt $\alpha := -\frac{\overline{(x,y)}}{\sqrt{(x,x)}}$ und $\beta := \sqrt{(x,x)}$ gesetzt haben. Gleichheit gilt genau dann, wenn $(\alpha x + \beta y, \alpha x + \beta y) = 0$, also nach (H2) genau dann wenn $\alpha x + \beta y = 0$ ist. Das kann nur eintreten, wenn x und y linear abhängig sind, und wird von den oben gesetzten Koeffizienten in diesem Fall genau erfüllt. QED

Die Cauchy-Schwarz'sche Ungleichung angewendet auf das Standard-Skalarprodukt ist ein Spezialfall der Hölder'schen Ungleichung (für den Fall $p = q = 2$):

Lemma 3.7 (Hölder'sche Ungleichung) *Sei $x, y \in \mathbb{K}^n$. Dann gilt*

$$\sum_{i=1}^{n} x_i y_i \leq \sum_{i=1}^{n} |x_i||y_i| \leq \|x\|_p \cdot \|y\|_q$$

falls entweder $1 < p, q < \infty$ und $\frac{1}{p} + \frac{1}{q} = 1$ oder falls $p = 1, q = \infty$ oder $p = \infty, q = 1$.

Mit Hilfe von Lemma 3.7 kann man die Minkowski-Ungleichung ableiten. Sie lautet wie folgt.

Lemma 3.8 (Minkowski-Ungleichung) *Für $x, y \in V$ und $1 \leq p \leq \infty$ gilt:* $\|x + y\|_p \leq \|x\|_p + \|y\|_p$

Die Minkowski-Ungleichung bestätigt, dass die p-Normen die Dreiecksungleichung erfüllen.

Normen und Skalarprodukte haben verschiedene wichtige Eigenschaften. Dazu zählt insbesondere, dass man aus jedem Skalarprodukt mittels

$$\|x\| := \sqrt{(x,x)}$$

eine Norm definieren kann und jeder Norm durch

$$d(x,y) = \|y - x\|$$

eine Metrik d zugeordnet ist. Man nennt diese Metrik dann auch die *von der Norm $\|\cdot\|$ induzierte Metrik*. Die Norm bzw. Metrik-Eigenschaften lassen sich leicht durch die Eigenschaften des zugrunde liegenden Skalarproduktes bzw. der

zugrunde liegenden Norm beweisen. Es folgt, dass jeder Prä-Hilbert Raum ein normierter Raum ist und jeder normierte Raum ist insbesondere ein metrischer Raum. Mit $\|x\| = \sqrt{(x,x)}$ können wir die Cauchy-Schwarz'sche Ungleichung auch als

$$(x,y) \leq \|x\|\|y\|$$

schreiben. Eine weitere nützlich Abschätzung ist die *Dreiecksungleichung andersherum*.

Lemma 3.9 *Sei $\|\cdot\|$ eine Norm auf V. Dann gilt für alle $x, y \in V$*

$$| \, \|x\| - \|y\| \, | \leq \|x - y\|.$$

Beweis: Für alle $x, y \in V$ gilt $\|x\| = \|x - y + y\| \leq \|x - y\| + \|y\|$. Daraus folgt

$$\|x\| - \|y\| \leq \|x - y\|.$$

Aus Symmetriegründen erhält man analog

$$\|y\| - \|x\| \leq \|x - y\|,$$

zusammen ergibt sich die Behauptung. QED

Wir haben erwähnt, wie man mit Hilfe einer Norm eine Metrik und damit Konvergenz bezüglich einer Norm definieren kann. Die Frage ist nun, in wie weit sich diese Konvergenz-Definitionen für verschiedene Normen unterscheiden. Dazu ist die folgende Definition hilfreich.

Definition 3.10 *Zwei Normen $\|\cdot\|_a$ und $\|\cdot\|_b$ auf einem Vektorraum V heißen* **äquivalent**, *wenn es positive reelle Zahlen c, C gibt, so dass für alle $x \in V$ gilt:*

$$c\|x\|_a \leq \|x\|_b \leq C\|x\|_a$$

Es lässt sich leicht zeigen, dass die in der Definition genannte Äquivalenz tatsächlich eine Äquivalenzrelation ist. Weiterhin gilt der folgende Satz:

Satz 3.11 *Zwei Normen $\|\cdot\|_a$ und $\|\cdot\|_b$ auf einem Vektorraum V sind genau dann äquivalent, wenn für jede Folge $(x_n) \subseteq V$ gilt: (x_N) konvergiert bezüglich der Norm $\|\cdot\|_a$ genau dann wenn (x_n) konvergiert bezüglich der Norm $\|\cdot\|_b$.*

Beweis:

- Nehmen wir zunächst an, dass die beiden Normen $\|\cdot\|_a$ und $\|\cdot\|_b$ äquivalent sind. Da eine Folge (x_n) genau dann gegen $\bar{x}$ konvergiert, wenn $x_n - \bar{x}$ eine Nullfolge ist, reicht es, die Aussage für Nullfolgen zu zeigen.

Sei dazu also $x_n \to 0$ eine Nullfolge bezüglich $\|\cdot\|_a$, d.h. zu jedem $\epsilon > 0$ existiert eine natürliche Zahl $N(\epsilon)$ so dass $\|x_n\|_a \leq \epsilon$ für alle $n \geq N(\epsilon)$. Wegen

$$\|x_n\|_b \leq C \cdot \|x_n\|_a \leq C\epsilon \quad \text{für alle } n \geq N(\epsilon)$$

folgt für jedes ϵ', dass $\|x_n\|_b \leq \epsilon'$ für alle $n \geq N(\frac{\epsilon'}{C})$, also ist x_n auch bezüglich $\|\cdot\|_b$ eine Nullfolge.

Die Umkehrung gilt analog.

- Gelte nun die Äquivalenz der Konvergenz-Definitionen. Durch Widerspruch zeigen wir zunächst, dass es eine Zahl $C > 0$ gibt mit

$$\|x\|_b \leq C \text{ für alle } x \in V \text{ mit } \|x\|_a = 1. \tag{3.1}$$

Angenommen also, eine solche Zahl C existiert nicht. Dann existiert zu jedem $C = C(n) := n^2$ ein x_n mit $\|x_n\|_a = 1$ und $\|x_n\|_b > n^2$. Die Folge

$$y_n := \frac{x_n}{n}$$

erfüllt also

$$\|y_n\|_a = \frac{1}{n} \text{ und } \|y_n\|_b > n.$$

Das heißt, (y_n) konvergiert gegen Null bezüglich $\|\cdot\|_a$, aber divergiert bezüglich $\|\cdot\|_b$, ein Widerspruch.

Somit gibt es ein $C > 0$, das (3.1) erfüllt. Damit ergibt sich für alle $x \in V \setminus \{0\}$:

$$\|x\|_b = \left\| \|x\|_a \frac{x}{\|x\|_a} \right\|_b = \|x\|_a \left\| \frac{x}{\|x\|_a} \right\|_b \leq C\|x\|_a,$$

die erste Ungleichung für die Normäquivalenz ist also erfüllt. Die zweite Ungleichung ergibt sich durch Vertauschen der Normen. QED

Die obige Aussage gilt für alle Vektorräume V. Wir diskutieren nun den Fall eines endlich-dimensionalen Raums.

Satz 3.12 *Sei V ein endlich-dimensionaler Vektorraum. Dann sind alle Normen über V äquivalent.*

Beweis: Sei $v_1, \ldots v_n$ eine Basis von V. Jedes Element $x \in V$ lässt sich also eindeutig darstellen durch

$$x = \sum_{k=1}^{n} \alpha_k v_k.$$

Wir konstruieren nun eine Norm (die *Maximum-Norm auf V*) und zeigen anschließend, dass jede weitere Norm auf V zu dieser Norm äquivalent ist. Wegen der Transitivität der Normäquivalenz folgt daraus die Behauptung des Satzes.

Man rechnet schnell nach, dass

$$\|x\|_\infty := \max_{k=1,\ldots,n} |\alpha_k|$$

eine Norm auf V definiert. Sei nun also $\|\cdot\|$ eine beliebige andere Norm auf V. Definiere

$$C := \sum_{k=1}^{n} \|v_k\|$$

als die Summe der Normen aller Basisvektoren. Dann folgt:

$$\begin{aligned} \|x\| &= \left\| \sum_{k=1}^{n} \alpha_k v_k \right\| \\ &\leq \sum_{k=1}^{n} |\alpha_k| \|v_k\| \text{ wegen (N2) und (N3)} \\ &\leq \sum_{k=1}^{n} \|x\|_\infty \|v_k\| \text{ weil } |\alpha_k| \leq \|x\|_\infty \\ &= C \cdot \|x\|_\infty. \end{aligned}$$

Für die andere Richtung definieren wir die gesuchte Konstante c durch

$$c := \inf\{\|x\| : x \in V \text{ und } \|x\|_\infty = 1\}.$$

Weil für alle $x \in V \setminus \{0\}$ gilt, dass

$$\left\| \frac{x}{\|x\|_\infty} \right\|_\infty = 1$$

folgt daraus, dass

$$\left\| \frac{x}{\|x\|_\infty} \right\| \geq c,$$

das heißt $\|x\| \geq c \cdot \|x\|_\infty$ für alle $x \neq 0$. Weil für $x = 0$ nichts zu zeigen ist, ergibt das also die Behauptung.

Allerdings bleibt noch zu zeigen, dass $c > 0$ gilt. Dazu führen wir einen Widerspruchsbeweis. Wir nehmen also an, dass $c = 0$. Dann gibt es eine Folge (y_m) mit $\|y_m\|_\infty = 1$ und $\|y_m\| \to 0$ für $m \to \infty$. Die Basisdarstellung in V liefert für jedes Folgenglied y_m

$$y_m = \sum_{k=1}^{n} \alpha_{km} v_k,$$

und damit n Folgen für die Koeffizienten $\alpha_{1m}, \alpha_{2m}, \ldots, \alpha_{nm}$ aus dem zugrunde liegenden Körper. Weil $\|y_m\|_\infty = 1$ für alle m gelten die folgenden beiden Aussagen für die Koeffizienten der Folgen:

$$\text{Für alle } m : |\alpha_{km}| \leq 1 \text{ für alle } k = 1, \ldots, n. \tag{3.2}$$

$$\text{Für alle } m \text{ existiert ein } k \in \{1, \ldots, n\} \text{ so dass } |\alpha_{km}| = 1. \tag{3.3}$$

Wegen (3.3) erfüllt mindestens eine der Koeffizienten-Folgen $\bar{k}$, dass $\#\{m : |\alpha_{\bar{k}m}| = 1\} = \infty$, d.h. es kommen unendlich viele Einsen (oder unendlich viele − Einsen) vor. Sei oBdA $\bar{k} = 1$, und $\#\{m : \alpha_{1m} = 1\} = \infty$. Wähle dann eine Teilfolge der $(y_m^{(1)}) \subseteq (y_m)$, in der die Koeffizienten-Teilfolge bezüglich der Koeffizienten α_{1m} des ersten Basisvektors nur aus Einsen besteht.
Weiterhin sind wegen (3.2) alle der n Koeffizienten-Folgen beschränkt. Nach dem Satz von Bolzano-Weierstrass wählen wir nun eine Teilfolge $(y_m^{(2)}) \subseteq (y_m^{(1)})$ für die die zweite Koeffizienten-Folge a_{2m} eine konvergente Teilfolge ist. Aus den Indizes dieser Folge wählen wir wiederum eine bezüglich der dritten Koeffizienten-Folge a_{3m} konvergente Teilfolge und so weiter, bis wir eine Teilfolge

$$y_m^{(n)} = \sum_{k=1}^{n} \alpha'_{km} v_k,$$

erhalten, für die alle Koeffizienten-Folgen konvergieren, d.h.

$$\begin{aligned} \alpha'_{1m} &\to \alpha_1 = 1 \\ \alpha'_{2m} &\to \alpha_2 \\ &\vdots \\ \alpha'_{nm} &\to \alpha_n. \end{aligned}$$

Nach Konstruktion wissen wir, dass (a'_{1m}) nur aus Einsen besteht und also gegen 1 konvergiert. Für

$$y = \sum_{k=1}^{n} \alpha_k v_k$$

gilt dann nach Teil 1 dieses Beweises

$$\|y_m^{(n)} - y\| \leq C\|y_m^{(n)} - y\|_\infty = \max_{k=1,\ldots,n} \{|\alpha'_{km} - \alpha_k|\} \to 0 \text{ für } m \to \infty.$$

Weil $\|y_m\| \to 0$ folgt daraus $y = 0$, ein Widerspruch zum Grenzwert $\alpha_1 = 1$ der ersten Koeffizienten-Folge. QED

Der gerade bewiesene Satz zeigt, dass es auf dem $\mathbb{K}^n$ nicht darauf ankommt, bezüglich welcher Norm man von Konvergenz redet. Genauso induzieren alle Normen auf dem $\mathbb{K}^n$ die gleiche Topologie: Begriffe wie Abgeschlossenheit, Beschränktheit und Kompaktheit hängen also nicht von der Wahl der Norm ab. Allerdings sollte man beachten, dass die Konstanten c, C nicht nur von den jeweiligen Normen, sondern auch von der Dimension des Raumes n abhängen. Weiterhin darf man nicht vergessen, dass der Satz nur für endlich-dimensionale Vektorräume gilt; auf Räume mit unendlicher Dimension (z.B. Funktionenräume) lässt er sich im allgemeinen nicht übertragen.

Als Beispiel wollen wir abschließend noch die p-Normen auf dem Raum der stetigen Funktionen $C[a,b]$ über einem Intervall $[a,b]$ angeben. Für eine stetige Funktion $f : [a,b] \to \mathbb{K}$ definiert man

$$\|f\|_{L_p[a,b]} := \begin{cases} \left(\int_a^b |f(x)|^p dx\right)^{\frac{1}{p}} & \text{falls } 1 \leq p < \infty \\ \max_{x\in[a,b]} |f(x)| & \text{falls } p = \infty \end{cases}$$

3.2 Normen für Abbildungen und Matrizen

Definition 3.13 *Es seien $(V, \|\cdot\|_V)$ und $(W, \|\cdot\|_W)$ zwei normierte Räume und $F : V \to W$ eine lineare Abbildung. Dann heißt F* **beschränkt**, *falls es eine Konstante $C > 0$ gibt, sodass für alle $v \in V$:*

$$\|F(v)\|_W \leq C\|v\|_V.$$

Wir untersuchen zunächst die Stetigkeit solcher linearen Abbildungen.

Lemma 3.14 *Sei $F : V \to W$ eine lineare Abbildung zwischen normierten Vektorräumen. Dann ist F genau dann beschränkt, wenn F stetig ist.*

Beweis:

- Ist F beschränkt, so folgt aus

  $$\|F(v) - F(w)\|_W = \|F(v-w)\|_W \leq C\|v-w\|_v$$

  direkt die Stetigkeit von F.

- Ist F stetig, so gibt es zu jedem $\epsilon > 0$ ein $\delta > 0$ so, dass $\|F(v)-F(w)\|_W < \epsilon$ für alle v, w mit $\|v-w\|_V \leq \delta$. Für $\epsilon = 1$ und $w = 0$ erhält man wegen $F(0) = 0$ also ein $\delta > 0$ so, dass

  $$\|F(v)\|_W < 1 \text{ für alle } \|v\|_V \leq \delta.$$

  Für jedes $v \in V \setminus \{0\}$ gilt

  $$\begin{aligned} & \left\| \delta \frac{v}{\|v\|_V} \right\|_V = \delta \\ \Longrightarrow \quad & \left\| F\left(\delta \frac{v}{\|v\|_V}\right) \right\|_W \leq 1 \\ \Longrightarrow \quad & \|F(v)\|_W = \frac{\|v\|_V}{\delta} \left\| F\left(\delta \frac{v}{\|v\|_V}\right) \right\|_W \leq \frac{1}{\delta}\|v\|_V, \end{aligned}$$

  also folgt die Beschränktheit mit $C = \frac{1}{\delta}$. QED

Zwischen endlich-dimensionalen Räumen stellt sich die Situation noch einfacher dar.

Lemma 3.15 *Sei $F : V \to W$ eine lineare Abbildung zwischen zwei normierten endlich-dimensionalen Vektorräumen. Dann ist F beschränkt und stetig.*

Beweis: Sei $F : V \to W$ linear und sei $v_1, \ldots, v_n$ eine Basis von V. Dann gilt

$$\begin{aligned}
& v = \sum_{k=1}^{n} \alpha_k v_k \\
\Longrightarrow \quad & F(v) = F\left(\sum_{k=1}^{n} \alpha_k v_k\right) = \sum_{k=1}^{n} \alpha_k F(v_k) \\
\Longrightarrow \quad & \|F(v)\|_W = \left\| \sum_{k=1}^{n} \alpha_k F(v_k) \right\|_W \leq \sum_{k=1}^{n} |\alpha_k| \|F(v_k)\|_W \\
& \qquad \leq \max_{k=1\ldots n} \|F(v_k)\|_W \sum_{k=1}^{n} |\alpha_k| = \max_{k=1\ldots n} \|F(v_k)\|_W \|v\|_1 \\
& \qquad \leq \underbrace{\max_{k=1\ldots n} \|F(v_k)\|_W \cdot C}_{=:C'} \|v\|_V,
\end{aligned}$$

wobei beim letzten Schritt ausgenutzt wurde, dass nach Satz 3.12 alle Normen auf V äquivalent sind. Damit ist F also beschränkt und nach Lemma 3.14 auch stetig. QED

Auf dem Raum der beschränkten linearen Abbildungen zwischen zwei normierten Vektorräumen definieren wir nun folgende Norm.

Definition 3.16 *Es seien $(V, \|\cdot\|_V)$ und $(W, \|\cdot\|_W)$ zwei normierte Räume. Für eine beschränkte lineare Abbildung $F : V \to W$ definiert man die zu $\|\cdot\|_V$ und $\|\cdot\|_W$* **zugeordnete Norm** *durch*

$$\|F\|_{V,W} := \sup_{v \in V \setminus \{0\}} \frac{\|F(v)\|_W}{\|v\|_V}.$$

Gilt $V = W$ und $\|\cdot\|_V = \|\cdot\|_W$ so schreiben wir auch $\|F\|_V$ statt $\|F\|_{V,W}$.

Weil F als beschränkt vorausgesetzt wurde gilt für alle $v \in V \setminus \{0\}$

$$\frac{\|F(v)\|_W}{\|v\|_V} \leq \frac{C\|v\|_V}{\|v\|_V} = C.$$

Wir erhalten also $\|F\|_{V,W} < C < \infty$. Die Norm der Abbildung F ist also die kleinstmögliche Konstante C, mit der man die Beschränktheit der Abbildung abschätzen kann.

Wir erwähnen noch, dass wir auch wirklich von Normen sprechen dürfen:

Satz 3.17 *Es seien $(V, \|\cdot\|_V)$ und $(W, \|\cdot\|_W)$ zwei normierte Räume. Dann ist $\|\cdot\|_{V,W}$ eine Norm auf dem Raum der beschränkten linearen Abbildungen von $V \to W$.*

Aufgabe: *Beweisen Sie Satz 3.17!*

Folgende Umformulierung erweist sich als nützlich.

Lemma 3.18 *Es seien $(V, \|\cdot\|_V)$ und $(W, \|\cdot\|_W)$ zwei normierte Räume und $F : V \to W$ eine beschränkte lineare Abbildung. Dann gilt*

$$\|F\|_{V,W} = \sup_{v \in V : \|v\|_V = 1} \|F(v)\|_W. \tag{3.4}$$

Beweis: Zunächst ist klar, dass

$$\sup_{v \in V : \|v\|_V = 1} \|F(v)\|_W \leq \|F\|_{V,W}.$$

Um $\sup_{v \in V : \|v\|_V = 1} \|F(v)\|_W \geq \|F\|_{V,W}$ zu zeigen, bemerken wir, dass wegen der Skalierbarkeit (Eigenschaft (N2)) der Norm $\|\cdot\|_W$ für alle $v \neq 0$, $v \in V$ gilt:

$$\frac{\|F(v)\|_W}{\|v\|_V} = \frac{1}{\|v\|_V} \|F(v)\|_W = \left\| F\left(\frac{v}{\|v\|_V} \right) \right\|_W.$$

Es gibt also zu jedem $v \neq 0$ ein u mit $\|u\|_V = 1$ so dass $\frac{\|F(v)\|_W}{\|v\|_V} = \|F(u)\|_W$. Entsprechend folgt

$$\sup_{v \neq 0} \frac{\|F(v)\|_W}{\|v\|_V} \leq \sup_{u : \|u\|_V = 1} \|F(u)\|_W$$

und zusammen ergibt sich die Behauptung. QED

Betrachte nun ein beliebiges $v \in V \setminus \{0\}$. Dann gilt:

$$\frac{\|F(v)\|_W}{\|v\|_V} \leq \sup_{v' \in V \setminus \{0\}} \frac{\|F(v')\|_W}{\|v'\|_V} = \|F\|_{V,W},$$

woraus wir wegen $F(0) = 0$ und $\|0\|_V = \|0\|_W = 0$

$$\|F(v)\|_W \leq \|F\|_{V,W} \cdot \|v\|_V \text{ für alle } v \in V \tag{3.5}$$

folgern. Wir sagen auch, $\|\cdot\|_{V,W}$ ist *passend* zu den Normen $\|\cdot\|_V$ und $\|\cdot\|_W$. Diese Eigenschaft wird später noch wichtig werden. Eine Verallgemeinerung ist die folgende.

Definition 3.19 *Es seien $(V, \|\cdot\|_V)$ und $(W, \|\cdot\|_W)$ zwei normierte Räume. Eine Norm $\|\cdot\|$ auf dem Raum der beschränkten, linearen Abbildungen von V nach W heißt* **zu den Normen $\|\cdot\|_V$ und $\|\cdot\|_W$ passend**, *oder* **mit den Normen $\|\cdot\|_V$ und $\|\cdot\|_W$ verträglich**, *falls für alle $v \in V$ gilt:*

$$\|F(v)\|_W \leq \|F\| \cdot \|v\|_V.$$

Gleichung (3.5) zeigt, dass die Norm $\|\cdot\|_{V,W}$ immer zu ihren natürlichen oder zugeordneten Normen $\|\cdot\|_V$ und $\|\cdot\|_W$ passt.

Aufgabe: *Seien $(U, \|\cdot\|_U)$,$(V, \|\cdot\|_V)$ und $(W, \|\cdot\|_W)$ normierte endlich-dimensionale Vektorräume und seien $F : U \to V$ und $G : V \to W$ beschränkte lineare Abbildungen. Zeigen Sie, dass dann für $G \circ F : U \to W$ gilt:*

$$\|G \circ F\|_{UW} \leq \|G\|_{VW} \|F\|_{UV}.$$

Wir möchten nun den Fall linearer Abbildungen zwischen den endlich-dimensionalen Vektorräumen

$$A : \mathbb{K}^n \to \mathbb{K}^m$$

genauer untersuchen. Jede lineare Abbildung kann dann durch eine Matrix A repräsentiert werden, so dass wir die zugehörige Norm $\|A\|_{V,W}$ in diesem Fall auch *Matrixnorm* nennen.

Im folgenden entwickeln wir Formeln für einige Matrixnormen, die aus den wichtigsten Normen auf dem $\mathbb{K}^n, \mathbb{K}^m$ entstehen.

Satz 3.20 *Sei $A \in \mathbb{K}^{m,n}$ eine lineare Abbildung vom $\mathbb{K}^n$ in den $\mathbb{K}^m$.*

1. *Betrachte $(\mathbb{K}^n, \|\cdot\|_1)$ und $(\mathbb{K}^m, \|\cdot\|_1)$ jeweils mit Manhattan-Norm. Dann heißt die zugehörige Matrixnorm Spaltensummennorm und sie ist gegeben durch*

$$\|A\|_1 = \sup_{x \in \mathbb{K}^n : \|x\|_1 = 1} \|Ax\|_1 = \max_{k=1,\ldots,n} \sum_{i=1}^{m} |a_{ik}|.$$

2. *Betrachte $(\mathbb{K}^n, \|\cdot\|_\infty)$ und $(\mathbb{K}^m, \|\cdot\|_\infty)$ jeweils mit Maximum-Norm. Dann heißt die die zugehörige Matrixnorm Zeilensummennorm und sie ist gegeben durch*

$$\|A\|_\infty = \sup_{x \in \mathbb{K}^n : \|x\|_\infty = 1} \|Ax\|_\infty = \max_{i=1,\ldots,m} \sum_{k=1}^{n} |a_{ik}|.$$

Beweis:

ad 1: Für alle $x \in \mathbb{K}^n$ gilt zunächst, dass

$$\begin{aligned}\|Ax\|_1 &= \sum_{i=1}^{m} |(Ax)_i| = \sum_{i=1}^{m} \left| \sum_{k=1}^{n} a_{ik} x_k \right| \\ &\leq \sum_{k=1}^{n} |x_k| \sum_{i=1}^{m} |a_{ik}| \leq \left(\max_{k=1,\ldots,n} \sum_{i=1}^{m} |a_{ik}| \right) \sum_{k=1}^{n} |x_k| \\ &= \left(\max_{k=1,\ldots,n} \sum_{i=1}^{n} |a_{ik}| \right) \|x\|_1.\end{aligned}$$

Damit gilt also

$$\|A\|_1 \leq \max_{k=1,\ldots,n} \sum_{i=1}^{m} |a_{ik}|.$$

Um $\|A\|_1 \geq \max_{k=1,\ldots,n} \sum_{i=1}^{m} |a_{ik}|$ zu zeigen, wählen wir j so dass

$$\sum_{i=1}^{m} |a_{ij}| = \max_{k=1,\ldots,n} \sum_{i=1}^{m} |a_{ik}|.$$

Für den $j$ten Einheitsvektor e_j gilt dann

$$\|Ae_j\|_1 = \|A_j\|_1 = \sum_{i=1}^{m} |a_{ij}| = \max_{k=1,\ldots,n} \sum_{i=1}^{m} |a_{ik}|.$$

Für die Norm von A folgt (mit (3.4)) daraus

$$\|A\|_1 = \sup_{x:\|x\|_1=1} \|Ax\|_1 \geq \|Ae_j\|_1 = \max_{k=1,\ldots,n} \sum_{i=1}^{m} |a_{ik}|.$$

ad 2: Für die Maximums-Norm erhalten wir analog für $x \in \mathbb{K}^n$

$$\begin{aligned}\|Ax\|_\infty &= \max_{i=1,\ldots,m} |(Ax)_i| = \max_{i=1,\ldots,m} \left| \sum_{k=1}^{n} a_{ik} x_k \right| \\ &\leq \max_{i=1,\ldots,m} \sum_{k=1}^{n} |a_{ik}||x_k| \leq \max_{i=1,\ldots,m} \sum_{k=1}^{n} |a_{ik}| \|x\|_\infty,\end{aligned}$$

also

$$\|A\|_\infty \leq \max_{i=1,\ldots,m} \sum_{k=1}^{n} |a_{ik}|.$$

Für die "$\geq$" Richtung wählen wir hier den Index j als den der Zeile mit maximaler Summe, d.h. so dass

$$\sum_{k=1}^{n} |a_{jk}| = \max_{i=1,\ldots,m} \sum_{k=1}^{n} |a_{ik}|.$$

Weiterhin wählen wir einen Vektor $z \in \mathbb{K}^n$ passend zum Index j durch

$$z_k = \begin{cases} \frac{\bar{a}_{jk}}{|a_{jk}|} & \text{falls } a_{jk} \neq 0 \\ 1 & \text{falls } a_{jk} = 0 \end{cases}$$

Dann gilt

a) $\|z\|_\infty = 1$, und

b) $a_{jk} z_k = \frac{a_{jk}\bar{a}_{jk}}{|a_{jk}|} = |a_{jk}|$ für $a_{jk} \neq 0$, insbesondere ist $a_{jk} z_k$ für alle $k = 1, \ldots, n$ positiv und reell.

Für die Norm von Az erhalten wir daraus, dass

$$\begin{aligned} \|Az\|_\infty &= \max_{i=1,\ldots,m} |(Az)_i| = \max_{i=1,\ldots,m} \left| \sum_{k=1}^{n} a_{ik} z_k \right| \\ &= \max_{i=1,\ldots,m} \sum_{k=1}^{n} |a_{ik} z_k| = \max_{i=1,\ldots,m} \sum_{k=1}^{n} |a_{ik}|. \end{aligned}$$

Wie für die Manhattan-Norm folgern wir daraus, dass

$$\|A\|_\infty \geq \max_{i=1,\ldots,m} \sum_{k=1}^{n} |a_{ik}|.$$

QED

Wir betrachten jetzt noch die Matrixnorm $\|A\|_2$. Dazu benötigen wir den auch in anderen Bereichen der Numerik wichtigen Begriff des *Spektralradius* einer Matrix.

Definition 3.21 *Sei $A \in \mathbb{K}^{n,n}$.*

- *$\lambda \in \mathbb{K}$ heißt* **Eigenwert** *von A falls es ein $v \in \mathbb{K}^n \setminus \{0\}$ gibt, so dass*

  $$Av = \lambda v.$$

  v heißt dann **Eigenvektor** *von A bezüglich des Eigenwertes λ.*

- *Der* **Spektralradius** *$\rho(A)$ einer Matrix A ist der betragsmäßig größte Eigenwert von A, d.h.*

  $$\rho(A) = \max\{|\lambda| : \lambda \in \mathbb{C} \text{ ist Eigenwert von } A\}.$$

Wir müssen zunächst an die folgenden Begriffe aus der linearen Algebra erinnern:

Notation 3.22

- *Eine Matrix $A \in \mathbb{R}^{n,n}$ heißt orthogonal, falls $A^T A = I$ beziehungsweise $A^{-1} = A^T$.*
- *Eine Matrix $A \in \mathbb{C}^{n,n}$ heißt unitär, falls $A^{-1} = \bar{A}^T$.*

Bemerkung: Die Spalten $A_1, \ldots, A_n$ von A von orthogonalen oder unitären Matrizen bilden eine Orthonormalbasis des $\mathbb{K}^n$. Das sieht man, indem man das Produkt $B = \bar{A}^T A$ durch Produkte der Spalten von A beschreibt. Weil B die Einheitsmatrix ist, gilt für das Element

$$b_{ij} = \bar{A}_i^T A_j = \begin{cases} 1 & \text{falls } i = j \\ 0 & \text{sonst,} \end{cases}$$

entsprechend folgt die Behauptung.
Folgenden Satz werden wir verwenden.

Satz 3.23 (Hauptachsentransformation) *Sei $A \in \mathbb{K}^{n,n}$ eine symmetrische (bzw. hermitesche) Matrix. Dann gibt es eine reguläre orthogonale (bzw. unitäre) Matrix $Q \in \mathbb{R}^{n,n}$ (bzw. $Q \in \mathbb{C}^{n,n}$) und eine Diagonalmatrix $D = diag(d_1, \ldots, d_n) \in \mathbb{R}^{n,n}$ so dass*

$$A = QDQ^{-1}.$$

Dabei sind $d_1, \ldots, d_n$ die Eigenwerte der Matrix A, und die Spalten von Q bilden eine Orthonormalbasis, die aus den zugehörigen Eigenvektoren besteht. Das heißt, es gilt

$$AQ_j = d_j Q_j \text{ für } j = 1, \ldots, n$$

Wir beweisen folgende Folgerung aus Satz 3.23.

Lemma 3.24 *Sei $A \in \mathbb{K}^{n,n}$ eine symmetrische (bzw. hermitesche) positiv semi-definite Matrix, und sei $\lambda^{\min}$ ihr betragsmäßig kleinster und $\rho(A) = \lambda^{\max}$ ihr betragsmäßig größter Eigenwert. Dann gilt $\lambda^{\min} \geq 0$ und alle $x \in \mathbb{K}^n$ erfüllen die folgende Abschätzung:*

$$\lambda^{\min} \|x\|_2^2 \leq \bar{x}^T A x \leq \lambda^{\max} \|x\|_2^2.$$

Beweis: Weil A positiv semi-definit ist, sind die Eigenwerte $\lambda_1, \ldots, \lambda_n$ von A nicht-negativ (siehe Lemma 2.18 auf Seite 37). Sei nach Satz 3.23 weiter $v_1, \ldots, v_n$ eine Orthonormalbasis des $\mathbb{K}^n$, die aus Eigenvektoren von A besteht. Wir schreiben $x = \sum_{i=1}^n \alpha_i v_i$ und rechnen wegen

$$Ax = A \sum_{i=1}^n \alpha_i v_i = \sum_{i=1}^n \alpha_i A(v_i) = \sum_{i=1}^n \alpha_i \lambda_i v_i$$

nach, dass

$$\bar{x}^T A x = \sum_{j,k=1}^n \bar{\alpha}_j \alpha_k \lambda_k \bar{v}_j^T v_k = \sum_{j=1}^n |\alpha_j|^2 \lambda_j,$$

wobei letztere Gleichheit aus der Orthogonalität der v_i folgt. Weiterhin gilt $\|x\|_2^2 = x^T x = \sum_{i=1}^n |\alpha_i|^2$ und entsprechend folgt

$$\lambda^{\min} \sum_{j=1}^n |\alpha_j|^2 \leq \sum_{j=1}^n |\alpha_j|^2 \lambda_j \leq \lambda^{\max} \sum_{j=1}^n |\alpha_j|^2,$$

zusammen also

$$\lambda^{\min} \|x\|_2^2 \leq \bar{x}^T A x \leq \lambda^{\max} \|x\|_2^2.$$

QED

Wir können nun endlich auch die Matrixnorm $\|A\|_2$ bezüglich der Euklidischen Norm berechnen.

Satz 3.25 *Für $A : \mathbb{K}^n \to \mathbb{K}^m$, also $A \in \mathbb{K}^{m,n}$ gilt*

$$\|A\|_2 = \sup_{x \in \mathbb{K}^n : x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sqrt{\rho(\bar{A}^T A)}$$

Man nennt $\|A\|_2$ auch die **Spektralnorm** *von A.*

Beweis: Zunächst gilt, dass $\bar{A}^T A \in \mathbb{K}^{n,n}$ eine hermitesche und positiv semidefinite Matrix ist. Daher sind alle ihre Eigenwerte größer oder gleich Null. Sei $\rho(\bar{A}^T A) = \lambda^{\max}$ der größte Eigenwert von $\bar{A}^T A$. Es gilt

$$\|Ax\|_2^2 = \overline{(Ax)}^T (Ax) = \bar{x}^T \bar{A}^T A x \leq \lambda^{\max} \|x\|_2^2,$$

wobei die letzte Ungleichung aus Lemma 3.24 folgt. Die Ungleichung ergibt also

$$\|A\|_2 \leq \sqrt{\rho(\bar{A}^T A)}.$$

Um Gleichheit zu zeigen, wählen wir z als Eigenvektor zu $\lambda^{\max}$ und erhalten

$$\|Az\|_2^2 = \bar{z}^T \bar{A}^T A z = \bar{z}^T \lambda^{\max} z = \lambda^{\max} \bar{z}^T z = \lambda^{\max} \|z\|_2^2.$$

Daraus ergibt sich analog zu dem Beweis von Satz 3.20, dass

$$\|A\|_2 = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} \geq \frac{\|Az\|_2}{\|z\|_2} = \sqrt{\lambda^{\max}} = \sqrt{\rho(\bar{A}^T A)}.$$

QED

Leider ist die Spektralnorm für größere Matrizen aufwändig zu berechnen. Daher ersetzt man sie manchmal durch eine der folgenden Normen:

Definition 3.26

$$\textit{Gesamtnorm:} \qquad \|A\|_G := n \max\{|a_{ij}| : 1 \leq i \leq m, 1 \leq j \leq n\}$$

$$\textit{Frobenius-Norm:} \qquad \|A\|_F := \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}$$

Beides sind wirklich Normen (da sie bis auf Vorfaktoren mit $\|\cdot\|_\infty$ beziehungsweise mit $\|\cdot\|_2$ auf dem $\mathbb{K}^{n \cdot m}$ übereinstimmen).

Lemma 3.27

1. *Die Norm $\|\cdot\|_G$ ist passend zu $\|\cdot\|_\infty$.*
2. *Die Norm $\|\cdot\|_F$ ist passend zu $\|\cdot\|_2$.*

Beweis: *Übung!*

3.3 Kondition

Zum Abschluss dieses Kapitels wollen wir die gewonnen Erkenntnisse anwenden, um die *Kondition* einer Matrix zu definieren. Diese wird uns helfen, die Übertragung von Fehlern abzuschätzen.

Betrachten wir dazu ein lineares Gleichungssystem $Ax = b$ mit folgenden Fehlern in den Eingangsdaten:

- ΔA sei der Fehler in der Matrix A,
- sowie Δb der Fehler im Ergebnisvektor b.

Lässt sich anhand dieser Daten der Fehler im Ergebnis abschätzen?

Um diese Frage zu beantworten, bemerken wir zunächst, dass

$$\Delta x = \tilde{x} - x$$

ist, wobei x als exakte Lösung des Gleichungssystems $Ax = b$ und $\tilde{x}$ durch die gewonnene Lösung

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

definiert ist. Es gilt also

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

Multipliziert man diese Gleichung aus und verwendet $Ax = b$ so ergibt sich

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

Nehmen wir nun zunächst an, dass die gestörte Matrix $A+\Delta A$ invertierbar wäre. Dann könnte man nach Δx auflösen und dadurch die Norm von x abschätzen, also

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

wobei wir eine submultiplikative (d.h. $\|AB\| \leq \|A\| \cdot \|B\|$) und zur Vektornorm passende Matrixnorm gewählt haben. Der relative Fehler ergibt sich entsprechend als

$$\begin{aligned} \frac{\|\Delta x\|}{\|x\|} &\leq \|(A+\Delta A)^{-1}\| \left(\frac{\|\Delta b\|}{\|x\|} + \|\Delta A\| \right) \\ &= \|(A+\Delta A)^{-1}\|\|A\| \left(\frac{\|\Delta b\|}{\|A\|\|x\|} + \frac{\|\Delta A\|}{\|A\|} \right) \\ &\leq \|(A+\Delta A)^{-1}\|\|A\| \left(\frac{\|\Delta b\|}{\|Ax\|} + \frac{\|\Delta A\|}{\|A\|} \right) \\ &= \underbrace{\|(A+\Delta A)^{-1}\|\|A\|}_{\text{Vergrößerungsfaktor}} \left(\underbrace{\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|}}_{\text{relative Fehler der Eingangsdaten}} \right) \end{aligned} \tag{3.6}$$

Bevor wir den Term des Vergrößerungsfaktors weiter abschätzen, beschäftigen wir uns mit der Frage, wann die Inverse von $A+\Delta A$ existiert.

Lemma 3.28 *Seien $A, \Delta A \in \mathbb{K}^{n,n}$, A regulär und $\|A^{-1}\|\|\Delta A\| < 1$, wobei $\|\cdot\|$ eine zu einer Vektornorm passende submultiplikative Matrixnorm ist, die $\|I\| = 1$ erfüllt. Dann ist $A+\Delta A$ regulär, und es gilt*

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

Beweis: Schreibe

$$\begin{aligned} & x = A^{-1}(A+\Delta A)x - A^{-1}(\Delta A)x \\ \Longrightarrow \quad & \|x\| \leq \|A^{-1}\|\|(A+\Delta A)x\| + \|A^{-1}\|\|\Delta A\|\|x\| \\ \Longrightarrow \quad & \|x\| \underbrace{(1 - \|A^{-1}\|\|\Delta A\|)}_{>0 \text{ nach Vor.}} \leq \|A^{-1}\|\|(A+\Delta A)x\|. \end{aligned}$$

Also folgt aus $(A+\Delta A)x = 0$ dass $\|x\| = 0$ und entsprechend auch $x = 0$. Die Abbildung $A+\Delta A$ ist somit injektiv und damit auch surjektiv, also ist die Matrix $A+\Delta A$ invertierbar.

Wir können also $B := (A + \Delta A)^{-1}$ definieren. Um die im Lemma genannte Abschätzung zu erhalten, rechnen wir nach

$$\begin{aligned} 1 &= \|I\| = \|B(A+\Delta A)\| = \|BA + BAA^{-1}\Delta A\| \\ &\geq \|BA\| - \|BA\|\|A^{-1}\|\|\Delta A\| \\ &= \|BA\| \underbrace{(1 - \|A^{-1}\|\|\Delta A\|)}_{>0}. \end{aligned}$$

Daraus erhalten wir

$$\|BA\| \leq \frac{1}{1 - \|A^{-1}\|\|\Delta A\|}$$

und schließlich

$$\|(A+\Delta A)^{-1}\| = \|BAA^{-1}\| \leq \|BA\|\|A^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|\Delta A\|}.$$

QED

Definition 3.29 *Für eine Matrix $A \in \mathbb{K}^{n,n}$ definieren wir*

$$\mathrm{cond}(A) := \|A\| \ \|A^{-1}\|$$

als die **Kondition** *von A.*

Wozu man diese Definition verwenden kann, zeigt der folgende Satz und das anschließende Korollar.

Satz 3.30 *Sei $\|\cdot\|$ eine Matrixnorm wie in Lemma 3.28. Sei $\|b\| \neq 0$ und $\|A^{-1}\| \ \|\Delta A\| < 1$. Sei $Ax = b$. Dann gilt für jede gestörte Lösung $\tilde{x} = x + \Delta x$ des gestörten Systems*

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

die folgende Abschätzung:

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

Zunächst bemerken wir, dass der Ausdruck wegen

$$1 > 1 - \mathrm{cond}(A)\frac{\|\Delta A\|}{\|A\|} = 1 - \|A^{-1}\| \ \|\Delta A\| > 0$$

wohldefiniert ist. Man sieht hier auch schon, dass eine kleinere Kondition zu kleineren relativen Fehlern führen wird.

Beweis: Aus (3.6) und der Abschätzung aus Lemma 3.28 folgt, dass

$$\begin{aligned}\frac{\|\Delta x\|}{\|x\|} &\leq \frac{\|A^{-1}\|}{1-\|A^{-1}\|\,\|\Delta A\|}\cdot\|A\|\left(\frac{\|\Delta A\|}{\|A\|}+\frac{\|\Delta b\|}{\|b\|}\right)\\ &= \frac{\mathrm{cond}(A)}{1-\mathrm{cond}(A)\frac{\|\Delta A\|}{\|A\|}}\left(\frac{\|\Delta A\|}{\|A\|}+\frac{\|\Delta b\|}{\|b\|}\right).\end{aligned}$$

QED

Eine einfache und oft betrachtete Anwendung dieses Ergebnisses ist das folgende Korollar, das man auch direkt aus (3.6) ohne die Voraussetzungen aus Lemma 3.28 herleiten kann.
Korollar: Hat man nur eine Störung in b (ist also $\Delta A = 0$), so übertragen sich die Fehler in b mit maximal der Kondition von A. Genauer:

$$\frac{\|\Delta x\|}{\|x\|} \leq \mathrm{cond}(A)\frac{\|\Delta b\|}{\|b\|}.$$

Diese Aussage ergibt sich direkt aus Satz 3.30, da die Voraussetzung $\|A^{-1}\|\,\|\Delta A\| = 0 < 1$ erfüllt ist.

Abschließend geben wir noch zwei nützliche Aussagen zur Bestimmung der Kondition einer Matrix an.

Lemma 3.31 *Für jede zu einer Vektornorm passend gewählte Matrixnorm und jede invertierbare Matrix A gilt: cond$(A) \geq 1$.*

Beweis: Für den Beweis verwenden wir die Definition für *passend* für A und A^{-1} in folgendem Sinn. Sei $x \neq 0$. Dann gilt $A^{-1}x \neq 0$ und entsprechend

$$\begin{aligned}\|A^{-1}x\| &\leq \|A^{-1}\|\,\|x\|\\ \|A(A^{-1}x)\| &\leq \|A\|\,\|(A^{-1}x)\|\end{aligned}$$

Zusammen ergibt sich

$$\|A\|\,\|A^{-1}\| \geq \|A\|\frac{\|A^{-1}x\|}{\|x\|} \geq \frac{\|A(A^{-1}x)\|}{\|A^{-1}x\|}\frac{\|A^{-1}x\|}{\|x\|} = \frac{\|x\|}{\|x\|} = 1.$$

QED

Für die der Euklidischen Norm zugeordnete Spektralnorm gelten die folgenden Aussagen.

Lemma 3.32 *Sei Q eine orthogonale (unitäre) Matrix. Dann gilt*

1. *$cond_2(Q) = 1$, und*

2. *$cond_2(QA) = cond_2(A) = cond_2(AQ)$ für alle Matrizen A, das heißt die Multiplikation mit Q ändert die Kondition der Matrix A nicht.*

Beweis:

1.

$$\begin{aligned} \mathrm{cond}_2(Q) &= \|Q\|_2\|Q^{-1}\|_2 = \sqrt{\rho(\bar{Q}^T Q)}\sqrt{\rho(\bar{Q}^{-1^T} Q^{-1})} \\ &= \sqrt{\rho(I)}\sqrt{\rho(Q\bar{Q}^T)} = 1 \cdot \sqrt{\rho(I)} = 1 \end{aligned}$$

2.

$$\begin{aligned} \|A\|_2 &= \|\bar{Q}^T QA\|_2 \leq \|\bar{Q}^T\|_2 \, \|QA\|_2 = \|QA\|_2 \\ &\leq \|Q\|_2 \, \|A\|_2 = \|A\|_2, \end{aligned}$$

also ist $\|A\|_2 = \|QA\|_2$. Analog ergibt sich $\|A\|_2 = \|AQ\|_2$, also erhält man $\mathrm{cond}_2(QA) = \mathrm{cond}_2(A) = \mathrm{cond}_2(AQ)$.

QED

Chapter 4

Lineare Gleichungssysteme: Orthogonalisierungsverfahren

4.1 Die QR-Zerlegung

Bisherige Lösung von Gleichungssystemen:

$$A \to L \cdot A = \begin{pmatrix} \ddots & & * \\ & \ddots & \\ 0 & & \ddots \end{pmatrix}$$

Dabei galt für die Kondition von $(L \cdot A)$:

$$\begin{aligned} \operatorname{cond}(L \cdot A) &\leq \|L\| \cdot \|L^{-1}\| \cdot \|A\| \cdot \|A^{-1}\| \\ &= \operatorname{cond}(A) \cdot \operatorname{cond}(L), \end{aligned}$$

die Kondition vergrößert sich also um bis zu $\operatorname{cond}(L)$.

Idee: *Die Kondition lässt sich verbessern, indem man A durch Multiplikation mit orthogonalen bzw. unitären Matrizen auf eine obere Δs-Gestalt bringt, denn für orthogonale/unitäre Matrizen gilt nach Lemma 3.32*

$$\operatorname{cond}(QA) = \operatorname{cond}(A)$$

sowohl für die Euklidische Norm als auch für $\|\cdot\|_F$.

Lemma 4.1 *Sei Q orthogonal (bzw. unitär). Dann gilt $\|Qx\|_2 = \|x\|_2$.*

Beweis:

$$\|Qx\|_2^2 = (Qx)^*(Qx) = x^* \underbrace{Q^*Q}_{I} x = x^*x = \|x\|_2^2$$

QED

Definition 4.2 *Die Zerlegung einer Matrix $A \in \mathbb{K}^{m,n}$ der Form $A = QR$ mit einer unitären Matrix $Q \in \mathbb{K}^{m,m}$ und einer oberen Δs-Matrix $R \in \mathbb{K}^{m,n}$ heißt* ***QR-Zerlegung*** *von A. Dabei hat die Matrix R folgende Gestalt:*

$$R = \left(\begin{array}{ccc} r_{11} & & * \\ & \ddots & \\ 0 & & r_{kk} \\ \hline & 0 & \ddots \\ \hline & 0 & \end{array} \right) \begin{array}{l} \Big\} n \\ \\ \Big\} m \end{array}$$

wobei $k \leq \min\{n, m\}$ und $r_{11}, r_{22}, ..., r_{kk} \neq 0$.

Definition 4.3 *Sei $h \in \mathbb{K}^n$ normiert, d.h. $h^*h = \|h\|^2 = 1$. Dann heißt*

$$H = I - 2hh^*$$

Householder-Matrix.

Bemerkung:

$$h = \begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix} \qquad h^* = (\overline{h}_1, \ldots, \overline{h_n}) \qquad hh^* = \begin{pmatrix} h_1\overline{h}_1 & h_1\overline{h}_2 & \cdots & h_1\overline{h}_n \\ h_2\overline{h}_1 & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ h_n\overline{h}_1 & \cdots & \cdots & h_n\overline{h}_n \end{pmatrix}$$

Lemma 4.4 *Sei H eine Householder-Matrix. Dann gilt $HH^* = H^*H = I$ und $H = H^*$.*

Beweis:

$$\begin{aligned} H^* &= (I - 2hh^*)^* = I - 2hh^* = H \\ HH^* &= (I - 2hh^*)(I - 2hh^*) = I - 4hh^* + 4h\underbrace{h^*h}_{1}h^* = I \end{aligned}$$

QED

Beispiel: Sei

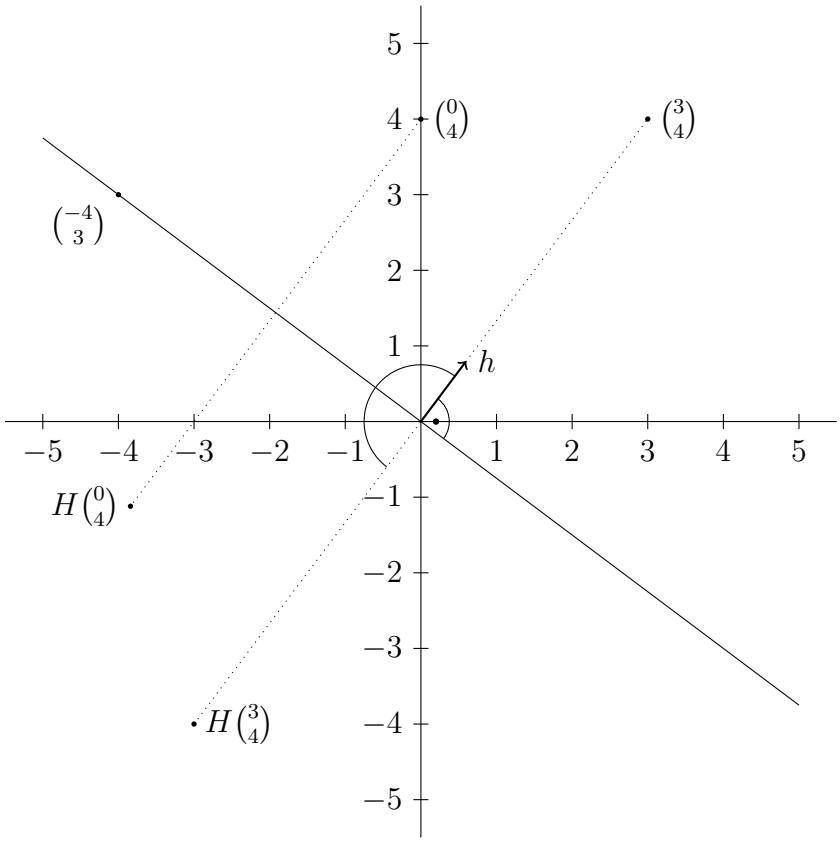
$$h = \frac{1}{5}\begin{pmatrix} 3 \\ 4 \end{pmatrix},$$

dann gilt $h^Th = 1$. Für die Householder-Matrix ergibt sich daraus:

$$H = I - 2hh^T = I - 2\begin{pmatrix} \frac{3}{5} \\ \frac{4}{5} \end{pmatrix}\begin{pmatrix} \frac{3}{5} & \frac{4}{5} \end{pmatrix} = \frac{1}{25}\begin{pmatrix} 7 & -24 \\ -24 & -7 \end{pmatrix}$$

wobei $H = H^T$ und $H^2 = I$. Nun kann man beliebige Punkte durch Multiplikation mit der Householder-Matrix auf andere abbilden, zum Beispiel:

$$H\begin{pmatrix}3\\4\end{pmatrix} = \begin{pmatrix}-3\\-4\end{pmatrix} \qquad H\left(\lambda\begin{pmatrix}3\\4\end{pmatrix}\right) = -\lambda\begin{pmatrix}3\\4\end{pmatrix}$$
$$H\begin{pmatrix}4\\-3\end{pmatrix} = \begin{pmatrix}4\\-3\end{pmatrix} \qquad H\left(\lambda\begin{pmatrix}4\\-3\end{pmatrix}\right) = \lambda\begin{pmatrix}4\\-3\end{pmatrix}$$
$$H\begin{pmatrix}0\\4\end{pmatrix} = \frac{1}{25}\begin{pmatrix}-96\\-28\end{pmatrix}$$



Aufgrund des Bildes scheint H also eine Spiegelung an der Geraden durch den Ursprung senkrecht zu h zu sein. Das wollen wir im folgenden begründen:

Lemma 4.5 *Die Abbildung $H : \mathbb{R}^n \to \mathbb{R}^n$ entspricht geometrisch einer Spiegelung an der zu h orthogonalen Ebene durch den Ursprung.*

Sei $x \in \mathbb{R}^n$. Wir zerlegen x in einen Anteil in Richtung h und einen Anteil orthogonal zu h.

$$x = \alpha h + \beta t \qquad \text{mit } t \perp h \quad (t^T h = 0)$$

$$\left[\text{Der Ansatz } x = (hh^T)x + y \text{ ist geeignet, da dann } y \perp h\right]$$

Im Beispiel:

$$\begin{pmatrix}0\\4\end{pmatrix} = \frac{16}{5} \cdot \frac{1}{5}\begin{pmatrix}3\\4\end{pmatrix} + \frac{12}{25}\begin{pmatrix}-4\\3\end{pmatrix}$$

Dann gilt:

$$\begin{aligned} H(x) &= (I - 2hh^T) \cdot (\alpha h + \beta t) \\ &= \alpha \cdot I \cdot h + \beta \cdot I \cdot t - 2\alpha h \underbrace{h^T h}_{1} - 2\beta h \underbrace{h^T t}_{0} \\ &= -\alpha h + \beta t \end{aligned}$$

Also ist H tatsächlich eine Spiegelung an der Ebene durch den Ursprung senkrecht zu h.

Unser Ziel ist es jetzt, die Kondition bei der Lösung von Gleichungssystemen zu verbessern, indem man A durch Multiplikation mit orthogonalen bzw. unitären Matrizen auf obere Dreiecksform bringt, also:

$$Q \cdot A = \begin{pmatrix} * & & \\ & \ddots & \\ 0 & & * \end{pmatrix}$$

wobei Q eine orthogonale bzw. unitäre Matrix darstellt. Dazu verwenden wir Householder-Matrizen H, für welche gilt:

$$H = I - 2hh^*, \text{ wobei } \|h^*\| = 1 \qquad H^* \cdot H = H \cdot H^* = I \qquad H^* = H$$

Unser Ziel ist es, eine Householdermatrix H so zu bestimmen, dass die Anwendung von H auf A zu einer Matrix führt, in der die erste Spalte ein Vielfaches des Einheitsvektors ist, also

$$H \cdot A_1 = \begin{pmatrix} * \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \lambda \cdot e_1 \qquad \text{und} \qquad HA = \left(\begin{array}{c|c} * & \\ 0 & \\ \vdots & * \\ 0 & \end{array}\right)$$

Das nächste Lemma zeigt, wie man so eine Matrix H wählen muss.

Lemma 4.6 *Sei $x \in \mathbb{K}^n \setminus \{0\}$. Für*

$$u := x \pm x_1 \cdot \frac{1}{|x_1|} \cdot \|x\|_2 \cdot e_1 \text{ und } H = I - 2 \cdot \frac{uu^*}{u^*u} \text{ gilt:}$$

$$Hx = \underbrace{-x_1 \cdot \frac{\|x\|_2}{|x_1|}}_{\in \mathbb{K}} \cdot e_1$$

Beweis: Ein stringenter Beweis lässt sich führen, indem man die Gleichung

$$Hx = -x_1 \cdot \frac{\|x\|_2}{|x_1|} \cdot e_1$$

direkt nachrechnet. Dazu empfiehlt es sich, eine Fallunterscheidung für

$$\begin{aligned} u &= x + x_1 \cdot \frac{1}{|x_1|} \cdot \|x\|_2 \cdot e_1 \text{ und} \\ u &= x - x_1 \cdot \frac{1}{|x_1|} \cdot \|x\|_2 \cdot e_1 \end{aligned}$$

durchzuführen. Stattdessen geben wir hier die "Herleitung" der im Lemma genannten Aussage an:

Wir suchen $u \in \mathbb{K}^n \setminus \{0\}$ mit $Hx = c \cdot e_1$, das heißt

$$Hx = x - 2 \cdot \frac{uu^* x}{u^* u} = c \cdot e_1.$$

Das ist erfüllt, falls die beiden folgenden Bedingungen gelten:

$$\frac{2u^* x}{u^* u} = 1 \tag{4.1}$$

$$u = x - c \cdot e_1 \tag{4.2}$$

Aus (4.1) folgt:

$$\begin{aligned} & 2u^* x = u^* u \in \mathbb{R} \\ \Rightarrow \quad & u^* x \in \mathbb{R} \\ \overset{(4.2)}{\Rightarrow} \quad & (x - ce_1)^* x \in \mathbb{R} \\ \Rightarrow \quad & x^* x - \overline{c} x_1 \in \mathbb{R} \\ \Rightarrow \quad & \overline{c} x_1 \in \mathbb{R} \qquad (*) \\ \Rightarrow \quad & c = \alpha \cdot x_1 \qquad \alpha \in \mathbb{R} \quad (**) \end{aligned}$$

Weiterhin gilt:

$$\begin{aligned} 0 &= 2u^* x - u^* u = u^*(2x - u) \\ &\overset{(4.2)}{=} (x - ce_1)^*(x + ce_1) \\ &= x^* x + x^* ce_1 - \overline{c} e_1^T x - \overline{c} c e_1^T e_1 \\ &= x^* x + c\overline{x}_1 - \underbrace{\overline{c} x_1}_{\substack{\in \mathbb{R}, \text{ nach } (**) \\ \Rightarrow \overline{c} x_1 = c\overline{x}_1}} - |c|^2 \\ &= \|x\|_2^2 - |c|^2 \end{aligned}$$

Zusammen mit (**) gilt:

$$\|x\|_2 = |c| \overset{(**)}{=} |\alpha||x_1|$$

Nun folgt:

$$|\alpha| = \frac{\|x\|_2}{|x_1|} \quad \text{also} \quad \alpha = \pm\frac{\|x\|_2}{|x_1|} \in \mathbb{R}$$

Daher ergeben sich als Lösung:

$$c \overset{(**)}{=} \pm x_1 \cdot \frac{\|x\|_2}{|x_1|} \quad \text{und} \quad u = x \mp x_1 \cdot \|x\|_2 \frac{1}{|x_1|} \cdot e_1.$$

QED

Die numerisch stabilere der beiden $\pm$ Variante ist $u = x + x_1 \cdot \frac{1}{|x_1|} \cdot \|x_2\| \cdot e_1$, weil es dann in der ersten Koordinate von u zu keiner Auslöschung kommen kann. Dieses funktioniert sogar, falls $x = \alpha \cdot e_1$.

Beispiel: Sei $x = \binom{3i}{4}$ gegeben. Gesucht sind nun $u(H)$ und c, so dass folgende Bedingungen erfüllt sind:

$$Hx = x - \frac{2}{u^*u}uu^*x = \begin{pmatrix} c \\ 0 \end{pmatrix}$$
$$u = \begin{pmatrix} u_1 + \tilde{u}_1 i \\ u_2 + \tilde{u}_2 i \end{pmatrix}$$

Es gilt:

$$2u^*x = 6\tilde{u}_1 + 8u_2 + i \cdot (6u_1 - 8\tilde{u}_2) \qquad u^*u = u_1^2 + \tilde{u}_2^2 + u_2^2 + \tilde{u}_2^2$$

Wir machen nun eine Fallunterscheidung. Im ersten Fall ist u reell, im zweiten Fall komplex. Es ergibt sich:

1. $2u^*x = u^*u$:
   (a) $6u_1 - 8\tilde{u}_2 = 0$
   (b) $6\tilde{u}_1 + 8u_2 = u_1^2 + \tilde{u}_1^2 + u_2^2 + \tilde{u}_2$
2. $\binom{3i}{4} - \binom{c_1 + \tilde{c}_1 i}{0} = \binom{u_1 + \tilde{u}_1 i}{u_2 + \tilde{u}_2 i}$
   (a) $-c_1 = u_1$
   (b) $3 - \tilde{c}_1 = \tilde{u}_1$
   (c) $4 = u_2$
   (d) $0 = \tilde{u}_2$

Aus 2(c) und 2(d) folgen folgende Werte:

$$\tilde{u}_2 = 0 \qquad u_2 = 4 \qquad \text{aus 1(a) folgt damit } u_1 = 0$$

Außerdem gilt:

$$6\tilde{u}_1^2 + 8 \cdot 4 = 0 + \tilde{u}_1^2 + 16 + 0 \quad \Rightarrow \tilde{u}_1^2 = \begin{cases} 8 \\ -2 \end{cases}$$

Daraus ergeben sich für u die Werte:

$$u = \begin{pmatrix} 8i \\ 4 \end{pmatrix} \quad \text{oder} \quad u = \begin{pmatrix} -2i \\ 4 \end{pmatrix}$$

Und auch nach Lemma 4.6 gilt:

$$u = x \pm x_1 \cdot \frac{\|x\|_2}{|x_1|} \cdot e_1 = \begin{pmatrix} 3i \\ 4 \end{pmatrix} \pm 3i \cdot \frac{5}{3} = \begin{pmatrix} 8i \\ 4 \end{pmatrix} \text{ oder } \begin{pmatrix} -2i \\ 4 \end{pmatrix}$$

Im ersten Fall erhält man die folgende Householder-Matrix:

$$\begin{aligned} H = I - 2\frac{uu^*}{u^*u} &= I - \frac{2}{80} \cdot \begin{pmatrix} 64 & 32i \\ -32i & 16 \end{pmatrix} \\ &= \frac{1}{10} \cdot \left[\begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix} - \begin{pmatrix} 16 & 8i \\ -8i & 4 \end{pmatrix} \right] = \frac{1}{5} \cdot \begin{pmatrix} -3 & -4i \\ 4i & 3 \end{pmatrix} \end{aligned}$$

Außerdem gilt:

$$Hx = \begin{pmatrix} -5i \\ 0 \end{pmatrix}$$

Satz 4.7 *Für eine Matrix $A \in \mathbb{K}^{m,n}$ mit dem Rang n (also $m \geq n$) existiert eine QR-Zerlegung.*

Beweis: Wir zeigen die folgende Aussage durch vollständige Induktion:

Für jedes $k = 1, \ldots, n$ gibt es Householder-Matrizen $H^{(1)}, \ldots, H^{(k)}$, so dass für $A^{(k)} = (a_{ij}^{(k)}) := H^{(k)} \cdot \ldots \cdot H^{(1)} \cdot A$ gilt: $a_{ij}^{(k)} = 0$ für alle $j \leq k$ und $i > j$.

Im Fall $k = n$ ergibt sich daraus, dass

$$\underbrace{H^{(n)} \cdot \ldots \cdot H^{(1)}}_{\text{unitäre Matrix}} \cdot A$$

eine obere Dreiecksmatrix ist; die Aussage des Satzes ist damit also bestätigt.

Für den Induktionsanfang wählt man in Lemma 4.6 $x = A_1$ und bestimmt anschließend eine Householder-Matrix $H^{(1)}$ so, dass die Gleichung $H^{(1)}x = \alpha \cdot e_1$ erfüllt ist. Es gilt also

$$H^{(1)}A = \left(\begin{array}{c|c} \alpha_1 & \\ \hline 0 & \\ \vdots & * \\ 0 & \end{array}\right)$$

und der Induktionsanfang ist gezeigt.

Seien nun nach $k < n$ Schritten die unitären Matrizen $H^{(1)}, \ldots, H^{(k)}$ so bestimmt, dass für $A^{(k)} \in \mathbb{K}^{m,n}$ gilt:

$$A^{(k)} = H^{(k)} \cdot \ldots \cdot H^{(1)} \cdot A = \left(\begin{array}{cc|c} * & & \\ & \ddots & B^{(k)} \\ & * & \\ \hline 0 & & C^{(k)} \end{array}\right) \Bigg\} k$$

wobei $B^{(k)} \in \mathbb{K}^{k,n-k}$ und $C^{(k)} \in \mathbb{K}^{m-k,n-k}$ und $a_{ij}^{(k)} = 0$ für alle $j \leq k$ und $i > j$ gilt. Wir bemerken, dass der Rang von $A^{(k)}$ gleich n ist, da A nach Voraussetzung Rang n hat und die Matrizen $H^{(i)}$ alle regulär sind. Sei $\tilde{x}^{(k+1)} = C_1 \in \mathbb{K}^{m-k}$. Zunächst gilt $\tilde{x}^{(k+1)} \neq 0$, denn wären die ersten $k+1$ Spalten linear abhängig, dann wäre der Rang von $A^{(k)}$ nicht n.
Nun benötigen wir einige Definitionen:

$$\tilde{u}^{(k+1)} := \tilde{x}^{(k+1)} + \frac{\tilde{x}_1^{(k+1)}}{|\tilde{x}_1^{(k+1)}|} \left\|\tilde{x}^{(k+1)}\right\|_2 e_1$$
$$\tilde{H}^{(k+1)} := I_{m-k} - 2\frac{\tilde{u}^{(k+1)}(\tilde{u}^{(k+1)})^*}{(\tilde{u}^{(k+1)})^*\tilde{u}^{(k+1)}}$$

Durch diese Definitionen folgt nach Lemma 4.6, dass

$$\tilde{H}^{(k+1)}C^{(k)} = \left(\begin{array}{c|c} \alpha & \\ 0 & \\ \vdots & * \\ 0 & \end{array}\right) \quad \text{mit } \alpha = \frac{-\tilde{x}_1^{(k+1)}}{|\tilde{x}_1^{(k+1)}|}\|\tilde{x}^{(k+1)}\|_2.$$

Um die Transformation auf ganz $A^{(k)}$ statt nur auf $C^{(k)}$ anwenden zu können, definieren wir

$$u^{(k+1)} := \left(\begin{array}{c} 0 \\ \vdots \\ 0 \\ \hline \tilde{u}^{(k+1)} \end{array}\right) \begin{array}{l} \Big\} k \\ \\ \Big\} m-k \end{array}$$

$$H^{(k+1)} := I_m - 2\frac{u^{(k+1)}(u^{(k+1)})^*}{(u^{(k+1)})^* u^{(k+1)}} = \left(\begin{array}{c|c} I_k & 0 \\ \hline 0 & \tilde{H}^{(k+1)} \end{array}\right)$$

wobei $\tilde{H}^{(k+1)}$ eine $m-k \times n-k$-Matrix ist. Mit diesen Definitionen erhält man jetzt für die Matrix $A^{(k+1)}$:

$$A^{(k+1)} = H^{(k)} \cdot A^{(k)} = \left(\begin{array}{c|c} \begin{pmatrix} * & & \\ & \ddots & \\ & & * \end{pmatrix} & B^{(k)} \\ \hline 0 & \tilde{H}^{(k+1)} C^{(k)} \end{array}\right)$$

mit der geforderten Eigenschaft, dass $a_{ij}^{(k+1)} = 0$ für alle $j \leq k+1$ und $i > j$.

QED

Algorithmus 6: QR-Verfahren (Matrixversion)

Input: $A \in \mathbb{K}^{m,n}$ `mit` $\mathrm{Rang}(A) = n$

Schritt 1: $A^{(0)} := A$

Schritt 2: **For** $k = 1, \ldots, n$ **do**

$$\begin{aligned} d &:= a_{kk}^{(k-1)} + a_{kk}^{(k-1)} \frac{1}{|a_{kk}^{(k-1)}|} \sqrt{\sum_{i=k}^{m} \left|a_{ik}^{(k-1)}\right|^2} \\ u^{(k)} &:= (0, \ldots 0, d, a_{k+1,k}^{(k-1)}, \ldots, a_{mk}^{(k-1)})^T \\ H^{(k)} &:= I_m - 2\frac{u^{(k)}(u^{(k)})^*}{(u^{(k)})^* u^{(k)}} \\ A^{(k)} &:= H^{(k)} A^{(k-1)} \end{aligned}$$

Ergebnis: QR-`Zerlegung mit`

$$\begin{aligned} A &:= QR \text{ mit} \\ R &:= A^{(n)} \text{ ist obere Dreiecksmatrix und} \\ Q &:= H^{(1)} \cdot \ldots \cdot H^{(n)} \text{ ist unitäre Matrix} \end{aligned}$$

Diese Berechnungen sind aufwändig, insbesondere die Anwendung von von $H^{(k)}$ auf alle Spalten von $A^{(k)}$. Wir suchen nun eine bessere Rechenvorschrift für die

Berechnung von dem Produkt Hv der Householder-Matrix H und einem beliebigen Vektor $v \in \mathbb{K}^m$. Es gilt:

$$H = I - \frac{2}{u^*u}uu^* = I - \beta uu^* \text{ mit } u = x + \frac{x_1}{|x_1|}\|x\|_2 e_1 \text{ und}$$

$$\beta = \frac{2}{u^*u} = \frac{1}{\|x\|_2(\|x\|_2 + |x_1|)} \tag{4.3}$$

Für beliebiges $v \in \mathbb{K}^m$ folgt:

$$\begin{aligned} Hv = (I - \beta uu^*)v &= v - \beta uu^*v \\ &= v - \beta u^*vu = v - su \qquad (4.4) \\ \text{mit } s &= \beta u^*v \in \mathbb{K} \qquad (4.5) \end{aligned}$$

Damit kann man HA_k für die Spalten A_k von A also effizient berechnen. Bei betragsmäßig kleinem β ergeben sich allerdings numerische Probleme. Sie kann man mit Hilfe des folgenden Lemmas vermeiden.

Lemma 4.8 *Sei H eine Householder-Matrix aus Lemma 4.6 zu $x \neq 0$ und H' die Householder-Matrix aus Lemma 4.6 zu $y = \alpha x$ mit $\alpha \in \mathbb{R}^+$ und $x, y \in \mathbb{K}^n$. Dann gilt $H = H'$.*

Beweis: Zunächst gelten folgende Gleichungen:

$$u = x + \frac{x_1}{|x_1|}\|x\|_2 e_1 \qquad H = I - \frac{2}{u^*u}uu^*$$

$$u' = y + \frac{y_1}{|y_1|}\|y\|_2 e_1 = \alpha x + \frac{\alpha x_1}{|\alpha x_1|}\|\alpha x_1\|_2 = \alpha u$$

Nun folgt daraus

$$H' = I - \frac{2}{(u')^*u'}u'(u')^* = I - \frac{2}{\alpha^2 u^*u}\alpha^2 uu^* = H$$

QED

Im folgenden Algorithmus wird statt x also $\frac{x}{\|x\|_\infty}$ verwendet, um die numerische Stabilität von β aus (4.3) zu gewährleisten.

Algorithmus 7: QR-Verfahren (Implementations-Variante)

Input: $A \in \mathbb{K}^{m,n}$ `mit` $\text{Rang}(A) = n$

Schritt 1: $u_{ik} := 0$ `für alle` $i = 1, \ldots, m$ `und` $k = 1, \ldots, n$

Schritt 2: For $k = 1, \ldots, n$ do

Schritt 2.1: ${A_k}^{\max} := \max_{i=k,\ldots,m} |a_{ik}|$

Schritt 2.2: $\alpha := 0$

Schritt 2.3: For $i = k, \ldots, m$ do

Schritt 2.3.1. $u_{ik} := \frac{a_{ik}}{{A_k}^{max}}$ (Normierung der k-ten Restspalte)

Schritt 2.3.2. $\alpha := \alpha + |u_{ik}|^2$ (Norm^2 der k-ten Restspalte)

Schritt 2.4: $\alpha := \sqrt{\alpha}$

Schritt 2.5: $\beta_k := \frac{1}{\alpha(\alpha+|u_{kk}|)}$ (β aus 4.3)

Schritt 2.6: $u_{kk} = u_{kk} + \frac{a_{kk}}{|a_{kk}|} \cdot \alpha$ (1. Komponente von u_k nach Lemma 4.6)

Schritt 2.7: $a_{kk} := -\frac{a_{kk}}{|a_{kk}|} \cdot {A_k}^{\max} \cdot \alpha$

Schritt 2.8: For $i = k+1, \ldots m$ do $a_{ik} := 0$ (erste Spalte von $H{A_k}^{\max}$)

Schritt 2.9: For $j = k+1, \ldots, n$ do

Schritt 2.9.1. $s := \beta_k \sum_{i=k}^{m} \overline{u_{ik}} a_{ij}$ (s aus 4.5)

Schritt 2.9.2. For $i = k, \ldots, m$ do $a_{ij} = a_{ij} - s \cdot u_{ik}$ (Berechnung von H $A_j^{(k)}$ nach 4.4)

Ergebnis: QR-Zerlegung der Originalmatrix A mit

$$R := A$$

$$Q := H^{(1)} \cdot \ldots \cdot H^{(n)} \text{ mit } H^{(k)} = I - \frac{2}{u_k^* u_k} u_k u_k^* \text{ und}$$

$$u_k = \begin{pmatrix} u_{1k} \\ \vdots \\ u_{nk} \end{pmatrix} \text{ für } k = 1, \ldots, n$$

Bemerkung:

- oft benötigt man Q nicht explizit und kann sich die Berechnung sparen
- U ist untere Dreiecksmatrix

Aufwand: Die QR-Zerlegung einer Matrix $A \in \mathbb{K}^{m,n}$ mit $\text{Rang}(A) = n$ erfordert

$$n^2(m) + O(mn) (= \text{O}(n^2 m))$$

wesentliche Operationen.

Zum Analysieren der Komplexität betrachten wir den teuersten Schritt. Das ist Schritt 2.9 mit

$$\begin{aligned}
\left(\sum_{k=1}^{n}\sum_{j=k+1}^{n} 2(m-k)\right) + O(mn) &= \sum_{k=1}^{n} 2(n-k)(m-k) + O(mn) \\
&= \left(\sum_{k=1}^{n} 2nm - 2nk - 2km + 2k^2\right) + O(mn) \\
&= 2n^2m + 2\sum_{k=1}^{n} k^2 - k(n+m) + O(mn) \\
&= 2n^2m + 2\underbrace{\frac{n(n+1)(2n+1)}{6}}_{\sum k^2} - 2(n+m)\underbrace{\frac{n(n+1)}{2}}_{\sum k} + O(mn) \\
&= 2n^2m + 2\frac{2n^3}{6} - n^2(n+m) + \underbrace{O(mn) + O(n^2)}_{=O(mn)} \\
&= n^2\left(m - \frac{1}{3}n\right) + O(mn)
\end{aligned}$$

Für $m = n$ ergibt sich also $\frac{2}{3}n^3 + O(n^2)$, das ist etwas höher als der Aufwand von $\frac{1}{3}n^3$ bei dem Gauss-Verfahren. Für schlecht konditionierte Matrizen lohnt es sich aber, diesen Aufwand in Kauf zu nehmen.

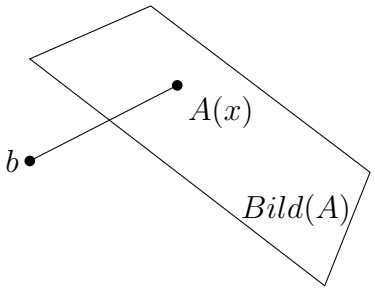
Bemerkung: Im Gegensatz zur LU-Zerlegung ist bei der QR-Zerlegung (bei $\mathrm{Rang}(A) = n$) keine Pivotisierung nötig. Das gilt allerdings nicht im Fall $\mathrm{Rang}(A) < n$. In diesem Fall tauscht man in jedem Schritt k vor der Berechnung der u_k und der β_k die Restspalte mit größter Euklidischer Norm an die k-te Position.

4.2 Lineare Ausgleichsprobleme

Beim Lösen linearer Gleichungssysteme bestand die Aufgabe darin, ein x zu finden, so dass $Ax = b$ gilt. Was passiert aber nun, wenn $Ax = b$ nicht lösbar ist? In diesem Fall versucht man, ein x zu finden, so dass der Ausdruck Ax die rechte Seite b möglichst gut annähert. Verwendet man zur Bewertung der Qualität der Annäherung die Euklidische Norm, führt das zu dem Minimierungsproblem

$$\text{minimiere}_{x \in \mathbb{K}^n} \|Ax - b\|_2,$$

in dem man unter allen Vektoren $x \in \mathbb{K}^n$ den sucht, der die Euklidische Norm von $Ax - b$ minimiert.



Da es äquivalent ist, statt $\|Ax-b\|_2$ die quadratische Funktion $\|Ax-b\|_2^2$, zu minimieren, definieren wir das **lineare Ausgleichsproblem** wie folgt:

(AuP) minimiere$_{x\in\mathbb{K}^n} F(x)$ mit $F(x)=\|Ax-b\|_2^2$, $A\in\mathbb{K}^{m,n}$, $b\in\mathbb{K}^m$

Beispiel:

$$A=\begin{pmatrix}2 & 1\\ 1 & -1\\ 1 & 1\end{pmatrix} \qquad b=\begin{pmatrix}1\\0\\1\end{pmatrix}$$

Zwei mögliche Lösungen für x werden im folgenden untersucht:

$$x=\begin{pmatrix}\frac{1}{3}\\ \frac{1}{3}\end{pmatrix} \text{ Lösung bzgl. } \begin{pmatrix}2 & 1\\ 1 & -1\end{pmatrix}=b \Rightarrow F(x)=\left\|\begin{pmatrix}1 & 0 & \frac{2}{3}\end{pmatrix}^T-b\right\|_2=(\tfrac{1}{3})^2$$

$$x=\begin{pmatrix}\frac{1}{2}\\ \frac{1}{2}\end{pmatrix} \text{ Lösung bzgl. } \begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix}=b \Rightarrow F(x)=\left\|\begin{pmatrix}\frac{3}{2} & 0 & 1\end{pmatrix}^T-b\right\|_2=(\tfrac{1}{2})^2$$

Im folgenden wollen wir untersuchen, wie man die beste Lösung für solche Ausgleichsprobleme findet. Zunächst erinnern wir daran, dass

$$\text{Bild}(A^*)=\{A^*z : z\in\mathbb{K}^m\}$$
$$\text{Bild}(A^*A)=\{A^*Ax : x\in\mathbb{K}^n\}.$$

Wir benötigen das folgende Lemma:

Lemma 4.9 *Sei $A\in\mathbb{K}^{m,n}$. Dann gilt* $\text{Bild}(A^*)=\text{Bild}(A^*A)$.

Beweis: *Übung!*

Nun können wir die Lösung des Ausgleichsproblems näher analysieren.

Satz 4.10 *Sei $A\in\mathbb{K}^{m,n}$, $b\in\mathbb{K}^m$ und $b\notin\{Ax : x\in\mathbb{K}^n\}$. Dann gilt*

1. *(AuP) ist lösbar.*

2. *$x \in \mathbb{K}^n$ ist Lösung von (AuP) genau dann, wenn*

$$A^*Ax = A^*b. \tag{N}$$

Man sagt ***"'x löst die Normalengleichung (N) bezüglich A und b."'***

3. *(AuP) ist eindeutig lösbar genau dann wenn* $\mathrm{Rang}(A) = n$.

Beweis:

1. Sei (x_k) eine so genannte Minimalfolge, d.h.

$$\|Ax_k - b\|_2 \to \alpha := \inf_{x \in \mathbb{K}^n} \|Ax - b\|_2.$$

Für jedes $\epsilon > 0$ gibt es also ein $K \in \mathbb{N}$, so dass für alle $k \geq K$ gilt: $\|Ax_k - b\|_2 < \alpha + \epsilon$. Für alle $k \geq K$ erhalten wir somit

$$\|Ax_k\|_2 = \|Ax_k - b + b\|_2 \leq \|Ax_k - b\|_2 + \|b\|_2 < \alpha + \epsilon + \|b\|_2.$$

Also ist die Folge $(Ax_k) \subseteq \mathrm{Bild}(A)$ beschränkt. Vom Satz von Bolzano-Weierstrass wissen wir daher, dass es eine konvergente Teilfolge von (Ax_k) gibt, die gegen ein $\tilde{y}$ konvergiert, insbesondere gilt also

$$\|Ax_k - b\|_2 \to \|\tilde{y} - b\|_2 = \alpha.$$

Aufgrund der Stetigkeit von A ist $\mathrm{Bild}(A)$ abgeschlossen. Daraus folgt, $\tilde{y} \in \mathrm{Bild}(A)$. Also gibt es $\tilde{x}$ mit $A\tilde{x} = \tilde{y}$ und daher gilt

$$\|A\tilde{x} - b\|_2 = \|\tilde{y} - b\|_2 = \alpha,$$

also ist $\tilde{x}$ Lösung des Ausgleichsproblems.

2. Für jede Lösung x_0 von (N) und für jedes $x \in \mathbb{K}^n$ gilt

$$\begin{aligned} F(x) - F(x_0) &= \|Ax - b\|_2^2 - \|Ax_0 - b\|_2^2 \\ &= (Ax - b)^*(Ax - b) - (Ax_0 - b)^*(Ax_0 - b) \\ &= x^*A^*Ax - x^*A^*b - b^*Ax + b^*b \\ &\quad - x_0^*A^*Ax_0 + x_0^*A^*b + b^*Ax_0 - b^*b \end{aligned}$$

Weil x_0 (N) erfüllt ist, gilt $A^*b = A^*Ax_0$ bzw. $b^*A = x_0^*A^*A$. Unter Verwendung dieser Gleichungen erhält man weiter

$$\begin{aligned} &= x^*A^*Ax - x^*A^*Ax_0 - x_0^*A^*Ax \\ &\quad - x_0^*A^*Ax_0 + x_0^*A^*Ax_0 + x_0^*A^*Ax_0 \\ &= x^*A^*Ax - x^*A^*Ax_0 - x_0^*A^*Ax + x_0^*A^*Ax_0 \\ &= (x - x_0)^*A^*A(x - x_0) \\ &= \|A(x - x_0)\|_2^2 \geq 0 \end{aligned} \tag{4.6}$$

"'$\Leftarrow$"' Sei x_0 eine Lösung von (N). Dann folgt $F(x) \geq F(x_0)$, $\forall x \in \mathbb{K}^n$, also ist x_0 eine Lösung von (AuP).

"'$\Rightarrow$"' Sei andererseits x Lösung von (AuP).

Wir wählen zunächst einen Punkt $x_0 \in \mathbb{K}^n$, der die Normalengleichung erfüllt: Da $A^*b \in \text{Bild}(A^*)$ ist auch $A^*b \in \text{Bild}(A^*A)$ (siehe Lemma 4.9) und daher gibt es

$$x_0 \in \mathbb{K}^n \text{ mit } A^*Ax_0 = A^*b. \tag{4.7}$$

Wegen (4.6) folgt $F(x) - F(x_0) \geq 0$. Andererseits gilt $F(x) \leq F(x_0)$, weil x eine Lösung von (AuP) ist. Zusammen folgt $F(x) = F(x_0)$ und daraus (wieder wegen (4.6)) $\|A(x - x_0)\|_2^2 = 0$, also auch $\|A(x - x_0) = 0\|$ Nach dem ersten Normaxiom erhalten wir $A(x - x_0) = 0$ und entsprechend $A^*Ax = A^*Ax_0 = A^*b$, also erfüllt x die Normalengleichung (N).

3. Falls $\text{Rang}(A) = n$ ist $A^*A \in \mathbb{K}^{n,n}$ regulär. Also ist $A^*Ax = A^*b$ eindeutig lösbar. Ist $\text{Rang}(A) < n$, so existiert wegen (4.7) eine Lösung x_0 von (AuP) sowie ein $z \in \text{Kern}(A)$ mit $z \neq 0$. Damit gilt:

$$F(x_0 + z) = \|A(x_0 + z) - b\|_2^2 = \|Ax_0 + Az - b\|_2^2 = \|Ax_0 - b\|_2^2 = F(x_0)$$

und $x_0 + z \neq x_0$, also gibt es zwei verschiedene Lösungen von (AuP)

Bemerkung: Ist die Lösung von (AuP) nicht eindeutig, so lässt sich aus

$$\text{Opt}^* = \{x \in \mathbb{K}^n : x \text{ ist Lösung von (AuP)}\}$$

ein eindeutiges $\tilde{x}$ mit minimaler Euklidischer Norm $\|\tilde{x}\|_2$ wählen. D.h.

$$\min_{x \in \text{Opt}^*} \|x\|_2$$

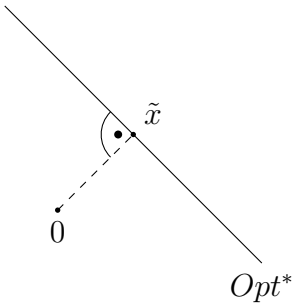
ist eindeutig lösbar.

Beweis: $\text{Opt}^* = \{x \in \mathbb{K}^n : A^*Ax = A^*b\}$ ist ein affin linearer Teilraum. Dieser enthält genau ein Element mit minimaler Euklidischer Länge, nämlich die orthogonale Projektion von 0 auf Opt^* QED

Aufgabe: Sei $L \in \mathbb{R}^n$ ein affin linearer Teilraum und $a \in \mathbb{R}^n$. Zeigen Sie, dass das Minimierungsproblem

$$\text{minimiere}_{x \in L} \|a - x\|$$

eindeutig lösbar ist, und zwar von der orthogonalen Projektion von a auf L.



Zwei Ansätze zum Lösen des Ausgleichsproblems

Sei $A \in \mathbb{K}^{m,n}$ mit $m \geq n$.

Idee 1 Nutze die Normalengleichung aus Satz 4.10 und löse das Gleichungssystem $A^T Ax = A^T b$ (im Reellen) durch das Cholesky-Verfahren. Das ist schnell, aber oft ungenau.

Idee 2 Führe eine QR-Zerlegung von A durch. Man erhält

$$A = QR = Q \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix}, \qquad \hat{R} \in \mathbb{K}^{n,n} \text{ obere Dreiecksmatrix}$$

Ist $\text{Rang}(A) = n$, so ist $\hat{R}$ regulär. Es gilt

$$\begin{aligned} \|Ax - b\|_2^2 &= \|QRx - b\|_2^2 = \|Q^*(QRx - b)\| \text{ (siehe den Beweis von Lemma 3.32)} \\ &= \|Rx - Q^*b\|_2^2 \\ &= \| \begin{pmatrix} \hat{R}x \\ 0 \end{pmatrix} - \begin{pmatrix} c \\ d \end{pmatrix} \|_2^2 = \|\hat{R}x - c\|_2^2 + \|d\|_2^2 \end{aligned} \tag{4.8}$$

wobei $\begin{pmatrix} c \\ d \end{pmatrix} = Q^*b$ eine Zerlegung des Vektors $Q^*b \in \mathbb{K}^m$ in $c \in \mathbb{K}^n$, $d \in \mathbb{K}^{m-n}$ ist.

Lemma 4.11 *Sei* $\|Ax - b\|_2^2 \to \min$ *ein lineares Ausgleichsproblem mit* $A \in \mathbb{K}^{m,n}, m \geq n$ *und* $\text{Rang}(A) = n$, *und* $A = QR$ *eine* QR*-Zerlegung von* A,

$$R = \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix} \text{ und } Q^*b = \begin{pmatrix} c \\ d \end{pmatrix}$$

Dann ist

$$x = \hat{R}^{-1}c$$

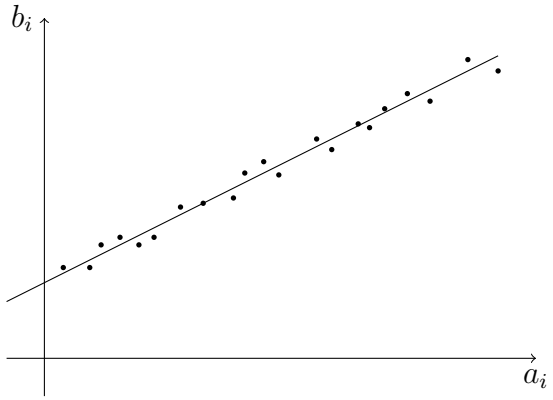
die eindeutige Lösung von (AuP) und $\|d\|_2^2$ *der zugehörige Zielfunktionswert.*

Beweis: Nach (4.8) wissen wir, dass $\|Ax - b\|_2^2 = \|\hat{R}x - c\|_2^2 + \|d\|_2^2$. Dieser Ausdruck wird minimal, falls $c = \hat{R}x$. Da $\hat{R}$ regulär, existiert so ein x, nämlich $\hat{R}^{-1}c$. Die Zielfunktion ergibt sich als

$$\|Ax - b\|_2^2 = 0 + \|d\|_2^2$$

also $\|Ax - b\|^2 = \|d\|_2^2$. QED

Anwendungsbeispiel (Statistik): Es seien Messdaten (a_i, b_i) mit $i = 1, \ldots, m$ gegeben, bei denen ein (unbekannter) linearer Zusammenhang besteht.



Gesucht sind die Parameter α, β, die diesen linearen Zusammenhang beschreiben. Dabei soll b_i durch die Gleichung $\alpha a_i + \beta$ in Abhängigkeit von a_i möglichst gut geschätzt werden können, d.h. $\alpha a_i + \beta$ soll möglichst nahe an b_i sein. Wir wollen die Qualität dieser Schätzung maximieren und versuchen dazu, die Summe aller quadrierten Schätzfehler zu minimieren. Das führt auf das folgende Problem:

$$\text{minimiere}_{\alpha,\beta} \sum_{i=1}^{m} |\alpha a_i + \beta - b_i|^2.$$

Mit

$$A = \begin{pmatrix} a_1 & 1 \\ \vdots & \\ a_n & 1 \end{pmatrix} \qquad x = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \qquad b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

erhält man

$$\|Ax - b\|_2^2 = \| \begin{pmatrix} a_1\alpha + \beta - b_1 \\ \vdots \\ a_n\alpha + \beta - b_n \end{pmatrix} \|_2^2 = \sum_{i=1}^{n} |\alpha a_i + \beta - b_i|^2,$$

also ein lineares Ausgleichsproblem. Man nennt dies "Methode der kleinsten Quadrate".

4.3 Singulärwertzerlegung

In diesem Abschnitt beschäftigen wir uns mit Orthogonalisierungsverfahren für nicht quadratische Matrizen. Sei $A = (a_{ij}) \in \mathbb{K}^{m,n}$ eine solche Matrix. Wir bezeichnen A als **Diagonalmatrix** falls $a_{ij} = 0$ für alle $i \neq j$ mit $i \in \{1, \ldots, m\}$ und $j \in \{1, \ldots, n\}$. Mit dieser Bezeichnung führen wir den Begriff der Singulärwertzerlegung ein.

Definition 4.12 *Sei $A \in \mathbb{K}^{m,n}$. Eine Zerlegung der Form $A = U\Sigma V^*$ mit unitären Matrizen $U \in \mathbb{K}^{m,m}$ und $V \in \mathbb{K}^{n,n}$ und einer Diagonalmatrix $\Sigma \in \mathbb{K}^{m,n}$ heißt eine* Singulärwertzerlegung *von A.*

Wir benutzen die Dimensionsformel: Für $A \in \mathbb{K}^{m,n}$

$$n = \mathrm{Rang}(A) + \dim(\mathrm{Kern}(A))$$

sowie die folgende Aussage aus der linearen Algebra.

Lemma 4.13 *Sei $A \in \mathbb{K}^{m,n}$. Dann gelten*

$$\mathrm{Kern}(A) = \mathrm{Kern}(A^*A)$$
$$\mathrm{Rang}(A) = \mathrm{Rang}(A^*) = \mathrm{Rang}(A^*A) = \mathrm{Rang}(AA^*)$$

Beweis: *Übung!*

Satz 4.14 *Jede Matrix $A \in \mathbb{K}^{m,n}$ besitzt eine Singulärwertzerlegung.*

Beweis: Seien $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ die (rellen) Eigenwerte von A^*A mit zugehörigen Eigenvektoren $v_1, \ldots, v_n$, so dass

$$A^*Av_j = \lambda_j v_j \text{ und } v_j^* v_k = \begin{cases} 1 & \text{falls } j = k \\ 0 & \text{falls } j \neq k \end{cases}$$

Sei $\mathrm{Rang}(A^*A) = r \leq \min\{m, n\}$. Dann sind genau r der Eigenwerte positiv und die restlichen Null. Weil ebenso $r = \mathrm{Rang}(AA^*)$, hat also auch AA^* genau r positive Eigenwerte. Definiere

$$\sigma_j = \sqrt{\lambda_j} \quad \text{und} \quad u_j = \frac{1}{\sigma_j} Av_j \qquad 1 \leq j \leq r \tag{4.9}$$

Dann gilt

$$AA^*u_j = \frac{1}{\sigma_j} A \underbrace{A^*Av_j}_{\lambda_j v_j} = \frac{1}{\sigma_j} \lambda_j Av_j = \lambda_j u_j \qquad 1 \leq j \leq r$$

$$u_j^* u_k = \frac{1}{\overline{\sigma}_j \sigma_k} v_j^* \underbrace{A^*Av_k}_{\lambda_k v_k} = \frac{\lambda_k}{\sigma_j \sigma_k} v_j^* v_k = \begin{cases} 1 & \text{falls } j = k \\ 0 & \text{sonst} \end{cases}$$

Also sind $u_1, \ldots, u_r$ ein Orthonormalsystem von Eigenvektoren zu den Eigenwerten $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ der Matrix AA^*. Ergänze $\{u_1, \ldots, u_r\}$ zu einer Orthonormalbasis $\{u_1, \ldots, u_m\}$ aus Eigenvektoren und setze

$$V := (v_1, \ldots, v_n) \in \mathbb{K}^{n,n} \quad \text{und} \quad U := (u_1, \ldots, u_m) \in \mathbb{K}^{m,m}$$

Dann gilt

$$Av_j = \begin{cases} \sigma_j u_j & \text{für} \quad 1 \leq j \leq r \text{ wegen (4.9)} \\ 0 & \text{für} \quad r+1 \leq j \leq n \text{ weil } v_j \in \text{Kern}(A^*A) = \text{Kern}(A) \end{cases}$$

Also erhält man

$$AV = U\Sigma \text{ mit } \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r, \underbrace{0, \ldots, 0}_{\min\{n,m\}-r}) \in \mathbb{K}^{m,n},$$

wobei die Matrizen V, U unitär sind, weil ihre Spalten jeweils ein Orthonormalsystem bilden. Es folgt $A = U\Sigma V^*$. QED

Bemerkung:

- Die Einträge von Σ sind eindeutig, wenn man Positivität verlangt.
- Ist A selber quadratisch und hermitesch, dann gilt $\sigma_j = |\mu_j|$ wenn μ_j Eigenwert von A ist.

Definition 4.15 *Die positiven Werte $\sigma_j > 0$, die in der Singulärwertzerlegung der Matrix A aus Satz 4.13 auftreten, heißen* Singulärwerte *von A.*

Eine effiziente Berechnung der Singulärwertzerlegung wird in Kapitel **??** besprochen.

4.4 Anwendung der Singulärwertzerlegung auf lineare Ausgleichsprobleme

Bisher hatten wir zwei Methoden kennen gelernt, um lineare Ausgleichsprobleme zu lösen. In diesem Abschnitt kommt eine weitere — nämlich durch Anwendung der Singulärwertzerlegung — dazu.

(AuP) minimiere$\|Ax - b\|_2^2$

Methode 1 Löse die Normalengleichung $A^*Ax = A^*b$ durch das Cholesky-Verfahren.

Methode 2 Bestimme eine QR-Zerlegung von A und löse

$$\|Ax - b\|_2^2 = \|\hat{R}x - c\|_2^2 + \|d\|_2^2$$

Methode 3 Die dritte Methode beruht auf der Singulärwertzerlegung und wird im folgenden erläutert.

Sei für eine Matrix $A \in \mathbb{K}^{m,n}$ eine Singulärwertzerlegung $A = U\Sigma V^*$ gegeben. Setze $y := V^*x \in \mathbb{K}^n$ und $c := U^*b \in \mathbb{K}^m$. Es folgt

$$\begin{aligned}
\|Ax - b\|_2^2 &= \|U\Sigma V^*x - UU^*b\|_2^2 \\
&= \|U(\Sigma y - c)\|_2^2 \\
&= \|\Sigma y - c\|_2^2 \qquad \text{nach Lemma 4.1 weil } U \text{ unitär} \\
&= \|(\sigma_1 y_1, \sigma_2 y_2, \ldots, \sigma_r y_r, 0, \ldots 0)^T - c\|_2^2 \\
&= \sum_{j=1}^{r} (\sigma_j y_j - c_j)^2 + \sum_{j=r+1}^{m} c_j^2
\end{aligned}$$

Satz 4.16 *Eine Lösung des linearen Ausgleichsproblems (AuP) ist gegeben durch*

$$x = \sum_{j=1}^{r} \frac{c_j}{\sigma_j} V_j + \sum_{j=r+1}^{n} \alpha_j V_j,$$

wobei $A = U\Sigma V^*$, σ_j die Singulärwerte, V_j die Spalten von V, und $c = U^*b$ sind. Die α_j können beliebig aus $\mathbb{R}$ gewählt werden. Für $\alpha_j = 0$ erhält man die Lösung von (AuP) mit minimaler Euklidischer Norm.

Beweis: Um $\|Ax - b\|_2^2$ zu minimieren, minimieren wir

$$\sum_{j=1}^{r} (\sigma_j \underbrace{y_j}_{\text{variabel}} - c_j)^2 + \sum_{j=r+1}^{m} c_j^2$$

Also wähle für beliebiges α_j, $j = r+1, \ldots, n$

$$y_j = \begin{cases} \frac{c_j}{\sigma_j} & \text{für } j = 1, \ldots, r \\ \alpha_j & \text{für } j = r+1, \ldots, n \end{cases}$$

x ergibt sich dann aus

$$x = Vy = \sum_{j=1}^{n} V_j y_j = \sum_{j=1}^{r} \frac{c_j}{\sigma_j} V_j + \sum_{j=r+1}^{n} \alpha_j V_j.$$

Die Norm von x berechnet man durch

$$\|x\|_2^2 = x^*x = \sum_{j=1}^{r} \left(\frac{c_j}{\sigma_j}\right)^2 + \sum_{j=r+1}^{n} \alpha_j^2$$

und dieser Ausdruck ist minimal für $\alpha_j = 0$, $j = r+1, \ldots, n$. QED

Zum Abschluss vergleichen wir die drei besprochenen Methoden. Sei eine Matrix A gegeben mit $\text{Rang}(A) = n$ und Singulärwerten $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$. Wir untersuchen die Kondition der drei möglichen Verfahren für das lineare Ausgleichsproblem.

Cholesky Löse $A^*Ax = A^*b$.

$$\begin{aligned}\text{cond}(A^*A) &= \|A^*A\|_2 \cdot \|(A^*A)^{-1}\|_2 \\ \|A^*A\|_2 &= \sqrt{\rho((A^*A)^*(A^*A))} = \sqrt{\rho(A^*AA^*A)} \\ &= \sqrt{\rho((A^*A)^2)} = \sqrt{\lambda_1^2} = \sigma_1^2,\end{aligned}$$

denn für eine beliebige Matrix B folgt aus $Bx = \lambda x$ dass $B^2x = \lambda^2 x$. Weiter gilt:

$$\begin{aligned}\|(A^*A)^{-1}\| &= \sqrt{\rho(((A^*A)^{-1}(A^*A)^{-1}))} = \sqrt{\rho(((A^*A)^2)^{-1})} \\ &= \sqrt{\frac{1}{\lambda_n^2}} = \frac{1}{\sigma_n^2},\end{aligned}$$

denn aus $Bx = \lambda x$ folgt $B^{-1}x = \frac{1}{\lambda}x$ und außerdem gilt $(B^2)^{-1} = (B^{-1})^2$. Zusammen erhalten wir

$$\text{cond}(A^*A) = \frac{\sigma_1^2}{\sigma_n^2} = \left(\frac{\sigma_1}{\sigma_n}\right)^2$$

Singulärwertzerlegung Löse $\Sigma y = c$ und $x = Vy$ (mit orthogonaler Matrix V)

$$\text{cond}(\Sigma) = \|\Sigma\|_2 \|\Sigma^{-1}\|_2 = \frac{\sigma_1}{\sigma_n}$$

QR-Zerlegung Löse $\hat{R}x = Q^*b$

$$\text{cond}(R) = \|R\|_2 \|R^{-1}\|_2$$

Weil $A^*A = (QR)^*(QR) = R^*Q^*QR = R^*R$ ist der größte (bzw. kleinste) Eigenwert von A^*A auch der größte (bzw. kleinste) Eigenwert von R^*R, also folgen

$$\begin{aligned}\|R\|_2 &= \sqrt{\lambda_1} = \sigma_1 \\ \|R^{-1}\|_2 &= \frac{1}{\sqrt{\lambda_n}} = \frac{1}{\sigma_n},\end{aligned}$$

weil $(R^{-1})^* R^{-1} = (RR^*)^{-1}$ Inverses von RR^* mit Eigenwerten $\lambda_1, \ldots, \lambda_n$. Also

$$\operatorname{cond}(R) = \frac{\sigma_1}{\sigma_n}$$

Die Kondition der Cholesky-Zerlegung ist also das Quadrat der Kondition aus QR-Verfahren oder Singulärwertzerlegung.

Chapter 5

Iterationsverfahren für Gleichungssysteme

5.1 Das Verfahren der sukzessiven Approximation

In diesem Kapitel betrachten wir nach den Eliminationsverfahren und den Orthogonalisierungsverfahren noch eine dritte Klasse von Verfahren, die man zur Lösung von linearen (und nichtlinearen) Gleichungssystemen verwenden kann, so genannte *iterative Verfahren*. Wir betrachten dazu gleich relativ allgemein Funktionen $f_1, \ldots, f_m$ mit

$$f_i : \mathbb{R}^n \to \mathbb{R},\ i = 1, \ldots, m$$

und bezeichnen das System

$$\begin{aligned} f_1(x_1, \ldots, x_n) &= 0 \\ f_2(x_1, \ldots, x_n) &= 0 \\ \vdots &\quad \vdots \\ f_m(x_1, \ldots, x_n) &= 0 \end{aligned} \tag{5.1}$$

als **nichtlineares Gleichungssystem** mit den Variablen $x_1, \ldots, x_n$. Definiert man

$$F(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_m(x) \end{pmatrix}, x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

so kann man das Gleichungssystem in Kurzform auch als

$$F(x) = 0$$

schreiben.

Gilt für $x \in \mathbb{R}^n$, dass $F(x) = 0$, so nennt man x eine Lösung des Gleichungssystems. Dass wir in dem Gleichungssystem die rechte Seite zu Null gesetzt haben, ist keine Einschränkung, weil man ein Gleichungssystem $F(x) = b$ mit $b = (b_1, \ldots, b_m)^T \in \mathbb{R}^m$ jederzeit zu $G(x) = F(x) - b = 0$ umformen kann.

Nichtlineare Gleichungssysteme lassen sich im Allgemeinen nicht durch algebraische Manipulationen exakt auflösen. Wir betrachten in diesem Kapitel daher *iterative Verfahren* bzw. *Iterationsverfahren*, die eine gegebene Lösung in jedem Schritt verbessern, bis eine vorgegebene Genauigkeit erreicht ist. Dazu betrachten wir Gleichungssysteme $F(x) = 0$, die als **Fixpunktgleichung** vorliegen.

Definition 5.1 *Sei $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ eine Funktion. Die Gleichung*

$$\Phi(x) = x$$

wird als **Fixpunktgleichung** *betrachtet. Jedes $x \in \mathbb{R}^n$, für das $\Phi(x) = x$ gilt, wird als* **Fixpunkt** *von Φ bezeichnet.*

Der Zusammenhang zwischen Fixpunktgleichungen und linearen Gleichungssystemen wird im folgenden Lemma beschrieben.

Lemma 5.2

1. *Sei $m \leq n$ und das Gleichungssystem $F(x) = 0$ wie in (5.1) gegeben. Sei $M : \mathbb{R}^m \to \mathbb{R}^n$ eine lineare, injektive Abbildung. Definiere*

$$\Phi(x) = M(F(x)) + x. \tag{5.2}$$

   Dann ist x ein Fixpunkt von Φ genau dann, wenn x das Gleichungssystem löst. Die Fixpunktgleichung $\Phi(x) = x$ ist also äquivalent zu dem Gleichungssystem $F(x) = 0$.

2. *Sei andererseits die Abbildung $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ gegeben. Definiere*

$$F(x) = \Phi(x) - x.$$

   Dann ist das Gleichungssystem $F(x) = 0$ äquivalent zu der Fixpunktgleichung $\Phi(x) = x$.

Beweis:

ad 1: Es gilt $\Phi(x) = x \iff M(F(x)) = 0$. Wegen der Injektivität der linearen Abbildung M ist das genau dann der Fall, wenn $F(x) = 0$ ist.

ad 2: Es gilt $F(x) = 0 \iff \Phi(x) = x$. QED

Gleichungssystem $F(x) = 0$ mit $m \leq n$ können wir also lösen, wenn wir Fixpunkte bestimmen können. Damit werden wir uns im folgenden beschäftigen. (Ausgleichsprobleme mit $m > n$ behandeln wir später.)

Die Idee der sukzessiven Approximation ist nun die folgende. Man betrachtet für ein gegebenes $x^{(0)}$ die Folge

$$x^{(k+1)} = \Phi(x^{(k)}),\ k = 0, 1, 2, \ldots.$$

Angenommen, Φ ist stetig und die Folge der $x^{(k)}$ konvergiert. Dann gibt es einen Grenzwert

$$y = \lim_{k \to \infty} x^{(k)},$$

für den gilt $y = \Phi(y)$, y ist also ein Fixpunkt von Φ.

Definition 5.3 *Die Iterationsvorschrift*

$$x^{(k+1)} = \Phi(x^{(k)})$$

nennt man **Verfahren der sukzessiven Approximation**.

Als Beispiel betrachten wir die Gleichung

$$f(x) = 2x - \tan(x) = 0.$$

Wir schreiben die Gleichung als Fixpunktgleichung um.

- In der ersten Variante schreiben wir

  $$\phi(x) = f(x) + x = 3x - \tan(x)$$

  und suchen einen Fixpunkt von ϕ mittels der Folge

  $$x^{(k+1)} = 3x^{(k)} - \tan(x^{(k)})$$

- In einer zweiten Variante schreiben wir

  $$x = \frac{\tan(x)}{2}$$

  und erhalten die Folge

  $$x^{(k+1)} = \frac{1}{2}\tan(x^{(k)}).$$

- Als drittes verwenden wir

  $$x = \arctan(2x),$$

  was zu der Folge

  $$x^{(k+1)} = \arctan(2x^{(k)})$$

  führt.

Implementiert man in allen drei Fällen das Verfahren der sukzessiven Iteration so ergeben sich unterschiedliche Verhalten der drei Formeln: Zum Beispiel für den Startwert 1.2 geht Iterationsvorschrift 1 gegen unendlich, Iterationsvorschrift 2 gegen Null und Iterationsvorschrift 3 gegen 1.1656...

Im folgenden wollen wir untersuchen, wann solche Iterationsvorschriften konvergieren. Zunächst beweisen wir den folgenden Satz für *skalare* Funktionen $f : \mathbb{R}^n \to \mathbb{R}^n$.

Satz 5.4 *Sei $I \subseteq \mathbb{R}$ ein abgeschlossenes Intervall, $q \in [0, 1)$ und $\phi : I \to I$ eine Funktion, die für alle $x, y \in I$*

$$|\phi(x) - \phi(y)| \leq q|x - y| \tag{5.3}$$

erfüllt. Besitzt ϕ einen Fixpunkt $x^* \in I$, so konvergiert die Folge

$$x^{(k+1)} = \phi(x^{(k)}), k = 0, 1, \ldots$$

für jeden Startwert $x^{(0)} \in I$ gegen x^* und es gilt

$$|x^{(k)} - x^*| \leq q^k |x^{(0)} - x^*| \text{ für } k = 0, 1, 2, \ldots$$

Beweis: Zunächst ist die Iterationsformel für $x^{(k)}$ ist wohldefiniert, weil $x^k \in I$ für alle k. Die Aussage lässt sich dann für alle $k \in \mathbb{N}_0$ durch Induktion zeigen. Der Induktionsanfang für $k = 0$ ist klar. Für den Induktionsschritt $k \to k + 1$ rechnet man

$$\begin{aligned} |x^{(k+1)} - x^*| &= |\phi(x^{(k)}) - \phi(x^*)| \leq q|x^{(k)} - x^*| \text{ wegen (5.3)} \\ &\leq qq^k|x^{(0)} - x^*| \text{ wegen der Induktionsannahme} \\ &= q^{(k+1)}|x^{(0)} - x^*|. \end{aligned}$$

Daraus folgt die Konvergenz. QED

Mit Hilfe dieses Satzes können wir erklären, warum die dritte Iterationsformel $x^{(k+1)} = \arctan(2x^{(k)})$ in unserem Beispiel für jeden Startwert $x^{(0)} \in I = [1, \infty)$ konvergiert:

- Dazu überlegt man zunächst, dass $\phi(x) \in [1, \infty) = I$ für alle $x \in I$, denn $\phi(1) > 1$ und ϕ ist monoton wachsend.
- Weiterhin besitzt ϕ einen Fixpunkt, denn die Funktion $H(x) = x - \phi(x)$ erfüllt $H(1) < 0$ und $H(x) \to \infty$ für $x \to \infty$. Also hat H nach dem Zwischenwertsatz eine Nullstelle in I und entsprechend hat ϕ in dem Intervall einen Fixpunkt.

- Jetzt muss noch die Kontraktions-Voraussetzung (5.3) nachgewiesen werden. Dazu verwenden wir den Mittelwertsatz, von dem wir wissen, dass für jedes $x, y \in I$ eine Zwischenstelle $\epsilon \in (x, y)$ existiert, so dass

  $$\phi(x) - \phi(y) = \phi'(\epsilon)(x - y),$$

  d.h. dass

  $$|\phi(x) - \phi(y)| = |\phi'(\epsilon)||x - y|.$$

  Kann man nun zeigen, dass $|\phi'(\epsilon)| \leq q < 1$ für alle $\epsilon \in I$ so ist die Kontraktionsbedingung (5.3) erfüllt. In unserem Beispiel rechnet man nach, dass

  $$|\phi'(x)| = \left| \frac{2}{1 + 4x^2} \right| \leq \phi'(1) = \frac{2}{5} < 1,$$

  weil ϕ' monoton fallend ist.

Also sind die Voraussetzungen von Satz 5.4 erfüllt und die Konvergenz der Iterationsformel ist bewiesen.

5.2 Der Banach'sche Fixpunktsatz

In diesem Abschnitt werden wir die Konvergenzeigenschaften der sukzessiven Approximation weiter untersuchen. Unser Ziel ist eine Verallgemeinerung von Satz 5.4 aus dem letzten Abschnitt, bei der wir die Existenz eines Fixpunktes nicht voraussetzen müssen. Außerdem gelingt es, den neuen Satz nicht nur für skalare Funktionen ϕ sondern für Operatoren Φ in beliebigen Banach-Räume X zu zeigen - d.h. die Unbekannte $x \in X$ kann nicht nur ein Vektor, sondern sogar eine Funktion sein.

Wir erinnern zunächst daran, dass jeder vollständige und normierte Raum ein **Banach-Raum** ist, d.h. also dass in einem Banach-Raum jede Cauchy-Folge konvergiert. Weiterhin übertragen wir (5.3) aus dem letzten Abschnitt auf normierte Räume.

Definition 5.5 *Sei X ein Banach-Raum mit Norm $\|\cdot\|$ und $U \subseteq X$ eine abgeschlossene Teilmenge von X. Eine Abbildung $\Phi : U \to X$ heißt* **kontrahierend***, falls es einen reellen Kontraktionsfaktor $q < 1$ gibt, so dass*

$$\|\Phi(x) - \Phi(y)\| \leq q\|x - y\| \text{ für alle } x, y \in U.$$

Wir können nun den Banach'schen Fixpunktsatz formulieren und beweisen.

Satz 5.6 (Banach'scher Fixpunktsatz) *Sei X ein Banach-Raum mit Norm $\|\cdot\|$ und $U \subseteq X$ eine abgeschlossene Teilmenge von X. Sei weiterhin $\Phi : U \to U$ eine kontrahierende Abbildung mit Kontraktionsfaktor $q < 1$. Dann gilt:*

1. Φ *besitzt einen eindeutig bestimmten Fixpunkt* x^*.

2. *Die Iterationsvorschrift der sukzessiven Approximation* $x^{(k+1)} = \Phi(x^{(k)})$, $k = 0, 1, \ldots$ *konvergiert gegen* x^* *für jeden Startwert* $x^{(0)} \in U$.

3. *Es gilt die* **a priori Fehlerschranke**

$$\|x^{(k)} - x^*\| \leq \frac{q^k}{1-q}\|x^{(1)} - x^{(0)}\| \quad \textit{für alle } k = 1, 2 \ldots \tag{5.4}$$

4. *Es gilt die* **a posteriori Fehlerschranke**

$$\|x^{(k)} - x^*\| \leq \frac{q}{1-q}\|x^{(k)} - x^{(k-1)}\| \quad \textit{für alle } k = 1, 2 \ldots \tag{5.5}$$

Beweis: Zunächst ist die Folge $x^{(k)}$ wohldefiniert weil $x^{(k)} \in U$ für alle $k \in \mathbb{N}_0$. Für den Beweis nutzen wir aus, dass in einem Banach-Raum alle Cauchy-Folgen konvergieren und zeigen daher als erstes, dass $x^{(k)}$ eine Cauchy-Folge ist.

Schritt 1: $x^{(k)}$ ist eine Cauchy-Folge: Es gilt

$$\begin{aligned} \|x^{(k)} - x^{(k-1)}\| &= \|\Phi(x^{(k-1)}) - \Phi(x^{(k-2)})\| \\ &\leq q\|x^{(k-1)} - x^{(k-2)}\| \leq \ \ldots \ \leq \\ &\leq q^j\|x^{(k-j)} - x^{(k-j-1)}\| \end{aligned} \tag{5.6}$$

für alle natürlichen Zahlen j mit $0 \leq j \leq k-1$. Mit Hilfe dieser Ungleichung rechnet man nun nach, dass

$$\begin{aligned} \|x^{(l)} - x^{(k)}\| &\leq \|x^{(l)} - x^{(l-1)}\| + \|x^{(l-1)} - x^{(l-2)}\| + \ldots + \|x^{(k+1)} - x^{(k)}\| \\ &\leq q^{l-k}\|x^{(k)} - x^{(k-1)}\| + q^{l-k-1}\|x^{(k)} - x^{(k-1)}\| + \ldots \\ &\qquad \ldots \quad + q\|x^{(k)} - x^{(k-1)}\| \\ &= \|x^{(k)} - x^{(k-1)}\| \sum_{j=1}^{l-k} q^j \\ &\leq \|x^{(k)} - x^{(k-1)}\| \sum_{j=1}^{\infty} q^j \\ &= \|x^{(k)} - x^{(k-1)}\| \frac{q}{1-q} \qquad (5.7) \\ &\leq q^{k-1}\|x^{(1)} - x^{(0)}\| \frac{q}{1-q} = \frac{q^k}{1-q}\|x^{(1)} - x^{(0)}\|. \qquad (5.8) \end{aligned}$$

Weil $\frac{q^k}{1-q} \to 0$ für $k \to \infty$ ist $x^{(k)}$ also eine Cauchy-Folge.

Schritt 2: Existenz des Fixpunktes. Weil $x^{(k)}$ eine Cauchy-Folge ist, gibt es $x^* = \lim_{k\to\infty} x^{(k)}$. Für x^* gilt dann

$$\|\Phi(x^*) - \Phi(x^{(k)})\| \leq q\|x^* - x^{(k)}\| \to 0 \text{ für } k \to \infty,$$

entsprechend haben wir

$$\Phi(x^*) = \lim_{k\to\infty} \Phi(x^{(k)}) = \lim_{k\to\infty} x^{(k+1)} = x^*$$

Schritt 3: Eindeutigkeit des Fixpunktes. Angenommen, $\tilde{x}$ sei ein weiterer Fixpunkt von Φ. Dann gilt

$$\|x^* - \tilde{x}\| = \|\Phi(x^*) - \Phi(\tilde{x})\| \leq q\|x^* - \tilde{x}\|.$$

Weil $q < 1$ folgt daraus, dass $\|x^* - \tilde{x}\| = 0$, also $x^* = \tilde{x}$.

Schritt 4: Fehlerschranken. Wir nutzen die in Schritt 1 aufgestellte Ungleichungskette für

$$\begin{aligned}
\|x^* - x^{(k)}\| &= \lim_{l\to\infty} \|x^{(l)} - x^{(k)}\| \\
&\leq \|x^{(k)} - x^{(k-1)}\| \frac{q}{1-q} \text{ wegen (5.7)} \\
&\leq \frac{q^k}{1-q}\|x^{(1)} - x^{(0)}\| \text{ wegen (5.8).}
\end{aligned}$$

Damit ist der Satz gezeigt. QED

Zum Nachweis der Kontraktion verallgemeinern wir noch das bereits für skalare Funktionen verwendete Kriterium auf den $\mathbb{R}^n$. Dabei bezeichnen wir für eine Funktion $F : \mathbb{R}^n \to \mathbb{R}^m$ die Jacobi-Matrix von F an der Stelle $x \in \mathbb{R}^n$ mit $DF(x)$, das heißt

$$DF(x) = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_1}{\partial x_2} & \cdots & \frac{\partial F_1}{\partial x_n} \\ \vdots & & & \vdots \\ \frac{\partial F_m}{\partial x_1} & \frac{\partial F_m}{\partial x_2} & \cdots & \frac{\partial F_m}{\partial x_n} \end{pmatrix}.$$

Für $F : \mathbb{R} \to \mathbb{R}^n$ bezeichnen wir den Tangentialvektor $DF(x)$ auch einfach mit $f'(x)$.

Lemma 5.7 *Sei $U \subseteq \mathbb{R}^n$ eine konvexe Menge und $\Phi : U \to \mathbb{R}^n$ stetig differenzierbar mit $\|D\Phi(x)\| \leq q < 1$ für alle $x \in U$ (wobei die Matrixnorm $\|\cdot\|$ die der Vektornorm zugeordnete Norm sein soll). Dann ist Φ kontrahierend mit Kontraktionsfaktor q.*

Beweis: Seien $x, y \in U$. Wir definieren eine Abbildung $f : \mathbb{R} \to \mathbb{R}^n$ durch

$$f(t) = \Phi(x + t(y - x)) \quad \text{für } t \in [0, 1].$$

Dann gilt

$$\begin{aligned}
\|\Phi(y) - \Phi(x)\| &= \|f(1) - f(0)\| = \| \int_0^1 f'(t)dt\| \\
&\qquad \text{nach dem Hauptsatz der Differential- und Integralrechnung} \\
&= \| \int_0^1 D\Phi(x + t(y - x))(y - x)dt\| \\
&\qquad \text{nach der multivariaten Kettenregel} \\
&\leq \int_0^1 \|D\Phi(x + t(y - x))\| \, \|y - x\|dt \\
&\leq \|y - x\| \int_0^1 q \; dt = q\|x - y\|.
\end{aligned}$$

QED

Abschließend untersuchen wir noch, wie wir bei der Approximation des Fixpunktes eine Genauigkeit von ε garantieren können. Wir wollen also erreichen, dass

$$\|x^{(k)} - x^*\| \leq \varepsilon$$

wenn k die Iteration ist, bei der wir abbrechen. Dazu können wir sowohl die a-priori als auch die a-posteriori Schranke aus Satz 5.6 nutzen.
Die a-priori Schranke sagt, dass

$$\|x^{(k)} - x^*\| \leq \frac{q^k}{1 - q}\|x^{(1)} - x^{(0)}\| \quad \text{für alle } k = 1, 2 \ldots .$$

$\|x^{(k)} - x^*\| \leq \varepsilon$ ist also gewährleistet, falls

$$\frac{q^k}{1 - q}\|x^{(1)} - x^{(0)}\| \leq \varepsilon,$$

und das lässt sich auflösen zu

$$k \geq \frac{\ln\left(\frac{(1-q)\varepsilon}{\|x^{(1)} - x^{(0)}\|}\right)}{\ln(q)}.$$

Ist der Kontraktionsfaktor q also klein, werden weniger Iterationsschritte benötigt als für einen großen Kontraktionsfaktor q.

Um während des Verfahrens ein Abbruchkriterium zu haben, nutzt man dagegen oft die (schärfere) a posteriori Fehlerschranke aus Satz 5.6, die besagt, dass

$$\|x^{(k)} - x^*\| \leq \frac{q}{1-q}\|x^{(k)} - x^{(k-1)}\| \text{ für alle } k = 1, 2 \ldots$$

und zu dem Abbruchkriterium

$$\frac{q}{1-q}\|x^{(k)} - x^{(k-1)}\| \leq \varepsilon$$

führt. Leider ist der Kontraktionsfaktor q oft nicht bekannt. In diesen Fällen behilft man sich mit folgender Abschätzung von q durch $\hat{q}_k$:

$$\hat{q}_k := \frac{\|x^{(k)} - x^{(k-1)}\|}{\|x^{(k-1)} - x^{(k-2)}\|}.$$

Es gilt $\hat{q}_k \leq q$, denn

$$\hat{q}_k = \frac{\|x^{(k)} - x^{(k-1)}\|}{\|x^{(k-1)} - x^{(k-2)}\|} = \frac{\|\Phi(x^{(k-1)}) - \Phi(x^{(k-2)})\|}{\|x^{(k-1)} - x^{(k-2)}\|} \leq q.$$

Dabei ist zu beachten, dass die Anzahl der Iterationen bei Verwendung von $\hat{q}_l$ möglicherweise zu klein gewählt wird, die Fehlerschranke $\|x^{(k)} - x^*\| \leq \varepsilon$ kann also bei Abbruch des Verfahrens nicht garantiert werden. In der Praxis funktioniert das Abbruchkriterium in der Regel aber ausreichend gut.

Algorithmus 8: Sukzessive Approximation mit heuristischem Abbruchkriterium

Input: `abgeschlossene Menge` $U \subseteq \mathbb{R}^n$, `Kontraktion` $\Phi : U \to U$, `Startwert` $x^{(0)} \in U$, `Toleranzwert` ε.

Schritt 1: $x^{(1)} := \Phi(x^{(0)})$

Schritt 2: $k := 1$

Schritt 3: `Repeat`

Schritt 3.1: $k := k + 1$

Schritt 3.2: $x^{(k)} := \Phi\left(x^{(k-1)}\right)$

Schritt 3.3: $q_k := \frac{\|x^{(k)} - x^{(k-1)}\|}{\|x^{(k-1)} - x^{(k-2)}\|}$

Schritt 3.4: `If` $q_k \geq 1$ `STOP: Φ ist keine Kontraktion.`

Until $\frac{q_k}{1-q_k}\|x^{(k)} - x^{(k-1)}\| \leq \varepsilon$

Ergebnis: `approximierter Fixpunkt` $x^* = x^{(k)}$

5.3 Iterative Verfahren für lineare Gleichungssysteme

Wir wenden nun den Banach'schen Fixpunktsatz auf lineare Operatoren an. Zunächst halten wir uns weiter in Banach-Räumen auf, kommen dann aber zur Lösung linearer Gleichungssysteme (also zum endlich dimensionalen Fall) zurück.

Satz 5.8 *Sei $B : X \to X$ ein linearer beschränkter Operator in einem Banach-Raum $(X, \|\cdot\|)$ mit $\|B\| < 1$ in der der Norm des Banach-Raumes zugeordneten Matrixnorm. Dann gilt*

1. *Der Operator $I - B$ ist invertierbar, das heißt das System $x - Bx = b$ hat genau eine Lösung x^* für jedes $b \in X$.*

2. *Der inverse Operator $(I - B)^{-1}$ ist beschränkt mit*

$$\|(I - B)^{-1}\| \leq \frac{1}{1 - \|B\|}.$$

3. *Die Iterationsvorschrift der sukzessiven Approximation* $x^{(k+1)} = Bx^{(k)} + b,\ k = 0, 1, \ldots$ *konvergiert gegen* x^* *für jeden Startwert* $x^{(0)} \in X$.

4. *Es gilt die* **a priori Fehlerschranke**

$$\|x^{(k)} - x^*\| \leq \frac{\|B\|^k}{1 - \|B\|} \|x^{(1)} - x^{(0)}\| \quad \textit{für alle } k = 1, 2 \ldots$$

5. *Es gilt die* **a posteriori Fehlerschranke**

$$\|x^{(k)} - x^*\| \leq \frac{\|B\|}{1 - \|B\|} \|x^{(k)} - x^{(k-1)}\| \quad \textit{für alle } k = 1, 2 \ldots$$

Beweis: Sei $b \in X$ beliebig aber fest. Definiere den linearen Operator Φ punktweise durch

$$\Phi x := Bx + b \text{ für alle } x \in X$$

Wegen $\|\Phi x - \Phi \tilde{x}\| = \|B(x - \tilde{x})\| \leq \|B\| \|x - \tilde{x}\|$ ist Φ kontrahierend mit $q := \|B\| < 1$. Satz 5.6 ergibt damit direkt die folgenden Aussagen:

ad 1. Es existiert ein eindeutiger Fixpunkt x^*, der $\Phi x^* = x^*$ erfüllt. Weil

$$\Phi x = x \iff Bx + b = x \iff (I - B)x = b$$

gibt es also eine eindeutige Lösung von $(I - B)x = x - Bx = b$ und $(I - B)$ ist invertierbar.

ad 3. $x^{(k+1)} = \Phi x^{(k)} = Bx^{(k)} + b$ konvergiert gegen x^* für jeden Startwert $x^{(0)}$.

ad 4. und 5. Hier folgt die Behauptung direkt mit $q := \|B\|$.

Als letzter Punkt bleibt noch die zweite Aussage zu zeigen, also die Beschränktheit der linearen Abbildung $(I - B)^{-1}$. Dazu definieren wir die Folge $x^{(k)}$ der sukzessiven Approximation mit Startwert $x^{(0)} := b$. Es ergibt sich

$$\begin{aligned} x^{(0)} &= b \\ x^{(1)} &= Bx^{(0)} + b = Bb + b \\ x^{(2)} &= Bx^{(1)} + b = B^2 b + Bb + b \\ &\vdots \\ x^{(k)} &= \sum_{j=0}^{k} B^j b. \end{aligned}$$

Deswegen gilt

$$\|x^{(k)}\| \leq \sum_{j=0}^{k} \|B^j b\| \leq \|b\| \sum_{j=0}^{k} \|B\|^j \leq \frac{\|b\|}{1 - \|B\|}$$

Weiter wissen wir von Aussage 3 und 1, dass $x^{(k)} \to x^* = (I-B)^{-1}b$. Daraus folgt, dass

$$\|(I-B)^{-1}b\| \leq \frac{\|b\|}{1-\|B\|}.$$

Weil b beliebig war, gilt diese Aussage für alle $b \in X$. Somit erhält man

$$\|(I-B)^{-1}\| = \sup_{b \in X} \frac{\|(I-B)^{-1}b\|}{\|b\|} \leq \frac{1}{1-\|B\|}.$$

QED

Im endlich-dimensionalen Fall sind alle Normen äquivalent, so dass aus der Konvergenz bezüglich einer Norm die Konvergenz in allen anderen Normen folgt. Da es unhandlich sein kann, das Kriterium für verschiedene Normen zu testen, wollen wir im folgenden ein notwendiges und hinreichendes Kriterium für die Konvergenz der sukzessiven Approximation herleiten. Dieses Kriterium wird über den Spektralradius $\rho(B)$ der obigen Matrix B formuliert werden. Um das Kriterium herleiten zu können, benötigen wir das folgende Resultat aus der Linearen Algebra.

Satz 5.9 (Lemma von Schur) *Sei $A \in \mathbb{K}^{n,n}$ eine Matrix. Dann gibt es eine unitäre Matrix Q so, dass*

$$Q^*AQ = R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ & r_{22} & \dots & r_{2n} \\ & & \ddots & \vdots \\ 0 & & & r_{nn} \end{pmatrix}$$

Mit Hilfe des Lemmas von Schur zeigen wir nun erst die folgende Aussage.

Lemma 5.10 *Sei $A \in \mathbb{K}^{n,n}$. Dann gilt $\rho(A) \leq \|A\|$ für jede zu einer Vektornorm auf $\mathbb{K}^n$ passende Matrixnorm $\|\cdot\|$.*
Andererseits gibt es zu jedem $\varepsilon > 0$ eine Norm $\|\cdot\|_\varepsilon$ auf $\mathbb{K}^n$ so dass

$$\|A\|_\varepsilon \leq \rho(A) + \varepsilon.$$

Beweis: Zum Beweis des ersten Teils der Aussage wählen wir einen Eigenwert λ von A mit zugehörigem normierten Eigenvektor $u \in \mathbb{K}^n$. Dann gilt, dass

$$\|A\| = \sup_{x:\|x\|=1} \|Ax\| \geq \|Au\| = \|\lambda u\| = |\lambda|.$$

Das gilt für alle Eigenwerte λ, also auch für den betragsmäßig größten.

Sei nun $\varepsilon > 0$ gegeben. Wir können ohne Beschränkung der Allgemeinheit annehmen, dass A nicht die Nullmatrix ist. Wir werden nun die gesuchte Norm $\|\cdot\|_\varepsilon$ konstruieren.

Nach dem Lemma von Schur (Satz 5.9) finden wir eine unitäre Matrix Q, so dass

$$R := Q^*AQ = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ & r_{22} & \dots & r_{2n} \\ & & \ddots & \vdots \\ 0 & & & r_{nn} \end{pmatrix}$$

eine obere Dreiecksmatrix (nicht die Nullmatrix) ist. Zunächst beobachten wir, dass

$$\begin{aligned} \det(\lambda I - A) &= \det(Q^*)\det(\lambda I - A)\det(Q) = \det(Q^*(\lambda I - A)Q) \\ &= \det(\lambda I - Q^*AQ) = \det(\lambda I - R) \\ &= (\lambda - r_{11})(\lambda - r_{22})\dots(\lambda - r_{nn}), \end{aligned}$$

also sind die Eigenwerte von A als Nullstellen des charakteristischen Polynoms genau die Diagonalelemente von R. Man definiert nun

$$\begin{aligned} r &:= \max_{i,j} |r_{ij}| > 0 \\ \delta &:= \min\left\{1, \frac{\varepsilon}{(n-1)r}\right\}, \text{ und} \\ D &:= \operatorname{diag}(1, \delta, \delta^2, \dots, \delta^{n-1}) \end{aligned}$$

Weil $\delta > 0$ ist D invertierbar und $D^{-1} = \operatorname{diag}(1, \delta^{-1}, \delta^{-2}, \dots, \delta^{-(n-1)})$. Wir berechnen nun

$$\begin{aligned} C: &= D^{-1}RD \\ &= D^{-1}\begin{pmatrix} r_{11} & \delta r_{12} & \delta^2 r_{13} & \dots & \delta^{n-1} r_{1n} \\ & \delta r_{22} & \delta^2 r_{23} & \dots & \delta^{n-1} r_{2n} \\ & & \delta^2 r_{33} & \dots & \delta^{n-1} r_{3n} \\ & & & \ddots & \vdots \\ 0 & & & & \delta^{n-1} r_{nn} \end{pmatrix} = \begin{pmatrix} r_{11} & \delta r_{12} & \delta^2 r_{13} & \dots & \delta^{n-1} r_{1n} \\ & r_{22} & \delta r_{23} & \dots & \delta^{n-2} r_{2n} \\ & & r_{33} & \dots & \delta^{n-3} r_{3n} \\ & & & \ddots & \vdots \\ 0 & & & & r_{nn} \end{pmatrix}. \end{aligned}$$

Satz 3.20 (siehe Seite 57) liefert, dass

$$\begin{aligned} \|C\|_\infty &= \max_{i=1,\dots,n} \sum_{j=1}^{n} |c_{ij}| \\ &\le \max_{i=1,\dots,n} r_{ii} + \delta r(n-1) \text{ weil } \delta^j \le \delta \\ &\le \rho(A) + \frac{\varepsilon}{(n-1)r} r(n-1) = \rho(A) + \varepsilon \end{aligned}$$

Setze

$$V := QD$$

und definiere damit

$$\|x\|_\varepsilon := \|V^{-1}x\|_\infty.$$

Das ist eine Norm, da V regulär. Um zu zeigen, dass diese Norm die gewünschte Eigenschaft hat, bemerken wir zunächst, dass

$$V^{-1}AV = D^{-1}Q^*AQD = D^{-1}RD = C$$

gilt. Damit erhält man

$$\begin{aligned}\|Ax\|_\varepsilon &= \|V^{-1}Ax\|_\infty \\ &= \|CV^{-1}x\|_\infty \\ &\leq \|C\|_\infty \|V^{-1}x\|_\infty \\ &= \|C\|_\infty \|x\|_\varepsilon,\end{aligned}$$

also ist

$$\|A\|_\varepsilon = \sup_{x\in\mathbb{K}^n\setminus\{0\}} \frac{\|Ax\|_\varepsilon}{\|x\|_\varepsilon} \leq \|C\|_\infty = \rho(A) + \varepsilon.$$

QED

Nun können wir (endlich!) das angekündigte Kriterium für die Konvergenz der sukzessiven Approximation formulieren.

Satz 5.11 *Sei $B \in \mathbb{K}^{n,n}$. Die Folge $x^{(k+1)} = Bx^{(k)} + b$ mit $k = 0, 1, 2, \ldots$ konvergiert für jedes $b \in \mathbb{K}^n$ und jeden Startwert $x^{(0)} \in \mathbb{K}^n$ genau dann wenn $\rho(B) < 1$.*

Beweis:

"'⇐"‘: Sei $\rho(B) < 1$. Nach Lemma 5.10 existiert eine Norm $\|\cdot\|_\varepsilon$ so dass $\|B\|_\varepsilon \leq \rho(B) + \varepsilon$ für jedes $\varepsilon > 0$. Wähle ε nun so, dass

$$\rho(B) + \varepsilon < 1,$$

dann konvergiert $x^{(k)}$ bezüglich der Norm $\|\cdot\|_\varepsilon$. Da in $\mathbb{K}^n$ alle Normen äquivalent (Satz 3.12) sind, folgt die Konvergenz in jeder Norm.

"'⇒"‘: Angenommen, $\rho(B) \geq 1$. Dann gibt es einen Eigenwert $\lambda \geq 1$ und einen zugehörigen Eigenvektor $v \neq 0$. Starte das Verfahren der sukzessiven Approximation für $b = v$ mit dem Startvektor $x^{(0)} = v$. Man erhält

$$\begin{aligned}x^{(0)} &= v \\ x^{(1)} &= Bv + v = \lambda v + v \\ x^{(2)} &= B(\lambda v + v) + v = \lambda^2 v + \lambda v + v \\ &\vdots \\ x^{(k)} &= \left(\sum_{j=0}^{k} \lambda^j\right) v \to \infty \quad \text{weil } \lambda \geq 1.\end{aligned}$$

Also konvergiert die Folge $x^{(k)}$ in diesem Fall nicht. QED

Wir möchten die Ergebnisse nun konkret auf die Lösung linearer Gleichungssysteme anwenden. Sei also ein lineares Gleichungssystem

$$Ax = b$$

mit $A \in \mathbb{K}^{n,n}, b \in \mathbb{K}^n$ gegeben. Wir bringen das Gleichungssystem mit Hilfe einer regulären Matrix M in Fixpunktform und erhalten die äquivalente Fixpunktgleichung

$$x + M^{-1}(b - Ax) = x,$$

die zur sukzessiven Approximation

$$x^{(k+1)} = x^{(k)} + M^{-1}(b - Ax^{(k)}) \text{ beziehungsweise } M(x^{(k+1)} - x^{(k)}) = b - Ax^{(k)}$$

führt. Numerisch kann man $x^{(k+1)}$ in jeder Iteration durch das sukzessive Lösen der beiden Systeme

$$Mw^{(k+1)} = b - Ax^{(k)} \text{ und } x^{(k+1)} = x^{(k)} + w^{(k+1)} \tag{5.9}$$

ermitteln. Das klappt besonders gut, wenn M z.B. eine Dreiecksmatrix ist, die gewährleistet, dass das System (5.9) effizient lösbar ist.

Wie soll man M wählen? Nach Satz 5.11 konvergiert die Folge $x^{(k+1)} = Bx^{(k)} + b$ genau dann, wenn $\rho(B) < 1$. Weil in unserem Fall

$$x^{(k+1)} = (I - M^{-1}A)x^{(k)} + M^{-1}b$$

muss also $\rho(I - M^{-1}A) < 1$ gelten.

Für die folgenden Verfahren zerlegen wir die gegebenen Matrix A in

$$A = A_D + A_L + A_R,$$

wobei $A_D = \text{diag}(a_{11}, a_{22}, \ldots, a_{nn})$ und

$$A_L = \begin{pmatrix} 0 & \ldots & 0 \\ & \ddots & \vdots \\ a_{ij} & & 0 \end{pmatrix}, \text{ und } A_R = \begin{pmatrix} 0 & & a_{ij} \\ \vdots & \ddots & \\ 0 & \ldots & 0 \end{pmatrix}$$

den Anteil des unteren und oberen Dreiecks aus A beinhalten. Weiterhin setzen wir voraus, dass (eventuell nach Pivotisierungs-Schritten) die Inverse

$$A_D^{-1} \text{ existiert,} \tag{5.10}$$

d.h. dass alle Elemente der Hauptdiagonalen von A nicht Null sind. (Wie wir vom Gauss-Verfahren wissen, lässt sich das bei regulären Matrizen immer erreichen.) Wir betrachten nun zunächst zwei vom Konzept her sehr ähnliche Verfahren, das *Gesamtschritt-Verfahren* und das *Einzelschritt-Verfahren*.

Gesamtschritt - oder Jacobi-Verfahren

Im so genannten Gesamtschritt-Verfahren (GSV) wählt man die nach unserer Voraussetzung (5.10) reguläre Matrix $M = A_D$. Als Fixpunktgleichung erhält man

$$\begin{aligned} x &= x + A_D^{-1}(b - Ax) = x - A_D^{-1}Ax + A_D^{-1}b \\ &= -A_D^{-1}(A - A_D)x + A_D^{-1}b \\ &= -A_D^{-1}(A_L + A_R)x + A_D^{-1}b \end{aligned} \tag{5.11}$$

Das Verfahren der sukzessiven Approximation ergibt sich folglich zu

$$x^{(k+1)} = -A_D^{-1}(A_L + A_R)x^{(k)} + A_D^{-1}b,\ k = 0, 1, 2, \ldots$$

mit der Iterationsmatrix

$$B = I - A_D^{-1}A = -A_D^{-1}(A_L + A_R) \tag{5.12}$$

Komponentenweise kann man schreiben

$$x_i^{(k+1)} = -\sum_{j \in \{1,\ldots,n\} \setminus \{i\}} \frac{a_{ij}}{a_{ii}} x_j^k + \frac{b_i}{a_{ii}},\ i = 1, \ldots, n.$$

Das Konvergenzverhalten analysiert der folgende Satz.

Satz 5.12 *Die Matrix $A = (a_{ij}) \in \mathbb{K}^n$ genüge einer der drei folgenden Bedingungen:*

Zeilensummenkriterium: $q_\infty = \max_{i=1,\ldots,n} \sum_{j \in \{1,\ldots,n\} \setminus \{i\}} \left| \frac{a_{ij}}{a_{ii}} \right| < 1$

Spaltensummenkriterium: $q_1 = \max_{j=1,\ldots,n} \sum_{i \in \{1,\ldots,n\} \setminus \{j\}} \left| \frac{a_{ij}}{a_{jj}} \right| < 1$

Quadratsummenkriterium: $q_2 = \sqrt{\sum_{i,j \in \{1,\ldots,n\}, i \neq j} \left| \frac{a_{ij}}{a_{ii}} \right|^2} < 1$

Dann konvergiert das Jacobi-Verfahren bezüglich jeder Norm im $\mathbb{K}^n$ für jede rechte Seite $b \in \mathbb{K}^n$ und für jeden Startwert $x^{(0)} \in \mathbb{K}^n$, und zwar gegen die eindeutig bestimmte Lösung x^* von $Ax^* = b$. Weiterhin gilt für $p \in \{1, 2, \infty\}$:

- *A priori Fehlerschranke:* $\|x^{(k)} - x^*\|_p \leq \frac{q_p^k}{1-q_p} \|x^{(1)} - x^{(0)}\|_p$
- *A posteriori-Fehlerschranke:* $\|x^{(k)} - x^*\|_p \leq \frac{q_p}{1-q_p} \|x^{(k)} - x^{(k-1)}\|_p$

Beweis: Wir untersuchen die Norm der Iterationsmatrix

$$B = -A_D^{-1}(A_L + A_R).$$

Nach Satz 3.20 gilt, dass

$$\| - A_D^{-1}(A_L + A_R)\|_p = q_p < 1 \text{ für } p \in \{\infty, 1\},$$

und aus Lemma 3.27 folgt, dass

$$\| - A_D^{-1}(A_L + A_R)\|_2 \leq \| - A_D^{-1}(A_L + A_R)\|_F = q_2 < 1.$$

Wir können also Satz 5.8 anwenden, aus dem sich der Rest der Behauptungen direkt ergibt. QED

Bemerkung: Die drei Konvergenzkriterien sind nicht äquivalent!

Der Algorithmus ergibt sich in kanonischer Weise:

Algorithmus 9: Jacobi-Verfahren

Input: `Reguläre Matrix` $A \in \mathbb{K}^{n,n}$ `mit` $a_{ii} \neq 0$ `für` $i = 1, \ldots, n$, $b \in \mathbb{K}^n$, $x^{(0)} \in \mathbb{K}^n$.

Schritt 1: $k := 0$

Schritt 2: `Repeat`

Schritt 2.1: `For` $i = 1, \ldots, n$ `do:` $x_i^{(k+1)} := \frac{1}{a_{ii}} \left(- \sum_{j \in \{1,\ldots,n\} \setminus \{i\}} a_{ij} x_j^{(k)} + b_j \right)$

Schritt 2.2: $k := k + 1$

Until `Abbruchkriterium`

Ergebnis: `Approximierte Lösung` x^* `von` $Ax^* = b$.

Das Verfahren kann man abbrechen, wenn die a posteriori Fehlerschranke, also $\frac{q_p}{1-q_p}\|x^{(k)} - x^{(k-1)}\|_p$ klein genug ist.

Einzelschritt - oder Gauß-Seidel-Verfahren

Im jetzt zu besprechenden Einzelschritt-Verfahren (ESV) wählt man $M = A_D + A_L$. Nach der Voraussetzung (5.10) ist M regulär. Als Fixpunktgleichung erhält man

$$\begin{aligned} x &= x + (A_D + A_L)^{-1}(b - Ax) \\ &= -(A_D + A_L)^{-1}(-(A_D + A_L) + A)x + (A_D + A_L)^{-1}b \\ &= -(A_D + A_L)^{-1}A_R x + (A_D + A_L)^{-1}b \end{aligned} \tag{5.13}$$

Das Verfahren der sukzessiven Approximation ergibt sich folglich zu

$$x^{(k+1)} = -(A_D + A_L)^{-1} A_R x^{(k)} + (A_D + A_L)^{-1} b,\ k = 0, 1, 2, \ldots$$

Die Iterationsmatrix ist entsprechend

$$C = I - (A_D + A_L)^{-1} A = -(A_D + A_L)^{-1} A_R.$$

Rechnerisch nutzt man die Umformulierung zu

$$(A_D + A_L) x^{(k+1)} = -A_R x^{(k)} + b,\ k = 0, 1, 2, \ldots \tag{5.14}$$

Um das komponentenweise zu schreiben löst man dieses System mittels Vorwärtselimination auf, um die Unbekannten $x_i^{(k+1)}$ für $i = 1, \ldots, n$ zu bestimmen. Man erhält

$$x_i^{(k+1)} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{(k+1)} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j^{(k)} + \frac{b_i}{a_{ii}},\ i = 1, \ldots, n.$$

Die Formel stimmt fast mit der entsprechenden komponentenweisen Iterationsformel des Jacobi-Verfahrens überein. Der Unterschied besteht lediglich darin, dass beim vorliegenden Gauß-Seidel-Verfahren zur Berechnung von $x_i^{(k+1)}$ die neuen (und hoffentlich besseren) Werte $x_j^{(k+1)}$ für $j = 1, \ldots, i-1$ herangezogen werden anstatt der Werte $x_j^{(k)}$ wie im Jacobi-Verfahren. Das ist der Grund, warum das Gauß-Seidel-Verfahren in den meisten Fällen schneller konvergiert als das Jacobi-Verfahren.
Über das Konvergenzverhalten gibt der folgende Satz Auskunft.

Satz 5.13 *Die Matrix $A = (a_{ij}) \in \mathbb{K}^{n,n}$ genüge dem Kriterium nach Sassenfeld:*

$$p := \max_{i=1,\ldots,n} p_i < 1$$

mit den Werten

$$\begin{aligned} p_1 &:= \sum_{j=2}^{n} \left| \frac{a_{1j}}{a_{11}} \right| \\ p_i &:= \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| p_j + \sum_{j=i+1}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| \quad \textit{für } i = 2, \ldots, n \end{aligned}$$

Dann konvergiert das Gauß-Seidel-Verfahren für jede rechte Seite $b \in \mathbb{K}^n$ und bei beliebigem $x^{(0)} \in \mathbb{K}^n$ gegen die eindeutig bestimmte Lösung x^* von $Ax^* = b$. Weiterhin gilt:

- *A priori-Fehlerschranke:* $\|x^{(k)} - x^*\|_\infty \leq \frac{p^k}{1-p} \|x^{(1)} - x^{(0)}\|_\infty$
- *A posteriori Fehlerschranke:* $\|x^{(k)} - x^*\|_\infty \leq \frac{p}{1-p} \|x^{(k)} - x^{(k-1)}\|_\infty$

Beweis: Wir wollen die Zeilensummennorm der Iterationsmatrix $(A_D + A_L)^{-1}A_R$ abschätzen, also zeigen, dass

$$\sup_{z:\|z\|_\infty=1} \| - (A_D + A_L)^{-1}A_R z\|_\infty < 1.$$

Dazu betrachten wir für beliebiges z mit $\|z\|_\infty = 1$

$$x := x(z) := -(A_D + A_L)^{-1}A_R z$$

und berechnen im folgenden $\|x\|_\infty$. Wir fassen x als Lösung des Gleichungssystems

$$(A_D + A_L)x = -A_R z$$

auf. Vorwärtselimination ergibt

$$x_i = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} z_j \text{ für } i = 1, \ldots, n$$

Wir zeigen nun, dass $|x_i| \le p_i$ für $i = 1, \ldots, n$:

Induktionsanfang: $i = 1$. Weil $|z_i| \le 1$ für alle i erhält man:

$$|x_1| = \left| \sum_{j=2}^{n} \frac{a_{1j}}{a_{11}} z_j \right| \le \sum_{j=2}^{n} \left| \frac{a_{1j}}{a_{11}} \right| |z_j| \le \sum_{j=2}^{n} \left| \frac{a_{1j}}{a_{11}} \right| = p_1$$

Induktionsschritt: $i - 1 \to i$.

$$\begin{aligned} |x_i| &= \left| -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} z_j \right| \\ &\le \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| \underbrace{|x_j|}_{\le p_j} + \sum_{j=i+1}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| \underbrace{|z_j|}_{\le 1} \\ &\le \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| p_j + \sum_{j=i+1}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| = p_i \end{aligned}$$

Also gilt $\|x\|_\infty \le p$ und damit

$$\|(A_D+A_L)^{-1}A_R\|_\infty = \sup_{z\in\mathbb{K}^n:\|z\|_\infty=1} \|(A_D+A_L)^{-1}A_R z\|_\infty = \sup_{z\in\mathbb{K}^n:\|z\|_\infty=1} \|x(z)\|_\infty \le p.$$

Weil $p < 1$ vorausgesetzt war, folgt die Behauptung nach Satz 5.8. QED

Übung: Zeigen Sie, dass die folgende Matrix das Sassenfeld-Kriterium erfüllt:

$$A = \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}$$

Bemerkung: In Satz 5.17 werden wir zeigen, dass das Gauss-Seidel-Verfahren bei Gleichungssystemen mit hermitescher und positiv definiter Koeffizientenmatrix konvergiert.

Der Vollständigkeit halber sei der Algorithmus des Einzelschrittverfahrens ebenfalls skizziert.

Algorithmus 10: Gauß-Seidel-Verfahren

Input: `Reguläre Matrix` $A \in \mathbb{K}^{n,n}$ `mit` $a_{ii} \neq 0$ `für` $i = 1, \ldots, n$, $b \in \mathbb{K}^n$, $x^{(0)} \in \mathbb{K}^n$.

Schritt 1: $k := 0$

Schritt 2: `Repeat`

Schritt 2.1: `For` $i = 1, \ldots, n$ `do:`
$x_i^{(k+1)} := \frac{1}{a_{ii}} \left(- \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} + b_j \right)$

Schritt 2.2: $k := k + 1$

Until `Abbruchkriterium`

Ergebnis: `Approximierte Lösung` x^* `von` $Ax^* = b$.

Relaxations-Verfahren

Die Idee der Relaxations-Verfahren besteht darin, die Konvergenz des Gesamtschrittverfahrens beziehungsweise des Einzelschrittverfahrens zu verbessern, indem man durch Einführen eines so genannten Relaxations-Parameters den Spektralradius der Iterationsmatrix verkleinert.

Wir betrachten zunächst das Gesamtschritt-Verfahren. Die Iterationsvorschrift ergibt sich nach (5.11) auf Seite 104:

$$x^{(k+1)} = x^{(k)} + A_D^{-1}(b - Ax^{(k)}).$$

In jedem Iterationsschritt wird also $x^{(k)}$ durch das A_D^{-1}-fache des **Residuums** $z^{(k)} = b - Ax^{(k)}$ korrigiert. Dabei ist oft zu beobachten, dass die Korrektur um einen festen Faktor zu klein ist. Deshalb kann es sinnvoll sein, den Wert um $\omega z^{(k)}$ statt um $z^{(k)}$ zu ändern, wobei ω ein beliebiger positiver Parameter sein darf. Das resultierende Verfahren ist das relaxierte Gesamtschrittverfahren.

Definition 5.14 *Das Iterationsverfahren*

$$x^{(k+1)} = x^{(k)} + \omega A_D^{-1}(b - Ax^{(k)})$$

heißt für $\omega > 0$ **Gesamtschritt-Relaxationsverfahren**.

Komponentenweise berechnen sich die Werte $x_i^{(k+1)}$ durch

$$x_i^{(k+1)} = x_i^{(k)} + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^{n} a_{ij} x_j^{(k)} \right), i = 1, \ldots, n$$

Es gilt

$$\begin{aligned} x^{(k+1)} &= x^{(k)} + \omega A_D^{-1}(b - Ax^{(k)}) \\ &= (I - \omega A_D^{-1} A)x^{(k)} + \omega A_D^{-1} b \\ &= \left[I - \omega I + \omega(-A_D^{-1} A + I) \right] x^{(k)} + \omega A_D^{-1} b \\ &= \left[(1-\omega) I + \omega B \right] x^{(k)} + \omega A_D^{-1} b, \end{aligned} \tag{5.15}$$

wobei im letzten Schritt $B = I - A_D^{-1} A$ die Iterationsmatrix des Gesamtschrittverfahrens aus (5.12) bezeichnet (siehe Seite 104). Die Iterationsmatrix des Gesamtschritt-Relaxationsverfahrens mit Relaxationsparameter ω bezeichnen wir im folgenden mit

$$B_\omega = (I - \omega A_D^{-1} A) = (1 - \omega) I + \omega B.$$

Wir bemerken, dass die Matrix des Gesamtschrittverfahrens gerade $B = B_1$ ist.

Satz 5.11 legt nahe, den Relaxationsparameter ω so zu wählen, dass der Spektralradius der Iterationsmatrix B_ω möglichst klein wird. Der folgende Satz gibt Auskunft darüber, wie dieses Ziel erreicht werden kann.

Satz 5.15 *Die zum Gesamtschrittverfahren gehörende Iterationsmatrix* $B = I - A_D^{-1} A$ *habe nur reelle Eigenwerte und einen Spektralradius* $\rho(B) < 1$. *Sei weiterhin* $-1 < \lambda_{\min}$ *der kleinste Eigenwert von* B *und* $\lambda_{\max} < 1$ *der größte. Für die Iterationsmatrix des Gesamtschritt-Relaxationsverfahrens*

$$B_\omega = (1 - \omega) I + \omega B$$

gilt dann:

$$\rho(B_\omega) \text{ wird minimal für } \omega^* = \frac{2}{2 - \lambda_{\min} - \lambda_{\max}}.$$

Im Fall $\lambda_{min} \neq -\lambda_{\max}$ *erhält man außerdem* $\rho(B_\omega) < \rho(B)$.

Beweis: Zunächst bemerken wir dass für $\omega \neq 0$

$$Bu = \lambda u \Longleftrightarrow [(1-\omega)I + \omega B]\, u = [(1-\omega) + \omega\lambda]\, u$$

gilt. Das heißt, λ ist Eigenwert von B genau dann wenn $(1-\omega)+\omega\lambda$ Eigenwert von B_ω ist. Weil $\omega > 0$ erhält man insbesondere, dass

$$\begin{array}{ll} (1-\omega)+\omega\lambda_{\min} & \text{der kleinste Eigenwert von } B_\omega \text{ ist, und} \\ (1-\omega)+\omega\lambda_{\max} & \text{der größte.} \end{array}$$

Bei gegebenem ω ist der Spektralradius der Matrix B_ω folglich

$$\rho(B_\omega) = \max\{-(1-\omega)-\omega\lambda_{\min}, (1-\omega)+\omega\lambda_{\max}\}$$

Jetzt möchten wir ω so bestimmen, dass dieser Ausdruck möglichst klein wird. Dazu überlegt man sich, dass die beiden Funktionen

$$\begin{aligned} -(1-\omega)-\omega\lambda_{\min} &= -1+\omega(1-\lambda_{\min}) \\ (1-\omega)+\omega\lambda_{\max} &= 1+\omega(\lambda_{\max}-1) \end{aligned}$$

Geraden sind. Das Maximum von zwei Geraden ist eine konvexe Funktion, die aus zwei linearen Abschnitten besteht und ihr eindeutiges Minimum genau am Schnittpunkt der beiden Geraden annimmt, falls dieser existiert. In unserem Fall ist das gegeben, da die beiden Steigungen der Geraden aufgrund der Bedingung $\lambda_{\min} < \lambda_{\max} < 1$

$$1-\lambda_{\min} \neq \lambda_{\max} - 1$$

erfüllen; die Geraden sind also nicht parallel. Ihr eindeutiger Schnittpunkt errechnet sich durch Auflösen der Gleichung

$$-(1-\omega)-\omega\lambda_{\min} = (1-\omega)+\omega\lambda_{\max}$$

und liegt entsprechend bei

$$w^* = \frac{2}{2-\lambda_{\min}-\lambda_{\max}}.$$

Für $\lambda_{\min} \neq -\lambda_{\max}$ folgt $\omega^* \neq 1$. Weiter ist das Minimum ω^* eindeutig. Also folgt in diesem Fall, dass

$$\rho(B_{\omega^*}) < \rho(B),$$

d.h., der Spektralradius von der Iterationsmatrix des Gesamtschritt-Relaxationsverfahrens B_{ω^*} ist in diesem Fall echt kleiner als der Spektralradius der Matrix B des (unrelaxierten) Gesamtschrittverfahrens. QED

Der optimale Relaxationskoeffizient ω^* liegt also im Bereich $(0, \infty)$.

- Ist $\omega^* < 1$ so spricht man von *Unterrelaxation*. Sie tritt auf, falls $-\lambda_{\min} > \lambda_{\max}$.
- Für $\omega^* = 1$ (also wenn $-\lambda_{\min} = \lambda_{\max}$) erhält man das normale Gesamtschrittverfahren.
- Ist $\omega^* > 1$ so spricht man von *Überrelaxation*. Sie tritt auf, falls $-\lambda_{\min} < \lambda_{\max}$.

Um ω^* zu berechnen sind scharfe Schranken für die Eigenwerte der Matrix A (inklusive Vorzeichen) nötig.

Das Gesamtschritt-Relaxationsverfahren hat noch eine andere Interpretation: Bezeichnet man die (unrelaxierte) Iterierte aus dem Gesamtschrittverfahren mit

$$z^{(k+1)} = Bx^{(k)} + A_D^{-1}b,$$

dann gilt nach (5.15), dass

$$x^{(k+1)} = (1-\omega)x^{(k)} + \omega z^{(k+1)}, \tag{5.16}$$

der neue Wert $x^{(k+1)}$ entsteht also, indem man zwischen dem letzten Wert $x^{(k)}$ und dem Wert $z^{(k+1)}$ aus dem Gesamtschrittverfahren linear interpoliert.

Wir untersuchen nun, wie man ein Relaxationsverfahren bezüglich des Einzelschrittverfahrens definieren kann. Im Einzelschrittverfahren (siehe (5.13)) hatten wir die Fixpunktgleichung

$$x = x + (A_D + A_L)^{-1}(b - Ax),$$

aus der sich die Iterationsmatrix

$$C = I - (A_D + A_L)^{-1}A = -(A_D + A_L)^{-1}A_R$$

ergibt. Zur numerischen Berechnung wurde die Umformulierung

$$(A_D + A_L)x^{(k+1)} = -A_R x^{(k)} + b, \;\; k = 0, 1, 2 \ldots$$

angegeben, die wir jetzt weiter zu

$$A_D x^{(k+1)} = b - A_L x^{(k+1)} - A_R x^{(k)}$$

umformulieren. Wir definieren das Relaxationsverfahren jetzt ähnlich wie für das Gesamtschrittverfahren, indem wir auch hier die auf der linken Seite im Einzelschrittverfahren auftretenden $x^{(k+1)}$ zu $z^{(k+1)}$ umbenennen. Das heißt, wir schreiben obige Gleichung als

$$A_D z^{(k+1)} = b - A_L x^{(k+1)} - A_R x^{(k)}. \tag{5.17}$$

Wie in (5.16) wählen wir den relaxierten Wert für $x^{(k+1)}$ als

$$x^{(k+1)} = (1-\omega)x^{(k)} + \omega z^{(k+1)}.$$

Diese Definition möchten wir nun in (5.17) einsetzen. Dazu multiplizieren wir (5.17) mit ω und substituieren $\omega z^{(k+1)}$ durch $x^{(k+1)} - (1-\omega)x^{(k)}$ wie in (5.16) gefordert. Man erhält

$$A_D x^{(k+1)} = (1-\omega)A_D x^{(k)} + \omega b - \omega A_L x^{(k+1)} - \omega A_R x^{(k)}, \tag{5.18}$$

was sich zu

$$(A_D + \omega A_L)x^{(k+1)} = \left[(1-\omega)A_D - \omega A_R\right] x^{(k)} + \omega b$$

und schließlich zu

$$x^{(k+1)} = (A_D + \omega A_L)^{-1} \left[(1-\omega)A_D - \omega A_R\right] x^{(k)} + \omega (A_D + \omega A_L)^{-1} b$$

umformulieren lässt. Daraus ergibt sich die Iterationsmatrix für das Einzelschritt-Relaxationsverfahren in Abhängigkeit von ω zu

$$C_w = (A_D + \omega A_L)^{-1} \left[(1-\omega)A_D - \omega A_R\right]. \tag{5.19}$$

Man sieht, dass auch hier $C_1 = C$ gilt, d.h. für den Relaxationsparameter $\omega = 1$ erhält man die Iterationsmatrix aus dem normalen Einzelschrittverfahren.

Um die Werte $x_i^{(k+1)}$ komponentenweise zu bestimmen, multipliziert man (5.18) von links mit A_D^{-1} und formuliert die entstehende Gleichung dann folgendermaßen um:

$$\begin{aligned} x^{(k+1)} &= (1-\omega)x^{(k)} + \omega A_D^{-1} b - \omega A_D^{-1} A_L x^{(k+1)} - \omega A_D^{-1} A_R x^{(k)} \\ &= x^{(k)} + \omega A_D^{-1}(b - A_L x^{(k+1)} - A_D x^{(k)} - A_R x^{(k)}) \\ &= x^{(k)} + \omega A_D^{-1}(b - A_L x^{(k+1)} - (A_D + A_R) x^{(k)}) \end{aligned}$$

Man bestimmt dann $x^{(k+1)}$ via

$$x_i^{(k+1)} = x_i^{(k)} + \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij} x_j^{(k)} \right) \quad i = 1, \ldots, n.$$

Das Verfahren nennt man auch *Successive overrelaxation*, abgekürzt als SOR-Verfahren (obwohl man streng genommen nur für $\omega > 1$ von einer Überrelaxation sprechen sollte).

Als nächstes untersuchen wir, für welche Relaxationsparameter w wir Konvergenz erwarten können. Zunächst geben wir ein negatives Ergebnis.

Satz 5.16 *Sei $A \in \mathbb{K}^{n,n}$ mit $a_{ii} \neq 0$ für $i = 1, \ldots, n$. Dann gilt*

$$\rho(C_w) \geq |\omega - 1|.$$

Insbesondere ist $\rho(C_w) \geq 1$ falls $\omega \notin (0, 2)$, d.h. das SOR-Verfahren konvergiert in diesen Fällen im allgemeinen nicht.

Beweis: Wir schreiben die Iterationsmatrix C_ω um zu

$$\begin{aligned} C_\omega &= (A_D + \omega A_L)^{-1} A_D A_D^{-1} \left[(1-\omega)A_D - \omega A_R\right] \\ &= \left[A_D^{-1}(A_D + \omega A_L)\right]^{-1} \left[(1-\omega)I - \omega A_D^{-1} A_R\right] \\ &= \left[(I + \omega A_D^{-1} A_L)\right]^{-1} \left[(1-\omega)I - \omega A_D^{-1} A_R\right], \end{aligned}$$

also dem Produkt von

- einer nach Satz 2.8 normierten unteren Dreiecksmatrix $(I + \omega A_D^{-1} A_L)^{-1}$, und
- einer oberen Dreiecksmatrix $(1-\omega)I - \omega A_D^{-1} A_R$ mit Diagonalelementen $(1-\omega)$.

Es gilt also

$$\det(C_\omega) = \det(I + \omega A_D^{-1} A_L)^{-1} \det\left[(1-\omega)I - \omega A_D^{-1} A_R\right] = (1-\omega)^n.$$

Weil die Determinante einer Matrix gleich dem Produkt ihrer Eigenwerte ist, gilt insbesondere $|\det(C_\omega)| \leq (\rho(C_\omega))^n$, also

$$|1-\omega|^n \leq (\rho(C_\omega))^n$$

und damit folgt die Behauptung. QED

Abschließend zeigen wir, dass die Rückrichtung der obigen Aussage zumindest für hermitesche und positiv definite Matrizen richtig ist: Für alle Werte $\omega \in (0, 2)$ des Relaxationsparameters konvergiert das Verfahren.

Satz 5.17 *Sei $A \in \mathbb{K}^{n,n}$ hermitesch und positiv definit. Dann konvergiert das Einzelschritt-Relaxationsverfahren (SOR-Verfahren) für jeden Relaxationsparameter $\omega \in (0, 2)$.*

Beweis: Wir berechnen den Spektralradius der Iterationsmatrix C_ω, und zeigen, dass $\rho(C_\omega) < 1$ gilt. Dazu sei also λ ein Eigenwert von C_ω mit zugehörigem Eigenvektor x. Unser Ziel ist, $|\lambda| < 1$ nachzuweisen. Nach (5.19) ist $C_\omega x = \lambda x$ gleichbedeutend mit

$$\left[(1-\omega)A_D - \omega A_R\right] x = \lambda (A_D + \omega A_L) x. \tag{5.20}$$

Wir nutzen nun folgende beide Aussagen, die sich direkt aus $A = A_L + A_D + A_R$ ergeben:

1. $(2-\omega)A_D - \omega A - \omega(A_R - A_L) = 2(1-\omega)A_D - 2\omega A_R$
2. $(2-\omega)A_D + \omega A - \omega(A_R - A_L) = 2A_D + 2\omega A_L$

Damit folgt aus (5.20), dass

$$[(2-\omega)A_D - \omega A - \omega(A_R - A_L)]\, x = \lambda\, [(2-\omega)A_D + \omega A - \omega(A_R - A_L)]\, x$$

Um diese Gleichung nach λ aufzulösen, bilden wir das Skalarprodukt durch die Multiplikation beider Seiten von links mit x^*. Um abzukürzen, führen wir die Bezeichnungen

$$\begin{aligned} d &:= x^* A_D x \\ a &:= x^* A x \end{aligned}$$

ein, und bemerken, dass $a > 0$ und $d > 0$ gilt, weil A positiv definit ist. Die Multiplikation von links mit x^* ergibt nun

$$(2-\omega)d - \omega a - \omega x^*(A_R - A_L)x = \lambda\, [(2-\omega)d + \omega a - \omega x^*(A_R - A_L)x] \quad (5.21)$$

Zunächst machen wir uns klar, dass

$$\begin{aligned} (ix^*(A_R - A_L)x)^* &= x^*(A_R^* - A_L^*)x\bar{i} \\ &= x^*(A_L - A_R)x(-i) \quad \text{weil } A = A^* \\ &= ix^*(A_R - A_L)x \end{aligned}$$

gilt, und daher $s := ix^*(A_R - A_L)x \in \mathbb{R}$ ist und wir (5.21) weiter umformulieren können zu

$$(2-\omega)d - \omega a + i\ \omega s = \lambda\, [(2-\omega)d + \omega a + i\ ws]\,,$$

in der bis auf λ alle auftretenden Werte $\omega, d, a, s \in \mathbb{R}$ sind. Mit

$$\begin{aligned} \alpha &:= (2-\omega)d - \omega a \in \mathbb{R} \\ \tilde{\alpha} &:= (2-\omega)d + \omega a \in \mathbb{R} \\ \beta &:= \omega s \in \mathbb{R} \end{aligned}$$

erhalten wir endlich

$$\alpha + i\beta = \lambda(\tilde{\alpha} + i\beta).$$

Dann gilt auch für die Beträge, dass

$$|\alpha + i\beta| = |\lambda||\tilde{\alpha} + i\beta|.$$

Wir nutzen noch aus, dass $\tilde{\alpha} > \alpha$ (weil $\omega \in (0,2)$) und erhalten

$$\alpha^2 + \beta^2 = |\lambda|(\tilde{\alpha}^2 + \beta^2) > |\lambda|(\alpha^2 + \beta^2),$$

also $|\lambda| < 1$. QED

Folge: Da das Einzelschrittverfahren ein Spezialfall des SOR-Verfahrens, nämlich mit $\omega = 1$ ist, haben wir mit dem vorliegenden Satz bewiesen, dass das Einzelschrittverfahren ESV für hermitesche und positiv definite Matrizen immer konvergiert.

5.4 Iterative Verfahren für nichtlineare Gleichungssysteme

In diesem Abschnitt betrachten wir nichtlineare Gleichungssysteme

$$F(x) = 0$$

mit einer reellen Funktion $F : \mathbb{R}^n \to \mathbb{R}^n$,

$$F(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_n(x) \end{pmatrix}$$

Wir nehmen zunächst an, dass unser Gleichungssystem bereits in Fixpunktform $G(x) = x$ vorliegt mit einer Funktion

$$G(x) = \begin{pmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_n(x) \end{pmatrix}.$$

Eine Fixpunktgleichung kann man z.B. durch die Funktion G mit

$$G(x) = x + M^{-1}(x)(F(x))$$

erzeugen. Dabei ist M ein linearer Operator, der von x abhängen darf. Wie wichtig es ist, die Funktion G sinnvoll zu wählen, zeigt das folgende Beispiel.

Wir betrachten die Funktion $f(x) = x - \cos x$ im Intervall $[0, 1]$.

- Wähle $g(x) = \cos x$. Dann gilt $f(x) = 0$ genau dann wenn $g(x) = x$. Weiterhin gilt $g : [0, 1] \to [0, 1]$. Jetzt wollen wir zeigen, dass g eine Kontraktion ist. Wir wenden Lemma 5.7 an und erhalten

  $$q = \sup_{0 \le x \le 1} |g'(x)| = \sup_{0 \le x \le 1} \sin x = \sin 1 < 1,$$

  also ist g eine Kontraktion. Nach Satz 5.6 konvergiert also das Verfahren $x^{(k+1)} = \cos x^{(k)}$. Allerdings ist die Konvergenzgeschwindigkeit unbefriedigend.

- Betrachten wir nun

  $$g(x) = x - \frac{x - \cos x}{1 + \sin x}$$

Auch dann gilt $f(x) = 0$ genau dann wenn $g(x) = x$, und man sieht nach kurzer Rechnung, dass $g : [0,1] \to [0,1]$, und dass

$$\begin{aligned} g'(x) &= 1 - \frac{(1+\sin x)^2 - (x - \cos x)\cos x}{(1+\sin x)^2} \\ &= 1 - 1 + \frac{(x-\cos x)\cos x}{(1+\sin x)^2} \\ &= 1 \quad \text{für } x = 0. \end{aligned}$$

Also ist g keine Kontraktion. Das Verfahren der sukzessiven Approximation konvergiert dennoch. Die Konvergenz mit Hilfe dieser Fixpunktgleichung ist sogar sehr schnell!

Um die schnelle Konvergenz zu erklären, schreibt man die Ableitung um zu

$$g'(x) = \frac{f(x)\cos x}{(1+\sin x)^2}.$$

Weil f eine Nullstelle x^* in $[0,1]$ hat, gilt $g'(x^*) = 0$, also gibt es eine Umgebung um x^*, in der g eine Kontraktion ist. Der Kontraktionsfaktor in dieser Umgebung ist nahe bei Null (also sehr klein), und das Konvergenzverhalten daher gut.

Bevor wir diese Beobachtung im Newton-Verfahren ausnutzen, verallgemeinern wir Satz 5.12 auf nichtlineare Funktionen und beweisen damit, dass das Verfahren der sukzessiven Approximation unter ähnlichen Bedingungen wie im Satz 5.12 auch im nichtlinearen Fall konvergiert.

Satz 5.18 *Sei $U \subseteq \mathbb{R}^n$ eine konvexe Menge und $G : U \to U$ eine stetig differenzierbare Abbildung (d.h. jedes der Elemente der Jacobi-Matrix DG ist stetig in U). Weiterhin gelte eine der folgenden Bedingungen:*

Zeilensummenkriterium: $q_\infty = \sup_{x\in U} \max_{i=1,\dots,n} \sum_{j=1}^n \left|\frac{\partial g_i}{\partial x_j}\right| < 1$

Spaltensummenkriterium: $q_1 = \sup_{x\in U} \max_{j=1,\dots,n} \sum_{i=1}^n \left|\frac{\partial g_i}{\partial x_j}\right| < 1$

Quadratsummenkriterium: $q_2 = \sup_{x\in U} \sqrt{\sum_{i,j=1}^n \left|\frac{\partial g_i}{\partial x_j}\right|^2} < 1$

Dann konvergiert das Verfahren der sukzessiven Approximation bezüglich jeder Norm im $\mathbb{R}^n$ für jeden Startwert $x^{(0)} \in \mathbb{R}^n$, und zwar gegen die eindeutig bestimmte Lösung x^* des nichtlinearen Gleichungssystems $G(x^*) = x^*$. Ist $q_p < 1$ für $p \in \{1, 2, \infty\}$ so gelten für dieses p außerdem die folgenden Schranken.

- *A priori Fehlerschranke:* $\|x^{(k)} - x^*\|_p \le \frac{q_p^k}{1-q_p}\|x^{(1)} - x^{(0)}\|_p$

- *A posteriori-Fehlerschranke:* $\|x^{(k)} - x^*\|_p \leq \frac{q_p}{1-q_p}\|x^{(k)} - x^{(k-1)}\|_p$

Beweis: Nach Satz 3.20 und Lemma 3.27 gilt, dass

$$\begin{aligned} \sup_{x\in U} \|DG(x)\|_p &= q_p \quad \text{für } p \in \{\infty, 1\} \\ \sup_{x\in U} \|DG(x)\|_2 &\leq q_2 \end{aligned}$$

Daher ist nach Lemma 5.7 die Abbildung $G : U \to U$ kontrahierend, falls $q_p < 1$ für ein $p \in \{\infty, 1, 2\}$. Satz 5.6 ergibt die Behauptung. QED

Unter den Voraussetzungen des letzten Satzes konvergiert also das Verfahren der sukzessiven Approximation auch im nichtlinearen Fall. Das Verfahren

$$x_i^{(k+1)} = g_i(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}) \; i = 1, \ldots, n, \; k = 0, 1, 2, \ldots$$

nennt man auch **nichtlineares Gesamtschrittverfahren**, während man das Verfahren

$$\begin{aligned} x_1^{(k+1)} &= g_1(x_1^{(k)}, \ldots, x_n^{(k)}) \\ x_i^{(k+1)} &= g_i(x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}, x_i^{(k)}, \ldots, x_n^{(k)}) \; i = 2, \ldots, n, \end{aligned}$$

als **nichtlineares Einzelschrittverfahren** bezeichnet. Die in Abschnitt 5.3 besprochenen Verfahren GSV und ESV sind Spezialfälle dieser Verfahren.

Das Newton-Verfahren für skalare Funktionen

Wir kommen nun wieder zurück zu dem originalen nichtlinearen Gleichungssystem

$$F(x) = 0$$

und entwickeln mit dem nun zu besprechenden Newton-Verfahren eine Fixpunktform, die — wenn sie konvergiert — zu einem schnelleren Konvergenzverhalten führt. Wir beginnen unsere Überlegung für eine reelle Funktion $f : \mathbb{R} \to \mathbb{R}$.

Gesucht ist eine Nullstelle x^* der Funktion f. Haben wir schon eine Schätzung der Nullstelle $x^{(0)}$ und ist f stetig differenzierbar, so besteht die Idee des Newton-Verfahrens darin, f durch seine Tangente durch den Punkt $x^{(0)}$

$$f \approx f(x^{(0)}) + f'(x^{(0)})(x - x^{(0)})$$

(also der Taylorreihe bis zum linearen Glied) zu ersetzen. Man sucht also die Nullstelle der Näherung anstatt der Nullstelle von f. Eine Nullstelle der Näherung $f(x^{(0)}) + f'(x^{(0)})(x - x^{(0)})$ existiert, falls $f'(x^{(0)}) \neq 0$ ist und ist in diesem Fall gegeben durch

$$x = x^{(0)} - \frac{f(x^{(0)})}{f'(x^{(0)})}$$

Wiederholt man das Vorgehen mit $x^{(1)} := x$ so erhält man das Newton-Verfahren. Erfüllt die Ableitung $f'(x) \neq 0$ so erhält man die Fixpunktgleichung

$$g(x) = x - \frac{f(x)}{f'(x)}.$$

Kommen wir kurz zu dem Beispiel $f(x) = x - \cos x$ von Seite 115 zurück: Hier wurde mit

$$g(x) = x - \frac{1}{1+\sin x}(x - \cos x) = x - \frac{f(x)}{f'(x)}$$

im zweiten Versuch genau die Fixpunktform des Newton-Verfahrens verwendet. Leider ist die Funktion g im allgemeinen keine Kontraktion auf dem gesamten zu betrachtenden Intervall, so dass Satz 5.18 nicht anwendbar ist. Dennoch gilt die folgende lokale Konvergenzaussage.

Satz 5.19 *Sei x^* eine einfache Nullstelle der $f : \mathbb{R} \to \mathbb{R}$. Sei weiterhin f in einer Umgebung von x^* zwei mal stetig differenzierbar. Dann konvergiert das Newton-Verfahren für jeden Startwert $x^{(0)}$, der hinreichend dicht bei x^* liegt.*

Beweis: Weil x^* einfache Nullstelle ist, gilt $f'(x^*) \neq 0$ und entsprechend gibt es eine Umgebung $U_1 := U(x^*)$ so dass $f'(x) \neq 0$ für alle $x \in U_1$. Die Verfahrensvorschrift $g(x) = x - \frac{f(x)}{f'(x)}$ ist damit für alle $x \in U_1$ definiert. Die Ableitung von g ist

$$g'(x) = 1 - \frac{[f'(x)]^2 - f(x)f''(x)}{[f'(x)]^2} = \frac{f''(x)}{[f'(x)]^2} f(x),$$

also $g'(x^*) = 0$. Wegen der Stetigkeit von g' gibt es eine Umgebung U_2, so dass für alle $x \in U = U_1 \cap U_2$ gilt $|g'(x)| \leq q < 1$. Nach dem Mittelwertsatz der Differential- und Integralrechnung erhalten wir daraus

$$\begin{aligned} \left| \frac{g(x) - g(y)}{x - y} \right| &= |g'(\xi)| \text{ mit einem } \xi \in [x, y] \subseteq U \\ &\leq q < 1 \end{aligned}$$

das heißt, $g : U \to U$ ist eine Kontraktion. Satz 5.6 liefert die Behauptung.

QED

Das Newton-Verfahren für mehrdimensionale Funktionen

Wir formulieren zunächst das Newton-Verfahren für den mehrdimensionalen Fall.

Definition 5.20 *Sei $U \subseteq \mathbb{R}^n$ offen und $F : U \to \mathbb{R}^n$ eine stetig differenzierbare Funktion mit einer für alle $x \in U$ regulären Jacobimatrix $DF(x)$. Dann heißt das Verfahren*

$$x^{(k+1)} = x^{(k)} - [DF(x^{(k)})]^{-1} F(x^{(k)}) \quad k = 0, 1, 2, \ldots$$

mit Startwert $x^{(0)} \in U$ **Newton-Verfahren**.

Die Motivation für das Newton-Verfahren ist die gleiche wie für skalare Funktionen: Anstatt die Nullstelle $F(x) = 0$ zu suchen, ersetzt man

$$F \approx F(x^{(0)}) + (DF(x^{(0)}))(x - x^{(0)}).$$

Existiert $[DF(x^{(0)})]^{-1}$, so kann man diese Gleichung nach x auflösen und erhält

$$x = x^{(0)} - [DF(x^{(0)})]^{-1} F(x^{(0)})$$

als nächste Iterierte. Das entspricht der Fixpunktgleichung (5.2)

$$G(x) = x + M \cdot F(x)$$

mit der regulären Matrix $M = -[DF(x)]^{-1}$.

Um das Newton-Verfahren numerisch zu realisieren wird zur Bestimmung von

$$x^{(k+1)} = x^{(k)} - [DF(x^{(k)})]^{-1} F(x^{(k)}) \quad k = 0, 1, 2, \ldots$$

in jedem Schritt das lineare Gleichungssystem

$$DF(x^{(k)})(x^{(k+1)} - x^{(k)}) = -F(x^{(k)})$$

gelöst. Das geschieht durch das Lösen des Systems

$$DF(x^{(k)}) w^{(k)} = -F(x^{(k)})$$

und anschließendes Berechnen von

$$x^{(k+1)} = x^{(k)} + w^{(k)}.$$

Bevor wir auf die Konvergenzeigenschaften näher eingehen, formulieren wir das Verfahren.

Algorithmus 11: Newton-Verfahren

Input: `Offene Menge` $U \subseteq \mathbb{R}^n$, `Differenzierbare Abbildung` $F : U \to \mathbb{R}^n$ `mit Jacobi-Matrix` $DF : U \to \mathbb{R}^{n,n}$. `Startwert` $x^{(0)} \in U$, `Toleranzwert` $\varepsilon > 0$.

Schritt 1: $k := 0$

Schritt 2: `Repeat`

Schritt 2.1: `Finde` $w^{(k)}$ `als Lösung des Gleichungssystems`

$$DF(x^{(k)})w^{(k)} = -F(x^{(k)}).$$

Schritt 2.2: $x^{(k+1)} := x^{(k)} + w^{(k)}$

Schritt 2.3: $q_k := \frac{\|w^{(k)}\|}{\|w^{(k-1)}\|}$

Schritt 2.4: `If` $q_k \geq 1$ `oder` $x^{(k+1)} \notin U$ `STOP: Das Verfahren scheint nicht zu konvergieren.`

Schritt 2.5: $k := k + 1$

Until $\frac{q_k}{1-q_k}\|w^{(k)}\| \leq \varepsilon$

Ergebnis: `Approximierte Nullstelle` $x^{(k)}$ `von` F.

Leider ist die Berechnung von $DF(x)$ für große n aufwändig, so dass man die Jacobi-Matrix in der Praxis nicht in jedem Schritt neu berechnet, sondern häufig die folgenden Varianten verwendet:

- Frozen Newton: Es wird nur einmal die Jacobi-Matrix berechnet, für die dann mittels LU-Zerlegung alle in den Iterationen auftretende Gleichungssysteme effizient lösbar sind.
- Man kann die Jacobi-Matrix auch alle K Schritte neu berechnen.
- Quasi-Newton: Die Jacobi-Matrix wird in jedem Schritt (approximativ) angepasst.

Um den Konvergenzbereich des Verfahrens zu vergrößern verwendet man auch das so genannte *gedämpfte* Newton-Verfahren, in dem man die Iterationsvorschrift

$$x^{(k+1)} = x^{(k)} + \lambda_k w^{(k)}, \lambda_k \in [0, 1]$$

verwendet. In der Fixpunktgleichung (5.2) wurde also $M = -\lambda_k[DF(x)]^{-1}$ gewählt.

Das Konvergenzverhalten des mehrdimensionalen Newton-Verfahrens lässt sich nicht ganz so einfach analysieren wie im eindimensionalen Fall. Daher ist die Verallgemeinerung von Satz 5.19 etwas schwieriger zu zeigen. Wir beweisen im folgenden Satz aber mehr, nämlich dass das Newton-Verfahren sogar **quadratisch konvergiert**.

Definition 5.21 *Sei $x^{(k)} \to x^*$ eine Folge im $\mathbb{K}^n$ mit $x^{(k)} \neq x^*$ für alle k. Wenn es eine Konstante q gibt, so dass*

$$\|x^{(k+1)} - x^*\| \leq q\|x^{(k)} - x^*\|^p \quad \text{für } \|x^{(k)} - x^*\| \to 0$$

so liegt eine Konvergenz der **Konvergenzordnung** *p gegen x^* vor. Den Fall $p = 1$ bezeichnet man für $q < 1$ als* **lineare Konvergenz**, *den Fall $p = 2$ für $q > 0$ als* **quadratische Konvergenz**.

Man beachte, dass sich die Anzahl der korrekt gefundenen Stellen einer Zahl bei quadratischer Konvergenz in jedem Schritt etwa verdoppelt. Wir formulieren jetzt den Satz zur Konvergenz des Newton-Verfahrens.

Satz 5.22 *Sei $U \subseteq \mathbb{R}^n$ offen und konvex und sei $F : U \to \mathbb{R}^n$ stetig differenzierbar. Sei $x^{(0)} \in U$. Weiter erfülle F und $x^{(0)}$ die folgenden vier Bedingungen für eine (beliebige) Norm $\|\cdot\|$ auf dem $\mathbb{R}^n$:*

[B1] : Es existiert eine Nullstelle $x^* \in U$ der Funktion F.

[B2] : $DF(x)$ ist regulär für alle $x \in U$.

[B3] : Es gibt $\omega > 0$ so dass

- *(a) $\|[DF(x)]^{-1}(DF(y) - DF(x))\| \leq \omega\|x - y\|$ für alle $x, y \in U$ und*
- *(b) für $\rho := \|x^* - x^{(0)}\|$ gilt $\frac{\omega}{2}\rho < 1$.*

[B4] : Die Kugel $B_\rho(x^*) := \{x \in \mathbb{R}^n : \|x - x^*\| < \rho\}$ mit Radius ρ um die Nullstelle x^* ist in U enthalten.

Für die im Newton-Verfahren definierte Folge

$$x^{(k+1)} := x^{(k)} - [DF(x^{(k)})]^{-1}F(x^{(k)})$$

gilt dann:

1. *$x^{(k)} \in B_\rho(x^*)$ für alle $k = 1, 2, \ldots$.*
2. *$x^{(k)}$ konvergiert gegen x^*.*
3. *Für $k = 0, 1, 2, \ldots$ gilt die folgende a priori Fehlerschranke*

$$\|x^{(k)} - x^*\| \leq \rho \left(\frac{\omega\rho}{2}\right)^{2^k - 1} \tag{5.22}$$

4. *Für $k = 0, 1, 2, \ldots$ gilt die folgende a posteriori Fehlerschranke:*

$$\|x^{(k+1)} - x^*\| \leq \frac{\omega}{2}\|x^{(k)} - x^*\|^2 \tag{5.23}$$

Beweis: Wir zeigen zunächst, dass aus $x^{(k)} \in U$ die a-posteriori Fehlerschranke (5.23) für k folgt. Danach beweisen wir, dass das Verfahren durchführbar ist sowie die Aussagen des Satzes per Induktion.

Teil 1: Wir benötigen zunächst eine Funktion $g : [0,1] \to \mathbb{R}^n$, die wir als

$$g(t) = F(x^{(k)} + t(x^* - x^{(k)}))$$

definieren. Durch die multivariate Kettenregel erhalten wir

$$g'(t) = DF(x^{(k)} + t(x^* - x^{(k)}))(x^* - x^{(k)}),$$

woraus nach dem Hauptsatz der Differential- und Integralrechnung wie in Lemma 5.7 folgt, dass

$$F(x^*) - F(x^{(k)}) = g(1) - g(0) = \int_0^1 g'(t)dt = \int_0^1 DF(x^{(k)} + t(x^* - x^{(k)}))(x^* - x^{(k)})dt.$$

Jetzt setzen wir wie oben beschrieben voraus, dass $x^{(k)} \in U$. Dann ist $DF(x^{(k)})$ regulär und daher $x^{(k+1)}$ definiert. Wir erhalten

$$\begin{aligned}
A &:= x^{(k+1)} - x^* \\
&= x^{(k)} - [DF(x^{(k)})]^{-1} F(x^{(k)}) - x^* \\
&= x^{(k)} - x^* - [DF(x^{(k)})]^{-1}(F(x^{(k)}) - F(x^*)) \\
&= [DF(x^{(k)})]^{-1} \left(F(x^*) - F(x^{(k)}) - DF(x^{(k)})(x^* - x^{(k)})\right) \\
&= [DF(x^{(k)})]^{-1} \left(\int_0^1 DF(x^{(k)} + t(x^* - x^{(k)}))(x^* - x^{(k)})dt - DF(x^{(k)})(x^* - x^{(k)})\right) \\
&= \int_0^1 [DF(x^{(k)})]^{-1} \left\{DF(x^{(k)} + t(x^* - x^{(k)})) - DF(x^{(k)})\right\} (x^* - x^{(k)})dt
\end{aligned}$$

Gehen wir zur Norm davon über, so erhalten wir

$$\begin{aligned}
\|A\| &= \|x^{(k+1)} - x^*\| \\
&\leq \int_0^1 \|[DF(x^{(k)})]^{-1} \left\{DF(x^{(k)} + t(x^* - x^{(k)})) - DF(x^{(k)})\right\}\| \, \|x^* - x^{(k)}\|dt
\end{aligned}$$

Nach Voraussetzung [B3] (a) gilt mit $x = x^{(k)}$ und $y = x^{(k)} + t(x^* - x^{(k)})$ dass

$$\begin{aligned}
& \|[DF(x^{(k)})]^{-1} \left(DF(x^{(k)} + t(x^* - x^{(k)})) - DF(x^{(k)})\right)\| \\
\leq\ & \omega \|x^{(k)} - (x^{(k)} + t(x^* - x^{(k)}))\| \\
=\ & \omega t \|x^{(k)} - x^*\|.
\end{aligned}$$

Setzen wir dieses Ergebnis in die obige Ungleichung ein, so ergibt sich

$$\begin{aligned}
\|A\| &= \|x^{(k+1)} - x^*\| \\
&\leq \int_0^1 \|[DF(x^{(k)})]^{-1} \left(DF(x^{(k)} + t(x^* - x^{(k)})) - DF(x^{(k)})\right)\| \, \|x^* - x^{(k)}\| dt \\
&\leq \int_0^1 \omega t \|x^{(k)} - x^*\| \, \|x^{(k)} - x^*\| dt = \int_0^1 \omega t \|x^{(k)} - x^*\|^2 dt \\
&= \frac{\omega}{2} \|x^{(k)} - x^*\|^2.
\end{aligned}$$

Damit ist also gezeigt, dass (5.23) gilt, falls $x^{(k)} \in U$.

Teil 2: Mit Hilfe der Aussage aus Teil 1 können wir nun per Induktion zeigen, dass $x^{(k)} \in U$ sowie die Aussagen 1,3 und 4 des Satzes.

Induktionsanfang Für $k = 0$ gilt:

- $x^{(0)} \in U$ nach Voraussetzung
- (5.22): Für $k = 0$ erhält man $\left(\frac{\omega\rho}{2}\right)^{2^k - 1} = 1$. Daher gilt $\|x^{(0)} - x^*\| = \rho$ nach der Definition von ρ.
- (5.23): Weil $x^{(0)} \in U$ können wir Teil 1 des Beweises verwenden. Wir erhalten:
$$\|x^{(1)} - x^*\| \leq \frac{\omega}{2} \|x^{(0)} - x^*\|^2.$$

Induktionsschritt: $k \to k+1$. Sei also der Satz richtig für k. Dann ist $x^{(k)} \in U$ nach der Induktionsannahme. Wir rechnen

$$\begin{aligned}
\|x^{(k+1)} - x^*\| &\leq \frac{\omega}{2} \|x^{(k)} - x^*\|^2 \text{ wegen Teil 1} \\
&\leq \frac{\omega}{2} \left(\frac{\omega\rho}{2}\right)^{2(2^k - 1)} \rho^2 \\
&\qquad \text{denn es gilt (5.22) nach Induktionsannahme} \\
&= \left(\frac{\omega\rho}{2}\right)^{2^{k+1} - 1} \rho < \rho \text{ wegen [B3], Teil (b).}
\end{aligned}$$

Aus der letzten Zeile folgt

- die a-priori Fehlerschranke (5.22) für $k + 1$,
- sowie $x^{(k+1)} \in B_\rho(x^*)$.
- Wegen [B4] ist also $x^{(k+1)} \in U$ und entsprechend ist die Folge wohldefiniert.

Nach Teil 1 erhalten wir auch

- die a posteriori-Schranke (5.23) für $k + 1$.

Wegen (5.22) und [B3], Teil (b) erhält man außerdem direkt die Konvergenz gegen x^*, und damit auch die 2. Aussage des Satzes.

QED

Zum Abschluss geben wir noch einen Satz an, der zeigt, dass bei zweimal stetigen Funktionen im wesentlichen die Voraussetzung [B1] und die Regularität der Jacobi-Matrix an der gesuchten Nullstelle nötig sind, um eine lokale quadratische Konvergenz in der Nähe der Nullstelle zu erreichen.

Satz 5.23 *Sei $U \subseteq \mathbb{R}^n$ offen und $F : U \to \mathbb{R}^n$ eine zweimal stetig differenzierbare Funktion. Sei $x^* \in U$ mit $F(x^*) = 0$ und $\det(DF(x^*)) \neq 0$. Dann existiert ein $\rho > 0$ so dass das Newton-Verfahren für alle Startwerte $x^{(0)} \in U := B_\rho(x^*)$ quadratisch konvergiert.*

Beweis: Wir untersuchen die Voraussetzungen von Satz 5.22.

- Zunächst gilt [B1] nach Voraussetzung.
- Weil $det(DF(x^*)) \neq 0$ und $det(DF(x))$ stetig ist, gibt es $\rho^1 > 0$ so dass $det(DF(x)) \neq 0$ für alle $x \in B_{\rho^1}(x^*)$. Dann gilt [B2] für $U \subseteq B_{\rho^1}(x^*)$.
- Um [B3] zu zeigen, betrachten wir die Funktion $h(x) = [DF(x)]^{-1}$. Als Matrixinversion ist sie stetig, daher gibt es für jedes $\varepsilon > 0$ ein $\rho > 0$ so dass

  $$\|h(x)\| - \|h(x^*)\| \leq \|h(x) - h(x^*)\| \leq \varepsilon \text{ für alle } x \in B_\rho(x^*).$$

  Mit $\varepsilon := \|[DF(x^*)]^{-1}\|$ existiert also ein $\rho^2 \leq \rho^1$ so dass

  $$\|[DF(x)]^{-1}\| - \|[DF(x^*)]^{-1}\| \leq \|[DF(x^*)]^{-1}\| \text{ für alle } x \in B_{\rho^2}(x^*),$$

  oder, äquivalent,

  $$\|[DF(x)]^{-1}\| \leq 2\|[DF(x^*)]^{-1}\| \text{ für alle } x \in B_{\rho^2}(x^*).$$

  Jetzt nutzten wir, dass DF nach Voraussetzung für $x \in U$ differenzierbar ist. Also gibt es eine Lipschitzkonstante $L > 0$ so dass

  $$\|[DF(x)] - [DF(y)]\| \leq L\|x - y\| \text{ für alle } x, y \in B_{\rho^2}.$$

  Wählt man $\omega := 2L\|[DF(x^*)]^{-1}\|$ so gilt

  $$\begin{aligned} \|[DF(x)]^{-1}([DF(y)] - [DF(x)])\| &\leq \underbrace{\|[DF(x)]^{-1}\|}\ \|\underbrace{([DF(y)] - [DF(x)])}\| \\ &\leq 2\|[DF(x^*)]^{-1}\|\ \ L\|x - y\| \\ &= \omega\|x - y\|, \end{aligned}$$

  also gilt [B3], Teil (a).
- Mit $\rho := \min\{\rho^2, \frac{2}{\omega}\}$ folgt wegen $\frac{\omega}{2}\rho < 1$ Teil (b) von [B3] und mit $U := B_\rho(x^*)$ auch [B4].

Damit sind die Voraussetzungen von Satz 5.22 erfüllt und es folgt die quadratische Konvergenz. QED

Chapter 6

Eigenwertprobleme

6.1 Grundlagen und Eigenschaften

Der Begriff des *Eigenwertes* einer Matrix ist in dieser Vorlesung schon ein paar Mal vorgekommen. Das betrifft die Charakterisierung einer positiv definiten Matrix, (Lemma 2.18) und den Spektralradius (Definition 3.21), der insbesondere zum Beweis der Konvergenz von Iterationsverfahren wichtig ist (siehe Satz 5.11). Außerdem waren Eigenwerte bei der Singulärwertzerlegung wichtig (siehe Abschnitt 4.14) wichtig. In diesem Kapitel beschäftigen wir uns nun damit, wie man die Eigenwerte einer quadratischen Matrix $A \in \mathbb{C}^{n,n}$ berechnet. (Die Eigenwerte einer reellen Matrix berechnet man genauso —- allerdings sind die Eigenwerte i. all. nicht alle reell.) Wir wiederholen zunächst die folgenden Bezeichnungen.

Definition 6.1 *Eine Zahl $\lambda \in \mathbb{K}$ heißt* **Eigenwert** *einer Matrix $A \in \mathbb{K}^{n,n}$ wenn es ein $v \in \mathbb{K}^n \setminus \{0\}$ gibt, so dass $Av = \lambda v$. v heißt dann* **Eigenvektor** *zum Eigenwert λ. Weiter nennt man*

$$N(A - \lambda I) := \{v \in \mathbb{K}^n : Av = \lambda v\}$$

den **Eigenraum** *zum Eigenwert λ und bezeichnet seine Dimension als die* **geometrische Vielfachheit** *des Eigenwertes.*

Weil das Gleichungssystem $Av = \lambda v$ oder äquivalent

$$(A - \lambda I)v = 0$$

immer die Lösung $v = 0$ hat (die zu keinem Eigenwert führt), muss man λ also so bestimmen, dass $A - \lambda I$ nicht regulär ist. Das ist äquivalent dazu, dass $det(A - \lambda I) = 0$, was auf das **charakteristische Polynoms** von A,

$$\chi_A(\lambda) := det(A - \lambda I),$$

führt. Die Eigenwerte einer Matrix $A \in \mathbb{K}^{n,n}$ sind dann genau die Nullstellen des charakteristischen Polynoms.

Definition 6.2 *Ist λ eine Nullstelle von $\chi_A(\lambda)$so bezeichnet man ihre Vielfachheit μ_i als die* **(algebraische) Vielfachheit** *des Eigenwertes λ.*

Ein einfaches Beispiel ist die Bestimmung der Eigenwerte einer oberen (oder einer unteren) Dreiecksmatrix $R = (r_{ij})$. Diese sind genau die Elemente auf ihrer Hauptdiagonalen, da

$$X_R(\lambda) = det(R - \lambda I) = (r_{11} - \lambda) \cdot \ldots \cdot (r_{nn} - \lambda)$$

als Nullstellen genau die Diagonalelemente von R besitzt.

Wie in vielen Lehrbüchern unterscheiden wir notationsmäßig wie folgt:

- Die *Eigenwerte der Matrix A* listet die Eigenwerte gemäß ihrer algebraischen Vielfachheiten mehrfach auf. Die Matrix $R = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 4 \\ 0 & 0 & 5 \end{pmatrix}$ besitzt also die Eigenwerte $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 5$.
- Schreiben wir andererseits *die verschiedenen Eigenwerte* einer Matrix A, so listen wir jeden Eigenwert (unabhängig von seiner algebraischen Vielfachheit) nur einmal auf.

Das charakteristische Polynom hat als Polynom vom Grad n über $\mathbb{C}$ nach dem Fundametalsatz der Algebra mindestens eine und höchstens n komplexe Nullstellen. Entsprechend erhält man den ersten Teil des folgenden Satzes. Der zweite Teil kann durch Induktion gezeigt werden.

Satz 6.3

- *Über $\mathbb{K} = \mathbb{C}$ besitzt jede $n \times n$ Matrix mindestens einen und höchstens n verschiedene Eigenwerte. Die Summe der algebraischen Vielfachheiten aller Eigenwerte von A ergibt n.*
- *Die Eigenvektoren zu verschiedenen Eigenwerten von A sind linear unabhängig.*

Wir führen außerdem noch das folgende Lemma an.

Lemma 6.4 *Sei $\lambda \in \mathbb{C}$ ein Eigenwert von $A \in \mathbb{C}^{n,n}$. Dann ist das konjugiert komplexe $\overline{\lambda}$ von λ ein Eigenwert von A^*.*

Beweis: Wir erinnern daran, dass $det(A^*) = \overline{det(A)}$ gilt. Sei nun λ ein Eigenwert von A, d.h. $det(\lambda I - A) = 0$. Dann gilt

$$det(\overline{\lambda} I - A^*) = det((\lambda I - A)^*) = \overline{det(\lambda I - A)} = 0,$$

also ist $\overline{\lambda}$ Eigenwert von A^*. QED

Durch die Ermittlung der Nullstellen des charakteristischen Polynoms kann man theoretisch also die Eigenwerte einer Matrix bestimmen. Das ist aber nur für kleine Werte von n sinnvoll und außerdem schlecht konditioniert. Man nutzt daher das folgende Lemma, das besagt, dass Ähnlichkeitstransformationen die Eigenwerte einer Matrix nicht verändern:

Lemma 6.5 *Sei $A \in \mathbb{C}^{n,n}$ und $Q \in \mathbb{C}^{n,n}$ regulär. Dann haben A und $Q^{-1}AQ$ die gleichen Eigenwerte $\lambda_1, \ldots, \lambda_n$.*

Beweis: Es gilt

$$det(A-\lambda I) = det(Q^{-1})det(A-\lambda I)det(Q) = det(Q^{-1}(A-\lambda I)Q) = det(Q^{-1}AQ-\lambda I).$$

Damit erhält man: λ ist ein Eigenwert von A genau dann wenn $det(A - \lambda I) = 0$ genau dann wenn $det(Q^{-1}AQ - \lambda I) = 0$ genau dann wenn λ ein Eigenwert von $Q^{-1}AQ$ ist. QED

Dieses Lemma liefert die folgende Idee zur algorithmischen Behandlung von Eigenwertproblemen: Man transformiert die Matrix A zu einer Matrix, für die man die Eigenwerte leicht berechnen kann.

Im weiteren verwenden wir das Lemma von Schur (siehe Satz 5.9). Es besagt, dass es zu jeder Matrix $A \in \mathbb{K}^{n,n}$ eine unitäre Matrix Q gibt, so dass

$$Q^*AQ = R$$

eine obere Dreiecksmatrix ist.

Übung: Sei $A \in \mathbb{C}^{n,n}$ eine Matrix mit den n Eigenwerten $\lambda_1, \ldots, \lambda_n$.

- *Ist A regulär, so gilt: λ ist Eigenwert von A genau dann wenn $\frac{1}{\lambda}$ Eigenwert von A^{-1} ist.*
- *Die Summe der n Eigenwerte von A ergibt die Spur von A.*
- *Das Produkt der n Eigenwerte von A ergibt $det(A)$. (Diese Aussage haben wir im Beweis von Satz 5.16 verwendet.)*

6.2 Der Spezialfall hermitescher Matrizen

Wir leiten nun zuerst einige Resultate her für den Fall von hermiteschen Matrizen, also Matrizen A, für die $A^* = A$ gilt. Hierfür wissen wir bereits aus Satz 3.23 die folgende Aussage:

Sei $A \in \mathbb{K}^{n,n}$ eine hermitesche Matrix. Dann gibt es eine unitäre Matrix $Q \in \mathbb{C}^{n,n}$ so dass

$$Q^*AQ = D = diag(d_1, \ldots, d_n) \in \mathbb{R}^{n,n}$$

eine Diagonalmatrix ist. Dabei sind $d_1, \ldots, d_n$ die Eigenwerte der Matrix A, und die Spalten von Q bilden eine Orthonormalbasis, die aus den zugehörigen Eigenvektoren besteht. Das heißt, es gilt

$$AQ_j = d_jQ_j \text{ für } j = 1, \ldots, n.$$

Weil $(Q^*AQ)^* = Q^*AQ$ folgt $D^* = D$, entsprechend sind die Eigenwerte von A in diesem Fall alle reell.

Übung: Folgern Sie Satz 3.23 aus dem Lemma von Schur!

Übung: Verwenden Sie Satz 3.23, um die folgenden beiden Aussagen über Eigenwerte zu zeigen:

- *Eine symmetrische Matrix A ist positiv definit genau dann wenn alle ihre Eigenwerte echt größer als Null sind. (Diese Aussage haben wir im Beweis von Lemma 2.18 verwendet.)*
- *A^*A hat nur reelle Eigenwerte, genauer: Ist $Rang(A^*A) = r$, dann sind genau r der Eigenwerte von A^*A positiv und die restlichen Eigenwerte sind Null. (Diese Aussage haben wir im Beweis von Satz 4.14 verwendet.)*

Die Eigenwerte hermitescher Matrizen in einem Prä-Hilbertraum (mit dem Skalarprodukt $(\cdot, \cdot)$ so dass $\|x\| = \sqrt{(x, x)}$) wollen wir im folgenden charakterisieren. Wir benötigen zunächst die folgende Eigenschaft von Orthonormalsystemen.

Lemma 6.6 *Sei $\{v_1, \ldots, v_n\}$ eine Orthonormalbasis eines Prä-Hilbert-Raumes mit Skalarprodukt $\|x\| = \sqrt{(x, x)}$. Sei $v = \sum_{i=1}^n \alpha_i v_i$. Dann gilt:*

1. $\alpha_i = (v, v_i)$, *i=1,. . . ,n*
2. $\|v\|^2 = \sum_{i=1}^n \alpha_i^2$.

Beweis: Beide Eigenschaften lassen sich unter Ausnutzung von $(v_i, v_j) = 0$ für $i \neq j$ und $(v_i, v_i) = 1$ wie folgt nachrechnen:

ad 1: $(v, v_i) = (\sum_{j=1}^n \alpha_j v_j, v_i) = \sum_{j=1}^n \alpha_j (v_j, v_i) = \alpha_i$

ad 2: $\|v\|^2 = (v, v) = \sum_{i,j \in \{1,\ldots,n\}} \alpha_i \overline{\alpha}_j (v_i, v_j) = \sum_{i=1}^n |\alpha_i|^2$

QED

Für die beiden folgenden Sätze betrachten wir $\max_{x \in M}(Ax, x)$ für kompakte Mengen M. Dieses Maximum existiert weil $h(x) = (Ax, x)$ eine stetige Abbildung ist.

Satz 6.7 (Satz von Rayleigh) *Sei A eine hermitesche Matrix mit den n Eigenwerten $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Seien v_i die aus Satz 3.23 zugehörigen Eigenvektoren zu λ_i, $i = 1, \ldots, n$, die eine Orthonormalbasis des $\mathbb{C}^n$ bilden. Wir definieren*

$$\begin{aligned} V_1 &:= \mathbb{C}^n \\ V_k &:= \{v \in \mathbb{C}^n : (v, v_i) = 0, i = 1, \ldots, k-1\}, \ k = 2, \ldots, n. \end{aligned}$$

Dann gilt:

$$\lambda_k = \max_{v \in V_k : \|v\| = 1} (Av, v), \ k = 1, \ldots, n.$$

Beweis: Zunächst sieht man direkt, dass

$$\lambda_k = \lambda_k (v_k, v_k) = (\lambda_k v_k, v_k) = (Av_k, v_k) \leq \max_{v \in V_k : \|v\| = 1} (Av, v),$$

da $v_k \in V_k$ und $(v_k, v_k) = \|v_k\|^2 = 1$.

Um $\lambda_k \geq \max_{v \in V_k : \|v\| = 1} (Av, v)$ zu zeigen, betrachten wir ein beliebiges $v \in V_k$ mit $\|v\| = 1$. Dann gibt es $\alpha_1, \ldots, \alpha_n$ so dass $v = \sum_{i=1}^n \alpha_i v_i$. Damit ergibt sich zunächst

$$Av = A(\sum_{i=k}^n \alpha_i v_i) = \sum_{i=1}^n \alpha_i A(v_i) = \sum_{i=1}^n \alpha_i \lambda_i v_i. \tag{6.1}$$

Nach Lemma 6.6 gilt:

1. $\alpha_i = (v, v_i)$, woraus wir für $v \in V_k$ insbesondere $\alpha_i = 0$ für $i = 1, \ldots, k-1$ folgern, sowie
2. $1 = \|v\|^2 = \sum_{i=1}^n |\alpha_i|^2 = \sum_{i=k}^n |(v, v_i)|^2$.

Damit erhalten wir

$$\begin{aligned} (Av, v) &= (\sum_{i=1}^n \alpha_i \lambda_i v_i, v) = \sum_{i=1}^n \alpha_i \lambda_i (v_i, v) \\ &= \sum_{i=k}^n (v, v_i) \lambda_i (v_i, v) = \sum_{i=k}^n (v, v_i) \lambda_i \overline{(v, v_i)} \\ &= \sum_{i=k}^n \lambda_i |(v, v_i)|^2 \leq \lambda_k \sum_{i=k}^n |(v, v_i)|^2 = \lambda_k \end{aligned}$$

QED

Diese Aussage erlaubt (weil man i. allg. die Eigenvektoren nicht kennt) nur eine Abschätzung für den größten Eigenwert. Für diesen erhält man untere Schranken durch

$$\rho(A) \geq \lambda_1 = \max_{v \in \mathbb{C}^n : \|v\| = 1} (Av, v) \geq (Au, u) \text{ für alle } u \text{ mit } \|u\| = 1.$$

Ohne Kenntnis der Eigenvektoren kann man aber den folgenden Satz verwenden.

Satz 6.8 (Minimum-Maximum Prinzip von Courant) *Sei A eine hermitesche Matrix mit den n Eigenwerten $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Sei weiter*

$$M_k = \{U : U \text{ ist Unterraum von } \mathbb{C}^n \text{ mit } dim(U) = n + 1 - k\}.$$

Dann gilt:

$$\lambda_k = \min_{U \in M_k} \max_{v \in U : \|v\| = 1} (Av, v), \ k = 1, \ldots, n.$$

Beweis: Aus dem vorhergehenden Satz 6.7 wissen wir, dass für alle $k = 1, \ldots, n$

$$\lambda_k = \max_{v \in V_k : \|v\| = 1} (Av, v) \geq \min_{U \in M_k} \max_{v \in U : \|v\| = 1} (Av, v).$$

Es muss also noch $\lambda_k \leq \min_{U \in M_k} \max_{v \in U : \|v\| = 1} (Av, v)$ gezeigt werden. Dazu konstruieren wir im folgenden für jeden Unterraum $U \in M_k$ einen Vektor $y \in U$ mit $\|y\| = 1$ und $\lambda_k \leq (Ay, y)$, woraus wir zunächst

$$\lambda_k \leq (Ay, y) \leq \max_{v \in U, \|v\| = 1} (Av, v)$$

für den Unterraum U folgern, woraus wir ableiten, dass

$$\lambda_k \leq \min_{U \in M_k} \max_{v \in U, \|v\| = 1} (Av, v).$$

Zur Konstruktion des gesuchten Vektors $y \in U \in M_k$ gehen wir wie folgt vor: Zunächst sei $\{u_1, \ldots, u_{n+1-k}\}$ eine Basis von U. Jedes $x \in U$ lässt sich also schreiben als

$$x = \sum_{j=1}^{n+1-k} \alpha_j u_j. \tag{6.2}$$

Sei nun $\{v_1, \ldots, v_n\}$ die aus Satz 3.23 bekannte Orthonormalbasis aus Eigenvektoren zu den n Eigenwerten $\lambda_1, \ldots, \lambda_n$ von A. Die Bedingung $(x, v_i) = 0$ für $i = k + 1, \ldots, n$ ist wegen (6.2) äquivalent zu

$$\sum_{j=1}^{n+1-k} \alpha_j (u_j, v_i) = 0 \text{ für } i = k + 1, \ldots, n \tag{6.3}$$

Das System (6.3) ist ein unterbestimmtes lineares Gleichungssystem mit $n - k$ Gleichungen und $n - k + 1$ Variablen. Es gibt also eine nicht-triviale Lösung $\bar{\alpha}_1, \ldots, \bar{\alpha}_{n+1-k}$ und

$$\bar{y} := \sum_{j=1}^{n+1-k} \bar{\alpha}_j u_j$$

erfüllt entsprechend $(\bar{y}, v_i) = 0,\ i = k+1, \ldots, n$. Wir normieren $\bar{y}$ und definieren dadurch

$$y := \tfrac{1}{\|\bar{y}\|}\bar{y}.$$

Nach Konstruktion ist $y \in U$ und $\|y\| = 1$. Um nachzurechnen, dass $(Ay, y) \geq \lambda_k$ stellen wir y (nach Lemma 6.6) dar als

$$y = \sum_{j=1}^{n}(y, v_j)v_j = \sum_{j=1}^{k}(y, v_j)v_j.$$

Damit können wir nun nachrechnen, dass

$$\begin{aligned}
(Ay, y) &= \left(\sum_{j=1}^{k}(y, v_j)Av_j, \sum_{i=1}^{k}(y, v_i)v_i\right) = \left(\sum_{j=1}^{k}\lambda_j(y, v_j)v_j, \sum_{i=1}^{k}(y, v_i)v_i\right) \\
&= \sum_{i,j=1}^{k}\lambda_j(y, v_j)\overline{(y, v_i)}(v_j, v_i) = \sum_{j=1}^{k}\lambda_j|(y, v_j)|^2 \\
&\geq \lambda_k\sum_{j=1}^{k}|(y, v_j)|^2 = \lambda_k\|y\| = \lambda_k
\end{aligned}$$

QED

6.3 Lokalisierungssätze

Die Ergebnisse dieses Abschnitts werden *Lokalisierungssätze* genannt, da sie die Lage der Eigenwerte anhand der vorliegenden Matrixdaten abschätzten. Wir beginnen mit einer Folgerung aus dem Courant'schen Minimum-Maximum-Prinzip, und setzen dabei wieder einen Prä-Hilbert-Raum mit Skalarprodukt $\|x\| = \sqrt{(x, x)}$ voraus.

Satz 6.9 *Seien A und B hermitesche Matrizen mit jeweils n Eigenwerten*

$$\lambda_1(A) \geq \lambda_2(A) \geq \ldots \geq \lambda_n(A) \text{ und } \lambda_1(B) \geq \lambda_2(B) \geq \ldots \geq \lambda_n(B).$$

Dann gilt für alle $k = 1, \ldots, n$:

$$|\lambda_k(A) - \lambda_k(B)| \leq \|A - B\|$$

Beweis: Sei $v \in \mathbb{C}^n$ beliebig. Nach der Cauchy-Schwarz'schen Ungleichung (siehe Lemma 3.6) erhalten wir für die von $\|\cdot\|$ induzierte Operatornorm

$$((A - B)v, v) \leq \|(A - B)v\| \cdot \|v\| \leq \|A - B\|\|v\|^2.$$

Daraus ergibt sich für $k = 1, \ldots, n$:

$$
\begin{aligned}
& (Av, v) \leq (Bv, v) + \|A - B\|\|v\|^2 \text{ für alle } v \in \mathbb{C}^n \\
\Longrightarrow \quad & (Av, v) \leq \max_{v' \in U : \|v'\| = 1} (Bv', v') + \|A - B\|\|v'\|^2 \text{ für alle } v \in \mathbb{C}^n \text{ mit } \|v\| = 1 \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{und alle } U \in M_k \\
\Longrightarrow \quad & \max_{v \in U : \|v\| = 1} (Av, v) \leq \max_{v' \in U : \|v'\| = 1} (Bv', v') + \|A - B\| \text{ für alle } U \in M_k \\
\Longrightarrow \quad & \min_{U' \in M_k} \max_{v \in U' : \|v\| = 1} (Av, v) \leq \max_{v \in U : \|v\| = 1} (Bv, v) + \|A - B\| \text{ für alle } U \in M_k \\
\Longrightarrow \quad & \min_{U \in M_k} \max_{v \in U : \|v\| = 1} (Av, v) \leq \min_{U \in M_k} \max_{v \in U : \|v\| = 1} (Bv, v) + \|A - B\| \\
\Longrightarrow \quad & \lambda_k(A) \leq \lambda_k(B) + \|A - B\|,
\end{aligned}
$$

wobei für die letzte Folgerung Satz 6.8 verwendet wurde.
Durch Vertauschen von A und B erhält man $\lambda_k(B) \leq \lambda_k(A) + \|A - B\|$. Zusammen folgt für $k = 1, \ldots, n$

$$|\lambda_k(A) - \lambda_k(B)| \leq \|A - B\|.$$

QED

Bemerkung: Da $A - B$ hermitesch ist, gilt

$$\|A - B\|_2 = \rho(A - B) \leq \|A - B\| \text{ für jede beliebige Norm } \|\cdot\|$$

Da obiger Satz insbesondere für das Standard-Skalarprodukt zusammen mit der Euklidischen Norm $\|\cdot\|_2$ richtig ist, können wir also folgern, dass

$$|\lambda_k(A) - \lambda_k(B)| \leq \|A - B\|_2 \leq \|A - B\| \text{ für alle Normen } \|\cdot\|,$$

d.h. in Satz 6.9 darf im $\mathbb{C}^n$ (unabhängig von der Wahl des Skalarproduktes) jede Norm verwendet werden.

Als einfache Folgerung führen wir das folgende Korollar an, mit dessen Hilfe man abschätzen kann, wie weit sich die Eigenwerte einer Matrix von ihren Diagonalelementen unterscheiden.

Korollar 6.10 *Sei $A = (a_{ij})$ eine hermitesche Matrix und seien $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ ihre Eigenwerte. Permutiere die Diagonalelemente von A so dass $a'_{11} \geq a'_{22} \geq \ldots \geq a'_{nn}$. Dann gilt*

$$|\lambda_j - a'_{jj}|^2 \leq \sum_{i,k : i \neq k} |a_{ik}|^2, \text{ und } j = 1, \ldots, n.$$

Beweis: Die Aussage folgt direkt aus Satz 6.9 mit $B = diag(a_{11}, \ldots, a_{nn})$ und $\|\cdot\| = \|\cdot\|_F$. QED

Abschließend geben wir noch einen wichtigen Satz an, der sich *nicht* mit hermiteschen Matrizen beschäftigt, sondern auf alle Matrizen anwendbar ist. Dazu benötigen wir folgende Definition:

Definition 6.11 *Sei $A \in \mathbb{C}^{n,n}$ eine Matrix. Dann bezeichnet man*

$$K_i = \left\{ \lambda \in \mathbb{C} : |\lambda - a_{ii}| \leq \sum_{k \in \{1,\ldots,n\} \setminus \{i\}} |a_{ik}| \right\}, \ i = 1, \ldots, n$$

$$K_i^* = \left\{ \lambda \in \mathbb{C} : |\lambda - a_{ii}| \leq \sum_{k \in \{1,\ldots,n\} \setminus \{i\}} |a_{ki}| \right\}, \ i = 1, \ldots, n$$

als **Gerschgorin-Kreise** *von A. Wenn nicht klar ist, bezüglich welcher Matrix die Mengen K_i, K_i^* gebildet werden, schreibt man $K_i(A)$ bzw. $K_i(A^*)$.*

Lemma 6.12 *Sei $A \in \mathbb{C}^{n,n}$ und seien $K_i(A), K_i^*(A)$ die Gerschgorin-Kreise von A. Seien weiter $K_i(A^*), K_i^*(A^*)$ die Gerschgorin-Kreise der Matrix A^*. Dann folgt für jedes $\lambda \in \mathbb{C}$:*

$$\lambda \in K_i(A) \Longleftrightarrow \overline{\lambda} \in K_i^*(A^*).$$

Beweis: Weil für $\lambda, a \in \mathbb{C}$ gilt $|\lambda - a| = |\overline{\lambda} - \overline{a}|$ folgt für $\overline{A} = (\overline{a}_{ij})_{i,j=\{1,\ldots,n\}}$, dass

$$\lambda \in K_i(A) \Longleftrightarrow \overline{\lambda} \in K_i(\overline{A}).$$

Weiter gilt durch Vertauschen von Zeilen und Spalten, dass

$$\lambda \in K_i(A) \Longleftrightarrow \lambda \in K_i^*(A^T).$$

Kombiniert man beide Aussagen (unter Verwendung von $A^* = \overline{A}^T$), erhält man die Aussage des Lemmas. QED

Was die Gerschgorin-Kreise mit der Lokalisation von Eigenwerten zu tun haben, zeigt der folgende Satz.

Satz 6.13 (Satz von Gerschgorin) *Sei $A \in \mathbb{C}^{n,n}$ eine Matrix und sei λ ein beliebiger Eigenwert von A. Dann gilt*

$$\lambda \in \left(\bigcup_{i=1,\ldots,n} K_i \right) \cap \left(\bigcup_{i=1,\ldots,n} K_i^* \right).$$

Beweis: Sei λ ein Eigenwert von A. Um nachzuweisen, dass

$$\lambda \in \left(\bigcup_{i=1,\ldots,n} K_i \right) \cap \left(\bigcup_{i=1,\ldots,n} K_i^* \right)$$

konstruieren wir Indizes $k, l \in \{1, \ldots, n\}$, so dass $\lambda \in K_k$ und $\lambda \in K_l^*$.

Wähle einen Eigenvektor v mit $\|v\|_\infty = 1$. Sei k so dass $|v_k| = 1$. Wir zeigen, dass $\lambda \in K_k$:

Es gilt $Av = \lambda v$. Wir schreiben die Multiplikation für Zeile k aus

$$\sum_{j=1}^{n} a_{kj} v_j = \lambda v_k$$

und erhalten daraus

$$\sum_{j \in \{1,\ldots,n\} \setminus \{k\}} a_{kj} v_j = \lambda v_k - a_{kk} v_k$$

Weiter ergibt sich

$$\begin{aligned} |\lambda - a_{kk}| &= |(\lambda - a_{kk}) v_k| = \left| \sum_{j \in \{1,\ldots,n\} \setminus \{k\}} a_{kj} v_j \right| \\ &\leq \sum_{j \in \{1,\ldots,n\} \setminus \{k\}} |a_{kj}| \, |v_j| \leq \sum_{j \in \{1,\ldots,n\} \setminus \{k\}} |a_{kj}|, \end{aligned}$$

also $\lambda \in K_k$.

Nun möchten wir noch zeigen, dass $\lambda \in K_l^*$ für ein $l \in \{1, \ldots, n\}$. Dazu betrachten wir die Matrix A^*. Nach Lemma 6.4 wissen wir, dass $\overline{\lambda}$ ein Eigenwert von A^* ist, insbesondere gibt es nach dem schon bewiesenen Teil ein l, so dass $\overline{\lambda} \in K_l(A^*)$. Aus Lemma 6.12 folgt $\lambda \in K_l^*(A)$. QED

6.4 Potenzmethode zur Bestimmung des Spektralradius

Die folgende Methode liefert eine Abschätzung an den betragsgrößten Eigenwert einer Matrix. Ihre Konvergenz kann man (unter weiteren Voraussetzungen) für diagonalisierbare Matrizen nachweisen.

Definition 6.14 *Eine Matrix $A \in \mathbb{K}^{n,n}$ heißt* **diagonalisierbar**, *falls es eine reguläre Matrix $Q \in \mathbb{K}^{n,n}$ gibt, so dass*

$$Q^{-1} A Q = D$$

eine Diagonalmatrix ist.

Wegen Satz 3.23 ist jede hermitesche Matrix A diagonalisierbar. Wir konnten daraus außerdem folgern, dass die Spalten von Q eine Orthonormalbasis des $\mathbb{K}^n$ bilden, die aus Eigenvektoren von A besteht. Wir erweitern diese Aussage nun für diagonalisierbare Matrizen wie folgt:

Lemma 6.15 *Sei $A \in \mathbb{K}^{n,n}$ eine Matrix und seien $\lambda_1, \ldots, \lambda_n$ die n Eigenwerte von A. Dann gilt:*
A ist diagonalisierbar $\Longleftrightarrow$ Es gibt eine Basis $\{v_1, \ldots, v_n\}$ des $\mathbb{K}^n$, die aus Eigenvektoren von A besteht.

Beweis: Sei A diagonalisierbar. Wir konstruieren eine Basis des $\mathbb{K}^n$, die aus Eigenvektoren besteht: Weil A diagonalisierbar gibt es eine reguläre Matrix Q, so dass $Q^{-1}AQ = D$ eine Diagonalmatrix ist. Nach Lemma 6.5 ist $D = diag(\lambda_1, \ldots, \lambda_n)$. D hat also die Eigenwerte $\lambda_1, \ldots, \lambda_n$ und die Einheitsvektoren $e_1, \ldots, e_n$ sind jeweils zugehörige Eigenvektoren. Wir definieren

$$v_i := Qe_i,\ i = 1, \ldots, n$$

und erhalten (weil Q regulär) daraus eine weitere Basis des $\mathbb{K}^n$. Wir rechnen nun nach, dass v_i Eigenvektor von A zum Eigenwert λ_i ist:

$$\begin{aligned} Av_i &= AQe_i = (QDQ^{-1})Qe_i = QDe_i \\ &= Q\lambda_i e_i = \lambda_i Qe_i = \lambda_i v_i \end{aligned}$$

Sei andererseits $\{v_1, \ldots, v_n\}$ eine aus Eigenvektoren bestehende Basis des $\mathbb{C}^n$. Wir definieren $Q = (v_1 \ldots v_n)$ als die Matrix, deren Spalten die Eigenvektoren sind. Dann gilt:

$$\begin{aligned} AQ &= A(v_1 \ldots v_n) = (Av_1 \ldots Av_n) = (\lambda_1 v_1 \ldots \lambda_n v_n) \\ &= (v_1 \ldots v_n) diag(\lambda_1, \ldots, \lambda_n) := QD, \end{aligned}$$

also $Q^{-1}AQ = D$ und A ist diagonalisierbar. QED

Die Idee der Potenzmethode besteht darin, mit einem Startvektor $v^{(0)}$ zu beginnen und in jedem Iterationsschritt

$$v^{(k)} := A(v^{(k-1)}) = A^k(v^{(0)}) \tag{6.4}$$

zu berechnen. Mit Hilfe dieser einfachen Iterationsvorschrift kann man den Spektralradius einer diagonalisierbaren Matrix A bestimmen, falls sie einen *dominanten* Eigenwert hat.

Notation 6.16 *Sei $A \in \mathbb{K}^{n,n}$ eine Matrix. Ein Eigenwert λ von A heißt* **dominanter Eigenwert** *falls $|\lambda| > |\lambda'|$ für alle Eigenwerte $\lambda' \neq \lambda$ von A.*

Hat eine Matrix einen dominanten Eigenwert λ so ist $\rho(A) = |\lambda| \neq 0$ (falls $n > 1$). Wir können nun die Konvergenz der Iterationsvorschrift (6.4) untersuchen.

Satz 6.17 *Sei A eine diagonalisierbare Matrix mit Eigenwerten $\lambda_1, \ldots, \lambda_n$ und sei λ_1 ein dominanter Eigenwert. Sei $\{v_1, \ldots, v_n\}$ eine aus den Eigenvektoren von A bestehende Basis des $\mathbb{C}^n$.*
Sei $v^{(0)} = \sum_{i=1}^n \alpha_i v_i \in \mathbb{C}^n$ so dass $\alpha_1 \neq 0$ und sei $v^{(k)} = A(v^{(k-1)})$ für alle $k \geq 1$. Dann konvergiert die Folge

$$\frac{v^{(k)}}{\lambda_1^k} \to \alpha_1 v_1,$$

also gegen einen Eigenvektor zu λ_1.

Beweis: Wir berechnen

$$\begin{aligned}
v^{(1)} &= Av^{(0)} = \sum_{i=1}^n \alpha_i A v_i = \sum_{i=1}^n \alpha_i \lambda_i v_i \\
v^{(2)} &= Av^{(1)} = \sum_{i=1}^n \alpha_i \lambda_i A v_i = \sum_{i=1}^n \alpha_i \lambda_i^2 v_i \\
&\vdots \qquad\qquad \vdots \\
v^{(k)} &= Av^{(k-1)} = A^k v^{(0)} = \sum_{i=1}^n \alpha_i \lambda_i^k v_i \\
&= \lambda_1^k \cdot \left(\alpha_1 v_1 + \sum_{i=2}^n \underbrace{\frac{\lambda_i^k}{\lambda_1^k} \alpha_i v_i}_{\to 0 \text{ für } k \to \infty} \right).
\end{aligned}$$

Das heißt,

$$\frac{v^{(k)}}{\lambda_1^k} = \frac{\lambda_1^k \cdot \left(\alpha_1 v_1 + \sum_{i=2}^n \frac{\lambda_i^k}{\lambda_1^k} \alpha_i v_i \right)}{\lambda_1^k} \to \alpha_1 v_1 \text{ für } k \to \infty.$$

QED

Die Folge $\frac{v^{(k)}}{\lambda_1^k}$ konvergiert also gegen einen Eigenvektor zum betragsgrößten Eigenwert von A. Damit ist also $\frac{v^{(k)}}{\lambda_1^k}$ und ebenso $v^{(k)}$ eine Approximation an den Eigenvektor v_1. Um nicht nur den Eigenvektor, sondern auch den Eigenwert λ_1 zu bestimmen, nutzt man aus, dass für den (approximierten) Eigenvektor $v^{(k)}$

$$Av^{(k)} \approx \lambda_1 v^{(k)}$$

gilt, also insbesondere

$$|\lambda_1| \approx \frac{\|Av^{(k)}\|}{\|v^{(k)}\|} = \frac{\|v^{(k+1)}\|}{\|v^{(k)}\|}.$$

Der Beweis von Satz 6.17 zeigt weiterhin, dass die Konvergenz des Verfahrens umso schneller ist, wenn die Quotienten $\left|\frac{\lambda_i}{\lambda_1}\right|$ klein sind. Sortiert man die Eigenwerte nach ihren Beträgen, so ist also $q := \frac{|\lambda_1|}{|\lambda_2|}$ der bestimmende Kontraktionsfaktor.

In der Praxis möchte man die durch die λ_1^k erzeugten sehr großen (falls $|\lambda_1| > 1$) oder sehr kleinen (falls $|\lambda_1| < 1$) Werte vermeiden. Daher verändert man die Iterationsvorschrift (6.4) und führt in jedem Schritt noch eine Normierung ein. Man erhält das auf von Mises zurückgehende Verfahren der Vektoriteration. Obwohl man das Verfahren allgemeiner anwenden kann, ist seine Konvergenz nur für diagonalisierbare Matrizen und den richtigen Startwert gesichert.

Algorithmus 12: Verfahren der Vektoriteration nach von Mises

Input: $A \in \mathbb{C}^{n,n}$, $v^{(0)} \in \mathbb{C}^n$, `Schranke` ε.

Schritt 1: $k := 0$

Schritt 2: `Repeat`

Schritt 2.1: $y^{(k+1)} := Av^{(k)}$

Schritt 2.2: $v^{(k+1)} := \frac{y^{(k+1)}}{\|y^{(k+1)}\|_2}$

Schritt 2.3: $k := k + 1$

Until $\left|\|Av^{(k)}\| - \|y^{(k)}\|\right| \leq \varepsilon$

Ergebnis: $\|y^k\|_2$ `ist Approximation an den betragsgrößten Eigenwert und` v^{k-1} `ist Approximation an den zugehörigen Eigenvektor.`

Die Vektoren $v^{(k)}$ sind in obigem Algorithmus also alle normiert, unterscheiden sich sonst aber nicht von den mit Hilfe der Iterationsvorschrift (6.4) erzeugten Vektoren. Damit folgt die Konvergenz der Folge $\frac{v^{(k)}}{\lambda_1^k}$ aus Satz 6.17. Der größte Eigenwert λ_1 ergibt sich unter der Bezeichnung von Algorithmus 12 aus

$$|\lambda_1| \approx \frac{\|Av^{(k)}\|}{\|v^{(k)}\|} = \|Av^{(k)}\| = \|y^{(k+1)}\|.$$

Das Abbruckkriterium lässt sich umformulieren zu

$$\left|\|Av^{(k)}\| - \|y^{(k)}\|\|v^k\|\right| \approx \left|\|Av^{(k)}\| - |\lambda_1|\|v^k\|\right|$$

und gibt somit an, wie gut die Approximation schon ist. Dabei wird in der vorliegenden Version nur der Betrag von λ_1, nicht aber sein Vorzeichen bestimmt. Das Vorzeichen kann jedoch leicht mitbestimmt werden, siehe dazu Übung (Blatt 11, Aufgabe 4).

Die Potenzmethode berechnet nur den betragsgrößten Eigenwert, lässt sich aber leicht modifizieren, so dass auch andere Eigenwerte bestimmt werden können. Sei dazu $A \in \mathbb{C}^{n,n}$ eine Matrix mit den Eigenwerten $\lambda_1, \ldots, \lambda_n$.

- Sucht man den betragsmäßig kleinsten Eigenwert von A, so nutzt man aus, dass die Eigenwerte von A^{-1} genau die Kehrwerte $\frac{1}{\lambda_i}$ der Eigenwerte λ_i der Matrix A sind. Verwendet man also die Matrix A^{-1} im von Mises-Verfahren, erhält man
$$\max_{i=1,\ldots,n} \left| \frac{1}{\lambda_i} \right| = \frac{1}{\min_{i=1,\ldots,n} |\lambda_i|},$$
also den Kehrwert des betragsmäßig kleinsten Eigenwertes von A. Dieses Verfahren wird auch *Verfahren der inversen Vektoriteration* genannt.

- Möchte man einen anderen Eigenwert λ_j von A bestimmen, für den man einen Näherungswert σ kennt, so ersetzt man A im von Mises-Verfahren durch die Matrix $(A - \sigma I)^{-1}$. Diese Matrix hat genau die Eigenwerte $(\lambda_i - \sigma)^{-1}$, $i = 1, \ldots, n$. Damit ergibt sich
$$\max_{i=1,\ldots,n} \left| \frac{1}{(\lambda_i - \sigma)} \right| = \frac{1}{\min_{i=1,\ldots,n} |\lambda_i - \sigma|},$$
d.h. ist der Output des von Mises-Verfahren ein Eigenwert μ von $(A - \sigma I)^{-1}$ so ist
$$\frac{1}{\mu} + \sigma$$
der Eigenwert von A, der am nächsten an σ liegt. Diese Variation des von Mises-Verfahrens ist auch als *Inverse Iteration mit Shift nach Wielandt* bekannt.

Übung: Sei $A \in \mathbb{C}^{n,n}$ eine Matrix und sei $\sigma \in \mathbb{R}$. Sei weiter $(A - \sigma I)$ regulär. Zeigen Sie: λ ist Eigenwert von A genau dann wenn $\frac{1}{\lambda - \sigma}$ Eigenwert ist von $(A - \sigma I)^{-1}$.

6.5 Das QR-Verfahren zur Bestimmung aller Eigenwerte

Das im folgenden vorgestellte Verfahren ist das wohl wichtigste zur Berechnung der Eigenwerte einer Matrix. Es beruht auf der QR-Zerlegung einer Matrix, die

wir in Kapitel 4.1 (für nicht notwendigerweise quadratische Matrizen) behandelt haben. Zur Bestimmung der Eigenwerte wenden wir die QR-Zerlegung auf quadratische Matrizen an. Zur Erinnerung:

Die Zerlegung einer Matrix $A \in \mathbb{K}^{n,n}$ der Form $A = QR$ mit einer unitären Matrix $Q \in \mathbb{K}^{n,n}$ und einer oberen Δs-Matrix $R \in \mathbb{K}^{n,n}$ heißt ***QR*-Zerlegung** von A.

Wir beginnen mit dem folgenden Lemma über die Eindeutigkeit der QR-Zerlegung.

Lemma 6.18 *Sei $A \in \mathbb{K}^{n,n}$ eine reguläre Matrix und seien $A = Q_1R_1 = Q_2R_2$ zwei QR-Zerlegungen von A mit unitären Matrizen $Q_1, Q_2 \in \mathbb{K}^{n,n}$ und oberen Dreiecksmatrizen $R_1, R_1 \in \mathbb{K}^{n,n}$. Dann gibt es eine unitäre Diagonalmatrix $S \in \mathbb{K}^{n,n}$ mit $Q_2 = Q_1S$ und $R_1 = SR_2$.*

Beweis: Weil alle beteiligten Matrizen regulär sind, folgt aus $Q_1R_1 = Q_2R_2$ dass $Q_1^*Q_2 = R_1R_2^{-1}$. Wir definieren

$$S := Q_1^*Q_2 = R_1R_2^{-1}.$$

Dann gilt zunächst $Q_2 = Q_1S$ und $R_1 = SR_2$ und S ist als Produkt von unitären Matrizen selbst unitär. Entsprechend gilt $S^* = S^{-1}$. Als Produkt von zwei oberen Dreiecksmatrizen ist S selbst eine obere Dreiecksmatrix, ebenso auch S^{-1} (siehe Satz 2.8). Andererseits ist S^* eine untere Dreiecksmatrix. Weil S unitär, gilt $S^{-1} = S^*$, also ist S eine Diagonalmatrix. QED

Die Idee des QR-Verfahrens, besteht darin, eine gegebene Matrix A durch iteratives Vertauschen von Q und R auf eine obere Dreiecksmatrix zu transformieren. Bevor wir untersuchen, wann das funktioniert, geben wir das Verfahren an:

Algorithmus 13: ***QR*-Verfahren zur Bestimmung der Eigenwerte einer Matrix**

Input: $A \in \mathbb{C}^{n,n}$

Schritt 1: $m := 0$, $A^{(0)} := A$

Schritt 2: `Repeat`

> **Schritt 2.1:** `Bestimme eine` QR-`Zerlegung von` $A^{(m)}$, `also` $A^{(m)} = Q_mR_m$
>
> **Schritt 2.2:** $A^{(m+1)} := R_mQ_m$
>
> **Schritt 2.3:** $m := m + 1$
>
> **Until** `Stop`

Ergebnis: `Approximierte Eigenwerte` $a_{ii}^{(m)}$, $i = 1, \ldots, n$ `auf der Hauptdiagonalen von` $A^{(m)}$.

Um dieses Verfahren zu untersuchen, zeigen wir zunächst, dass $A^{(m)}$ für alle $m \in \mathbb{N}$ die gleichen Eigenwerte hat wie die Originalmatrix A. Anschließend untersuchen wir, wann $A^{(m)}$ gegen eine obere Dreiecksmatrix konvergiert.

Notation 6.19 *Mit den Bezeichnungen aus Algorithmus 13 definieren wir für alle $m \in \mathbb{N}$ die beiden folgenden Matrizen:*

$$\begin{aligned} \mathcal{Q}_m &:= Q_0 Q_1 \cdots Q_{m-1} \\ \mathcal{R}_m &:= R_{m-1} \cdots R_1 R_0 \end{aligned}$$

Lemma 6.20 *Mit den Bezeichnungen aus Algorithmus 13 gilt für alle $m \in \mathbb{N}$:*

1. $A^{(m)} = Q_{m-1}^* A^{(m-1)} Q_{m-1}$
2. $A^{(m)} = \mathcal{Q}_m^* A \mathcal{Q}_m$
3. $A^m = \mathcal{Q}_m \mathcal{R}_m$

Beweis:

1. Wir rechnen

$$A^{(m)} = R_{m-1} Q_{m-1} = Q_{m-1}^* \underbrace{Q_{m-1} R_{m-1}}_{A^{(m-1)}} Q_{m-1} = Q_{m-1}^* A^{(m-1)} Q_{m-1}$$

2. Diese Aussage zeigen wir per Induktion. Der Induktionsanfang ergibt sich wegen $A^{(0)} = Q_0 R_0$ zu

$$A^{(1)} = R_0 Q_0 = \underbrace{Q_0^* A^{(0)}}_{R_0} Q_0 = \mathcal{Q}_1^* A \mathcal{Q}_1.$$

   Für den Induktionsschritt nehmen wir an, dass $A^{(m-1)} = \mathcal{Q}_{m-1}^* A \mathcal{Q}_{m-1}$. Dann rechnen wir

$$\begin{aligned} A^{(m)} &= Q_{m-1}^* A^{(m-1)} Q_{m-1} \quad \text{nach Teil 1 des Lemmas} \\ &= Q_{m-1}^* \mathcal{Q}_{m-1}^* A \mathcal{Q}_{m-1} Q_{m-1} \quad \text{nach der Induktionsannahme} \\ &= \mathcal{Q}_m^* A \mathcal{Q}_m. \end{aligned}$$

3. Auch diesen Teil zeigen wir durch Induktion. Der Induktionsanfang folgt wegen

$$A^1 = A = A^{(0)} = Q_0 R_0 = \mathcal{Q}_1 \mathcal{R}_1.$$

Für den Induktionsschritt nehmen wir an, dass $A^{m-1} = \mathcal{Q}_{m-1}\mathcal{R}_{m-1}$ und rechnen

$$\begin{aligned} A^m &= AA^{m-1} = \mathcal{Q}_{m-1}A^{(m-1)}\mathcal{Q}_{m-1}^* A^{m-1} \quad \text{nach Teil 2} \\ &= \mathcal{Q}_{m-1}A^{(m-1)}\mathcal{Q}_{m-1}^*\mathcal{Q}_{m-1}\mathcal{R}_{m-1} \quad \text{nach der Induktionsannahme} \\ &= \mathcal{Q}_{m-1}A^{(m-1)}\mathcal{R}_{m-1} = \mathcal{Q}_{m-1}Q_{m-1}R_{m-1}\mathcal{R}_{m-1} = \mathcal{Q}_m\mathcal{R}_m. \end{aligned}$$

QED

Der erste Teil des Lemmas zeigt, dass im QR-Algorithmus tatsächlich in jedem Schritt eine Ähnlichkeitstransformation ausgeführt wird, und entsprechend alle $A^{(m)}$ die gleichen Eigenwerte besitzen. Um die Eigenwerte von A zu identifizieren, können wir also die Eigenwerte von $A^{(m)}$ berechnen. Das hilft uns jedoch nur, wenn $A^{(m)}$ gegen eine Form konvergiert, für die wir die Eigenwerte auch leicht ausrechnen können. Das folgende Ergebnis zeigt, unter welchen Voraussetzungen $A^{(m)}$ gegen eine obere Dreiecksmatrix konvergiert, die Zerlegung

$$A = \mathcal{Q}_m A^{(m)} \mathcal{Q}_m^*$$

also nach Teil 2 des Lemmas gegen eine Schur-Zerlegung konvergiert.

Satz 6.21 *Sei $A \in \mathbb{C}^{n,n}$ eine diagonalisierbare Matrix mit Eigenwerten $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n| > 0$ und zugehörigen Eigenvektoren $v_1, \ldots, v_n$. Sei*

$$A = VDV^{-1}$$

mit $D = diag(\lambda_1, \ldots, \lambda_n)$ und $V := (v_1, \ldots, v_n)$. Wir nehmen weiter an, dass V^{-1} eine LU-Zerlegung besitzt.
Dann konvergieren die Matrizen $A^{(m)}$ aus Algorithmus 13 gegen eine obere Dreiecksmatrix. Die Diagonaleinträge $a_{ii}^{(m)}$ von $A^{(m)}$ konvergieren dabei linear gegen die Eigenwerte.

Beweis: In dem nun folgenden (technischen) Beweis formen wir $A^{(m)}$ um zu einer oberen Dreiecksmatrix $S_m U_m D U_m^{-1} S_m^*$ und einem Restterm, dessen Norm mit $O(q^m)$ für ein q mit $|q| < 1$ gegen Null geht. Um diese Zerlegung zu erzeugen, gehen wir in folgenden Schritten vor:

1. Zu der aus Teil 3 von Lemma **??** schon bekannten QR-Zerlegung vonA^m erzeugen wir eine weitere:

   Weil $V^{-1} = LU$ und $D = diag(\lambda_1, \ldots, \lambda_n)$ regulär ist, definieren wir

   $$V_m := VD^m LD^{-m} \text{ mit einer } QR\text{-Zerlegung } V_m = P_m U_m. \qquad (6.5)$$

Dann ist

$$\begin{aligned} A^m &= (VDV^{-1})^m = VD^mV^{-1} \\ &= VD^mLU = VD^mLD^{-m}D^mU \\ &= V_mD^mU = P_mU_mD^mU \\ &= \underbrace{P_m}_{\text{unitär}} \quad \underbrace{U_mD^mU}_{\text{obere } \Delta\text{-Matrix}} \end{aligned}$$

eine QR-Zerlegung von A^m. Diese vergleichen wir mit $A^m = \mathcal{Q}_m\mathcal{R}_m$ und erhalten aus Lemma 6.18 eine unitäre Diagonalmatrix S mit

$$\mathcal{Q}_m = P_mS_m^* \text{ und } \mathcal{R}_m = S_mU_mD^mU. \tag{6.6}$$

2. Daraus gewinnen wir eine neue Darstellung von $A^{(m)}$:

   Wir möchten mit $A^{(m)} = Q_mR_m$ rechnen und nutzen zunächst (6.6) um Q_m und R_m wie folgt auszudrücken.

$$\begin{aligned} Q_m &= Q_{m-1}^{-1} \cdot \ldots \cdot Q_0^{-1} \cdot Q_0 \cdot \ldots \cdot Q_m \\ &= \mathcal{Q}_m^{-1}\mathcal{Q}_{m+1} \\ &= S_mP_m^{-1}P_{m+1}S_{m+1}^* \text{ siehe (6.6)} \\ R_m &= R_m \cdot \ldots R_0 \cdot R_0^{-1} \cdot \ldots \cdot R_{m-1}^{-1} \\ &= \mathcal{R}_{m+1}\mathcal{R}_m^{-1} \\ &= S_{m+1}U_{m+1}D^{m+1}UU^{-1}D^{-m}U_m^{-1}S_m^* \text{ siehe (6.6)} \\ &= S_{m+1}U_{m+1}DU_m^{-1}S_m^* \end{aligned}$$

   Daraus erhalten wir zunächst folgende Darstellung für $A^{(m)}$:

$$\begin{aligned} A^{(m)} &= Q_mR_m = S_mP_m^{-1}P_{m+1}S_{m+1}^*S_{m+1}U_{m+1}DU_m^{-1}S_m^* \\ &= S_m(U_mU_m^{-1})P_m^{-1}P_{m+1}U_{m+1}DU_m^{-1}S_m^* \\ &= S_mU_mV_m^{-1}V_{m+1}DU_m^{-1}S_m^* \end{aligned} \tag{6.7}$$

3. An der eben erzielten Darstellung von $A^{(m)}$ "stört" uns der Term $V_m^{-1}V_{m+1}$, denn ohne ihn wäre $A^{(m)}$ als das Produkt von Diagonalmatrizen und oberen Dreiecksmatrizen selbst eine obere Dreiecksmatrix. Daher beschäftigen wir uns nun mit der Assymptotik von V_m und $V_m^{-1}V_{m+1}$.

   Zunächst betrachten wir $V_m = VD^mLD^{-m}$ für $m \to \infty$. L ist eine normierte Dreiecksmatrix und D enthält die der Größe nach sortierten Eigenwerte auf der Hauptdiagonalen. Wir definieren

$$q := \max_{i=2,\ldots,n} \frac{|\lambda_i|}{|\lambda_{i-1}|} \in (0,1).$$

Sei $\tilde{L}$ eine beliebige normierte untere Dreiecksmatrix. Wir berechnen die Elemente des Produktes

$$(D^m \tilde{L} D^{-m})_{ij} = \lambda_i^m \tilde{l}_{ij} \frac{1}{\lambda_j^m} = \begin{cases} 0 & \text{falls } i < j \\ 1 & \text{falls } i = j \\ \tilde{l}_{ij} \left(\frac{\lambda_i}{\lambda_j}\right)^m & \text{falls } i > j \end{cases}$$

Dabei gilt, dass für $i > j$

$$\frac{\lambda_i}{\lambda_j} = \frac{\lambda_i \cdot \lambda_{i-1} \cdot \ldots \cdot \lambda_{j+1}}{\lambda_{i-1} \cdot \lambda_{i-2} \cdot \ldots \cdot \lambda_j} \leq q^{i-j} < q,$$

also

$$\tilde{l}_{ij} \left(\frac{\lambda_i}{\lambda_j}\right)^m \in \mathrm{O}(q^m).$$

$D^m \tilde{L} D^{-m}$ ist also eine normierte, untere Dreiecksmatrix, deren Nichtnull-Einträge linear gegen Null konvergieren.

Wählen wir $\tilde{L} = L$ und multiplizieren die "fast" Diagonalmatrix $D^m L D^{-m}$ von rechts mit V so erhalten wir

$$V_m = V D^m L D^{-m} = V + E_m \text{ mit } \|E_m\|_2 \in \mathrm{O}(q^m), m \to \infty.$$

Für V_m^{-1} ergibt sich mit $\tilde{L} = L^{-1}$, dass

$$V_m^{-1} = D^m L^{-1} D^{-m} V^{-1} = V^{-1} + E'_m \text{ mit } \|E'_m\|_2 \in \mathrm{O}(q^m), m \to \infty.$$

Als letztes untersuchen wir noch das Produkt $V_m^{-1} V_{m+1}$ und erhalten

$$\begin{aligned} V_m^{-1} V_{m+1} &= D^m L^{-1} D^{-m} V^{-1} V D^{m+1} L D^{-m-1} \\ &= D^m L^{-1} D L D^{-1} D^{-m}. \end{aligned}$$

Mit der normierten unteren Dreiecksmatrix $\tilde{L} = L^{-1} D L D^{-1}$ ergibt sich

$$V_m^{-1} V_{m+1} = I + F_m \text{ mit } \|F_m\|_2 \in \mathrm{O}(q^m), m \to \infty.$$

4. Nun setzen wir unsere Analyse für $m \to \infty$ in die für $A^{(m)}$ hergeleitete Formel (6.7) ein und erhalten

$$A^{(m)} = S_m U_m V_m^{-1} V_{m+1} D U_m^{-1} S_m^* = S_m U_m D U_m^{-1} S_m^* + S_m U_m F_m D U_m^{-1} S_m^*$$

Wir untersuchen die Norm des zweiten Summanden und schätzen dazu zunächst ab:

$$\begin{aligned} \|S_m\|_2 = \|S_m^*\|_2 &= 1 \text{ weil } S_m \text{ unitär} \\ \|U_m\|_2 &= \|P_m^* V_m\|_2 = \|V_m\|_2 \text{ weil } P_m \text{ unitär} \\ &= \|V + E_m\|_2 \leq \|V\|_2 + \|E_m\|_2 \\ \|U_m^{-1}\|_2 &= \|V^{-1} + E'_m\|_2 \leq \|V^{-1}\|_2 + \|E'_m\|_2 \\ \|D\|_2 &= \sqrt{\rho(D^* D)} = \sqrt{\lambda_1^2} = |\lambda_1| \end{aligned}$$

Damit ergibt sich

$$\begin{aligned}\|S_m U_m F_m D U_m^{-1} S_m^*\|_2 &\leq \|S_m\|_2 \, \|U_m\|_2 \, \|F_m\|_2 \, \|D\|_2 \, \|U_m^{-1}\|_2 \, \|S_m^*\|_2 \\ &\leq (\|V\|_2 + \|E_m\|_2)\|F_m\|_2|\lambda_1|(\|V^{-1}\|_2 + \|E'_m\|_2) \in \mathrm{O}(q^m)\end{aligned}$$

Das heißt,

$$A^{(m)} \approx S_m U_m D U_m^{-1} S_m^*,$$

konvergiert also gegen eine obere Dreiecksmatrix.

QED

Abschließend geben wir noch ein Kriterium für die Existenz einer LU-Zerlegung von V^{-1} an.

Lemma 6.22 *Für die Matrix V^{-1} existiert genau dann eine LU-Zerlegung, wenn*

$$span\{e_1, \ldots, e_k\} \cap span\{v_{k+1}, \ldots, v_n\} = \{0\} \text{ für alle } k = 1, \ldots, n.$$

Beweis:
Ein Vektor x ist in $span\{e_1, \ldots, e_k\} \cap span\{v_{k+1}, \ldots, v_n\}$ genau dann, wenn es eindeutige Koeffizienten α_j, β_j gibt, so dass

$$x = \sum_{j=1}^{k} \alpha_j e_j = \sum_{j=k+1}^{n} \beta_j v_j$$

Multiplikation mit V^{-1} führt zu dem äquivalenten System

$$V^{-1}x = \sum_{j=1}^{k} \alpha_j V^{-1} e_j = \sum_{j=k+1}^{n} \beta_j e_j.$$

Es liegt also ein Gleichungssystem vor mit n Unbekannten $\alpha_1, \ldots, \alpha_k, \beta_{k+1}, \ldots, b_n$ und Koeffizientenmatrix

$$C(k) = \left(\begin{array}{ccc|ccc} & & & & 0 & \\ V^{-1}e_1 & \ldots & V^{-1}e_k & & & \\ & & & -1 & & \\ & & & & \ddots & \\ & & & & & -1 \end{array} \right)$$

Dabei stimmt $|det(C(k))|$ genau mit dem k.ten Hauptminor von V^{-1} überein. Folglich gilt:

$$span\{e_1, \ldots, e_k\} \cap span\{v_{k+1}, \ldots, v_n\} = \{0\} \text{ für alle } k = 1, \ldots, n$$

genau dann wenn das Gleichungssystem mit Koeffizientenmatrix $C(k)$ eindeutig lösbar ist. Das ist genau dann der Fall, wenn $det(C(k)) \neq 0$ gilt, und das wiederum ist genau die Bedingung, dass der k.te Hauptminor von V^{-1} nicht Null ist. Von Satz 2.12 wissen wir, dass das äquivalent dazu ist, dass eine LU-Zerlegung von V^{-1} existiert. QED

Das Verfahren in der hier vorgestellten Basisvariante ist allerdings nicht sonderlich effizient. Es kann durch verschiedene Tricks noch erheblich beschleunigt werden. Bei dem *Shift QR-Verfahren* führt man in einem Reduktionsschritt die gegebene Matrix zuerst durch Ähnlichkeitstransformationen in eine obere Hessenbergmatrix über und arbeitet in den dann folgenden Operationen mit einem variablen Shift Parameter.

6.6 Das Jakobiverfahren

In diesem Abschnitt beschäftigen wir uns mit der Bestimmung aller Eigenwerte und Eigenvektoren einer symmetrischen, reellen Matrix $A \in \mathbb{R}^{n,n}$. Das Jacobi-Verfahren ist zwar im allgemeinen langsamer als der (optimierte) QR-Algorithmus, es kann aber kleine Eigenwerte und die zugehörigen Eigenwerte mit einem geringen Fehler berechnen und ist daher dennoch von Interesse.

Sei $A \in \mathbb{R}^{m \times m}$ symmetrisch. Wir betrachten die Frobenius-Norm

$$\|A\|_{\mathrm{F}} = \left[\sum_{i,j=1}^{n} |a_{ij}|^2 \right]^{\frac{1}{2}}$$

und geben für sie zunächst ein paar Eigenschaften an.

Lemma 6.23

1. $\|A\|_F = Spur(A^T A) = Spur(AA^T)$
2. *Für jede orthogonale Matrix Q gilt:* $\|A\|_F = \|Q^T AQ\|_F$.
3. *Ist A symmetrisch und hat die Eigenwerte $\lambda_1, \ldots, \lambda_n$, so gilt* $\|A\|_F^2 = \sum_{i=1}^{n} |\lambda_i|^2$.

Beweis:

1. Für die Spur eines Matrixproduktes AB mit $A, B \in \mathbb{R}^{m \times m}$ gilt

$$\mathrm{Spur}(AB) = \sum_{i,j=1}^{n} a_{ij} b_{ji} = \sum_{i,j=1}^{n} b_{ij} a_{ji} = \mathrm{Spur}(BA),$$

insbesondere

$$\mathrm{Spur}(A^T A) = \mathrm{Spur}(AA^T) = \sum_{i,j=1}^{n} |a_{ij}|^2.$$

2. Sei $Q \in \mathbb{R}^{m\times m}$ orthogonal. Dann gilt $Q^{-1} = Q^T$. Wir verwenden Teil 1 des Lemmas und rechnen

$$\begin{aligned}\|Q^TAQ\|_{\mathrm{F}} &= \mathrm{Spur}(Q^TAQQ^TA^TQ) = \mathrm{Spur}(Q^TAA^TQ) = \mathrm{Spur}(A^TQQ^TA)\\ &= \mathrm{Spur}(A^TA) = \|A\|_{\mathrm{F}}.\end{aligned}$$

3. Ist $A \in \mathbb{R}^{m\times m}$ symmetrisch, so gibt es nach Satz 3.23 eine orthogonale Matrix Q, so dass $Q^TAQ = diag(\lambda_1, \ldots, \lambda_n)$ eine Diagonalmatrix ist. Nach Teil 2 des Lemmas ist dann

$$\|A\|_F = \|Q^TAQ\|_F = \|diag(\lambda_1, \ldots, \lambda_n)\|_F = \sum_{i=1}^{n} |\lambda_i|^2.$$

QED

Definition 6.24 *Die* **Außennorm** *einer Matrix $A \in \mathbb{K}^{n,n}$ ist definiert als*

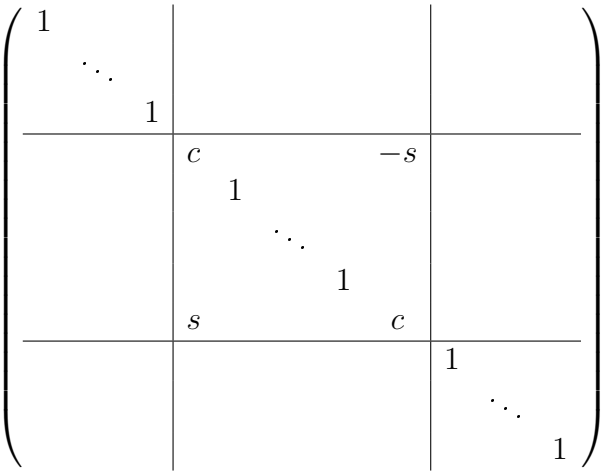
$$N(A) := \sum_{\substack{i,j=1\\ i\neq j}}^{n} |a_{ij}|^2 = \|A\|_F^2 - \sum_{i=1}^{n} |a_{ii}|^2.$$

Vorsicht: Die Bezeichnung *Außennorm* von A ist irreführend, denn die Abbildung $N(A)$ ist keine Norm.

Sei $A = diag(b_1, \ldots, b_n)$ eine Diagonalmatrix. Dann gilt $N(A) = 0$, die Außennorm verschwindet also für Diagonalmatrizen. Für beliebige Matrizen ist die Außennorm ein Maß dafür, wie groß die Nicht-Diagonalelemente sind. Geht die Außennorm gegen Null, so tendiert die Matrix zu einer Diagonalmatrix und man kann die Einträge auf der Hauptdiagonalen als Abschätzung für die Eigenwerte verwenden.

Die Idee des Jacobi-Verfahrens besteht darin, Nichtnull-Elemente der Matrix A nach und nach verschwinden zu lassen. Das geschieht durch Ähnlichkeitstransformationen, in denen A in $G_{ij}AG_{ij}^T$ überführt wird. Als Matrix G_{ij} wählt man so genannte *Givens-Rotationen*, die wie folgt aussehen.

Definition 6.25 *Sei $1 \leq i < j \leq n$. Dann nennen wir die Matrix*



mit $g_{ii} = g_{jj} = c = \cos\varphi$, $g_{ij} = -s = \sin\varphi$ und $g_{ji} = s = -\sin\varphi$ eine **Givens-Rotation** *bzw. eine* **Jakobi-Transformation**.

Lemma 6.26 *Sei A eine* Givens-Rotation. *Dann ist A eine orthogonale Matrix.*

Beweis: Offensichtlich sind die Spalten G_k für $k \in \{1, \ldots, n\} \setminus \{i, j\}$ alle paarweise orthogonal und erfüllen $\|G_j\| = 1$. Weiterhin ist Spalte i (bzw. Spalte j) zu allen $k \in \{1, \ldots, n\} \setminus \{i, j\}$ orthogonal. Wir rechnen noch nach, dass

$$(G_i, G_j) = c(-s) + cs = 0$$

und dass für $l \in \{i, j\}$ gilt

$$\|G_l\|^2 = c^2 + s^2 = \cos^2\varphi + \sin^2\varphi = 1.$$

QED

Nun betrachten wir, was passiert, wenn man eine Givens-Transformation auf eine symmetrische Matrix A anwendet.

Lemma 6.27 *Sei $A \in \mathbb{R}^{n \times n}$ symmetrisch, $i \neq j$ mit $a_{ij} \neq 0$. Sei*

$$B = G_{ij} A G_{ij}^T$$

1. *B ist symmetrisch.*
2. *$b_{kl} = a_{kl}$ für $k, l \notin \{i, j\}$, d.h. die Givens-Rotation verändert keine Elemente außerhalb der Zeilen i, j und der Spalten i, j.*
3. *Speziell für die Diagonalelemente gilt:*

$$\begin{aligned} b_{ii} &= c^2 a_{ii} - 2csa_{ij} + s^2 a_{jj} \\ b_{jj} &= s^2 a_{ii} + 2csa_{ij} + c^2 a_{jj} \\ b_{kk} &= a_{kk} \; sonst \end{aligned}$$

4. Das Element $b_{ij} = b_{ji}$ erfüllt: $b_{ij} = cs(a_{ii} - a_{jj}) + (c^2 - s^2)a_{ij}$

Beweis: Die Symmetrie sieht man aufgrund der Orthogonalität der G_{ij} direkt durch

$$B^T = \left(G_{ij}AG_{ij}^T\right)^T = G_{ij}A^TG_{ij}^T = G_{ij}AG_{ij}^T = B.$$

Die anderen Aussagen muss man nachrechnen.

QED

Satz 6.28 *Sei $B = G_{ij}AG_{ij}^T$. Dann gilt*

$$N(B) = N(A) - 2a_{ij}^2 + 2b_{ij}^2.$$

Beweis: Etwas mühsam rechnet man zuerst anhand der Formeln in Lemma 6.27 nach, dass

$$a_{ii}^2 + a_{jj}^2 + 2a_{ij}^2 = b_{ii}^2 + b_{jj}^2 + 2b_{ij}^2. \tag{6.8}$$

Wegen Lemma 6.26 ist G_{ij}^T orthogonal, so dass nach dem zweiten Teil von Lemma 6.23 $\|A\|_F = \|G_{ij}AG_{ij}^T\|_F = \|B\|_F$ gilt. Mit

$$\begin{aligned}
N(B) &= \sum_{\substack{i,j=1\\ i\neq j}}^{n} |b_{ij}|^2 = \|B\|_F^2 - \sum_{i=1}^{n} |b_{ii}|^2 \\
&= \|A\|_F^2 - \sum_{i=1}^{n} |b_{ii}|^2 \quad \text{(Lemma 6.23)} \\
&= \|A\|_F^2 - \sum_{i=1}^{n} |a_{ii}|^2 + a_{ii}^2 + a_{jj}^2 - b_{ii}^2 - b_{jj}^2 \quad \text{(Teil 3 aus Lemma 6.27)} \\
&= \|A\|_F^2 - \sum_{i=1}^{n} |a_{ii}|^2 + 2b_{ij}^2 - 2a_{ij}^2 \quad \text{siehe (6.8)}
\end{aligned}$$

erhält man das gewünschte Ergebnis. QED

Wenn die Givens-Rotation G_{ij} eine Matrix B mit $b_{ij} = 0$ erzeugt, minimieren wir also die Außennorm der Matrix $B = G_{ij}AG_{ij}^T$. Wir suchen nun eine Givens-Rotation, die genau das leistet, die nach Teil 4 von Lemma 6.27 also

$$b_{ij} = cs(a_{ii} - a_{jj}) + (c^2 - s^2)a_{ij}$$

erfüllt. Das kann man auflösen zu

$$\cot(2\varphi) = \frac{a_{jj} - a_{ii}}{2a_{ij}} \tag{6.9}$$

und diese Gleichung ist in dem Intervall $(-\frac{\pi}{4}, \frac{\pi}{4}]$ eindeutig lösbar (wobei wir hier keine Eindeutigkeit brauchen).

Mit der Kenntnis dieser Ergebnisse formulieren wir das Jakobi-Verfahren:

Algorithmus 14: Jakobi-Verfahren zur Eigenwertbestimmung

Input: $A \in \mathbb{R}^{n,n}$

Schritt 1: $k := 0$, $A^{(0)} := A$

Schritt 2: `Repeat`

> **Schritt 2.1:** `Bestimme ein Indexpaar` (i,j) `mit` $i \neq 0$ `and` $a_{ij} \neq 0$ `nach einer Auswahlvorschrift.`
>
> **Schritt 2.2:** $G_m := G_{ij}$ `wobei` ϕ `nach (6.9) bestimmt wird.`
>
> **Schritt 2.3:** $A^{(m+1)} = G_m A^{(m)} {G_m}^T$.
>
> **Schritt 2.3:** $k := k + 1$
>
> **Until** `stop`

Ergebnis: `Approximierte Eigenwerte` $a_{ii}^{(m)}$, $i = 1, \ldots, n$ `auf der Hauptdiagonalen von` $A^{(m)}$.

Es gibt verschiedene Auswahlvorschriften, die man in Schritt 2.1 des Jakobi-Verfahrens verwenden kann. Im **klassischen Jakobi-Verfahren** wird die folgende Regel verwendet. (Dabei bezeichne wie üblich $a_{ij}^{(m)}$ das Matrixelement (i,j) in der Matrix $A^{(m)}$.)

Schritt 2.1: Bestimme ein Indexpaar (i,j) so dass $a_{ij}^{(m)} = \max_{k,l \in \{1,\ldots,n\}, k \neq l} a_{kl}^{(m)}$.

Das Verfahren sucht also das größte Element aus der Matrix heraus und transformiert es durch geschickte Wahl der Givens Rotation auf Null. (Genauer: $a_{ij}^{(m+1)} = 0$.) Weil wir mit symmetrischen Matrizen arbeiten, müssen wir zur Bestimmung des Indexpaares (i,j) nur eine obere Dreiecksmatrix durchsuchen. Obwohl bei jeder Transformation einmal erzeugte Nullen wieder verschwinden können, liegt Konvergenz vor.

Satz 6.29 *Das klassische Jakobi-Verfahren konvergiert linear in der Außennorm.*

Sei $A^{(m)}$ die von dem klassischen Jakobi-Verfahren in Schritt m erzeugte Matrix. Da das in Schritt 2.1 gewählte Indexpaar $|a_{ij}^{(m)}| = \max_{l \neq k} |a_{lk}^{(m)}|$ erfüllt, können wir die Außennorm $N(A^{(m)})$ durch

$$N(A^{(m)}) = \sum_{\substack{l,k=1 \\ l \neq k}}^{n} |a_{lk}^{(m)}|^2 \leq n(n-1)|a_{ij}^{(m)}|^2$$

abschätzen und erhalten daraus

$$a_{ij}^{(m)} \geq \frac{N(A^{(m)})}{n(n-1)}.$$

Jetzt betrachten wir die Außennorm der Matrix $A^{(m+1)}$ aus dem nächsten Iterationsschritt. Nach Satz 6.28 wissen wir, dass

$$N(A^{(m+1)} = N(A^{(m)}) - 2(a_{ij}^{(m)})^2 + 2(a_{ij}^{(m+1)})^2 = N(A^{(m)}) - 2(a_{ij}^{(m)})^2,$$

weil die Givens-Rotation nach (6.9) so gewählt wurde, dass $a_{ij}^{(m+1)} = 0$. Damit erhalten wir

$$N(A^{(m+1)}) = N(A^{(m)}) - 2|a_{ij}^{(m)}|^2 \leq \underbrace{\left(1 - \frac{2}{n(n-1)}\right)^1}_{=:q} N(A^{(m)}).$$

Es gilt $q < 1$ und $q < 0$ für $m \geq 3$, also konvergiert die Außennorm $N(A(m) \to 0$ für $m \to \infty$. Die beobachtete Konvergenz ist linear, da der Exponent von q gleich 1 ist. QED

Die Konvergenz der Außennormen beim Jakobi-Verfahren motiviert ein Abbruchkriterium, das aus dem Satz von Gerschgorin (Satz 6.13) gewonnen wird. Da die Summe der Nicht-Diagonalelemente während der Iteration immer kleiner werden, werden auch die Radien der Gershgorin-Kreise immer kleiner, so dass man diese als Abbruchkriterium verwenden kann.

Die Konvergenz der Einträge auf der Hauptdiagonalen untersucht das folgende Korollar noch etwas genauer.

Korollar 6.30 *Sind $\lambda_1 \geq \cdots \geq \lambda_n$ die Eigenwerte der symmetrischen Matrix $A \in \mathbb{R}^{n \times n}$ und ist $\tilde{a}_{11}^{(m)} \geq \cdots \geq \tilde{a}_{nn}^{(m)}$ eine Umsortierung der Diagonalelemente von $A^{(m)}$, so gilt*

$$|\lambda_i - \tilde{a}_{ii}^{(m)}| \leq \sqrt{N(A^{(m)})} \to 0 \text{ für } m \to \infty.$$

Beweis: Aus Korollar 6.10 mit $A = A^{(m)}$ und $B = \mathrm{diag}(a_{11}^{(m)}, \ldots, a_{nn}^{(m)})$ sowie der euklidischen Norm erhalten wir, da A und $A^{(m)}$ die gleichen Eigenwerte besitzen:

$$|\lambda_i - \tilde{a}_{ii}^{(m)}| = |\lambda_i(A_m) - \lambda_i(B)| \leq \|A^{(m)} - B\|_2 \leq \|A^{(m)} - B\|_{\mathrm{F}} = \sqrt{N(A^{(m)}}.$$

QED

Auf die Eigenvektoren können wir schließen, da sich $A^{(m)}$ schreiben lässt als

$$A^{(m+1)} = G_m A^{(m)} {G_m}^T = \cdots = G_m \cdot \ldots \cdot G_1 \cdot A \cdot {G_1}^T \cdot \ldots \cdot {G_m}^T =: Q_m A {Q_m}^T,$$

wobei $Q^{(m)}$ orthogonal ist und $A^{(m+1)}$ näherungsweise diagonal. Also bestehen die Zeilen von $Q^{(m)}$ näherungsweise aus Eigenvektoren von A.

Es gibt viele Modifikationen des Jakobi-Verfahrens, insbesondere in Bezug auf die in in Schritt 2.1 benötigte Auswahlregel. Der Grund liegt darin, dass die Auswahlregel des klassischen Jakobi-Verfahrens mit einem Aufwand von $\mathcal{O}(n^2)$ schon recht aufwändig ist. Für die beiden folgenden Modifikationen kann Konvergenz nachgewiesen werden.

- *Zyklisches Jakobi-Verfahren:* Hier wird keine aufwändige Auswahlregel berechnet sondern die Reihenfolge der Indizes vorher festgelegt, z.B. durch $(1,2),(1,2),\ldots,(1,2),(2,3)$ $1,n)$.

- *Zyklischen Jakobi-Verfahren mit Schwellenwert:* Die Idee entspricht zunächst dem zyklischen Jakobi-Verfahren in dem eine feste Reihenfolge der Indizes festgelegt ist. Allerdings überspringt man ein Element (i,j), wenn $\left|a_{ij}^{(k)}| < \gamma\right|$, da die Iteration in diesem Fall keine oder eine sehr geringe Verbesserung der Außennorm bringt. Der *Schwellenwert* γ wird im Laufe des Verfahrens reduziert. Dabei prüft man, ob alle Elemente der aktuellen Matrix unterhalb des Schwellenwerts liegen, und halbiert γ, sobald das der Fall ist.