Numerik I

Vorlesung von

Prof. Dr. Harald Garcke

im Wintersemester 2012/13

Überarbeitung und Textsatz in LyX von

Andreas Völklein

Stand: 19. April 2013

ACHTUNG

Diese Mitschrift ersetzt *nicht* die Vorlesung.
Es wird daher *dringend* empfohlen, die Vorlesung zu besuchen.

Copyright Notice

Copyright © 2012-2013 Andreas Völklein
Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation;
with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.
A copy of the license is included in the section entitled "GNU Free Documentation License".

Disclaimer of Warranty

Unless otherwise mutually agreed to by the parties in writing and to the extent not prohibited by applicable law, **the Copyright Holders and any other party, who may distribute the Document as permitted above, provide the Document "as is", without warranty of any kind**, expressed, implied, statutory or otherwise, including, but not limited to, the implied warranties of merchantability, fitness for a particular purpose, non-infringement, the absence of latent or other defects, accuracy, or the absence of errors, whether or not discoverable.

Limitation of Liability

In no event unless required by applicable law or agreed to in writing **will the Copyright Holders, or any other party, who may distribute the Document as permitted above, be liable to you for any damages**, including, but not limited to, any general, special, incidental, consequential, punitive or exemplary damages, however caused, regardless of the theory of liability, arising out of or related to this license or any use of or inability to use the Document, even if they have been advised of the possibility of such damages.
In no event will the Copyright Holders'/Distributor's liability to you, whether in contract, tort (including negligence), or otherwise, **exceed the amount you paid the Copyright Holders/Distributor** for the document under this agreement.

Links

Der Text der „GNU Free Documentation License" kann auch auf der Seite

`https://www.gnu.org/licenses/fdl-1.3.de.html`

nachgelesen werden.
Eine transparente Kopie der aktuellen Version dieses Dokuments kann von

`https://github.com/andiv/NumerikI`

heruntergeladen werden.

Literatur

Differentialtopologie:

- Dahmen, Reusken: *Numerik für Ingenieure und Naturwissenschaftler*; Springer; doi: 10.1007/978-3-540-76493-9
- Deuflhard, Hohmann: *Numerische Mathematik I, Eine algorithmisch orientierte Einführung*; de Gruyter; ISBN: 978-3-11-020354-7
- Freund, Hoppe: *Stoer/Bulirsch: Numerische Mathematik 1*; Springer; doi: 10.1007/978-3-540-45390-1
- Gene Golub, James Ortega: *Scientific Computing*; 1993, Academic Press ISBN: 0-12-289253-4
- Hämmerlin, Hoffmann: *Numerische Mathematik*; 1994, Springer ISBN: 3-540-58033-6
- Press, Teukolsky, Vertterling, Flannery: *Numerical Recipes in C++*; 2002, Cambridge University Press ISBN: 0-521-75033-4
- Stoer, Bulirsch: *Numerische Mathematik 2*; Springer doi: 10.1007/b137272

Inhaltsverzeichnis

1 Einführung

1.1 Aufgaben der numerischen Mathematik

- Entwicklung und Analyse von Rechenmethoden zur angenäherten Lösung von Problemen innerhalb der Mathematik und in den Anwendungen
- Rechenmethoden $\rightsquigarrow$ Algorithmen $\rightsquigarrow$ Programm

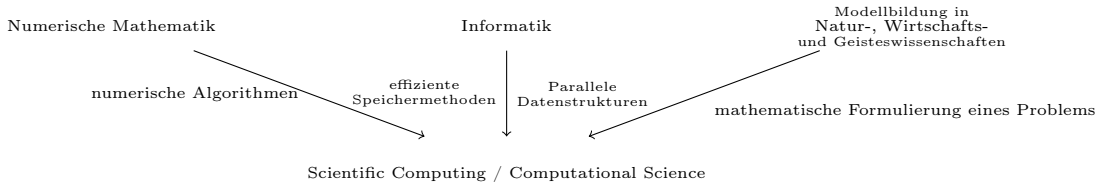


Abbildung 1.1: Numerik im Zusammenspiel verschiedener Disziplinen

Betrachten wir ein klassisches Problem und studieren die obigen Aspekte.

1. Phänomen in der Natur
   Beispiel: Bewegung von Himmelskörpern (z.B. Dreikörperproblem: Erde, Sonne, Mond oder allgemeinere Mehrkörperprobleme)
2. Mathematisches Modell
   Sei $x_i(t) \in \mathbb{R}^3$ die Position von Himmelskörper i ($i \in \{1,2,3\}$) zur Zeit t (hier drei Körper).
   Newtonsche Bewegungsgleichungen:

$$m_i \cdot \ddot{x}_i = -\nabla_{x_i} U(x_1(t), x_2(t), x_3(t))$$

   m_i: Masse des Körpers
   $\ddot{x}_i$: zweite zeitliche Ableitung von x_i
   U: potentielle Energie

$$U(x_1, x_2, x_3) = -G \sum_{\substack{i \neq j \\ i,j=1}}^{3} \frac{m_i m_j}{|x_i - x_j|}$$

   G: Gravitationskonstante
3. Mathematische Theorie
   Kennt man zum Anfangszeitpunkt Position und Geschwindigkeit der Himmelskörper, dann hat die Differentialgleichung aus 2. genau eine (lokale) Lösung nach dem Satz von Picard-Lindelöf.

4. Numerische Mathematik
   Ersetze das (in t) kontinuierliche Problem aus 2. durch ein diskretes Problem. Statt x_i zu berechnen, berechnen wir eine Funktion

$$x_i^h : \underbrace{h \cdot \mathbb{N}_0}_{=\{0,h,2h,\ldots\}} \to \mathbb{R}^3$$

   für kleines $h \in \mathbb{R}_{>0}$, die nur zu den Zeiten h_n ($n \in \mathbb{N}$) definiert ist und die Differentialgleichung näherungsweise erfüllt.
   z. B.: Berechne $x_i^h(nh)$, sodass für $i \in \{1,2,3\}$ gilt:

$$m_i \frac{x_i^h((n+1)h) - 2x_i^h(nh) + x_i^h((n-1)h)}{h^2} = -\nabla_{x_i} U\left(\left(x_1^h, x_2^h, x_3^h\right)(nh)\right)$$

   Verwende dabei die Näherung:

$$\ddot{x}_i(nh) \approx \frac{1}{h}\left(\frac{x_i^h((n+1)h) - x_i^h(nh)}{h} - \frac{x_i^h(nh) - x_i^h((n-1)h)}{h}\right) = \\ = \frac{x_i^h((n+1)h) - 2x_i^h(nh) + x_i^h((n-1)h)}{h^2}$$

   Frage: Konvergiert $x_i^h \to x_i$?
5. Numerische Berechnung
   Entwickle Algorithmus zur Berechnung von x_i^h. $\rightsquigarrow$ Computerprogramm $\rightsquigarrow$ Ergebnisse
   (Oben ist dies einfach.)
6. Analyse der Ergebnisse
   Vergleichen der Computerergebnisse mit Messungen und mathematischen Resultaten.
   Falls Abweichungen auftreten, stellen sich die folgenden Fragen:
   a) Stimmt das mathematische Modell?
   b) Ist meine numerische Methode gut? (Nähert 4. die Situation 2. gut an?)
   c) Wie groß ist der Einfluss von Mess- und Rundungsfehlern?

1.2 Mit welchen Fragestellungen wollen wir uns beschäftigen?

1. Lösen linearer Gleichungssysteme
   Seien A eine $(n \times n)$-Matrix (Notation: $A \in \mathbb{R}^{n\times n}$) und $b \in \mathbb{R}^n$.
   Gesucht ist eine Lösung $x \in \mathbb{R}^n$ mit $Ax = b$.
   Beispiel: Wettervorhersage/Strömungen mit mehr als 10^6 Unbekannten.
   a) *Cramersche Regel:*

$$x_j = \frac{\det(A)_j}{\det A}$$

   Die Matrix $(A)_j$ ergibt sich folgendermaßen: Ersetze die j-te Spalte von A durch b.
   Die Determinanten ergeben lassen sich mit den Permutationen der symmetrischen Gruppe $\mathfrak{S}(n)$ berechnen:

$$\det A = \sum_{\pi \in \mathfrak{S}(n)} \operatorname{sign}(\pi)\, a_{1,\pi(1)} \cdot \ldots \cdot a_{n,\pi(n)}$$

Diese Berechnung einer Determinante benötigt etwa $n \cdot n!$ Multiplikationen und Additionen. Nun sind aber die n Determinanten $\det(A)_j$ und die Determinante $\det(A)$ zu berechnen. Insgesamt benötigt man also $n \cdot (n+1)!$ Operationen.
Bei einer (20×20)-Matrix (kleine Matrix!) sind das in etwa 10^{21} Operationen. Falls jede arithmetische Operation 10^{-6} Sekunden braucht, benötigen wir Rechenzeiten von 10^{15} s, also in etwa 30 Millionen Jahre! Ein Supercomputer mit $3 \cdot 10^{15}$ Flops bräuchte fast vier Tage.

b) *Gaußsches Eliminationsverfahren:* Benötigt etwa n^3 Operationen, bei $n = 20$ sind das ungefähr 8000 Operationen und die Rechenzeit beträgt weniger als 0,005 Sekunden!

2. Lösen nichtlinearer Gleichungen
Seien $a_i \in \mathbb{R}$ und $p(x) = \sum_{i=0}^{n} a_i x^i$. Gesucht ist ein $x \in \mathbb{R}$ mit $p(x) = 0$.
Beispiel: $p(x) = x^2 - 5$
Gesucht: Verfahren zur Berechnung von $\sqrt{5}$
Starte mit $x_0 > 0$ beliebig. Berechne iterativ für $i \in \mathbb{N}_0$:

$$x_{i+1} = \frac{1}{2}\left(x_i + \frac{5}{x_i}\right)$$

Behauptung: $x_i \xrightarrow{i \to \infty} \sqrt{5}$

3. Approximation
Beispiel: lineare Ausgleichsprobleme
Zusammenhang zwischen Spannung und Stromfluss
U: Spannung, I: Stromstärke, R: Widerstand (unbekannt), Ohmsches Gesetz: $U = RI$
Messungen: (U_i, I_i) für $i \in \{1, \ldots, N\}$
Vermutung: Es gibt Messfehler.
TODO: Abb1 einfügen
Finde R so, dass

$$f(z) = \sum_{i=1}^{N} (U_i - zI_i)^2$$

bei R minimal wird, also $f(R) = \min_{z \in \mathbb{R}} f(z)$.

4. Interpolation
Ein Flugzeug soll durch die Punkte (x_i, y_i) fliegen. Der Flug soll möglichst glatt verlaufen.
TODO: Abb2 einfügen
Gesucht: Eine Kurve (möglichst glatt), die durch die Punkte $(x_i, y_i)_{i \in \{1,\ldots,N\}}$ geht. Die Kurve soll einfach zu berechnen sein.

5. Flächenberechnung/Integralberechnung
TODO: Abb3 einfügen
Berechne $\int_a^b f(x)\,\mathrm{d}x$. Aus Erfahrung wissen wir: Es ist nicht immer möglich Integrale exakt auszurechnen. Benötige ein Näherungsverfahren.

6. Optimierungsverfahren (Numerik II)
Eine bayerische Brauerei produziert Pils und Weißbier.
Zur Produktion benötigen wir:

	Pils	Weißbier
Malz	2 Einheiten	4 Einheiten
Hopfen	5 Einheiten	2 Einheiten
Hefe	5 Einheiten	4 Einheiten

Zur Verfügung stehen: 120 Einheiten Malz, 150 Einheiten Hopfen und 130 Einheiten Hefe
Ziel: Maximiere Gewinn. Gewinn ist für beide Biere gleich.
x_1: Liter Pils, x_2: Liter Weißbier; Maximiere $x_1 + x_2$
Aufgabe: Maximiere $f(x_1, x_2) = x_1 + x_2$ unter den Nebenbedingungen:

$$\begin{aligned} 2x_1 + 4x_2 &\leq 120 & x_1 &\geq 0 \\ 5x_1 + 2x_2 &\leq 150 & x_2 &\geq 0 \\ 5x_1 + 4x_2 &\leq 120 \end{aligned}$$

7. Verhalten bei Störungen, Stabilität des Verfahrens
(Eingabefehler, Rundungsfehler, Diskretisierungsfehler)
Beispiele:

   i) $\frac{1}{10^{-8}} = 10^8$; Störung des Nenners um $10^{-8} \rightsquigarrow \frac{1}{2 \cdot 10^{-8}} = 5 \cdot 10^7$

   ii) $x^2 + 314x - 2 = 0$ (allgemein: $ax^2 + bx + c = 0$);
   Löse Gleichung mit p-q-Formel (Mitternachtsformel). Gerechnet wird auf 5 signifikante Stellen genau, wobei sich ein relativer Fehler $= \frac{\text{Fehler}}{\text{Größe der Lösung}}$ von 57% ergibt. Eine einfache aber geschickte Umformulierung ergibt einen relativen Fehler von nur circa $1,5 \cdot 10^{-5}$. Diese ist wie folgt:

$$\begin{aligned} x_1 &= \frac{1}{2a}\left(-b - \operatorname{sign}(b)\sqrt{b^2 - 4ac}\right) \\ x_2 &= \frac{c}{ax_1} = \frac{2c}{-b - \operatorname{sign}(b)\sqrt{b^2 - 4ac}} \end{aligned}$$

   Exakte Lösung: 0,0063693
   Mit Rundung: Mitternachtsformel: $x_2 \approx 0,01$; Umformulierung: $x_2 \approx$ 0,0063692

2 Gaußsches Eliminationsverfahren

Sei $A \in \mathbb{R}^{n\times n}$ eine reelle $n \times n$-Matrix. Verwende die Schreibweise $A = (a_{ij})_{i,j\in\{1,\dots,n\}}$ mit den Einträgen a_{ij}.

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix}$$

Es sei $b \in \mathbb{R}^n$. Gesucht ist $x \in \mathbb{R}^n$ mit $Ax = b$.

2.1 Existenz und Eindeutigkeit von Lösungen

Existenz: Eine Lösung $x \in \mathbb{R}^n$ existiert genau dann, wenn $\operatorname{rg}(A) = \operatorname{rg}(A, b)$ gilt, das heißt b eine Linearkombination der Spalten von A ist.
Sei $\operatorname{Rg}(A) = m < n$. Dann gilt:

1. $\dim(\ker(A)) = n - m \neq 0$
2. $Ax = b$ hat entweder keine Lösung oder es existiert eine Lösung x und für alle $y \in \ker(A)$ ist auch $x + y$ eine Lösung von $Ax + b$.

Gilt andererseits $\operatorname{rg}(A) = n$, so ist A surjektiv und wegen $A \in \mathbb{R}^{n\times n}$ gilt äquivalent die Injektivität von A, das heißt $\det(A) \neq 0$, und es gibt genau eine Lösung von $Ax = b$. Eine solche Matrix heißt *regulär* oder nicht *singulär*.

Im Folgenden sei A nicht singulär.

2.2 Gleichungssysteme mit Dreiecksmatrizen

Im Falle von Dreiecksmatrizen

$$A = \begin{pmatrix} a_{11} & \dots & \dots & a_{1n} \\ 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & a_{nn} \end{pmatrix}$$

ergibt sich ein gestaffeltes Gleichungssystem.

$$Ax = b \quad \Leftrightarrow \quad \begin{aligned} a_{11}x_1 + \dots + a_{nn}x_n &= b_1 \\ \vdots \qquad\qquad & \quad \vdots \\ a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n &= b_{n-1} \\ a_{nn}x_n &= b_n \end{aligned}$$

Auflösung	Multiplikation/Division	Addition/Subtraktion
$x_n = \frac{b_n}{a_n}$	1	0
$x_{n-1} = \frac{(b_{n-1} - a_{n-1,n} x_n)}{a_{n-1,n-1}}$	2	1
$\vdots$	$\vdots$	$\vdots$
$x_n = \frac{(b_1 - a_{12} x_2 - \ldots - a_{1n} x_n)}{a_{11}}$	n	$n-1$

Aufwand in flops (floating point Operations, Gleitkommaoperationen):

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

Sei nun $A = LR$ mit einer unteren Dreiecksmatrix L und einer oberen Dreiecksmatrix R. Löse $Ax = b$ in zwei Schritten:

1. Bestimme y mit $Ly = b$.
2. Bestimme x mit $Rx = y$.

Ziel: Bringe jede Matrix A auf die Form $A = LR$.

Ist das möglich? Nahezu ja.

2.3 Gaußsches Eliminationsverfahren (ohne Pivotisierung)

Zunächst ein Beispiel:

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \qquad\qquad \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} x = b$$

$$\begin{aligned} x_1 + 2x_2 &= b_1 \\ 3x_1 + 4x_2 &= b_2 \end{aligned}$$

Multipliziere erste Gleichung mit (-3) und addiere das Ergebnis zur zweiten Zeile:

$$\begin{aligned} x_1 + 2x_2 &= b_1 \\ -2x_2 &= b_2 - 3b_1 \end{aligned}$$

Das Gleichungssystem hat nun Dreiecksgestalt.

In Matrixform:

$$\begin{pmatrix} 1 & 0 \\ -3 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} x = \begin{pmatrix} 1 & 0 \\ -3 & 1 \end{pmatrix} b$$

$$\begin{pmatrix} 1 & 2 \\ 0 & -2 \end{pmatrix} x = \begin{pmatrix} 1 & 0 \\ -3 & 1 \end{pmatrix} b$$

Nun gilt:

$$\begin{pmatrix} 1 & 0 \\ -3 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\Leftrightarrow \qquad \begin{pmatrix} 1 & 0 \\ -3 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix}$$

Also gilt:

$$\begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 0 & -2 \end{pmatrix} x = b$$

Allgemein: Transformiere beliebiges invertierbares $A \in \mathbb{R}^{n \times n}$ auf Dreiecksgestalt.

- 1. Schritt: (Neue Zeile i) = (Alte Zeile i) $- l_{i1} \cdot$ (Alte Zeile 1) für $i \in \{2, \ldots, n\}$
  Dabei sei $l_{i1} = \frac{a_{i1}}{a_{11}}$ $(i \in \{2, \ldots, n\})$. Voraussetzung: $a_{11} \neq 0$

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

- Iteriere diesen Prozess bis man die Dreiecksgestalt erhält. Im k-ten Schritt ($a_{ij}^{(1)} := a_{ij}$):

$$\begin{aligned} l_{ij} &= a_{ik}^{(k)} && i \in \{k+1, \ldots, n\} \\ a_{ij}^{(k+1)} &= a_{ij}^{(k)} - l_{ik}a_{kj}^{(k)} && i, j \in \{k+1, \ldots, n\} \\ b_i^{(k+1)} &= b_i^{(k)} - l_{ik}b_k^{(k)} && i \in \{k+1, \ldots, n\} \end{aligned}$$

  Stelle dies in Matrixform dar:

$$A^{(k)} := \begin{pmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \ldots & \ldots & \ldots & a_{1n}^{(k)} \\ 0 & \ddots & & & & \vdots \\ \vdots & \ddots & \ddots & & & \vdots \\ \vdots & & 0 & a_{kk}^{(k)} & \ldots & a_{kn}^{(k)} \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & a_{nk}^{(k)} & \ldots & a_{nn}^{(k)} \end{pmatrix}$$

$$A = A^{(1)} \to A^{(2)} \to \ldots \to A^{(n)} =: R$$

  Die Operationen von $A^{(k-1)} \to A^{(k)}$ drücken wir durch Frobenius-Matrizen aus.
  Eine Frobenius-Matrix hat (nach Definition) die Gestalt:

$$L_k := \mathbb{1} - l^k e_k^{\mathrm{T}} = \begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & \ddots & & & & \\ & & & 1 & & & \\ & & & -l_{k+1,k} & \ddots & & \\ & & & \vdots & & \ddots & \\ 0 & & & -l_{nk} & 0 & & 1 \end{pmatrix}$$

e_k ist der k-te Basisvektor und e_k^{T} der transponierte Vektor.

$$l^k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ l_{k+1,k} \\ \vdots \\ l_{nk} \end{pmatrix}$$

2.4 Lemma (LR-Zerlegung)

1. $$A^{(k+1)} = L_k \cdot A^{(k)}$$
2. $$b^{(k-1)} = L_k \cdot b^{(k)}$$
3. $$R := A^{(n)} = L_{n-1} \cdot \ldots \cdot L_1 \cdot A$$
4. $$L_k^{-1} = \mathbb{1} + l^k \cdot e_k^{\mathrm{T}}$$
5. $$L := (L_{n-1} \cdot \cdots \cdot L_1)^{-1} = \mathbb{1} + \sum_{k=1}^{n-1} l^k \cdot e_k^{\mathrm{T}} = \begin{pmatrix} 1 & & & 0 \\ l_{2,1} & \ddots & & \\ \vdots & \ddots & \ddots & \\ l_{n1} & \ldots & l_{n,n-1} & 1 \end{pmatrix}$$
6. R ist eine obere Dreiecksmatrix und es gilt:
$$A = L \cdot R$$

Beweis

1. Für $i, j \leq k$ ist $A_{ij}^{(k+1)} = A_{ij}^{(k)}$ und sonst $a_{ij}^{(k+1)} = a_{ij}^{(k)} - l_{ik} a_{kj}^{(k)}$. Es gilt:

$$L_k \cdot A^{(k)} = \begin{pmatrix} 1 & & & & & & 0 \\ & \ddots & & & & & \\ & & \ddots & & & & \\ & & & 1 & & & \\ & & & -l_{k+1,k} & \ddots & & \\ & & & \vdots & & \ddots & \\ 0 & & & -l_{nk} & 0 & & 1 \end{pmatrix} A^{(k)} =$$

$$= \begin{pmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \dots & \dots & \dots & a_{1n}^{(k)} \\ 0 & \ddots & & & & \vdots \\ \vdots & \ddots & \ddots & & & \vdots \\ \vdots & & 0 & a_{kk}^{(k)} & \dots & a_{kn}^{(k)} \\ \vdots & & \vdots & a_{k+1,k}^{(k)} - l_{k+1,k} a_{kk}^{(k)} & & a_{k+1,n}^{(k)} - l_{k+1,k} a_{kn}^{(k)} \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & a_{nk}^{(k)} - l_{nk} a_{kk}^{(k)} & \dots & a_{nn}^{(k)} - l_{nk} a_{kn}^{(k)} \end{pmatrix} = A^{(k+1)}$$

2. Analog zu 1. folgt dies.
3. Induktion ergibt:

$$A^{(n)} = L_{n-1} A^{(n-1)} = L_{n-1} \cdot L_{n-2} \cdot A^{(n-2)} = \dots = L_{n-1} \cdot \dots \cdot L_1 \cdot A$$

4. Es gilt:

$$L_k = \mathbb{1} - l^k \cdot e_k^{\mathrm{T}}$$
$$\left(\mathbb{1} + l^k \cdot e_k^{\mathrm{T}}\right) \cdot \left(\mathbb{1} - l^k \cdot e_k^{\mathrm{T}}\right) = \mathbb{1} - l^k \cdot e_k^{\mathrm{T}} + l^k e_k^{\mathrm{T}} - \underbrace{l^k \cdot e_k^{\mathrm{T}} \cdot l^k \cdot e_k^{\mathrm{T}}}_{=0} = \mathbb{1}$$
$$\Rightarrow \qquad \left(\mathbb{1} - l^k \cdot e_k^{\mathrm{T}}\right)^{-1} = \mathbb{1} + l^k \cdot e_k^{\mathrm{T}}$$

5. Damit folgt:

$$(L_{n-1} \cdot \dots \cdot L_1)^{-1} = L_1^{-1} \cdot \dots \cdot L_{n-1}^{-1} = \left(\mathbb{1} + l^1 \cdot e_1^{\mathrm{T}}\right) \cdot \dots \cdot \left(\mathbb{1} + l^{n-1} \cdot e_{n-1}^{\mathrm{T}}\right)$$

Behauptung:

$$\prod_{i=1}^{n-1} \left(\mathbb{1} + l^i \cdot e_i^{\mathrm{T}}\right) = \mathbb{1} + \sum_{i=1}^{n-1} l^i e_i^{\mathrm{T}}$$

Beweis: Induktionsschritt $k \rightsquigarrow k+1$:

$$\prod_{i=1}^{k+1} \left(\mathbb{1} + l^i \cdot e_i^{\mathrm{T}}\right) = \left(\mathbb{1} + l^{k+1} \cdot e_{k+1}^{\mathrm{T}}\right) \left(\mathbb{1} + \sum_{i=1}^{k} l^k e_k^{\mathrm{T}}\right) =$$
$$= \mathbb{1} + \sum_{i=1}^{k} l^k e_k^{\mathrm{T}} + l^{k+1} \cdot e_{k+1}^{\mathrm{T}} + l^{k+1} \cdot \underbrace{e_{k+1}^{\mathrm{T}} \sum_{i=1}^{k} l^k e_k^{\mathrm{T}}}_{=0} =$$
$$= \mathbb{1} + \sum_{i=1}^{k+1} l^k e_k^{\mathrm{T}}$$

$\square_{\text{Behauptung}}$

6. Dies gilt nach Konstruktion von R und L.

$\square_{2.4}$

2.5 Algorithmus zum Eliminationsverfahren

Vorgehen beim Lösen des Gleichungssystems $A \cdot x = b$:

1. Bilde die LR-Zerlegung von A.
2. Löse das Gleichungssystem $L \cdot z = b$.
3. Löse das Gleichungssystem $R \cdot x = z$.

Vorteil: Hat man die LR-Zerlegung, so kann man das Gleichungssystem für $A \cdot x = b$ für viele rechte Seiten lösen. Falls wir A nicht mehr brauchen, speichert man die Matrizen L und R auf der alten Matrix A, um Speicherplatz zu sparen.

$$A \mapsto \begin{pmatrix} r_{1,1} & \cdots & \cdots & r_{1,n} \\ l_{2,1} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ l_{n,1} & \cdots & l_{n,n-1} & r_{n,n} \end{pmatrix}$$

2.6 Landau-Symbole

Es seien $f, g : D \to \mathbb{R}^m$, $D \subseteq \mathbb{R}^n$, $n, m \in \mathbb{N}$.

i) Wir sagen $f = \mathcal{O}_{x \to x_0}(g)$, falls gilt:

$$\lim_{\substack{x \to x_0 \\ x \in D \setminus \{x_0\}}} \sup \frac{\|f(x)\|}{\|g(x)\|} < \infty$$

(d. h. es gibt ein $K > 0$ mit $\|f(x)\| \leq K \|g(x)\|$ für x nahe x_0.)

ii) Wir sagen $f = o_{x \to x_0}(g)$, falls gilt:

$$\lim_{\substack{x \to x_0 \\ x \in D \setminus \{x_0\}}} \frac{\|f(x)\|}{\|g(x)\|} = 0$$

Bemerkung: $x_0 = \pm\infty$ ist zugelassen.

Bemerkung

Hängt g nur von einer Variablen ab, gibt es also keine Parameter, so schreibe zur Einfachheit $\mathcal{O}_{x_0}(g)$ statt $\mathcal{O}_{x \to x_0}(g)$. Wenn zudem klar ist, welche x_0 gemeint ist, schreibt man häufig nur $\mathcal{O}(g)$. Analoges gilt für $o(g)$.

2.7 Aufwand des Gaußschen Eliminationsverfahrens

Zähle Gleitkommaoperationen ($+$, $\cdot$, $-$, $:$):

Es ergibt sich:

$$\sum_{k=1}^{n-1} (n-k)(1 + (n-k) \cdot 2) = \sum_{l=1}^{n-1} l(1 + 2l) = \sum_{l=1}^{n-1} \left(l + 2l^2\right) =$$
$$= \frac{n(n-1)}{2} + 2\frac{n(n-1)(2n-1)}{6} = \frac{2}{3}n^3 + \mathcal{O}_\infty\left(n^2\right)$$

Für $n = 20$ brauchen wir weniger als 6000 Operationen, im Gegensatz zu 10^{21} für die Cramersche Regel!

2.8 LR-Zerlegung nicht immer möglich/sinnvoll

1. Beispiel:

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

Hier ist $a_{11} = 0$ und die LR-Zerlegung deshalb nicht anwendbar. Als Ausweg kann man jedoch die Zeilen vertauschen.

2. Eine LR-Zerlegung kann große Fehler generieren.

$$A = \begin{pmatrix} 10^{-m} & 1 \\ 1 & 1 \end{pmatrix} \qquad b = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

Löse $Ax = b$.

$$A^{(2)} = \begin{pmatrix} 10^{-m} & 1 \\ 0 & 1 - 10^m \end{pmatrix} \qquad b^{(2)} = \begin{pmatrix} 1 \\ 2 - 10^m \end{pmatrix}$$

$1 - 10^m$: zur Darstellung brauche $m - 1$-Stellen: 9999999...999
Hat der Rechner nur $n < m-1$ Stellen zur Verfügung, wird abgeschnitten (oder gerundet) (vergleiche Kapitel 3):

$$\tilde{A}^{(2)} = \begin{pmatrix} 10^{-m} & 1 \\ 0 & -10^m \end{pmatrix} \qquad b^{(2)} = \begin{pmatrix} 1 \\ -10^m \end{pmatrix}$$

Die Lösung auf dem Rechner ist daher:

$$\tilde{x}_2 = 1 \qquad \tilde{x}_1 = 0$$

Die tatsächliche Lösung hingegen ist:

$$(1 - 10^m)\, x_2 = 2 - 10^m \qquad \Rightarrow \quad x_2 = \frac{2 - 10^m}{1 - 10^m} \approx 1$$

$$x_1 = \frac{10^m}{10^m - 1} \approx 1$$

Die schlechte Lösung kommt durch Rundungsfehler zustande. Die Vertauschung der Zeilen rettet uns!

$$\begin{pmatrix} 1 & 1 \\ 10^{-m} & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

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

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \approx \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

$$\Rightarrow \quad x_1 \approx 1$$

$$x_2 \approx 1$$

2.9 Gauß-Elimination mit Pivotisierung

Pivotstrategie zur Elimination in der k-ten Spalte:

1. Wähle $p \in \{k, \dots, n\}$, sodass gilt:

$$\left|a_{pp}^{(k)}\right| = \max_{j \in \{k,\dots,n\}} \left|a_{jk}^{(k)}\right|$$

2. Definiere

$$\tilde{A}^{(k)} = \left(\tilde{a}^{(k)}\right)_{ij}$$

   mit für $j \in \{1, \dots, n\}$:

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

   Dies vertauscht die Zeilen k und p.
3. Eliminiere wie bisher $\tilde{A}^{(k)} \rightsquigarrow A^{(k+1)}$.

Ziel: Schreibe dies in Matrixform:

Der Austausch von zwei Zeilen kann mit Hilfe von Permutationen beschrieben werden. Sei

$$\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$$

eine Permutation in $\mathfrak{S}_n$. Definiere die Matrix:

$$P_\sigma = \left(e_{\sigma(1)} \dots e_{\sigma(n)}\right) \in \mathbb{R}^{n \times n}$$

Es gilt:

$$\begin{aligned} & P_\sigma e_i = e_{\sigma(i)} \\ \Rightarrow \qquad & P_\sigma A = \begin{pmatrix} a_{\sigma^{-1}(1)} \\ \vdots \\ a_{\sigma^{-1}(n)} \end{pmatrix} \end{aligned}$$

Dabei ist $a_i = (a_{i1}, \dots, a_{in})$ die i-te Zeile von A. Damit lässt sich das Vorgehen 1. bis 3. wie folgt beschreiben:

$$R = A^{(n)} = L_{n-1} P_{n-1} \cdot \ldots \cdot L_2 P_2 L_1 P_1 A$$

Dabei vertauscht P_i zwei Zeilen mit Index $\geq i$.

2.10 Lemma über Permutationen

Sei $\sigma : \{1, \dots, n\} \to \{1, \dots, n\}$ eine Permutation. Dann ist P_σ orthogonal, das heißt es gilt $P_\sigma^{-1} = P_\sigma^{\mathrm{T}}$. Außerdem gilt:

$$P_\sigma^{-1} = P_{\sigma^{-1}}$$

Beweis

$$\left(P_{\sigma^{-1}} \circ P_\sigma\right)(e_i) = e_{\sigma^{-1}(\sigma(i))} = e_i$$

Also gilt $P_\sigma^{-1} = P_{\sigma^{-1}}$. Sei nun $x, y \in \mathbb{R}^n$.

$$x \cdot y = x^{\mathrm{T}} y = \sum_{i=1}^{n} x_i y_i$$

Um $P_\sigma^{\mathrm{T}} = P_\sigma^{-1}$ zu zeigen, zeige:

$$x \cdot \left(P_\sigma^{-1} y\right) = P_\sigma x \cdot y$$

$$e_i \left(P_\sigma^{-1} e_j\right) = e_i \cdot \left(P_{\sigma^{-1}} e_j\right) = e_i \cdot e_{\sigma^{-1}(j)} = \delta_{i,\sigma^{-1}(j)} = \overset{i=\sigma^{-1}(j) \Leftrightarrow \sigma(i)=j}{=} \delta_{\sigma(i),j} = e_{\sigma(i)} \cdot e_j = P_\sigma e_i \cdot e_j$$

$\square_{2.10}$

2.11 **Satz** (Existenz einer LR-Zerlegung mit Spalten-Pivotisierung)

Für jede invertierbare Matrix A existiert eine Permutationsmatrix P, sodass die LR-Zerlegung $PA = LR$ möglich ist. P kann so gewählt werden, dass gilt:

$$\max_{1 \le i,j \le n} |l_{ij}| \le 1$$

Beweis

1. Schritt: Da A invertierbar ist, folgt:

$$\left|a_{p1}^{(1)}\right| = \max_{1 \le i \le n} \underbrace{\left|a_{i1}^{(1)}\right|}_{=|a_{i1}|} > 0$$

Vertauschung liefert ein $\tilde{a}_{11}^{(1)} \neq 0$. Die Elimination ist also möglich und

$$A^{(2)} = L_1 P_1 A = \begin{pmatrix} a_{11}^{(2)} & * & \dots & * \\ 0 & & & \\ \vdots & & B^{(2)} & \\ 0 & & & \end{pmatrix}$$

ist als Produkt invertierbarer Matrizen invertierbar und somit auch $B^{(2)}$.
Dieses Verfahren kann iteriert werden. Damit ergibt sich:

$$R = L_{n-1} P_{n-1} L_{n-2} P_{n-2} \dots L_1 P_1 A$$

Dabei vertauscht P_i zwei Zeilen mit Index $\ge i$ und:

$$L_i = \mathbb{1} - l^i e_i^{\mathrm{T}} \qquad\qquad l_{ik} = \begin{cases} \frac{\tilde{a}_{ik}^{(k)}}{\tilde{a}_{kk}^{(k)}} & \text{für } i > k \\ 0 & \text{für } i \le k \end{cases}$$

Ziel: Vertausche sukzessive $L_i$und Permutationsmatrizen P_j.
Behauptung: Für $i > k-1$ gilt

$$P_i L_{k-1} = \hat{L}_{k-1} P_i$$

mit:

$$\hat{L}_{k-1} = \left(\mathbb{1} - \left(P_i l^{k-1}\right) e_{k-1}^{\mathrm{T}}\right)$$

(Das heißt, im Vektor l^i werden zwei Einträge gemäß Permutation P_i vertauscht.)

$$\begin{aligned} P_i L_{k-1} &= P_i \left(\mathbb{1} - l^{k-1} e_{k-1}^{\mathrm{T}}\right) P_i^{-1} P_i = \\ &= \left(\mathbb{1} - P_i l^{k-1} e_{k-1}^{\mathrm{T}} P_i^{-1}\right) P_i \end{aligned}$$

Es gilt:

$$e_{k-1}^{\mathrm{T}} P_i^{-1} = e_{k-1}^{\mathrm{T}} P_i^{\mathrm{T}} = (P_i e_{k-1})^{\mathrm{T}} = e_{k-1}^{\mathrm{T}}$$

Denn P_i vertauscht Zeilen mit Index $\geq i$ und es ist $i > k-1$. Es folgt:

$$P_i L_{k-1} = \left(\mathbb{1} - P_i l^{k-1} e_{k-1}^{\mathrm{T}}\right) P_i$$

Per Induktion folgt:

$$P_{n-1} \cdot \ldots \cdot P_k L_{k-1} = \left(\mathbb{1} - \left(P_{n-1} \cdot \ldots \cdot P_k\right) l^{k-1} e_{k-1}^{\mathrm{T}}\right) P_{n-1} \cdot \ldots \cdot P_k$$

Behauptung: Für $k \in \{n-1, n-2, \ldots, 1\}$ gilt:

$$L_{n-1} P_{n-1} \cdot \ldots \cdot L_k P_k = \prod_{j=k}^{n-1} \left(\mathbb{1} - \hat{l}^j e_j^{\mathrm{T}}\right) P_{n-1} \cdot \ldots \cdot P_k$$

$$\hat{l}^j := P_{n-1} \cdot \ldots \cdot P_{j+1} l^j$$

Beweis: Induktion: Induktionsanfang bei $k = n-1$ ist klar.
Induktionsschritt $k \rightsquigarrow k-1$:

$$\begin{aligned} &L_{n-1} P_{n-1} \cdot \ldots \cdot L_{k-1} P_{k-1} = \\ &\quad = \prod_{j=k}^{n-1} \left(\mathbb{1} - \hat{l}^j e_j^{\mathrm{T}}\right) P_{n-1} \cdot \ldots \cdot P_k L_{k-1} P_{k-1} = \\ &\quad = \prod_{j=k}^{n-1} \left(\mathbb{1} - \hat{l}^j e_j^{\mathrm{T}}\right) \left(\mathbb{1} - \left(P_{n-1} \cdot \ldots \cdot P_k\right) l^{k-1} e_{k-1}^{\mathrm{T}}\right) P_{n-1} \cdot \ldots \cdot P_k P_{k-1} = \\ &\quad = \prod_{j=k-1}^{n-1} \left(\mathbb{1} - \hat{l}^j e_j^{\mathrm{T}}\right) P_{n-1} \cdot \ldots \cdot P_{k-1} \end{aligned}$$

$\square_{\text{Behauptung}}$

Damit ergibt sich:

$$\begin{aligned} R &= \hat{L}PA \\ \hat{L} &:= \prod_{j=1}^{n-1} \left(\mathbb{1} - \hat{l}^j e_j^{\mathrm{T}} \right) = \mathbb{1} - \sum_{j=1}^{n-1} \hat{l}^j e_j^{\mathrm{T}} \\ L &:= \hat{L}^{-1} = \prod_{j=1}^{n-1} \left(\mathbb{1} + \hat{l}^j e_j^{\mathrm{T}} \right) \end{aligned}$$

Es gilt $PA = LR$ und die Pivotisierungsstrategie sorgt dafür, dass $|l_{ij}| \leq 1$ ist. $\square_{2.11}$

2.12 Bemerkungen

1. L ergibt sich aus Frobeniusmatrizen durch Vertauschen von Elementen l_{ik}. Die Vertauschung ist analog wie in $A^{(k)}$. Diese Tatsache wird im Algorithmus ausgenutzt.
2. Permutationsmatrizen P können beschrieben (und gespeichert) werden, indem man die Permutation der Indizes $\{1, \ldots, n\}$ angibt.
   Auf dem Rechner nutze einen Vektor $s = (s_1, \ldots, s_n)$ mit $s_i \in \{1, \ldots, n\}$. s_i gibt an, welche alte Zeile nach Vertauschung in Zeile i steht. (vergleiche PA)
3. Die Zerlegung $PA = LR$ kann genutzt werden, um $\det(A)$ zu berechnen. Es gilt:

$$\det(A) = \det\left(P^{-1}LR\right) = \underbrace{\det\left(P^{\mathrm{T}}\right)}_{=\det(P)} \cdot \underbrace{\det(L)}_{=1} \cdot \underbrace{\det(R)}_{=r_{11}\cdot\ldots\cdot r_{nn}}$$

   Falls keine Permutation nötig war gilt:

$$\det(A) = r_{11} \cdot \ldots \cdot r_{nn}$$

   Außerdem gilt:

$$\det(P_\sigma) = \operatorname{sign}(\sigma)$$

2.13 Der Gauß-Algorithmus mit Pivotisierung

Dies ist einen Algorithmus zur Bestimmung von P, L und R, sodass $PA = LR$ gilt.

```
input((a_ij)_{i,j∈{1,...,n}})
for (k = 1,...,n) // s_k = k
{    p = k
     for (i = k+1,...,n)
          {if |a_ik| > |a_pk| then p = i}
     if (p ≠ k) then
     { // Vertausche s_k und s_p
          for (j = 1,...,n)
               {Vertausche a_pj und a_kj} // Vertauscht auch die l_ij!
     }
     for (i = k+1,...,n)
```

```
{     a_ik = a_ik / a_kk
      for (j = k+1, ..., n)
            {a_ij = a_ij − a_ik · a_kj}
      }
}
output ((a_ij)_{i,j∈{1,...,n}}, (s_k)_{k∈{1,...,n}})
```

2.14 Eindeutigkeit der LR-Zerlegung

Zu einer nicht-singulären Matrix A gibt es höchstens eine Zerlegung $A = LR$ mit einer unteren Dreiecksmatrix L mit $l_{ii} = 1$ und einer unteren Dreiecksmatrix R.

Beweis

Sei $A = (a_{ij})_{ij}$, $L = (l_{ij})_{ij}$ und $R = (r_{ij})_{ij}$ mit:

$$A = LR$$

$$\Rightarrow \quad a_{ij} = \sum_{k=1}^{\min(i,j)} l_{ik} r_{kj}$$

Für $i \leq j$ gilt:

$$r_{ij} = a_{ij} - \sum_{k<i} l_{ik} r_{kj}$$

Für $i > j$ gilt:

$$a_{ij} = \sum_{k=1}^{j-1} l_{ik} r_{kj} + l_{ij} \underbrace{r_{jj}}_{\neq 0}$$

$$\Rightarrow \quad l_{ij} = \frac{1}{r_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik} r_{kj} \right)$$

Wir können l_{ij} und r_{ij} rekursiv berechnen. Für $i \in \{2, \ldots, n\}$ berechne abwechselnd

$$r_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik} r_{kj} \qquad\qquad j \in \{i, \ldots, n\}$$

und:

$$l_{ki} = \frac{1}{r_{ii}} \left(a_{ki} - \sum_{j=1}^{i-1} l_{kj} r_{ji} \right) \qquad\qquad k \in \{i, \ldots, n\}$$

$$\begin{pmatrix} r_{11} & \cdots & \cdots & r_{1n} \\ l_{21} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ l_{n1} & \cdots & l_{n,n-1} & r_{nn} \end{pmatrix} \xrightarrow[\text{Reihenfolge berechnen}]{\text{werden in dieser}} \begin{pmatrix} \rightarrow & 1 & \rightarrow & \\ \downarrow & \rightarrow & 3 & \rightarrow \\ 2 & \downarrow & & 5 \\ \downarrow & 4 & 6 & \\ & \downarrow & & \cdots \end{pmatrix}$$

Berechne in der Reihenfolge 1, 2, 3, ... Dies zeigt, dass sich die (l_{ij}) und (r_{ij}) eindeutig aus den (a_{ij}) berechnen lassen. Es folgt die Eindeutigkeit der LR-Zerlegung und ein Berechnungsverfahren. $\square_{2.14}$

3 Zahlendarstellung und Fehleranalyse

3.1 Ursachen von Fehlern

Es gibt verschiedene Fehlerursachen:

1. Fehler im mathematischen Modell
2. Messfehler
3. Rundungsfehler
4. Approximationsfehler

Die Punkte 1. und 2. behandeln wir hier nicht.
Zum 4. Punkt ein Beispiel: Ein Integral

$$\int_a^b f(x)\,\mathrm{d}x$$

ist eventuell nicht exakt berechenbar.

$$\int_a^b f(x)\,\mathrm{d}x \overset{\text{Fehler!}}{\approx} \sum_{i=1}^{N} f(x_i)\cdot h$$

$$x_i = a + i\cdot h = a + i\cdot\frac{b-a}{N}$$

In diesem Kapitel studieren wir den Einfluss von Fehlern auf die Qualität von numerischen Problemen. Zunächst behandeln wir die Frage, wie Fehler auf dem Rechner dargestellt werden?

3.2 Zahlendarstellung

Jede Zahl $x \neq 0$ die auf dem Rechner dargestellt wird hat die Form:

$$x = (-1)^s \cdot m \cdot d^e$$

Dabei bestimmt $s \in \{0,1\}$ das Vorzeichen. $e \in \mathbb{Z}$ ist der Exponent, m die Mantisse und $d \in \mathbb{N}_{\geq 2}$ die Basis (meist $d = 2$). Es gilt

$$e_{\min} \leq e \leq e_{\max}$$

und:

$$m = m_0 + \sum_{i=1}^{l} m_i\, d^{-i}$$

$$m_i \in \{0, 1, \ldots, d-1\}$$

$$m_0 \neq 0$$

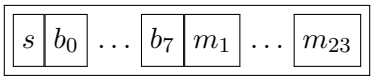
Hier sind $d,\ e_{\min},\ e_{\max}$ und l fest.

3.3 Beispiel (Maschinenzahlen mit einfacher/doppelter Genauigkeit)

Wir betrachten Zahlen mit je 32 Bit also 4 Byte Speicherplatz. Es sei $d = 2$ (Dualdarstellung), also $m_0 = 1$, $e_{\min} = -126$, $e_{\max} = 127$ und $l = 23$. Wie speichern wir $x = (-1)^s \cdot m \cdot d^e$? Definiere $b := e + 127 \in \{1, \ldots, 254\}$ und schreibe b als Dualzahl:

$$b = \sum_{i=0}^{7} b_i 2^i \qquad\qquad b_i \in \{0, 1\}$$

Die 32 Bit sind wie folgt belegt:



$m_0 = 1$ muss nicht abgespeichert werden. Die Null wird durch

$$b_1 = ... = b_7 = m_1 = ... = m_{23} = 0$$

dargestellt. Unendlich entspricht:

$$b_1 = ... = b_7 = 1$$
$$m_1 = ... = m_{23} = 0$$

Plus unendlich hat $s = 0$, minus unendlich $s = 1$. Die Folge $b_1 = ... = b_7 = 1$, $m_i \neq 0$ für mindestens ein i ergibt NaN (Not a Number). Letzteres ist zum Beispiel das Ergebnis von $\sqrt{-5}$. Bei doppelter Genauigkeit (8 Byte) gilt $l = 52$, $e_{\min} = -1022$, $e_{\max} = 1023$.

3.4 Einige wichtige Zahlen

Es gibt eine größte Zahl:

$$x_{\max} \in \mathbb{F}\left(d, e_{\min}, e_{\max}, l\right)$$

Sie ist:

$$x = (d-1) \cdot \left(1 + \sum_{i=1}^{l} d^{-i}\right) \cdot d^{e_{\max}} = (d-1)\,\frac{1 - d^{-(l+1)}}{1 - d^{-1}} d^{e_{\max}} =$$
$$= \left(1 - d^{-(l+1)}\right) \cdot d^{e_{\max}+1} = d^{e_{\max}+1} - d^{e_{\max}-l}$$

Bei Float wäre das ungefähr $3,4 \cdot 10^{38}$. Heute werden auch nicht normalisierte Zahlen benutzt. Beispiel: Float (einfache Genauigkeit in C). Falls $b_0 = ... = b_7 = 0$, so entspricht dies der Zahl

3.5 Relative und absolute Fehler

1. Seien nun $x, \tilde{x} \in \mathbb{R}^n$ und $x \neq 0$, so definieren wir:

$$\varrho_{\tilde{x}}\left(x\right) := \frac{\|\tilde{x} - x\|}{\|x\|}$$
$$\alpha_{\tilde{x}}\left(x\right) := \|\tilde{x} - x\|$$

2. Sei $x \in \mathbb{R}$ und sei $\tilde{x} \in \mathbb{R}$ eine Näherung von x. Dann ist

$$\alpha_{\tilde{x}}(x) := |\tilde{x} - x|$$

der *absolute* und für $x \neq 0$

$$\varrho_{\tilde{x}}(x) := \frac{|\tilde{x} - x|}{|x|}$$

der *relative*. Der relative Fehler ist skalierungsinvariant:

$$\varrho_{v\tilde{x}}(vx) = \varrho_{\tilde{x}}(x)$$

Die Grundoperationen $+, -, \cdot, /$ werden durch $\oplus, \ominus, \odot, \oslash$ auf dem Rechner realisiert. Für diese Grundoperationen gibt es Maschinenzahlen, sodass $x, y \in \mathbb{F}$, sodass gilt:

$$\varrho_{x \oplus y}(x + y) \leq \varepsilon$$

Entsprechende Ungleichungen gelten für $-, \cdot, /$.

3.6 Kondition von Problemen

Wir wollen verstehen, wie sensitiv die Lösung eines Problems von den Eingabedaten abhängig ist. Dabei betrachten wir nicht, wie wir das Problem konkret lösen, zum Beispiel durch einen Algorithmus. Alle Probleme, die wir behandeln, lassen sich mit Abbildungen formulieren.
Seien E die Menge der Eingabedaten, L die Menge der Lösungen und:

$$f : E \to L$$

Gesucht ist $f(z)$!
Beispiel: Wir suchen eine Lösung des linearen Gleichungssystems $Ax = b$ mit $A \in \mathbb{R}^{n \times n}$ und $b \in \mathbb{R}^n$. Mit $E_1 \in \mathrm{GL}(n, \mathbb{R})$ schreiben wir:

$$E = E_1 \times \mathbb{R}^n$$

Dann ist:

$$f(A, b) = A^{-1}b$$

Frage: Wie stark ändert sich $f(z)$, falls z fehlerbehaftet ist?
Um dies zu diskutieren, brauchen wir neue Begriffe.

3.7 Definition (Operatornorm)

Auf $\mathbb{R}^n$ beziehungsweise $\mathbb{R}^m$ seien Normen $\|.\|_a$ beziehungsweise $\|.\|_b$ gegeben. Für $A \in \mathbb{R}^{n \times m}$ ist die *Operatornorm* gegeben durch:

$$|||A|||_{a,b} := \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_b}{\|x\|_a}$$

Bemerkung

1. Für alle $x \in \mathbb{R}^n$ gilt:
$$\|Ax\|_b \leq |||A||| \cdot \|x\|_a$$
2. $|||A|||_{a,b}$ hängt von der Wahl von $\|.\|_a$ und $\|.\|_b$ ab.
3. Seien $A \in \mathbb{R}^{n\times m}$ und $B \in \mathbb{R}^{m\times l}$. $\|.\|_a$, $\|.\|_b$ und $\|.\|_c$ seien die Normen auf $\mathbb{R}^n$, $\mathbb{R}^m$ beziehungsweise $\mathbb{R}^l$. Dann gilt:
$$|||A \cdot B|||_{ac} \leq |||A|||_{ab} \cdot |||B|||_{bc}$$

Beweis

$$|||AB|||_{ac} = \sup_{x\in\mathbb{R}\setminus\{0\}} \frac{\|ABx\|}{\|x\|} \leq \sup |||A||| \cdot \frac{\|Bx\|}{\|x\|} \leq |||A||| \cdot |||B||| \cdot \|x\|$$

$\square_{3.}$

3.8 Satz und Definition (Kondition)

Sei $E \subseteq \mathbb{R}^{k_1} \times \ldots \times \mathbb{R}^{k_l}$ offen und $f : E \to \mathbb{R}^n \setminus \{0\}$ differenzierbar.
Dann gilt für alle $z \in \mathbb{R}^{k_1} \times \ldots \times \mathbb{R}^{k_l}$ mit $z_j \neq 0$ für $j \in \{1, \ldots, l\}$.

$$\varrho_{f(\overline{z})}(f(z)) \leq \sum_{j=1}^{l} \kappa_j(f, z)\, \varrho_{\overline{z}}(z_j) + R(z - \overline{z}) \qquad \text{für } \overline{z} \to z$$

Dabei ist

$$\varrho_{\overline{z}_j}(z_j) := \frac{\|z_j - \overline{z}_j\|_j}{\|z_j\|_j} \qquad \text{der relative Fehler der Daten,}$$

$$\varrho_{f(\overline{z})_j}(f(z)) := \frac{\|f(z) - f(\overline{z})\|}{\|f(z)\|} \qquad \text{der relative Fehler der Lösung,}$$

$$\kappa_j(f, z) := \frac{|||\mathrm{D}_j f(z)|||_j \cdot \|z_j\|_j}{\|f(z)\|} \qquad \text{der Verstärkungsfaktor}$$

und $\|.\|_j$ und $\|.\|$ sind Normen auf $\mathbb{R}^{k_j}$ beziehungsweise $\mathbb{R}^n$ für $j \in \{1, \ldots, l\}$. Für R gilt:

$$R(z - \overline{z}) = o_0\left(\sum_{j=1}^{l} \|z_j - \overline{z}_j\|_j\right)$$

Der Ausdruck $\kappa_j(f, z)$ heißt *Kondition von f im Punkte z in Richtung j.*

Beweis

Die Differenzierbarkeit von f liefert:

$$f(\overline{z}) = f(z) + \sum_{j=1}^{l} \mathrm{D}_j f(z)(\overline{z}_j - z_j) + \overline{R}$$

Damit folgt:

$$\overline{R} = o_0\left(\sum_{j=1}^{l} \|\overline{z}_j - z_j\|_j\right)$$

$$\frac{f(\overline{z}) - f(z)}{\|f(z)\|} = \frac{1}{\|f(z)\|}\sum_{j=1}^{l} \mathrm{D}_j f(z)(\overline{z}_j - z_j) + \frac{1}{\|f(z)\|}\overline{R}$$

Es folgt:

$$\varrho_{f(\overline{z})}(f(z)) = \frac{\|f(\overline{z}) - f(z)\|}{\|f(z)\|} \leq \frac{1}{\|f(z)\|}\sum_{j=1}^{l} \|\mathrm{D}_j f(z)(\overline{z}_j - z_j)\| + \underbrace{\frac{1}{\|f(z)\|}\left\|\overline{R}\right\|}_{=:R} \leq$$

$$\leq \frac{1}{\|f(z)\|}\sum_{j=1}^{l} \frac{\|z_j\|_j}{\|z_j\|_j} \||\mathrm{D}_j f(z)|\|_j \|\overline{z}_j - z_j\|_j + R =$$

$$= \sum_{j=1}^{l} \kappa_j(f,z)\,\varrho_{\overline{z}}(z_j) + R$$

Dabei erfüllt R die geforderten Eigenschaften. $\square_{3.8}$

3.9 Bemerkung

Die Kondition wird auch als Verstärkungsfaktor bezeichnet. Eine hohe Kondition bewirkt, dass ein Fehler in den Daten zu großen Fehlern in der Lösung $f(z)$ führen kann.
Achtung: Die Ungleichung liefert nur eine Abschätzung, die manchmal zu pessimistisch ist.

3.10 Kondition der elementaren Operationen

a) *Multiplikation*: Betrachte:

$$f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$$
$$(a,b) \mapsto ab$$

Wende den Satz 3.8 an ($k_1 = k_2 = 1$). Alle Normen sind der Betrag. Für alle $a, b \in \mathbb{R} \setminus \{0\}$ gilt:

$$\kappa_1(f,(a,b)) = \frac{|D_1 f(a,b)| \cdot |a|}{|ab|} = \frac{|b|}{|b|} = 1$$

Es folgt:

$$\varrho_{\overline{a}\cdot\overline{b}}(a \cdot b) \leq \varrho_{\overline{a}}(a) + \varrho_{\overline{b}}(b) + R$$

Also ist die Multiplikation (und analog die Division) gut konditioniert. Fehler werden nicht wesentlich verstärkt.

b) *Addition*: Betrachte:

$$\begin{aligned} f : \mathbb{R} \times \mathbb{R} &\to \mathbb{R} \\ (a, b) &\mapsto a + b \end{aligned}$$

Wende den Satz 3.8 an ($k_1 = k_2 = 1$). Alle Normen sind der Betrag. Für alle $a, b \in \mathbb{R}$ gilt:

$$\begin{aligned} \kappa_1(f, (a, b)) &= \frac{|D_1 f(a, b)| \cdot |a|}{|a + b|} = \frac{|1| \cdot |a|}{|a + b|} = \frac{|a|}{|a + b|} \\ \kappa_2(f, (a, b)) &= \frac{|b|}{|a + b|} \end{aligned}$$

Es folgt:

$$\varrho_{\overline{a}+\overline{b}}(a + b) \leq \frac{|a|}{|a + b|}\varrho_{\overline{a}}(a) + \frac{|b|}{|a + b|}\varrho_{\overline{b}}(b) + R$$

κ_1 und κ_2 werden groß, falls $a + b$ klein ist. Also ist die Addition schlecht konditioniert, wenn $a \approx -b$ ist. Fehler werden dann wesentlich verstärkt.

Auf dem Rechner macht sich dies durch Auslöschung führender Zahlen bemerkbar.
Beispiel: Dezimalzahlen im Rechner mit höchstens 6 Ziffern.
Daten: $a = 1{,}2346789$; $b = 1{,}2345678$

$$a - b = 0{,}000111$$

Auf dem Rechner:

$$\overline{a} = 1{,}23467 \qquad \overline{b} = 1{,}23456$$

$$\overline{a} - \overline{b} = 0{,}00011$$

Also sind nur zwei Ziffern genau. Dies bedeutet einen großen relativen Fehler.

3.11 Kondition des Skalarprodukt

$$\begin{aligned} f : \mathbb{R}^n \times \mathbb{R}^n &\to \mathbb{R} \\ (x, y) &\mapsto x \cdot y = \sum_{i=1}^{n} x_i y_i \end{aligned}$$

Der $\mathbb{R}^n$ sei versehen mit der euklidischen Norm. Es gilt:

$$\mathrm{D}_1 f(x, y) = (y_1, \ldots, y_n)$$

$$\|\|\|\mathrm{D}_1 f(x,y)\|\| = \sup_{x\in\mathbb{R}\setminus\{0\}} \frac{|x\cdot y|}{\|x\|} \overset{\text{max. bei } x=y}{=} \|y\|$$

$$\Rightarrow \quad \kappa_1(f,(x,y)) = \frac{\|y\|\cdot\|x\|}{|x\cdot y|} = \frac{1}{|\cos(x,y)|}$$

Dabei ist $\cos(x,y)$ der Kosinus des Winkels zwischen x und y.

Mit Satz 3.8 folgt, dass das Skalarprodukt schlecht konditioniert ist, wenn x und y nahezu senkrecht aufeinander stehen.

TODO: Abb1 einfügen

3.12 Kondition linearer Gleichungssysteme

Gesucht ist x, sodass $Ax = b$ ist, das heißt $x = A^{-1}b$.
Frage: Wie stark ändert sich x, wenn A und b leicht gestört sind?
Betrachte:

$$\begin{aligned} f : \mathrm{GL}(n)\times\mathbb{R}^n &\mapsto \mathbb{R}^n \\ (A,b) &\mapsto A^{-1}b \end{aligned}$$

$\|.\|$ sei Norm auf $\mathbb{R}^n$ und $\|\|.\|\|$ die zugehörige Operatornorm. Es gilt:

$$\kappa_2(A,b) = \frac{\|\|\mathrm{D}_2 f\|\|\cdot\|b\|}{\|A^{-1}b\|} = \frac{\left\|\left\|A^{-1}\right\|\right\|\cdot\|Ax\|}{\|x\|} \le \frac{\left\|\left\|A^{-1}\right\|\right\|\cdot\|\|A\|\|\cdot\|x\|}{\|x\|} = \left\|\left\|A^{-1}\right\|\right\|\cdot\|\|A\|\|$$

Weiter gilt:

$$\kappa_1(A,b) = \frac{\|\|\mathrm{D}_1 f\|\|\cdot\|\|A\|\|}{\|A^{-1}b\|}$$

Es gilt:

$$\mathrm{D}_1 f = \mathrm{D}_A\left(A^{-1}b\right)$$

Frage: Wie leite ich $A \mapsto A^{-1}$ ab?

3.12.1 **Lemma** (Neumannsche Reihe)

Sei $\|\|.\|\|$ eine Operatornorm und sei $B \in \mathrm{GL}(n)$ mit $\|\|B\|\| < 1$. Dann gilt:

$$(\mathbb{1} - B)^{-1} = \sum_{k=0}^{\infty} B^k$$

Beweis

Zunächst zeigen wir:

$$\sum_{k=0}^{\infty} B^k < \infty$$

Sei $q := |||B||| < 1$.

$$\left|\left|\left|\sum_{k=m}^{l} B^k\right|\right|\right| \leq \sum_{k=m}^{l} \left|\left|\left|B^k\right|\right|\right| \overset{3.7}{\leq} \sum_{k=m}^{l} |||B|||^k \leq \sum_{k=m}^{l} q^k \xrightarrow[\text{geo. R.}]{m,l\to\infty} 0$$

Das Cauchysche Konvergenz-Kriterium liefert:

$$\lim_{l\to\infty} \sum_{k=0}^{l} B^k < \infty$$

Es gilt:

$$\left|\left|\left|B^k\right|\right|\right| \leq |||B|||^k = q^k \xrightarrow{k\to\infty} 0$$

$$\left(\sum_{k=0}^{\infty} B^k\right)(\mathbb{1} - B) = \lim_{l\to\infty} \left(\sum_{k=0}^{l} B^k\right)(\mathbb{1} - B) =$$
$$= \lim_{l\to\infty} \left(\sum_{k=0}^{l} B^k - B^{k+1}\right) = \lim_{l\to\infty} \left(\mathbb{1} - B^{l+1}\right) = \mathbb{1}$$

$\square_{3.12.1}$

3.12.2 Lemma

Die Abbildung

$$h : \mathrm{GL}(n) \to \mathrm{GL}(n)$$
$$A \mapsto A^{-1}$$

ist differenzierbar und es gilt:

$$\mathrm{D}h(A) : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$$
$$B \mapsto -A^{-1}BA^{-1}$$

Beweis

Sei $|||B|||$ klein.

$$(A+B)^{-1} = \left(A\left(\mathbb{1} + A^{-1}B\right)\right)^{-1} = \left(\mathbb{1} + A^{-1}B\right)^{-1} A^{-1} =$$
$$\overset{\text{Neumannsche Reihe}}{=} \left(\sum_{k=0}^{\infty} \left(-A^{-1}B\right)^k\right) A^{-1} = A^{-1} - A^{-1}BA^{-1} + \mathcal{O}_0\left(|||B|||^2\right)$$

Es folgt:

$$(A+B)^{-1} - A^{-1} = -A^{-1}BA^{-1} + \mathcal{O}_0\left(|||B|||^2\right)$$

Damit erfüllt die Abbildung $B \mapsto -A^{-1}BA^{-1}$ die Definition der Ableitung. $\square_{3.12.2}$

Es war zu berechnen:

$$\mathrm{D}_1 f = \mathrm{D}_A \left(A^{-1} b\right)$$

Nach dem Lemma gilt:

$$\begin{aligned} \mathrm{D}_1 f\,(A,b) : \mathbb{R}^{n\times n} &\to \mathbb{R}^n \\ B &\mapsto -A^{-1}BA^{-1}b \end{aligned}$$

$$\left|\left|\left|\mathrm{D}_1 f\,(A,b)\right|\right|\right| = \sup_B \frac{\left\|A^{-1}BA^{1-}b\right\|}{|||B|||} \leq \sup_B \frac{\left|\left|\left|A^{-1}\right|\right|\right| \cdot |||B||| \cdot \|x\|}{|||B|||} = \left|\left|\left|A^{-1}\right|\right|\right| \cdot \|x\|$$

Insgesamt folgt:

$$\kappa_1\,(A,b) = \frac{\left|\left|\left|\mathrm{D}_1 f\right|\right|\right| \cdot |||A|||}{\|A^{-1}b\|} \leq \frac{\left|\left|\left|A^{-1}\right|\right|\right| \cdot \|x\| \cdot |||A|||}{\|x\|} = \left|\left|\left|A^{-1}\right|\right|\right| \cdot |||A|||$$

□??

3.13 Kondition einer Matrix

Definition: Sei $A \in \mathrm{GL}\,(\mathbb{R}, n)$ und $\|.\|$ eine Norm auf $\mathbb{R}^n$. Dann definieren wir die *Kondition* von A wie folgt:

$$K\,(A) := |||A||| \cdot \left|\left|\left|A^{-1}\right|\right|\right|$$

(Dabei wird die gleiche Norm in Bild und Urbild verwendet.)

Die Kondition einer Matrix A gibt an, wie stark Eingangsfehler die Lösung des Gleichungssystems $Ax = b$ beeinflussen können. Die Aussage gilt für Fehler in A und b!

Lemma: $K\,(A) \geq 1$

Beweis: Es gilt:

$$1 = |||\mathbb{1}||| = \left|\left|\left|A \cdot A^{-1}\right|\right|\right| \leq |||A||| \cdot \left|\left|\left|A^{-1}\right|\right|\right| = K\,(A)$$

□Lemma

Ist $K\,(A)$ sehr groß, so sagen wir, A ist schlecht konditioniert.

3.14 Kondition von nichtlinearen Gleichungen

Sei $g : \mathbb{R}^n \to \mathbb{R}^n$ stetig differenzierbar und $x \in \mathbb{R}^n$.

Problem (P): Gesucht ist $y \in \mathbb{R}^n$, sodass $y = g\,(x)$ gilt.

Es gelte:

$$\det\left(\mathrm{D}g\,(y)\right) \neq 0$$

Dann ist $\mathrm{D}g\,(y)$ invertierbar und es existiert nach dem Satz über die lokale Umkehrbarkeit (siehe Analysis II) eine Inverse $f = g^{-1}$ in einer lokalen Umgebung von x.

Die Abbildung, die die Lösung von (P) beschreibt, lautet:

$$x \mapsto g^{-1}(x) =: f(x)$$

Die Kondition dieses Problems ist gegeben durch:

$$K(f,x) = \frac{|||\mathrm{D}f(x)|||}{\|f(x)\|} \cdot \|x\| = \frac{\left|\left|\left|\left(\mathrm{D}g(f(x))\right)^{-1}\right|\right|\right|}{\|f(x)\|} \|x\| = \frac{\left|\left|\left|\left(\mathrm{D}g(y)\right)^{-1}\right|\right|\right| \cdot \|x\|}{\|y\|}$$

Beispiel $n = 1$:

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

TODO: Abbildung einfügen

3.15 Stabilität von Algorithmen

Sei $E \subseteq \mathbb{R}^n$ und $f : E \to \mathbb{R}^n$.
Gesucht ist $y = f(x)$ für $x \in E$.
Auf dem Rechner stehen elementare Operationen wie $+$, $-$, $\cdot$, $/$ zur Verfügung, die mit einer relativen Genauigkeit ε realisiert sind. Für alle $x \in \mathbb{F}$ gilt für jede elementare Operation φ also:

$$\varrho_{\overline{\varphi(x)}}(\varphi(x)) \leq \varepsilon$$

Definition (Algorithmus)

Eine Zerlegung der Abbildung $f : E \to \mathbb{R}^n$ der Form $f = f^{(l)} \circ \ldots \circ f^{(1)}$ mit $l \in \mathbb{N}$, $f^{(i)} U_i \to U_{i+1}$ ($U_i \subseteq \mathbb{R}^{k_i}$, $\mathbb{R}^{k_{i+1}}$) und $k_i \in \mathbb{N}$, $k_1 = k$ sowie $k_{l+1} = n$ heißt *Algorithmus*, falls alle $f^{(i)}$ durch elementare Operationen ausführbar sind.
Beispiel:

$$\begin{aligned} f : \mathbb{R} &\to \mathbb{R} \\ x &\mapsto \frac{x^2 - 1}{x^2 + 1} \end{aligned}$$

$$\begin{aligned} f^{(1)} : \mathbb{R} &\to \mathbb{R} \\ x &\mapsto x^2 \end{aligned}$$

$$\begin{aligned} f^{(2)} : \mathbb{R} &\to \mathbb{R}^2 \\ x &\mapsto (y-1, y+1) \\ f^{(3)} : \mathbb{R}^2 &\to \mathbb{R} \\ (x,y) &\mapsto \frac{x}{y} \end{aligned}$$

$$f(x) = \left(f^{(3)} \circ f^{(2)} \circ f^{(1)}\right)(x)$$

Die Umsetzung auf dem Rechner

$$\overline{f} = \overline{f}^{(b)} \circ \ldots \circ \overline{f}^{(1)} : \mathbb{F}^k \to \mathbb{F}^n$$

heißt *Implementation* von f.

Im Allgemeinen gibt es viele Möglichkeiten, f in Elementare Operationen aufzuspalten, das heißt viele Algorithmen.

Ziel: Verstehe den Einfluss eines Algorithmus auf die Fehlerverstärkung.

3.16 Lemma

Sei $x \in \mathbb{R}^k$ und $\overline{x} \in \mathbb{F}^k$, sodass $\varrho_{\overline{x}_i}(x_i) \leq \varepsilon$ für $i \in \{1, \ldots, k\}$ gilt. Sei weiter $f = f^{(l)} \circ \ldots \circ f^{(1)}$ ein Algorithmus zur Berechnung von f und sei $\overline{f} = \overline{f}^{(l)} \circ \ldots \circ \overline{f}^{(1)}$ eine Implementierung. Mit den Abkürzungen

$$x^{(j+1)} := f^{(j)} \circ \ldots \circ f^{(1)}(x) \qquad x^{(1)} := x$$
$$\overline{x}^{(j+1)} := \overline{f}^{(j)} \circ \ldots \circ \overline{f}^{(1)}(\overline{x}) \qquad \overline{x}^{(1)} := \overline{x}$$

gilt:

$$\varrho_{\overline{x}_i^{(j+1)}}\left(x_i^{j+1}\right) \leq \varepsilon + \sum_m \kappa_m\left(f_i^{(j)}, x^{(j)}\right) \varrho_{\overline{x}_m^{(j)}}\left(x_m^{(j)}\right) + R$$

Dabei ist R ein Restterm, der quadratisch in den Fehlern und ε ist. Es sei stets $f_i^{(j)}\left(x^{(j)}\right) \neq 0$ für alle i, j.

Beweis

$$\begin{aligned}
\varrho_{\overline{x}_i^{(j+1)}}\left(x_i^{(j+1)}\right) &= \frac{\left|f_i^{(j)}\left(x^{(j)}\right) - \overline{f}_i^{(j)}\left(\overline{x}^{(j)}\right)\right|}{\left|f_i^{(j)}\left(x^{(j)}\right)\right|} \leq \\
&\leq \frac{\left|f_i^{(j)}\left(x^{(j)}\right) - f_i^{(j)}\left(\overline{x}^{(j)}\right)\right|}{\left|f_i^{(j)}\left(x^{(j)}\right)\right|} + \frac{\left|f_i^{(j)}\left(\overline{x}^{(j)}\right) - \overline{f}_i^{(j)}\left(\overline{x}^{(j)}\right)\right|}{\left|f_i^{(j)}\left(x^{(j)}\right)\right|} \leq \\
&\leq \frac{\left|f_i^{(j)}\left(x^{(j)}\right) - f_i^{(j)}\left(\overline{x}^{(j)}\right)\right|}{\left|f_i^{(j)}\left(x^{(j)}\right)\right|} + \underbrace{\frac{\left|f_i^{(j)}\left(\overline{x}^{(j)}\right) - \overline{f}_i^{(j)}\left(\overline{x}^{(j)}\right)\right|}{\left|f_i^{(j)}\left(\overline{x}^{(j)}\right)\right|}}_{\leq \mathrm{eps}} \cdot \underbrace{\frac{\left|f_i^{(j)}\left(\overline{x}^{(j)}\right)\right|}{\left|f_i^{(j)}\left(x^{(j)}\right)\right|}}_{\leq 1 + \frac{\left|f_i^{(j)}\left(x^{(j)}\right) - f_i^{(j)}\left(\overline{x}^{(j)}\right)\right|}{\left|f_i^{(j)}\left(x^{(j)}\right)\right|}} \leq \\
&\leq \varrho_{f(\overline{x})}(f(x)) + \mathrm{eps} + R \leq \\
&\leq \sum_m \kappa_m\left(f_i^{(j)}, x^{(j)}\right) \varrho_{\overline{x}_m^{(j)}}\left(x_m^{(j)}\right) + \mathrm{eps} + R
\end{aligned}$$

$\square_{3.16}$

3.17 Bemerkungen

Nutze obiges Lemma rekursiv, um entstehende Fehler zu analysieren. Insgesamt erhalte eine Abschätzung der Form:

$$\varrho_{\overline{f}(x)}\left(f\left(x\right)\right) \leq K\left(x\right)\varepsilon + R$$

Dabei ist K eine Größe, die sich aus Produkten der Konditionen $\kappa_m\left(f_i^{(j)}, x^{(j)}\right)$ zusammensetzt und R ist quadratisch in ε.
Der Quotient $\frac{K(x)}{K(f,x)}$ ist ein Indikator für die Güte (Stabilität) eines Algorithmus. Der Quotient sagt aus, wie stark die Fehler durch den Algorithmus zusätzlich verstärkt werden. Je kleiner der Quotient ist, desto besser ist der Algorithmus.

3.18 Beispiel (quadratische Gleichung)

Gesucht ist ein x mit $x^2 + 2px - q = 0$. Nehme $p, q \geq 0$ und $p \gg q$ an. Die Nullstellen sind:

$$x_\pm = -p \pm \sqrt{p^2 + q}$$

Ziel: Berechne die größere Nullstelle. Sei $f\left(p,q\right) = -p + \sqrt{p^2+q}$ und $w := \sqrt{p^2+q}$. Betrachte zwei Algorithmen zur Berechnung von f. Nutze die Identität $x_+ \cdot x_- = -1$, mit der folgt:

$$f\left(p,q\right) = \frac{-q}{-p - \sqrt{p^2+q}}$$

Setze voraus, dass die Wurzelfunktion eine elementare Operation ist.

Algorithmus A	Algorithmus B
$f_1 := p \cdot p$	$q_1 := p \cdot p$
$f_2 := f_1 + q$	$q_2 := q_1 + q$
$f_3 := \sqrt{f_m}$	$q_3 := \sqrt{q_2}$
$f_4 := -p + f_3$	$q_4 := -p - g_3$
	$q_5 := \frac{-q}{g_4}$

Zur Vereinfachung seien p und q Maschinenzahlen, also $\varrho_{\overline{p}}\left(p\right) = 0$ und $\varrho_{\overline{q}}\left(q\right) = 0$. Definiere:

$$\varrho_i := \varrho_{\overline{f}_i}\left(f_i\right) \qquad \delta_i := \varrho_{\overline{g}_i}\left(g_i\right)$$

Es gilt bei Vernachlässigung quadratischer Fehlerterme:

$\varrho_1 \leq \varepsilon + K\left(f_1, p\right)\varrho_{\overline{p}}\left(p\right) = \varepsilon$	$\delta_1 \leq \varepsilon$
$\varrho_2 \leq \varepsilon + \underbrace{K\left(f_2, f_1\right)}_{=\frac{\lvert f_1\rvert}{\lvert f_1+q\rvert}\leq 1}\varrho_1 \leq 2\varepsilon$	$\delta_2 \leq 2\varepsilon$
$\varrho_3 \leq \varepsilon + \underbrace{K\left(f_3, f_2\right)}_{=\frac{\frac{1}{2}\frac{1}{\sqrt{f_2}}f_2}{\sqrt{f_2}}=\frac{1}{2}}\varrho_2 \leq 2\varepsilon$	$\delta_3 \leq 2\varepsilon$

$$\varrho_4 \leq \varepsilon + \underbrace{K(f_4, f_3)}_{=\frac{f_3}{-p+f_3}} \varrho_3 \leq \left(1 + 2\frac{\sqrt{pq}}{\sqrt{p^2+q}-p}\right)\varepsilon \qquad \delta_4 \leq \varepsilon + K(g_4, g_3)\,\delta_3 \leq 3\varepsilon$$

$$\delta_5 \leq 4\varepsilon$$

$$f(p,q) = -p + \sqrt{p^2+q}$$

Die Kondition von f ist:

$$\kappa_1(f,(p,q)) = \frac{|\partial_p f| \cdot |p|}{|f|} = \left|\frac{\left(-1 + \frac{p}{\sqrt{p^2+q}}\right)p}{-p+\sqrt{p^2+q}}\right| = \left|\frac{-p}{\sqrt{p^2+q}}\frac{\sqrt{p^2+q}-p}{-p+\sqrt{p^2+q}}\right| \leq 1$$

$$\kappa_1(f,(p,q)) = \frac{|\partial_q f| \cdot |q|}{|f|} = \frac{1}{2}\left(\frac{|p|}{\sqrt{p^2+q}} + 1\right) \leq 1$$

Es folgt, dass das Berechnen von Nullstellen gut konditioniert ist.
Algorithmus A ist schlecht, da er die Fehler eines gut konditionierten Problems stark verstärkt. ($p \gg q$)
Algorithmus B ist gut konditioniert. (Gesamtfehler ist gegen 4ε beschränkt.)

3.19 Rückwärtsanalyse

Bisher haben wir stets die Vorwärtsanalyse betrachtet:
Gesucht ist $f(x)$. Statt mit x rechnen wir mit $\overline{x}\,(\in \mathbb{F})$.
Berechne $f(\overline{x})$ durch Implementation eines Algorithmus:

$$\overline{f} = \overline{f}^{(l)} \circ \ldots \circ \overline{f}^{(1)}$$

Dann haben wir den relativen Fehler

$$\frac{\left\|\overline{f}(\overline{x}) - f(x)\right\|}{\|f(x)\|}$$

untersucht.
Rückwärtsanalyse: Finde $\tilde{x}$, sodass $f(\tilde{x}) = \overline{f}(\overline{x})$ und schätze $\|x - \tilde{x}\|$ ab, das heißt:
Interpretiere das Ergebnis als exakte Lösung einer gestörten Eingabe.
Dazu ein Beispiel:

3.20 Satz

Zur Lösung eines linearen Gleichungssystems $Ax = b$ mit $A \in \mathbb{R}^{n\times n}$ und $b,\ x \in \mathbb{R}^n$, sei die Gauß-Elimination (siehe 2.5) und die anschließende Auflösung des gestaffelten Gleichungssystems (wie in 2.13) auf einem Rechner mit Maschinengenauigkeit eps implementiert.

Das Programm berechnet ein $\overline{x} \in \mathbb{R}^n$. Es gilt, dass $\overline{x}$ Lösung von

$$\overline{A}\overline{x} = b$$

ist ($\overline{A} = (\overline{a}_{ij})\,;\ A = (a_{ij})$) mit:

$$|\overline{a}_{ij}| \leq \gamma_n \,(2 + \gamma_n)\, \left|\overline{L}\right| \left|\overline{R}\right|$$

Dabei sind $\overline{L}$ und $\overline{R}$ die berechneten Dreiecksmatrizen ($\overline{L} = \left(\overline{l}_{ij}\right)\,;\ \overline{R} = (\overline{r}_{ij})$) und:

$$\left|\overline{L}\right| = \max_{i,j} \left|\overline{l}_{ij}\right| \qquad \left|\overline{R}\right| = \max_{i,j} \left|\overline{r}_{ij}\right| \qquad \gamma_n := \frac{n \cdot \text{eps}}{1 - n \cdot \text{eps}}$$

Dabei sei vorausgesetzt, dass $n \cdot \text{eps} < 1$ gilt.

(ohne Beweis)

Der Beweis steht im Buch von Deuflhard und Hohmann in Kapitel 2.4.

Üblicherweise ist n sehr viel kleiner als $\frac{1}{\text{eps}}$.

Falls $\overline{L}$ und $\overline{R}$ große Einträge haben, löst $\overline{x}$ ein Gleichungssystem mit möglicherweise stark veränderter Matrix $\overline{A}$. (Formuliere dies positiv: Falls $\overline{L}$ und $\overline{R}$ kleine Einträge haben, löst $\overline{x}$ ein Gleichungssystem mit schwach veränderter Matrix $\overline{A}$.)

3.21 Bemerkung

Die Gauß-Elimination (LR-Zerlegung) angewandt auf die Hilbert-Matrix

$$H = \left(\frac{1}{i + j - 1}\right)_{i,j \in \{1,\ldots,n\}}$$

liefert für große n sehr schlechte Ergebnisse. Die Ursache ist, dass die Kondition

$$\kappa\,(H) = |||H||| \left|||H^{-1}|||\right.$$

sehr groß ist, also gibt es eine große Fehlerverstärkung (vergleiche 3.12).

Satz 3.20 sagt: Wenn $\overline{L}$ und $\overline{R}$ keine zu großen Einträge haben, löst $\overline{x}$ eine Gleichung $\overline{H} \cdot \overline{x} = b$, wobei $\overline{H}$ nahe H ist.

Es ist leider eine Tatsache, dass bei der Hilbert-Matrix $\overline{L}$ und $\overline{R}$ große Einträge haben.

3.22 Einige Grundregeln, die sich aus diesem Kapitel ergeben.

- Kenntnisse über die Kondition eines Problems sind entscheidend für die Bewertung der Ergebnisse.
- Multiplikationen und Divisionen sind gut konditioniert.
- Subtraktion zweier annähernd gleicher Zahlen ist schlecht konditioniert (Auslöschung). Vermeide dies nach Möglichkeit!
- Addition von Zahlen mit gleichem Vorzeichen ist gut konditioniert.

- Unvermeidliche, schlecht konditionierte Elementaroperationen sollte man meist möglichst früh im Algorithmus durchführen, zum Beispiel unvermeidliche Subtraktionen. Dann wirkt die Fehlerverstärkung noch auf kleinere Fehler.
- Bei einem stabilen Algorithmus bleiben die im Laufe der Rechnung erzeugten Fehler in der Größenordnung der durch die Kondition bedingten unvermeidlichen Fehler.
- Vermeide Abfragen, ob eine Größe gleich Null ist, oder ob zwei Größen gleich sind. (Dies ist wegen Rundungsfehlern nicht aussagekräftig.)
- Beachte vorteilhafte Reihenfolge bei Summation: erst die kleinen, dann die großen.

4 QR-Zerlegung, lineare Ausgleichsprobleme

4.1 Einführende Bemerkungen zur QR-Zerlegung

Eine $(n \times n)$-Matrix A besitze eine Zerlegung $A = QR$ mit einer orthogonalen $(n \times n)$-Matrix Q und einer oberen Dreiecksmatrix R. Q orthogonal heißt:

$$\begin{aligned} & QQ^{\mathrm{T}} = \mathbb{1} \\ \Leftrightarrow \quad & Q^{-1} = Q^{\mathrm{T}} \end{aligned}$$

Löse das Gleichungssystem $Ax = b$ wie folgt:

1. Löse $Qz = b$, oder äquivalent $z = Q^{-1}b = Q^{\mathrm{T}}b$. Also ist z hier einfach berechenbar.
2. Löse $Rx = z$. (gestaffeltes Gleichungssystem)

Einfache orthogonale Abbildungen (Matrizen) sind Drehungen und Spiegelungen.

Ziel: Multipliziere A sukzessive mit Drehungen oder Spiegelungen, bis eine obere Dreiecksmatrix vorliegt.

$$A \rightsquigarrow Q^{(1)}A \rightsquigarrow Q^{(2)}Q^{(1)}A \rightsquigarrow \ldots \rightsquigarrow \underbrace{Q^{(m)}\ldots Q^{(1)}}_{=Q^{\mathrm{T}}} A = R$$

4.2 Hyperebenen-Spiegelungen

Sei $u \in \mathbb{R}^n$ mit $\|u\| = 1$.

$$H_u = \left\{ x \in \mathbb{R}^n \middle| x \cdot u = 0 \right\}$$

sei die Hyperebene, die zu u senkrecht steht.

TODO: Abbildung 2D-Hyperebene

Gesucht ist eine Matrix Q, die die Spiegelung an H_u beschreibt.

Eine Spiegelung Q an der Hyperebene H_u ist definiert durch die Eigenschaften:

$$\begin{aligned} Qx &= x \qquad \underset{x \in H_u}{\forall} \\ Qu &= -u \end{aligned}$$

Sei $x \in \mathbb{R}^n$ beliebig, so zerlege in einen Anteil parallel zu u und einen senkrechten Anteil:

$$x = \alpha u + (x - \alpha u)$$

Bestimme α so, dass $(x - \alpha u) \cdot u = 0$ ist, das heißt $\alpha = x \cdot u$, womit folgt:

$$x = (x \cdot u)\, u + \underbrace{(x - (x \cdot u)\, u)}_{\in H_u}$$

Also gilt:

$$\begin{aligned} Qx &= -(x \cdot u)\, u + (x - (x \cdot u)\, u) = x - 2\,(x \cdot u)\, u = \\ &= \mathbb{1}x - 2\left(uu^{\mathrm{T}}\right) x \end{aligned}$$

In Matrixform ist das:

$$Q = \mathbb{1} - 2\left(uu^{\mathrm{T}}\right)$$

Matrizen dieser Form heißen Householder-Matrizen.
Bemerkung: u und $-u$ definieren die gleiche Spiegelung.

4.3 Lemma

Für $u \in \mathbb{R}^n \setminus \{0\}$ mit $\|u\| = 1$ und $Q = \mathbb{1} - 2uu^{\mathrm{T}}$ gilt:

$$\begin{aligned} Q &= Q^{\mathrm{T}} && \text{(symmetrisch)} \\ Q^{-1} &= Q^{\mathrm{T}} && \text{(orthogonal)} \\ Q^2 &= \mathbb{1} && \text{(involutorisch)} \end{aligned}$$

Beweis

$$\begin{aligned} Q^{\mathrm{T}} &= \mathbb{1} - 2\left(uu^{\mathrm{T}}\right)^{\mathrm{T}} = \mathbb{1} - 2\left(uu^{\mathrm{T}}\right) = Q \\ Q^2 = QQ^{\mathrm{T}} &= \left(\mathbb{1} - 2\left(uu^{\mathrm{T}}\right)\right)\left(\mathbb{1} - 2\left(uu^{\mathrm{T}}\right)\right) = \\ &= \mathbb{1} - 2\left(uu^{\mathrm{T}}\right) - 2\left(uu^{\mathrm{T}}\right) + 4u\underbrace{u^{\mathrm{T}}u}_{=1}u^{\mathrm{T}} = \mathbb{1} \end{aligned}$$

$\square_{4.3}$

4.4 QR-Zerlegung mit Householder-Matrizen

Sei $A \in \mathbb{R}^{m \times n}$, das heißt:

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix}$$

Ziel: Finde Spiegelung $Q^{(1)} \in \mathbb{R}^{m \times m}$, sodass gilt:

$$Q^{(1)}A = \begin{pmatrix} * & \dots & \dots & * \\ 0 & * & & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & * & \dots & * \end{pmatrix}$$

Sei

$$a_j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{pmatrix}$$

die j-te Spalte der Matrix A. Dann muss gelten:

$$Q^{(1)} a_1 = \pm \|a_1\| e_1$$

a_1 muss auf einen Vektor der Länge $\|a_1\|$ abgebildet werden, da $Q^{(1)}$ orthogonal sein soll.
TODO: Abb2 einfügen
Definiere:

$$u_1 = \pm \frac{a_1 \pm \|a_1\| e_1}{\|a_1 \pm \|a_1\| e_1\|}$$

Wähle:

$$v_1 = \frac{a_1 + \operatorname{sgn}(a_{11}) \|a_1\| e_1}{\|a_1 + \operatorname{sgn}(a_{11}) \|a_1\| e_1\|}$$

Diese Vorzeichenwahl dient der Vermeidung einer möglicherweise schlechten Konditionierung.
Nun definiere:

$$Q^{(1)} = \mathbb{1} - 2 v_1 v_1^{\mathrm{T}}$$

$$\begin{aligned}
Q^{(1)} a_1 = a_1 - 2 v_1 v_1^{\mathrm{T}} a_1 &= a_1 - 2 \frac{a_1 + \operatorname{sgn}(a_{11}) \|a_1\| e_1}{\|a_1 + \operatorname{sgn}(a_{11}) \|a_1\| e_1\|} \left(\frac{a_1 + \operatorname{sgn}(a_{11}) \|a_1\| e_1}{\|a_1 + \operatorname{sgn}(a_{11}) \|a_1\| e_1\|} \right)^{\mathrm{T}} a_1 = \\
&= a_1 - 2 \left(a_1 + \operatorname{sgn}(a_{11}) \|a_1\| e_1\right) \cdot \frac{\|a_1\|^2 + |a_{11}| \cdot \|a_1\|}{\left\| \left(a_{11} + \operatorname{sgn}(a_{11}) \|a_1\|\right)^2 + \|a_1\|^2 - a_{11}^2 \right\|^2} = \\
&= a_1 - 2 \left(a_1 + \operatorname{sgn}(a_{11}) \|a_1\| e_1\right) \cdot \frac{\|a_1\|^2 + |a_{11}| \cdot \|a_1\|}{2 \left(|a_{11}| \|a_1\| + \|a_1\|^2 \right)} = \\
&= a_1 - \left(a_1 + \operatorname{sgn}(a_{11}) \|a_1\| e_1\right) = -\operatorname{sgn}(a_{11}) \|a_1\| e_1
\end{aligned}$$

Wir erhalten: $A^{(2)} := Q^{(1)} \cdot A^{(1)}$ mit $A^{(1)} = A$. Mit

$$\alpha_1 = -\operatorname{sgn}(a_{11}) \|a_1\| e_1$$

gilt:

$$A^{(2)} = \begin{pmatrix} \alpha_1 & & & \\ 0 & & & \\ \vdots & Q^{(1)} a_2 & \dots & Q^{(1)} a_n \\ 0 & & & \end{pmatrix}$$

Iteration liefert:

$$A^{(k)} = \begin{pmatrix} \alpha_1 & & & & & & \\ 0 & \ddots & & & & & \\ \vdots & \ddots & \ddots & & & & \\ \vdots & & \ddots & \alpha_{k-1} & & & \\ \vdots & & & 0 & & & \\ \vdots & & & \vdots & \tilde{a}_{k+1}^{(k)} & \dots & \tilde{a}_n^{(k)} \\ 0 & \dots & \dots & 0 & & & \end{pmatrix} = \begin{pmatrix} \alpha_1 & & & * \\ & \ddots & & \\ & & \alpha_{k-1} & \\ 0 & & & B^{(k)} \end{pmatrix}$$

$B^{(k)}$ ist eine $(m-k+1)\times(n-k+1)$-Matrix. Wähle $(m-k+1)\times(m-k+1)$-Spiegelungs-Matrix $\tilde{Q}^{(k)}$ sodass gilt:

$$\tilde{Q}^{(k)}B^{(k)} = \begin{pmatrix} * & \dots & \dots & * \\ 0 & * & \dots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & * & \dots & * \end{pmatrix}$$

Dazu sei $v_k \in \mathbb{R}^{m-k+1}$ mit:

$$\begin{aligned} v_k &:= \frac{\tilde{a}_k^{(k)} - \alpha_k e_k^{(k)}}{\left\|\tilde{a}_k^{(k)} - \alpha_k e_k^{(k)}\right\|} \\ \alpha_k &:= -\operatorname{sgn}\left(\tilde{a}_{kk}^{(k)}\right)\left\|\tilde{a}_k^{(k)}\right\| \\ e_1^{(k)} &:= (1,0,\dots,0)^{\mathrm{T}} \in \mathbb{R}^{m-k+1} \\ \tilde{Q}^{(k)} &:= 1 - 2v_k v_k^{\mathrm{T}} \end{aligned}$$

Definiere nun:

$$Q^{(k)} = \begin{pmatrix} \mathbb{1}_{k-1} & 0 \\ 0 & \underbrace{\tilde{Q}^{(k)}}_{n-k+1 \text{ Spalten}} \end{pmatrix}$$

Das Verfahren endet, wenn $k = p := \min(n, m-1)$ ist. Wir erhalten $R = A^{(p)} = Q^{(p)}\cdot\ldots\cdot Q^{(1)}A$. Es gilt:

$$\begin{aligned} &\left(Q^{(k)}\right)^{-1} = Q^k \\ \Rightarrow \qquad &A = Q^{(1)}\cdot\ldots\cdot Q^{(p)}R = QR \end{aligned}$$

4.5 Algorithmus

Das Vorgehen bei der Berechnung der QR-Zerlegung ist im k-ten Schritt:

1. $\alpha_k = -\operatorname{sgn}\left(\tilde{a}_{kk}^{(k)}\right)\left\|\tilde{a}_k^{(k)}\right\|$
2. $h_k = \tilde{a}_k^{(k)} - \alpha_k e_k$
3. $\varepsilon := \|h_k\|^2 = 2\left(\left\|\tilde{a}_k^{(k)}\right\|^2 + \left|\tilde{a}_{kk}^{(k)}\right|\left\|\tilde{a}_k^{(k)}\right\|\right)$
4. $\tilde{Q}^{(k)}\tilde{a}_j^{(k)} = \tilde{a}_j^{(k)} - \frac{2}{\varepsilon}h_k h_k^T \cdot a_j^{(k)}$

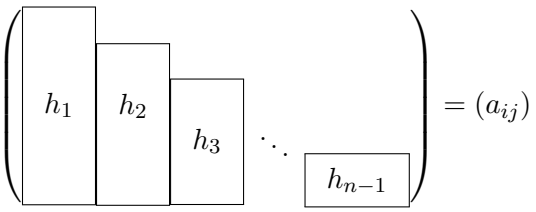
Statt ε berechne für den Algorithmus $\beta = \alpha_k\left(a_{kk}^{(k)} - \alpha_k\right)$ und damit $\tilde{Q}^{(k)}\tilde{a}_j^{(k)} = a_j^{(k)} + \frac{h_k \cdot a_j^{(k)}}{\beta}h_k$.

4.6 Algorithmus

`input(`$(a_{ij})_{i\in\{1,\dots,m\},j\in\{1,\dots,n\}}$`)`
`for (`$k = 1,\dots,\min(n, m-1)$`) {`
    $\alpha_k = a_{kk}^2$

```
for (j = k+1, ..., m)
        {α_k = α_k + a_jk^2}
α_k = −sgn(a_kk) √α_k
a_kk = a_kk − α_k
β = α_k a_kk
for (i = k+1, ..., n) {
        γ = a_kk · a_ki
        for (j = k+1, ..., m)
                {γ = γ + a_jk · a_ji}
        δ = γ/β
        for (j = k, ..., m)
                {a_ji = a_ji + δ · a_jk}
        }
}
output ((a_ij)_{i∈{1,...,m}, j∈{1,...,n}}, (α)_{k=1,...,min(n,m−1)})
```

Die QR-Zerlegung nach dem Algorithmus kann wie folgt abgespeichert werden:



h_k geht bis zur Diagonale hinauf. Über der Diagonale wir R in die Matrix (a_{ij}) gespeichert. Die Diagonale von R wird in einem Extravektor $(\alpha_k)_{k\in\{1,\ldots,\min(n,m-1)\}}$ gespeichert, also $\alpha_k = r_{kk}$.

4.7 Aufwand der QR-Zerlegung

Sei $p = \min(n, m-1)$. Der Aufwand ist:

- Additionen:

$$\sum_{k=1}^{p} ((m-k-1)+1+(n-k-1)((m-k-1)+(m-k))) =$$
$$= \sum_{k=1}^{p} (m-k+(n-k-1)(2(m-k)-1))$$

- Multiplikationen:

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

- Wurzeln: p

Für $n = m$, also $p = m - 1 = n - 1$, lässt sich der Aufwand mit Hilfe von Summenformeln berechnen. Die Anzahl von Additionen und Multiplikationen ist:

$$\begin{aligned}
&\sum_{k=1}^{p} \left(2\left(m-k\right)+3+4\left(n-k-1\right)\left(m-k\right)\right) = \\
&\qquad = \sum_{k=1}^{n-1} \left(3+2\left(2\left(n-k\right)-1\right)\left(n-k\right)\right) = \\
&\qquad = 3\left(n-2\right) + \sum_{k=1}^{n-1} 4\left(n-k\right)^2 - 2\left(n-k\right) \overset{l=n-k}{=} 3\left(n-2\right) + \sum_{l=1}^{n-1} 4l^2 - 2l = \\
&\qquad = 3\left(n-2\right) + 4\frac{\left(n-1\right)n\left(2\left(n-1\right)+1\right)}{6} - 2\frac{\left(n-1\right)\left(n-2\right)}{2} = \\
&\qquad = \left(4-n\right)\left(n-2\right) + \frac{2}{3}\left(n-1\right)n\left(2n-1\right) = \\
&\qquad = \left(4n - n^2 - 8 + 2n\right) + \frac{2}{3}\left(2n^2 - 2n - n + 1\right)n = \\
&\qquad = -n^2 + 6n - 8 + \frac{4}{3}n^3 - 2n^2 + \frac{2}{3}n = \\
&\qquad = \frac{4}{3}n^3 - 3n^2 + \frac{20}{3}n - 8 = \frac{4}{3}n^3 + \mathcal{O}_\infty\left(n^2\right)
\end{aligned}$$

Die Anzahl der Wurzeln ist $n - 1$.

Ist $m \gg n$, so gilt für den Aufwand in flops, also Multiplikationen und Additionen:

$$\begin{aligned}
&\sum_{k=1}^{p} \left(2\left(m-k\right)+3+4\left(n-k-1\right)\left(m-k\right)\right) = \\
&\qquad = \sum_{k=1}^{n} \left(3+2\left(2\left(n-k\right)-1\right)\left(m-k\right)\right) = \\
&\qquad = 3\left(n-1\right) + \sum_{k=1}^{n} 4\left(n-k\right)\left(m-k\right) - 2\left(m-k\right) = \\
&\qquad = 3\left(n-1\right) + \sum_{k=1}^{n} 4nm - 4k\left(m+n\right) + 4k^2 - 2m - 2k = \\
&\qquad = 3\left(n-1\right) + \sum_{k=1}^{n} 4k^2 - 2k\left(2\left(m+n\right)+1\right) + 4nm - 2m = \\
&\qquad = \left(3+2m\left(2n-1\right)\right)\left(n-1\right) + 4\frac{n\left(n+1\right)\left(2n+1\right)}{6} - 2\frac{n\left(n+1\right)}{2}\left(2\left(m+n\right)+1\right) = \\
&\qquad\qquad = 2m\left(2n-1\right)\left(n-1\right) - 2n\left(n+1\right)m - 2n\left(n+1\right)n + \\
&\qquad\qquad\quad + \frac{2}{3}n\left(n+1\right)\left(2n+1\right) + 3\left(n-1\right) - n\left(n+1\right) = \\
&\qquad\qquad = 2m\left(2n^2 - n - 2n + 1 - n^2 - n\right) - 2n^3 + 2n^2 + \\
&\qquad\qquad\quad + \frac{4}{3}n^3 + \frac{6}{3}n^2 + \frac{2}{3}n + 3n - 3 - n^2 - n = \\
&\qquad\qquad = 2m\left(n^2 - 4n + 1\right) - \frac{2}{3}n^3 + 3n^2 + \frac{8}{3}n - 3 = 2mn^2 + o_\infty\left(mn^2\right)
\end{aligned}$$

4.8 Konstruktion orthogonaler Matrizen

Sei $Q \in \mathbb{R}^{n\times n}$ orthogonal. Dann gilt $K_2(Q) = 1$. Dabei sei:

$$K_2(Q) = |||Q|||_2 \cdot \left|\left|\left|Q^{-1}\right|\right|\right|_2$$

Beweis

Da Q orthogonal ist, gilt dies auch für Q^{-1}. Für $x \in \mathbb{R}^n$ gilt somit $\|Qx\|_2 = \|x\|_2 = \left\|Q^{-1}x\right\|_2$. Damit folgt:

$$|||Q|||_2 = \max_{x\in\mathbb{R}^n} \frac{\|Qx\|_2}{\|x\|_2} = \max_{x\in\mathbb{R}^n} \frac{\|x\|_2}{\|x\|_2} = 1$$

$$\left|\left|\left|Q^{-1}\right|\right|\right|_2 = \max_{x\in\mathbb{R}^n} \frac{\left\|Q^{-1}x\right\|_2}{\|x\|_2} = \max_{x\in\mathbb{R}^n} \frac{\|x\|_2}{\|x\|_2} = 1$$

Also gilt $K_2(Q) = 1 \cdot 1 = 1$. $\square_{4.8}$

Orthogonale Matrizen sind also sehr gut konditioniert. Die QR-Zerlegung ist daher stabiler als die LR-Zerlegung.

4.9 Methode der kleinsten Fehlerquadrate

Gegeben seien m Messpunkte $(t_i, b_i) \in \mathbb{R}^2$ für $i \in \{1, \ldots, m\}$, die Zustände eines Objektes in verschiedenen Situationen t_i beschreibt.

Annahme: Den Messungen liegt eine Gesetzmäßigkeit zu Grunde:

$$b(t) = \varphi(t; x_1, \ldots, x_n) \tag{4.1}$$

Dabei sind $x_1, \ldots, x_n \in \mathbb{R}$ die n unbekannten freien Parameter.

Falls es Messfehler gibt, wird für $i \in \{1, \ldots, m\}$ nur gelten:

$$b_i \approx \varphi(t_i; x_1, \ldots, x_n)$$

Ziel: Bestimme die Parameter $x_1, \ldots, x_n$ so, dass die Abstände zwischen den b_i und $\varphi(t_i; x_1, \ldots, x_n)$ möglichst klein werden.

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \qquad f(x_1, \ldots, x_n) = \begin{pmatrix} f_1 \\ \vdots \\ f_m \end{pmatrix} := \begin{pmatrix} \varphi(t_1; x_1, \ldots, x_n) \\ \vdots \\ \varphi(t_m; x_1, \ldots, x_n) \end{pmatrix}$$

Bestimme nun $x_1, \ldots, x_n$ so, dass $\|b - f(x_1, \ldots, x_n)\|$ minimal wird.

Frage: Welche Norm auf $\mathbb{R}^n$ sollen wir wählen?

Wir nehmen die euklidische, da die Wahrscheinlichkeitstheorie zeigt, dass dies eine gute Wahl ist, falls die Fehler normalverteilt sind. Zudem lässt sich

$$\|b - f(x_1, \ldots, x_n)\|^2$$

einfach differenzieren.

Betrachte die Funktion:

$$b(t) = \varphi(t; x_1, \ldots, x_n) = a_1(t)\, x_1 + \ldots + a_n(t)\, x_n$$

Die a_k können auch nichtlinear in t sein.

4.10 Lineare Ausgleichsprobleme

Es sei $\varphi(t; x_1, \ldots, x_n) = a_1(t)\, x_1 + \ldots + a_n(t)\, x_n$ mit $a_i : \mathbb{R} \to \mathbb{R}$ für $i \in \{1, \ldots, n\}$.
Problem (P): Gesucht sind $x_1, \ldots, x_n \in \mathbb{R}$, sodass

$$\|b - f(x_1, \ldots, x_n)\|^2 = \sum_{i=1}^{m} \left(b_i - \sum_{j=1}^{n} a_j(t_i)\, x_j \right)^2$$

minimal wird. (Dabei ist $\|.\|$ die euklidische Norm.)
Wir führen die Matrix $A = (a_{ij})_{i \in \{1,\ldots,m\}, j \in \{1,\ldots,n\}} \in \mathbb{R}^{m \times n}$ mit $a_{ij} := a_j(t_i)$ ein. (P) kann umformuliert werden:
Problem (P'): Gesucht ist ein Vektor $x \in \mathbb{R}^n$, sodass $\|b - Ax\|^2$ minimal ist. Rechnung dazu:

$$\|b - Ax\|^2 = \sum_{i=1}^{m} \left(b_i - \sum_{j=1}^{n} a_{ij} x_j \right)^2 = \sum_{i=1}^{m} \left(b_i - \sum_{j=1}^{n} a_j(t_i)\, x_j \right)^2$$

Ab jetzt sei $m \geq n$, das heißt es liegen mehr Messungen vor, als es Parameter gibt.

4.11 Geometrische Interpretation des linearen Ausgleichsproblems

(P') heißt: Suche den Punkt im Bild von A, der zu b minimalen Abstand hat.
Beispiel für $m = 2$ und $n = 1$:

$$\begin{aligned} A : \mathbb{R} &\to \mathbb{R}^2 \\ x &\mapsto \begin{pmatrix} a_{11} x \\ a_{21} x \end{pmatrix} \end{aligned}$$

Das Bild von A ist eine Gerade. Die Geometrische Anschauung ist, dass $b - Ax$ senkrecht auf $\operatorname{im}(A)$ steht.
TODO: Abb3 einfügen
Betrachte dies im allgemeinen Kontext.

4.12 Satz (Projektionssatz)

Sei V ein reeller Vektorraum mit Skalarprodukt $\langle ., . \rangle$, sei $U \subseteq V$ ein endlich-dimensionaler Unterraum und sei

$$U^{\perp} := \left\{ v \in V \,\middle|\, \langle v, u \rangle = 0 \quad \underset{u \in U}{\forall} \right\}$$

sein orthogonales Komplement in V. Weiter sei $\|v\| := \sqrt{\langle v, v \rangle}$ die vom Skalarprodukt induzierte Norm.

1. Dann existiert zu jedem $v \in V$ genau ein $Pv \in U$, sodass gilt:

$$\|v - Pv\| = \min_{u \in U} \|v - u\|$$

TODO: Abb4 einfügen

2. $P : V \to U$ ist linear und $v - Pv \in U^\perp$.
3. Zu jedem $v \in V$ gibt es genau ein $\overline{u} \in U$ mit der Eigenschaft $v - \overline{u} \in U^\perp$, das heißt es gilt $Pv = \overline{u}$.

Bemerkung: Zur Anschauung betrachte $V = \mathbb{R}^n$ mit euklidischem Skalarprodukt.

Beweis

Sei $v \in V$.

a) **Behauptung:** Falls für $\overline{u} \in U$ schon

$$\langle v - \overline{u}, u \rangle = 0 \qquad \underset{u \in U}{\forall}$$

gilt, so ist $\overline{u}$ eindeutig bestimmtes Minimum:

$$\min_{u \in U} \|v - u\| = \|v - \overline{u}\|$$

Beweis: Für alle $\tilde{u} \in U$ gilt:

$$\begin{aligned} \|v - \tilde{u}\|^2 = \|v - \overline{u} + \overline{u} - \tilde{u}\|^2 = \langle v - \overline{u} + \overline{u} - \tilde{u}, v - \overline{u} + \overline{u} - \tilde{u} \rangle = \\ = \|v - \overline{u}\|^2 + 2 \underbrace{\langle v - \overline{u}, \overline{u} - \tilde{u} \rangle}_{=0} + \|\tilde{u} - \overline{u}\|^2 = \|v - \overline{u}\|^2 + \|\tilde{u} - \overline{u}\|^2 \end{aligned}$$

(Dies ist der Satz des Pythagoras.) Es folgt:

$$\|v - \tilde{u}\|^2 - \|v - \overline{u}\|^2 = \|\tilde{u} - \overline{u}\|^2 \geq 0$$

Also ist $\overline{u}$ das Minimum und eindeutig. $\square_{\text{Behauptung}}$

TODO: Abb5 einfügen

$$(v - \tilde{u}) \perp (\overline{u} - \tilde{u})$$

Existiert ein $\overline{u}$ mit obiger Eigenschaft?

b) **Behauptung:** Es existiert ein $\overline{u} \in U$ sodass für alle $u \in U$ gilt:

$$\langle v - \overline{u}, u \rangle = 0$$

Beweis: Sei $\{u_1, \ldots, u_n\}$ eine Orthonormalbasis von U. Setze

$$\overline{u} = \sum_{i=1}^{n} \alpha_i u_i$$

mit $\alpha_i \in \mathbb{R}$, womit folgt:

$$\begin{aligned} & \langle v - \overline{u}, u \rangle = 0 \qquad \underset{u \in U}{\forall} \\ \Leftrightarrow \quad & \langle v, u_j \rangle = \sum_{i=1}^{n} \alpha_i \underbrace{\langle u_i, u_j \rangle}_{=\delta_{ij}} \qquad \underset{j}{\forall} \\ & \langle v, u_j \rangle = \alpha_j \end{aligned}$$

Also gilt:

$$\overline{u} = \sum_{i=1}^{n} \langle v, u_i \rangle \, u_i$$

Definiere nun $Pv = \overline{u}$, sodass gilt:

$$\langle v - Pv, u \rangle = 0 \quad \underset{u \in U}{\forall}$$

Das heißt $v - Pv \in U^{\perp}$.
Zu zeigen ist die Linearität von P. Seien v_1 und $v_2 \in V$.

$$\begin{aligned} & \langle v_i - Pv_i, u \rangle = 0 \quad \underset{u \in U}{\forall}; \; i \in \{1, 2\} \\ \Rightarrow \quad & \langle v_1 + v_2 - (Pv_1 + Pv_2), u \rangle = 0 \quad \underset{u \in U}{\forall} \\ \Rightarrow \quad & P(v_1 + v_2) = Pv_1 + Pv_2 \end{aligned}$$

Dies folgt aus der Eindeutigkeit von $P(v_1 + v_2)$. $\square_{\text{Behauptung}}$

$\square_{4.12}$

4.13 **Bemerkung** (orthogonale Projektion)

Die Abbildung $P : V \to U$ heißt *orthogonale Projektion* von V auf U. Es gilt $P^2 = P$.

4.14 **Lemma und Definition** (Normalengleichungen)

Die Voraussetzungen seien wie in 4.12.
Sei $v \in V$ und sei $\overline{u} = Pv \in U$. Sei nun $\{w_1, \ldots, w_n\}$ eine Basis (nicht notwendig orthogonal) von U und sei:

$$\overline{u} = \sum_{i=1}^{n} \beta_i w_i$$

Dann erfüllen die β_i folgendes lineare Gleichungssystem, *Normalengleichungen* genannt:

$$\sum_{i=1}^{n} \beta_i \langle w_i, w_j \rangle = \langle v, w_j \rangle$$

Hintergrund des Begriffs:
TODO: Abb6 einfügen

Beweis

$\overline{u}$ erfüllt für alle $j \in \{1, \ldots, n\}$:

$$0 = \langle v - \overline{u}, w_j \rangle = \langle v, w_j \rangle - \sum_{i=1}^{n} \beta_i \langle w_i, w_j \rangle$$

$\square_{4.14}$

4.15 Satz (Existenz von Lösungen zum linearen Ausgleichsproblem)

Sei $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ und $m \geq n$. Dann existiert eine Lösung $\overline{x} \in \mathbb{R}^n$ des linearen Ausgleichsproblems:

$$\|A\overline{x} - b\|^2 = \min_{x\in\mathbb{R}^n} \|Ax - b\|^2$$

x erfüllt die Normalengleichungen:

$$A^{\mathrm{T}}A\overline{x} = A^{\mathrm{T}}b$$

Die Lösung ist genau dann eindeutig bestimmt, wenn A vollen Rang hat.

Beweis

$$\min_{x\in\mathbb{R}^n} \|Ax - b\|^2 = \min_{y\in\mathrm{im}(A)} \|y - b\|^2$$

Der Projektionssatz mit $U = \mathrm{im}\,(A)$ bedeutet, dass das Minimum-Problem eine eindeutige Lösung $\overline{y} \in \mathrm{im}\,(A)$ hat mit folgender Eigenschaft:

$$\langle \overline{y} - b, y\rangle = 0 \qquad \forall_{y\in\mathrm{im}(A)}$$

Da $\overline{y} \in \mathrm{im}\,(A)$ ist, gibt es ein $\overline{x} \in \mathbb{R}^n$ mit $A\overline{x} = \overline{y}$. Wegen $y \in \mathrm{im}\,(A)$ gibt es ein $x \in \mathbb{R}^n$ mit $Ax = y$.
Wegen $\langle \overline{y} - b, y\rangle = 0$ gilt:

$$\begin{aligned} & \langle A\overline{x} - b, Ax\rangle = 0 \qquad \forall_{x\in\mathbb{R}^n} \\ \Leftrightarrow \quad & \left\langle A^{\mathrm{T}}A\overline{x} - A^{\mathrm{T}}b, x\right\rangle = 0 \qquad \forall_{x\in\mathbb{R}^n} \\ \Leftrightarrow \quad & A^{\mathrm{T}}A\overline{x} = A^{\mathrm{T}}b \end{aligned}$$

Hat die Matrix A vollen Rang, so ist A injektiv. Also gibt es genau ein $\overline{x} \in \mathbb{R}^n$ mit $A\overline{x} = \overline{y}$ und damit folgt die Eindeutigkeit. $\square_{4.15}$

4.16 Lösung linearer Ausgleichsprobleme mittels QR-Zerlegung

Sei $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ und $m \geq n$. Gesucht ist ein $\overline{x} \in \mathbb{R}^n$ mit:

$$\|A\overline{x} - b\|^2 = \min_{x\in\mathbb{R}^n} \|Ax - b\|^2$$

Sei nun $A = QR$ die QR-zerlegung von A. R hat die Form:

$$\begin{pmatrix} * & \dots & \dots & * \\ 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & * \\ 0 & \dots & \dots & 0 \\ \vdots & & & \vdots \\ 0 & \dots & \dots & 0 \end{pmatrix} = \begin{pmatrix} & \overbrace{\qquad}^{\in\mathbb{R}^{n\times n}} & \\ & R_1 & \\ & & \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{pmatrix}$$

Dann gilt

$$\|Ax-b\|^2 = \|QRx-b\|^2 \overset{Q^{\mathrm{T}} \text{ orthogonal}}{=} \left\|Q^{\mathrm{T}}\left(QRx-b\right)\right\|^2 = \left\|Rx-Q^{\mathrm{T}}b\right\|^2$$

und:

$$Rx = \begin{pmatrix} R_1 x \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Wähle $b_1 \in \mathbb{R}^n$ und $b_2 \in \mathbb{R}^{m-n}$ mit:

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = Q^{\mathrm{T}}b$$

Es folgt:

$$\|Ax-b\|^2 = \underbrace{\|R_1x-b_1\|^2}_{\text{Norm auf } \mathbb{R}^n} + \underbrace{\|b_2\|^2}_{\text{Norm auf } \mathbb{R}^{m-n}}$$

Falls A maximalen Rang hat, so hat auch R_1 maximalen Rang und somit ist die $n \times n$-Matrix R_1 invertierbar. Also ist

$$x = R_1^{-1}b_1$$

ein Minimum für $\|Ax-b\|$.

Satz: Hat die Matrix A vollen Rang und besitzt

$$A = QR = Q\begin{pmatrix} R_1 \\ 0 \end{pmatrix}$$

als QR-Zerlegung, so besitzt das lineare Ausgleichsproblem bezüglich $b \in \mathbb{R}^m$ die eindeutige Lösung:

$$x = R_1^{-1}b_1$$

Dabei ist mit $b_1 \in \mathbb{R}^n$ und $b_2 \in \mathbb{R}^{m-n}$:

$$Q^{\mathrm{T}}b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

4.17 Vorgehen bei der Lösung von linearen Ausgleichsproblemen mit QR-Zerlegung

Seien $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ gegeben.

1. Bestimme QR-Zerlegung von A: Führe den Algorithmus 4.6 für $R = \begin{pmatrix} R_1 \\ 0 \end{pmatrix}$ mit $R_1 \in \mathbb{R}^{n\times n}$ durch.
2. Berechne $(b_1)_{i\in\{1,\dots,n\}} = \left(Q^{\mathrm{T}}b\right)_{i\in\{1,\dots,n\}}$.
3. Löse das gestaffelte Gleichungssystem $R_1x = b_1$. (vergleiche Kapitel 2)

x ist die gesuchte Lösung.

4.18 Bemerkung

Es gibt noch eine andere Möglichkeit $\overline{x}$ zu berechnen:
Löse die Normalengleichungen ($(n \times n)$-Gleichungssystem):

$$A^{\mathrm{T}} A\overline{x} = A^{\mathrm{T}} b$$

$A^{\mathrm{T}}A$ ist symmetrisch und das Cholesky-Verfahren kann angewandt werden. Die LR-Zerleung liefert ebenfalls eine Lösung.
Aber: Zur Berechnung von $A^{\mathrm{T}}A$ müssen Skalarprodukte von Spaltenvektoren berechnet werden, was oft schlecht konditioniert ist. Das Verfahren mit QR-Zerlegung ist stabiler („weniger Fehler“).

4.19 QR-Zerlegung mit Givens-Rotation

Ziel: Leite eine QR-Zerlegung mit Hilfe von Drehungen her.
Finde eine Drehungsmatrix Q, sodass gilt:

$$QA = Q \begin{pmatrix} a & b \\ e & d \end{pmatrix} = \begin{pmatrix} r & * \\ 0 & * \end{pmatrix} \tag{4.2}$$

Drehungen lassen sich wie folgt schreiben:

$$Q = \begin{pmatrix} \cos(\varphi) & \sin(\varphi) \\ -\sin(\varphi) & \cos(\varphi) \end{pmatrix}$$

Setzte nun $c := \cos(\varphi)$ und $s := \sin(\varphi)$. Es gilt $c^2 + s^2 = 1$. Es muss laut (4.2) gelten:

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} a \\ e \end{pmatrix} = \begin{pmatrix} r \\ 0 \end{pmatrix}$$

Wegen $r^2 = a^2 + e^2$ (Länge bleibt erhalten) folgt:

$$r = \pm\sqrt{a^2 + e^2} \qquad c = \frac{a}{r} \qquad s = \frac{e}{r}$$

Im allgemeiner Fall ist die Matrix von der Form $A \in \mathbb{R}^{m \times n}$. Wie rotiere ich?
Antwort: Bette die Givens-Rotation in den $\mathbb{R}^n$ ein.

$$G_{i,k} = \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & c & \dots & s & & & \\ & & & \vdots & 1 & \vdots & & & \\ & & & -s & \dots & c & & & \\ & & & & & & 1 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 1 \end{pmatrix}$$

$G_{i,k}$ realisiert Drehung in der von e_1 und e_k aufgespannten Ebene. Es gilt:

$$G_{i,k}\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_{i-1} \\ r \\ x_{i+1} \\ \vdots \\ x_{k-1} \\ 0 \\ x_{k+1} \\ \vdots \\ x_n \end{pmatrix}$$

$$r = \pm\sqrt{x_i^2 + x_k^2} \qquad c = \frac{x_i}{r} \qquad s = \frac{x_k}{r}$$

Beispiel

$$G_{1,2} = \begin{pmatrix} \frac{4}{5} & -\frac{3}{5} & 0 \\ \frac{3}{5} & \frac{4}{5} & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad G_{1,2}\begin{pmatrix} 4 \\ -3 \\ 1 \end{pmatrix} = \begin{pmatrix} 5 \\ 0 \\ 1 \end{pmatrix}$$

$$G_{1,3} = \begin{pmatrix} \frac{5}{\sqrt{26}} & 0 & \frac{1}{\sqrt{26}} \\ 0 & 1 & 0 \\ -\frac{1}{\sqrt{26}} & 0 & \frac{5}{\sqrt{26}} \end{pmatrix} \qquad G_{1,3}\begin{pmatrix} 5 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \sqrt{26} \\ 0 \\ 0 \end{pmatrix}$$

Wir schreiben diese Vorgehen wie folgt:

$$\begin{pmatrix} 4 \\ -1 \\ 3 \end{pmatrix} \xrightarrow{G_{1,2}} \begin{pmatrix} 5 \\ 0 \\ 1 \end{pmatrix} \xrightarrow{G_{1,3}} \begin{pmatrix} \sqrt{26} \\ 0 \\ 0 \end{pmatrix}$$

Im Hinblick auf eine speicherschonende Implementierung ist es wichtig, $G_{i,k}$ durch eine Zahl kodieren zu können.

$$\varrho = \varrho_{i,k} = \begin{cases} 1 & c = 0 \\ \frac{1}{2}\operatorname{sgn}(c)\, s & |s| < |c| \\ 2 \cdot \frac{\operatorname{sgn}(s)}{c} & |c| \leq |s| \end{cases}$$

Dekodierung:

$$(s, c) = \begin{cases} (1, 0) & \varrho = 1 \\ \left(2\varrho, \sqrt{1 - (2\varrho)^2}\right) & |\varrho| < 1 \\ \left(\sqrt{1 - \left(\frac{2}{\varrho}\right)^2}, \frac{2}{\varrho}\right) & |\varrho| > 1 \end{cases}$$

Wende nacheinander Givens-Rotationen an, um die Einträge zu erzeugen.
Achtung: Zerstöre keine Null-Einträge die schon vorher erzeugt wurden.
Wir bekommen Givens-Rotationen $G_{i_N,k_N}, \ldots, G_{i_1,k_1}$, sodass gilt:

$$\begin{aligned} G_{i_N,k_N} \cdot \ldots \cdot G_{i_1,k_1} &= R \\ A &= \underbrace{G_{i_1,k_1}^T \cdot \ldots \cdot G_{i_N,k_N}^T}_{=Q} \cdot R \end{aligned}$$

Bemerkung

Q nie explizit ausrechnen!
Viele Matrizen in der Anwendung sind dünn besetzt, das heißt ganz viele Einträge sind Null. In diesem Fall braucht man nur noch wenige Rotationen um eine obere Dreiecksgestalt zu erhalten.

Bemerkung zu Givens-Rotationen

Speicherung von

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix}$$

mit Hilfe von ϱ kann nach Dekodierung auch

$$-\begin{pmatrix} c & s \\ -s & c \end{pmatrix}$$

liefern. Dies ist kein Problem. Auch diese Matrix sorgt für die richtige Rotation.

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} r \\ 0 \end{pmatrix}$$

Die andere Matrix ändert nur das Vorzeichen von r.

5 Numerische Lösung nichtlinearer Gleichungssysteme

5.1 Problemstellung

Beispiele

1. Nullstellenbestimmung
   Gegeben: Eine Definitionsmenge $D \subset \mathbb{R}^n$ und eine stetige Funktion $f : D \to \mathbb{R}^m$.
   Gesucht: Eine Nullstelle $x \in D$ mit $f(x) = 0$.
2. Fixpunktprobleme
   Gegeben: Eine Definitionsmenge $D \subset \mathbb{R}^n$ und eine stetige Funktion $g : D \to \mathbb{R}^n$.
   Gesucht: Ein Fixpunkt$x \in D$ mit $g(x) = x$.
   Jede Fixpunktgleichung kann in eine Nullstellengleichung überführt werden und umgekehrt, indem man setzt:

$$f(x) = g(x) - x$$

5.2 Beispiele

a) Bestimme die Nullstellen eines Polynoms $f(x) = \sum_{i=0}^{m} a_i x^i$ mit $a_i \in \mathbb{R}$. Suche $x \in \mathbb{R}$ mit:

$$f(x) = \sum_{i=0}^{m} a_i \cdot x^i = 0$$

Zum Beispiel bestimme $x \in \mathbb{R}$ mit $x^2 - 5 = 0$.

b) Sei $h : \mathbb{R}^n \to \mathbb{R}$ eine Funktion. Gesucht ist das Minimum x_0 von h. Dann erfüllt x_0 die Gleichung:

$$f(x_0) := \nabla h(x) = 0$$

c) Keplersche Gleichung
Gesucht: Ein $x \in \mathbb{R}$ mit $x = e \cdot \sin(x) + a$.
Gesucht: Ein Fixpunkt x mit $x = g(x) := e \cdot \sin(x) + a$.

d) Diskretisierung von Randwertproblemen
Sei $I = [a, b]$ ein Intervall.
Gesucht: $u : [a, b] \to \mathbb{R}$ mit den Eigenschaften:

$$u''(x) = f(u(x)) \qquad \underset{x \in [a,b]}{\forall}$$

$$u(a) = u_a$$
$$u(b) = u_b$$

Dabei seien $u_a, u_b \in \mathbb{R}$ sowie die stetige Funktion $f : \mathbb{R} \to \mathbb{R}$ gegeben. Sei $N \in \mathbb{N}$. Definiere für $i \in \{0, \ldots, N+1\}$:

$$h := \frac{b-a}{N+1}$$
$$x_i := a + i \cdot h$$

Für $i \in \{1, \ldots, N\}$ kann man die zweite Ableitung näherungsweise berechnen:

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

Ersetze die Differentialgleichung $u''(x) = f(u(x_i))$ mit den Punkten x_i durch folgendes Gleichungssystem für $i \in \{1, \ldots, N\}$:

$$\frac{u(x_{i+1}) - 2u(x_1) + u(x_{i-1})}{h^2} \approx f(u(x_i))$$

Führe den Vektor

$$u = \begin{pmatrix} u_1 \\ \vdots \\ u_N \end{pmatrix}$$

ein und betrachte das Gleichungssystem:

$$\begin{pmatrix} 2 & -1 & 0 & \ldots & 0 \\ -1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \ldots & 0 & -1 & 2 \end{pmatrix} u = - \begin{pmatrix} h^2 f(u_1) - u_a \\ h^2 f(u_2) \\ \vdots \\ h^2 f(u_{n-1}) \\ h^2 f(u_n) - u_b \end{pmatrix}$$

Dies ist eine nichtlineares Gleichungssystem für die u_i, die $u(x_i)$ annähern. (siehe Vorlesung über Numerik von Differentialgleichungen).

5.3 Bisektionsverfahren

Sei $f : [a, b] \to \mathbb{R}$ stetig. Es gelte $f(a) \cdot f(b) < 0$. Nach dem Mittelwertsatz existiert mindestens eine Nullstellen $\bar{x} \in \mathbb{R}$ mit $f(\bar{x}) = 0$. Ohne Beschränkung der Allgemeinheit sei $f(a) < 0$ und $f(b) > 0$.
Iterationsverfahren: Halbiere in jedem Schritt das Intervall. Setze $x_-^{(0)} = a$ und $x_+^{(0)} = b$. Für $k = \mathbb{N}_{\geq 0}$ berechne:

$$x_m^{(k)} = \frac{x_+^{(k)} + x_-^{(k)}}{2}$$

Ist $f\left(x_m^{(k)}\right) = 0$ ist die Nullstelle gefunden und man kann aufhören. Sonst berechne:

$$x_+^{(k+1)} := \begin{cases} x_+^{(k)} & \text{falls } f\left(x_m^{(k)}\right) < 0 \\ x_m^{(k)} & \text{falls } f\left(x_m^{(k)}\right) > 0 \end{cases}$$
$$x_-^{(k+1)} := \begin{cases} x_m^{(k)} & \text{falls } f\left(x_m^{(k)}\right) < 0 \\ x_-^{(k)} & \text{falls } f\left(x_m^{(k)}\right) > 0 \end{cases}$$

Die Folgen $\left(x_+^{(k)}\right)_{k\in\mathbb{N}}$ und $\left(x_-^{(k)}\right)_{k\in\mathbb{N}}$ sind monoton und beschränkt, weshalb beide konvergieren. Aus dem Zwischenwertsatz folgt, dass die Folgen $\left(x_+^{(k)}\right)_{k\in\mathbb{N}}$ und $\left(x_-^{(k)}\right)_{k\in\mathbb{N}}$ denselben Grenzwert $\bar{x} \in \mathbb{R}$ haben. Weiter gilt

$$\left|x_\pm^{(k)} - \overline{x}\right| \leq x_+^{(k)} - x_-^{(k)} \leq 2^{-k}\,|b-a|$$

und:

$$f(\bar{x}) = \lim_{k\to\infty} f\left(x_+^{(k)}\right) = \lim_{k\to\infty} f\left(x_-^{(k)}\right)$$

Wegen $f\left(x_+^{(k)}\right) > 0$ und $f\left(x_-^{(k)}\right) < 0$ bedeutet dies $f(\overline{x}) = 0$:

$$f(\overline{x}) \overset{f \text{ stetig}}{=} \lim_{k\to\infty} \underbrace{f\left(x_+^{(k)}\right)}_{>0} \geq 0$$
$$\overset{f \text{ stetig}}{=} \lim_{k\to\infty} \underbrace{f\left(x_-^{(k)}\right)}_{<0} \leq 0$$
$$\Rightarrow \quad f(\overline{x}) = 0$$

5.4 Konvergenzordnung

Eine Folge $\left(x^{(k)}\right)_{k\in\mathbb{N}}$ mit $x^{(k)} \in \mathbb{R}^n$ konvergiert mit der Ordnung $p \in \mathbb{R}_{\geq 1}$ gegen $\overline{x}$, falls $x^{(k)} \to \overline{x}$ für $k \to \infty$ konvergiert und falls ein $c \in \mathbb{R}_{>0}$ existiert, sodass gilt:

$$\left\|x^{(k+1)} - \overline{x}\right\| \leq c \left\|x^{(k)} - \overline{x}\right\|^p$$

Im Fall $p = 1$ setzen wir zusätzlich $c < 1$ voraus und nennen die Konvergenz *linear*.
Im Fall $p = 2$ sprechen wir von quadratischer Konvergenz.
Falls

$$\lim_{k\to\infty} \frac{\left\|x^{(k+1)} - \overline{x}\right\|}{\left\|x^{(k)} - \overline{x}\right\|} = 0$$

gilt, so sprechen wir von *superlinearer* Konvergenz.

5.5 Fixpunktiteration

Gesucht sei ein Fixpunkt $x \in D \subseteq \mathbb{R}^n$ der stetigen Funktion $g : D \to \mathbb{R}^n$, das heißt $g(x) = x$.
Ansatz: Benutze einfache Iteration, das heißt wähle ein $x^{(0)} \in D$ und definiere $x^{(k+1)} = g\left(x^{(k)}\right)$ für $k \in \mathbb{N}_0$.
Falls $x^{(k)} \to \overline{x} \in D$ konvergiert, so gilt:

$$\overline{x} = \lim_{k\to\infty} x^{(k+1)} = \lim_{k\to\infty} g\left(x^{(k)}\right) = g(\overline{x})$$

Beispiel 1:
TODO: Abb7 einfügen
Beispiel 2:
TODO: Abb8 einfügen

5.6 Kontraktion

Sei $D \subseteq \mathbb{R}^n$ abgeschlossen und $\|.\|$ eine Norm auf $\mathbb{R}^n$. Eine Abbildung $g : D \to \mathbb{R}^n$ heißt *Kontraktion* (bezüglich $\|.\|$), falls ein $\kappa \in (0,1)$ existiert mit:

$$\|g(u) - g(v)\| \leq \kappa \|u - v\| \qquad \underset{u,v\in D}{\forall}$$

Die Zahl κ heißt *Kontraktionszahl* von g.
TODO: Abb9 einfügen

5.7 Banachscher Fixpunktsatz

Zusätzlich zu den Eigenschaften in 5.6 gelte $g(D) \subseteq D$. Dann gilt:

1. g hat genau einen Fixpunkt.
2. Für alle $x^{(0)} \in D$ gilt: Die Folge $\left(x^{(k)}\right)_{k\in\mathbb{N}}$ definiert durch $x^{(k+1)} = g\left(x^{(k)}\right)$ konvergiert gegen $\overline{x}$.
3. Es gilt:

$$\left\|x^{(k)} - \overline{x}\right\| \leq \frac{\kappa}{1-\kappa} \cdot \left\|x^{(k)} - x^{(k-1)}\right\|$$

   Dies ist eine *a posteriori* Fehlerabschätzung. Es gibt auch eine *a priori* Abschätzung:

$$\left\|x^{(k)} - \overline{x}\right\| \leq \frac{\kappa^k}{1-\kappa} \left\|x^{(1)} - x^{(0)}\right\|$$

Beweis

Es gilt:

$$\left\|x^{(k+1)} - x^{(k)}\right\| = \left\|g\left(x^{(k)}\right) - g\left(x^{(k-1)}\right)\right\| \overset{\text{Kontraktion}}{\leq} \kappa \left\|x^{(k)} - x^{(k-1)}\right\| \leq \kappa^k \left\|x^{(1)} - x^{(0)}\right\|$$

$$\left\|x^{(k+l)} - x^{(k)}\right\| \leq \sum_{i=0}^{l-1} \left\|x^{(k+i+1)} - x^{(k+i)}\right\| \leq \sum_{i=0}^{l-1} \kappa^i \left\|x^{(k+1)} - x^{(k)}\right\| \overset{\text{geo.R.}}{\leq} \frac{\kappa^k}{1-\kappa} \left\|x^{(1)} - x^{(0)}\right\|$$

Also ist $x^{(k)}$ eine Cauchy-Folge und konvergiert daher gegen $\overline{x} \in D$, da D abgeschlossen ist.

$$\overline{x} = \lim_{k\to\infty} x^{(k+1)} = \lim_{k\to\infty} g\left(x^{(k)}\right) \overset{\text{stetig}}{=} g\left(\overline{x}\right)$$

Eindeutigkeit: Seien $\overline{x}$ und $\overline{y}$ zwei Fixpunkte.

$$\|\overline{x} - \overline{y}\| \leq \kappa \|g\left(\overline{x}\right) - g\left(\overline{x}\right)\| \underset{g(\overline{y})=\overline{y}}{\overset{g(\overline{x})=\overline{x}}{=}} \leq \kappa \|\overline{x} - \overline{y}\|$$

Wegen $\kappa < 1$ folgt $\overline{x} = \overline{y}$.

A priori Abschätzung: Es gilt:

$$\left\|x^{(k)} - \overline{x}\right\| = \lim_{k\to\infty} \left\|x^{(k)} - x^{(k+l)}\right\| \leq \frac{\kappa^k}{1-k} \left\|x^{(1)} - x^{(0)}\right\|$$

A posteriori Abschätzung: Es gilt:

$$\left\|x^{(k)} - \overline{x}\right\| = \lim_{k\to\infty} \left\|x^{(k)} - x^{(k+l)}\right\| \leq \left\|x^{(k+1)} - x^{(k)}\right\| \cdot \underbrace{\sum_{i=0}^{l-1} \kappa^i}_{\leq \frac{1}{1-\kappa}} \leq \frac{\kappa}{1-k} \left\|x^{(k+1)} - x^{(k)}\right\|$$

$\square_{5.7}$

5.8 Bemerkungen

1. Das Iterationsverfahren aus dem Banachschen Fixpunktsatz konvergiert linear, denn es gilt:

   $$\left\|x^{(k+1)} - \overline{x}\right\| = \left\|g\left(x^{(k)}\right) - g\left(\overline{x}\right)\right\| \overset{g(\overline{x})=\overline{x}}{\leq} \kappa \left\|x^{(k)} - \overline{x}\right\|$$

   Da $\kappa < 1$ ist, folgt die lineare Konvergenz.

2. Sei $D = \overline{\Omega}$ mit $\Omega \subseteq \mathbb{R}^n$ offen und konvex. Dann gilt:
   Eine stets differenzierbare Abbildung $g : D \to \mathbb{R}^n$ ist kontrahierend, falls gilt:

   $$\sup_{x\in D} |||\mathrm{D}g\left(x\right)||| < 1$$

   Beweis: Sei $h\left(t\right) = g\left(x + t\left(y - x\right)\right)$, das heißt $h\left(1\right) = g\left(y\right)$ und $h\left(0\right) = g\left(x\right)$.

   $$\begin{aligned}
   \|g\left(y\right) - g\left(x\right)\| = \|h\left(1\right) - h\left(0\right)\| = \left\|\int_0^1 h'\left(t\right) \mathrm{d}t\right\| = \left\|\int_0^1 \mathrm{D}g\left(x + t\left(y - x\right)\right)\left(y - x\right) \mathrm{d}t\right\| \leq \\
   \leq \int_0^1 \|\mathrm{D}g\left(x + t\left(y - x\right)\right)\left(y - x\right)\| \,\mathrm{d}t \leq \\
   \leq \|\left(y - x\right)\| \int_0^1 |||\mathrm{D}g\left(x + t\left(y - x\right)\right)||| \,\mathrm{d}t \leq \\
   \leq \|\left(y - x\right)\| \underbrace{\sup_{z\in D} |||\mathrm{D}g\left(z\right)|||}_{<1}
   \end{aligned}$$

   Also folgt, dass g kontrahierend. $\square_{5.8.2}$

3. Das Problem bei der Anwendung des Banachschen Fixpunktsatzes ist zu zeigen, dass $g\left(D\right) \subseteq D$ gilt.

5.9 Praktische Formulierung des Banachschen Fixpunktsatzes

Es sei $g : D \to \mathbb{R}^n$ kontrahierend mit Kontraktionszahl κ (und damit Lipschitz-stetig mit Lipschitz-Konstante κ).
Es seien $x_0 \in D$ und ein $r > 0$ mit:

1. $D_0 := \overline{B_r(x_0)} \subseteq D$
2. $\|g(x_0) - x_0\| \leq (1-\kappa)\, r$

Dann gilt $g(D_0) \subseteq D_0$ und der Banachsche Fixpunktsatz ist mit der Menge D_0 anwendbar, das heißt g besitzt einen Fixpunkt in D_0.

Beweis

Sei $y \in D_0$, so gilt:

$$\|g(y) - x_0\| \leq \|g(y) - g(x_0)\| + \|g(x_0) - x_0\| < \kappa \underbrace{\|y - x_0\|}_{\leq r} + (1-\kappa)\, r \leq r$$

$\square_{5.9}$

5.10 Das Newton-Verfahren in einer Dimension

Sei $f : \mathbb{R} \to \mathbb{R}$ stetig differenzierbar. Dann liegt folgendes Vorgehen nahe, um eine Nullstelle zu bestimmen:

TODO: Abb10 einfügen

- Wähle einen Startwert $x^{(0)}$.
- Lege eine Tangente an die Funktion f im Punkt $\left(x^{(0)}, f\left(x^{(0)}\right)\right)$.
- Berechne die Nullstelle $x^{(1)}$ der Tangente.
- Iteriere dieses Verfahren

Bemerkung: Die Tangente hat genau dann eine Nullstelle, wenn $f'\left(x^{(0)}\right) \neq 0$ ist.

Wie berechnen sich die Iterierten $x^{(k)}$? Die Tangente ist der Graph der Geraden:

$$T(x) = f\left(x^{(0)}\right) + f'\left(x^{(0)}\right)\left(x - x^{(0)}\right)$$

Löse $T(x) = 0$. Die Nullstelle $x^{(1)}$ ist der nächste Schätzwert für die Nullstelle von f:

$$x^{(1)} = x^{(0)} - \left(f'\left(x^{(0)}\right)\right)^{-1} f\left(x^{(0)}\right)$$

Das Verfahren lautet nun:

1. Wähle ein $x^{(0)} \in \mathbb{R}$.
   Für $r \in \mathbb{N}_0$:
2. Berechne $f'\left(x^{(k)}\right)$. Falls $f'\left(x^{(k)}\right) = 0$: Stoppe!
3. Setze $x^{(k+1)} = x^{(k)} - \left(f'\left(x^{(0)}\right)\right)^{-1} f\left(x^{(k)}\right)$ und gehe zu 2.

Bemerkungen

1. Der gewählte Startwert sollte schon eine einigermaßen gute Näherung sein. Betrachte zum Beispiel den Graphen und schätze Nullstelle. Andere Möglichkeit: Verwende erst Bisektionsverfahren bis eine gute Näherung als Startwert für das Newton-Verfahren gefunden ist.
2. Das Newton-Verfahren konvergiert nicht immer. Es können zum Beispiel Oszillationen auftreten.

5.11 Satz

Sei $f \in \mathcal{C}^2((a,b),\mathbb{R})$ und $x^* \in (a,b)$ eine einfache Nullstelle von f, das heißt $f(x^*) = 0$ und $f'(x^*) \neq 0$. Dann gibt es ein $\varepsilon > 0$, sodass für jedes $x^{(0)} \in \overline{B_\varepsilon(x^*)}$ das Newton-Verfahren quadratisch gegen x^* konvergiert .

Beweis

Definiere:

$$g(x) = x - \frac{f(x)}{f'(x)}$$

Das Newton-Verfahren angewendet auf $x^{(0)}$ liefert:

$$x^{(k+1)} = g\left(x^{(k)}\right) = g^{k+1}\left(x^{(0)}\right)$$

Weiter gilt:

$$g'(x) = 1 - \frac{f'(x)\,f'(x) - f(x)\,f''(x)}{f'(x)^2} = \frac{f(x)\,f''(x)}{|f'(x)|^2}$$

Also ist $g'(x^*) = 0$. Weiterhin gibt es eine Umgebung U_0 von x^*, sodass $f'(x) \neq 0$ für alle $x \in U_0$ ist. Da f'' stetig ist, existiert also ein $\varepsilon \in \mathbb{R}_{>0}$ und ein $\kappa \in (0,1)$, sodass $|g'(x)| \leq \kappa < 1$ für alle $x \in \overline{B_\varepsilon(x^*)}$ gilt. Da $g(x^*) = x^*$ gilt, liefert der praktische Fixpunktsatz von Banach 5.9 zunächst

$$g\left(\overline{B_\varepsilon(x^*)}\right) \subset \overline{B_\varepsilon(x^*)}$$

und damit auch die Konvergenz für alle $x^{(0)} \in \overline{B_\varepsilon(x^*)}$.
Quadratische Konvergenz:

$$\begin{aligned}
\left|x^{(k+1)} - x^*\right| &= \left|x^{(k)} - \left(f'\left(x^{(k)}\right)\right)^{-1} f\left(x^{(k)}\right) - x^*\right| = \\
&\overset{f(x^*)=0}{=} \left|f'\left(x^{(k)}\right)\right|^{-1} \left|f(x^*) - f\left(x^{(k)}\right) - f'\left(x^{(k)}\right)\left(x^* - x^{(k)}\right)\right| = \\
&\overset{\text{Taylorreihe}}{=} \left|f'\left(x^{(k)}\right)\right|^{-1} \left|\frac{f''(x)}{2} \cdot \left(x^{(k)} - x^*\right)^2 + \mathcal{O}\left(\left(x^{(k)} - x^*\right)^3\right)\right| = \\
&\overset{\text{Restgliedabschätzung}}{\leq} \sup_{x \in B_\varepsilon(x^*)} \left(\left|f'(x)\right|^{-1}\right) \sup_{x \in B_\varepsilon(x^*)} \left(\left|f''(x)\right| \frac{1}{2} \left|x^{(k)} - x^*\right|^2\right)
\end{aligned}$$

$\square_{5.11}$

5.12 Bemerkung

a) Bei m-fachen Nullstellen

$$f(x^*) = f'(x^*) = ... = f^{(m-1)}(x^*) = 0 \qquad f^{(m)}(x^*) \neq 0$$

liegt im Allgemeinen keine quadratische Konvergenz vor, sondern nur noch lineare. Man kann quadratische Konvergenz erhalten, indem man das Newton-Verfahren modifiziert:

$$x^{(k+1)} = x^{(k)} - m\frac{f\left(x^{(k)}\right)}{f'\left(x^{(k)}\right)}$$

b) Die Ableitung f' muss bekannt sein.

c) Die Lage und Größe des Konvergenzintervalls ist a priori häufig unbekannt.
Mögliches Vorgehen: Wende erst das Bisektionsverfahren an, bis eine halbwegs gute Näherung gefunden wurde. Verwende dann das Newtonverfahren, um schnell bessere Näherungen zu bekommen.

d) Beispiele

i) Die Iteration kann den Definitionsbereich verlassen.
TODO: Abbildung einfügen

ii) Die Folge kann oszillieren.

iii) Probleme gibt es oft, falls die Ableitung Nullstelle in der Nähe hat.
TODO: Abbildung einfügen

5.13 Newton-Verfahren für Systeme

Betrachte eine zweimal stetig differenzierbare Funktion $f : D \to \mathbb{R}^n$ mit offenem Definitionsbereich $D \subset \mathbb{R}^n$. Suche $x \in D$ mit $f(x) = 0$. Führe dazu eine Taylorentwicklung um $\overline{x} \in D$ durch:

$$f(x) = f(\bar{x}) + \mathrm{D}f(\bar{x})(x - \bar{x}) + \mathcal{O}_0\left(\|x - \bar{x}\|^2\right)$$

Die Ableitung ist:

$$\mathrm{D}f(\bar{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots & \frac{\partial f_n}{\partial x_n} \end{pmatrix}$$

In Komponenten erhalten wir für $x \to \overline{x}$:

$$f_i(x) = f_i(\bar{x}) + \sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j}(x_j - \bar{x}_j) + \mathcal{O}_0\left(\|x - \bar{x}\|^2\right)$$

Suche jetzt Nullstelle der (affin-)linearen Näherung. Ist $x^{(k)}$ schon berechnet, so löse

$$0 = f\left(x^{(k)}\right) + \mathrm{D}f\left(x^{(k)}\right)\left(x^{(k+1)} - x^{(k)}\right)$$

nach $x^{(k+1)}$. Falls $\mathrm{D}f\left(x^{(k)}\right)$ invertierbar ist, so erhalten wir:

$$x^{(k+1)} = x^{(k)} - \left(\mathrm{D}f\left(x^{(k)}\right)\right)^{-1}\left(f\left(x^{(k)}\right)\right)$$

Hier wird jedoch die Inverse gebraucht.

Für $n = 1$ entspricht dies dem Verfahren das wir schon kennen. Für gegebenes $x^{(0)}$ bestimme iterativ $x^{(1)}, x^{(2)}, ...$

In der Praxis gehen wir wie folgt vor: Sei $x^{(k)}$ gegeben, so löse:

$$\mathrm{D}f\left(x^{(k)}\right)s^{(k)} = -f\left(x^{(k)}\right)$$

Setze:

$$x^{(k+1)} = x^{(k)} + s^{(k)}$$

Dies benötigt nur eine Lösung eines linearen Gleichungssystems, während obige Formel die sehr aufwändige Berechnung einer Inversen benötigt.

Ziel: Zeige die quadratische Konvergenz.

5.14 Satz

Es sei $\Omega \subset \mathbb{R}^n$, $f : \Omega \to \mathbb{R}^n$ eine stetig differenzierbare Funktion mit invertierbarer Jacobi-Matrix $\mathrm{D}f(x)$ für alle $x \in \Omega$. Sei $\beta \in \mathbb{R}_{>0}$ mit:

$$\left|\left|\left|\left(\mathrm{D}f(x)\right)^{-1}\right|\right|\right| \leq \beta \qquad \underset{x\in\Omega}{\forall}$$

Weiter sei $\mathrm{D}f(x)$ Lipschitz-stetig mit Konstanten γ, das heißt:

$$|||\mathrm{D}f(x) - \mathrm{D}f(y)||| \leq \|x - y\| \qquad \underset{x,y\in\Omega}{\forall}$$

Weiter existiert eine Lösung x^* von $f(x^*) = 0$ in Ω. Der Startwert $x^{(0)}$ erfülle

$$x^{(0)} \in B_\omega(x^*) := \left\{x \in \mathbb{R}^n \,\middle|\, \|x^* - x\| < \omega\right\}$$

Dabei ist ω so gewählt, dass $B_\omega(x^*) \subset \Omega$ ist und $\omega \leq \frac{2}{\gamma\beta}$ gilt.

Dann bleibt die durch die Newton-Verfahren definierte Folge $\left(x^{(k)}\right)_{k\in\mathbb{N}}$ in der Kugel $B_\omega(x^*)$ und konvergiert quadratisch gegen x^*. Genauer gilt:

$$\left\|x^{(k+1)} - x^*\right\| \leq \frac{\beta\gamma}{2}\left\|x^{(k)} - x^*\right\|$$

Beweis

Für $x^{(k)} \in \Omega$ gilt:

$$\begin{aligned} x^{(k+1)} - x^* &= x^{(k)} - x^* - \left(\mathrm{D}f\left(x^{(k)}\right)\right)^{-1}\left(f\left(x^{(k)}\right) - f(x^*)\right) \\ &= \left(\mathrm{D}f\left(x^{(k)}\right)\right)^{-1}\left(f(x^*) - f\left(x^{(k)}\right) - \mathrm{D}f\left(x^{(k)}\right)\left(x^* - x^{(k)}\right)\right) \end{aligned}$$

Aus der Abschätzung $\left|\left|\left|\left(\mathrm{D}f\left(x\right)\right)^{-1}\right|\right|\right| \leq \beta$ folgt:

$$\left\|x^{(k+1)} - x^*\right\| \leq \beta \left\|f\left(x^*\right) - f\left(x^{(k)}\right) - \mathrm{D}f\left(x^{(k)}\right)\left(x^* - x^{(k)}\right)\right\|$$

Betrachte:

$$\phi\left(t\right) = f\left(y + t\left(x - y\right)\right) \qquad \underset{x,y \in B_\omega(x^*),\, t\in[0,1]}{\forall}$$

Es gilt dann:

$$\phi\left(1\right) = f\left(x\right) \qquad\qquad \phi\left(0\right) = f\left(y\right)$$

Weiter ist:

$$\phi'\left(t\right) = \mathrm{D}f\left(y + t\left(x - y\right)\right)\left(x - y\right)$$

Da $\mathrm{D}f$ Lipschitz ist, folgt:

$$\begin{aligned}\left\|\phi'\left(t\right) - \phi'\left(0\right)\right\| &= \left\|\left(\mathrm{D}f\left(y + t\left(x - y\right)\right) - \mathrm{D}f\left(y\right)\right)\left(x - y\right)\right\| \\ &= \left|\left|\left|\mathrm{D}f\left(y + t\left(x - y\right)\right) - \mathrm{D}f\left(y\right)\right|\right|\right| \cdot \left\|x - y\right\| \\ &\leq \gamma \left\|x - y\right\|^2\end{aligned}$$

Außerdem gilt:

$$f\left(x\right) - f\left(y\right) - \mathrm{D}f\left(y\right)\left(x - y\right) = \phi\left(1\right) - \phi\left(0\right) - \phi'\left(0\right) \overset{\text{HDI}}{=} \int_0^1 \left(\phi'\left(t\right) - \phi'\left(0\right)\right) \mathrm{d}t$$

Es folgt:

$$\left\|f\left(x\right) - f\left(y\right) - \mathrm{D}f\left(y\right)\left(x - y\right)\right\| \leq \gamma \left\|x - y\right\|^2 \int_0^1 t \mathrm{d}t = \frac{\gamma}{2}\left\|x - y\right\|^2$$

Damit folgt jetzt:

$$\left\|x^{(k+1)} - x^*\right\| \leq \frac{\beta\gamma}{2}\left\|x^* - x^{(k)}\right\|^2$$

Dies zeigt die quadratische Konvergenz. Es bleibt jedoch noch zu zeigen, dass für alle $k \in \mathbb{N}$ gilt:

$$\left\|x^{(k)} - x^*\right\| < \omega$$

Für $k = 0$ gilt dies nach Voraussetzung.
Induktionsschritt $k \rightsquigarrow k + 1$:

$$\left\|x^{(k+1)} - x^*\right\| \leq \frac{\beta\gamma}{2}\left\|x^* - x^{(k)}\right\| \cdot \left\|x^* - x^{(k)}\right\| \leq \frac{\beta\gamma\omega}{2}\omega$$

Wegen $\omega < \frac{2}{\beta\gamma}$ folgt die Behauptung. $\square_{5.14}$

Bemerkung

Die Parameter β und γ sind durch das Problem gegeben. Damit die Voraussetzungen des Satzes erfüllt sind, muss ω genügend klein gewählt werden.

5.15 Beispiel: Diskretisierung eines nicht-linearen Randwertproblems

Betrachte die Situation aus 5.2 d). Seien $I = [a, b]$, $u_a, u_b \in \mathbb{R}$ und $f : \mathbb{R} \to \mathbb{R}$ stetig differenzierbar. Suche eine Lösung $u : [a, b] \to \mathbb{R}$ von:

$$u''(x) = f(u(x)) \qquad \underset{x \in I}{\forall} \qquad u(a) = u_a \qquad u(b) = u_b$$

– Zerlege I in $N + 1$ äquidistante Teilintervalle mit Länge:

$$h = \frac{b-a}{N+1}$$

–

$$u''(x) \approx \frac{1}{h^2}\left(u(x_{i+1}) - 2u(x_i) + u(x_{i-1})\right)$$

– Verlange:

$$\frac{1}{h^2}\left(u_{i+1} - 2u_i + u_{i-1}\right) = f(u_i)$$

Gesucht sind $u_1, ..., u_N$.
Dann erfüllt der Vektor $u = (u_1, ..., u_N)^T$ die nichtlineare Gleichung:

$$H(u) := \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix} u + \begin{pmatrix} h^2 f(u_1) - u_a \\ h^2 f(u_2) \\ \vdots \\ h^2 f(u_{n-1}) \\ h^2 f(u_n) - u_b \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

Ziel: Löse diese Gleichung mit dem Newton-Verfahren.
Vorgehen: Bestimme iterativ:

$$u^{(k+1)} = u^{(k)} - \left(\mathrm{D}H\left(u^{(k)}\right)\right)^{-1} H\left(u^{(k)}\right)$$

Berechne dafür $v^{(k+1)}$ als Lösung von

$$\mathrm{D}H\left(u^{(k)}\right) v^{(k+1)} = -H\left(u^{(k)}\right)$$

und setze:

$$u^{(k+1)} = u^{(k)} + v^{(k+1)}$$

Es gilt:

$$\mathrm{D}H\left(u^{(k)}\right) = \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix} + h^2 \begin{pmatrix} f'\left(u_1^{(k)}\right) & 0 & \dots & \dots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & f'\left(u_n^{(k)}\right) \end{pmatrix}$$

Das zu lösende Gleichungssystem ist:

$$\begin{pmatrix} 2+h^2 f'\left(u_1^{(k)}\right) & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2+h^2 f'\left(u_n^{(k)}\right) \end{pmatrix} v^{(k+1)} = \underbrace{\begin{pmatrix} u_a - 2u_1^{(k)} + u_2^{(k)} - h^2 f\left(u_1^{(k)}\right) \\ u_1^{(k)} - 2u_2^{(k)} + u_3^{(k)} - h^2 f\left(u_2^{(k)}\right) \\ \vdots \\ u_{n-1}^{(k)} - 2u_n^{(k)} + u_b - h^2 f\left(u_n^{(k)}\right) \end{pmatrix}}_{=:b^{(k)}}$$

Für $k \ll 1$ ist $\mathrm{D}H\left(u^{(k)}\right)$ positiv definit und dann existiert eine eindeutige Lösung dieser Gleichung. (Wie kann man das zeigen?) Mittels Gauß-Elimination sehen wir, dass $v^{(k+1)}$ durch ein lineares Gleichungssystem der folgenden Form berechnet werden kann:

$$\underbrace{\begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ l_2 & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & l_n & 1 \end{pmatrix}}_{=:L} \cdot \underbrace{\begin{pmatrix} r_1 & -1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & -1 \\ 0 & \dots & \dots & 0 & r_n \end{pmatrix}}_{=:R} v^{(k+1)} = b^{(k)}$$

Dabei ist:

$$\begin{aligned} r_1 &= 2 + h^2 f'\left(u_1^{(k)}\right) \\ r_i &= 2 + h^2 f'\left(u_i^{(k)}\right) - \frac{1}{r_{i-1}} \\ l_i &= -\frac{1}{r_{i-1}} \end{aligned}$$

Wir lösen nun

$$LRv^{(k+1)} = b^{(k)}$$

indem wir erst $Ly = b^{(k)}$ lösen, das heißt:

$$\begin{aligned} y_1 &= u_a - 2u_1^{(k)} + u_2^{(k)} - h^2 f\left(u_1^{(k)}\right) \\ y_i &= u_{i-1}^{(k)} - 2u_i^{(k)} + u_{i+1}^{(k)} - h^2 f\left(u_i^{(k)}\right) + \frac{1}{r_{i-1}} y_{i-1} \\ y_n &= u_{n-1}^{(k)} - 2u_n^{(k)} + u_b - h^2 f\left(u_n^{(k)}\right) + \frac{1}{r_{n-1}} y_{n-1} \end{aligned}$$

Danach lösen wir $Rv^{(k+1)} = y$ durch Rückwärtssubstitution.
Achtung: Speichere **auf keinen Fall** die komplette $(n \times n)$-Matrix, da dies eine enorme Speicherplatzverschwendung ist!

5.16 Abbruchkriterien beim Newton-Verfahren

1. Limitiere die Anzahl der Iterationen, schon um Endlosschleifen durch fehlerhafte Programmierung zu vermeiden.

2. Breche ab, wenn das Verfahren nicht konvergiert, also zum Beispiel wenn $x^{(k)}$ nicht im Definitionsbereich liegt.
3. Breche ab, wenn das Newton-Verfahren genau genug ist, das heißt der Fehler

$$e^{(k)} := \left\| x^* - x^{(k)} \right\|$$

klein genug ist. Realistisch ist es, abzubrechen, wenn $\left\| x^{(k+1)} - (k) \right\| < \text{tol}$ ist. Dabei ist tol (Toleranz) eine vorgegebene Zahl.

5.17 Das gedämpfte Newton-Verfahren

Beobachtung: Im Newton-Verfahren liefert die „Korrektur“

$$S^{(k)} = -\left(\mathrm{D}f\left(x^{(k)}\right)\right)^{-1} f\left(x^{(k)}\right)$$

eine Richtung, in der $|f|$ abnimmt. Wähle nun $\lambda \in (0, 1]$ und definiere:

$$x^{(k+1)} = x^{(k)} + \lambda S^{(k)}$$

Die Strategie ist, λ so gedämpft zu wählen, dass $|f|$ abnimmt. Dahinter steckt die Idee, dass weniger manchmal mehr ist.

TODO: Abbildung einfügen

Betrachte zunächst $n = 1$. Für $f\left(x^{(k)}\right) \neq 0 \neq f'\left(x^{(k)}\right)$ gilt:

$$g(\lambda) = f\left(x^{(k)} + \lambda S^{(k)}\right)$$

$$g'(0) = f'\left(x^{(k)}\right) \cdot \underbrace{S^{(k)}}_{= \frac{f\left(x^{(k)}\right)}{f'\left(x^{(k)}\right)}} = -f\left(x^{(k)}\right) \begin{cases} < 0 & \text{falls } f\left(x^{(k)}\right) > 0 \\ > 0 & \text{falls } f\left(x^{(k)}\right) < 0 \end{cases}$$

Das heißt für eine kleines $\lambda \in (0, 1]$ gilt:

$$\left| f\left(x^{(k)} + \lambda S^{(k)}\right) \right| < \left| f\left(x^{(k)}\right) \right|$$

Ziel: Verallgemeinere dies auf $\mathbb{R}^n$. Es seien $f : \mathbb{R}^n \to \mathbb{R}^n$ und $x^{(k)}$ gegeben. Definiere:

$$S^{(k)} = -\left(\mathrm{D}f\left(x^{(k)}\right)\right)^{-1} f\left(x^{(k)}\right)$$

Wir wollen $f(x) = 0$ erreichen. Ziel ist es $\left\| f\left(x^{(k)}\right) \right\|$ kleiner zu machen (in einer geeigneten Norm). Versuche λ in

$$x^{(k+1)} = x^{(k)} + \lambda S^{(k)}$$

so zu wählen, dass gilt:

$$\left\| f\left(x^{(k+1)}\right) \right\| < \left\| f\left(x^{(k)}\right) \right\| \qquad \text{(M)}$$

Die Wahl der Norm ist nicht trivial!
Problem: Das Newton-Verfahren ist affin-invariant. Das Nullstellenproblem $f(x) = 0$ ist für eine invertierbare Matrix A äquivalent zu:

$$f_A(x) = Af(x) = 0$$

Zum Beispiel eine Diagonalmatrix skaliert die Zeilen. Das Newton-Verfahren bemerkt A gar nicht!

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - \left(\mathrm{D}f_A\left(x^{(k)}\right)\right)^{-1} f_A\left(x^{(k)}\right) = \\ &= x^{(k)} - \left(A\mathrm{D}f\left(x^{(k+1)}\right)\right)^{-1} Af\left(x^{(k)}\right) = \\ &= x^{(k)} - \left(\mathrm{D}f\left(x^{(k)}\right)\right)^{-1} f\left(x^{(k)}\right) \end{aligned}$$

Verlange daher einen modifizierten Monotonietest (M) der affin-invariant ist.
(AIM): Teste, ob gilt:

$$\left\|\left(\mathrm{D}f(\hat{x})\right)^{-1} f\left(x^{(k+1)}\right)\right\|_2 \leq \left\|\left(\mathrm{D}f(\hat{x})\right)^{-1} f\left(x^{(k)}\right)\right\|_2$$

Bemerkung

i) Die rechte Seite enthält für $\hat{x} = x^{(k)}$ gerade die Newton-Korrektur $S^{(k)}$.
ii) Die linke Seite benötigt die Lösung eines linearen Gleichungssystems. Jetzt sei stets $\hat{x} = x^{(k)}$ und definiere:

$$\|z\|_{(k)} := \left\|\left(\mathrm{D}f\left(x^{(k)}\right)\right)^{-1} z\right\|_2$$

5.18 Lemma

Es sei $x^{(k)} \in \mathbb{R}^n$ mit $f\left(x^{(k)}\right) \neq 0$ gegeben, f sei in einer Umgebung von $x^{(k)}$ zweimal stetig differenzierbar und $\mathrm{D}f\left(x^{(k)}\right)$ sei invertierbar. Betrachte

$$\begin{aligned} S^{(k)} &:= -\left(\mathrm{D}f\left(x^{(k)}\right)\right)^{-1} f\left(x^{(k)}\right) \\ x^{(k+1)} &:= x^{(k)} + \lambda S^{(k)} \end{aligned}$$

und $\|\cdot\|_{(k)}$ wie oben. Dann existiert ein $x \in (0,1)$ mit:

$$\left\|f\left(x^{(k+1)}\right)\right\|_{(k)} < \left\|f\left(x^{(k)}\right)\right\|_{(k)}$$

Beweis

Definiere:

$$\begin{aligned} B &:= \left(\mathrm{D}f\left(x^{(k)}\right)\right)^{-1} \\ f_B(x) &:= Bf(x) \end{aligned}$$

$$g(x) := (f_B(x))^{\mathrm{T}} \cdot f_B(x) = \left\| \left(\mathrm{D}f\left(x^{(k)}\right) \right)^{-1} f(x) \right\|_2^2 = \|f(x)\|_{(k)}^2$$

Die Taylor-Entwicklung von g um $\lambda = 0$ ist:

$$g\left(x^{(k)} + \lambda S^{(k)}\right) = g\left(x^{(k)}\right) + \lambda \nabla g\left(x^{(k)}\right) \cdot S^{(k)} + \mathcal{O}_0\left(\lambda^2\right)$$

Wegen

$$\begin{aligned} \nabla g(x) &= 2 (f_B(x))^{\mathrm{T}} \mathrm{D}(f_B(x)) \\ \mathrm{D}f_B(x) &= B\mathrm{D}f(x) \end{aligned}$$

folgt:

$$\begin{aligned} \left(\nabla g\left(x^k\right)\right) \circ S^{(k)} &= -2 \left(f_B\left(x^k\right)\right)^{\mathrm{T}} \mathrm{D}\left(f_B\left(x^k\right)\right) \cdot \left(\mathrm{D}f\left(x^{(k)}\right)\right)^{-1} f\left(x^{(k)}\right) = \\ &= -2 \left(Bf\left(x^k\right)\right)^{\mathrm{T}} B\mathrm{D}f\left(x^{(k)}\right) \cdot \left(\mathrm{D}f\left(x^{(k)}\right)\right)^{-1} f\left(x^{(k)}\right) = \\ &= -2 \left(f\left(x^k\right)\right)^{\mathrm{T}} B^{\mathrm{T}} B f\left(x^{(k)}\right) = -2 \left\| Bf\left(x^{(k)}\right) \right\|_2^2 < 0 \end{aligned}$$

Daraus folgt für genügend kleines $\lambda \in \mathbb{R}_{>0}$:

$$g\left(x^{(k)} + \lambda S^{(k)}\right) < g\left(x^{(k)}\right)$$

$\square_{5.18}$

5.19 Algorithmisches Vorgehen beim gedämpften Newton-Verfahren

Lasse maximal $k_{\max}$ Schleifen laufen, wähle als kleinsten Dämpfungsparameter $\lambda_{\min}$ und gebe eine kleine Toleranz tol vor.

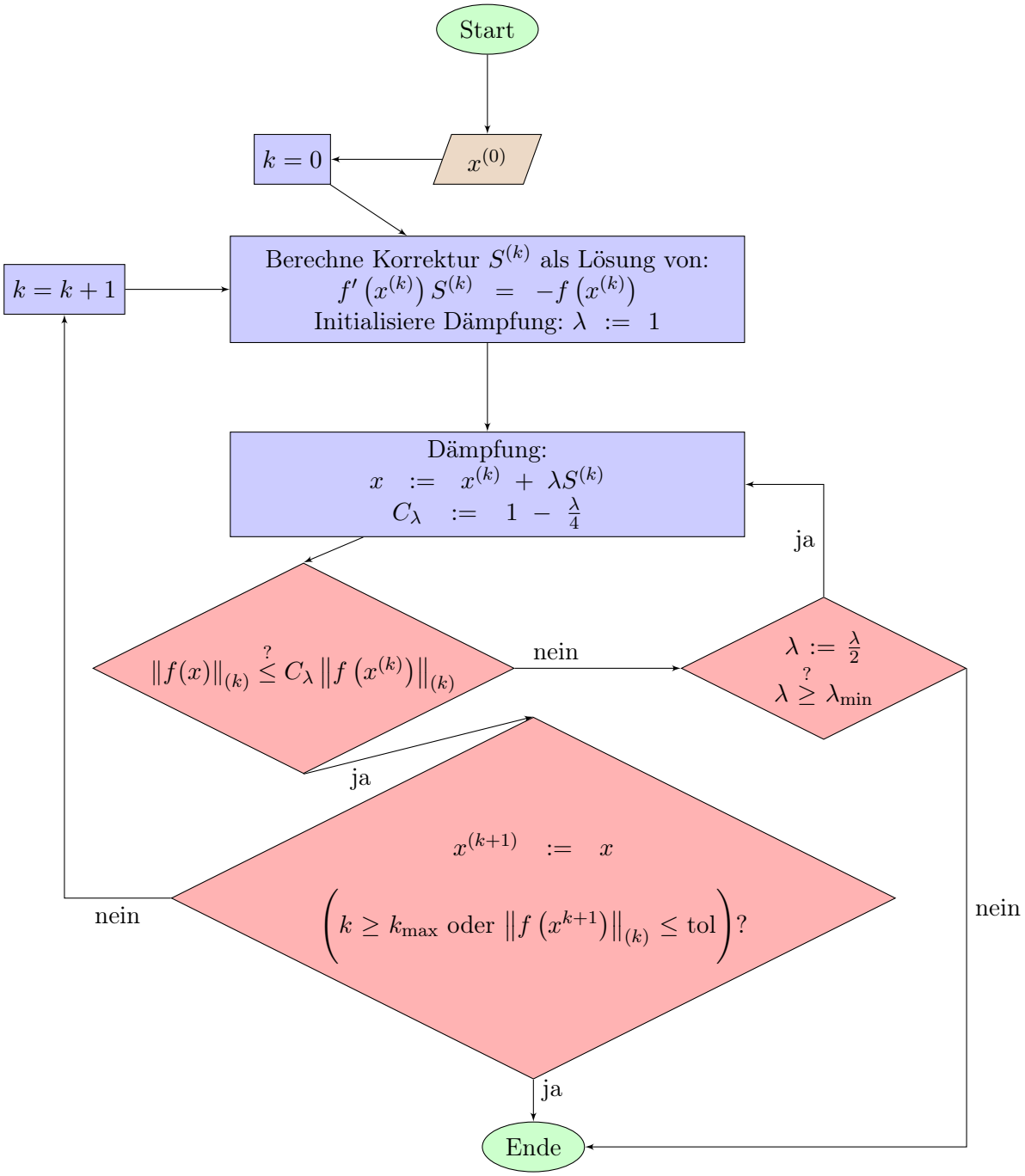


Abbildung 5.1: Flussdiagramm des Algorithmus

5.20 Das Sekantenverfahren

Idee: Zu zwei Schätzwerten $x, y \in \mathbb{R}$ einer Nullstelle von f wollen wir eine bessere Näherung bestimmen. Dabei wählen wir z als Schnittpunkt der Geraden durch $(x, f(x))$ und $(y, f(y))$ mit der x-Achse. Die Bestimmung von z geht daher wie folgt:

$$m := \frac{f(x) - f(y)}{x - y}$$

$$m = \frac{f(z) - f(x)}{z - x} \approx \frac{-f(x)}{z - x}$$
$$z \approx x - \frac{f(x)}{m} = x - \frac{x - y}{f(x) - f(y)} \cdot f(x)$$

Sekantenverfahren:

1. Wähle Startwerte $x^{(0)}$ und $x^{(1)}$.
2. Setze für $k \in \mathbb{N}$, falls $f\left(x^{(k)}\right) \neq f\left(x^{(k-1)}\right)$ ist:

$$x^{(k+1)} := x^{(k)} - \frac{x^{(k)} - x^{(k-1)}}{f\left(x^{(k)}\right) - f\left(x^{(k-1)}\right)} \cdot f\left(x^{(k)}\right)$$

5.21 Bemerkung

a) Das Sekantenverfahren lässt sich aus dem Newton-Verfahren ableiten, indem man Ableitung durch Differenzenquotient ersetzt.

b) Dies liefert auch Verfahren für Systeme: Ersetze in

$$\mathrm{D}f\left(x^{(k)}\right) = \left(\frac{\partial f_i}{\partial x_j}\right)_{i,i\in\{1,\ldots,n\}}\left(x^{(k)}\right)$$

die partielle Ableitung $\frac{\partial f_i}{\partial x_j}\left(x^{(k)}\right)$ durch den Differenzenquotient:

$$\frac{f_i\left(x^{(k)}\right) - f_i\left(x^{(k)} - \left(0,\ldots,0,x_j^{(k)} - x_j^{(k-1)},0,\ldots,0\right)\right)}{x_j^{(k)} - x_j^{(k-1)}}$$

c) Das Sekantenverfahren konvergiert superlinear mit Konvergenzordnung:

$$p = \frac{1+\sqrt{5}}{2} \approx 1,618$$

$$\Rightarrow \qquad \left\|x^{(k-1)} - x^*\right\| \leq C \left\|x^{(k)} - x^*\right\|^p$$

d) Trotz der langsameren Konvergenz als beim Newton-Verfahren ist das Sekantenverfahren oft effizienter, da keine Ableitung berechnet werden muss.

6 Interpolation

6.1 Einführung

Schon in 1.2 4. wurde folgendes Problem formuliert:
Gegeben seien Punkte (x_i, f_i) für $i \in \{0, \ldots, n\}$ mit paarweise verschiedenen x_i.
Finde eine Funktion p so, dass $p(x_i) = f_i$ ist.
Dieses Problem taucht zum Beispiel auf, wenn (x_i, f_i) Messdaten sind. Dann ist man daran interessiert, was an allen x. In diesem Fall sucht man eine interpolierende Funktion p mit $p(x_i) = f_i$. Dann wertet man p an der Stelle x aus. Weiter Anwendungsfelder:

- Computer-Aided Design (CAD)
- Computer-Aided Geometric Design (CAGD)

In diesem Fall soll die Funktion möglichst glatt, das heißt möglichst oft differenzierbar sein.

6.2 Anforderungen an die Interpolationsaufgabe

1. Die Funktion soll sich aus einfachen Funktionen zusammensetzen, also zum Beispiel aus Polynome, stückweisen Polynome, sin, cos, etc. oder Potenzen von exp bestehen.
2. Zu gegebenen Werten (x_i, f_i) sollte die Interpolierende leicht zu berechnen sein.
3. Wenn ich die Daten (x_i, f_i) leicht ändere, sollte sich p nur leicht ändern.
4. Eindeutige Lösung

6.3 Allgemeine Interpolationsaufgabe

Sei $I \subseteq \mathbb{R}$ ein Intervall und $g_i : I \to \mathbb{R}$ für $i \in \{0, \ldots, n\}$ seien $n+1$ gegebene Funktionen, so definiere $V := \operatorname{span}\{g_i | i \in \{0, \ldots, n\}\}$.
Zu Punkten (x_i, f_i) für $i \in \{0, \ldots, n\}$ mit paarweise verschiedenen x_i suchen wir ein $p \in V$, sodass für alle $i \in \{0, \ldots, n\}$ gilt:

$$p(x_i) = f_i$$

Da jedes $p \in V$ eine Darstellung

$$p = \sum_{j=0}^{n} a_j g_j$$

besitzt, ist die Aufgabe im Allgemeinen äquivalent dazu, Zahlen $a_0, ..., a_n \in \mathbb{R}$ mit

$$\sum_{j=0}^{n} a_j g_j(x_i) = f_i \tag{6.1}$$

für $i \in \{0, \ldots, n\}$ zu finden. Dabei handelt es sich um ein lineares Gleichungssystem. Für die Koeffizienten a_j definieren wir die Matrix:

$$f = \begin{pmatrix} g_0(x_0) & \ldots & g_n(x_0) \\ \vdots & & \vdots \\ g_0(x_n) & \ldots & g_n(x_n) \end{pmatrix} = (g_j(x_i))_{i,j \in \{0,\ldots,n\}}$$

Dann ist (6.1) äquivalent zu:

$$A \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} f_0 \\ \vdots \\ f_n \end{pmatrix} \tag{6.2}$$

6.4 Satz

Die Interpolationsaufgabe hat genau dann eine Lösung, wenn $\det A \neq 0$ ist.

Beweis

Im Allgemeinen ist das lineare Gleichungssystem (6.2) genau dann eindeutig Lösbar, wenn $\det(A) \neq 0$ ist. $\square_{6.4}$

6.5 Polynominterpolation

Mit der Notation von oben wähle $g_j(x) := x^j$ für $j \in \{0, \ldots, n\}$. Das heißt p ist ein Polynom n-ten Grades. Der Raum aller Polynome maximale n-ten Grades wird mit $\mathbb{P}_n$ bezeichnet.

6.6 Satz

Zu gegebenen Punkten (x_i, f_i) für $i \in \{0, \ldots, n\}$ mit paarweise verschiedenen x_i gibt es genau ein $p \in \mathbb{P}_n$, das die Interpolationsaufgabe löst, das heißt:

$$p(x_i) = f_i$$

Beweis

Eindeutigkeit: Seien $p, q \in \mathbb{P}_n$ mit $p(x_i) = f_i = q(x_i)$ für alle $i \in \{0, ..., n\}$. Dann wäre $p - q$ ein Polynom n-ten Grades mit $n+1$ vielen Nullstellen. Aus dem Fundamentalsatz der Algebra folgt dann bereits $p = q$ (betrachte Zerlegung in Linearfaktoren).
Existenz: Satz 6.4 besagt, dass die eindeutige Lösbarkeit äquivalent ist zu:

$$\det \begin{pmatrix} 1 & x_0 & \ldots & x_0^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \ldots & x_n^n \end{pmatrix} \overset{\text{Vandermonde-Determinante}}{=} \prod_{i=0}^{n} \prod_{j=i+1}^{n} (x_i - x_j) \overset{x_i \neq x_j}{\neq} 0$$

Daraus folgt die Behauptung. $\square_{6.6}$

Also löse das Gleichungssystem mit der Vandermonde-Matrix.

6.7 Bemerkung

Will man zu Punkten (x_i, y_i) für $i \in \{0, \ldots, n\}$ mit paarweise verschiedenen x_i das Interpolationspolynom

$$p(x) = a_n x^n + \ldots + a_0$$

berechnen, so löst man:

$$\begin{pmatrix} 1 & x_0 & \ldots & x_0^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \ldots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} f_0 \\ \vdots \\ f_n \end{pmatrix}$$

Problem: Mit dem Gauß-Verfahren benötigt man $\mathcal{O}\left(n^3\right)$ Rechenoperationen und dies ist schlecht. Andere Verfahren benötigen nur $\mathcal{O}\left(n^2\right)$ Operationen.

6.8 Lagrange-Interpolationsformel

Idee: Wähle die Basis statt aus Monomen aus Polynomen, sodass

$$A = \left(w_j(x_i)\right)_{i,j}$$

die Einheitsmatrix ist. Dann gilt:

$$\sum_{j=0}^{n} a_j w_j(x_i) = f_i \quad \Leftrightarrow \quad a_i = f_i$$

Wie sind die w_j zu wählen? Es muss für $i, j \in \{0, ..., n\}$ gelten:

$$w_j(x_i) = \delta_{i,j}$$

Daher hat w_j eine Nullstelle in x_i für $i \neq j$ und es folgt:

$$w_j(x) = k_j \prod_{\substack{i=0 \\ i \neq j}} (x - x_i)$$

Damit $w_j(x_j) = 1$ ist, muss gelten:

$$k_j \prod_{\substack{i=0 \\ i \neq j}}^{n} (x_j - x_i) = 1$$

Daraus folgt:

$$w_j(x) = \prod_{\substack{i=0 \\ i \neq j}}^{n} \frac{x - x_i}{x_j - x_i}$$

6.9 **Satz** (Lagrange-Interpolationspolynom)

Das Interpolationspolynom $p \in \mathbb{P}_n$ zu den Punkten aus $\left\{ (x_i, f_i) \,\middle|\, i \in \{0, \ldots, n\} \right\}$ mit paarweise verschiedenen x_i hat die Darstellung:

$$p(x) = \sum_{j=0}^{n} f_j w_j(x)$$

Beweis

Da A die Einheitsmatrix ist, gilt:

$$\sum_{j=0}^{n} a_j w_j(x_i) = f_i \quad \Leftrightarrow \quad a_i = f_i$$

$\square_{6.9}$

6.10 Bemerkung

Es gibt effizientere Verfahren, aber zur theoretischen Untersuchung ist diese Darstellung gut geeignet.

6.11 **Definition** (Interpolationspolynom)

Zu vorgegebenen Punkten (x_i, f_i) für $i \in \{0, \ldots, n\}$ mit paarweise verschiedenen x_i bezeichne

$$\left(Pf \middle| x_0, \ldots, x_n\right) \in \mathbb{P}_n$$

das eindeutig bestimmte *Interpolationspolynom* vom Grad $\leq n$. Später schreiben wir oft

$$f_i := f(x_i)$$

für eine Funktion f. Das Interpolationspolynom ist nach Satz 6.9:

$$\left(Pf \middle| x_0, \ldots, x_n\right)(x) = \sum_{j=0}^{n} f_j \underbrace{\prod_{\substack{i=0 \\ i \neq j}}^{n} \frac{x - x_i}{x_j - x_i}}_{= w_j(x)}$$

Dies ist anfällig gegenüber Auslöschung.

6.12 Rekursionsformel von Neville und Aitken

Will man die Interpolierende nur an einer Stelle x auswerten, so bietet sich eine rekursive Berechnung an.

6.13 **Lemma** (Rekursionsformel)

$$\left(Pf\middle|x_0,\dots,x_n\right)(x) = \frac{x-x_0}{x_n-x_0}\left(Pf\middle|x_1,\dots,x_n\right)(x) + \frac{x_n-x}{x_n-x_0}\left(Pf\middle|x_0,\dots,x_{n-1}\right)(x) =: \varphi(x)$$

Beweis

Wegen der Eindeutigkeit des Interpolationspolynoms genügt es, dass für alle $i \in \{0,\dots,n\}$ schon $\varphi(x_i) = f_i$ gilt. Dazu berechne:

$$\varphi(x_0) = 0 + \frac{x_n-x_0}{x_n-x_0}\underbrace{\left(Pf\middle|x_0,\dots,x_{n-1}\right)(x_0)}_{=f_0} = f_0$$

Analog folgt $\varphi(x_n) = f_n$. Für $1 < i < n$ gilt:

$$\begin{aligned}\varphi(x_i) &= \frac{x_i-x_0}{x_n-x_0}\underbrace{\left(P_f\big|_{x_1,\dots,x_n}\right)(x_i)}_{=f_i} + \frac{x_n-x_i}{x_n-x_0}\underbrace{\left(P_f\big|_{x_0,\dots,x_{n-1}}\right)(x_i)}_{=f_i} = \\ &= \frac{f_i}{x_n-x_0}(x_i - x_0 + x_n - x_i) = f_i\end{aligned}$$

$\square_{6.13}$

6.14 Algorithmus von Neville-Aitken

Sei $x \in \mathbb{R}$ fest und seien $i_0,\dots,i_k \in \{0,\dots,n\}$ paarweise verschieden. Setze:

$$P_{i_0,\dots,i_k}(x) := \left(Pf\middle|x_{i_0},\dots,x_{i_k}\right)(x)$$

Dann gilt nach Lemma 6.13:

$$\begin{aligned}P_{i_0,\dots,i_k} &= \frac{x-x_{i_0}}{x_{i_k}-x_{i_0}}P_{i_1,\dots,i_k} + \frac{x_{i_k}-x}{x_{i_k}-x_{i_0}}P_{i_0,\dots,i_{k-1}} = \\ &= P_{i_1,\dots,i_k} + \frac{x_{i_k}-x}{x_{i_k}-x_{i_0}}\left(P_{i_0,\dots,i_{k-1}} - P_{i_1,\dots,i_k}\right)\end{aligned}$$

Damit folgt das Neville-Aitken-Schema:

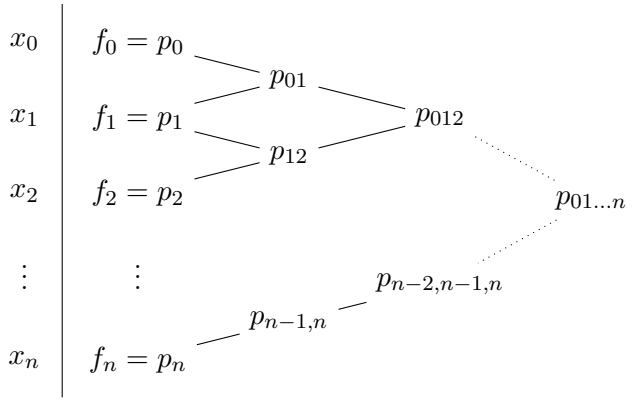


Abbildung 6.1: Neville-Aitken-Schema

Beispiel: $n = 3$

x_i	0	1	2	4
f_i	0	1	8	64

6.15 Newtonsche Interpolationsformel

Idee: Finde bei bekanntem Interpolationspolynom $\left(Pf\middle|x_0,\ldots,x_{n-1}\right)$ einen Korrekturterm, durch dessen Ergänzung $\left(Pf\middle|x_0,\ldots,x_n\right)$ entsteht. Dazu betrachte zunächst folgende Darstellung:

6.16 Satz

$p_n = \left(Pf\middle|x_0,\ldots,x_n\right) \in \mathbb{P}_n$ lässt sich schreiben als:

$$p_n(x) = d_0 + d_1(x-x_0) + d_2(x-x_0)(x-x_1) + \ldots + d_n(x-x_0)\cdot\ldots\cdot(x-x_{n-1})$$

Dabei sind die Koeffizienten $d_k \in \mathbb{R}$ eindeutig bestimmt durch:

$$d_k = \frac{f_k - p_{k-1}(x_k)}{(x_k-x_0)\cdot\ldots\cdot(x_k-x_{k-1})}$$

Beweis

Führe eine Induktion über n durch:
Induktionsanfang bei $n=0$: Notwendigerweise gilt $d_0 = f_0$.
Induktionsschritt $n-1 \rightsquigarrow n$: Sei

$$p_{n-1}(x) = d_0 + d_1(x-x_0) + d_2(x-x_0)(x-x_1) + \ldots + d_{n-1}(x-x_0)\cdot\ldots\cdot(x-x_{n-2}) \in \mathbb{P}_{n-1}$$

mit $p_{n-1}(x_i) = f_i$ für alle $i \in \{0,\ldots,n-1\}$, so verwende den Ansatz:

$$p_n(x) = p_{n-1}(x) + d_n(x-x_0)\cdot\ldots\cdot(x-x_{n-1})$$

Da $p_n(x_n) = f_n$ sein muss, folgt:

$$d_n = \frac{f_n - p_{n-1}(x_n)}{(x_n-x_0)\cdot\ldots\cdot(x_n-x_{n-1})}$$

$\square_{6.16}$

6.17 Bemerkung

Obige Darstellung hat den Vorteil, dass bei Hinzunahme eines weiteren Knotens x_{n+1} nur ein Koeffizient neu berechnet werden muss.

6.18 Newtonsche dividierte Differenzen

Wir definieren die Newtonsche dividierte Differenz durch:

$$[x_0,\ldots,x_n]\,f := d_n$$

Dabei ist d_n wie in Satz 6.16 der Koeffizient von x^n des Interpolationspolynoms. Rekursive Anwendung liefert folgende Form der Newtonschen Interpolationsformel:

$$\left(Pf\middle|x_0,\ldots,x_n\right)(x) = [x_0]\,f + (x-x_0)\,[x_0,x_1]\,f + (x-x_0)\,(x-x_1)\,[x_0,x_1,x_2]\,f + \ldots + \\ + (x-x_0)\cdot\ldots\cdot(x-x_{n-1})\,[x_0,\ldots,x_n]\,f$$

Für eine Funktion $f : [a,b] \to \mathbb{R}$ und paarweise verschiedene $x_i \in [a,b]$ für $i \in \{0,\ldots,n\}$ setzen wir immer $f_i := f(x_i)$.
Frage: Wie berechnen wir $[x_0,\ldots,x_n]\,f$ und woher kommt der Name?

6.19 Korollar

Es gilt:

$$[x_0,\ldots,x_n]\,f = \frac{[x_1,\ldots,x_n]\,f - [x_0,\ldots,x_{n-1}]\,f}{x_n - x_0}$$

Beweis

Nach Lemma 6.13 gilt:

$$\left(Pf\middle|x_0,\ldots,x_n\right)(x) = \frac{x-x_0}{x_n-x_0}\left(Pf\middle|x_1,\ldots,x_n\right)(x) + \frac{x_n-x}{x_n-x_0}\left(Pf\middle|x_0,\ldots,x_{n-1}\right)(x) =: \varphi(x)$$

$[x_0,\ldots,x_n]\,f$ ist der Koeffizient vor x^n in $\varphi(x)$.
$[x_1,\ldots,x_n]\,f$ ist der Koeffizient vor x^{n-1} in $\left(Pf\middle|x_1,\ldots,x_n\right)$.
$[x_0,\ldots,x_{n-1}]\,f$ ist der Koeffizient vor x^{n-1} in $\left(Pf\middle|x_0,\ldots,x_{n-1}\right)$.
Ein Vergleich der Koeffizienten liefert die Behauptung. $\square_{6.19}$

6.20 Dreiecksschema zur Berechnung der Newtonschen dividierten Differenzen

Mit $[x_i]\,f = f_i$ (Interpolationspolynom ist eine Konstante) und Korollar 6.19 können wir die dividierten Differenzen nach folgendem rekursivem Schema berechnen:

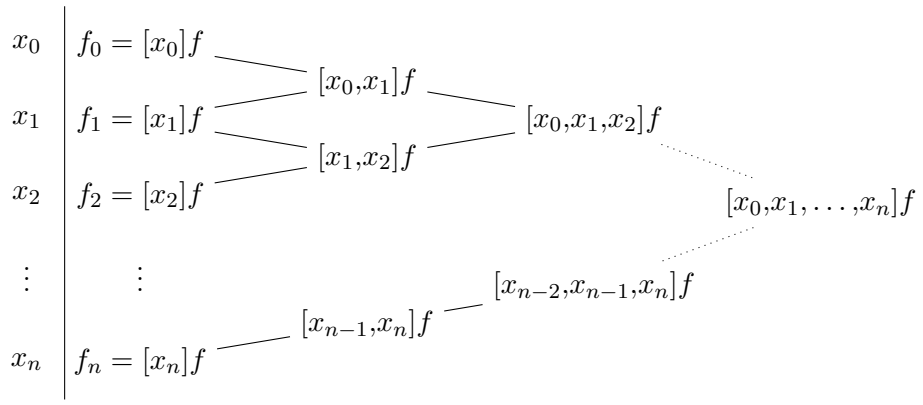


Abbildung 6.2: Newtonsche dividierte Differenzen

Der obere Rand liefert die Koeffizienten in der Newton-Darstellung aus 6.16 beziehungsweise 6.19. $\square_{6.20}$

6.21 Das Horner-Schema

Wir wollen die Polynome

$$p(x) = \sum_{k=0}^{n} a_k x^k$$

beziehungsweise

$$q(x) = \sum_{k=0}^{n} d_k \cdot \prod_{j=0}^{k-1} (x - x_j)$$

auswerten.

Vorwärtssubstitution

Berechne die Potenzen x^k, die Produkte $a_k x^k$ und summiere auf. Berechne also

$$u_0 = 1 \qquad\qquad v_0 = a_0$$

und für $k \in \{1, \ldots, n\}$:

$$u_k := x \cdot u_{k-1} \qquad\qquad v_k := v_{k-1} + a_k u_k$$

Der Wert v_n ist dann gerade $p(x)$.
Hier werden $2n$ Multiplikationen und n Additionen benötigt.

Rückwärtssubsitution (Horner-Schema)

Beispiel:

$$p(x) = 1 + 2x + 3x^2 + 4x^3 = 1 + x(2 + x(3 + 4x))$$

Allgemein gilt:

$$p(x) = \sum_{k=0}^{n} a_k x^k = a_0 + x(a_1 + x(a_2 + \ldots + x(a_{n-1} + a_n x)\ldots))$$

Definieren wir:

$$\begin{aligned} w_{n+1} &= 0 \\ w_k &= a_k + x \cdot w_{k+1} \end{aligned}$$

So liefert w_0 den Wert $p(x)$.
Dieses Vorgehen ist ein Algorithmus zur Auswertung von p an der Stelle x und benötigt $n+1$ Additionen und $n+1$ Multiplikationen. Er spart also $n-2$ Operationen im Vergleich zur Vorwärtssubstitution.
Will man $q(x)$ auswerten, so geht man entsprechend vor:

$$\begin{aligned} q(x) &= d_0 + d_1(x - x_0) + \ldots + d_n(x - x_0) \cdot \ldots \cdot (x - x_{n-1}) = \\ &= d_0 + (x - x_0)(d_1 + (x - x_1)(d_2 + \ldots + (x - x_{n-2})(d_{n-1} + (x - x_{n-1})\, d_n)\ldots)) \end{aligned}$$

In diesem Fall lautet das Horner-Schema für $k \in \{n, n-1, \ldots, 0\}$:

$$\begin{aligned} w_{n+1} &:= 0 \\ w_k &= d_k + (x - x_n)\, w_{k+1} \end{aligned}$$

6.22 Satz

Die Newtonschen dividierten Differenzen, haben die folgenden Eigenschaften:

1. Die Abbildung $[x_0, \ldots, x_n] : \mathcal{C}^0(\mathbb{R}) \to \mathbb{R}$ mit $f \mapsto [x_0, \ldots, x_n] f$ ist linear.
2. Ist $f(x) = x^k$ für $k \in \{0, \ldots, n\}$, so gilt $[x_0, \ldots, x_n] f = \delta_{k,n}$.
3. $[x_0, \ldots, x_n] f$ ist unabhängig von der Reihenfolge der Argumente $x_0, \ldots, x_n$.
4. Ist $f \in \mathcal{C}^n([a, b], \mathbb{R})$, so gibt es ein $\xi \in (\min(x_0, \ldots, x_n), \max(x_0, \ldots, x_n))$ mit:

$$[x_0, \ldots, x_n] f = \frac{f^{(n)}(\xi)}{n!}$$

Beweis

1. Seien $f, g \in \mathcal{C}^0(\mathbb{R})$ und $\alpha \in \mathbb{R}$. Wähle $p, q \in \mathbb{P}_n$ so, dass für alle $i \in \{0, \ldots, n\}$ gilt:

$$p(x_i) = f(x_i)$$
$$q(x_i) = g(x_i)$$

   Dann folgt für alle $i \in \{0, \ldots, n\}$:

$$(\alpha p + q)(x_i) = (\alpha f + g)(x_i)$$

   Das bedeutet, dass $\alpha p + q$ das Interpolationspolynom zu $\alpha f + g$ interpoliert an den Stellen $x_0, \ldots, x_n$ ist. Nach Definition ist $[x_0, \ldots, x_n] f$ der Koeffizient von x^n des Polynoms $\alpha p + q$. Da $\deg(p) = \deg(g)$ ist, ergibt sich dieser als das α-fache des Koeffizient von x^n in p addiert zum Koeffizient von x^n in q Daraus folgt:

$$[x_0, \ldots, x_n](\alpha f + g) = \alpha [x_0, \ldots, x_n] f + (x_0, \ldots, x_n) g$$

   Also ist $[x_0, \ldots, x_n]$ eine lineare Abbildung. $\square_{1)}$
2. Wegen $k \in \{0, \ldots, n\}$ ist das Interpolationspolynom von f:

$$(Pf | x_0, \ldots, n) = x^k$$

   Daher ist der Koeffizient von x^n nur von Null verschieden, wenn $k = n$ gilt, das heißt:

$$[x_0, \ldots, x_k] f = \begin{cases} 0 & k < n \\ 1 & k = n \end{cases}$$

   $\square_{2)}$
3. Die Eigenschaft Interpolationspolynom zu sein ist unabhängig von der Reihenfolge der $x_0, \ldots, x_n$. Also ist der Koeffizient von x^n unabhängig von der Reihenfolge. $\square_{3)}$
4. Betrachte $p(x) \in \mathbb{P}_n$ mit $p(x_i) = f(x_i)$ für $i \in \{0, \ldots, n\}$ und der Darstellung:

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

   Dann hat $q := p - f$ die $n + 1$ Nullstellen $x_0, \ldots, x_n$. Ohne Einschränkung (vergleiche 3.) sei $x_0 < x_1 < \ldots < x_n$. Nach dem Satz von Rolle existieren mindestens n Nullstellen von q'. Mindestens eine liegt in jedem offenen Intervall (x_{i-1}, x_i) für $i \in \{0, \ldots, n\}$.

Somit hat q'' mindestens $n-1$ Nullstellen, und iterativ folgt, dass $q^{(n)}$ mindestens eine Nullstelle hat. Das heißt, es existiert ein $\xi \in (x_0, x_n)$ mit:

$$0 = q^{(n)}(\xi) = p^{(n)}(\xi) - f^{(n)}(\xi) = a_n n! - f^{(n)}(\xi)$$

Wegen $a_n = [x_0, ..., x_n] f$ folgt:

$$[x_0, ..., x_n] f = \frac{f^{(n)}(\xi)}{n!}$$

$\square_{4)}$

$\square_{6.22}$

6.23 Fehlerdarstellung

Sei $I \subseteq \mathbb{R}$ ein Intervall, $f \in \mathcal{C}^{n+1}(I)$ und $p \in \mathbb{P}_n$ das Interpolationspolynom zu $(x_i, f(x_i))$ mit paarweise verschiedenen x_i, das heißt $p(x_i) = f(x_i)$ für $i \in \{0, \ldots, n\}$.
Dann existiert zu jedem Punkt $\bar{x} \in I$ ein Punkt $\xi(\bar{x}) \in (\min(x_0, ..., x_n), \max(x_0, ..., x_n))$ mit:

$$f(\bar{x}) - p(\bar{x}) = \left(\prod_{j=0}^{n} (\bar{x} - x_j) \right) \cdot \frac{1}{(n+1)!} f^{(n+1)}(\xi(\bar{x}))$$

Beweis

Falls $\bar{x} = x_i$ für ein $i \in \{0, \ldots, n\}$ gilt, so ist die Aussage trivialerweise erfüllt. Sei nun $\bar{x} \neq x_i$ für alle $i \in \{0, \ldots, n\}$. Dann sei $\bar{p} \in \mathbb{P}_{n+1}$ das Interpolationspolynom mit folgenden Eigenschaften:

$$\underset{i \in \{0,...,n\}}{\forall} : \bar{p}(x_i) = f(x_i) \qquad\qquad p(\bar{x}) = f(\bar{x})$$

Daraus folgt:

$$\bar{p}(x) = p(x) + \prod_{j=0}^{n} (x - x_j) [\bar{x}, x_0, ..., x_n] f \tag{6.3}$$

Aus 6.22 4. folgt die Existenz eines $\xi \in (\min(\bar{x}, x_0, ..., x_n), \max(\bar{x}, x_0, ..., x_n))$ mit:

$$[\bar{x}, x_0, ..., x_n] f = \frac{f^{(n+1)}(\xi)}{(n+1)!} \tag{6.4}$$

Wegen $\bar{p}(\bar{x}) = f(\bar{x})$ folgt aus (6.3) und (6.4):

$$f(\bar{x}) = \bar{p}(\bar{x}) = p(\bar{x}) + \prod_{j=0}^{n} (x - x_j) \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

$\square_{6.23}$

6.24 Fehlerabschätzung

Sei $f \in \mathcal{C}^{n+1}([a,b])$ und $p \in \mathbb{P}_n$ das Interpolationspolynom zu paarweise verschiedenen Stützstellen $x_0, \ldots, x_n \in [a,b]$, das heißt für $i \in \{0, \ldots, n\}$ gilt:

$$p(x_i) = f(x_i)$$

Dann gilt:

$$\|f - p\|_{\infty,[a,b]} := \sup_{x \in [a,b]} |f(x) - p(x)| \leq \frac{\left\|f^{(n+1)}\right\|_{\infty,[a,b]}}{(n+1)!} (b-a)^{n+1}$$

Beweis

Mit 6.23 folgt:

$$|f(x) - p(x)| = \left| \prod_{j=0}^{n} \underbrace{(x - x_i)}_{\leq |b-a|} \cdot \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \right| \leq (b-a)^{n+1} \frac{\left\|f^{n+1}\right\|_{\infty,[a,b]}}{(n+1)!}$$

$\square_{6.24}$

6.25 Bemerkungen

1. In der Fehlerdarstellung wurde folgendes gezeigt:

   $$|f(x) - p(x)| = \left| \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \cdot \prod_{j=0}^{n} (x - x_i) \right|$$

   Versuche die rechte Seite klein zu machen!
   Den Wert $f^{(n+1)}(\xi(x))$ kennen wir in der Regel nicht!
   Ziel: Mache den Ausdruck $\prod_{j=0}^{n}(x - x_j)$ klein, das heißt wähle $x_j \in [a,b]$ geeignet.
   Betrachte zunächst $[a,b] = [-1,1]$ und wähle x_j als die Nullstellen des Tschebyschow-Polynoms:

   $$p_{n+1}(x) = \cos((n+1)\arccos(x))$$

   Also sind die Nullstellen:

   $$x_j = \cos\left(\frac{2j+1}{n+1} \cdot \frac{\pi}{2}\right)$$

   Für allgemeine Intervalle $[a,b]$ müssen die Nullstellen transformiert werden:

   $$\bar{x}_j = 2 \cdot \frac{x_j - a}{b-a} - 1$$

2. Seien zu jedem $n \in \mathbb{N}$ die paarweise verschiedenen Knoten

   $$\left\{x_0^{(n)}, \ldots, x_n^{(n)}\right\} \subseteq [a,b]$$

gegeben.
TODO: Abbildungen
Das heißt für jedes n wähle $n+1$ Stützstellen für die Interpolation. Sei $f \in \mathcal{C}([a,b],\mathbb{R})$ gegeben und sei $p \in \mathbb{P}_n$ so gewählt, dass für alle $i \in \{0,\dots,n\}$ gilt:

$$p_n\left(x_i^{(n)}\right) = f\left(x_i^{(n)}\right)$$

Frage: Konvergiert $p_n \to f_n$ für $n \to \infty$?
Es gelten folgende Sätze:

Satz von Marcinkiewicz
Zu jedem $f \in \mathcal{C}([a,b],\mathbb{R})$ können Knoten $\left(x_0^{(n)},\dots,x_n^{(n)}\right) \subseteq [a,b]$ gewählt werden, sodass die Folge der Interpolationspolynome p_n gegen f konvergiert, das heißt:

$$\|f - p_n\|_{\infty,[a,b]} \xrightarrow{n\to\infty} 0$$

Satz von Faber
Zu jeder vorgegeben Folge von Knotenmengen $\left\{x_0^{(n)},\dots,x_n^{(n)}\right\} \subseteq [a,b]$ existiert eine Funktion $f \in \mathcal{C}([a,b],\mathbb{R})$, sodass gilt:

$$\|f - p_n\|_{\infty,[a,b]} \xrightarrow{n\to\infty} \infty$$

6.26 Definition (Hermite-Interpolation)

Gegeben seien die Knoten $x_0 < x_1 < \ldots < x_m$ und Werte $f_i^{(k)}$ für $k \in \{0,\dots,n_i-1\}$ und $i \in \{0,\dots,m\}$. Damit sind

$$n+1 = \sum_{i=0}^{m} n_i$$

Werte gegeben.
Ein Polynom $p \in \mathbb{P}_n$ mit der Eigenschaft

$$p^{(k)}(x_i) = f_i^{(k)} \tag{6.5}$$

für alle $i \in \{0,\dots,m\}$ und $k \in \{0,\dots,n_i-1\}$ heißt *Hermite-Interpolationspolynom.*

6.27 Satz

Das Hermite-Interpolationspolynom existiert und ist eindeutig bestimmt.

Beweis

Eindeutigkeit: Seien $p_1, p_2 \in \mathbb{P}_n$ mit der Eigenschaft (6.5). Dann gilt

$$q := p_1 - p_2 \in \mathbb{P}_n$$

und q hat $n+1$ Nullstellen (mit Vielfachheit gezählt). Daraus folgt $q = 0$ und somit $p_1 = p_2$.

Existenz: Betrachte die lineare Abbildung:

$$L : \mathbb{P}_n \to \mathbb{R}^{n+1}$$

$$p \mapsto L(p) := \underbrace{\begin{pmatrix} p(x_0) \\ p'(x_0) \\ \vdots \\ p^{(n_i-1)}(x_0) \\ p(x_1) \\ \vdots \\ p(x_m) \\ \vdots \\ p^{(n_m-1)}(x_m) \end{pmatrix}}_{=:v}$$

Aus den obigen Überlegungen folgt, dass L injektiv ist. Wegen $\dim(\mathbb{P}_n) = n+1$ ist L auch surjektiv. Dies zeigt die eindeutige Existenz eines Interpolationspolynoms als Urbild $p = L^{-1}(v)$. $\square_{6.27}$

Jetzt führen wir folgende Notation ein:

$$\overbrace{\underbrace{x_0}_{=:t_0} = \ldots = \underbrace{x_0}_{=:t_{n_0-1}}}^{n_0\text{-mal}} < \overbrace{\underbrace{x_1}_{=:t_{n_0}} = \ldots = \underbrace{x_1}_{=:t_{n_0+n_1-1}}}^{n_1\text{-mal}} < x_2 \ldots < \overbrace{x_m = \ldots = \underbrace{x_1}_{=:t_n}}^{n_m\text{-mal}}$$

Die Interpolationsbedingungen sind für $j \in \{0, \ldots, n\}$:

$$p^{(s_j)}(t_j) = f^{(s_j)}(t_j) \tag{6.6}$$

$$s_j = \max\left\{x \middle| t_j = t_{j-x}\right\}$$

Wir bezeichnen mit $P(f|t_0, \ldots, t_n)$ das Interpolationspolynom in $\mathbb{P}_n$, welches (6.6) erfüllt.

6.28 Bestimmung des Hermite-Interpolationspolynoms

Definition

Wir bezeichnen für beliebige Stützstellen t_i mit $i \in \{0, \ldots, n\}$ jeweils den Koeffizienten von x^{k-i} des Hermite-Polynoms $P\left(f\middle|t_i, \ldots, t_k\right) \in \mathbb{P}_{k-i}$ wieder mit $[t_i, \ldots, t_k] f$. Für $t_0 = \ldots = t_k$ ergibt sich als Folgerung:

$$[t_0, \ldots, t_k] f = \frac{f^{(k)}(t_0)}{k!}$$

Betrachte dafür das Taylor-Polynom:

$$f(t_0) + f'(t_0)(t - t_0) + \ldots + \frac{f^{(k)}(t_0)}{k!}(t - t_0)^k = P\left(f\middle|t_0, \ldots, t_k\right)$$

Lemma

Gegeben seien $t_0, \ldots, t_k \in \mathbb{R}$. Dann gilt für $t_i, t_j \in \{t_0, \ldots, t_k\}$:

$$[t_0, \ldots, t_k] f = \begin{cases} \dfrac{[t_0, \ldots, t_{i-1}, t_{i+1}, \ldots, t_k] f - [t_0, \ldots, t_{j-1}, t_{j+1}, \ldots, t_k] f}{t_i - t_j} & \text{für } t_i \neq t_j \\ \dfrac{f^{(k)}(t_0)}{k!} & \text{falls } t_0 = \ldots = t_k \end{cases}$$

Beweis

Dies kann unmittelbar nachgerechnet werden: Für $t_i \neq t_j$ gilt nach dem Lemma (6.13) von Aitken:

$$P\left(f \middle| t_0, \ldots, t_k\right)(x) = \frac{(x - t_i) P\left(f \middle| t_0, \ldots, t_{i-1}, t_{i+1}, \ldots, t_k\right)(x)}{t_j - t_i} - \\ - \frac{(x - t_j) P\left(f \middle| t_0, \ldots, t_{j-1}, t_{j+1}, \ldots, t_k\right)(x)}{t_j - t_i}$$

Ein Vergleich der Koeffizienten von x^k liefert die Behauptung. $\square_{\text{Lemma}}$

Beispiel

Für $t_0 = 0$, $t_1 = t_2 = t_3 = \frac{1}{2}$, $t_4 = 1$ bestimme $P \in \mathbb{P}_4$ so, dass gilt:

$$P(0) = 1 \qquad P(1) = \frac{5}{2}$$
$$P\left(\frac{1}{2}\right) = \frac{3}{2} \qquad P'\left(\frac{1}{2}\right) = \frac{1}{2} \qquad P''\left(\frac{1}{2}\right) = 0$$

t_i	$f_i = [t_i]f$	$[t_i,t_{i+1}]f$	$[t_i,t_{i+1},t_{i+2}]f$	$[t_i,t_{i+1},t_{i+2},t_{i+3}]f$	$[t_i,t_{i+1},t_{i+2},t_{i+3},t_{i+4}]f$
0	1	$\frac{\frac{3}{2}-1}{\frac{1}{2}-0} = 1$	$\frac{\frac{1}{2}-1}{\frac{1}{2}-0} = -1$	$\frac{0-(-1)}{\frac{1}{2}-0} = 2$	$\frac{6-2}{1-0} = 4$
$\frac{1}{2}$	$\frac{3}{2}$	$f'\left(\frac{1}{2}\right) = \frac{1}{2}$	$\frac{1}{2} f''\left(\frac{1}{2}\right) = 0$	$\frac{3-0}{1-\frac{1}{2}} = 6$	
$\frac{1}{2}$	$\frac{3}{2}$	$f'\left(\frac{1}{2}\right) = \frac{1}{2}$	$\frac{2-\frac{1}{2}}{1-\frac{1}{2}} = 3$		
$\frac{1}{2}$	$\frac{3}{2}$	$\frac{\frac{5}{2}-\frac{3}{2}}{1-\frac{1}{2}} = 2$			
1	$\frac{5}{2}$				

Abbildung 6.3: Hermite-Interpolationskoeffizienten

Das Interpolationspolynom ist also:

$$P(x) = \underline{1} + \underline{1} \cdot (x - 0) \underline{-1} \cdot x \left(x - \frac{1}{2}\right) + \underline{2} \cdot x \left(x - \frac{1}{2}\right)^2 + \underline{4} \cdot x \left(x - \frac{1}{2}\right)^3$$

7 Numerische Integration

7.1 Einführung

Häufig sind Integrale $\int_a^b f(x)\,\mathrm{d}x =: I(f)$ nicht analytisch lösbar (keine explizite Stammfunktion berechenbar), oder aber die Stammfunktion ist sehr kompliziert.

Ziel: Nähere $\int_a^b f(x)\,\mathrm{d}x$ durch einen einfachen Ausdruck an. Diese Aufgabe heißt *numerische Quadratur*.

Wir wollen also die Abbildung

$$I : f \mapsto \int_a^b f(x)\,\mathrm{d}x$$

durch eine Abbildung

$$J : f \mapsto J(f)$$

ersetzen. J sollte

- elementar zu berechnen sein.
- möglichst viele Eigenschaften von I haben.
- die Werte $I(f)$ möglichst gut annähern.

7.2 Eigenschaften des Integrals

f, g seien integrierbare Funktionen, zum Beispiel $f, g : [a, b] \to \mathbb{R}$ Riemann integrierbar.

a) I ist linear:

$$I(\alpha f + \beta g) = \alpha I(f) + \beta I(g)$$

b) I ist monoton:

$$f \leq g \quad \Rightarrow \quad I(f) \leq I(g)$$

Nach Möglichkeit soll J auch die Eigenschaften a) und b) haben.

7.3 Einfache Quadraturformeln

Zerlege das Intervall $[a, b]$ in n Teilintervalle $[x_i, x_{i+1}]$ mit $i \in \{0, \ldots, n-1\}$ und Länge $h_i := x_{i+1} - x_i$:

$$a = x_0 < x_1 < \ldots < x_n = b$$

Ersetze dann auf jeden Teilintervall die Funktion f durch eine einfach zu integrierende Funktion.

a) *Rechteckregel:* Wähle eine konstante Funktion auf jedem Teilintervall.
TODO: Abb14 einfügen
Übliche Wahlen für die Konstante sind der linke/rechte/mittlere Intervallpunkt.
Wählen wir den mittleren Punkt, so ergibt sich:

$$R_M(f) := \sum_{j=0}^{n-1} f\left(\frac{x_{j+1}+x_j}{2}\right)(x_{j+1}-x_j)$$

Sind die Intervalle gleich groß, so definiere:

$$h := \frac{b-a}{n} = x_{j+1} - x_j$$

Dann gilt:

$$R_M(f) = h\sum_{j=0}^{n-1} f\left(x_j + \frac{h}{2}\right)$$

b) *Trapezregel:* Wähle auf jedem Teilintervall eine (affin-)lineare Funktion. Eine (übliche) Möglichkeit ist die Wahl einer affin-linearen Funktion, die die Werte an den Randpunkten verbindet.
TODO: Abb15 einfügen
Auf jedem Teilintervall ergibt sich als Integral:

$$\frac{f(x_i)+f(x_{i+1})}{2}(x_{i+1}-x_i)$$

Damit folgt:

$$T(f) = \sum_{i=0}^{n-1} \frac{f(x_i)+f(x_{i+1})}{2}(x_{i+1}-x_i)$$

Ist die Breite

$$h = x_{i+1} - x_i$$

der Intervalle konstant, so folgt:

$$T(f) = h\left(\frac{1}{2}f(x_0) + \sum_{i=1}^{n-1} f(x_i) + \frac{1}{2}f(x_n)\right)$$

7.4 Interpolatorische Integrationsformeln

Nähere f jetzt durch Polynome höherer Ordnung an.
Seien $x_0, \ldots, x_n$ Stützstellen mit $a \le x_0 < x_1 < \ldots < x_n \le b$ und sei $f := [a,b] \to \mathbb{R}$.
Sei nun $p \in \mathbb{P}_n$ die Funktion, die f an den Stellen $x_0, \ldots, x_n$ interpoliert, das heißt:

$$p(x_i) = f(x_i)$$

In Lagrange-Darstellung heißt das:

$$p(x) = \sum_{j=0}^{n} f(x_j)\, w_j(x)$$

Dabei ist:

$$w_j(x) = \prod_{\substack{i=0 \\ i \neq j}} \frac{x - x_i}{x_j - x_i}$$

Statt

$$\int_a^b f(x)\,\mathrm{d}x$$

berechne:

$$J(f) = \int_a^b p(x)\,\mathrm{d}x = \int_a^b \sum_{j=0}^{n} f(x_j)\, w_j(x)\,\mathrm{d}x = \sum_{j=0}^{n} f(x_j) \underbrace{\int_a^b w_j(x)\,\mathrm{d}x}_{=:\alpha_j}$$

Die $\alpha_0, \ldots, \alpha_n$ sind unabhängig von f und müssen nur einmal berechnet werden.

7.5 **Definition** (Integrationsformel, Quadraturformel)

i) Eine Abbildung $J : C^0([a,b];\mathbb{R}) \to \mathbb{R}$ heißt *Integrationsformel* (*Quadraturformel*), falls

$$J(f) = \sum_{j=0}^{n} \alpha_j f(x_j)$$

mit $\alpha_j \in \mathbb{R}$ und $a \leq x_0 < x_1 < \ldots < x_n \leq b$ gilt.

ii) J heißt *exakt* für eine Menge von Funktionen K, falls für alle $f \in K$ gilt:

$$J(f) = \int_a^b f(x)\,\mathrm{d}x$$

iii) J heißt genau dann *abgeschlossen* (*offen*), wenn $a = x_0$ und $x_n = b$ ($a < x_0$ und $x_n < b$) gilt.

7.6 **Satz** (Charakterisierung interpolatorischer Quadraturformeln)

Sei $J(f) = \sum_{j=0}^{n} a_j f(x_j)$ eine Quadraturformel. Dann sind äquivalent:

i) J ist exakt für Polynome vom Grad n.

ii) Es gilt $a_i = \int_a^b \omega_i(x)\,\mathrm{d}x$ mit:

$$\omega_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{(x - x_j)}{(x_i - x_j)}$$

iii) Es gilt $J = I \circ P$ mit

$$I(f) = \int_a^b f(x)\,\mathrm{d}x$$
$$P : C^0([a,b]) \to P_n$$
$$f \mapsto P(f)$$

und $P(f)$ ist das Interpolationspolynom zu $x_1, \ldots, x_n$.

Beweis

„i) ⇒ ii)“: Da $\omega_i \in \mathbb{P}_n$ ist, ist die Quadraturformel exakt und es folgt:

$$\alpha_i = \sum_{j=0}^{n} \alpha_j \underbrace{\omega_j(x_i)}_{=\delta_{ij}} = J(\omega_i) = I(\omega_i) = \int_a^b \omega_i(x)\,\mathrm{d}x$$

„ii) ⇒ iii)“:

$$\begin{aligned}(I \circ P)(f) &= \int_a^b \sum_{i=0}^{n} f(x_i)\,\omega_i(x)\,\mathrm{d}x = \\ &= \sum_{i=0}^{n} f(x_i)\,f(x_i) \underbrace{\int_a^b \omega_i(x)\,\mathrm{d}x}_{=\alpha_i} = \sum_{i=0}^{n} \alpha_i f(x_i) = J(f)\end{aligned}$$

„iii) ⇒ i)“: Sei $p \in \mathbb{P}_n$, so gilt $P(p) = p$. Es folgt:

$$J(p) = (I \circ P)(p) = I(P(p)) = I(p)$$

□7.6

Bemerkung

Es gibt bei *festen* Stützstellen nur eine Quadraturformel der Eigenschaft i).

7.7 Abgeschlossene Newton-Cotes Formeln

Eine abgeschlossene, interpolatorische Quadraturformel, bei denen die Stützstellen äquidistant gewählt sind, heißen *Newton-Cotes Formel*.
Wir setzen $h = \frac{b-a}{n}$ und $x_j = a + jh$ mit $j \in \{0, \ldots, n\}$. Dann ist

$$Q_n(f) = \sum_{j=0}^{n} \alpha_j f(x_j)$$

mit

$$\alpha_j = \int_a^b \omega_j(x)\,\mathrm{d}x$$

die eindeutig bestimmte interpolatorische Quadraturformel, die Polynome vom Grad n exakt integriert. Diese nennen wir Newton-Cotes Formeln der Ordnung n. Es gilt für $k \in \{0, \ldots, n\}$:

$$\begin{aligned}Q_n\left(x^k\right) &= \int_a^b x^k \mathrm{d}x = \frac{1}{k+1}\left(b^{k+1} - a^{k+1}\right) = \\ &= \sum_{j=0}^{n} \alpha_j^{(n)} x_j^k\end{aligned}$$

Die $\alpha_j^{(n)}$ können also als Lösung des folgenden Gleichungssystems berechnet werden:

$$\begin{pmatrix} 1 & \ldots & 1 \\ x_0 & \ldots & x_n \\ \vdots & & \vdots \\ x_0^n & \ldots & x_n^n \end{pmatrix} \begin{pmatrix} \alpha_0^{(n)} \\ \vdots \\ \vdots \\ \alpha_n^{(n)} \end{pmatrix} = \begin{pmatrix} b-a \\ \frac{b^2-a^2}{2} \\ \vdots \\ \frac{b^{n+1}-a^{n+1}}{n+1} \end{pmatrix}$$

7.8 Bemerkung

i) Quadraturformeln sind linear.

ii) Bei abgeschlossenen Newton-Cotes-Formeln sind einige Werte $a_i^{(n)}$ negativ, falls n groß ist, das heißt aus $f \leq g$ folgt nicht immer $Q_n(f) \leq Q_n(g)$.

7.9 Satz

Sei $Q_n(f) = \sum_{i=0}^{n} \alpha_i f(x_i)$ eine abgeschlossene Newton-Cotes Formel der Ordnung n. Dann gilt:

i) $\sum_{i=0}^{n} \alpha_i = b - a$

ii) $\alpha_i = \alpha_{n-i}$

iii) Ist n gerade, so integriert Q_n Polynome bis zum Grad $n+1$ exakt.

Beweis

i) Es gilt:

$$b - a = \int_a^b 1 \mathrm{d}x = Q_n(1) = \sum_{i=0}^{n} \alpha_i$$

$\square_{\text{i)}}$

ii) Es sei $x_i = a + ih$ und $h = \frac{b-a}{n}$.

Behauptung: $\omega_i(x) = \omega_{n-1}(a + b - x)$

Beweis: Es gilt:

$$a + b - x_j = a + b - a - jh = a + n \cdot \underbrace{\frac{(b-a)}{n}}_{=h} - jh =$$

$$= a + (n - j) \cdot h = x_{n-j}$$

$$\omega_{n-i}(a + b - x_j) = \omega_{n-i}(x_{n-j}) = \delta_{n-i,n-j} = \delta_{ij}$$

Außerdem gilt $\omega_{n-i}(a + b - x) \in \mathbb{P}_n(x)$. Die Eindeutigkeit des Interpolationspolynoms liefert:

$$\omega_{n-i}(a + b - x) = \omega_i(x)$$

$\square_{\text{Behauptung}}$

Es folgt:

$$\alpha_i = \int_a^b \omega_i(x) \, \mathrm{d}x = \int_a^b \omega_{n-i}(\underbrace{a + b - x}_{=:z}) \mathrm{d}x =$$

$$= -\int_b^a \omega_{n-i}(z) \, \mathrm{d}z = \int_a^b \omega_{n-i}(z) \, \mathrm{d}z = \alpha_{n-i}$$

$\square_{\text{ii)}}$

iii) Sei n gerade und $p \in \mathbb{P}_{n+1}$.

$$x_{\frac{n}{2}} = a + \frac{n}{2} \cdot h = a + \frac{n}{2} \cdot \frac{b-a}{n} = a + \frac{b-a}{2} = \frac{a+b}{2}$$

Zerlege p folgendermaßen mit einem Restpolynom $q \in \mathbb{P}_n$.

$$p(x) = a_{n+1} \left(x - x_{\frac{n}{2}}\right)^{n+1} + q(x)$$

Dann gilt:

$$\begin{aligned}\int_a^b p(x)\,\mathrm{d}x &= \int_a^b a_{n+1} \left(x - x_{\frac{n}{2}}\right)^{n+1} \mathrm{d}x + \int q(x)\,\mathrm{d}x = \\ &\overset{z:=x-x_{\frac{n}{2}}}{=} a_{n+1} \underbrace{\int_{-\frac{b-a}{2}}^{\frac{b-a}{2}} z^{n+1} \mathrm{d}z}_{=0} + Q_n(q) = Q_n(p)\end{aligned}$$

Nutze für den letzten Schritt die Symmetrie der α_i, also $\alpha_i = \alpha_{n-i}$ und die Tatsache, dass $n+1$ ungerade ist.

$$\begin{aligned}Q_n\left(\left(x - x_{\frac{n}{2}}\right)^{n+1}\right) &= \sum_{i=0}^{n} \alpha_i \left(x_i - x_{\frac{n}{2}}\right)^{n+1} = \\ &= \sum_{i=0}^{\frac{n}{2}-1} \left(\alpha_i \left(\left(i - \frac{n}{2}\right) h\right)^{n+1} + \alpha_{n-i} \left(\left(n - i - \frac{n}{2}\right) h\right)^{n+1}\right) = \\ &= \sum_{i=0}^{\frac{n}{2}-1} \alpha_i h^{n+1} \left(\left(i - \frac{n}{2}\right)^{n+1} + \left(\frac{n}{2} - i\right)^{n+1}\right) = \\ &= \sum_{i=0}^{\frac{n}{2}-1} \alpha_i h^{n+1} \left(\left(i - \frac{n}{2}\right)^{n+1} - \left(i - \frac{n}{2}\right)^{n+1}\right) = 0\end{aligned}$$

$\square_{\text{iii)}}$

$\square_{7.9}$

7.10 Darstellungsformel für den Integrationsfehler

Sei $J(f) = \sum_{i=0}^{n} \alpha_i f(x_i)$ mit $\alpha_i \in \mathbb{R}$ und $0 \leq x_0 < x_1 < \ldots < x_n \leq b$. Dann ist der Integrationsfehler:

$$R(f) := J(f) - \int_a^b f(x)\,\mathrm{d}x$$

eine lineare Abbildung. Es gilt folgende Fehlerabschätzung.

Satz (Peano)

Für alle $p \in \mathbb{P}_n$ gelte $R(p) = 0$, das heißt alle Polynome $p \in \mathbb{P}_n$ werden exakt integriert. Dann gilt für $f \in C^{n+1}([a,b])$:

$$\begin{aligned} R(f) &= \int_a^b f^{(n+1)}(t)\, K(t)\, \mathrm{d}t \\ K(t) &:= \frac{1}{n!} R\left[(.-t)^n\, \Theta(.-t)\right] = \\ &= \frac{1}{n!}\left(\sum_{i=1}^{n} \alpha_i (x_i - t)^n\, \Theta(x_i - t) - \int_a^b (x-t)^n\, \Theta(x-t)\, \mathrm{d}x\right) \end{aligned}$$

Hierbei ist Θ die Heaviside-Funktion:

$$\Theta(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Verwende im Folgenden die Kurzschreibweise:

$$(.-t)_+^n := (.-t)^n\, \Theta(.-t)$$

Beweis

Wir berechnen die Taylor-Entwicklung um $x = a$.

$$f(x) = f(a) + f'(a) \cdot (x-a) + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n + r_n(x)$$

Für das Restglied gilt:

$$\begin{aligned} r_n(x) &= \frac{1}{n!}\int_a^x f^{(n+1)}(t)\,(x-t)^n\, \mathrm{d}t = \\ &= \frac{1}{n!}\int_a^b f^{(n+1)}(t)\,(x-t)^n\, \Theta(x-t)\, \mathrm{d}t = \\ &= \frac{1}{n!}\int_a^b f^{(n+1)}(t)\,(x-t)_+^n\, \mathrm{d}t \end{aligned}$$

Wende nun R auf die Taylor-Entwicklung an. Da $R(p) = 0$ für alle $p \in \mathbb{P}_n$ gilt, folgt:

$$\begin{aligned} R(f) &= R(r_n) = \\ &= \frac{1}{n!}\left(\int_a^b f^{(n+1)}(t)\,(.-t)_+^n\, \mathrm{d}t\right) = \\ &= \frac{1}{n!}\left(\sum_{i=0}^{n} \alpha_i \int_a^b f^{(n+1)}(t)\,(x_i-t)_+^n\, \mathrm{d}t - \int_a^b \int_a^b f^{(n+1)}(t)\,(x-t)_+^n\, \mathrm{d}t\mathrm{d}x\right) = \\ &\overset{\text{Fubini}}{=} \frac{1}{n!}\left(\int_a^b f^{(n+1)}(t)\left(\sum_{i=0}^{n} \alpha_i (x_i-t)_+^n - \int_a^b (x-t)_+^n\, \mathrm{d}x\right)\mathrm{d}t\right) = \\ &= \frac{1}{n!}\int_a^b f^{(n+1)}(t)\, R\left((.-t)_+^n\right)\mathrm{d}t = \\ &= \int_a^b f^{(n+1)}(t)\, K(t)\, \mathrm{d}t \end{aligned}$$

$\square_{7.10}$

7.11 Fehlerdarstellung bei den Newton-Cotes Formeln

Sei $Q_n(f) = \sum_{i=0}^{n} \alpha_i f(x_i)$ die Newton-Cotes Formel der Ordnung n.

$$R_n(f) := Q_n(f) - \int_a^b f(x)\,\mathrm{d}x$$

Ist n ungerade (gerade), so werden Polynome bis Grad n $(n+1)$ exakt integriert.
Für $Q_n(f)$ sei der Kern K_n:

$$R_n(f) = \int_a^b f^{(n+1+l)}(t)\,K_n(t)\,\mathrm{d}t \tag{7.1}$$

Dabei ist:

$$l := \begin{cases} 0 & \text{falls } n \text{ ungerade} \\ 1 & \text{falls } n \text{ gerade} \end{cases}$$

Dann gilt:

$$K_n(t) = \frac{1}{(n+l)!} R_n\left((.-t)_+^{n+l}\right)$$

Bei den Newton-Cotes-Formeln hat K_n ein konstantes Vorzeichen. Der verallgemeinerte Mittelwertsatz der Integralrechnung liefert (vergleiche (7.1)):

$$R_n(f) = f^{(n+1+l)}(\xi) \cdot \int_a^b K_n(t)\,\mathrm{d}t$$

Hierbei ist $\xi \in (a,b)$. Für die Wahl

$$f = x^{n+1+l}$$

ergibt sich:

$$R_n\left(x^{n+1+l}\right) = (n+1+l)! \int_a^b K_n(t)\,\mathrm{d}t$$

Satz

Im Fall der Newton-Cotes-Formeln haben die Peano-Kerne K_n ein konstantes Vorzeichen. Somit existiert für $f \in C^{n+l+1}([a,b])$ mit l wie oben ein $\xi \in (a,b)$, sodass gilt:

$$R_n(f) = Q_n(f) - \int_a^b f(x)\,\mathrm{d}x = \frac{R_n\left(x^{n+l+1}\right)}{(n+l+1)!} f^{(n+l+1)}(\xi)$$

Beweis

Im Buch *Interpolation* von Johan Steffensen (1965) wird $K_n \geq 0$ gezeigt. Der Rest folgt unter Ausnutzung der Diskussion oben. Siehe dazu Stoer oder Hämmerlin, Hoffmann.

7.12 Beispiel: Die Newton-Cotes Formel für $n = 1$

Für $n = 1$ gilt:

$$Q_n(f) = Q_1(f) = \frac{f(a) + f(b)}{2}(b-a)$$

Jetzt sei $t \in [a, b]$.

$$\begin{aligned} K(t) &= \frac{1}{n!} R\left[(.-t)_+^n\right] = R_x\left((x-t)_+\right) = Q_{1,x}\left((x-t)_+\right) - \int_a^b (x-t)_+ \,\mathrm{d}x = \\ &= (b-a) \cdot \frac{\overbrace{(a-t)_+}^{=0} + (b-t)_+}{2} - \int_t^b (x-t)\,\mathrm{d}x = \\ &= \frac{(b-a)(b-t)}{2} - \left(\frac{b^2}{2} - \frac{t^2}{2} - (b-t)\,t\right) = \frac{1}{2}(b-t)(t-a) \geq 0 \end{aligned}$$

Es folgt, dass der Peano-Kern ein konstantes Vorzeichen hat, und somit ergibt sich für ein $\xi \in (a, b)$:

$$\begin{aligned} R_1(f) &= Q_1(f) - \int_a^b f(x)\,\mathrm{d}x = f^{(2)}(\xi)\,\frac{R_1\left(x^2\right)}{2} \\ R_1\left(x^2\right) &= (b-a)\,\frac{b^2+a^2}{2} - \int_a^b x^2 \mathrm{d}x = \frac{(b-a)^3}{6} \end{aligned}$$

Es folgt:

$$R_1(f) = f^{(2)}(\xi)\,\frac{(b-a)^3}{12} = f^{(2)}(\xi)\,\frac{h^3}{12}$$

Weitere Fehlerdarstellungen sind:

$R_2(f) = \frac{h^5}{90} f^{(4)}(\xi)$	Keplersche Fassregel (Simpson Regel)
$R_3(f) = \frac{3h^5}{80} f^{(4)}(\xi)$	Newtonsche $\frac{3}{8}$-Regel

7.13 Iterierte Newton-Cotes Formeln

Idee: Verwende die Newton-Cotes Formel (oder verwandte Formeln) nicht auf $[a, b]$, sondern unterteile $[a, b]$ in Teilintervalle. (vergleiche 7.3)

Trapezregel: Sei $a = x_0 < x_1 < \ldots < x_n = b$ mit $x_i = a + ih$ und $h = \frac{b-a}{n}$ eine äquidistante Unterteilung von $[a, b]$. Für $i \in \{1, \ldots, n\}$ ist der Wert der Trapezregel auf $[x_{i-1}, x_i]$:

$$T_i(f) = \frac{h}{2}\left(f(x_{i-1}) + f(x_i)\right)$$

Dann gilt für geeignete $\xi_i \in (x_{i-1}, x_i)$:

$$R_i(f) = T_i(f) - \int_{x_{i-1}}^{x_i} f(x)\,\mathrm{d}x = \frac{h^3}{12} f^{(2)}(\xi_i)$$

Die Iterierte Trapezregel für $f \in C^2([a,b])$ ist:

$$\hat{T}_n(f) = \sum_{i=1}^{n} T_i(f)$$

Der dabei gemachte Fehler besitzt die Darstellung:

$$\hat{T}_n(f) - \underbrace{\int_a^b f(x)\,\mathrm{d}x}_{=\sum_{i=1}^{n}\int_{x_{i-1}}^{x_i} f(x)\mathrm{d}x} = \sum_{i=1}^{n} \frac{h^3}{12} f^{(2)}(\xi_i) = (b-a)\frac{h^2}{12}\frac{1}{n}\sum_{i=1}^{n} f^{(2)}(\xi_i) =$$

$$= (b-a)\frac{h^2}{12} f^{(2)}(\xi)$$

Hier ist $\xi \in (a,b)$ geeignet und:

$$h = \frac{b-a}{n}$$

Die letzte Identität folgt aus dem Zwischenwertsatz:

$$\min\left(f^{(2)}\right) \leq \frac{1}{n}\sum_{i=1}^{n} f^{(2)}(\xi_i) \leq \max\left(f^{(2)}\right)$$

Die iterierte Trapezregel konvergiert für $h = \frac{b-a}{n} \to 0$ quadratisch gegen Null.

7.14 Euler-MacLaurinsche Summenformel

Für Funktionen $f \in C^{2m+2}([a,b])$ gilt die folgende Entwicklung für die iterierte Trapezregel:

Satz

Es sei $f \in C^{2m+2}([a,b])$ und $\hat{T}_n(f)$ die iterierte Trapezregel für äquidistante Knoten $x_i = a+ih$ mit $h = \frac{b-a}{n}$, das heißt:

$$\hat{T}_n(f) = h\left(\frac{1}{2}f(a) + \sum_{i=1}^{n-1} f(a+ih) + \frac{1}{2}f(b)\right)$$

Dann gilt:

$$\hat{T}(f) = \underbrace{\int_a^b f(x)\,\mathrm{d}x}_{=:\tau_0} + \tau_1 h^2 + \tau_2 h^4 + \ldots + \tau_m h^{2m} + \alpha_{m+1}(h)\,h^{2m+2}$$

Dabei ist für $k \in \{1,\ldots,m\}$ nun

$$\tau_k = \frac{B_{2k}}{(2k)!}\left(f^{(2k-1)}(b) - f^{(2k-1)}(a)\right)$$

mit den Bernulli-Zahlen B_{2k} und

$$\alpha_{m+1}(h) = \frac{B_{2m+2}}{(2m+2)!}(b-a)\,f^{(2m+2)}(\xi(h)) \tag{7.2}$$

mit $\xi(h) \in (a,b)$.

Beweis

Siehe Stoer oder Hämmerlin, Hoffmann.

Beispiel

$$\hat{T}_h(f) = \tau_0 + \tau_1 h^2 + \tau_2 h^4 + \dots$$
$$\hat{T}_{\frac{h}{2}}(f) = \tau_0 + \tau_1 \left(\frac{h}{2}\right)^2 + \tau_2 \left(\frac{h}{2}\right)^4 + \dots$$

Versuche eine bessere Näherung durch eine Linearkombination zu erhalten:

$$\alpha_0 \hat{T}_h + \alpha_1 \hat{T}_{\frac{h}{2}} = (\alpha_0 + \alpha_1)\tau_0 + \tau_1 h^2 \left(\alpha_0 + \frac{\alpha_1}{4}\right) + \dots$$

Wähle dazu:

$$\alpha_0 + \alpha_1 = 1$$
$$\alpha_0 + \frac{\alpha_1}{4} = 0$$

$$\Rightarrow \quad \alpha_0 \hat{T}_h + \alpha_1 \hat{T}_{\frac{h}{2}} = \int_a^b f(x)\,\mathrm{d}x + Ch^4 + \dots$$

7.15 Idee der Extrapolation

Definiere:

$$q(x) = \tau_0 + \tau_1 x + \tau_2 x^2 + \dots + \tau_m x^m$$

Es gilt:

i) $\hat{T}_h(f) = q\left(h^2\right) + \mathcal{O}_0\left(h^{2m+2}\right)$

ii) $q(0) = \tau_0 = \int_a^b f(x)\,\mathrm{d}x$

Anmerkung: Angenommen wir kennen $\hat{T}_h$ (als Funktion von h) an den Stellen

$$h_0 = \frac{b-a}{n_0} > h_1 = \frac{b-a}{n_1} > \dots > h_m = \frac{b-a}{n_m}$$

mit $n_0, \dots, n_m \in \mathbb{N}$. Dann existiert genau ein Polynom $p \in \mathbb{P}_m$, sodass für $i \in \{0, \dots, m\}$ gilt:

$$p\left(h_i^2\right) = \hat{T}_{h_i}(f)$$

Erwartung:

- p und q liegen nahe beieinander, da sie an den Stellen $h_0, \dots, h_m$ bis auf $\mathcal{O}_0\left(h_j^{2m+2}\right)$ für $j \in \{0, \dots, m\}$ übereinstimmen.
- $p(0) \approx \int_a^b f(x)\,\mathrm{d}x$

Wir interessieren uns für $p(0)$. Nutze das Neville-Aitken Verfahren, um $p(0)$ zu berechnen.

Bemerkung

$m = 1$ entspricht obigem Vorgehen mit α_0, α_1.
TODO: Abb17 einfügen

7.16 Romberg-Verfahren

Zur Erinnerung: $P(g|x_j, \ldots, x_k)$ ist das Interpolationspolynom zu den Punkten (x_i, g_i) mit $g_i = g(x_i)$. Dann gilt:

$$P(g|x_j, \ldots, x_k)(x) = \frac{(x - x_j) P(g|x_{j+1}, \ldots, x_k)(x) + (x_k - x) P(g|x_j, \ldots, x_{k-1})(x)}{x_k - x_j}$$

Hier gilt:

$$x_i = h_i^2 \qquad g_i = T_i := \hat{T}_{h_i}(f) \qquad x = 0$$

Zur Abkürzung:

$$T_{k,i} := P(g|x_{k-i}, \ldots, x_k)(0)$$

Uns interessiert $T_{m,m}$. Neville-Aitken liefert (oben $j = k - i$):

$$\begin{aligned} T_{k,i} &= \frac{-x_{k-i} T_{k,i-1} + x_k T_{k-i,i-1}}{x_k - x_{k-i}} = \\ &= T_{k,i-1} + \frac{-x_k}{x_k - x_{k-i}} (T_{k,i-1} - T_{k-i,i-1}) = \\ &= T_{k,i-1} + \frac{1}{\frac{x_{k-i}}{x_k} - 1} (T_{k,i-1} - T_{k-i,i-1}) \end{aligned}$$

Es folgt für $i \in \{1, \ldots, m\}$ und $k \in \{1, \ldots, m\}$:

$$T_{k,i} = T_{k,i-1} + \frac{1}{\left(\frac{h_{k-i}}{h_k}\right)^2 - 1} (T_{k,i-1} - T_{k-1,i-1})$$

Nutze das Dreiecksschema:

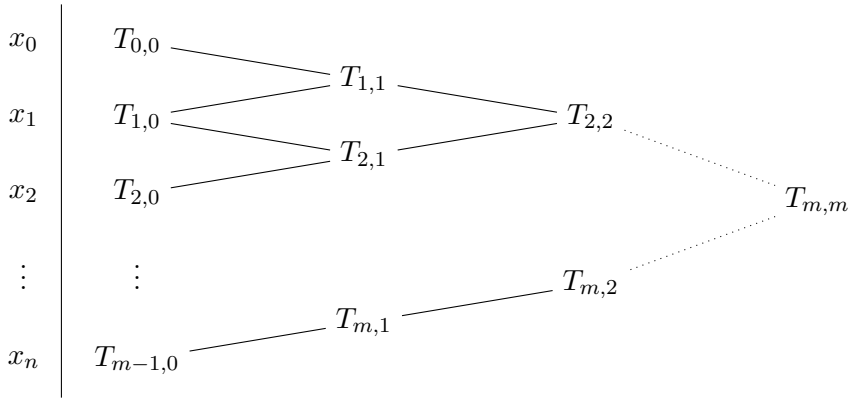


Abbildung 7.1: Extrapolationsschema

7.17 Wahl der Schrittweite

a) Romberg Folge: Die einfachste Wahl ist die iterative Intervallhalbierung in 2^j gleichgroße Teilintervalle für $j \in \{0, \ldots, m\}$. Dann hat man die *Romberg-Folge*:

$$h_j = \frac{b-a}{2^j}$$

$$T_{k,i} = T_{k,i-1} + \underbrace{\frac{1}{2^{2i}-1}}_{=\frac{1}{4^i-1}} \left(T_{k,i-1} - T_{k-1,i-1}\right)$$

Für T_{h_j} nutze $T_{h_{j-1}}$.
TODO: Abb18 einfügen

$$\begin{aligned}
\hat{T}_{\frac{h}{2}} &= \frac{h}{4}\left(f(a) + 2\sum_{i=1}^{2n-1} f\left(a + i\frac{h}{2}\right) + f(b)\right) = \\
&= \frac{h}{4}\left(f(a) + 2\sum_{k=1}^{n-1} f(a+kh) + f(b)\right) + \frac{h}{2}\sum_{k=1}^{n} f\left(a + \frac{2k-1}{2}h\right) = \\
&= \frac{1}{2}\hat{T}_h + \frac{h}{2}\sum_{k=1}^{n} f\left(a + \frac{2k-1}{2}h\right)
\end{aligned}$$

$\hat{T}_h$ ist schon bekannt!

b) Bulirsch-Folge: Die zweite übliche Wahl ist:

$$\begin{aligned}
h_{2j} &= 2^{-j}(b-a) \\
h_{2j+1} &= \frac{1}{3} \cdot 2^{-j} \cdot (b-a)
\end{aligned}$$

Die Folge lautet:

$$(b-a), \frac{b-a}{3}, \frac{b-a}{4}, \frac{b-a}{6}, \frac{b-a}{8}, \frac{b-a}{12}, \ldots$$

Der Vorteil dabei ist, dass die Schrittweite nicht so schnell klein wird. Bei der Berechnung von $T(h_j)$ kann $T(h_{j-2})$ benutzt werden (vergleiche a)).

c) Für glatte Funktionen f wird in der Praxis das Romberg-Verfahren benutzt, um Integrale näherungsweise zu berechnen. $T_{m,m}$ ist in der Regel deutlich näher an dem Integral $\int_a^b f(x)\,\mathrm{d}x$, als $T_{m,0}$.

7.18 Verfahren und Abbruch

In der Praxis ist f häufig nicht beliebig oft differenzierbar. In die Fehlerdarstellung (7.2) geht aber die Differenzierbarkeitsordnung $2m+2$ ein. Falls m unbekannt oder falls f nicht so oft differenzierbar ist, gehe wie folgt vor: Wähle $m_{\max}$, das heißt wir bestimmen iterativ $T_{0,0}, T_{1,1}, \ldots, T_{m_{\max},m_{\max}}, \ldots, T_{i,m_{\max}}$.

$$\begin{array}{cccc} T_{0,0} & & & \\ T_{1,0} & T_{1,1} & & \\ \vdots & & \ddots & \\ T_{m_{\max},0} & \cdots & \cdots & T_{m_{\max},m_{\max}} \\ T_{m_{\max}+1,0} & \cdots & \cdots & T_{m_{\max}+1,m_{\max}} \\ \vdots & & & \vdots \\ T_{i,0} & \cdots & \cdots & T_{i,m_{\max}} \end{array}$$

$T_{i,m_{\max}}$ wird deutlich besser sein, als $T_{m_{\max},m_{\max}}$, da kleinere h_j-Werte genutzt werden.
Zum Beispiel: Setze $m_{\max} = 7$ und breche ab, falls:

a) i zu groß wird.

b) genügend viele Ziffern „stehen“, das heißt genügend viele führende Ziffern sich nicht mehr ändern, oder

$$|T_{i,m_{\max}} - T_{i+1,m_{\max}}| \leq \varepsilon \cdot |\tau|$$

für sehr kleines $\varepsilon \in \mathbb{R}_{>0}$ gilt. Dabei ist τ ein Schätzwert für das Integral, zum Beispiel $\tau = T_{i+1,m_{\max}}$.

7.19 Idee der Gauß-Quadratur

Bisher waren die Stützstellen x_j äquidistant.
Jetzt wählen wir in $J(f) = \sum_{j=0}^{n} \alpha_j f(x_j)$ auch die $x_0, \ldots, x_j$ geeignet. Die Gaußquadratur wählt $\alpha_0, \ldots, \alpha_n$ und $x_0, \ldots, x_n$, sodass Polynome möglichst hohen Grades exakt integriert werden. Wir hoffen, Polynome des Grades $2n + 1$ exakt integrieren zu können.

Verallgemeinere die Aufgabenstellung

Betrachte gewichtete Integrale:

$$I(f) := \int_a^b f(x)\,\omega(x)\,\mathrm{d}x$$

Dabei können a und b auch $-\infty$ oder ∞ sein. ω ist eine nicht negative Gewichtsfunktion.

7.20 Gewichtsfunktionen

Wir stellen folgende Voraussetzungen an ω:

a) ω ist auf (a, b) nicht negativ und stückweise stetig (messbar).

b) Alle Momente

$$\mu_k := \int_a^b x^k \omega(x)\,\mathrm{d}x$$

für $k \in \mathbb{N}_{\geq 0}$ sind endlich.

c) Für jedes Polynom p mit $P(x) \geq 0$ für alle $x \in [a, b]$ und

$$\int_a^b \omega(x)\,P(x)\,\mathrm{d}x = 0$$

folgt $P(x) = 0$ für alle $x \in \mathbb{R}$.

Diese Voraussetzungen sind für positives und stetiges ω auf einem endlichen Intervall erfüllt.
Falls a) und b) erfüllt sind, lässt sich c) äquivalent durch $\mu_0 > 0$ ausdrücken.
Typische Gewichtsfunktionen sind:

$\omega(x)$	Intervall
1	$[-1,1]$
e^{-x}	$[0,\infty)$
$\frac{1}{\sqrt{1-x^2}}$	$(-1,1)$
e^{-x^2}	$(-\infty,\infty)$

Zu einem gegebenen ω führen wir das Skalarprodukt

$$\langle f,g\rangle_\omega := \int_a^b \omega(x)\, f(x)\, g(x)\,\mathrm{d}x$$

ein. Dies ist definiert für alle messbaren Funktionen $f,g:[a,b]\to\mathbb{R}$, für die die Norm

$$\|f\| := \sqrt{\int_a^b \omega(x)\, f^2(x)\,\mathrm{d}x}$$

endlich ist. Funktionen f,g heißen *orthogonal*, falls $\langle f,g\rangle_\omega = 0$ gilt.
Ziel: Finde zu gegebenen ω und n die Werte $(x_i,\alpha_i)_{i\in\{0,\dots,n\}}$, sodass Polynome möglichst hohen Grades durch $\sum_{i=0}^n \alpha_i f(x_i)$ exakt integriert werden.
Das Mittel dazu sind Orthogonalpolynome.

7.21 Lemma

Es sei

$$\hat{I}_n(f) = \sum_{i=0}^n \alpha_{i,n} f(x_{i,n})$$

für alle $p\in\mathbb{P}_{2n+1}$ exakt, das heißt für alle $p\in\mathbb{P}_{2n+1}$ gilt:

$$I(p) = \int_a^b p(x)\,\omega(x)\,\mathrm{d}x = \hat{I}_n(p)$$

Dann ist

$$p_{n+1}(x) := (x-x_{0,n})\cdot\ldots\cdot(x-x_{n,n}) \in \mathbb{P}_{n+1}$$

orthogonal zu allen $p\in\mathbb{P}_n$ bezüglich $\langle .,.\rangle_\omega$.

Beweis

Sei $p \in \mathbb{P}_n$, also $p \cdot p_{n+1} \in \mathbb{P}_{2n+1}$, dann folgt:

$$\langle p, p_{n+1} \rangle_\omega = I\left(p \cdot p_{n+1}\right) = \hat{I}_n\left(p \cdot p_{n+1}\right) = \sum_{i=0}^{n} \alpha_{i,n} p\left(x_{i,n}\right) \underbrace{p_{n+1}\left(x_{i,n}\right)}_{=0} = 0$$

$\square_{7.21}$

Die gesuchten Knoten $(x_{i,n})_{n\in\mathbb{N},i\in\{1,\ldots,n\}}$ müssen also Nullstellen zueinander orthogonaler Polynome sein.

Wir suchen also orthogonale Polynome $p_0, p_1, p_2, \ldots$ mit $\deg\left(P_j\right) = j$.

Fragen:

1. Existieren solche P_j?
2. Sind die Nullstellen einfach und reell?
3. Wie sind die $\alpha_{i,n}$ zu wählen?
4. Ist dann bei guter Wahl $\hat{I}_n\left(p\right)$ exakt für alle $\mathbb{P}_{2n+1}$?

7.22 **Satz** (Existenz von Orthogonalpolynomen)

Es gibt für $j \in \mathbb{N}_{\geq 0}$ eindeutig bestimmte, normierte Polynome $p_j \in \mathbb{P}_j$ mit der Eigenschaft $\langle p_j, p_k \rangle_\omega = 0$ für $j \neq k$.

Diese Polynome genügen der 3-Term Rekursionsformel mit $p_{-1} := 0$ und $\gamma_0 := 1$:

1. $p_0\left(x\right) = 1$
2. $p_{j+1}\left(x\right) = \left(x - \delta_i\right) p_i\left(x\right) - \gamma_i p_{i-1}\left(x\right)$

$$\delta_i := \frac{\langle x \cdot p_i, p_i \rangle_\omega}{\langle p_i, p_i \rangle_\omega} \qquad\qquad \gamma_i := \frac{\langle p_i, p_i \rangle_\omega}{\langle p_{i-1}, p_{i-1} \rangle_\omega}$$

Beweis

Bestimme die Polynome rekursiv analog zum Gram-Schmidschen Orthogonalisierungsverfahren. Offensichtlich ist $p_0 = 1$.

Nehme an, dass $p_j \in \mathbb{P}_j$ normiert sind mit $j \geq i$ und den gewünschten Eigenschaften existiert.

Zeige, dass ein normiertes $p_{i+1} \in \mathbb{P}_{i+1}$ existiert, sodass für $j \leq i$ nun

$$\langle p_{i+1}, p_j \rangle_\omega = 0$$

und die Rekursionsformel gilt. Da die Polynome $p_0, \ldots, p_i$ normiert sind, gilt

$$p_{i+1}\left(x\right) = \left(x - \delta_{i+1}\right) p_i\left(x\right) + c_{i-1} p_{i-1}\left(x\right) + \ldots + c_0 p_0\left(x\right)$$

mit eindeutig bestimmten δ_{i+1} und c_k für $k \leq i-1$. Da für alle $j \neq k$ und $j, k \leq i$ nun $\langle p_j, p_k \rangle_\omega = 0$ gilt $\langle p_{i+1}, p_j \rangle_\omega = 0$ für $j \leq i$ genau dann ,wenn gilt:

$$\langle p_{i+1}, p_i \rangle_\omega = \langle x p_i, p_i \rangle_\omega - \delta_{i+1} \langle p_i, p_i \rangle_\omega = 0$$

Für $j \leq i$ gilt:

$$\langle p_{i+1}, p_{j-1} \rangle_\omega = \underbrace{\langle x \cdot p_{j-1}, p_i \rangle_\omega}_{=0 \text{ für } j<i} + c_{j-1} \langle p_{j-1}, p_{j-1} \rangle_\omega \overset{!}{=} 0$$

Damit folgt für $j < i$:

$$c_{j-1} = 0$$

Außerdem gilt:

$$\langle xp_{i-1}, p_i \rangle_\omega = -c_{i-1} \langle p_{i-1}, p_{i-1} \rangle_\omega$$

Es folgt die Formel für $\gamma_i = -c_{i-1}$ und somit:

$$p_{i+1} = (x - \delta_{i+1})\, p_i - \gamma_i p_{i-1}$$

Dabei ist δ_{i+1} wie in der Behauptung. $\square_{7.22}$

Bemerkung

i) Da die Polynome $p_0, \ldots, p_n$ orthogonal und ungleich Null sind, sind sie auch linear unabhängig und bilden eine Basis von $\mathbb{P}_n$.

ii) Für alle $p \in \mathbb{P}_n$ gilt $\langle p, p_{n+1} \rangle_\omega$. Dies folgt aus i).

7.23 Satz

Die Nullstellen $x_0, \ldots, x_{n-1}$ von p_n sind reell, einfach und liegen im offenen Intervall (a, b).

Beweis

TODO: Abb19 einfügen

Seien $a < x_0 < x_1 < \ldots < x_l < b$ die Nullstellen von p_n, an denen p_n das Vorzeichen wechselt, das heißt wir betrachten reelle Nullstellen mit ungerader Vielfachheit in (a, b). Zu zeigen ist $l = n - 1$.

Annahme: $l = n - 1$

Definiere:

$$q(x) := \prod_{j=0}^{l} (x - x_j) \in \mathbb{P}_l$$

Wegen $\deg(q) < n$ folgt $\langle p_n, q \rangle_\omega = 0$. Andererseits ändert $p_n \cdot q$ das Vorzeichen nicht. Es folgt:

$$\langle p_n, q \rangle_\omega = \int_a^b \underbrace{\omega(x)}_{\geq 0} \underbrace{p_n(x)\, q(x)}_{\text{sgn}(\ldots) = \text{konst.}} \, \mathrm{d}x \neq 0$$

Dies ist ein Widerspruch zu 7.20 c). $\square_{7.23}$

7.24 Beispiele für orthogonale Polynomsysteme

a) Für $\omega(x) = 1$ und $[a, b] = [-1, 1]$ sind die Legendre-Polynome orthogonal. Sie sind durch folgende Formel definiert:

$$p_k(x) = \frac{k!}{(2k)!} \frac{\mathrm{d}^k}{\mathrm{d}x^k} \left(x^2 - 1\right)^k$$

Es gilt die Rekursionsformel:

$$p_{k+1}(x) = xp_k(x) - \frac{k^2}{4k^2-1}p_{k-1}(x)$$

$$p_0(x) = 1 \qquad p_1(x) = x$$

b) Für $\omega(x) = \frac{1}{\sqrt{1-x^2}}$ auf $[-1,1]$ erhält man die Tschebyscheff-Polynome:

$$T_n(x) = \cos(n \arccos(x))$$

Rekursion:

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$$

c) Für $\omega(x) = e^{-x}$ ist und $[a,b) = [0,\infty)$ erhält man die Laguerre-Polynome:

$$\Lambda_n(x) = (-1)^n e^x \frac{\mathrm{d}^n}{\mathrm{d}x^n}\left(x^n e^{-x}\right)$$

d) Für $\omega(x) = e^{-x^2}$ und $(a,b) = (-\infty,\infty)$ erhalten wir die Hermite-Polynome:

$$H_n(x) = \frac{(-1)^n}{2^n} e^{x^2} \frac{\mathrm{d}^n}{\mathrm{d}x^n}\left(e^{-x^2}\right)$$

Betrachten wir im Fall a) die Nullstellen des Legendre-Polynoms für die Integrationsformel, so spricht man von Gauß-Legendre Quadratur. (Analog: b) Gauß-Tschebyscheff, c) Gauß-Laguerre, d) Gauß-Tschebyscheff)

Bemerkung

i) Für $\omega(x) = 1$ und ein endliches Intervall (a,b) ist das Romberg-Verfahren für glatte Funktionen ohne Singularität in der Regel besser als das Gauß-Legendre-Verfahren.

ii) Das Gauß-Tschebyscheff-Verfahren bietet sich an, wenn der Integrand bei $x = \pm 1$ (schwach) singulär ist. Zerlege dazu die Funktion g in $g = f\omega$.

iii) Das Gauß-Laguerre-Verfahren und das Gauß-Tschebyscheff-Verfahren sind wichtig für Integrale auf unbeschränkten Intervallen.

7.25 Bestimmung der Gewichte α_i

Die Gewichte $(\alpha_i)_{i\in\{0,\dots,n\}}$ für die Gauß-Quadratur

$$\sum_{i=0}^{n} \alpha_i f(x_i)$$

werden für die Nullstellen $x_0,\dots,x_n$ des Orthogonalpolynoms eindeutig bestimmt durch:

$$\alpha_i = \int_a^b \omega(x) L_i(x)\,\mathrm{d}x \tag{7.3}$$

Dabei sind die $(L_i)_{i\in\{0,\dots,n\}}$ die zu $x_0,\dots,x_n$ gehörenden Lagrange-Interpolationspolynome. (Setze L_i in die Gauß-Quadraturformel ein und nutze aus, dass die $L_i \in \mathbb{P}_n$ exakt integrierbar sind.) Genaue Werte für die Gewichte finden sie in der Literatur. (Stoer oder Hämmerlin, Hoffmann)

Wir bestimmen die α_i nun etwas anders.

Bemerkungen

Für die Quadratur auf dem Intervall $[a, b]$ mit $a, b \in \mathbb{R}$ transformiere die Stützstellen auf $[-1, 1]$ durch eine affin-lineare Abbildung nach $[a, b]$. Entsprechend transformiere $[a, \infty)$ nach $[0, \infty)$.

7.26 Satz

Es sei $(p_i)_{i\in\mathbb{N}}$ ein normiertes orthogonales Polynomsystem.

1. Seien $x_0, \ldots, x_n$ die Nullstellen von p_{n+1} und $\alpha_0, \ldots, \alpha_n$ seien die Lösungen des linearen Gleichungssystems:

$$\sum_{i=0}^{n} \alpha_i p_k(x_i) = \begin{cases} \langle p_0, p_0 \rangle_\omega & \text{falls } k = 0 \\ 0 & \text{falls } k \in \{1, \ldots, n\} \end{cases} \tag{7.4}$$

   Dann gilt für $i \in \{0, \ldots, n\}$

$$\alpha_i > 0$$

   und für alle $p \in \mathbb{P}_{2n+1}$:

$$\int_a^b \omega(x)\, p(x)\, \mathrm{d}x = \sum_{i=0}^{n} \alpha_i p(x_i) \tag{7.5}$$

2. Gilt umgekehrt (7.5) für Zahlen x_i, α_i für $i \in \{0, \ldots, n\}$, so sind die x_i Nullstellen von p_{n+1} und die α_i erfüllen (7.4).
3. Es gibt keine reellen Zahlen x_i, α_i, sodass (7.5) für alle $p \in \mathbb{P}_{2n+2}$ gilt.

Beweis

1. Es gilt:

$$I(p_k) = \langle p_k, p_0 \rangle_\omega = \begin{cases} \langle p_0, p_0 \rangle_\omega & \text{für } k = 0 \\ 0 & \text{für } k \neq 0 \end{cases}$$

$$\hat{I}_n(p_k) = \sum_{i=0}^{n} \alpha_i p_k(x_i)$$

   Da $(p_0, \ldots, p_n)$ eine Basis von $\mathbb{P}_n$ ist, folgt:
   (7.4) ist äquivalent dazu, dass $\hat{I}_n$ für alle $p \in \mathbb{P}_n$ exakt ist.
   Damit ist (7.4) äquivalent zu (7.3) und eindeutig lösbar. Sei nun $p \in \mathbb{P}_{2n+1}$, so folgt durch Polynomdivision die Restklassendarstellung mit geeigneten $q, r \in \mathbb{P}_n$:

$$p = p_{n+1} q + r$$

   Daher existieren $\alpha_k, \beta_k \in \mathbb{R}$ mit:

$$q(x) = \sum_{k=0}^{n} \alpha_k p_k(x)$$

$$r(x) = \sum_{k=0}^{n} \beta_k p_k(x)$$

Wegen $p_0 = 1$ und der Orthogonalität folgt:

$$\int_a^b \omega(x)\, p(x)\, \mathrm{d}x = \underbrace{\langle p_{n+1}, q\rangle_\omega}_{=0} + \langle r, p_0\rangle_\omega = \beta_0 \langle p_0, p_0\rangle_\omega$$

Andererseits gilt wegen $p_{n+1}(x_i) = 0$ für $i \in \{0, \ldots, n\}$ nun:

$$\sum_{i=0}^{n} \alpha_i p(x_i) = \sum_{i=0}^{n} \alpha_i r(x_i) = \sum_{k=0}^{n} \beta_k \sum_{i=0}^{n} \alpha_i p_k(x_i) \overset{(7.4)}{=} \beta_0 \langle p_0, p_0\rangle_\omega$$

Dies zeigt (7.5).
Sei nun für $l \in \{0, \ldots, n\}$:

$$0 \leq \overline{p}_l(x) := \prod_{\substack{j=0 \\ j\neq l}}^{n} (x - x_j)^2 \in \mathbb{P}_{2n}(x)$$

Es folgt:

$$0 \overset{\substack{\text{vergleiche Voraussetzung c)} \\ \text{an die Gewichte}}}{<} \int_a^b \omega(x)\, \overline{p}_l(x)\, \mathrm{d}x = \sum_{i=0}^{n} \alpha_i \overline{p}_l(x_i) = \alpha_l \underbrace{\overline{p}_l(x_l)}_{>0}$$

$\square_{1.}$

3. Nehme an, es existieren x_i, α_i so, dass die Quadratur für alle $p \in \mathbb{P}_{2n+2}$ exakt ist. Wähle:

$$0 \leq \overline{p}(x) := \prod_{j=0}^{n} (x - x_j)^2 \in \mathbb{P}_{2n+2}(x)$$

Es folgt:

$$0 < I(\overline{p}) = \hat{I}_n(\overline{p}) = \sum_{i=0}^{n} \alpha \overline{p}(x_i) = 0$$

Dies ist ein Widerspruch. $\square_{1.}$

2. Da $p_k \in \mathbb{P}_{2n+1}$ für $k \in \{0, \ldots, n\}$ gilt wie in 1.), dass (7.4) richtig ist.
Außerdem folgt wie in 1. $\alpha_i > 0$. Zu zeigen ist , dass die x_i Nullstellen von p_{n+1} sind. Es gilt:

$$0 \overset{k\in\{0,\ldots,n\}}{=} \langle p_{n+1}, p_k\rangle_\omega = \int_a^b \omega(x) \underbrace{p_{n+1}(x)\, p_k(x)}_{\in \mathbb{P}_{2n+1}} \mathrm{d}x = \sum_{i=0}^{n} \alpha_i p_{n+1}(x_i)\, p_k(x_i) \qquad (7.6)$$

Setze:

$$A = (p_k(x_i))_{k,i\in\{0,\ldots,n\}}$$
$$c = \begin{pmatrix} \alpha_0 p_{n+1}(x_0) \\ \vdots \\ \alpha_n p_{n+1}(x_n) \end{pmatrix}$$

Also ist (7.4) äquivalent zu $Ac = 0$. Falls A regulär ist, folgt $c = 0$ und wegen $\alpha_i > 0$ folgt $p_{n+1}(x_i) = 0$.
Beobachte zunächst $x_0 < x_1 < \ldots < x_n$ (siehe oben). Dass A regulär ist, folgt aus dem folgenden Satz. $\square_{1.}$

$\square_{7.26}$

7.27 Satz

Seien $\{p_0, \ldots, p_n\}$ eine Basis von $\mathbb{P}_n$, so ist für beliebige $x_0 < \ldots < x_n$ die Matrix

$$A = \begin{pmatrix} p_0(x_0) & \ldots & p_0(x_n) \\ \vdots & & \vdots \\ p_n(x_0) & \ldots & p_n(x_n) \end{pmatrix}$$

nicht singulär ist.

Beweis

TODO: Rest (Beweis) einfügen (evtl. siehe Zentralübung)

8 Iterationsverfahren zur Lösung linearer Gleichungssysteme

8.1 Einführung

Bisher lösten wir $Ax = b$ für $A \in \mathbb{R}^{n\times n}$ und $b \in \mathbb{R}^n$ mittels einer direkten Methode (zum Beispiel der LR-Zerlegung, oder der QR-Zerlegung). Die Lösung x wird nach einer bestimmten Anzahl von Schritten durch einen Algorithmus explizit berechnet.

In vielen Anwendungen ist A eine sehr große, dafür aber dünn besetzte Matrix, das heißt viele Einträge sind Null. Direkte Verfahren zerstören häufig diese Eigenschaft. Dies führt zu einem unnötig hohen Aufwand.

Auswege:

a) Passe direkte Verfahren an, zum Beispiel pivotisiere so geschickt, dass wenig neue Nichtnullelemente auftreten, oder nutze Givens-Rotationen bei QR-Zerlegung.

b) Nutze iterative Verfahren und berechne eine Näherungslösung (vergleiche mit der Lösung nichtlinearer Gleichungen).

Betrachte zunächst ein Beispiel einer Matrix mit vielen Null-Einträgen:

8.2 Diskretisierung der Poisson-Gleichung

Für eine offene Teilmenge $\Omega \subseteq \mathbb{R}^d$ und eine gegebene Funktion $v : \Omega \to \mathbb{R}$ definiere den Laplace-Operator:

$$\Delta v = \sum_{i=1}^{d} \frac{\partial^2}{(\partial x_i)^2} v$$

Gesucht ist eine Funktion $u : \overline{\Omega} \to \mathbb{R}$, für die die *Poisson-Gleichung*

$$\begin{aligned} -\Delta u &= f && \text{auf } \Omega \\ u &= g && \text{auf } \partial\Omega \end{aligned}$$

erfüllt ist. Dabei sind $f : \Omega \to \mathbb{R}$ und $g : \partial\Omega \to \mathbb{R}$ gegebene Funktionen.

Ziel: Bestimme u näherungsweise.

Zur Vereinfachung sei nun $\Omega = (0,1)^2$ das Einheitsquadrat. Führe ein regelmäßiges quadratisches Gitter mit einer Schrittweite $h = \frac{1}{n}$ für $n \in \mathbb{N}_{\geq 1}$ ein.

$$\Omega_h := \left\{ (ih, jh) \in (0,1)^2 \,\middle|\, 1 \leq i, j \leq n-1 \right\}$$

$$\overline{\Omega_h} := \left\{ (ih, jh) \in (0,1)^2 \,|0 \leq i,j \leq n \right\}$$

TODO: Abb20 einfügen

Wie komme ich an Gleichungen?
Eine Taylor-Entwicklung in x für eine glatte Funktion $u : \Omega \to \mathbb{R}$ liefert:

$$\frac{\partial^2}{\partial x^2} u(x,y) = \frac{u(x-h,y) - 2u(x,y) + u(x+h,y)}{h^2} + \mathcal{O}\left(h^2\right)$$

Daraus ergibt sich für $(x,y) \in \Omega_h$:

$$\begin{aligned} -\Delta u(x,y) &= -\left(\frac{\partial^2}{\partial x^2} u(x,y) + \frac{\partial^2}{\partial y^2} u(x,y) \right) = \\ &= \underbrace{\frac{u(x-h,y) + u(x+h,y) + u(x,y-h) + u(x,y+h) - 4u(x,y)}{h^2}}_{=:\Delta_h u(x,y)} + \mathcal{O}\left(h^2\right) \end{aligned}$$

TODO: Abb21 einfügen

Es seien nun $\ell^2(\Omega_h)$ und $\ell^2\left(\overline{\Omega_h}\right)$ die Menge der Gitterfunktionen $u_h : \Omega_h \to \mathbb{R}$ beziehungsweise $u_h : \overline{\Omega}_h \to \mathbb{R}$.

Diskretes Poisson-Problem (DP)

Gesucht ist eine Funktion $u_h \in \ell^2\left(\overline{\Omega_h}\right)$ mit:

$$\begin{aligned} -\Delta_h u_h(\xi) &= f(\xi) && \text{für } \xi \in \Omega_h \\ -\Delta_h u_h(\xi) &= g(\xi) && \text{für } \xi \in \overline{\Omega}_h \setminus \Omega_h \end{aligned}$$

Dies ist ein lineares Gleichungssystem für die Werte $(u_h(\xi))_{\xi \in \overline{\Omega}_h}$ der Gitterfunktion u_h. Die Werte $u_h(\xi)$ mit $\xi \in \overline{\Omega}_h \setminus \Omega_h$ sind direkt bekannt. Bringe jetzt die Gleichungen für $u_h(\xi)$ mit $\xi \in \Omega_h$ auf die Form eines linearen Gleichungssystems:

$$Ax = b$$

TODO: Abb22 einfügen

Dafür brauchen wir eine Nummerierung der Gitterpunkte.

Beispiel

Betrachte $n = 5$, also $h = \frac{1}{5}$, und wähle folgende Nummerierung:
TODO: Abb23 einfügen

$$\begin{matrix} (h,4h) & \dots & (4h,4h) \\ \vdots & & \vdots \\ (h,h) & \dots & (4h,h) \end{matrix}$$

Mit dieser Anordnung lässt sich das diskrete Poisson-Problem wie folgt schreiben:

$$A_1 x = b$$

Dabei sind $A_1 \in \mathbb{R}^{m\times m}$, $T \in \mathbb{R}^{(n-1)\times(n-1)}$ und $b \in \mathbb{R}^m$ mit:

$$A_1 = \frac{1}{h^2}\begin{pmatrix} T & -\mathbb{1} & 0 & \dots & 0 \\ -\mathbb{1} & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -\mathbb{1} \\ 0 & \dots & 0 & -\mathbb{1} & T \end{pmatrix} \quad T = \begin{pmatrix} 4 & -1 & 0 & \dots & 0 \\ -1 & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \dots & 0 & -1 & 4 \end{pmatrix} \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_{n-1} \end{pmatrix}$$

$$b_1 = \begin{pmatrix} f(h,h) + \frac{1}{h^2}\left(g(h,0) + g(0,h)\right) \\ f(2h,h) + \frac{1}{h^2} g(2h,0) \\ \vdots \\ f(1-2h,h) + \frac{1}{h^2} g(1-2h,0) \\ f(1-h,h) + \frac{1}{h^2}\left(g(1-h,0) + g(1,h)\right) \end{pmatrix} \quad b_j = \begin{pmatrix} f(h,jh) + \frac{1}{h^2} g(0,jh) \\ f(2h,jh) \\ \vdots \\ f(1-2h,jh) \\ f(1-h,jh) + \frac{1}{h^2} g(1,jh) \end{pmatrix}$$

$$b_{n-1} = \begin{pmatrix} f(h,1-h) + \frac{1}{h^2}\left(g(h,1) + g(0,1-h)\right) \\ f(2h,1-h) + \frac{1}{h^2} g(2h,1) \\ \vdots \\ f(1-2h,1-h) + \frac{1}{h^2} g(1-2h,1) \\ f(1-h,1-h) + \frac{1}{h^2}\left(g(1-h,1) + g(1,1-h)\right) \end{pmatrix}$$

Beispiel:

i) Sei $f = 1$, $g = 0$ und $h = 2^{-6}$, also $n = 64$. Die Matrix A_1 hat die Dimension 3969×3969. Wichtig ist, dass die Dimension der Matrix für $d = 3$ deutlich größer wird.
TODO: Lösung plotten; Label: Lösung der diskreten Poisson-Gleichung
Die obige Darstellung von A_1 zeigt, dass viele Einträge Null sind, denn nur etwa $5 \times (n-1)^2 = 19845$ der $(n-1)^4 = 15\,752\,961$ Einträge von A_1 sind von Null verschieden.

ii) Jetzt betrachte $h = \frac{1}{16}$.
TODO: Folie S. 3; Nichtnullelemente der Matrix A_1
Cholesky Zerlegung (Symmetrische LR-Zerlegung):
TODO: Folie S. 6; Nichtnullelemente von L bei der Cholesky-Zerlegung $A = LL^{\mathrm{T}}$
Setzen wir $\mathrm{nz}(A_1)$ als die Anzahl der Nichtnullelemente von A_1, so folgt:

$$\mathrm{nz}(A_1) \approx 5 \cdot h^{-2} \qquad \mathrm{nz}(L) \approx h^{-3}$$

h	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	$\frac{1}{128}$	
$\mathrm{nz}(A_1)$	1065	4681	19593	80137	$\approx 5h^{-2}$
$\mathrm{nz}(L)$	3389	29821	250109	2048509	$\approx h^{-3}$

8.3 Allgemeine Iterationsverfahren

Schreibe $Ax = b$ in eine Fixpunktgleichung um.

$$\begin{aligned} & Ax = b \\ \Leftrightarrow \quad & B^{-1}(b - Ax) = 0 \\ \Leftrightarrow \quad & \underbrace{\left(\mathbb{1} - B^{-1}A\right)x + B^{-1}b}_{=:g(x)} = x \end{aligned}$$

Definiere

$$G = \mathbb{1} - B^{-1}A \qquad\qquad c := B^{-1}b$$

und die Fixpunktiteration:

$$x^{(k+1)} := g\left(x^{(k)}\right) = Gx^{(k)} + c = B^{-1}\left(\left(B - A\right)x^{(k)} + b\right)$$

Der Banachsche Fixpunktsatz sagt, dass $x^{(k+1)} = g\left(x^{(k)}\right)$ konvergiert, falls g eine Kontraktion ist. Die Norm auf $\mathbb{R}^n$ ist zunächst beliebig, sie ist möglichst geschickt zu wählen.

8.4 Konvergenzsatz

Sei $A \in \mathrm{GL}\left(n, \mathbb{R}\right)$ invertierbar, $b \in \mathbb{R}^n$ und $x = A^{-1}b$. Weiter sei $B \in \mathrm{GL}\left(n, \mathbb{R}\right)$ invertierbar, $x^{(0)} \in \mathbb{R}^n$ der Startvektor und $x^{(k+1)} = \left(\mathbb{1} - B^{-1}A\right)x^{(k)} + B^{-1}b$.

1. Die Fixpunktiteration konvergiert für alle $b \in \mathbb{R}^n$ und alle $x_0 \in \mathbb{R}^n$ genau dann, wenn gilt:

   $$\varrho\left(\mathbb{1} - B^{-1}A\right) < 1$$

   Dabei ist ϱ der Spektralradius, der für $G \in \mathbb{R}^{n\times n}$ wie folgt über die Eigenwerte $\lambda_j\left(G\right)$ von G definiert ist:

   $$\varrho\left(G\right) := \max_j \left|\lambda_j\left(G\right)\right|$$

2. Hinreichend für die Konvergenz ist die Bedingung:

   $$\left|\left|\left|\mathbb{1} - B^{-1}A\right|\right|\right| < 1$$

   Hier ist $|||.|||$ eine beliebige Operatornorm.

Beweis

1. Nach Definition gilt:

   $$\begin{aligned} x^{(k+1)} &= \left(\mathbb{1} - B^{-1}A\right)x^{(k)} + B^{-1}b \\ x &= \left(\mathbb{1} - B^{-1}A\right)x + B^{-1}b \end{aligned}$$

   Für

   $$f^{(k+1)} := x^{(k+1)} - x$$

   gilt:

   $$\begin{aligned} f^{(k+1)} &= \left(\mathbb{1} - B^{-1}A\right)f^{(k)} \\ \Rightarrow \qquad f^{(k)} &= \left(\mathbb{1} - B^{-1}A\right)^k f^{(0)} \end{aligned}$$

   a) „⇒“: Es gelte für alle $f^{(0)} \in \mathbb{C}^n$:

   $$f^{(k)} \to 0$$

Wähle $f^{(0)} \neq 0$ als Eigenvektor zum Eigenwert λ, das heißt:

$$\left(\mathbb{1} - B^{-1}A\right) f^{(0)} = \lambda f^{(0)}$$

Es folgt:

$$f^{(k)} = \lambda^k f^{(0)}$$
$$|\lambda|^k \left\| f^{(0)} \right\| = \left\| f^{(k)} \right\| \xrightarrow{k\to\infty} 0$$

Es folgt $|\lambda| < 1$ und somit:

$$\varrho\left(\mathbb{1} - B^{-1}A\right) < 1$$

$\square_{\text{„}\Rightarrow\text{“}}$

b) „$\Leftarrow$“: Es gelte $\varrho\left(\mathbb{1} - B^{-1}A\right) < 1$.
Es existiert eine Matrix C, sodass

$$J = C^{-1}\left(\mathbb{1} - B^{-1}A\right)C$$

in Jordanscher Normalform ist:

$$J = \begin{pmatrix} \boxed{J_1} & & & 0 \\ & \boxed{J_2} & & \\ & & \ddots & \\ 0 & & & \boxed{J_m} \end{pmatrix} \qquad J_j = \begin{pmatrix} \lambda_j & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_j \end{pmatrix}$$

Dabei sind die $\lambda_i \neq \lambda_j$ für $i \neq j$ die Eigenwerte von $\mathbb{1} - B^{-1}A$ mit $|\lambda_j| < 1$ und J_m ist eine $m_j \times m_j$-Matrix.

$$g^{(k)} := C^{-1} f^{(k)}$$

$$\underbrace{Cg^{(k+1)}}_{=f^{(k+1)}} = \left(\mathbb{1} - B^{-1}A\right) \underbrace{Cg^{(k)}}_{=f^{(k)}}$$

$$g^{(k+1)} = C^{-1}\left(\mathbb{1} - B^{-1}A\right)Cg^{(k)} = Jg^{(k)} = J^{k+1}g^{(0)}$$

$$J^{k+1} = \begin{pmatrix} \boxed{J_1^{k+1}} & & & 0 \\ & \boxed{J_2^{k+1}} & & \\ & & \ddots & \\ 0 & & & \boxed{J_m^{k+1}} \end{pmatrix}$$

$$J_j^k = (\lambda_j \mathbb{1} + J_j - \lambda_j \mathbb{1})^k = \sum_{i=0}^{k} \begin{pmatrix} k \\ i \end{pmatrix} \lambda_j^{k-i} (J_j - \lambda_j \mathbb{1})^i$$

$$J_j - \lambda_j \mathbb{1} = \underbrace{\begin{pmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{pmatrix}}_{(m_j \times m_j)\text{-Matrix}}$$

Mit

$$(J_j - \lambda_j \mathbb{1})^{m_j - 1} = 0$$

folgt:

$$J_j^k = \sum_{i=0}^{m_j - 1} \underbrace{\binom{k}{i} \lambda_j^{k-i}}_{|.| \leq k^{m_j} |\lambda_j|^{k-i} \xrightarrow{k\to\infty} 0} \qquad (J_j - \lambda_j \mathbb{1})^i \xrightarrow{k\to\infty} 0$$

Es folgt $J^k \to 0$ und somit:

$$\begin{aligned} & g^{(k)} = J^k g^{(0)} \xrightarrow{k\to\infty} 0 \\ \Rightarrow \quad & f^{(k)} = C g^{(k)} \xrightarrow{k\to\infty} 0 \\ \Rightarrow \quad & x^{(k)} \xrightarrow{k\to\infty} x \end{aligned}$$

$\square_{\text{„}\Leftarrow\text{“}}$

2. Es sei $\left|\left|\left| \mathbb{1} - B^{-1}A \right|\right|\right| < 1$. Betrachte die Abbildung:

$$g : x \mapsto \left(\mathbb{1} - B^{-1}A\right) x + B^{-1} b$$

Behauptung: g ist kontrahierend auf $\mathbb{R}^n$.
Beweis:

$$\|g(x) - g(y)\| = \left\| \left(\mathbb{1} - B^{-1}A\right)(x - y) \right\| \leq \underbrace{\left|\left|\left| \mathbb{1} - B^{-1}A \right|\right|\right|}_{<1} \cdot \|x - y\|$$

Also ist g kontrahierend. $\square_{\text{Behaputung}}$

Daher konvergiert die Folge $\left(x^{(k)}\right)_{k\in\mathbb{N}}$ gegen einen Fixpunkt von g.

$\square_{8.4}$

Bemerkung

Die Konvergenz des Fehlers $x^{(k)} - x$ gegen Null ist umso schneller, je kleiner $\varrho\left(\mathbb{1} - B^{-1}A\right) < 1$ ist.

Für diagonalisierbare (zum Beispiel symmetrische) Matrizen $\mathbb{1} - B^{-1}A$ gibt $\varrho\left(\mathbb{1} - B^{-1}A\right)$ an, um welchen Faktor sich der Fehler pro Schritt mindestens reduziert.

8.5 Einfache (klassische) Iterationsverfahren

Iteration zum Lösen von $Ax = b$:

$$x^{(k+1)} = \left(\mathbb{1} - B^{-1}A\right) x^{(k)} + B^{-1}b$$

Wähle $B = A$, so gilt:

$$x^{(1)} = B^{-1}b = A^{-1}b = x$$

Dies bringt nichts, da die Berechnung der Inversen vermieden werden soll.
Nächste Strategie: Wähle B so, dass B^{-1} einfach zu berechnen ist.

a) Die einfachste Wahl ist $B = \mathbb{1}$. Dies ergibt das *Richardson-Verfahren*:

$$x^{(k+1)} = x^{(k)} + \left(b - Ax^{(k)}\right)$$

Sei $G := \mathbb{1} - A$. Die Konvergenz für alle Startwerte ist äquivalent zu:

$$\varrho\left(G\right) = \max_{i\in\{1,\dots,n\}} |1 - \lambda_i| < 1$$

TODO: Abb24 einfügen

b) Für bessere Wahl schreiben wir $A = L + D + R$.

$$A = (a_{ij})_{i,j\in\{1,\dots,n\}} \qquad D = \begin{pmatrix} a_{11} & & 0 \\ & \ddots & \\ 0 & & a_{nn} \end{pmatrix}$$

$$L = \begin{pmatrix} 0 & & & 0 \\ a_{21} & \ddots & & \\ \vdots & \ddots & \ddots & \\ a_{n1} & \dots & a_{n,n-1} & 0 \end{pmatrix} \qquad R = \begin{pmatrix} 0 & a_{12} & \dots & a_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & a_{n-1,n} \\ 0 & & & 0 \end{pmatrix}$$

Wähle $B = D$. Dies liefert das *Gesamtschrittverfahren* (Jacobi-Verfahren).
Bemerkung: Falls $a_{ii} = 0$ ist für gewisse i, so permutiere Zeilen und Spalten, sodass die Diagonalelemente nicht Null sind.

$$x^{(k+1)} = x^{(k)} - D^{-1}Ax^{(k)} + D^{-1}b$$

Dies ist äquivalent zu:

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - D^{-1}\left(D + L + R\right)x^{(k)} + D^{-1}b \\ Dx^{(k+1)} &= \left(L + R\right)x^{(k)} + b \end{aligned}$$

Dies ist das Jacobi-Verfahren (J), das in Komponenten geschrieben wie folgt ist:

$$a_{ii}x_i^{(k+1)} = -\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j^{(k)} + b_i$$

Die Iterationsmatrix ist:

$$G = -D^{-1}\left(L + R\right)$$

c) *Einzelschritt-Verfahren* (Gauß-Seidel-Verfahren): In (J) stehen bei der Berechnung von $x_i^{(k+1)}$ die Werte $x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}$ schon zur Verfügung, was genutzt werden soll.
Für $i \in \{1, \ldots, n\}$ erhalte:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(-\sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} + b_i \right)$$

Dies ist äquivalent zu:

$$\begin{aligned} Dx^{(k+1)} &= -Lx^{(k+1)} - Rx^{(k)} + b \\ (D+L)\, x^{(k+1)} &= -Rx^{(k)} + b \\ x^{(k+1)} &= x^{(k)} + (D+L)^{-1} \left(-Ax^{(k)} + b \right) \end{aligned}$$

Mit

$$B = D + L$$

erhalten wir die Iterationsmatrix:

$$G = \mathbb{1} - (D+L)^{-1} A = -(D+L)^{-1} R$$

8.6 **Definition** (Zeilensummenbedingungen)

Eine $(n \times n)$-Matrix A erfüllt die *starke Zeilensummenbedingung*, falls für $i \in \{1, \ldots, n\}$ gilt:

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|$$

A erfüllt die *schwache Zeilensummenbedingung*, falls neben

$$|a_{ii}| \geq \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|$$

für $i \in \{1, \ldots, n\}$ auch für mindestens ein i die strenge Ungleichung „$>$“ gilt.

8.7 **Definition** (zerfallend)

Eine $(n \times n)$-Matrix A heißt *zerfallend*, wenn es eine echte Teilmenge $J \subsetneq \{1, 2, \ldots, n\}$ mit $J \neq \emptyset$ gibt, sodass $a_{ij} = 0$ für $i \in J$ und $j \notin J$ gilt.
Nach Umnummerierung hat eine zerfallende Matrix die Struktur:

$$\begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}$$

Hier sind A_{11}, A_{21} und A_{22} Blockmatrizen.

8.8 Satz (Zeilensummenkriterium)

i) Die $(n \times n)$-Matrix A erfülle die schwache Zeilensummenbedingung und sei *nicht* zerfallend. Dann konvergieren sowohl Gesamtschritt- als auch Einzelschrittverfahren.

ii) Wenn die starke Zeilensummenbedingung erfüllt ist, kann auf die Voraussetzung nicht zerfallend verzichtet werden. Die Spektralradien der Iterationsmatrizen sind kleiner als:

$$\max_{i \in \{1,\ldots,n\}} \frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}| < 1$$

Beweis

Es sei $x \neq 0$. Das Gesamtschrittverfahren hat die Iterationsmatrix:

$$G = -D^{-1}(L + R)$$

Für die Maximumsnorm erhalten wir die Zeilensummennorm

$$|||A|||_\infty = \max_i \sum_{k=1}^{n} |a_{ik}|$$

als Operatornorm. Wegen der starken Zeilensummenbedingung gilt:

$$\left|\left|\left|-D^{-1}(L + R)\right|\right|\right| = \max_i \frac{1}{|a_{ii}|} \sum_{\substack{j=1 \\ j \neq i}} |a_{ij}| < 1$$

Satz 8.4 liefert die Konvergenz.
Die schwache Zeilensummenbedingung liefert:

$$\varrho(G) \leq |||G|||_\infty \leq 1$$

Falls $Gx = \lambda x$ ist, folgt:

$$|\lambda| \, \|x\| = \|Gx\| \leq |||G||| \cdot \|x\|$$

Zu zeigen ist, dass es keinen Eigenwert λ von G mit $|\lambda| = 1$ gibt.
Nehme also an, dass ein Eigenwert λ von G mit $|\lambda| = 1$ existiert. Ein zugehörige normiertem Eigenvektor mit $\|x\|_\infty = 1$ sei $x \in \mathbb{R}^n$.

$$Gx = \lambda x$$

Es gilt:

$$|(Gx)_i| = |\lambda x_i| = |x_i|$$

Definiere:

$$N_1 := \left\{ l \in \{1, \ldots, n\} \,\middle|\, |x_l| = 1 \right\}$$
$$N_2 := \left\{ k \in \{1, \ldots, n\} \,\middle|\, |x_k| < 1 \right\}$$

Also gilt:

$$|x_i| = |(Gx)_i| \leq \frac{1}{|a_{ii}|} \cdot \left(\sum_{j \neq i} |a_{ij}| \cdot |x_j| \right) =$$

$$= \frac{1}{|a_{ii}|} \cdot \left(\sum_{\substack{j \in N_1 \\ j \neq i}} |a_{ij}| \cdot |x_j| + \sum_{\substack{j \in N_2 \\ j \neq i}} |a_{ij}| \cdot \underbrace{|x_j|}_{<1} \right)$$

Falls für ein i ein $j \in N_2$ mit $i \neq j$ und $a_{ij} \neq 0$ existiert, so folgt:

$$|x_i| \leq \frac{1}{|a_{ii}|} \left(\sum_{j \neq i} |a_{ij}| \cdot |x_j| \right) < \overline{a} \leq 1$$

$$\overline{a} = \max_i \frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}|$$

Somit muss für $i \in N_1$ nun $a_{ij} = 0$ für alle $j \in N_2$ gelten, da wir sonst einen Widerspruch zu $|x_i| = 1$ erhalten.

Dies ist ein Widerspruch zur Tatsache, dass A nicht zerfällt.

Für den Rest des Beweises siehe: HACKBUSCH: *Iterationsverfahren großer schwach besetzter Gleichungssysteme*; Teubner. $\square_{8.8}$

8.9 Kontraktionskriterium für spd-Matrizen

spd steht für symmetrisch und positiv definit.

Ziel: Formuliere eine Bedingung für die Tatsache $\varrho(G) < 1$, wobei $G = \mathbb{1} - B^{-1}A$ ist.

Für eine symmetrische und positiv definite Matrix A definiere das Skalarprodukt:

$$\langle x, y \rangle_A := \langle x, Ay \rangle = x \cdot Ay = \sum_{i,j} x_i a_{ij} y_j$$

Dabei ist $\langle ., . \rangle$ das euklidische Skalarprodukt. Für eine Matrix $F \in \mathbb{R}^{n \times n}$ ist:

$$F^* := A^{-1} F^{\mathrm{T}} A$$

Es gilt:

$$\langle Fx, y \rangle_A = \langle Fx, Ay \rangle = \left\langle x, F^{\mathrm{T}} Ay \right\rangle = \left\langle x, AA^{-1} F^{\mathrm{T}} Ay \right\rangle = \langle x, F^* y \rangle_A$$

Eine (bezüglich $\langle ., . \rangle_A$) selbstadjungierte Matrix $B = B^*$ heißt positiv bezüglich $\langle ., . \rangle_A$, falls $\langle Bx, x \rangle_A > 0$ für alle $x \neq 0$ gilt.

8.10 Lemma

Sei $G \in \mathbb{R}^{n \times n}$ und G^* sei die bezüglich eines Skalarprodukts $\langle ., . \rangle$ adjungierte Matrix zu G. Ist dann $F := \mathbb{1} - G^* G$ eine bezüglich $\langle ., . \rangle$ positive Matrix, so folgt $\varrho(G) < 1$.

Beweis

Da F positiv ist, gilt für alle $x \in \mathbb{R}^n \setminus \{0\}$:

$$\langle Fx, x\rangle = \langle x, x\rangle - \langle G^*Gx, x\rangle = \langle x, x\rangle - \langle Gx, Gx\rangle > 0$$

Somit folgt $\|x\| > \|Gx\|$ mit der vom Skalarprodukt induzierten Norm $\|.\| := \sqrt{\langle ., .\rangle}$. Da

$$S^{n-1} = \left\{x \in \mathbb{R}^n \middle| \, \|x\| = 1\right\}$$

kompakt ist, folgt:

$$\varrho\,(G) \leq |||G||| = \sup_{\|x\|=1} \|Gx\| < 1$$

$\square_{8.10}$

8.11 Satz

Das Gauß-Seidel-Verfahren konvergiert für jede symmetrische, positiv definite Matrix A.

Beweis

Für

$$G := \mathbb{1} - (D + L)^{-1} A$$

ist zu zeigen, dass

$$F = \mathbb{1} - G^*G$$

positiv definit bezüglich des Skalarprodukts $\langle x, y\rangle_A := \langle x, Ay\rangle = x \cdot Ay$.

$$\langle Gx, Ay\rangle = \langle Gx, y\rangle_A = \langle x, G^*y\rangle_A = \langle x, AG^*y\rangle$$

Wegen $R^* = L$ folgt:

$$G^* = \mathbb{1} - \underbrace{A^{-1}A^*}_{=\mathbb{1}} (D^* + L^*)^{-1} A = \mathbb{1} - (D + R)^{-1} A$$

Somit erhält man:

$$\begin{aligned}
F = \mathbb{1} - G^*G &= \mathbb{1} - \left(\mathbb{1} - (D + R)^{-1} A\right)\left(\mathbb{1} - (D + L)^{-1} A\right) = \\
&= \mathbb{1} - \mathbb{1} + (D + L)^{-1} A + (D + R)^{-1} A - (D + R)^{-1} \underbrace{A}_{=D+L+R} (D + L)^{-1} A = \\
&= \left((D + L)^{-1} + (D + R)^{-1} - (D + R)^{-1} (D + L + R) (D + L)^{-1}\right) A = \\
&= \left((D + L)^{-1} + (D + R)^{-1} - \left(\mathbb{1} + (D + R)^{-1} L\right) (D + L)^{-1}\right) A = \\
&= \left((D + R)^{-1} - (D + R)^{-1} L (D + L)^{-1}\right) A = \\
&= (D + R)^{-1} \left(\mathbb{1} - L (D + L)^{-1}\right) A =
\end{aligned}$$

$$= (D+R)^{-1}(D+L-L)(D+L)^{-1}A = (D+R)^{-1}D(D+L)^{-1}A$$

Für $x \neq 0$ gilt:

$$\langle Fx, x\rangle_A = \left\langle (D+R)^{-1}D(D+L)^{-1}Ax, Ax\right\rangle = \left\langle D(D+L)^{-1}Ax, (D+L)^{-1}Ax\right\rangle$$

Da A positiv definit ist, gilt $a_{ii} > 0$ für $i \in \{1,\dots,n\}$ und somit ist

$$D^{\frac{1}{2}} = \begin{pmatrix} \sqrt{a_{11}} & & \\ & \ddots & \\ & & \sqrt{a_{nn}} \end{pmatrix}$$

reell und selbstadjungiert. Es folgt:

$$\langle Fx, x\rangle_A = \left\langle D^{\frac{1}{2}}(D+L)^{-1}Ax, D^{\frac{1}{2}}(D+L)^{-1}Ax\right\rangle \overset{x\neq 0}{>} 0$$

Daher ist F positiv und somit $\varrho(G) < 1$. Also konvergiert das Gauß-Seidel-Verfahren. $\square_{8.11}$

8.12 Beispiel (Poisson-Gleichung)

Betrachte das Gleichungssystem aus 8.2: Die Matrix $A_1 \in \mathbb{R}^{m\times m}$ mit $m = (n-1)^2$ hat maximal fünf Elemente pro Zeile.
Der Aufwand pro Iteration beim Gesamtschritt- oder Einzelschritt-Verfahren ist ca. $10n^2$ flops. Der benötigte Speicherplatz ist ebenfalls von der Ordnung n^2.

Konvergenz: Die schwache Zeilensummenbedingung ist erfüllt. Außerdem ist A_1 nicht zerfallend. Dies liefert die Konvergenz des Verfahrens (siehe Satz 8.4).
Da A_1 positiv definit und symmetrisch ist, konvergiert das Verfahren nach Satz 8.11.

Die Konvergenzgeschwindigkeit des Gesamtschritt-Verfahrens mit Iterationsmatrix G_G und des Einzelschritt-Verfahrens mit G_E sind wie folgt:

- Einzelschritt-Verfahren: Wegen

$$\varrho(G_G) = \cos\left(\frac{\pi}{n}\right) = 1 - \frac{1}{2}\left(\frac{\pi}{n}\right)^2 + \mathcal{O}_\infty\left(\frac{1}{n^4}\right)$$

folgt:

$$\varrho(G_E) = (\varrho(G_G))^2 = \left(\cos\left(\frac{\pi}{n}\right)\right)^2 = 1 - \left(\frac{\pi}{n}\right)^2 + \mathcal{O}_\infty\left(\frac{1}{n^4}\right)$$

Dies wird von Hackbusch genauer gezeigt.
Wir hatten uns überlegt, dass $\varrho(G)$ angibt, um welchen Faktor sich der Fehler verringert. Für G_E und G_G gilt aber $\varrho(G_E), \varrho(G_G) \xrightarrow{n\to\infty} 1$. Also ist die Konvergenzgeschwindigkeit bei kleinem $h = \frac{1}{n}$ sehr langsam.

8.13 SOR-Verfahren

Ziel: Beschleunige das Einzelschritt-Verfahren!
Beim Einzelschritt-Verfahren ist die Iterationsvorschrift:

$$x^{(k+1)} = x^{(k)} \underbrace{-D^{-1}\left(Lx^{(k+1)} + (D+R)x^{(k)} - b\right)}_{\text{diese Größe wird zu } x^{(k)} \text{addiert}}$$

Oft ist es günstiger, die Aufaddierung zu verstärken oder abzuschwächen.

Definition

Das SOR-Verfahren zur Lösung der Gleichung $Ax = b$ lautet für einen Relaxationsparameter $\omega > 0$ wie folgt:

$$x^{(k+1)} = x^{(k)} - \omega D^{-1} \left(L x^{(k+1)} + (D + R)\, x^{(k)} - b \right)$$

In Komponenten ist dies:

$$x_i^{(k+1)} = x_i^{(k)} - \omega \frac{1}{a_{ii}} \left(\sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} + \sum_{j=i}^{n} a_{ij} x_j^{(k)} - b_i \right)$$

Die englische Bezeichnung „sucessive overrelaxation“ ist für die Abkürzung SOR verantwortlich. Für $\omega = 1$ erhält man wieder das Einzelschrittverfahren.

8.14 Lemma

Für das SOR-Verfahren ist B aus Abschnitt 8.3 als $\left(\frac{1}{\omega} D + L\right)$ zu wählen, das heißt:

$$\begin{aligned} x^{(k+1)} &= \left(\mathbb{1} - B^{-1} A \right) x^{(k)} + B^{-1} b = \\ &= \left(\mathbb{1} - \left(\frac{1}{\omega} D + L \right)^{-1} A \right) x^{(k)} + \left(\frac{1}{\omega} D + L \right)^{-1} b \end{aligned}$$

Beweis

Aus der Definition des SOR-Verfahrens folgt:

$$\begin{aligned} (D + \omega L)\, x^{(k+1)} &= (D + \omega L)\, x^{(k)} - \omega \left(\mathbb{1} + \omega L D^{-1} \right) \left(L x^{(k+1)} + (D + R)\, x^{(k)} - b \right) = \\ &= D x^{(k)} - \omega \Bigg(L \left(x^{(k+1)} - x^{(k)} \right) + (D + R)\, x^{(k)} - b + \\ &\qquad + L \underbrace{\omega D^{-1} \left(L x^{(k+1)} + (D + R)\, x^{(k)} - b \right)}_{= -\left(x^{(k+1)} - x^{(k)} \right)} \Bigg) = \\ &= D x^{(k)} - \omega\, (D + R)\, x^{(k)} + \omega b \end{aligned}$$

Damit folgt:

$$\left(\frac{1}{\omega} D + L \right) x^{(k+1)} = \left(\frac{1}{\omega} D + L \right) x^{(k)} - A x^{(k)} + b$$

Das liefert:

$$x^{(k+1)} = x^{(k)} - \left(\frac{1}{\omega} D + L \right)^{-1} A x^{(k)} + \left(\frac{1}{\omega} D + L \right)^{-1} b$$

$\square_{8.14}$

Es soll nun ω so bestimmt werden, dass die Konvergenzgeschwindigkeit zunimmt. (Den besten Wert zu bestimmen ist allerdings schwierig.)

8.15 Satz

Die Diagonalelemente von $A \in \mathbb{R}^{n\times n}$ seien von Null verschieden. Dann gilt für die Iterationsmatrix des SOR-Verfahrens:

$$\varrho\left(\mathbb{1} - \left(\frac{1}{\omega}D + L\right)^{-1} A\right) \geq |\omega - 1|$$

Beweis

Es gilt:

$$\begin{aligned} C_E(\omega) &:= \mathbb{1} - \left(\frac{1}{\omega}D + L\right)^{-1} A = \\ &= \left(\frac{1}{\omega}D + L\right)^{-1} \left(\frac{1}{\omega}D + L - L - D - R\right) = \\ &= (D + \omega L)^{-1} \left((\omega - 1) D - \omega R\right) = \\ &= \underbrace{\left(\mathbb{1} + \omega D^{-1} L^{-1}\right)^{-1}}_{\substack{\text{untere Dreiecksmatrix} \\ \text{mit 1 auf der Diagonalen}}} \underbrace{\left((\omega - 1)\,\mathbb{1} - \omega D^{-1} R\right)}_{\substack{\text{obere Dreiecksmatrix} \\ \text{mit } (1-\omega) \text{ auf der Diagonalen}}} \end{aligned}$$

Es folgt somit:

$$|\det(C_E(\omega))| = |1 - \omega|^n$$

Wegen $|\det(C_E(\omega))| \leq (\varrho(C_E(\omega)))^n$ folgt $|1 - \omega| \leq \varrho(C_E(\omega))$. □8.15

Dieser Satz zeigt, dass nur Werte $\omega \in (0, 2)$ konvergente Verfahren liefern.

8.16 Satz

Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Dann konvergiert das SOR-Verfahren für alle $\omega \in (0, 2)$.

Beweis

$$C_E(\omega) = \mathbb{1} - B^{-1}A = \mathbb{1} - \left(\frac{1}{\omega}D + L\right)^{-1} A$$

Zu zeigen ist $\varrho(C_E(\omega)) < 1$: Sei $\lambda \in \mathbb{C}$ ein Eigenwert von $C_E(\omega)$ mit zugehörigem Eigenvektor $x \in \mathbb{C}^n$, sodass $\|x\|_2 = 1$ ist. Dann gilt:

$$C_E(\omega)\, x = \left(\mathbb{1} - \left(\frac{1}{\omega}D + L\right)^{-1} A\right) x = \lambda x$$

Das heißt es gilt:

$$Ax = (1 - \lambda)\, Bx$$

Dabei ist $\lambda \neq 1$, da A positiv definit ist. Mit dem komplex konjugierten Vektor

$$\overline{x} = \begin{pmatrix} \overline{x}_1 \\ \vdots \\ \overline{x}_n \end{pmatrix}$$

erhalten wir somit:

$$\frac{1}{1-\lambda} = \frac{\overline{x}^{\mathrm{T}} B x}{\overline{x}^{\mathrm{T}} A x}$$

Da A symmetrisch ist, gilt wegen $B^{\mathrm{T}} = \frac{1}{\omega} D + L^{\mathrm{T}} = \frac{1}{\omega} D + R$:

$$B + B^{\mathrm{T}} = \left(\frac{2}{\omega} - 1 \right) D + A$$

Für den Realteil von $(1-\lambda)^{-1}$ gilt:

$$\begin{aligned} \mathrm{Re}\left(\frac{1}{1-\lambda}\right) = \mathrm{Re}\left(\frac{\overline{x}^{\mathrm{T}} B x}{\overline{x}^{\mathrm{T}} A x}\right) &= \frac{1}{2}\left(\frac{\overline{x}^{\mathrm{T}} B x}{\overline{x}^{\mathrm{T}} A x} + \overline{\frac{\overline{x}^{\mathrm{T}} B x}{\overline{x}^{\mathrm{T}} A x}}\right) \overset{B \text{ reell}}{=} \frac{1}{2}\left(\frac{\overline{x}^{\mathrm{T}} B x}{\overline{x}^{\mathrm{T}} A x} + \frac{x^{\mathrm{T}} B \overline{x}}{x^{\mathrm{T}} A \overline{x}}\right) = \\ &= \frac{1}{2}\left(\frac{\overline{x}^{\mathrm{T}} B x}{\overline{x}^{\mathrm{T}} A x} + \frac{\overline{x}^{\mathrm{T}} B^{\mathrm{T}} x}{\overline{x}^{\mathrm{T}} A^{\mathrm{T}} x}\right) \overset{A \text{ symmetrisch}}{=} \frac{1}{2} \cdot \frac{\overline{x}^{\mathrm{T}} \left(B + B^{\mathrm{T}}\right) x}{\overline{x}^{\mathrm{T}} A x} = \\ &= \frac{1}{2}\Bigg(\underbrace{\left(\frac{2}{\omega} - 1\right)}_{>0} \cdot \underbrace{\frac{\overline{x}^{\mathrm{T}} D x}{\overline{x}^{\mathrm{T}} A x}}_{>0 \text{da } A \text{ positiv definit}} + 1 \Bigg) > \frac{1}{2} \end{aligned}$$

Mit $\lambda = u + iv$ für $u, v \in \mathbb{R}$ folgt:

$$\begin{aligned} \frac{1}{2} < \mathrm{Re}\left(\frac{1}{1-\lambda}\right) &= \frac{1-u}{(1-u)^2 + v^2} \\ (1-u)^2 + v^2 &< 2 - 2u \\ 1 - 2u + u^2 + v^2 &< 2 - 2u \\ u^2 + v^2 &< 1 \end{aligned}$$

$\square_{8.16}$

8.17 Fehlerreduktion bei iterativen Verfahren

Der Konvergenzsatz 8.4 liefert, dass $\varrho\left(\mathbb{1} - B^{-1}A\right)$ entscheidend für die Konvergenz ist. Diese Größe ist auch entscheidend für die Konvergenzgeschwindigkeit.

Zur Vereinfachung sei $\left(\mathbb{1} - B^{-1}A\right)$ diagonalisierbar und habe die Eigenwerten $\lambda_1, \ldots, \lambda_n$ mit $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n| > 0$ zu den Eigenvektoren $v_1, \ldots, v_n$, die eine Basis bilden. Betrachte $b, x \in \mathbb{R}^n$ mit $Ax = b$. Wähle einen Startwert x^0 und definiere:

$$x^{(0)} - x =: e^{(0)} =: \sum_{i=0}^{n} c_i v_i$$

Es gilt nun (vergleiche Satz 8.4):

$$e^{(k)} = x^{(k)} - x = \left(\mathbb{1} - B^{-1}A\right)^k e^{(0)} =$$

$$= \left(\mathbb{1} - B^{-1}A\right)^k \left(\sum_{i=1}^{n} c_i v_i\right) = \sum_{i=1}^{n} c_i \lambda_i^k v_i =$$
$$= \lambda_1^k \Bigg(c_1 v_1 + \underbrace{\sum_{i=2}^{n} c_i \left(\frac{\lambda_i}{\lambda_1}\right)^k v_i}_{=:r^{(k)}} \Bigg) = \lambda_1^k \left(c_1 v_1 + r^{(k)}\right)$$

Ist nun $c_1 \neq 0$, so existieren $c_{\min}, c_{\max} \in \mathbb{R}_{>0}$, sodass wegen $\left|\frac{\lambda_i}{\lambda_1}\right| \leq 1$ gilt:

$$0 < c_{\min} \leq \left\| c_1 v_1 + r^{(k)} \right\| \leq c_{\max}$$

Insgesamt folgt:

$$\sigma_k := \left(\frac{\left\|e^{(k)}\right\|}{\left\|e^{(0)}\right\|} \right)^{\frac{1}{k}} = \frac{|\lambda_1| \cdot \left\| c_1 v_1 + r^{(k)} \right\|^{\frac{1}{k}}}{\left\|e^{(0)}\right\|^{\frac{1}{k}}} \xrightarrow{k\to\infty} |\lambda_1| = \varrho\left(\mathbb{1} - B^{-1}A\right)$$

Daher ist $\left\|e^{(k)}\right\|$ ungefähr gleich $|\lambda_1|^k \cdot \left\|e^{(0)}\right\|$ für große k.
Um den Anfangsfehler $\left\|e^{(0)}\right\|$ um den Faktor $R > 1$ zu erniedrigen, muss (im Grenzfall) gelten:

$$\begin{aligned} & |\lambda_1|^k < \frac{1}{R} \\ \Leftrightarrow \quad & k \ln\left(|\lambda_1|\right) < -\ln(R) \\ \overset{0<|\lambda_1|<1}{\Leftrightarrow} \quad & k > \frac{\ln(R)}{-\ln\left(|\lambda_1|\right)} \end{aligned}$$

Für $R = e$ erhalte die Bedingung:

$$k > \left(-\ln\left(|\lambda_1|\right)\right)^{-1}$$

Wir nennen daher $-\ln\left(\varrho\left(\mathbb{1} - B^{-1}A\right)\right)$ die *asymptotische Konvergenzrate.*

8.18 Fehlerreduktion für das diskrete Poisson-Problem

Das diskrete Poisson-Problem wurde bereits in 8.2 beschrieben. Für $n \in \mathbb{N}_{\geq 1}$ setze:

$$h := \frac{1}{n}$$

Für das Gesamtschrittverfahren mit

$$\varrho\left(\mathbb{1} - B^{-1}A_1\right) \approx 1 - \frac{1}{2}\pi^2 h^2$$

ist die asymptotische Konvergenzrate:

$$-\ln\left(\varrho\left(\mathbb{1} - D^{-1}A_1\right)\right) \approx -\ln\left(1 - \frac{(\pi h)^2}{2}\right) \overset{h\ll 1}{\approx} \frac{1}{2}(\pi h)^2$$

Um den Startfehler um den Faktor R zu vermindern, sind (asymptotisch)

$$K(R) := \frac{-\ln(R)}{\ln\left(\varrho\left(\mathbb{1} - D^{-1}A_1\right)\right)} = \frac{-\ln(R)}{\ln(\cos \pi h)} \approx \frac{2}{\pi^2 h^2} \ln R$$

Iterationsschritte nötig.

Sei $L(R)$ die tatsächliche Anzahl von Iterationsschritten, die zur Reduktion des Startfehlers um einen Faktor R benötigt werden und $K(R)$ die theoretische Schätzung von L. Für $R = 10^3$ ergibt sich:

h	$\frac{1}{40}$	$\frac{1}{80}$	$\frac{1}{160}$	$\frac{1}{320}$
$L\left(10^3\right)$	2092	8345	33332	133227
$K\left(10^3\right)$	2237	8965	35833	143338

Bemerkung

- K ist eine gute Schätzung für L.
- Halbieren von h vervierfacht den Aufwand.

Für das Einzelschrittverfahren ist der Fehler etwa halb so groß:

$$\frac{-\ln(R)}{\ln\left(\varrho\left(\mathbb{1} - (D+L)^{-1} A_1\right)\right)} \approx \frac{\ln R}{\pi^2 h^2}$$

h	$\frac{1}{40}$	$\frac{1}{80}$	$\frac{1}{160}$	$\frac{1}{320}$
$L\left(10^3\right)$	1056	4193	16706	66694
$K\left(10^3\right)$	1119	4478	17916	71669

Es gilt $m = (n-1)^2 \approx n^2$ und mit $h = n^{-1}$ folgt $m \approx h^{-2}$.

Ein Schritt im Gesamtschritt- (oder Einzelschrittverfahren) kostet stets konst. $\cdot\, m$ flops. Der Aufwand um den Fehler um den Faktor R zu vermindern ist proportional zu m^2.

8.19 Optimaler Relaxationsparameter für das diskrete Poissonproblem

Stelle die Anzahl der nötigen Iteration dar um den Fehler um den Faktor 10^3 zu vermindern.

TODO: Folie S. 26; Anzahl der Iterationen in Abhängigkeit von ω

Bemerkung

- Das SOR-Verfahren mit optimalen ω ist deutlich besser als das Einzelschrittverfahren, für das $\omega = 1$ gilt.
- Der optimale Wert von ω ist stark abhängig vom vorliegenden Problem.

Für das Poisson-Problem ist es möglich, ein optimales ω zu bestimmen.

8.20 Satz

Wir betrachten die diskretisierte Poisson-Gleichung

$$A_1 x = b$$

wie in 8.2. Es sei $\mu := \varrho\left(\mathbb{1} - D^{-1}A_1\right) < 1$ der Spektralradius des Jacobi-Verfahren, also des Gauß-Seidel-Verfahrens, und sei M_ω die Iterationsmatrix des SOR-Verfahrens. Dann ist $\varrho\left(M_\omega\right)$ für folgenden Relaxationsparameter minimal:

$$\omega_{\text{opt}} = \frac{2}{1+\sqrt{1-\mu^2}} = 1 + \left(\frac{\mu}{1+\sqrt{1-\mu^2}}\right)^2 = 2\left(1+\sin\left(\frac{\pi}{n}\right)\right)$$

Weiter gilt mit $h = \frac{1}{n}$:

$$\varrho\left(M_\omega\right) = \omega_{\text{opt}} - 1 = \frac{\left(\cos\left(\pi h\right)\right)^2}{\left(1+\sin\left(\pi h\right)\right)^2} = \frac{1-\sin\left(\pi h\right)}{1+\sin\left(\pi h\right)} \approx 1 - 2\pi h$$

Um einen Startfehler um einen Faktor R zu reduzieren, sind K Iterationen nötig und es gilt:

$$K = -\frac{\ln\left(R\right)}{\ln\left(\varrho\left(M_{\omega_{\text{opt}}}\right)\right)} \approx \frac{1}{2\pi h}\ln\left(R\right) \approx \frac{\ln\left(R\right)}{2\pi}\sqrt{m}$$

Der Aufwand pro Iteration ist also proportional zu m. Der Gesamtaufwand zur Reduktion um den Faktor ist also proportional zu $m^{\frac{3}{2}}$.

Bemerkung

- Noch deutlich besser ist das vorkonditionierte cg-Verfahren („conjugate gradient“), auch pcg-Verfahren genannt.
- Am schnellsten sind Mehrgitterverfahren, deren Aufwand proportional zu m ist. (Jacobi-Verfahren und Gauß-Seidel-Verfahren gehen hier ein.)

8.21 SSOR-Verfahren

Die symmetrische SOR-Methode (SSOR) ist eine Variante des SOR-Verfahrens. Die Idee ist, dass ein erster Teilschritt die Neuberechnung der Vektorkomponenten mit kleinen Indizes beginnt und ein zweiter Teilschritt mit großen Indizes. Konkret heißt das, man berechnet

$$x_i^{\left(k+\frac{1}{2}\right)} = x_i^{(k)} - \frac{\omega}{a_{ii}}\left(\sum_{j=1}^{i-1} a_{ij}x^{\left(k+\frac{1}{2}\right)} + \sum_{j=i}^{n} a_{ij}x_j^{(k)} - b_i\right)$$

für aufsteigendes $i \in \{1,\ldots,n\}$ und dann

$$x_i^{(k+1)} = x_i^{\left(k+\frac{1}{2}\right)} - \frac{\omega}{a_{ii}}\left(\sum_{j=i+1}^{n} a_{ij}x^{(k+1)} + \sum_{j=1}^{i} a_{ij}x_j^{\left(k+\frac{1}{2}\right)} - b_i\right)$$

für absteigendes $i \in \{n,\ldots,1\}$. Die Iterationsmatrix ist $\mathbb{1} - B^{-1}A$ mit:

$$B^{-1} = \omega\left(2-\omega\right)\left(D+\omega R\right)^{-1} D \left(D+\omega L\right)^{-1}$$

9 Eigenwertaufgaben

9.1 Einleitung

Bestimmung von Eigenwerten ist in der Praxis von großer Bedeutung:

- Analyse des Schwingungsverhaltens von mechanischen oder elektrischen Systemen (Eigenfrequenzen, Resonanzen, die z.B. einen Brückeneinsturz auslösen können)
- Die Reihenfolge der Auflistung von Suchtreffern bei Google wird mit Hilfe von Eigenwertaufgaben bestimmt, deren Lösung nummerisch berechnet wird.
  (siehe Deufelhard/Hohmann)

Zunächst wollen wir die Eigenwerte theoretisch lokalisieren.

9.2 Satz von Gerschgorin

Die Eigenwerte einer Matrix $A = (a_{ij})_{ij} \in \mathbb{C}^{n\times n}$ liegen in der Vereinigung aller Gerschgorinkreise

$$K_i := \left\{ z \in \mathbb{C} \middle| \, |z - a_{ii}| \leq r_i \right\}$$

mit den Radien:

$$r_i := \sum_{\substack{k=1 \\ k\neq i}}^{n} |a_{jk}|$$

Beweis

Sei $v \neq 0$ mit $Av = \lambda v$, das heißt für $l \in \{1, \ldots, n\}$:

$$(\lambda - a_{ll})\, v_l = \sum_{\substack{k=1 \\ k\neq l}}^{n} a_{lk} v_k$$

Sei nun l so, dass gilt:

$$|v_l| = \|v\|_\infty = \max_{i\in\{1,\ldots,n\}} |v_i| > 0$$

Dann folgt wegen $\left|\frac{v_k}{v_l}\right| \leq 1$:

$$|\lambda - a_{ll}| \leq \sum_{\substack{k=1 \\ k\neq l}}^{n} |a_{lk}| = r_l$$

$\square_{9.2}$

Die obige Aussage lässt sich verschärfen, indem man A^{T} oder $D^{-1}AD$ etc. betrachtet. Je weniger sich A von einer Diagonalmatrix unterscheidet, desto schärfer ist die Abschätzung.

Beispiel

$$A = \begin{pmatrix} 1+\frac{\mathrm{i}}{2} & 0{,}5 & 0{,}1 \\ 0{,}3 & 1-\frac{\mathrm{i}}{1} & 0{,}5 \\ 0{,}4 & 0 & -0{,}5 \end{pmatrix}$$

Die Gerschgorin-Radien sind:

$$r_1 = 0{,}6 \qquad r_2 = 0{,}8 \qquad r_3 = 0{,}4$$

TODO: Rest von Folie

9.3 Satz

Bilden k Gerschgorinkreise eine Menge G, die zu den restlichen Kreisen K_j disjunkt ist, dann liegen in G genau k Eigenwerte der Matrix.

Beweis:

Verwende eine Homotopiemethode. Betrachte:

$$A(t) := \underbrace{\begin{pmatrix} a_{11} & & 0 \\ & \ddots & \\ 0 & & a_{nn} \end{pmatrix}}_{=:D} + t(A - D)$$

Dann gilt $A(0) = D$ und $A(1) = A$. Seien $\lambda_i(t)$ die Eigenwerte von $A(t)$ für $i \in \{1, \ldots, n\}$. Für $t = 0$ gilt $\lambda_i(0) = a_{ii}$ und es liegen genau k Eigenwerte in G und $(n-k)$ in den restlichen K_i. $\lambda_i(t)$ ist stetig in t, da die Nullstellen des charakteristischen Polynoms stetig von den Koeffizienten abhängen. Weiter ist $K_i(t) \subseteq K_i$ für $t \in [0,1]$. Aus der Stetigkeit folgt dann die Behauptung, da ein Eigenwert nicht aus G heraus springen kann. $\square_{9.3}$

Beispiel

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

Es folgt:

$$K_1 = \{\mu \mid |\mu - 1| \leq 0{,}2\}$$
$$K_2 = \{\mu \mid |\mu - 2| \leq 0{,}4\}$$
$$K_3 = \{\mu \mid |\mu - 3| \leq 0{,}2\}$$

TODO: Plot der Gerschgorin-Kreise

Je ein Eigenwert liegt in jedem Kreis, also müssen alle Eigenwerte reell sein, da sonst immer ein paar von komplex-konjugierten Eigenwerten existieren würde.

9.4 Potenzmethode (Vektoriteration)

Sei $A \in \mathbb{R}^{n\times n}$ eine in $\mathbb{C}$ diagonalisierbare Matrix und sei der dominante Eigenwert einfach, das heißt:

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$$

Seien $v_1, \ldots, v_n$ die zugehörigen Eigenvektoren, ohne Einschränkung mit $\|v_i\|_2 = 1$ für alle $i \in \{1, \ldots, n\}$. Die Vektoriteration ist gegeben durch $x^{(k+1)} = Ax^{(k)}$. Da A reell ist, folgt $\lambda_1 \in \mathbb{R}$, da sonst $\overline{\lambda_1}$ ebenfalls ein Eigenwert wäre, aber $|\overline{\lambda_1}| = |\lambda_1|$ gilt, im Widerspruch dazu, dass λ_1 der betragsmäßig größte Eigenwert ist. Dann kann auch $v_1 \in \mathbb{R}^n$ gewählt werden, da die Eigenwertgleichung getrennt für den Real- und den Imaginärteil von v gilt, weil A reell ist. Schreibe nun den Startwert $x^{(0)}$ wie folgt:

$$x^{(0)} = \sum_{i=1}^{n} \alpha_i v_i$$

Wir nehmen an, dass $x^{(0)}$ nicht senkrecht zu v_1 ist, also $\alpha_1 \neq 0$ gilt.
Für die Iterierten $x^{(k)} \in \mathbb{R}^n$ gilt:

$$\begin{aligned} x^{(k)} = A^k x^{(0)} &= \sum_{i=1}^{n} \alpha_i \lambda_i^k v_i = \\ &= \lambda_1^k \Bigg(\alpha_1 v_1 + \underbrace{\sum_{i=2}^{n} \alpha_i \left(\frac{\lambda_i}{\lambda_1}\right)^k v_i}_{=:r^{(k)}} \Bigg) = \lambda_1^k \left(\alpha_1 v_1 + r^{(k)} \right) \end{aligned}$$

Außerdem gilt:

$$\left\| r^{(k)} \right\|_2 \leq \sum_{i=2}^{n} |\alpha_i| \cdot \left| \frac{\lambda_i}{\lambda_1} \right|^k \cdot \underbrace{\|v_i\|_2}_{=1} \leq c\left(\alpha_2, \ldots, \alpha_n\right) \cdot \left| \frac{\lambda_2}{\lambda_1} \right|^k \xrightarrow{k\to\infty} 0$$

Für große k zeigt $x^{(k)}$ also in Richtung von v_1.
Wir wollen die Aussage, dass die durch $x^{(k)} = A^k x^{(0)}$ definierte Richtung gegen die durch v_1 definierte Richtung konvergiert, präzisieren.
Zwischen zwei eindimensionalen Unterräumen

$$\begin{aligned} T_v &= \left\{ \alpha v \middle| \alpha \in \mathbb{R} \right\} & \|v\|_2 = 1 \\ T_w &= \left\{ \alpha w \middle| \alpha \in \mathbb{R} \right\} & \|w\|_2 = 1 \end{aligned}$$

führen wir folgenden Abstand ein:

$$d\left(T_v, T_w\right) := \min_{z \in T_w} \|z - v\|_2 = \min_{\alpha \in \mathbb{R}} \|\alpha w - v\|_2$$

TODO: Abb25
Setze $v = v_1$ und $w = \frac{x^{(k)}}{\|x^{(k)}\|}$. Wir wissen:

$$\frac{1}{\lambda_1^k \alpha_1} x^{(k)} = v_1 + \frac{1}{\alpha_1} r^{(k)}$$

$$d\left(T_{\frac{x^{(k)}}{\left\|x^{(k)}\right\|}}, T_{v_1}\right) \leq \left\|\frac{1}{\lambda_1^k \alpha_1} x^{(k)} - v_1\right\|_2 = \frac{1}{|\alpha_1|} \cdot \left\|r^{(k)}\right\|_2 = \mathcal{O}_{k\to\infty}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right) \xrightarrow{k\to\infty} 0$$

Zur Annäherung von λ_1 benutze:

$$\lambda^{(k)} = \frac{\left(x^{(k)}\right)^{\mathrm{T}} A x^{(k)}}{\left\|x^{(k)}\right\|_2^2} = \frac{\left(x^{(k)}\right)^{\mathrm{T}} x^{(k+1)}}{\left\|x^{(k)}\right\|_2^2}$$

Denn es gilt:

$$\frac{v_1^{\mathrm{T}} A v_1}{\|v_1\|_2^2} = \frac{\lambda_1 v_1^{\mathrm{T}} v_1}{\|v_1\|_2^2} = \lambda_1$$

Setze nun:

$$c_k := \frac{1}{\lambda_1^k \alpha_1}$$

Es gilt:

$$c_{k+1} = \frac{1}{\lambda_1^{k+1}\alpha_1} = \frac{1}{\lambda_1} c_k$$

Wir erhalten:

$$\lambda^{(k)} = \lambda_1 \frac{\left(c_k x^{(k)}\right)^{\mathrm{T}} \left(c_{k+1} x^{(k+1)}\right)}{\left\|c_k x^{(k)}\right\|_2^2} = \lambda_1 \frac{\left(v_1 + \frac{1}{\alpha_1} r^{(k)}\right)^{\mathrm{T}} \left(v_1 + \frac{1}{\alpha_1} r^{(k+1)}\right)}{\left\|v_1 + \frac{1}{\alpha_1} r^{(k)}\right\|_2^2} =$$

$$\underset{\text{Cauchy-Schwarz-Ug.}}{\overset{\|v_1\|_2=1}{=}} \lambda_1 \frac{1 + \mathcal{O}_{k\to\infty}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)}{1 + \mathcal{O}_{k\to\infty}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)} = \lambda_1\left(1 + \mathcal{O}_{k\to\infty}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right)\right)$$

Dies zeigt $\lambda^{(k)} \xrightarrow{k\to\infty} \lambda_1$.
Da $\left\|x^{(k)}\right\|_2 \to \infty$ divergiert, falls $|\lambda_1| > 1$ gilt, und $\left\|x^{(k)}\right\|_2 \to 0$ konvergiert, falls $|\lambda_1| < 1$ gilt, ist es zweckmäßig, die Iterierten zu normieren. Betrachte also $\frac{x^{(k)}}{\left\|x^{(k)}\right\|}$ statt $x^{(k)}$. (Skalierung)

9.5 **Algorithmus** (Vektoriteration, Potenzmethode)

Wähle zunächst einen Startvektor $y^{(0)}$ mit $\left\|y^{(0)}\right\|_2 = 1$. Für aufsteigendes $k \in \mathbb{N}_{\geq 0}$ berechne dann:

$$\begin{aligned}
\tilde{y}^{(k+1)} &= A y^{(k)} \\
\lambda^{(k)} &= \left(y^{(k)}\right)^{\mathrm{T}} \tilde{y}^{(k+1)} \\
y^{(k+1)} &= \frac{\tilde{y}^{(k+1)}}{\left\|\tilde{y}^{(k+1)}\right\|_2}
\end{aligned}$$

Bemerkung

Mit $x^{(0)} = y^{(0)}$ folgt per Induktion:

$$y^{(k)} = \frac{x^{(k)}}{\left\|x^{(k)}\right\|_2} = \frac{A^k x^{(0)}}{\left\|A^k x^{(0)}\right\|_2}$$

Der Induktionsanfang $k = 0$ ist klar.

$$y^{(k+1)} = \frac{Ay^{(k)}}{\left\|Ay^{(k)}\right\|} \overset{\text{Induktions-}}{\underset{\text{voraussetzung}}{=}} \frac{AA^k x^{(0)}}{\left\|AA^k x^{(0)}\right\|_2}$$

Also liefert der obige Algorithmus bis auf einen Faktor in $x^{(k)}$, die Folgen $\left(x^{(k)}\right)$ und $\left(\lambda^{(k)}\right)$.

9.6 Satz

Unter den Voraussetzungen in 9.4 approximiert die Vektoriteration in 9.5 den dominanten Eigenwert $\lambda_1 \in \mathbb{R}$ und einen zugehörigen normierten Eigenvektor $v_1 \in \mathbb{R}^n$. Es gilt

$$\left|\lambda^{(k)} - \lambda_1\right| = \mathcal{O}_{k\to\infty}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right) \xrightarrow{k\to\infty} 0$$

und:

$$\mathrm{d}\left(T_{y^{(k)}}, T_{v_1}\right) = \mathcal{O}_{k\to\infty}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^k\right) \xrightarrow{k\to\infty} 0$$

Beweis

Obige Überlegungen beweisen diesen Satz. $\square_{9.6}$

Bemerkung

i) Die Effizienz, das heißt die Konvergenzgeschwindigkeit, hängt von $\left|\frac{\lambda_2}{\lambda_1}\right|$ ab.

ii) Für eine mehrfache Nullstelle $\lambda_1 = \ldots = \lambda_m$ mit $|\lambda_m| > |\lambda_{m+1}|$ konvergiert das Verfahren weiterhin gegen λ_1.
Gilt $|\lambda_1| = |\lambda_2|$, aber $\lambda_1 \neq \lambda_2$, so erhält man im Allgemeinen keine Konvergenz.

iii) Nachteile des Verfahrens:

- Es kann nur der betragsmäßig größte Eigenwert berechnet werden.
- Falls $|\lambda_1| \approx |\lambda_2|$ gilt, ist die Konvergenz langsam.

9.7 Inverse Vektoriteration

Es sei A eine invertierbare Matrix mit Eigenwerten λ_i. Weiter sei v ein Eigenvektor zum Eigenwert λ. Für ein beliebiges $\tilde{\lambda} \in \mathbb{R}$ gilt:

$$\left(A - \tilde{\lambda}\mathbb{1}\right) v = \left(\lambda - \tilde{\lambda}\right) v$$

$$\Leftrightarrow \qquad \left(A - \tilde{\lambda}\mathbb{1}\right)^{-1} v = \frac{1}{\lambda - \tilde{\lambda}} v$$

Sei $\tilde{\lambda}$ ein guter Schätzwert von λ_i ist, das heißt für alle $j \neq i$ gilt:

$$\left|\tilde{\lambda} - \lambda_i\right| < \left|\tilde{\lambda} - \lambda_j\right|$$

Dann ist $\frac{1}{\lambda_i - \tilde{\lambda}}$ der betragsmäßig maximale Eigenwert von $\left(A - \tilde{\lambda}\mathbb{1}\right)^{-1}$ und hat den Eigenvektor v. Die Potenzmethode angewandt auf $\left(A - \tilde{\lambda}\mathbb{1}\right)^{-1}$ liefert:

$$\mu = \frac{1}{\lambda_i - \tilde{\lambda}}$$

Somit folgt:

$$\lambda_i = \frac{1}{\mu} - \tilde{\lambda}$$

Der Eigenvektor ist jeweils v.
Die reine inverse Vektoriteration verwendet $\tilde{\lambda} = 0$. Verwendet man $\tilde{\lambda} \neq 0$, so sagt man dazu *Spektralverschiebung.*

9.8 Inverse Vektoriteration mit Spektralverschiebung

Algorithmus:
Wähle $x^{(0)} \in \mathbb{R}^n \setminus \{0\}$ und setze:

$$v^{(0)} := \frac{x^{(0)}}{\left\|x^{(0)}\right\|_2}$$

Für aufsteigendes $k \in \mathbb{N}_{\geq 0}$ löse folgendes Gleichungssystem und berechne:

$$\left(A - \tilde{\lambda}\mathbb{1}\right) \tilde{x}^{(k+1)} = v^{(k)} \tag{9.1}$$

$$\lambda^{(k)} = \frac{1}{\left(v^{(k)}\right)^{\mathrm{T}} \tilde{x}^{(k+1)}} + \tilde{\lambda}$$

$$v^{(k+1)} = \frac{\tilde{x}^{(k+1)}}{\left\|\tilde{x}^{(k+1)}\right\|}$$

Bei der Implementierung muss man noch ein Abbruchkriterium hinzufügen.
In (9.1) muss je Iteration ein lineares Gleichungssystem gelöst werden. Der Aufwand ist also höher, als bei der Potenzmethode. Wird einmal eine LR- beziehungsweise eine QR-Zerlegung von $A - \tilde{\lambda}\mathbb{1}$ berechnet (Aufwand $\mathcal{O}_\infty\left(n^3\right)$), so braucht (9.1) noch eine Vorwärts- und eine Rückwärtssubstitution (Aufwand $\mathcal{O}_\infty\left(n^2\right)$). Als Schätzung für $\tilde{\lambda}$ können die Gerschgorin-Kreise benutzt werden.
Konvergenzgeschwindigkeit:

$$\frac{\left|\lambda_i - \tilde{\lambda}\right|}{\min_{j \neq i}\left|\lambda_j - \tilde{\lambda}\right|}$$

Dies folgt aus dem Resultat für die Potenzmethode.
Jetzt suchen wir eine Methode, die es erlaubt, *alle* Eigenwerte zu berechnen.

9.9 QR-Algorithmus

Es sei $A \in \mathbb{R}^{n\times n}$ mit QR-Zerlegung $A = QR$, das heißt Q orthogonal und R eine obere Dreiecksmatrix. Wir definiere:

$$A_0 := A$$

Für aufsteigendes $k \in \mathbb{N}_{\geq 0}$ berechne:

$$A_k = Q_k R_k$$
$$A_{k+1} = R_k Q_k$$

Dabei ist $Q_k R_k$ die QR-Zerlegung von A_k.

9.10 Lemma

Die Matrizen A_k haben folgende Eigenschaften:

i) Die Matrizen A_k sind alle konjugiert zu A, das heißt gehen durch eine Ähnlichkeitstransformation aus A hervor.

ii) Ist A symmetrisch, so sind alle A_k symmetrisch.

iii) Ist A symmetrisch und tridiagonal, so auch alle A_k.

Beweis

i) Sei $A = QR$ und $A' := RQ$. Dann gilt, da Q orthogonal ist, also $QQ^{\mathrm{T}} = \mathbb{1}$ gilt:

$$QA'Q^{\mathrm{T}} = QRQQ^{\mathrm{T}} = QR = A$$

ii) Es gilt:

$$\left(A'\right)^{\mathrm{T}} = \left(A'\right)^{\mathrm{T}} Q^{\mathrm{T}}Q = Q^{\mathrm{T}}R^{\mathrm{T}}Q^{\mathrm{T}}Q = Q^{\mathrm{T}}\left(QR\right)^{\mathrm{T}}Q = Q^{\mathrm{T}}A^{\mathrm{T}}Q = Q^{\mathrm{T}}AQ = A'$$

iii) Realisiere die QR-Zerlegung von A durch Givens-Rotationen.

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

Wähle $(n-1)$ Givens-Rotationen $G_{1,2}, \ldots, G_{n-1,n}$, sodass gilt:

$$G_{n-1,n} \cdot \ldots \cdot G_{2,3}G_{1,2}A = R$$
$$Q = G_{1,2}^{\mathrm{T}} \cdot \ldots \cdot G_{n-1,n}^{\mathrm{T}}$$

$$A = \begin{pmatrix} * & * & & & 0 \\ \otimes & * & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & * \\ 0 & & & \otimes & * \end{pmatrix} \xrightarrow[\text{Rotationen an}]{\text{wende Givens-}} \begin{pmatrix} * & * & * & & 0 \\ & * & \ddots & \ddots & \\ & & \ddots & \ddots & * \\ & & & \ddots & * \\ 0 & & & & * \end{pmatrix} = R$$

$$\longrightarrow \begin{pmatrix} * & * & \oplus & & 0 \\ * & * & \ddots & \ddots & \\ & \ddots & \ddots & \ddots & \oplus \\ & & \ddots & \ddots & * \\ 0 & & & * & * \end{pmatrix} = RQ$$

$\otimes$ sind die zu eliminierenden Positionen und $\oplus$ sind neu erzeugte „fill in“ Elemente.
Die Spalten $(1,2), \ldots, (n-1,n)$ werden nacheinander linear kombiniert.
Wegen

$$A' = RQ = Q^{\mathrm{T}} AQ = \left(Q^{\mathrm{T}} AQ\right)^{\mathrm{T}}$$

ist A' symmetrisch und somit verschwinden alle $\oplus$. Dies zeigt die Behauptung.

$\square_{9.10}$

9.11 **Definition** (Hessenbergform)

Eine Matrix $B \in \mathbb{R}^{n \times n}$ hat *(obere) Hessenbergform*, wenn $b_{ij} = 0$ für $i > j + 1$ gilt, das heißt:

$$B = \begin{pmatrix} * & \cdots & \cdots & * \\ * & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ 0 & & * & * \end{pmatrix}$$

Eine solche Matrix nennt man *Hessenbergmatrix*.

9.12 Satz

Zu jeder Matrix $A \in \mathbb{R}^{n \times n}$ existieren $(n-2)$ Householder-Matrizen H_j für $j \in \{1, \ldots, n-2\}$ mit

$$Q = H_{n-2} \cdot \ldots \cdot H_1$$

für die

$$B = QAQ^{\mathrm{T}}$$

Hessenberg-Gestalt hat.

Beweis

Schreibe die Matrix in der Form

$$A = \begin{pmatrix} a_{11} & * \\ \tilde{a_1} & \tilde{A} \end{pmatrix}$$

mit $\tilde{A} \in \mathbb{R}^{(n-1)\times(n-1)}$ und $\tilde{a}_1 \in \mathbb{R}^{n-1}$. Wähle eine Householder-Matrix $\tilde{H}_1 \in \mathbb{R}^{(n-1)\times(n-1)}$ mit:

$$\tilde{H}_1 \tilde{a}_1 = \|\tilde{a}_1\|_2 \cdot e_1$$

$e_1 \in \mathbb{R}^{n-1}$ ist der erste Einheitsvektor. Mit

$$H_1 = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \tilde{H}_1 & \\ 0 & & & \end{pmatrix}$$

erhalten wir:

$$\begin{aligned} H_1 A H_1^{\mathrm{T}} &= \begin{pmatrix} 1 & 0 \\ 0 & \tilde{H}_1 \end{pmatrix} \begin{pmatrix} a_{11} & * \\ \tilde{a}_1 & \tilde{A} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \tilde{H}_1^{\mathrm{T}} \end{pmatrix} = \\ &= \begin{pmatrix} 1 & 0 \\ 0 & \tilde{H}_1 \end{pmatrix} \begin{pmatrix} a_{11} & * \\ \tilde{a}_1 & \tilde{A}\tilde{H}_1^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} a_{11} & * \\ \tilde{H}_1\tilde{a}_1 & \tilde{H}_1\tilde{A}\tilde{H}_1^{\mathrm{T}} \end{pmatrix} = \\ &= \begin{pmatrix} a_{11} & * \\ \|\tilde{a}_1\|_2\, e_1 & \tilde{H}_1\tilde{A}\tilde{H}_1^{\mathrm{T}} \end{pmatrix} \end{aligned}$$

Damit hat die erste Spalte die gewünschte Form. Wende nun dieses Verfahren rekursiv auf $\tilde{H}_1 A \tilde{H}_1^{\mathrm{T}}$ etc. an. $\square_{9.12}$

Bemerkung

i) Eine symmetrische Hessenberg-Matrix hat Tridiagonalform.

ii) Die Überlegungen zum Beweis von Lemma 9.10 iii) liefern, dass der QR-Algorithmus die Hessenbergform erhält.

Jetzt zur *Konvergenz* des QR-Algorithmus. (Die Eigenwerte sind paarweise verschieden!)

9.13 Satz

Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch mit den Eigenwerten $\lambda_1, \dots, \lambda_n$ mit

$$|\lambda_1| > |\lambda_2| > \dots > |\lambda_n| > 0$$

und es gelte $A = T^{-1}\Lambda T$ mit

$$\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

und einer Matrix T, die eine LR-Zerlegung besitzt. Dann gilt mit:

$$A_k = \left(a_{ij}^{(k)}\right)$$

a) $\lim\limits_{k\to\infty} Q_k = \mathbb{1}$

b) $\lim\limits_{k\to\infty} R_k = \Lambda$

c) Es gilt $a_{ij}^{(k)} = o_{k\to\infty}\left(\left|\frac{\lambda_i}{\lambda_j}\right|^k\right)$ für $i > j$.

Beweis

Behauptung: Für $k \in \mathbb{N}_{\geq 1}$ gilt:

$$A^k = \underbrace{Q_1 \cdot \ldots \cdot Q_k}_{=:P_k} \cdot \underbrace{R_k \cdot \ldots \cdot R_1}_{=:U_k}$$

Beweis: Für $k = 1$ gilt dies nach Voraussetzung.
Induktionsschritt $k \rightsquigarrow k+1$:

$$Q_{k+1}R_{k+1} = A_{k+1} = Q_k^{\mathrm{T}} \cdot \ldots \cdot Q_1^{\mathrm{T}} \cdot A \cdot Q_1 \cdot \ldots \cdot Q_k = P_k^{-1} A P_k$$

Damit folgt:

$$A^{k+1} = A \cdot A^k \overset{\text{Induktions-}}{\underset{\text{voraussetzung}}{=}} A \cdot P_k \cdot U_k = P_k \cdot Q_{k+1} \cdot R_{k+1} \cdot U_k = P_{k+1} \cdot U_{k+1}$$

$\square_{\text{Behauptung}}$

Es folgt:

$$A^k = T^{-1} \Lambda^k T$$

$$\Lambda^k = \begin{pmatrix} \lambda_1^k & & 0 \\ & \ddots & \\ 0 & & \lambda_n^k \end{pmatrix}$$

Mit $A = T^{-1}\Lambda T$ und $T = LR$ folgt:

$$A^k = T^{-1}\Lambda LR = T^{-1}\left(\Lambda^k L \Lambda^{-k}\right)\left(\Lambda^k R\right)$$

Es gilt $l_{ii} = 1$ und:

$$\left(\Lambda^k L \Lambda^{-k}\right)_{ij} = \left(l_{ij}\left(\frac{\lambda_i}{\lambda_j}\right)^k\right)$$

Für $k \to \infty$ ergibt sich

$$\Lambda^k L \Lambda^{-k} = \mathbb{1} + E_k$$

mit $E_k \xrightarrow{k\to\infty} 0$. Somit gilt:

$$A^k = T^{-1}\left(\mathbb{1} + E_k\right)\Lambda^k R$$

Für $\mathbb{1} + E_k$ mache eine QR-Zerlegung:

$$\mathbb{1} + E_k = \tilde{Q}_k \tilde{R}_k$$

Dabei sollen alle Diagonalelemente von $\tilde{R}_k$ positiv sein.

Beobachtung: Die QR-Zerlegung mit positiven Einträgen auf Diagonaleinträgen von R ist eindeutig.

Die Einträge von Q und R hängen stetig von den Einträgen der Matrix ab, wie man im Berechnungsverfahren sieht. Es gilt:

$$A^k = \underbrace{T^{-1}\tilde{Q}_k}_{\text{orthogonal}} \cdot \underbrace{\tilde{R}_k \Lambda^k R}_{\text{obere Dreiecksmatix}}$$

Wähle T orthogonal, sodass in den Spalten eine Orthonormalbasis aus den Eigenvektoren. Die Eindeutigkeit der QR-Zerlegung liefert aber bis auf Vorzeichen auf der Diagonale von R.

$$P_k = T^{-1}\tilde{Q}_k$$
$$U_k = \tilde{R}_k \Lambda^k R$$

Für $k \to \infty$ folgt:

$$Q_k = \underbrace{P_{k-1}^{\mathrm{T}}}_{=P_{k-1}^{-1}} P_k = \tilde{Q}_{k-1}^{\mathrm{T}} \cdot T \cdot T^{-1}\tilde{Q}_k = \tilde{Q}_{k-1}^{\mathrm{T}}\tilde{Q}_k \xrightarrow{k\to\infty} \mathbb{1}$$

Dies folgt wegen:

$$\tilde{Q}_k\tilde{R}_k = \mathbb{1} + E_k \to \mathbb{1} = \underbrace{\mathbb{1}\cdot\mathbb{1}}_{QR\text{-Zerlegung von } \mathbb{1}}$$

$$\Rightarrow \qquad \tilde{Q}_k\tilde{R}_k \to \mathbb{1}$$

$$R_k = U_k \cdot U_{k-1}^{-1} = \tilde{R}_k\Lambda^k R \cdot R^{-1}\Lambda^{-(k-1)}\tilde{R}_{k-1}^{-1} = \tilde{R}_k \Lambda \tilde{R}_{k-1}^{-1} \xrightarrow{k\to\infty} \Lambda$$

Es gilt:

$$\lim_{k\to\infty} A_k = \lim_{k\to\infty} Q_k R_k = \lim_{k\to\infty} R_k = \Lambda$$

$\square_{9.13}$

Bemerkung

i) Eine genauere Analyse zeigt, dass dieses Verfahren in der Tat auch für mehrfache Eigenwerte $\lambda_1 = \ldots = \lambda_j \neq 0$ konvergiert. (Falls $\lambda_i = -\lambda_k$ gilt, so konvergiert das Verfahren für die anderen Eigenwerte.) Für $\lambda_i = -\lambda_{i+1}$ konvergiert das Verfahren nicht.

ii) In der Praxis geht man wie folgt:

   I) Bringe die Matrix auf die Hessenberg-Form.

   II) Wende den QR-Algorithmus an. (Falls es Probleme mit der Konvergenz gibt, ersetze A durch $A - \mu\mathbb{1}$.)
   Ist A symmetrisch, so ist die Hessenberg-Form tridiagonal, was einen geringen Aufwand im QR-Verfahren bewirkt.

10 Das Verfahren der konjugierten Gradienten (*cg*-Verfahren)

Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit, sowie $b \in \mathbb{R}^n$. Gesucht ist eine Lösung $x \in \mathbb{R}^n$ von $Ax = b$.

Idee: Nähere x sukzessive in Unterräumen möglichst genau an.

Wähle einen Abstandsbegriff, der von A abhängt. Definiere das Skalarprodukt wie folgt:

$$\langle y, z\rangle_A := y \cdot Az = \sum_{i,j} a_{ij} y_i z_j \qquad \underset{y,z\in\mathbb{R}^n}{\forall}$$

Die Norm

$$\|y\|_A = \sqrt{\langle y, y\rangle_A}$$

heißt *Energienorm*.

Sei $x_0 \in \mathbb{R}^n$ und $U_1 \subseteq U_2 \subseteq \ldots \subseteq U_n = \mathbb{R}^n$ eine Folge linearer Unterräumen mit $\dim U_k = k$. Gesucht ist $x_k \in x_0 + U_k = \{x_0 + u \,|\, u \in U_k\}$, sodass gilt:

$$\|x_k - x\|_A = \min_{y\in x_0+U_k} \|y - x\|_A$$

Das heißt, wir suchen die beste Approximation im affinen Unterraum $x_0 + U_k$. Hierbei sei $\dim(x_0 + U_k) = k$.

Der Projektionssatz liefert:

$$\langle x_k - x, y\rangle = 0 \qquad \underset{y\in U_k}{\forall}$$

Sei nun $\{p_1, \ldots, p_k\}$ eine bezüglich des $\langle ., .\rangle_A$-Skalarprodukts orthogonale Basis von U_k. Es gilt:

$$\langle p_i, p_j\rangle_A = \delta_{ij} \langle p_j, p_j\rangle_A$$

Definiere das Residuum:

$$r_k = b - Ax_k = A(x - x_k)$$

Jetzt gilt (vergleiche Projektionssatz 4.12):

$$x_k = x_0 + \sum_{j=1}^{k} \frac{\langle p_j, x - x_0\rangle_A}{\langle p_j, p_j\rangle_A} p_j = x_0 + \sum_{j=1}^{k} \frac{p_j \cdot A(x - x_0)}{\langle p_j, p_j\rangle_A} p_j =$$

$$= x_0 + \sum_{j=1}^{k} \frac{p_j \cdot r_0}{\langle p_j, p_j\rangle_A} p_j \tag{10.1}$$

Bemerkung: Die Unbekannte x taucht rechts nicht mehr auf.
Frage: Wie wähle ich die Unterräume U_k?
Wähle einen Krylov-Raum ($k \in \{1, 2, \ldots, n\}$):

$$\begin{aligned} U_0 &= \{0\} \\ U_k &= \operatorname{span}\left\{r_0, Ar_0, \ldots, A^{k-1}r_0\right\} \end{aligned}$$

Definiere:

$$\alpha_k := \frac{\langle p_k, x - x_0\rangle_A}{\langle p_k, p_k\rangle_A}$$

Aus (10.1) folgt:

$$\begin{aligned} x_k &= x_{k-1} + \alpha_k p_k \\ r_k &= b - Ax_k = b - Ax_{k-1} - A\left(x_k - x_{k-1}\right) = \\ &= r_{k-1} - \alpha_k A p_k \end{aligned}$$

Es folgt durch iterierte Anwendung:

$$\begin{aligned} r_k \cdot p_l &= r_{l-1} \cdot p_l - \sum_{j=l}^{k} \alpha_j \langle p_j, p_l\rangle_A = \\ &= \langle x - x_{l-1}, p_l\rangle_A - \alpha_l \langle p_l, p_l\rangle_A = \langle x - x_0, p_l\rangle_A - \alpha_l \langle p_l, p_l\rangle_A \end{aligned}$$

Die letzte Zeile gilt wegen:

$$\begin{aligned} & x_{l-1} = x_0 + \sum_{j=1}^{l-1} \alpha_j p_j \\ \Rightarrow \qquad & \langle x_{l-1}, p_l\rangle_A = \langle x_0, p_l\rangle_A \end{aligned} \tag{10.2}$$

Insgesamt folgt:

$$r_k \cdot p_l = \langle x - x_0, p_l\rangle_A - \alpha_l \langle p_l, p_k\rangle_A \overset{\text{Definition}}{=} \alpha_l \langle p_l, p_l\rangle_A - \alpha_l \langle p_l, p_l\rangle_A = 0 \tag{10.3}$$

Jetzt folgt ein Lemma, dass uns eine einfache Berechnungsvorschrift liefert.

10.1 Lemma

Sei $x_0 \in \mathbb{R}^n$, $\{p_1, \ldots, p_k\}$ sei eine A-orthogonale Basis von $U_k = \operatorname{span}\left\{r_0, Ar_0, \ldots, A^{k-1}r_0\right\}$ mit $p_1 = r_0$. Außerdem seien α_k, r_k wie oben definiert und $r_k \neq 0$.
Dann sind die Residuen $r_0, \ldots, r_k$ paarweise orthogonal, das heißt für $i, j \in \{0, \ldots, k\}$ gilt:

$$r_i \cdot r_j = \delta_{ij} r_i \cdot r_i$$

Außerdem spannen sie U_{k+1} auf, das heißt $U_k = \operatorname{span}\{r_0, r_1, \ldots, r_k\}$, und für $l \in \{1, \ldots, k\}$ gilt:

$$r_k \cdot p_l = 0$$

Beweis

Wir zeigen nur die letzte Aussage, da der Rest sich von oben ergibt. Führe hierzu eine vollständige Induktion über k durch.
Induktionsanfang bei $k = 0$: Ist klar, da es kein $l \in \{1, \ldots, 0\} = \emptyset$ gibt.
Induktionsschritt $k - 1 \rightsquigarrow k$: Es gilt:

$$r_k - r_{l-1} = -A\left(x_k - x_{l-1}\right) = -A\sum_{j=l}^{k} \alpha_j p_j$$

$$\begin{aligned} r_k \cdot p_l &= r_{l-1} \cdot p_l - \sum_{j=l}^{k} \alpha_j \left\langle p_j, p_l \right\rangle_A = \left\langle x - x_{l-1}, p_l \right\rangle_A = \\ &= \left\langle x - x_{l-1}, p_l \right\rangle_A - \alpha_l \left\langle p_l, p_l \right\rangle_A \overset{(10.2)}{=} \left\langle x - x_0, p_l \right\rangle_A - \alpha_l \left\langle p_l, p_l \right\rangle_A \end{aligned}$$

Wie oben folgt:

$$r_k \cdot p_l \overset{(10.3)}{=} 0$$

$\square_{10.1}$

Aus dem Lemma folgt:
Entweder gilt

$$r_k = 0 \qquad \Rightarrow \qquad b - Ax_k = 0 \qquad \Rightarrow \qquad x = x_k \quad \text{fertig}$$

oder die $p_1, \ldots, p_k, r_k$ sind linear unabhängig und spannen U_{k+1} auf.
Wir wählen:

$$p_{k+1} = r_k - \sum_{j=1}^{k} \frac{\left\langle r_k, p_j \right\rangle_A}{\left\langle p_j, p_j \right\rangle_A} p_j \overset{Ap_j = \frac{1}{\alpha_j}(r_j - r_{j-1})}{=} r_k - \frac{\left\langle r_k, p_k \right\rangle_A}{\left\langle p_k, p_k \right\rangle_A} p_k$$

$$\begin{aligned} x_k &= x_{k-1} + \alpha_k p_k \\ p_{k+1} &= r_k + \beta_{k+1} p_k \\ \beta_{k+1} &= -\frac{\left\langle r_k, p_k \right\rangle_A}{\left\langle p_k, p_k \right\rangle_A} \end{aligned}$$

Es gilt:

$$\begin{aligned} \alpha_k &= \frac{\left\langle x - x_0, p_k \right\rangle_A}{\left\langle p_k, p_k \right\rangle_A} = \frac{\left\langle x - x_{k-1}, p_k \right\rangle_A}{\left\langle p_k, p_k \right\rangle_A} = \frac{r_{k-1} \cdot p_k}{\left\langle p_k, p_k \right\rangle_A} = \frac{r_{k-1} \cdot r_{k-1}}{\left\langle p_k, p_k \right\rangle_A} \\ \beta_{k+1} &= -\frac{\left\langle r_k, p_k \right\rangle_A}{\left\langle p_k, p_k \right\rangle_A} = -\alpha_k \cdot \frac{\left\langle r_k, p_k \right\rangle_A}{r_{k-1} \cdot r_{k-1}} = \frac{r_k \cdot (-\alpha_k A p_k)}{r_{k-1} \cdot r_{k-1}} = \\ &= \frac{r_k \cdot (r_k - r_{k-1})}{r_{k-1} \cdot r_{k-1}} = \frac{r_k \cdot r_k}{r_{k-1} \cdot r_{k-1}} \end{aligned}$$

10.2 *cg*-Algorithmus

Wähle Startwert x_0 und setzte $r_0 := b - Ax_0$ und $p_1 = r_0$. Für $k \in \mathbb{N}_{\geq 1}$ definiere:

$$\alpha_k = \frac{r_{k-1} \cdot r_{k-1}}{p_k \cdot Ap_k} \qquad x_k = x_{k-1} + \alpha_k p_k \qquad r_k = r_{k-1} - \alpha A p_k$$

Falls $\|r_k\|_2$ klein genug ist, stoppe den Algorithmus.

$$\beta_{k+1} = \frac{r_k \cdot r_k}{r_{k-1} \cdot r_{k-1}} \qquad p_{k+1} = r_k + \beta_{k+1} p_k$$

Bemerkung

i) Theoretisch hat man nach n Schritten die exakte Lösung. In der Praxis hört man oft schon vorher auf. Nach n Schritten ist die Lösung oft noch durch Rundungsfehler gestört.

ii) Konvergenzgeschwindigkeit, ist gegeben durch folgende Abschätzung:

$$\|x_k - x\|_A \leq 2 \left(\frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1} \right)^k \|x_0 - x\|_A$$

Dabei ist $\kappa_2(A)$ die euklidische Kondition. Die Konvergenz ist schnell, falls $\kappa_2(A)$ nahe bei 1 ist. In der Praxis ist $\kappa_2(A)$ sehr groß.
Ausweg: Löse statt $Ax = b$ die Gleichung

$$W^{-1}Ax = W^{-1}b$$

mit einer einfach zu invertierenden Matrix W, sodass gilt:

$$\kappa_2\left(W^{-1}A\right) \ll \kappa_2(A)$$

TODO: Abb Effizienz

Verfahren	Anzahl der Operationen	Zeit ($N = 3 \cdot 10^4$, $t_{\text{flop}} = 56\,\mu\text{s}$)
Cramersche Regel	$\sim N!$	$\approx 10^{10^5}$ Jahre = „∞“
Gaußsches Eliminationsverfahren für Bandmatrizen	$\sim N^2$	14 h
Überrelaxationsverfahren, SOR (1960)	$\sim N^{1,5}$	4 min 51 s
Mehrgitter-Verfahren (1980)	$\sim N$	1,68 s

Beispiele von Anwendungen numerischer Verfahren:

1. Wetterprognose
2. Klimaprognose
3. Numerischer Windkanal

TODO: Folien

Schnelle lineare Löser in der Wettervorhersage:
Heutzutage:

- Maschenweite von ca. 40 km, ergibt ca. 400000 Gitterpunkte
- In der Höhe 40 Schichten.
- Insgesamt ca. 16 Millionen Gitterpunkte

Zu lösen ist in jedem Zeitschritt ein Gleichungssystem mit 16 Millionen Unbekannten.
Es werden für 10 Tage-Vorhersage ca. 6500 Zeitschritte der Länge ca. 2 Minuten verwendet.
Zukunft:

- Maschenweite von ca. 20 km, ergibt ca. 2 Millionen Gitterpunkte
- In der Höhe 100 Schichten.
- Insgesamt ca. 200 Millionen Gitterpunkte

Anhang

Danksagungen

Mein besonderer Dank geht an Professor Garcke, der diese Vorlesung hielt und es mir gestattete, diese Vorlesungsmitschrift zu veröffentlichen.

Außerdem möchte ich mich ganz herzlich bei allen bedanken, die durch aufmerksames Lesen Fehler gefunden und mir diese mitgeteilt haben.

Andreas Völklein

GNU Free Documentation License

Version 1.3, 3 November 2008
Copyright © 2000, 2001, 2002, 2007, 2008 Free Software Foundation, Inc.

<https://fsf.org/>

Everyone is permitted to copy and distribute verbatim copies of this license document, but changing it is not allowed

0. Preamble

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. Applicability and definitions

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "**Document**", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "**you**". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "**Modified Version**" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "**Secondary Section**" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "**Invariant Sections**" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "**Cover Texts**" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "**Transparent**" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "**Opaque**".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "**Title Page**" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

The "**publisher**" means any person or entity that distributes copies of the Document to the public.

A section "**Entitled XYZ**" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "**Acknowledgements**", "**Dedications**", "**Endorsements**", or "**History**".) To "**Preserve the Title**" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that

these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. Verbatim copying

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. Copying in quantity

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. Modifications

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution

and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.

C. State on the Title page the name of the publisher of the Modified Version, as the publisher.

D. Preserve all the copyright notices of the Document.

E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.

F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.

G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.

H. Include an unaltered copy of this License.

I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.

J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.

K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.

L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.

M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.

N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.

O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties—for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. Combining documents

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

6. Collections of documents

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. Aggregation with independent works

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate"

if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. Translation

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. Termination

You may not copy, modify, sublicense, or distribute the Document except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, or distribute it is void, and will automatically terminate your rights under this License.

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, receipt of a copy of some or all of the same material does not give you any rights to use it.

10. Future revisions of this license

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the

present version, but may differ in detail to address new problems or concerns. See `https://www.gnu.org/copyleft/`.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation. If the Document specifies that a proxy can decide which future versions of this License can be used, that proxy's public statement of acceptance of a version permanently authorizes you to choose that version for the Document.

11. Relicensing

"Massive Multiauthor Collaboration Site" (or "MMC Site") means any World Wide Web server that publishes copyrightable works and also provides prominent facilities for anybody to edit those works. A public wiki that anybody can edit is an example of such a server. A "Massive Multiauthor Collaboration" (or "MMC") contained in the site means any set of copyrightable works thus published on the MMC site.

"CC-BY-SA" means the Creative Commons Attribution-Share Alike 3.0 license published by Creative Commons Corporation, a not-for-profit corporation with a principal place of business in San Francisco, California, as well as future copyleft versions of that license published by that same organization.

"Incorporate" means to publish or republish a Document, in whole or in part, as part of another Document.

An MMC is "eligible for relicensing" if it is licensed under this License, and if all works that were first published under this License somewhere other than this MMC, and subsequently incorporated in whole or in part into the MMC, (1) had no cover texts or invariant sections, and (2) were thus incorporated prior to November 1, 2008.

The operator of an MMC Site may republish an MMC contained in the site under CC-BY-SA on the same site at any time before August 1, 2009, provided the MMC is eligible for relicensing.

Addendum: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

> Copyright © YEAR YOUR NAME.
> Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation;
>
> with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts.
>
> A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with ... Texts." line with this:

with the Invariant Sections being LIST THEIR TITLES, with the Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.